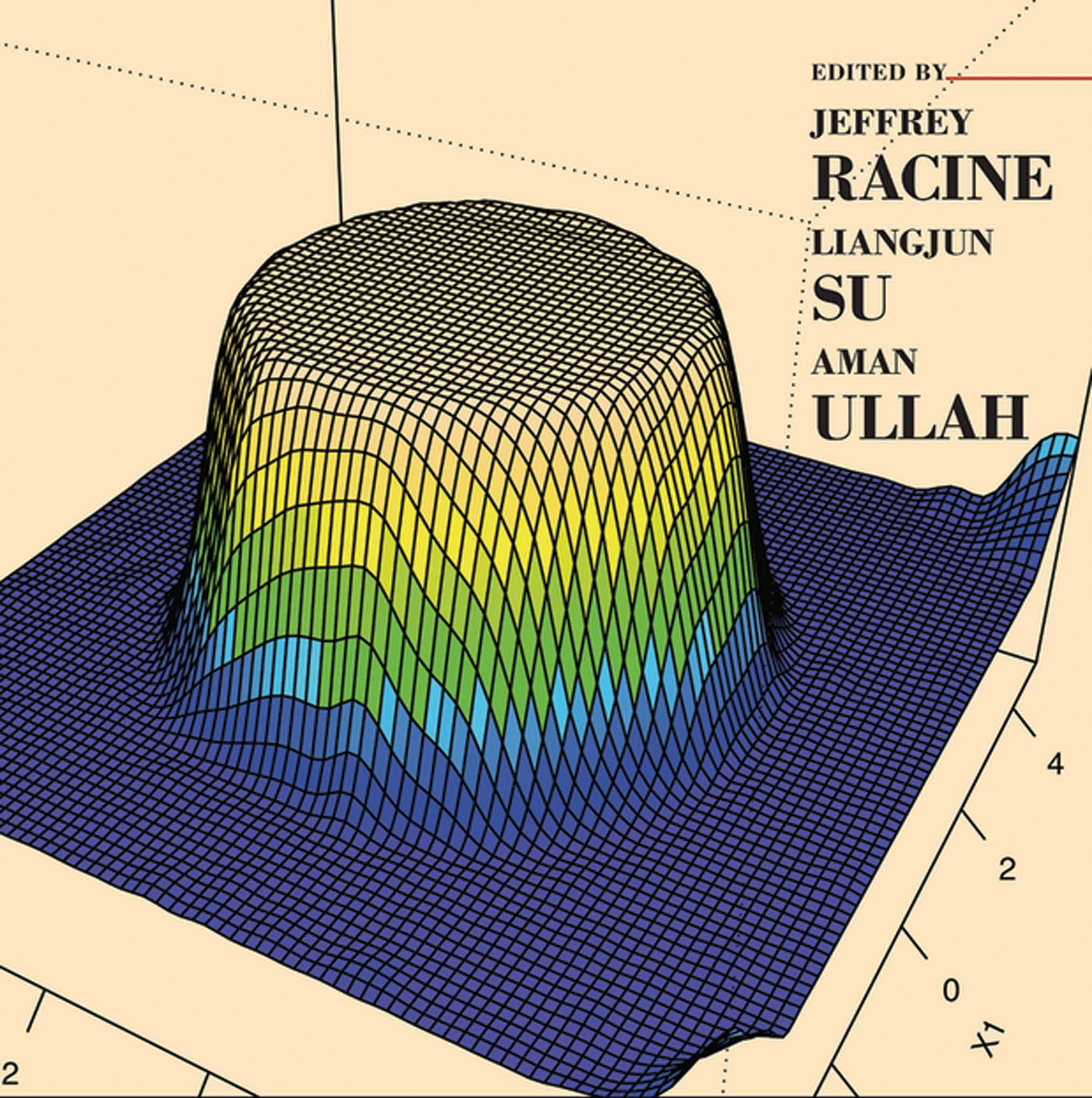
EDITED BY

JEFFREY
RACINE

LIANGJUN
SU

AMAN
ULLAH



The Oxford Handbook *of* APPLIED NONPARAMETRIC AND SEMIPARAMETRIC ECONOMETRICS AND STATISTICS

THE OXFORD HANDBOOK OF

APPLIED NONPARAMETRIC AND SEMIPARAMETRIC ECONOMETRICS AND STATISTICS

CONSULTING EDITORS

THE OXFORD HANDBOOK OF

APPLIED NONPARAMETRIC AND SEMIPARAMETRIC ECONOMETRICS AND STATISTICS

Edited by

JEFFREY S. RACINE, LIANGJUN SU,

and

AMAN ULLAH

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics / edited by
Jeffrey Racine, Liangjun Su, and Aman Ullah.
pages cm
Includes bibliographical references and index.
ISBN 978–0–19–985794–4
1. Econometrics. 2. Nonparametric statistics. I. Racine, Jeffrey Scott, 1962– editor of compilation.
II. Su, Liangjun, editor of compilation. III. Ullah, Aman, editor of compilation.
HB139.O96 2013
330.01′5195—dc23
2013002107

1 3 5 7 9 8 6 4 2
Printed in the United States of America
on acid-free paper

Contents

PART I METHODOLOGY

PART II INVERSE PROBLEMS

PART III ADDITIVE MODELS

PART IV MODEL SELECTION AND AVERAGING

PART V TIME SERIES

PART VI CROSS SECTION

Preface

Since the birth of Econometrics almost eight decades ago, theoretical and applied Econometrics and Statistics has, for the most part, proceeded along 'Classical lines which typically invokes the use of rigid user-specified parametric models, often linear. However, during the past three decades a growing awareness has emerged that results based on poorly specified parametric models could lead to misleading policy and forecasting results. In light of this, around three decades ago the subject of nonparametric Econometrics and nonparametric Statistics emerged as a field with the defining feature that models can be 'data-driven'—hence tailored to the data set at hand. Many of these approaches are described in the books by Prakasa Rao (1983), Härdle (1990), Fan and Gijbels (1996), Pagan and Ullah (1999), Yatchew (2003), Li and Racine (2007), and Horowitz (2009), and they appear in a wide range of journal outlets. The recognition of the importance of this subject along with advances in computer technology has fueled research in this area, and the literature continues to increase at an exponential rate. This pace of innovation makes it difficult for specialists and nonspecialists alike to keep abreast of recent developments. There is no single source available for those seeking an informed overview of these developments.

This handbook contains chapters that cover recent advances and major themes in the nonparametric and semiparametric domain. The chapters contained herein provide an up-to-date reference source for students and researchers who require definitive discussions of the cutting-edge developments in applied Econometrics and Statistics. Contributors have been chosen on the basis of their expertise, their international reputation, and their experience in exposing new and technical material. This handbook highlights the interface between econometric and statistical methods for nonparametric and semiparametric procedures; it is comprised of new, previously unpublished research papers/chapters by leading international econometricians and statisticians. This handbook provides a balanced viewpoint of recent developments in applied sciences with chapters covering advances in methodology, inverse problems, additive models, model selection and averaging, time series, and cross-section analysis.

Methodology

Semi-nonparametric (SNP) models are models where only a part of the model is parameterized, and the nonspecified part is an unknown function that is represented by an infinite series expansion. SNP models are, in essence, models with infinitely many parameters. In Chapter 1, Herman J. Bierens shows how orthonormal functions can be constructed along with how to construct general series representations of density and distribution functions in a SNP framework. Bierens reviews the necessary Hilbert space theory involved as well.

The term 'special regressor' originates in Lewbel (1998) and has been employed in a wide variety of limited dependent variable models including binary, ordered, and multinomial choice as well as censored regression, selection, and treatment models and truncated regression models, among others (a special regressor is an observed covariate with properties that facilitate identification and estimation of a latent variable model). In Chapter 2, Arthur Lewbel provides necessary background for understanding how and why special regressor methods work, and he details their application to identification and estimation of latent variable moments and parameters.

Inverse Problems

Ill-posed problems surface in a range of econometric models (a problem is 'well-posed' if its solution exists, is unique, and is stable, while it is 'ill-posed' if any of these conditions are violated). In Chapter 3, Marine Carrasco, Jean-Pierre Florens and Eric Renault study the estimation of a function φ in linear inverse problems of the form $T\varphi = r$, where r is only observed with error and T may be given or estimated. Four examples are relevant for Econometrics, namely, (i) density estimation, (ii) deconvolution problems, (iii) linear regression with an infinite number of possibly endogenous explanatory variables, and (iv) nonparametric instrumental variables estimation. In the first two cases T is given, whereas it is estimated in the two other cases, respectively at a parametric or nonparametric rate. This chapter reviews some main results for these models such as concepts of degree of ill-posedness, regularity of φ, regularized estimation, and the rates of convergence typically obtained. Asymptotic normality results of the regularized solution $\hat{\varphi}_\alpha$ are obtained and can be used to construct (asymptotic) tests on φ.

In Chapter 4, Victoria Zinde-Walsh provides a nonparametric analysis for several classes of models, with cases such as classical measurement error, regression with errors in variables, and other models that may be represented in a form involving convolution equations. The focus here is on conditions for existence of solutions, nonparametric identification, and well-posedness in the space of generalized functions (tempered distributions). This space provides advantages over working in function spaces by relaxing assumptions and extending the results to include a wider variety of models, for example by not requiring existence of and underlying density. Classes of (generalized) functions for which solutions exist are defined; identification conditions, partial identification, and its implications are discussed. Conditions for well-posedness are given, and the related issues of plug-in estimation and regularization are examined.

Additive Models

Additive semiparametric models are frequently adopted in applied settings to mitigate the curse of dimensionality. They have proven to be extremely popular and tend to be simpler to interpret than fully nonparametric models. In Chapter 5, Joel L. Horowitz considers estimation of nonparametric additive models. The author describes methods for estimating standard additive models along with additive models with a known or unknown link function. Tests of additivity are reviewed along with an empirical example that illustrates the use of additive models in practice.

In Chapter 6, Shujie Ma and Lijian Yang present an overview of additive regression where the models are fit by spline-backfitted kernel smoothing (SBK), and they focus on improvements relative to existing methods (i.e., Linton (1997)). The SBK estimation method has several advantages compared to most existing methods. First, as pointed out in Sperlich et al. (2002), the estimator of Linton (1997) mixed up different projections, making it uninterpretable if the real data generating process deviates from additivity, while the projections in both steps of the SBK estimator are with respect to the same measure. Second, the SBK method is computationally expedient, since the pilot spline estimator is much faster computationally than the pilot kernel estimator proposed in Linton (1997). Third, the SBK estimator is shown to be as efficient as the "oracle smoother" uniformly over any compact range, whereas Linton (1997) proved such 'oracle efficiency' only at a single point. Moreover, the regularity conditions needed by the SBK estimation procedure are natural and appealing and close to being minimal. In contrast, higher-order smoothness is needed with growing dimensionality of the regressors in Linton and Nielsen (1995). Stronger and more obscure conditions are assumed for the two-stage estimation proposed by Horowitz and Mammen (2004).

In Chapter 7, Enno Mammen, Byeong U. Park and Melanie Schienle give an overview of smooth backfitting estimators in additive models. They illustrate their wide applicability in models closely related to additive models such as (i) nonparametric regression with dependent errors where the errors can be transformed to white noise by a linear transformation, (ii) nonparametric regression with repeatedly measured data, (iii) nonparametric panels with fixed effects, (iv) simultaneous nonparametric equation models, and (v) non- and semiparametric autoregression and GARCH-models. They review extensions to varying coefficient models, additive models with missing observations, and the case of nonstationary covariates.

Model Selection and Averaging

"Sieve estimators" are a class of nonparametric estimator where model complexity increases with the sample size. In Chapter 8, Bruce Hansen considers "model selection" and "model averaging" of nonparametric sieve regression estimators. The concepts of series and sieve approximations are reviewed along with least squares estimates of sieve approximations and measurement of estimator accuracy by integrated mean-squared error (IMSE). The author demonstrates that the critical issue in applications is selection of the order of the sieve, because the IMSE greatly varies across the choice. The author adopts the cross-validation criterion as an estimator of mean-squared forecast error and IMSE. The author extends existing optimality theory by showing that cross-validation selection is asymptotically IMSE equivalent to the infeasible best sieve approximation, introduces weighted averages of sieve regression estimators, and demonstrates how averaging estimators have lower IMSE than selection estimators.

In Chapter 9, Liangjun Su and Yonghui Zhang review the literature on variable selection in nonparametric and semiparametric regression models via shrinkage. The survey includes simultaneous variable selection and estimation through the methods of least absolute shrinkage and selection operator (Lasso), smoothly clipped absolute deviation (SCAD), or their variants, with attention restricted to nonparametric and semiparametric regression models. In particular, the author considers variable selection in additive models, partially linear models, functional/varying coefficient models, single index models, general nonparametric regression models, and semiparametric/nonparametric quantile regression models.

In Chapter 10, Jeffrey S. Racine and Christopher F. Parmeter propose a data-driven approach for testing whether or not two competing approximate models are equivalent in terms of their expected true error (i.e., their expected performance on unseen data drawn from the same DGP). The test they consider is applicable in cross-sectional and time-series settings, furthermore, in time-series settings their method overcomes two of the drawbacks associated with dominant approaches, namely, their reliance on only one split of the data and the need to have a sufficiently large 'hold-out' sample for these tests to possess adequate power. They assess the finite-sample performance of the test via Monte Carlo simulation and consider a number of empirical applications that highlight the utility of the approach.

Default probability (the probability that a borrower will fail to serve its obligation) is central to the study of risk management. Bonds and other tradable debt instruments are the main source of default for most individual and institutional investors. In contrast, loans are the largest and most obvious source of default for banks. Default prediction is becoming more and more important for banks, especially in risk management, in order to measure their clients degree of risk. In Chapter 11, Wolfgang Härdle, Dedy Dwi Prastyo and Christian Hafner consider the use of Support Vector Machines (SVM) for modeling default probability. SVM is a state-of-the-art nonlinear classification technique that is well-suited to the study of default risk. This chapter emphasizes SVM-based default prediction applied to the CreditReform database. The SVM parameters are optimized by using an evolutionary algorithm (the so-called "Genetic Algorithm") and show how the "imbalanced problem" may be overcome by the use of "down-sampling" and "oversampling."

Time Series

In Chapter 12, Peter C. B. Phillips and Zhipeng Liao consider an overview of recent developments in series estimation of stochastic processes and some of their applications in Econometrics. They emphasize the idea that a stochastic process may, under certain conditions, be represented in terms of a set of orthonormal basis functions, giving a series representation that involves deterministic functions. Several applications of this series approximation method are discussed. The first shows how a continuous function can be approximated by a linear combination of Brownian motions (BMs), which is useful in the study of spurious regression. The second application utilizes the series representation of BM to investigate the effect of the presence of deterministic trends in a regression on traditional unit-root tests. The third uses basis functions in the series approximation as instrumental variables to perform efficient estimation of the parameters in cointegrated systems. The fourth application proposes alternative estimators of long-run variances in some econometric models with dependent data, thereby providing autocorrelation robust inference methods in these models. The authors review work related to these applications and ongoing research involving series approximation methods.

In Chapter 13, Jiti Gao considers some identification, estimation, and specification problems in a class of semilinear time series models. Existing studies for the stationary time series case are reviewed and discussed, and Gao also establishes some new results for the integrated time series case. The author also proposes a new estimation method and establishes a new theory for a class of semilinear nonstationary autoregressive models.

Nonparametric and semiparametric estimation and hypothesis testing methods have been intensively studied for cross-sectional independent data and weakly dependent time series data. However, many important macroeconomics and financial data are found to exhibit stochastic and/or deterministic trends, and the trends can be nonlinear in nature. While a linear model may provide a decent approximation to a nonlinear model for weakly dependent data, the linearization can result in severely biased approximation to a nonlinear model with nonstationary data. In Chapter 14, Yiguo Sun and Qi Li review some recent theoretical developments in nonparametric and semiparametric techniques applied to nonstationary or near nonstationary variables. First, this chapter reviews some of the existing works on extending the I(0), I(1), and cointegrating relation concepts defined in a linear model to a nonlinear framework, and it points out some difficulties in providing satisfactory answers to extend the concepts of I(0), I(1), and cointegration to nonlinear models with persistent time series data. Second, the chapter reviews kernel estimation and hypothesis testing for nonparametric and semiparametric autoregressive and cointegrating models to explore unknown nonlinear relations among I(1) or near I(1) process(es). The asymptotic mixed normal results of kernel estimation generally replace asymptotic normality results usually obtained for weakly dependent data. The authors also discuss kernel estimation of semiparametric varying coefficient regression models with correlated but not cointegrated data. Finally, the authors discuss the concept of co-summability introduced by Berengner-Rico and Gonzalo (2012), which provides an extension of cointegration concepts to nonlinear time series data.

Cross Section

Sets of regression equations (SREs) play a central role in Econometrics. In Chapter 15, Aman Ullah and Yun Wang review some of the recent developments for the estimation of SRE within semi- and nonparametric frameworks. Estimation procedures for various nonparametric and semiparametric SRE models are presented including those for partially linear semiparametric models, models with nonparametric autocorrelated errors, additive nonparametric models, varying coefficient models, and models with endogeneity.

In Chapter 16, Daniel J. Henderson and Esfandiar Maasoumi suggest some new directions in the analysis of nonparametric models with exogenous treatment assignment. The nonparametric approach opens the door to the examination of potentially different distributed outcomes. When combined with cross-validation, it also identifies potentially irrelevant variables and linear versus nonlinear effects. Examination of the distribution of effects requires distribution metrics, such as stochastic dominance tests for ranking based on a wide range of criterion functions, including dollar valuations. They can identify subgroups with different treatment outcomes, and they offer an empirical demonstration based on the GAIN data. In the case of one covariate (English as the primary language), there is support for a statistical inference of uniform first-order dominant treatment effects. The authors also find several others that indicate second- and higher-order dominance rankings to a statistical degree of confidence.

Jeffrey S. Racine
Liangjun Su
Aman Ullah

References

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman and Hall, London.

Härdle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, New York.

Horowitz, J. L. (2009), *Semiparametric and Nonparametric Methods in Econometrics*, Springer-Verlag.

Li, Q. & Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Pagan, A. & Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge University Press, New York.

Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*, Academic Press, Orlando, FL.

Yatchew, A. J. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge University Press, New York.

Acknowledgments

We would like to express our deep gratitude to each of the the contributors without whom this volume would not be possible. We would also like to acknowledge the guidance and encouragement of the staff at Oxford University Press—in particular, Terry Vaughn and Cathryn Vaulman, whose tireless efforts and keen eye have smoothed the production of this handbook. Finally, we would like to thank Damaris Carlos, who helped us with many different types of tasks.

List of Contributors

Herman J. Bierens, Professor Emeritus of Economics, Pennsylvania State University

Marine Carrasco, Professor of Economics, University of Montreal

Jean-Pierre Florens, Professor of Mathematics, Toulouse School of Economics; Research Director, Institut D'Economie Industrielle

Jiti Gao, Professor, Department of Econometrics and Business Statistics, Monash University

Christian M. Hafner, Professor, School of Economics, Catholic University of Louvain

Bruce E. Hansen, Trygve Haavelmo Professor, Department of Economics, University of Wisconsin, Madison

Wolfgang Karl Härdle, Ladislaus von Bortkiewicz Chair of Statistics, Department of Economics and Business Administration, Humboldt University

Daniel J. Henderson, J. Weldon and Delores Cole Faculty Fellow of Finance and Legal, Department of Economics, University of Alabama

Joel L. Horowitz, Charles E. and Emma H. Morrison Professor, Department of Economics, Northwestern University

Arthur Lewbel, Professor, Department of Economics, Boston College

Qi Li, Hugh Roy Cullen Professor in Liberal Arts, Department of Economics, Texas A&M University

Zhipeng Liao, Assistant Professor, Department of Economics, University of California, Los Angeles

Shujie Ma, Assistant Professor, Department of Statistics, University of California, Riverside

Esfandiar Maasoumi, Arts and Sciences Distinguished Professor, Department of Economics, Emory University

Enno Mammen, Professor, Department of Economics, University of Mannheim

Byeong U. Park, Professor, Department of Statistics, Seoul National University

Christopher F. Parmeter, Assistant Professor, Department of Economics, University of Miami

Peter C. B. Phillips, Sterling Professor of Economics and Professor of Statistics, Yale University; Alumni Distinguished Professor of Economics, University of Auckland; Adjunct professor, University of Southampton and Singapore Management University; Co-Director, Centre for Financial Econometrics, Singapore Management University

Dedy Dwi Prastyo, PhD student, School of Business & Economics, Humboldt University

Jeffrey S. Racine, Senator William McMaster Chair in Econometrics and Professor of Economics and Statistics, McMaster University

Eric Renault, Professor, Department of Economics, University of Montreal for Carrasco, Toulouse School of Economics for Florens, Brown University for Renault.

Melanie Schienle, Professor, School of Economics & Management, Leibniz University Hannover

Liangjun Su, Professor, School of Economics, Singapore Management University

Yiguo Sun, Professor, Department of Economics & Finance, University of Guelph

Aman Ullah, Distinguished Professor, Department of Economics, University of California, Riverside

Yun Wang, Assistant Professor, School of International Trade and Economics, University of International Business and Economics

Lijian Yang, Professor, Department of Statistics and Probability and an Adjunct Professor for the Center for Global Change and Earth Observations, Michigan State University

Yonghui Zhang, Assistant Professor, School of Economics, Renmin University of China, China

Victoria Zinde-Walsh, Professor, Department of Economics, McGill University

PART I

METHODOLOGY

CHAPTER 1

THE HILBERT SPACE THEORETICAL FOUNDATION OF SEMI-NONPARAMETRIC MODELING

HERMAN J. BIERENS

1.1. Introduction

Semi-nonparametric (SNP) models are models where only a part of the model is parameterized, and the nonspecified part is an unknown function that is represented by an infinite series expansion. Therefore, SNP models are, in essence, models with infinitely many parameters. The parametric part of the model is often specified as a linear index, that is, a linear combination of conditioning and/or endogenous variables, with the coefficients involved the parameters of interests, which we will call the structural parameters. Although the unknown function involved is of interest as well, the parameters in its series expansion are only of interest insofar as they determine the shape of this function.

The theoretical foundation of series expansions of functions is Hilbert space theory, in particular the properties of Hilbert spaces of square integrable real functions. Loosely speaking, Hilbert spaces are vector spaces with properties similar to those of Euclidean spaces. As is well known, any vector in the Euclidean space $\mathbb{R}^k$ can be represented by a linear combination of k orthonormal vectors. Similarly, in Hilbert spaces of functions, there exist sequences of orthonormal functions such that any function in this space can be represented by a linear combination of these orthonormal functions. Such orthonormal sequences are called complete.

The main purpose of this chapter is to show how these orthonormal functions can be constructed and how to construct general series representations of density and distribution functions. Moreover, in order to explain why this can be done, I will review the necessary Hilbert space theory involved as well.

The standard approach to estimate SNP models is sieve estimation, proposed by Grenander (1981). Loosely speaking, sieve estimation is like standard parameter estimation, except that the dimension of the parameter space involved increases to infinity with the sample size. See Chen (2007) for a review of sieve estimation. However, the main focus of this chapter is on SNP modeling rather than on estimation.

Gallant (1981) was the first econometrician to propose Fourier series expansions as a way to model unknown functions. See also Eastwood and Gallant (1991) and the references therein. However, the use of Fourier series expansions to model unknown functions has been proposed earlier in the statistics literature. See, for example, Kronmal and Tarter (1968).

Gallant and Nychka (1987) consider SNP modeling and sieve estimation of Heckman's (1979) sample selection model, where the bivariate error distribution of the latent variable equations is modeled semi-nonparametrically using a bivariate Hermite polynomial expansion of the error density.

Another example of an SNP model is the mixed proportional hazard (MPH) model proposed by Lancaster (1979), which is a proportional hazard model with unobserved heterogeneity. Elbers and Ridder (1982) and Heckman and Singer (1984) have shown that under mild conditions the MPH model is nonparametrically identified. The latter authors propose to model the distribution function of the unobserved heterogeneity variable by a discrete distribution. Bierens (2008) and Bierens and Carvalho (2007) use orthonormal Legendre polynomials to model semi-nonparametrically the unobserved heterogeneity distribution of interval-censored mixed proportional hazard models and bivariate mixed proportional hazard models, respectively.

However, an issue with the single-spell MPH model is that for particular specifications of the baseline hazard, its efficiency bound is singular, which implies that any consistent estimator of the Euclidean parameters in the MPH model involved converges at a slower rate than the square root of the sample size. See Newey (1990) for a general review of efficiency bounds, and see Hahn (1994) and Ridder and Woutersen (2003) for the efficiency bound of the MPH model. On the other hand, Hahn (1994) also shows that in general the multiple-spell MPH model does not suffer from this problem, which is confirmed by the estimation results of Bierens and Carvalho (2007).

This chapter is organized as follows. In Section 1.2 I will discuss three examples of SNP models,[1] with focus on semiparametric identification. The SNP index regression model is chosen as an example because it is one of the few SNP models where the unknown function involved is not a density or distribution function. The two other examples are the bivariate MPH model in Bierens and Carvalho (2007) and the first-price auction model in Bierens and Song (2012, 2013), which have been chosen because these papers demonstrate how to do SNP modeling and estimation in practice, and in both models the unknown function involved is a distribution function. Section 1.3 reviews Hilbert space theory. In Section 1.4 it will be shown how to generate various sequences of orthonormal polynomials, along with what kind of Hilbert spaces they span. Moreover, it will also be shown how these results can be applied to the SNP index regression model. In Section 1.5 various nonpolynomial complete orthonormal sequences of functions will be derived. In Section 1.6 it will be shown how arbitrary density and distribution functions can be represented by series expansions in terms of complete orthonormal sequences of functions, along with how these results can be applied to the bivariate MPH model in Bierens and Carvalho (2007) and to the first-price auction model in Bierens and Song (2012, 2013). In Section 1.7 I will briefly discuss the sieve estimation approach, and in Section 1.8 I will make a few concluding remarks.

Throughout this chapter I will use the following notations. The well-known indicator function will be denoted by $1(\cdot)$, the set of positive integers will be denoted by $\mathbb{N}$, and the set of non-negative integers, $\mathbb{N} \cup \{0\}$, by $\mathbb{N}_0$. The abbreviation "a.s." stands for "almost surely"—that is, the property involved holds with probability 1—and "a.e." stands for "almost everywhere," which means that the property involved holds except perhaps on a set with Lebesgue measure zero.

1.2. Examples of SNP Models

1.2.1. The SNP Index Regression Model

Let Y be a dependent variable satisfying $E[Y^2] < \infty$, and let $X \in \mathbb{R}^k$ be a vector of explanatory variables. As is well known, the conditional expectation $E[Y|X]$ can be written as $E[Y|X] = g_0(X)$, where $g_0(x)$ is a Borel measurable real function on $\mathbb{R}^k$.[2] Newey (1997) proposed to estimate $g_0(x)$ by sieve estimation via a multivariate series expansion. However, because there are no parameters involved, the resulting estimate of $g_0(x)$ can only be displayed and interpreted graphically, which in practice is only possible for $k \leq 2$. Moreover, to approximate a bivariate function $g_0(x)$ by a series expansion of order n requires n^2 parameters.[3] Therefore, a more practical approach is the following.

Suppose that there exists a $\beta_0 \in \mathbb{R}^k$ such that $E[Y|X] = E[Y|\beta_0'X]$ a.s. Then there exists a Borel measurable real function $f(x)$ on $\mathbb{R}$ such that $E[Y|X] = f(\beta_0'X)$ a.s. Because for any nonzero constant c, $E[Y|\beta_0'X] = E[Y|c\beta_0'X]$ a.s., identification of f requires to normalize β_0 in some way, for example by setting one component of β_0 to 1. Thus, in the case $k \geq 2$, let $X = (X_1, X_2')'$ with $X_2 \in \mathbb{R}^{k-1}$, and $\beta_0 = (1, \theta_0')'$ with $\theta_0 \in \mathbb{R}^{k-1}$, so that

$$E[Y|X] = f(X_1 + \theta_0'X_2) \qquad \text{a.s.} \tag{1.1}$$

To derive further conditions for the identification of f and θ_0, suppose that for some $\theta_* \neq \theta_0$ there exists a function f_* such that $f(X_1 + \theta_0'X_2) = f_*(X_1 + \theta_*'X_2)$ a.s. Moreover, suppose that the conditional distribution of X_1 given X_2 is absolutely continuous with support $\mathbb{R}$. Then conditional on X_2, $f(x_1 + \theta_0'X_2) = f_*(x_1 + \theta_0'X_2 + (\theta_* - \theta_0)'X_2)$ a.s.

for all $x_1 \in \mathbb{R}$. Consequently, for arbitrary $z \in \mathbb{R}$ we may choose $x_1 = z - \theta_0' X_2$, so that

$$f(z) = f_*(z + (\theta_* - \theta_0)' X_2) \qquad \text{a.s. for all } z \in \mathbb{R}. \tag{1.2}$$

If $f(z)$ is constant, then $E[Y|X] = E[Y]$ a.s., so let us exclude this case. Then (1.2) is only possible if $(\theta_* - \theta_0)' X_2$ is a.s. constant, which in turn implies that $(\theta_* - \theta_0)'(X_2 - E[X_2]) = 0$ a.s. and thus $(\theta_* - \theta_0)' E[(X_2 - E[X_2])(X_2 - E[X_2])'](\theta_* - \theta_0) = 0$. Therefore, if $\mathrm{Var}[X_2]$ is nonsingular, then $\theta_* = \theta_0$.

Summarizing, it has been shown that the following results hold.

Theorem 1.1. *The function $f(z)$ and the parameter vector θ_0 in the index regression model* (1.1) *are identified if*

(*a*) $\Pr[E(Y|X) = E(Y)] < 1$;
(*b*) *The conditional distribution of X_1 given X_2 is absolutely continuous with support $\mathbb{R}$;*
(*c*) *The variance matrix of X_2 is finite and nonsingular.*
Moreover, in the case $X \in \mathbb{R}$ the regression function $f(z)$ is identified for all $z \in \mathbb{R}$ if the distribution of X is absolutely continuous with support $\mathbb{R}$.

The problem how to model $f(z)$ semi-nonparametrically and how to estimate f and θ_0 will be addressed in Section 1.4.4.

1.2.2. The MPH Competing Risks Model

Consider two durations, T_1 and T_2. Suppose that conditional on a vector X of covariates and a common unobserved (heterogeneity) variable V, which is assumed to be independent of X, the durations T_1 and T_2 are independent, that is, $\Pr[T_1 \leq t_1, T_2 \leq t_2 | X, V] = \Pr[T_1 \leq t_1 | X, V].\Pr[T_2 \leq t_2 | X, V]$. This is a common assumption in bivariate survival analysis. See van den Berg (2000). If the conditional distributions of the durations T_1 and T_2 are of the mixed proportional hazard type, then their survival functions conditional on X and V take the form $S_i(t|X,V) = \Pr[T_i > t|X,V] = \exp(-V \exp(\beta_i' X) \Lambda_i(t|\alpha_i))$, $i = 1,2$, where $\Lambda_i(t|\alpha_i) = \int_0^t \lambda_i(\tau|\alpha_i) d\tau$, $i = 1,2$, are the integrated baseline hazards depending on parameter vectors α_i.

This model is also known as the competing risks model. It is used in Bierens and Carvalho (2007) to model two types of recidivism durations of ex-convicts, namely (a) the time T_1 between release from prison and the first arrest for a misdemeanor and (b) the time T_2 between release from prison and the first arrest for a felony, with Weibull baseline hazards, that is,

$$\begin{aligned} &\lambda(t|\alpha_i) = \alpha_{i,1}\alpha_{i,2} t^{\alpha_{i,2}-1}, \quad \Lambda(t|\alpha_i) = \alpha_{i,1} t^{\alpha_{i,2}}, \qquad \alpha_{i,1} > 0,\ \alpha_{i,2} > 0, \\ &\text{with } \alpha_i = (\alpha_{i,1}, \alpha_{1,2})', \qquad i = 1,2, \end{aligned} \tag{1.3}$$

where $\alpha_{i,1}$ is a scale factor.

In this recidivism case we only observe $T = \min(T_1, T_2)$ together with a discrete variable D that is 1 if $T_2 > T_1$ and 2 if $T_2 \leq T_1$. Thus, $D = 1$ corresponds to rearrest for a misdemeanor and $D = 2$ corresponds to rearrests for a felony. Then conditional on X and V, $\Pr[T > t, D = i|X, V] = \int_t^\infty V \exp(-V(\exp(\beta_1' X)\Lambda(\tau|\alpha_1) + \exp(\beta_2' X)\Lambda(\tau|\alpha_2))) \cdot \exp(\beta_i' X)\lambda(\tau|\alpha_i)\, d\tau$, $i = 1, 2$, which is not hard to verify. Integrating V out now yields

$$\begin{aligned}
&\Pr[T > t, D = i|X] \\
&= \int_t^\infty \int_0^\infty v \exp\big(-v\big(\exp\big(\beta_1' X\big)\Lambda(\tau|\alpha_1) + \exp\big(\beta_2' X\big)\Lambda(\tau|\alpha_2)\big)\big)\, dG(v) \\
&\qquad \times \exp\big(\beta_i' X\big)\lambda(\tau|\alpha_i)\, d\tau, \qquad i = 1, 2, \qquad (1.4)
\end{aligned}$$

where $G(v)$ is the (unknown) distribution function of V.

It has been shown in Bierens and Carvalho (2006), by specializing the more general identification results of Heckman and Honore (1989) and Abbring and van den Berg (2003), that under two mild conditions the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ and the distribution function G are identified. One of these conditions is that the variance matrix of X is finite and nonsingular. The other condition is that $E[V] = 1$,[4] so that (1.4) can be written as

$$\begin{aligned}
&\Pr[T > t, D = d|X] \\
&= \int_t^\infty H\big(\exp\big(-\big(\exp\big(\beta_1' X\big)\Lambda(\tau|\alpha_1) + \exp\big(\beta_2' X\big)\Lambda(\tau|\alpha_2)\big)\big)\big) \\
&\quad \times \exp\big(\beta_d' X\big)\lambda(\tau|\alpha_d)\, d\tau, \qquad d = 1, 2, \qquad (1.5)
\end{aligned}$$

where

$$H(u) = \int_0^\infty v u^v\, dG(v) \qquad (1.6)$$

is a distribution function on the unit interval $[0, 1]$. Thus,

Theorem 1.2. *If the variance matrix of X is finite and nonsingular, then the parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ and the distribution function $H(u)$ in the MPH competing risks Weibull model* (1.5) *are identified.*

Proof. (Bierens and Carvalho, 2006, 2007). ■

It follows now straightforwardly from (1.5) that, given a random sample $\{T_j, D_j, X_j\}_{j=1}^N$ from (T, D, X), the log-likelihood function involved takes the form $\ln(L_N(\alpha_1, \alpha_2, \beta_1, \beta_2, H)) = \sum_{j=1}^N \ell(T_j, D_j, X_j|\alpha_1, \alpha_2, \beta_1, \beta_2, H)$, where

$$\begin{aligned}
&\ell(T, D, X|\alpha_1, \alpha_2, \beta_1, \beta_2, H) \\
&= \ln\big(H\big(\exp\big(-\big(\exp\big(\beta_1' X\big)\Lambda(T|\alpha_1) + \exp\big(\beta_2' X\big)\Lambda(T|\alpha_2)\big)\big)\big)\big) \\
&\quad + (2 - D)\big(\beta_1' X + \ln(\lambda(T|\alpha_1))\big) + (D - 1)\big(\beta_2' X + \ln(\lambda(T|\alpha_2))\big). \qquad (1.7)
\end{aligned}$$

At this point the distribution function $H(u)$ representing the distribution of the unobserved heterogeneity is treated as a parameter. The problem of how to model $H(u)$ semi-nonparametrically will be addressed in Section 1.6.

Note that the duration $T = \min(T_1, T_2)$ in Bierens and Carvalho (2007) is only observed over a period $[0, \overline{T}]$, where $\overline{T}$ varies only slightly per ex-inmate, so that T is right-censored. Therefore, the actual log-likelihood in Bierens and Carvalho (2007) is more complicated than displayed in (1.7).

1.2.3. First-Price Auctions

A first price-sealed bids auction (henceforth called *first-price auction*) is an auction with $I \geq 2$ potential bidders, where the potential bidder's values for the item to be auctioned off are independent and private, and the bidders are symmetric and risk neutral. The reservation price p_0, if any, is announced in advance and the number I of potential bidders is known to each potential bidder.

As is well known, the equilibrium bid function of a first-price auction takes the form

$$\beta(v|F,I) = v - \frac{1}{F(v)^{I-1}} \int_{p_0}^{v} F(x)^{I-1}\,dx \qquad \text{for } v > p_0 > \underline{v}, \tag{1.8}$$

if the reservation price p_0 is binding, and

$$\beta(v|F,I) = v - \frac{1}{F(v)^{I-1}} \int_{0}^{v} F(x)^{I-1}\,dx \qquad \text{for } v > \underline{v}, \tag{1.9}$$

if the reservation price p_0 is nonbinding, where $F(v)$ is the value distribution, $I \geq 2$ is the number of potential bidders, and $\underline{v} \geq 0$ is the lower bound of the support of $F(v)$. See, for example, Riley and Samuelson (1981) or Krishna (2002). Thus, if the reservation price p_0 is binding, then, with V_j the value for bidder j for the item to be auctioned off, this potential bidder issues a bid $B_j = \beta(V_j|F,I)$ according to bid function (1.8) if $V_j > p_0$ and does not issue a bid if $V_j \leq p_0$, whereas if the reservation price p_0 is not binding, each potential bidder j issues a bid $B_j = \beta(V_j|F,I)$ according to bid function (1.9). In the first-price auction model the individual values $V_j, j = 1, \ldots, I$, are assumed to be independent random drawing from the value distribution F. The latter is known to each potential bidder j, and so is the number of potential bidders, I.

Guerre et al. (2000) have shown that if the value distribution $F(v)$ is absolutely continuous with density $f(v)$ and bounded support $[\underline{v}, \overline{v}]$, $\overline{v} < \infty$, then $f(v)$ is nonparametrically identified from the distribution of the bids. In particular, if the reservation price is nonbinding, then the inverse bid function is $v = b + (I-1)^{-1}\Lambda(b)/\lambda(b)$, where v is a private value, b is the corresponding bid, and $\Lambda(b)$ is the distribution function of the bids with density $\lambda(b)$. Guerre et al. (2000) propose to estimate the latter two functions via nonparametric kernel methods, as $\hat{\Lambda}(b)$ and $\hat{\lambda}(b)$, respectively. Using the pseudo-private values $\widetilde{V} = B + (I-1)^{-1}\hat{\Lambda}(B)/\hat{\lambda}(B)$, where each B is an observed bid, the density $f(v)$ of the private value distribution can now be estimated by kernel density estimation.

Bierens and Song (2012) have shown that the first-price auction model is also nonparametrically identified if instead of the bounded support condition, the value distribution F in (1.8) and (1.9) is absolutely continuous on $(0,\infty)$ with connected support[5] and finite expectation. As an alternative to the two-step nonparametric approach of Guerre et al. (2000), Bierens and Song (2012) propose a simulated method of moments sieve estimation approach to estimate the true value distribution $F_0(v)$, as follows. For each SNP candidate value distribution F, generate simulated bids according to the bid functions (1.8) or (1.9) and then minimize the integrated squared difference of the empirical characteristic functions of the actual bids and the simulated bids to the SNP candidate value distributions involved.

This approach has been extended in Bierens and Song (2013) to first-price auctions with auction-specific observed heterogeneity. In particular, given a vector X of auction-specific covariates, Bierens and Song (2013) assume that $\ln(V) = \theta' X + \varepsilon$, where X and ε are independent. Denoting the distribution function of $\exp(\varepsilon)$ by F, the conditional distribution of V given X then takes the form $F(v\exp(-\theta' X))$.

1.3. Hilbert Spaces

1.3.1. Inner Products

As is well known, in a Euclidean space $\mathbb{R}^k$ the inner product of a pair of vectors $x = (x_1,\dots,x_k)'$ and $y = (y_1,\dots,y_k)'$ is defined as $x'y = \sum_{m=1}^{k} x_m y_m$, which is a mapping $\mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ satisfying $x'y = y'x$, $(cx)'y = c(x'y)$ for arbitrary $c \in \mathbb{R}$, $(x+y)'z = x'z + y'z$, and $x'x > 0$ if and only if $x \neq 0$. Moreover, the norm of a vector $x \in \mathbb{R}^k$ is defined as $||x|| = \sqrt{x'x}$, with associated metric $||x-y||$. Of course, in $\mathbb{R}$ the inner product is the ordinary product $x \cdot y$.

Mimicking these properties of inner product, we can define more general inner products with associated norms and metrics as follows.

Definition 1.1. *An inner product on a real vector space $\mathcal{V}$ is a real function $\langle x,y\rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ such that for all x, y, z in $\mathcal{V}$ and all c in $\mathbb{R}$, we obtain the following:*

1. $\langle x,y\rangle = \langle y,x\rangle$.
2. $\langle cx,y\rangle = c\langle x,y\rangle$.
3. $\langle x+y,z\rangle = \langle x,z\rangle + \langle y,z\rangle$.
4. $\langle x,x\rangle > 0$ *if and only if* $x \neq 0$.
   Given an inner product, the associated norm and metric are defined as $||x|| = \sqrt{\langle x,x\rangle}$ *and* $||x-y||$, *respectively.*

As is well known from linear algebra, for vectors $x, y \in \mathbb{R}^k$, $|x'y| \leq ||x||.||y||$, which is known as the Cauchy–Schwarz inequality. This inequality carries straightforwardly over to general inner products:

Theorem 1.3. *(Cauchy–Schwarz inequality)* $|\langle x, y\rangle| \leq ||x||.||y||$.

1.3.2. Convergence of Cauchy Sequences

Another well-known property of a Euclidean space is that every Cauchy sequence has a limit in the Euclidean space involved.[6] Recall that a sequence of elements x_n of a metric space with metric $||x - y||$ is called a Cauchy sequence if $\lim_{\min(k,m)\to\infty} ||x_k - x_m|| = 0$.

Definition 1.2. *A Hilbert space $\mathcal{H}$ is a vector space endowed with an inner product and associated norm and metric such that every Cauchy sequence has a limit in $\mathcal{H}$.*

Thus, a Euclidean space is a Hilbert space, but Hilbert spaces are much more general than Euclidean spaces.

To demonstrate the role of the Cauchy convergence property, consider the vector space $C[0,1]$ of continuous real functions on $[0,1]$. Endow this space with the inner product $\langle f, g\rangle = \int_0^1 f(u)g(u)\, du$ and associated norm $||f|| = \sqrt{\langle f, f\rangle}$ and metric $||f - g||$. Now consider the following sequence of functions in $C[0,1]$:

$$f_n(u) = \begin{cases} 0 & \text{for} \quad 0 \leq u < 0.5, \\ 2^n(u - 0.5) & \text{for} \quad 0.5 \leq u < 0.5 + 2^{-n}, \\ 1 & \text{for} \quad 0.5 + 2^{-n} \leq u \leq 1, \end{cases}$$

for $n \in \mathbb{N}$. It is an easy calculus exercise to verify that f_n is a Cauchy sequence in $C[0,1]$. Moreover, it follows from the bounded convergence theorem that $\lim_{n\to\infty} ||f_n - f|| = 0$, where $f(u) = \mathbf{1}(u > 0.5)$. However, this limit $f(u)$ is discontinuous in $u = 0.5$, and thus $f \notin C[0,1]$. Therefore, the space $C[0,1]$ is not a Hilbert space.

1.3.3. Hilbert Spaces Spanned by a Sequence

Let $\mathcal{H}$ be a Hilbert space and let $\{x_k\}_{k=1}^{\infty}$ be a sequence of elements of $\mathcal{H}$. Denote by

$$\mathcal{M}_m = \text{span}(\{x_j\}_{j=1}^m)$$

the subspace spanned by $x_1, \ldots, x_m$; that is, $\mathcal{M}_m$ consists of all linear combinations of $x_1, \ldots, x_m$. Because every Cauchy sequence in $\mathcal{M}_m$ takes the form $z_n = \sum_{i=1}^m c_{i,n} x_i$, where the $c_{i,n}$'s are Cauchy sequences in $\mathbb{R}$ with limits $c_i = \lim_{n\to\infty} c_{i,n}$, it follows trivially that $\lim_{n\to\infty} ||z_n - z|| = 0$, where $z = \sum_{i=1}^m c_i x_i \in \mathcal{M}_m$. Thus, $\mathcal{M}_m$ is a Hilbert space.

Definition 1.3. *The space* $\mathcal{M}_\infty = \overline{\cup_{m=1}^{\infty}\mathcal{M}_m}$[7] *is called the space spanned by* $\{x_j\}_{j=1}^{\infty}$, *which is also denoted by span*$(\{x_j\}_{j=1}^{\infty})$.

Let x_n be a Cauchy sequence in $\mathcal{M}_\infty$. Then x_n has a limit $\overline{x} \in \mathcal{H}$, that is, $\lim_{n\to\infty}||x_n - \overline{x}|| = 0$. Suppose that $\overline{x} \notin \mathcal{M}_\infty$. Because $\mathcal{M}_\infty$ is closed, there exists an $\varepsilon > 0$ such that the set $\mathcal{N}(\overline{x},\varepsilon) = \{x \in \mathcal{H} : ||x - \overline{x}|| < \varepsilon\}$ is completely outside $\mathcal{M}_\infty$, that is, $\mathcal{N}(\overline{x},\varepsilon) \cap \mathcal{M}_\infty = \emptyset$. But $\lim_{n\to\infty}||x_n - \overline{x}|| = 0$ implies that there exists an $\underline{n}(\varepsilon)$ such that $x_n \in \mathcal{N}(\overline{x},\varepsilon)$ for all $n > \underline{n}(\varepsilon)$, hence $x_n \notin \mathcal{M}_\infty$ for all $n > \underline{n}(\varepsilon)$, which contradicts $x_n \in \mathcal{M}_\infty$ for all n. Thus,

Theorem 1.4. $\mathcal{M}_\infty$ *is a Hilbert space.*

In general, $\mathcal{M}_\infty$ is smaller than $\mathcal{H}$, but as we will see there exist Hilbert spaces $\mathcal{H}$ containing a sequence $\{x_j\}_{j=1}^{\infty}$ for which $\mathcal{M}_\infty = \mathcal{H}$. Such a sequence is called complete:

Definition 1.4. *A sequence* $\{x_k\}_{k=1}^{\infty}$ *in a Hilbert space* $\mathcal{H}$ *is called complete if* $\mathcal{H} =$ *span*$(\{x_j\}_{j=1}^{\infty})$.

Of particular importance for SNP modeling are Hilbert spaces spanned by a complete orthonormal sequence, because in that case the following approximation result holds.

Theorem 1.5. *Let* $\{x_j\}_{j=1}^{\infty}$ *a complete orthonormal sequence in a Hilbert space* $\mathcal{H}$, *that is,* $\langle x_i, x_j\rangle = 1(i = j)$ *and* $\mathcal{H} = \mathrm{span}(\{x_j\}_{j=1}^{\infty})$. *For an arbitrary* $y \in \mathcal{H}$, *let* $\widehat{y}_n = \sum_{j=1}^{n}\langle y, x_j\rangle x_j$. *Then* $\lim_{n\to\infty}||y - \widehat{y}_n|| = 0$ *and* $\sum_{j=1}^{\infty}\langle y, x_j\rangle^2 = ||y||^2$.

This result is a corollary of the fundamental projection theorem:

Theorem 1.6. *Let* $\mathcal{S}$ *be a sub-Hilbert space of a Hilbert space* $\mathcal{H}$. *Then for any* $y \in \mathcal{H}$ *there exists a* $\widehat{y} \in \mathcal{S}$ *(called the projection of* y *on* $\mathcal{S}$*) such that* $||y - \widehat{y}|| = \inf_{z\in\mathcal{S}}||y - z||$. *Moreover, the projection residual* $u = y - \widehat{y}$ *satisfies* $\langle u, z\rangle = 0$ *for all* $z \in \mathcal{S}$.[8]

Now observe that $\widehat{y}_n$ in Theorem 1.5 is the projection of y on $\mathcal{M}_n = \mathrm{span}(\{x_j\}_{j=1}^{n})$, with residual $u_n = \underline{y - \widehat{y}_n}$ satisfying $\langle u_n, y_n\rangle = 0$ for all $y_n \in \mathcal{M}_n$, and that due to $y \in \mathrm{span}(\{x_j\}_{j=1}^{\infty}) = \cup_{m=1}^{\infty}\mathcal{M}_m$ there exists a sequence $y_n \in \mathcal{M}_n$ such that $\lim_{n\to\infty}||y - y_n|| = 0$. Then $||y - \widehat{y}_n||^2 = \langle u_n, y - \widehat{y}_n\rangle = \langle u_n, y\rangle = \langle u_n, y - y_n\rangle \leq ||u_n||.||y - y_n|| \leq ||y||.||y - y_n|| \to 0$, where the first inequality follows from the Cauchy–Schwarz inequality while the second inequality follows from the fact that $||u_n||^2 \leq ||y||^2$. Moreover, the result $\sum_{j=1}^{\infty}\langle y, x_j\rangle^2 = ||y||^2$ in Theorem 1.5 follows from the fact that $||y||^2 = \langle y, y\rangle = \lim_{n\to\infty}\langle\widehat{y}_n, y\rangle = \lim_{n\to\infty}\sum_{j=1}^{n}\langle y, x_j\rangle^2$.

1.3.4. Examples of Non-Euclidean Hilbert Spaces

Consider the space $\mathcal{R}$ of random variables defined on a common probability space $\{\Omega, \mathcal{F}, P\}$ with finite second moments, endowed with the inner product $\langle X, Y\rangle =$

$E[X.Y]$ and associated norm $||X|| = \sqrt{\langle X, X\rangle} = \sqrt{E[X^2]}$ and metric $||X - Y||$. Then we have the following theorem.

Theorem 1.7. *The space $\mathcal{R}$ is a Hilbert space.*[9]

This result is the basis for the famous Wold (1938) decomposition theorem, which in turn is the basis for time series analysis.

In the rest of this chapter the following function spaces play a key role.

Definition 1.5. *Given a probability density $w(x)$ on $\mathbb{R}$, the space $L^2(w)$ is the space of Borel measurable real functions f on $\mathbb{R}$ satisfying $\int_{-\infty}^{\infty} f(x)^2 w(x)\, dx < \infty$, endowed with the inner product $\langle f, g\rangle = \int_{-\infty}^{\infty} f(x)g(x)w(x)\, dx$ and associated norm $||f|| = \sqrt{\langle f, f\rangle}$ and metric $||f - g||$. Moreover, $L^2(a, b)$, $-\infty \le a < b \le \infty$, is the space of Borel measurable real functions on (a, b) satisfying $\int_a^b f(x)^2\, dx$, with inner product $\langle f, g\rangle = \int_a^b f(x)g(x)\, dx$ and associated norm and metric.*

Then for $f, g \in L^2(w)$, we have $\langle f, g\rangle = E[f(X)g(X)]$, where X is a random drawing from the distribution with density $w(x)$; hence from Theorem 1.7 we obtain the following theorem.

Theorem 1.8. *The space $L^2(w)$ is a Hilbert space.*

Also $L^2(a, b)$ is a Hilbert space, as will be shown in Section 1.5.

In general the result $\lim_{n\to\infty} ||y - \widehat{y}_n|| = 0$ in Theorem 1.5 does not imply that $\lim_{n\to\infty} \widehat{y}_n = y$, as the latter limit may not be defined, and even if so, $\lim_{n\to\infty} \widehat{y}_n$ may not be equal to y. However, in the case $\mathcal{H} = L^2(w)$ the result $\lim_{n\to\infty} ||y - \widehat{y}_n|| = 0$ implies $\lim_{n\to\infty} \widehat{y}_n = y$, in the following sense.

Theorem 1.9. *Let $\{\rho_m(x)\}_{m=0}^{\infty}$ be a complete orthonormal sequence in $L^2(w)$,*[10] *and let X be a random drawing from the density w. Then for every function $f \in L^2(w)$, $\Pr[f(X) = \lim_{n\to\infty} \sum_{m=0}^{n} \gamma_m \rho_m(X)] = 1$, where $\gamma_m = \int_{-\infty}^{\infty} \rho_m(x) f(x) w(x)\, dx$ with $\sum_{m=0}^{\infty} \gamma_m^2 = \int_{-\infty}^{\infty} f(x)^2 w(x)\, dx$.*

Proof. Denote $f_n(x) = \sum_{m=0}^{n} \gamma_m \rho_m(x)$, and recall from Theorem 1.5 that $\sum_{m=0}^{\infty} \gamma_m^2 = ||f||^2 < \infty$. It follows now that

$$E[(f(X) - f_n(X))^2] = \int_{-\infty}^{\infty} (f(x) - f_n(x))^2 w(x)\, dx = \sum_{m=n+1}^{\infty} \gamma_m^2 \to 0$$

as $n \to \infty$; hence by Chebyshev's inequality, $p\lim_{n\to\infty} f_n(X) = f(X)$. As is well known,[11] the latter is equivalent to the statement that for every subsequence of n there exists a further subsequence m_k, for example, such that $\Pr[\lim_{k\to\infty} f_{m_k}(X) = f(X)] = 1$, and the same applies to any further subsequence m_{k_n} of m_k:

$$\Pr\left[\lim_{n\to\infty} f_{m_{k_n}}(X) = f(X)\right] = 1. \tag{1.10}$$

Given n, there exists a natural number k_n such that $m_{k_n-1} < n \leq m_{k_n}$, and for such a k_n we obtain

$$E\left[\left(f_{m_{k_n}}(X) - f_n(X)\right)^2\right] = E\left[\left(\sum_{j=n+1}^{m_{k_n}} \gamma_m \rho_m(X)\right)^2\right] = \sum_{j=n+1}^{m_{k_n}} \gamma_m^2 \leq \sum_{j=m_{k_n-1}+1}^{m_{k_n}} \gamma_m^2,$$

hence

$$\sum_{n=0}^{\infty} E[(f_{m_{k_n}}(X) - f_n(X))^2] \leq \sum_{n=0}^{\infty} \sum_{j=m_{k_n-1}+1}^{m_{k_n}} \gamma_m^2 \leq \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

By Chebyshev's inequality and the Borel–Cantelli lemma,[12] the latter implies

$$\Pr\left[\lim_{n\to\infty} (f_{m_{k_n}}(X) - f_n(X))\right] = 1. \tag{1.11}$$

Combining (1.10) and (1.11), the theorem follows. ■

1.4. Orthonormal Polynomials and the Hilbert Spaces They Span

1.4.1. Orthonormal Polynomials

Let $w(x)$ be a density function on $\mathbb{R}$ satisfying

$$\int_{-\infty}^{\infty} |x|^k w(x)\, dx < \infty \qquad \text{for all } k \in \mathbb{N}, \tag{1.12}$$

and let $p_k(x|w)$ be a sequence of polynomials in $x \in \mathbb{R}$ of order $k \in \mathbb{N}_0$ such that $\int_{-\infty}^{\infty} p_k(x|w) p_m(x|w) w(x)\, dx = 0$ if $k \neq m$. In words, the polynomials $p_k(x|w)$ are *orthogonal* with respect to the density function $w(x)$. These orthogonal polynomials can be generated recursively by the three-term recurrence relation (hereafter referred to as TTRR)

$$p_{k+1}(x|w) + (b_k - x)\, p_k(x|w) + c_k p_{k-1}(x|w) = 0, \qquad k \geq 1, \tag{1.13}$$

starting from $p_0(x|w) = 1$ and $p_1(x|w) = x - \int_0^1 z.w(z)\, dz$, for example, where

$$b_k = \frac{\int_{-\infty}^{\infty} x \cdot p_k(x|w)^2 w(x)\, dx}{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x)\, dx}, \qquad c_k = \frac{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x)\, dx}{\int_{-\infty}^{\infty} p_{k-1}(x|w)^2 w(x)\, dx}. \tag{1.14}$$

See, for example, Hamming (1973).

Defining

$$\overline{p}_k(x|w) = \frac{p_k(x|w)}{\sqrt{\int_{-\infty}^{\infty} p_k(y|w)^2 w(y)\ dy}} \tag{1.15}$$

yields a sequence of *orthonormal* polynomials with respect to $w(x)$:

$$\int_{-\infty}^{\infty} \overline{p}_k(x|w)\overline{p}_m(x|w)w(x)\ dx = \mathbf{1}(k=m). \tag{1.16}$$

It follows straightforwardly from (1.13) and (1.15) that these orthonormal polynomials can be generated recursively by the TTRR

$$a_{k+1}.\overline{p}_{k+1}(x|w) + (b_k - x)\overline{p}_k(x|w) + a_k.\overline{p}_{k-1}(x|w) = 0, \qquad k \in \mathbb{N}, \tag{1.17}$$

starting from $\overline{p}_0(x|w) = 1$ and

$$\overline{p}_1(x|w) = \frac{x - \int_{-\infty}^{\infty} z \cdot w(z)\ dz}{\sqrt{\int_{-\infty}^{\infty} \left(y - \int_{-\infty}^{\infty} z \cdot w(z)\ dz\right)^2 w(y)\ dy}},$$

where b_k is the same as in (1.14) and

$$a_k = \frac{\sqrt{\int_{-\infty}^{\infty} p_k(x|w)^2 w(x)\ dx}}{\sqrt{\int_{-\infty}^{\infty} p_{k-1}(x|w)^2 w(x)\ dx}}.$$

The sequence is $\overline{p}_k(x|w)$ uniquely determined by $w(x)$, except for signs. In other words, $|\overline{p}_k(x|w)|$ is unique. To show this, suppose that there exists another sequence $\overline{p}_k^*(x|w)$ of orthonormal polynomials w.r.t. $w(x)$. Since $\overline{p}_k^*(x|w)$ is a polynomial of order k, we can write $\overline{p}_k^*(x|w) = \sum_{m=0}^{k} \beta_{m,k}\overline{p}_m(x|w)$. Similarly, we can write $\overline{p}_k(x|w) = \sum_{m=0}^{k} \alpha_{m,.k}\overline{p}_m^*(x|w)$. Then for $j < k$, we have

$$\int_{-\infty}^{\infty} \overline{p}_k^*(x|w)\overline{p}_j(x|w)w(x)\ dx = \sum_{m=0}^{j} \alpha_{m,j} \int_{-\infty}^{\infty} \overline{p}_k^*(x|w)\overline{p}_m^*(x|w)w(x)\ dx = 0$$

and

$$\begin{aligned}\int_{-\infty}^{\infty} \overline{p}_k^*(x|w)\overline{p}_j(x|w)w(x)\ dx &= \sum_{m=0}^{k} \beta_{m,k} \int_{-\infty}^{\infty} \overline{p}_m(x|w)\overline{p}_j(x|w)w(x)\ dx \\ &= \beta_{j,k} \int_{-\infty}^{\infty} \overline{p}_j(x|w)^2 w(x)\ dx = \beta_{j,k};\end{aligned}$$

hence $\beta_{j,.k} = 0$ for $j < k$ and thus $\overline{p}_k^*(x|w) = \beta_{k,k}\overline{p}_k(x|w)$. Moreover, by normality,

$$1 = \int_{-\infty}^{\infty} \overline{p}_k^*(x|w)^2 w(x)\ dx = \beta_{k,k}^2 \int_{-\infty}^{\infty} \overline{p}_k(x|w)^2 w(x)\ dx = \beta_{k,k}^2,$$

so that $\overline{p}_k^*(x|w) = \pm\overline{p}_k(x|w)$. Consequently, $|\overline{p}_k(x|w)|$ is unique. Thus, we have the following theorem.

Theorem 1.10. *Any density function $w(x)$ on $\mathbb{R}$ satisfying the moment conditions* (1.12) *generates a unique sequence of orthonormal polynomials, up to signs. Consequently, the sequences a_k and b_k in the TTRR* (1.17) *are unique.*

1.4.2. Examples of Orthonormal Polynomials

1.4.2.1. Hermite Polynomials

If $w(x)$ is the density of the standard normal distribution,

$$w_{\mathcal{N}[0,1]}(x) = \exp\left(-x^2/2\right)/\sqrt{2\pi},$$

the orthonormal polynomials involved satisfy the TTRR

$$\sqrt{k+1}\overline{p}_{k+1}(x|w_{\mathcal{N}[0,1]}) - x.\overline{p}_k(x|w_{\mathcal{N}[0,1]}) + \sqrt{k}\overline{p}_{k-1}(x|w_{\mathcal{N}[0,1]}) = 0, \qquad x \in \mathbb{R},$$

for $k \in \mathbb{N}$, starting from $\overline{p}_0(x|w_{\mathcal{N}[0,1]}) = 1$, $\overline{p}_1(x|w_{\mathcal{N}[0,1]}) = x$. These polynomials are known as Hermite[13] polynomials.

The Hermite polynomials are plotted in Figure 1.1, for orders $k = 2, 5, 8$.

1.4.2.2. Laguerre Polynomials

The standard exponential density function

$$w_{\text{Exp}}(x) = \mathbf{1}(x \geq 0)\exp(-x) \qquad (1.18)$$

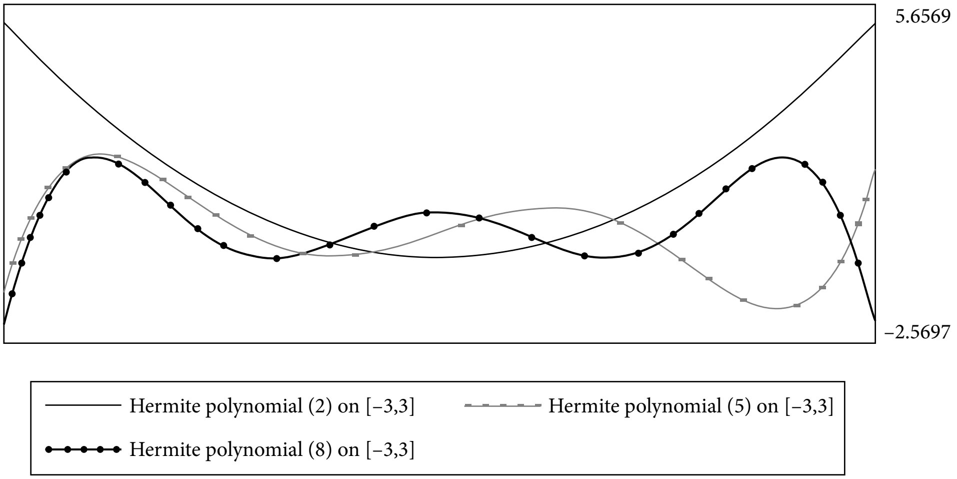


FIGURE 1.1 Hermite polynomials.

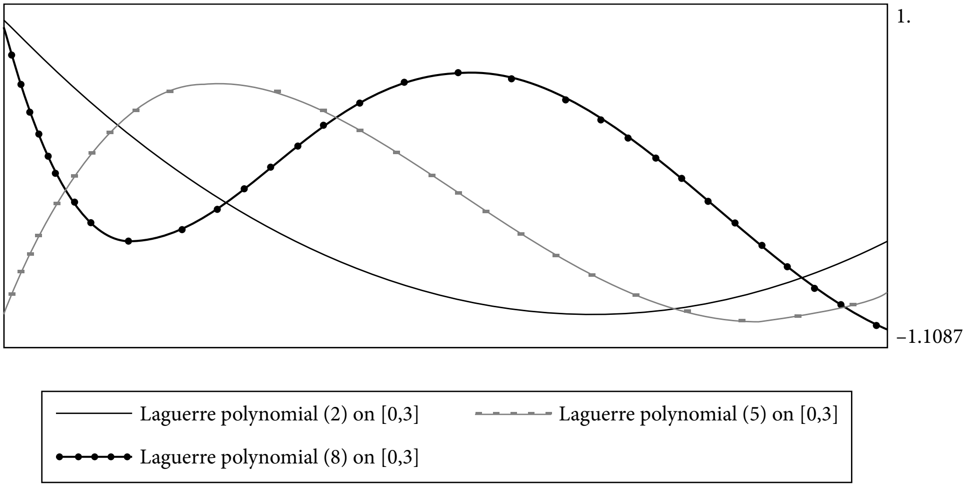


FIGURE 1.2 Laguerre polynomials.

gives rise to the orthonormal Laguerre[14] polynomials, with TTRR

$$(k+1)\,\overline{p}_{k+1}(x|w_{\mathrm{Exp}})+(2k+1-x)\,\overline{p}_{k}(x|w_{\mathrm{Exp}})+k.\overline{p}_{k-1}(x|w_{\mathrm{Exp}})=0,\ x\in[0,\infty).$$

for $k\in\mathbb{N}$, starting from $\overline{p}_0(x|w_{\mathrm{Exp}})=1$, $\overline{p}_1(x|w_{\mathrm{Exp}})=x-1$.

These polynomials are plotted in Figure 1.2, for orders $k=2,5,8$.

1.4.2.3. Legendre Polynomials

The uniform density on $[-1,1]$,

$$w_{\mathcal{U}[-1,1]}(x)=\tfrac{1}{2}\mathbf{1}(|x|\leq 1),$$

generates the orthonormal Legendre[15] polynomials on $[-1,1]$, with TTRR

$$\frac{k+1}{\sqrt{2k+3}\sqrt{2k+1}}\overline{p}_{k+1}(x|w_{\mathcal{U}[-1,1]})-x\cdot\overline{p}_{k}(x|w_{\mathcal{U}[-1,1]})$$
$$+\frac{k}{\sqrt{2k+1}\sqrt{2k-1}}\overline{p}_{k-1}(x|w_{\mathcal{U}[-1,1]})=0,\qquad |x|\leq 1,$$

for $k\in\mathbb{N}$, starting from $\overline{p}_0(x|w_{\mathcal{U}[-1,1]})=1$, $\overline{p}_1(x|w_{\mathcal{U}[-1,1]})=\sqrt{3}x$.

Moreover, substituting $x=2u-1$, it is easy to verify that the uniform density

$$w_{\mathcal{U}[0,1]}(u)=\mathbf{1}(0\leq u\leq 1)$$

on $[0,1]$ generates the orthonormal polynomials

$$\overline{p}_k(u|w_{\mathcal{U}[0,1]})=\overline{p}_k(2u-1|w_{\mathcal{U}[-1,1]}),$$

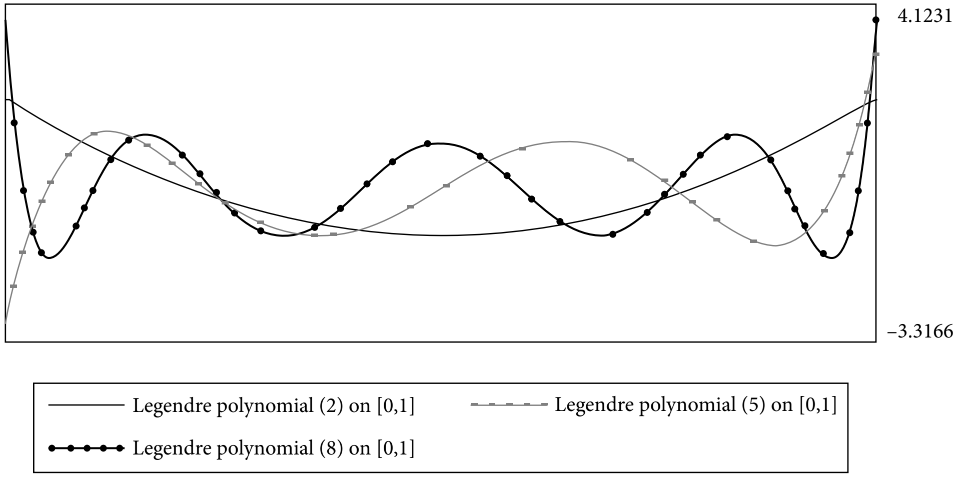


FIGURE 1.3 Shifted Legendre polynomials.

which are known as the shifted Legendre polynomials, also called the Legendre polynomials on the unit interval. The TTRR involved is

$$\frac{(k+1)/2}{\sqrt{2k+3}\sqrt{2k+1}}\overline{p}_{k+1}(u|w_{\mathcal{U}[0,1]})+(0.5-u)\cdot\overline{p}_k(u|w_{\mathcal{U}[0,1]})$$
$$+\frac{k/2}{\sqrt{2k+1}\sqrt{2k-1}}\overline{p}_{k-1}(u|w_{\mathcal{U}[0,1]})=0, \qquad 0\le u\le 1,$$

for $k\in\mathbb{N}$, starting from $\overline{p}_0(u|w_{\mathcal{U}[0,1]})=1$, $\overline{p}_1(u|w_{\mathcal{U}[0,1]})=\sqrt{3}(2u-1)$.

The latter Legendre polynomials are plotted in Figure 1.3, for orders $k=2,5,8$.

1.4.2.4. Chebyshev Polynomials

Chebyshev polynomials are generated by the density function

$$w_{\mathcal{C}[-1,1]}(x)=\frac{1}{\pi\sqrt{1-x^2}}\mathbf{1}(|x|<1), \tag{1.19}$$

with corresponding distribution function

$$W_{\mathcal{C}[-1,1]}(x)=1-\pi^{-1}\arccos(x), \qquad x\in[-1,1]. \tag{1.20}$$

The orthogonal (but not orthonormal) Chebyshev polynomials $p_k(x|w_{\mathcal{C}[-1,1]})$ satisfy the TTRR

$$p_{k+1}(x|w_{\mathcal{C}[-1,1]})-2xp_k(x|w_{\mathcal{C}[-1,1]})+p_{k-1}(x|w_{\mathcal{C}[-1,1]})=0, \qquad |x|<1, \tag{1.21}$$

for $k \in \mathbb{N}$, starting from $p_0(x|w_{\mathcal{C}[-1,1]}) = 1$, $p_1(x|w_{\mathcal{C}[-1,1]}) = x$, with orthogonality properties

$$\int_{-1}^{1} \frac{p_k(x|w_{\mathcal{C}[-1,1]})p_m(x|w_{\mathcal{C}[-1,1]})}{\pi\sqrt{1-x^2}}\,dx = \begin{cases} 0 & \text{if} \quad k \neq m, \\ 1/2 & \text{if} \quad k = m \in \mathbb{N}, \\ 1 & \text{if} \quad k = m = 0. \end{cases}$$

An important practical difference with the other polynomials discussed so far is that Chebyshev polynomials have the closed form:

$$p_k(x|w_{\mathcal{C}[-1,1]}) = \cos\left(k \cdot \arccos(x)\right). \tag{1.22}$$

To see this, observe from (1.20) and the well-known sine–cosine formulas that

$$\begin{aligned} &\int_{-1}^{1} \frac{\cos(k \cdot \arccos(x))\cos(m \cdot \arccos(x))}{\pi\sqrt{1-x^2}}\,dx \\ &\quad = -\frac{1}{\pi}\int_{-1}^{1} \cos(k \cdot \arccos(x))\cos(m \cdot \arccos(x))\,d\arccos(x) \\ &\quad = \frac{1}{\pi}\int_{0}^{\pi} \cos(k \cdot \theta)\cos(m \cdot \theta)\,d\theta = \begin{cases} 0 & \text{if} \quad k \neq m, \\ 1/2 & \text{if} \quad k = m \in \mathbb{N}, \\ 1 & \text{if} \quad k = m = 0. \end{cases} \end{aligned}$$

Moreover, it follows from the easy equality $\cos((k+1)\theta) - 2\cos(\theta)\cos(k \cdot \theta) + \cos((k-1)\theta) = 0$ that the functions (1.22) satisfy the TTRR (1.21) and are therefore genuine polynomials, and so are the orthonormal Chebyshev polynomials

$$\overline{p}_k(x|w_{\mathcal{C}[-1,1]}) = \begin{cases} 1 & \text{for} \quad k = 0, \\ \sqrt{2}\cos(k \cdot \arccos(x)) & \text{for} \quad k \in \mathbb{N}. \end{cases}$$

Substituting $x = 2u - 1$ for $u \in [0,1]$ in (1.20) yields

$$W_{\mathcal{C}[0,1]}(u) = 1 - \pi^{-1}\arccos(2u-1) \tag{1.23}$$

with density function

$$w_{\mathcal{C}[0,1]}(u) = \frac{1}{\pi\sqrt{u(1-u)}} \tag{1.24}$$

and shifted orthonormal Chebyshev polynomials

$$\overline{p}_k(u|w_{\mathcal{C}[0,1]}) = \begin{cases} 1 & \text{for} \quad k = 0, \\ \sqrt{2}\cos(k \cdot \arccos(2u-1)) & \text{for} \quad k \in \mathbb{N}. \end{cases} \tag{1.25}$$

The polynomials (1.25) are plotted in Figure 1.4, for orders $k = 2, 5, 8$.

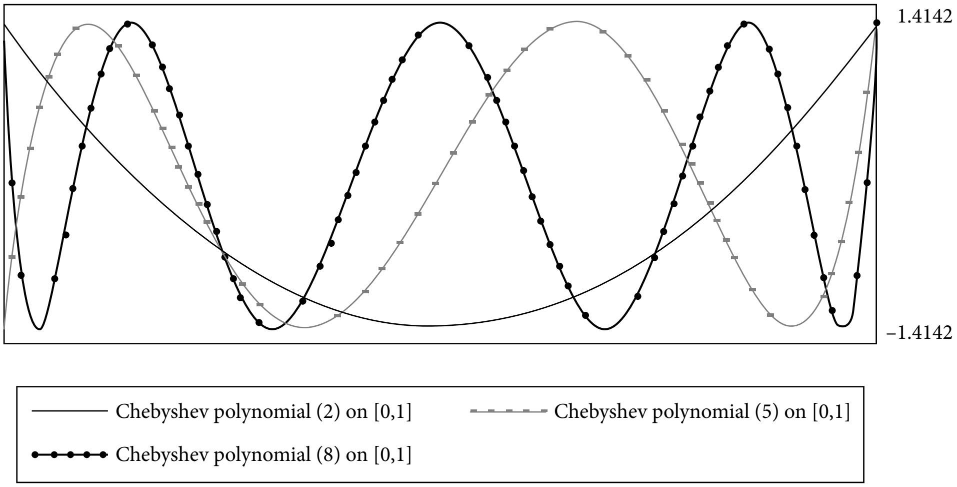


FIGURE 1.4 Shifted Chebyshev polynomials.

1.4.3. Completeness

The reason for considering orthonormal polynomials is the following.

Theorem 1.11. *Let $w(x)$ be a density function on $\mathbb{R}$ satisfying the moment conditions (1.12). Then the orthonormal polynomials $\overline{p}_k(x|w)$ generated by w form a complete orthonormal sequence in the Hilbert space $L^2(w)$. In particular, for any function $f \in L^2(w)$ and with X a random drawing from w,*

$$f(X) = \sum_{k=0}^{\infty} \gamma_k \overline{p}_k(X|w) \qquad \text{a.s.,} \tag{1.26}$$

where $\gamma_k = \int_{-\infty}^{\infty} \overline{p}_m(x|w) f(x) w(x)\, dx$ with $\sum_{k=0}^{\infty} \gamma_k^2 = \int_{-\infty}^{\infty} f(x)^2 w(x)\, dx$.

Proof. Let $f_n(x) = \sum_{m=0}^{n} \gamma_m \overline{p}_m(x|w)$. Then $||f - f_n||^2 = ||f||^2 - \sum_{m=0}^{n} \gamma_m^2$, which is not hard to verify, hence $\sum_{m=0}^{\infty} \gamma_m^2 \leq ||f||^2 < \infty$ and thus $\lim_{n\to\infty} \sum_{m=n+1}^{\infty} \gamma_m^2 = 0$. The latter implies that f_n is a Cauchy sequence in $L^2(w)$, with limit $\overline{f} \in$ span $(\{\overline{p}_m(\cdot\,|w)\}_{m=0}^{\infty}) \subset L^2(w)$. Thus, $\lim_{n\to\infty} ||\overline{f} - f_n|| = 0$.

To prove the completeness of the sequence $\overline{p}_m(\cdot\,|w)$, we need to show that $||\overline{f} - f|| = 0$, because then $f \in \text{span}(\{\overline{p}_m(\cdot\,|w)\}_{m=0}^{\infty})$, which by the arbitrariness of $f \in L^2(w)$ implies that $L^2(w) = \text{span}(\{\overline{p}_m(\cdot\,|w)\}_{m=0}^{\infty})$. This will be done by showing that for a random drawing X from $w(x)$, we obtain

$$\Pr[\overline{f}(X) = f(X)] = 1, \tag{1.27}$$

because then $||\overline{f}-f||^2 = E[(\overline{f}(X)-f(X))^2] = 0$. In turn, (1.27) is true if for all $t \in \mathbb{R}$, we obtain

$$E[(\overline{f}(X)-f(X))\exp(i\cdot t\cdot X)] = 0, \tag{1.28}$$

because of the uniqueness of the Fourier transform.[16]

To prove (1.28), note first that the limit function $\overline{f}$ can be written as $\overline{f}(x) = \sum_{m=0}^{n}\gamma_m\overline{p}_m(x|w)+\varepsilon_n(x)$, where $\lim_{n\to\infty}\int_{-\infty}^{\infty}\varepsilon_n(x)^2w(x)\ dx = 0$. Therefore,

$$\begin{aligned}&\left|\int_{-\infty}^{\infty}(\overline{f}(x)-f(x))\overline{p}_m(x|w)w(x)\ dx\right| = \left|\int_{-\infty}^{\infty}\varepsilon_n(x)\overline{p}_m(x|w)w(x)\ dx\right|\\ &\le\sqrt{\int_{-\infty}^{\infty}\varepsilon_n(x)^2w(x)\ dx}\sqrt{\int_{-\infty}^{\infty}\overline{p}_m(x|w)^2w(x)\ dx}=\sqrt{\int_{-\infty}^{\infty}\varepsilon_n(x)^2w(x)\ dx}\\ &\to 0\end{aligned}$$

for $n\to\infty$, which implies that for any $g\in\text{span}(\{\overline{p}_m(\cdot|w)\}_{m=0}^{\infty})$, we have $\int_{-\infty}^{\infty}(\overline{f}(x)-f(x))g(x)w(x)\ dx = 0$. Consequently, $E[(\overline{f}(X)-f(X))\exp(i\cdot t\cdot X)] = \int_{-\infty}^{\infty}(\overline{f}(x)-f(x))\exp(i\cdot t\cdot x)w(x)\ dx = 0$ for all $t\in\mathbb{R}$, because it follows from the well-known series expansions of $\cos(t\cdot x) = \text{Re}[\exp(i\cdot t\cdot x)]$ and $\sin(t\cdot x) = \text{Im}[\exp(i\cdot t\cdot x)]$ that these functions are elements of $\text{span}(\{\overline{p}_m(\cdot|w)\}_{m=0}^{\infty})$. Thus, $\{\overline{p}_m(\cdot|w)\}_{m=0}^{\infty}$ is complete in $L^2(w)$. The result (1.26) now follows from Theorem 1.9. ■

1.4.4. Application to the SNP Index Regression Model

Suppose that the response function $f(x)$ in the index regression model (1.1) satisfies

$$\sup_x|f(x)|\cdot\exp(-t_0\cdot|x|) = M(t_0) < \infty \qquad \text{for some } t_0 > 0. \tag{1.29}$$

so that $-M(t_0)\exp(t_0\cdot|x|)\le f(x)\le M(t_0)\exp(t_0\cdot|x|)$. Then for the standard normal density $w_{\mathcal{N}[0,1]}(x)$, we have $\int_{-\infty}^{\infty}f(x)^2w_{\mathcal{N}[0,1]}(x)\ dx < 2M(t_0)\exp(t_0^2/2) < \infty$; hence $f\in L^2(w_{\mathcal{N}[0,1]})$, so that $f(x)$ has the Hermite series expansion

$$f(x) = \sum_{m=0}^{\infty}\delta_{0,m}\overline{p}_m(x|w_{\mathcal{N}[0,1]}) = \delta_{0,0}+\delta_{0,1}x+\sum_{k=2}^{\infty}\delta_{0,k}\overline{p}_k(x|w_{\mathcal{N}[0,1]}) \qquad \text{a.e. on } \mathbb{R},$$

with $\delta_{0,m} = \int_{-\infty}^{\infty}f(x)\overline{p}_m(x|w_{\mathcal{N}[0,1]})w_{\mathcal{N}[0,1]}(x)\ dx$ for $m = 0,1,2,\ldots$. Thus, model (1.1) now reads

$$E[Y|X] = \lim_{n\to\infty}f_n(X_1+\theta_0'X_2|\boldsymbol{\delta}_n^0) \qquad \text{a.s,} \tag{1.30}$$

where

$$f_n(x|\boldsymbol{\delta}_n) = \delta_0+\delta_1x+\sum_{k=2}^{n}\delta_k\overline{p}_k(x|w_{\mathcal{N}[0,1]}) \tag{1.31}$$

with $\boldsymbol{\delta}_n = (\delta_0,\delta_1,\ldots,\delta_n,0,0,0,\ldots)$ and $\boldsymbol{\delta}_n^0 = (\delta_{0,0},\delta_{0,1},\ldots,\delta_{0,n},0,0,0,\ldots)$.

For fixed $n \in \mathbb{N}$ the parameters involved can be approximated by weighted nonlinear regression of Y on $f_n(X_1 + \theta' X_2 | \boldsymbol{\delta}_n)$, given a random sample $\{(Y_j, X_j)\}_{j=1}^{N}$from (Y, X) and given predefined compact parameter spaces Δ_n and Θ for $\boldsymbol{\delta}_n^0$ and θ_0, respectively. Then the weighted NLLS sieve estimator of $(\theta_0, \boldsymbol{\delta}_n^0)$ is

$$\left(\widehat{\theta}_n, \widehat{\boldsymbol{\delta}}_n\right) = \arg \min_{(\theta, \boldsymbol{\delta}_n) \in \Theta \times \Delta_n} \frac{1}{N} \sum_{j=1}^{N} \left(Y_j - f_n(X_{1,j} + \theta' X_{2,j} | \boldsymbol{\delta}_n)\right)^2 K(||X_j||), \tag{1.32}$$

where $K(x)$ is a positive weight function on $(0, \infty)$ satisfying $\sup_{x>0} x^n K(x) < \infty$ for all $n \geq 0$. The reason for this weight function is to guarantee that

$$E\left[\sup_{(\theta, \boldsymbol{\delta}_n) \in \Theta \times \Delta_n} \left(Y - f_n(X_1 + \theta' X_2 | \boldsymbol{\delta}_n)\right)^2 K(||X||)\right] < \infty$$

without requiring that $E[||X||^{2n}] < \infty$. Then by Jennrich's (1969) uniform law of large numbers and for fixed n, we have

$$\sup_{(\theta, \boldsymbol{\delta}_n) \in \Theta \times \Delta_n} \left| \frac{1}{N} \sum_{j=1}^{N} \left(Y_j - f_n(X_{1,j} + \theta' X_{2,j} | \boldsymbol{\delta}_n)\right)^2 K(||X_j||) - g_n(\theta, \boldsymbol{\delta}_n) \right|$$
$$= o_p(1),$$

where $g_n(\theta, \boldsymbol{\delta}_n) = E[(Y - f_n(X_1 + \theta' X_2 | \boldsymbol{\delta}_n))^2 K(||X||)]$, so that

$$p \lim_{N \to \infty} \left(\widehat{\theta}_n, \widehat{\boldsymbol{\delta}}_n\right) = \left(\overline{\theta}_n, \overline{\boldsymbol{\delta}}_n\right) = \arg \min_{(\theta, \boldsymbol{\delta}_n) \in \Theta \times \Delta_n} g_n(\theta, \boldsymbol{\delta}_n).$$

Under some alternative conditions the same result can be obtained by using the Wald (1949) consistency result in van der Vaart (1998, Theorem 5.14), which does not require that the expectation of the objective function is finite for all values of the parameters, so that in that case there is no need for the weight function $K(x)$.

Note that, in general, $\overline{\theta}_n \neq \theta_0$. Nevertheless, it can be shown that under some additional regularity conditions,[17] and with $n = n_N$ an arbitrary subsequence of N, $p\lim_{N \to \infty} \widehat{\theta}_{n_N} = \theta_0$ and $p\lim_{N \to \infty} \int_{-\infty}^{\infty} (f_{n_N}(x | \widehat{\boldsymbol{\delta}}_{n_N}) - f(x))^2 w_{\mathcal{N}[0,1]}(x)\, dx = 0$.

1.5. Non-Polynomial Complete Orthonormal Sequences

Recall that the support of a density $w(x)$ on $\mathbb{R}$ is defined as the set $\{x \in \mathbb{R} : w(x) > 0\}$. For example, the support of the standard exponential density (1.18) is the interval $[0, \infty)$. In this chapter I will only consider densities $w(x)$ with connected support—that is, the support is an interval—and for notational convenience this support will be denoted by an *open* interval (a, b), where $a = \inf_{w(x)>0} x \geq -\infty$ and $b = \sup_{w(x)>0} x \leq \infty$, even if for finite a and/or b, $w(a) > 0$ or $w(b) > 0$.

1.5.1. Nonpolynomial Sequences Derived from Polynomials

For every density $w(x)$ with support (a,b), $\int_a^b f(x)^2\,dx < \infty$ implies that $f(x)/\sqrt{w(x)} \in L^2(w)$. Therefore, the following corollary of Theorem 1.11 holds trivially.

Theorem 1.12. *Every function $f \in L^2(a,b)$ can be written as*

$$f(x) = \sqrt{w(x)}\left(\sum_{k=0}^{\infty}\gamma_k\overline{p}_k(x|w)\right) \qquad \textit{a.e. on } (a,b),$$

where w is a density with support (a,b) satisfying the moment conditions (1.12) and $\gamma_k = \int_a^b f(x)\overline{p}_k(x|w)\sqrt{w(x)}\,dx$. Consequently, $L^2(a,b)$ is a Hilbert space with complete orthonormal sequence $\psi_k(x|w) = \overline{p}_k(x|w)\sqrt{w(x)}$, $k \in \mathbb{N}_0$.

If (a,b) is bounded, then there is another way to construct a complete orthonormal sequence in $L^2(a,b)$, as follows. Let $W(x)$ be the distribution function of a density w with bounded support (a,b). Then $G(x) = a + (b-a)W(x)$ is a one-to-one mapping of (a,b) onto (a,b), with inverse $G^{-1}(y) = W^{-1}((y-a)/(b-a))$, where W^{-1} is the inverse of $W(x)$. For every $f \in L^2(a,b)$, we have

$$(b-a)\int_a^b f(G(x))^2\,w(x)\,dx = \int_a^b f(G(x))^2\,dG(x) = \int_a^b f(x)^2\,dx < \infty.$$

Hence $f(G(x)) \in L^2(w)$ and thus by Theorem 1.11 we have $f(G(x)) = \sum_{k=0}^{\infty}\gamma_k\overline{p}_k(x|w)$ a.e. on (a,b), where $\gamma_k = \int_a^b f(G(x))\overline{p}_k(x|w)w(x)\,dx$. Consequently,

$$f(x) = f\left(G\left(G^{-1}(x)\right)\right) = \sum_{k=0}^{\infty}\gamma_k\overline{p}_k\left(G^{-1}(x)\,|w\right) \qquad \text{a.e. on } (a,b).$$

Note that $dG^{-1}(x)/dx = dG^{-1}(x)/dG(G^{-1}(x)) = 1/G'(G^{-1}(x))$, so that

$$\begin{aligned}
&\int_a^b \overline{p}_k\left(G^{-1}(x)\,|w\right)\overline{p}_m\left(G^{-1}(x)\,|w\right)dx \\
&= \int_a^b \overline{p}_k\left(G^{-1}(x)\,|w\right)\overline{p}_m\left(G^{-1}(x)\,|w\right)G'\left(G^{-1}(x)\right)dG^{-1}(x) \\
&= \int_a^b \overline{p}_k(x|w)\,\overline{p}_m(x|w)\,G'(x)\,dx \\
&= (b-a)\int_a^b \overline{p}_k(x|w)\,\overline{p}_m(x|w)\,w(x)\,dx = (b-a)\,\mathbf{1}(k=m)\,.
\end{aligned}$$

Thus, we have the following theorem.

Theorem 1.13. *Let w be a density with bounded support (a, b) and let W be the c.d.f. of w, with inverse W^{-1}. Then the functions*

$$\psi_k(x|w) = \overline{p}_k\left(W^{-1}((x-a)/(b-a))\,|w\right)/\sqrt{(b-a)}, \qquad k \in \mathbb{N}_0$$

form a complete orthonormal sequence in $L^2(a, b)$. Hence, every function $f \in L^2(a, b)$ has the series representation $f(x) = \sum_{k=0}^{\infty} \gamma_k \psi_k(x|w)$ a.e. on (a, b), with $\gamma_k = \int_a^b \psi_k(x|w) f(x)\, dx$.

1.5.2. Trigonometric Sequences

Let us specialize the result in Theorem 1.13 to the case of the Chebyshev polynomials on $[0, 1]$, with $a = 0$, $b = 1$ and W, w and $\overline{p}_k(u|w)$ given by (1.23), (1.24), and (1.25), respectively. Observe that in this case $W_{C[0,1]}^{-1}(u) = (1 - \cos(\pi u))/2$. It follows now straightforwardly from (1.25) and the easy equality $\arccos(-x) = \pi - \arccos(x)$ that for $k \in \mathbb{N}$, $\overline{p}_k(W_{C[0,1]}^{-1}(u)|w_{C[0,1]}) = \sqrt{2}\cos(k\pi)\cos(k\pi u) = \sqrt{2}(-1)^k \cos(k\pi u)$, which by Theorem 1.13 implies the following.

Theorem 1.14. *The cosine sequence*

$$\psi_k(u) = \begin{cases} 1 & \text{for} \quad k = 0, \\ \sqrt{2}\cos(k\pi u) & \text{for} \quad k \in \mathbb{N}, \end{cases}$$

is a complete orthonormal sequence in $L^2(0, 1)$. Hence, every function $f \in L^2(0, 1)$ has the series representation $f(u) = \gamma_0 + \sum_{k=1}^{\infty} \gamma_k \sqrt{2}\cos(k\pi u)$ a.e. on $(0, 1)$, with $\gamma_0 = \int_0^1 f(u)\, du$, $\gamma_k = \sqrt{2}\int_0^1 \cos(k\pi u) f(u)\, du$ for $k \in \mathbb{N}$.

This result is related to classical Fourier analysis. Consider the following sequence of functions on $[-1, 1]$:

$$\begin{aligned} &\varphi_0(x) = 1, \\ &\varphi_{2k-1}(x) = \sqrt{2}\sin(k\pi x), \; \varphi_{2k}(x) = \sqrt{2}\cos(k\pi x), \qquad k \in \mathbb{N}. \end{aligned} \tag{1.33}$$

These functions are know as the Fourier series on $[-1, 1]$. It is easy to verify that these functions are orthonormal with respect to the uniform density $w_{\mathcal{U}[-1,1]}(x) = \frac{1}{2}\mathbf{1}(|x| \leq 1)$ on $[-1, 1]$, that is, $\frac{1}{2}\int_{-1}^{1} \varphi_m(x)\varphi_k(x)\, dx = \mathbf{1}(m = k)$. The following theorem is a classical Fourier analysis result.

Theorem 1.15. *The Fourier sequence $\{\varphi_n\}_{n=0}^{\infty}$ is complete in $L^2(w_{\mathcal{U}[-1,1]})$.*[18]

Now Theorem 1.14 is a corollary of Theorem 1.15. To see this, let $f \in L^2(0, 1)$ be arbitrary. Then $g(x) = f(|x|) \in L^2(w_{\mathcal{U}[-1,1]})$; hence

$$g(x) = \alpha + \sum_{k=1}^{\infty} \beta_k \sqrt{2}\cos(k\pi x) + \sum_{m=1}^{\infty} \gamma_m \sqrt{2}\sin(k\pi x)$$

a.e. on $[-1,1]$, where

$$\alpha = \frac{1}{2}\int_{-1}^{1} g(x)\ dx = \int_{0}^{1} f(u)\ du$$

$$\beta_k = \frac{1}{2}\int_{-1}^{1} g(x)\sqrt{2}\cos(k\pi x)\ dx = \int_{0}^{1} f(u)\sqrt{2}\cos(k\pi u)\ du$$

$$\gamma_m = \frac{1}{2}\int_{-1}^{1} g(x)\sqrt{2}\sin(k\pi x)\ dx = 0$$

so that $f(u) = \alpha + \sum_{k=1}^{\infty} \beta_k \sqrt{2}\cos(k\pi u)$ a.e. on $[0,1]$.

Similarly, given an arbitrary $f \in L^2(0,1)$, let $g(x) = (\mathbf{1}(x \geq 0) - \mathbf{1}(x < 0))f(|x|)$. Then $g(x) = \sum_{m=1}^{\infty} \gamma_m \sqrt{2}\sin(k\pi x)$ a.e. on $[-1,1]$; hence $f(u) = \sum_{m=1}^{\infty} \gamma_m \sqrt{2}\sin(k\pi u)$ a.e. on $(0,1)$, where $\gamma_m = \int_0^1 f(u)\sqrt{2}\sin(m\pi u)\ du$. Therefore, we have the following corollary.

Corollary 1.1. *The sine sequence* $\sqrt{2}\sin(m\pi u)$, $m \in \mathbb{N}$, *is complete in* $L^2(0,1)$.

Although this result implies that for every $f \in L^2(0,1)$, $\lim_{n\to\infty} f_n(u) = f(u)$ a.e. on $(0,1)$, where $f_n(u) = \sum_{m=1}^{n} \gamma_m \sqrt{2}\sin(k\pi u)$ with $\gamma_m = \sqrt{2}\int_0^1 f(u)\sin(m\pi u)\ du$, the approximation $f_n(u)$ may be very poor in the tails of $f(u)$ if $f(0) \neq 0$ and $f(1) \neq 0$, because, in general, $\lim_{u\downarrow 0}\lim_{n\to\infty} f_n(u) \neq \lim_{n\to\infty}\lim_{u\downarrow 0} f_n(u)$, and similarly for $u \uparrow 1$. Therefore, the result of Corollary 1.1 is of limited practical significance.

1.6. Density and Distribution Functions

1.6.1. General Univariate SNP Density Functions

Let $w(x)$ be a density function with support (a,b). Then for any density $f(x)$ on (a,b), we obtain

$$g(x) = \sqrt{f(x)}/\sqrt{w(x)} \in L^2(w), \tag{1.34}$$

with $\int_a^b g(x)^2 w(x)\ dx = \int_a^b f(x)\ dx = 1$. Therefore, given a complete orthonormal sequence $\{\rho_m\}_{m=0}^{\infty}$ in $L^2(w)$ with $\rho_0(x) \equiv 1$ and denoting $\gamma_m = \int_a^b \rho_m(x) g(x) w(x)\ dx$, any density $f(x)$ on (a,b) can be written as

$$f(x) = w(x)\left(\sum_{m=0}^{\infty} \gamma_m \rho_m(x)\right)^2 \quad \text{a.e. on } (a,b), \qquad \text{with } \sum_{m=0}^{\infty} \gamma_m^2 = \int_a^b f(x)\ dx = 1. \tag{1.35}$$

The reason for the square in (1.35) is to guarantee that $f(x)$ is non-negative.

A problem with the series representation (1.35) is that in general the parameters involved are not unique. To see this, note that if we replace the function $g(x)$ in (1.34) by $g_B(x) = (\mathbf{1}(x \in B) - \mathbf{1}(x \in (a,b)\backslash B))\sqrt{f(x)}/\sqrt{w(x)}$, where B is an arbitrary Borel set, then $g_B(x) \in L^2(w)$ and $\int_a^b g_B(x)^2 w(x)\ dx = \int_a^b f(x)\ dx = 1$, so that (1.35) also holds for the sequence

$$\begin{aligned}\gamma_m &= \int_a^b \rho_m(x) g_B(x) w(x)\ dx \\ &= \int_{(a,b)\cap B} \rho_m(x)\sqrt{f(x)}\sqrt{w(x)}\ dx - \int_{(a,b)\backslash B} \rho_m(x)\sqrt{f(x)}\sqrt{w(x)}\ dx.\end{aligned}$$

In particular, using the fact that $\rho_0(x) \equiv 1$, we obtain

$$\gamma_0 = \int_{(a,b)\cap B} \sqrt{f(x)}\sqrt{w(x)}\ dx - \int_{(a,b)\backslash B} \sqrt{f(x)}\sqrt{w(x)}\ dx,$$

so that the sequence γ_m in (1.35) is unique if γ_0 is maximal. In any case we may without loss of generality assume that $\gamma_0 \in (0,1)$, so that without loss of generality the γ_m's can be reparameterized as

$$\gamma_0 = \frac{1}{\sqrt{1+\sum_{k=1}^\infty \delta_k^2}}, \qquad \gamma_m = \frac{\delta_m}{\sqrt{1+\sum_{k=1}^\infty \delta_k^2}},$$

where $\sum_{k=1}^\infty \delta_k^2 < \infty$. This reparameterization does not solve the lack of uniqueness problem, of course, but is convenient in enforcing the restriction $\sum_{m=0}^\infty \gamma_m^2 = 1$.

On the other hand, under certain conditions on $f(x)$ the δ_m's are unique, as will be shown in Section 1.6.4.

Summarizing, the following result has been shown.

Theorem 1.16. *Let $w(x)$ be a univariate density function with support (a,b), and let $\{\rho_m\}_{m=0}^\infty$ be a complete orthonormal sequence in $L^2(w)$, with $\rho_0(x) \equiv 1$. Then for any density $f(x)$ on (a,b) there exist possibly uncountably many sequences $\{\delta_m\}_{m=1}^\infty$ satisfying $\sum_{m=1}^\infty \delta_m^2 < \infty$ such that*

$$f(x) = \frac{w(x)\left(1+\sum_{m=1}^\infty \delta_m \rho_m(x)\right)^2}{1+\sum_{m=1}^\infty \delta_m^2} \qquad \text{a.e. on } (a,b). \tag{1.36}$$

Moreover, for the sequence $\{\delta_m\}_{m=1}^\infty$ for which $\sum_{m=1}^\infty \delta_m^2$ is minimal,

$$\sqrt{f(x)} = \frac{\sqrt{w(x)}\left(1+\sum_{m=1}^\infty \delta_m \rho_m(x)\right)}{\sqrt{1+\sum_{m=1}^\infty \delta_m^2}} \qquad \text{a.e. on } (a,b);$$

hence

$$\delta_m = \frac{\int_a^b \rho_m(x)\sqrt{f(x)}\sqrt{w(x)}\ dx}{\int_a^b \sqrt{f(x)}\sqrt{w(x)}\ dx}, \qquad m \in \mathbb{N}. \tag{1.37}$$

In practice, the result of Theorem 1.16 cannot be used directly in SNP modeling, because it is impossible to estimate infinitely many parameters. Therefore, the density (1.36) is usually approximated by

$$f_n(x) = \frac{w(x)\left(1+\sum_{m=1}^{n}\delta_m\rho_m(x)\right)^2}{1+\sum_{m=1}^{n}\delta_m^2} \tag{1.38}$$

for some natural number n, possibly converging to infinity with the sample size. Following Gallant and Nychka (1987), I will refer to truncated densities of the type (1.38) as *SNP densities.*

Obviously,

Corollary 1.2. *Under the conditions of Theorem 1.16,* $\lim_{n\to\infty} f_n(x) = f(x)$ *a.e. on* (a,b)*. Moreover, it is not hard to verify that*

$$\int_a^b |f(x)-f_n(x)|\, dx \le 4\sqrt{\sum_{m=n+1}^{\infty}\delta_m^2} + 2\sum_{m=n+1}^{\infty}\delta_m^2 = o(1), \tag{1.39}$$

where the δ_m*'s are given by* (1.37)*, so that with* $F(x)$ *the c.d.f. of* $f(x)$ *and* $F_n(x)$ *the c.d.f. of* $f_n(x)$*, we obtain*

$$\lim_{n\to\infty}\sup_x |F(x)-F_n(x)| = 0.$$

Remarks

1. The rate of convergence to zero of the tail sum $\sum_{m=n+1}^{\infty}\delta_m^2$ depends on the smoothness, or the lack thereof, of the density $f(x)$. Therefore, the question of how to choose the truncation order n given an *a priori* chosen approximation error cannot be answered in general.
2. In the case that the $\rho_m(x)$'s are polynomials, the SNP density $f_n(x)$ has to be computed recursively via the corresponding TTRR (1.17), except in the case of Chebyshev polynomials, but that is not much of a computational burden. However, the computation of the corresponding SNP distribution function $F_n(x)$ is more complicated. See, for example, Stewart (2004) for SNP distribution functions on $\mathbb{R}$ based on Hermite polynomials, and see Bierens (2008) for SNP distribution functions on $[0,1]$ based on Legendre polynomials. Both cases require to recover the coefficients $\ell_{m,k}$ of the polynomials $\overline{p}_k(x|w) = \sum_{m=0}^{k}\ell_{m,k}x^m$, which can be done using the TTRR involved. Then with $P_n(x|w) = (1,\overline{p}_1(x|w),\ldots,\overline{p}_n(x|w))'$, $Q_n(x) = (1,x,\ldots,x^n)'$, $\delta = (1,\delta_1,\ldots,\delta_n)$, and L_n the lower-triangular matrix consisting of the coefficients $\ell_{m,k}$, we can write $f_n(x) = (\delta'\delta)^{-1}w(x)(\delta' P_n(x|w))^2 = (\delta'\delta)^{-1}\delta' L_n Q_n(x)Q_n(x)' w(x)L_n'\delta$; hence

$$F_n(x) = \frac{1}{\delta'\delta}\delta' L_n\left(\int_{-\infty}^{x} Q_n(z)Q_n(z)'w(z)\, dz\right)L_n'\delta = \frac{\delta' L_n M_n(x)L_n'\delta}{\delta'\delta},$$

where $M_n(x)$ is the $(n+1)\times(n+1)$ matrix with typical elements $\int_{-\infty}^{x} z^{i+j}w(z)\,dz$ for $i,j = 0,1,\dots,n$. This is the approach proposed by Bierens (2008). The approach in Stewart (2004) is in essence the same and is therefore equally cumbersome.

1.6.2. Bivariate SNP Density Functions

Now let $w_1(x)$ and $w_2(y)$ be a pair of density functions on $\mathbb{R}$ with supports (a_1,b_1) and (a_2,b_2), respectively, and let $\{\rho_{1,m}\}_{m=0}^{\infty}$ and $\{\rho_{2,m}\}_{m=0}^{\infty}$ be complete orthonormal sequences in $L^2(w_1)$ and $L^2(w_2)$, respectively. Moreover, let $g(x,y)$ be a Borel measurable real function on $(a_1,b_1)\times(a_2,b_2)$ satisfying

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} g(x,y)^2 w_1(x)w_2(y)\,dx\,dy < \infty. \tag{1.40}$$

The latter implies that $g_2(y) = \int_{a_1}^{b_1} g(x,y)^2 w_1(x)\ dx < \infty$ a.e. on (a_2,b_2), so that for each $y \in (a_2,b_2)$ for which $g_2(y) < \infty$ we have $g(x,y) \in L^2(w_1)$. Then $g(x,y) = \sum_{m=0}^{\infty}\gamma_m(y)\rho_{1,m}(x)$ a.e. on (a_1,b_1), where $\gamma_m(y) = \int_{a_1}^{b_1} g(x,y)\rho_{1,m}(x)w_1(x)\,dx$ with $\sum_{m=0}^{\infty}\gamma_m(y)^2 = \int_{a_1}^{b_1} g(x,y)^2\cdot w_1(x)\ dx = g_2(y)$. Because by (1.40) we have $\int_{a_2}^{b_2} g_2(y)w_2(y)\ dy < \infty$, it follows now that for each $y \in (a_2,b_2)$ for which $g_2(y) < \infty$ and all integers $m \geq 0$ we have $\gamma_m(y) \in L^2(w_2)$, so that $\gamma_m(y) = \sum_{k=0}^{\infty}\gamma_{m,k}\rho_{2,k}(y)$ a.e. on (a_2,b_2), where $\gamma_{m,k} = \int_{a_1}^{b_1}\int_{a_2}^{b_2} g(x,y)\rho_{1,m}(x)\rho_{2,k}(y)w_1(x)w_2(y)\,dxdy$ with $\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\gamma_{m,k}^2 < \infty$. Hence,

$$g(x,y) = \sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\gamma_{m,k}\rho_{1,m}(x)\rho_{2,k}(y) \qquad \text{a.e. on } (a_1,b_1)\times(a_2,b_2). \tag{1.41}$$

Therefore, it follows similar to Theorem 1.16 that the next theorem holds.

Theorem 1.17. *Given a pair of density functions $w_1(x)$ and $w_2(y)$ with supports (a_1,b_1) and (a_2,b_2), respectively, and given complete orthonormal sequences $\{\rho_{1,m}\}_{m=0}^{\infty}$ and $\{\rho_{2,m}\}_{m=0}^{\infty}$ in $L^2(w_1)$ and $L^2(w_2)$, respectively, with $\rho_{1,0}(x) = \rho_{2,0}(y) \equiv 1$, for every bivariate density $f(x,y)$ on $(a_1,b_1)\times(a_2,b_2)$ there exist possibly uncountably many double arrays $\delta_{m,k}$ satisfying $\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\delta_{m,k}^2 < \infty$, with $\delta_{0,0} = 1$ by normalization, such that a.e. on $(a_1,b_1)\times(a_2,b_2)$, we obtain*

$$f(x,y) = \frac{w_1(x)w_2(y)\left(\sum_{m=0}^{\infty}\sum_{k=0}^{\infty}\delta_{m,k}\rho_{1,m}(x)\rho_{2,k}(y)\right)^2}{\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\delta_{k,m}^2}.$$

For example, let $w_1(x)$ and $w_2(y)$ be standard normal densities and $\rho_{1,m}(x)$ and $\rho_{2,k}(y)$ Hermite polynomials, that is, $\rho_{1,k}(x) = \rho_{2,k}(x) = \overline{p}_k(x|w_{\mathcal{N}[0,1]})$. Then for any density function $f(x,y)$ on $\mathbb{R}^2$ there exists a double array $\delta_{m,k}$ and associated sequence of SNP densities

$$f_n(x,y) = \frac{\exp(-(x^2+y^2)/2)}{2\pi\sum_{k=0}^{n}\sum_{m=0}^{n}\delta_{k,m}^2}\left(\sum_{m=0}^{n}\sum_{k=0}^{n}\delta_{m,k}\overline{p}_m(x|w_{\mathcal{N}[0,1]})\overline{p}_k(y|w_{\mathcal{N}[0,1]})\right)^2$$

such that $\lim_{n\to\infty} f_n(x,y) = f(x,y)$ a.e. on $\mathbb{R}^2$.

This result is used by Gallant and Nychka (1987) to approximate the bivariate error density of the latent variable equations in Heckman's (1979) sample selection model.

1.6.3. SNP Densities and Distribution Functions on $[0,1]$

Since the seminal paper by Gallant and Nychka (1987), SNP modeling of density and distribution functions on $\mathbb{R}$ via the Hermite expansion has become the standard approach in econometrics, despite the computational burden of computing the SNP distribution function involved.

However, there is an easy trick to avoid this computational burden, by mapping one-to-one any absolutely continuous distribution function $F(x)$ on (a,b) with density $f(x)$ to an absolutely continuous distribution function $H(u)$ with density $h(u)$ on the unit interval, as follows. Let $G(x)$ be an *a priori* chosen absolutely continuous distribution function with density $g(x)$ and support (a,b). Then we can write

$$F(x) = H(G(x)) \quad \text{and} \quad f(x) = h(G(x))\cdot g(x), \tag{1.42}$$

where

$$H(u) = F(G^{-1}(u)) \quad \text{and} \quad h(u) = f(G^{-1}(u))/g(G^{-1}(u)) \tag{1.43}$$

with $G^{-1}(u)$ the inverse of $G(x)$.

For example, let $(a,b) = \mathbb{R}$ and choose for $G(x)$ the logistic distribution function $G(x) = 1/(1+\exp(-x))$. Then $g(x) = G(x)(1-G(x))$ and $G^{-1}(u) = \ln(u/(1-u))$, hence $h(u) = f(\ln(u/(1-u)))/(u(1-u))$. Similarly, if $(a,b) = (0,\infty)$ and $G(x) = 1-\exp(-x)$, then any density $f(x)$ on $(0,\infty)$ corresponds uniquely to $h(u) = f(\ln(1/(1-u)))/(1-u)$.

The reason for this transformation is that there exist closed-form expressions for SNP densities on the unit interval and their distribution functions. In particular, Theorem 1.18 follows from (1.23)–(1.25) and Corollary 1.2.

Theorem 1.18. *For every density $h(u)$ on $[0,1]$ with corresponding c.d.f. $H(u)$ there exist possibly uncountably many sequences $\{\delta_m\}_{m=1}^{\infty}$ satisfying $\sum_{m=1}^{\infty}\delta_m^2 < \infty$ such that $h(u) = \lim_{n\to\infty} h_n(u)$ a.e. on $[0,1]$, where*

$$h_n(u) = \frac{1}{\pi\sqrt{u(1-u)}}\frac{\left(1+\sum_{m=1}^{n}(-1)^m\delta_m\sqrt{2}\cos(m\cdot\arccos(2u-1))\right)^2}{1+\sum_{m=1}^{n}\delta_m^2}, \tag{1.44}$$

and $\lim_{n\to\infty}\sup_{0\leq u\leq 1}|\underline{H}_n(1-\pi^{-1}\arccos(2u-1))-H(u)|=0$, *where*

$$\underline{H}_n(u)=u+\frac{1}{1+\sum_{m=1}^{n}\delta_m^2}\left[2\sqrt{2}\sum_{k=1}^{n}\delta_k\frac{\sin(k\pi u)}{k\pi}+\sum_{m=1}^{n}\delta_m^2\frac{\sin(2m\pi u)}{2m\pi}\right.$$
$$+2\sum_{k=2}^{n}\sum_{m=1}^{k-1}\delta_k\delta_m\frac{\sin((k+m)\pi u)}{(k+m)\pi}$$
$$\left.+2\sum_{k=2}^{n}\sum_{m=1}^{k-1}\delta_k\delta_m\frac{\sin((k-m)\pi u)}{(k-m)\pi}\right]. \quad (1.45)$$

Moreover, with $w(x)$ being the uniform density on $[0,1]$ and $\rho_m(x)$ being the cosine sequence, it follows from Corollary 1.2 that the next theorem holds.

Theorem 1.19. *For every density $h(u)$ on $[0,1]$ with corresponding c.d.f. $H(u)$ there exist possibly uncountably many sequences $\{\delta_m\}_{m=1}^{\infty}$ satisfying $\sum_{m=1}^{\infty}\delta_m^2<\infty$ such that a.e. on $[0,1]$, $h(u)=\lim_{n\to\infty}h_n(u)$, where*

$$h_n(u)=\frac{\left(1+\sum_{m=1}^{n}\delta_m\sqrt{2}\cos(m\pi u)\right)^2}{1+\sum_{m=1}^{n}\delta_m^2}, \quad (1.46)$$

and $\lim_{n\to\infty}\sup_{0\leq u\leq 1}|\underline{H}_n(u)-H(u)|=0$, *where* $\underline{H}_n(u)$ *is defined by* (1.45).

The latter follows straightforwardly from (1.46) and the well-known equality $\cos(a)\cos(b)=(\cos(a+b)+\cos(a-b))/2$, and the same applies to the result for $H(u)$ in Theorem 1.18.

1.6.4. Uniqueness of the Series Representation

The density $h(u)$ in Theorem 1.19 can be written as $h(u)=\eta(u)^2/\int_0^1\eta(v)^2\,dv$, where

$$\eta(u)=1+\sum_{m=1}^{\infty}\delta_m\sqrt{2}\cos(m\pi u) \qquad \text{a.e. on } (0,1). \quad (1.47)$$

Moreover, recall that in general we have

$$\delta_m=\frac{\int_0^1(\mathbf{1}(u\in B)-\mathbf{1}(u\in[0,1]\backslash B))\sqrt{2}\cos(m\pi u)\sqrt{h(u)}\,du}{\int_0^1(\mathbf{1}(u\in B)-\mathbf{1}(u\in[0,1]\backslash B))\sqrt{h(u)}\,du},$$

$$\frac{1}{\sqrt{1+\sum_{m=1}^{\infty}\delta_m^2}}=\int_0^1(\mathbf{1}(u\in B)-\mathbf{1}(u\in[0,1]\backslash B))\sqrt{h(u)}\,du.$$

for some Borel set B satisfying $\int_0^1 (\mathbf{1}(u \in B) - \mathbf{1}(u \in [0,1]\backslash B))\sqrt{h(u)}\, du > 0$; hence

$$\eta(u) = (\mathbf{1}(u \in B) - \mathbf{1}(u \in [0,1]\backslash B))\sqrt{h(u)}\sqrt{1 + \sum_{m=1}^{\infty} \delta_m^2} \tag{1.48}$$

Similarly, given this Borel set B and the corresponding δ_m's, the SNP density (1.46) can be written as $h_n(u) = \eta_n(u)^2 / \int_0^1 \eta_n(v)^2 dv$, where

$$\begin{aligned} \eta_n(u) &= 1 + \sum_{m=1}^{n} \delta_m \sqrt{2} \cos(m\pi u) \\ &= (\mathbf{1}(u \in B) - \mathbf{1}(u \in [0,1]\backslash B))\sqrt{h_n(u)}\sqrt{1 + \sum_{m=1}^{n} \delta_m^2}. \end{aligned} \tag{1.49}$$

Now suppose that $h(u)$ is continuous and positive on $(0,1)$. Moreover, let $S \subset [0,1]$ be the set with Lebesgue measure zero on which $h(u) = \lim_{n\to\infty} h_n(u)$ fails to hold. Then for any $u_0 \in (0,1)\backslash S$ we have $\lim_{n\to\infty} h_n(u_0) = h(u_0) > 0$; hence for sufficient large n we have $h_n(u_0) > 0$. Because obviously $h_n(u)$ and $\eta_n(u)$ are continuous on $(0,1)$, for such an n there exists a small $\varepsilon_n(u_0) > 0$ such that $h_n(u) > 0$ for all $u \in (u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0,1)$, and therefore

$$\mathbf{1}(u \in B) - \mathbf{1}(u \in [0,1]\backslash B) = \frac{\eta_n(u)}{\sqrt{h_n(u)}\sqrt{1 + \sum_{m=1}^{n} \delta_m^2}} \tag{1.50}$$

is continuous on $(u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0,1)$. Substituting (1.50) in (1.48), it follows now that $\eta(u)$ is continuous on $(u_0 - \varepsilon_n(u_0), u_0 + \varepsilon_n(u_0)) \cap (0,1)$; hence by the arbitrariness of $u_0 \in (0,1)/S$, $\eta(u)$ is continuous on $(0,1)$.

Next, suppose that $\eta(u)$ takes positive and negative values on $(0,1)$. Then by the continuity of $\eta(u)$ on $(0,1)$ there exists a $u_0 \in (0,1)$ for which $\eta(u_0) = 0$ and thus $h(u_0) = 0$, which, however, is excluded by the condition that $h(u) > 0$ on $(0,1)$. Therefore, either $\eta(u) > 0$ for all $u \in (0,1)$ or $\eta(u) < 0$ for all $u \in (0,1)$. However, the latter is excluded because by (1.47) we have $\int_0^1 \eta(u)\, du = 1$. Thus, $\eta(u) > 0$ on $(0,1)$, so that by (1.48), $\mathbf{1}(u \in B) - \mathbf{1}(u \in [0,1]\backslash B) = 1$ on $(0,1)$.

Consequently, we have the following theorem.

Theorem 1.20. *For every continuous density $h(u)$ on $(0,1)$ with support $(0,1)$ the sequence $\{\delta_m\}_{m=1}^{\infty}$ in Theorem 1.19 is unique, with*

$$\delta_m = \frac{\int_0^1 \sqrt{2}\cos(m\pi u)\sqrt{h(u)}\, du}{\int_0^1 \sqrt{h(u)}\, du}.$$

As is easy to verify, the same argument applies to the more general densities considered in Theorem 1.16:

Theorem 1.21. *Let the conditions of Theorem 1.16 be satisfied. In addition, let the density $w(x)$ and the orthonormal functions $\rho_m(x)$ be continuous on (a,b).*[19] *Then every continuous and positive density $f(x)$ on (a,b) has a unique series representation (1.36), with*

$$\delta_m = \frac{\int_a^b \rho_m(x)\sqrt{w(x)}\sqrt{f(x)}\,dx}{\int_0^1 \sqrt{w(x)}\sqrt{f(x)}\,dx}.$$

Moreover, note that Theorem 1.18 is a special case of Theorem 1.16. Therefore, the following corollary holds.

Corollary 1.3. *For every continuous and positive density $h(u)$ on $(0,1)$ the δ_m's in Theorem 1.18 are unique and given by*

$$\delta_m = (-1)^m \frac{\int_0^1 \sqrt{2}\cos(m\cdot\arccos(2u-1))\,(u(1-u))^{-1/4}\sqrt{h(u)}\,du}{\int_0^1 (u(1-u))^{-1/4}\sqrt{h(u)}\,du}.$$

1.6.5. Application to the MPH Competing Risks Model

Note that the distribution (1.6) in the MPH competing risks Weibull model (1.5) has density

$$h(u) = \int_0^\infty v^2 u^{v-1}\,dG(v),$$

which is obviously positive and continuous on $(0,1)$. However, if $G(1) > 0$, then $h(0) = \infty$; and if $E[V^2] = \int_0^\infty v^2\,dG(v) = \infty$, then $h(1) = \infty$. To allow for these possibilities, the series representation in Theorem 1.18 on the basis of Chebyshev polynomials seems an appropriate way of modeling $H(u)$ semi-nonparametrically, because then $h_n(0) = h_n(1) = \infty$ if $1 + \sqrt{2}\sum_{m=1}^\infty(-1)^m\delta_m \neq 0$ and $1 + \sqrt{2}\sum_{m=1}^\infty \delta_m \neq 0$. However, the approach in Theorem 1.19 is asymptotically applicable as well, because the condition $\sum_{m=1}^\infty \delta_m^2 < \infty$ does not preclude the possibilities that $\sum_{m=1}^\infty \delta_m = \infty$ and/or $\sum_{m=1}^\infty(-1)^m\delta_m = \infty$, which imply that $\lim_{n\to\infty} h_n(0) = \lim_{n\to\infty} h_n(1) = \infty$.

As said before, the actual log-likelihood in Bierens and Carvalho (2007) is more complicated than displayed in (1.7), due to right-censoring. In their case the log-likelihood involves the distribution function $H(u) = \int_0^\infty u^v\,dG(v)$ next to its density $h(u) = \int_0^\infty vu^{v-1}\,dG(v)$, where $h(1) = \int_0^\infty v\,dG(v) = 1$ due to the condition $E[V] = 1$. Note also that $G(1) > 0$ implies $h(0) = \infty$. Bierens and Carvalho (2007) use a series representation of $h(u)$ in terms of Legendre polynomials with SNP density $h_n(u)$ satisfying the restriction $h_n(1) = 1$. However, as argued in Section 1.6.1, the computation of the corresponding SNP distribution function $H_n(u)$ is complicated.

Due to the restriction $h_n(1) = 1$, the approach in Theorem 1.18 is not applicable as an alternative to the Legendre polynomial representation of $h(u) = \int_0^\infty vu^{v-1}\,dG(v)$, whereas the approach in Theorem 1.19 does not allow for $h_n(0) = \infty$. On the other hand, Bierens and Carvalho (2007) could have used $H_n(u) = \underline{H}_n(\sqrt{u})$, for example, where $\underline{H}_n$ is defined by (1.45), with density

$$h_n(u) = \frac{\left(1 + \sum_{m=1}^{n} \delta_m \sqrt{2} \cos\left(m\pi\sqrt{u}\right)\right)^2}{2\left(1 + \sum_{m=1}^{n} \delta_m^2\right)\sqrt{u}}$$

and δ_1 chosen such that

$$1 = \frac{\left(1 + \sqrt{2}\sum_{m=1}^{n} (-1)^m \delta_m\right)^2}{2\left(1 + \sum_{m=1}^{n} \delta_m^2\right)} \qquad (1.51)$$

to enforce the restriction $h_n(1) = 1$.

1.6.6. Application to the First-Price Auction Model

In the first-price auction model, the value distribution $F(v)$ is defined on $(0,\infty)$, so at first sight a series expansion of the value density $f(v)$ in terms of Laguerre polynomials seems appropriate. However, any distribution function $F(v)$ on $(0,\infty)$ can be written as $F(v) = H(G(v))$, where $G(v)$ is an *a priori* chosen absolutely continuous distribution function with support $(0,\infty)$, so that $H(u) = F(G^{-1}(u))$ with density $h(u) = f((G^{-1}(u))/g(G^{-1}(u))$, where G^{-1} and g are the inverse and density of G, respectively. For example, choose $G(v) = 1 - \exp(-v)$, so that $g(v) = \exp(-v)$ and $G^{-1}(u) = \ln(1/(1-u))$.

The equilibrium bid function (1.8) can now be written as

$$\beta(v|H) = v - \frac{\int_{p_0}^{v} H(G(x))^{I-1}\, dx}{H(G(v))^{I-1}}, \qquad v \geq p_0. \qquad (1.52)$$

Bierens and Song (2012) use the SNP approximation of $H(u)$ on the basis of Legendre polynomials, but using the results in Theorem 1.19 would have been much more convenient. In any case the integral in (1.52) has to be computed numerically.

Similarly, the conditional value distribution $F(v\exp(-\theta'X))$ in Bierens and Song (2013) can be written as $H(G(v\exp(-\theta'X)))$, where now H is modeled semi-nonparametrically according the results in Theorem 1.19. In this case the number of potential bidders $I = I(X)$ and the reservation price $p_0 = p_0(X)$ also depend on the auction-specific covariates X; but as shown in Bierens and Song (2013), $I(X)$ can be estimated nonparametrically and therefore may be treated as being observable, whereas $p_0(X)$ is directly observable. Then in the binding reservation price case the auction-specific equilibrium bid function becomes

$$\beta(v|H,\theta,X) = v - \frac{\int_{p_0(X)}^{v} H(G(x\cdot\exp(-\theta'X)))^{I(X)-1}\, dx}{H(G(v\exp(-\theta'X)))^{I(X)-1}}, \qquad v \geq p_0(X). \qquad (1.53)$$

1.7. A Brief Review of Sieve Estimation

Recall from (1.30)–(1.32) that in the SNP index regression case the objective function takes the form

$$\widehat{Q}_N(\theta,\boldsymbol{\delta}_\infty)=\frac{1}{N}\sum_{j=1}^{N}\left(Y_j-\sum_{m=0}^{\infty}\delta_m\overline{p}_m(X_{1,j}+\theta' X_{2,j}|w_{\mathcal{N}[0,1]})\right)^2 K(||X_j||),$$

where $\boldsymbol{\delta}_\infty=(\delta_1,\delta_2,\delta_3,\dots)\in\mathbb{R}^\infty$ satisfies $\sum_{m=0}^{\infty}\delta_m^2<\infty$, with true parameters θ_0 and $\boldsymbol{\delta}_\infty^0=(\delta_{0,1},\delta_{0,2},\delta_{0,3},\dots)$ satisfying

$$(\theta_0,\boldsymbol{\delta}_\infty^0)=\arg\min_{\theta,\boldsymbol{\delta}_\infty}\overline{Q}(\theta,\boldsymbol{\delta}_\infty) \tag{1.54}$$

subject to $\sum_{m=0}^{\infty}\delta_m^2<\infty$, where $\overline{Q}(\theta,\boldsymbol{\delta}_\infty)=E[\widehat{Q}_N(\theta,\boldsymbol{\delta}_\infty)]$.

Similarly, in the MPH competing risk model with $H(u)$ modeled semi-nonparametrically as, for example, $H(\sqrt{u}|\boldsymbol{\delta}_\infty)=\lim_{n\to\infty}\underline{H}_n(\sqrt{u})$ with $\underline{H}_n$ defined by (1.45), and subject to the restriction (1.51), the objective function is

$$\widehat{Q}_N(\theta,\boldsymbol{\delta}_\infty)=-\frac{1}{N}\ln(L_N(\alpha_1,\alpha_2,\beta_1,\beta_2,H(\sqrt{u}|\boldsymbol{\delta}_\infty))),$$
$$\theta=\left(\alpha_1',\alpha_2',\beta_1',\beta_2'\right)'.$$

with true parameters given by (1.54) with $\overline{Q}(\theta,\boldsymbol{\delta}_\infty)=E[\widehat{Q}_N(\theta,\boldsymbol{\delta}_\infty)]$.

In the first-price auction model with auction-specific covariates the function $\overline{Q}(\theta,\boldsymbol{\delta}_\infty)$ is the probability limit of the objective function $\widehat{Q}_N(\theta,\boldsymbol{\delta}_\infty)$ involved rather than the expectation. See Bierens and Song (2013).

Now let Θ be a compact parameter space for θ_0, and for each $n\geq 1$, let Δ_n be a compact space of nuisance parameters $\boldsymbol{\delta}_n=(\delta_1,\delta_2,\delta_3,\dots,\delta_n,0,0,0,\dots)$, endowed with metric $d(.,.)$, such that $\boldsymbol{\delta}_n^0=(\delta_{0,1},\delta_{0,2},\delta_{0,3},\dots,\delta_{0,n},0,0,0,\dots)\in\Delta_n$. Note that $\boldsymbol{\delta}_\infty^0\in\overline{\cup_{n=1}^{\infty}\Delta_n}$, where the bar denotes the closure.

The sieve estimator of $(\theta_0,\boldsymbol{\delta}_\infty^0)$ is defined as

$$\left(\widehat{\theta}_n,\widehat{\boldsymbol{\delta}}_n\right)=\arg\min_{(\theta,\boldsymbol{\delta}_n)\in\Theta\times\Delta_n}\widehat{Q}_N(\theta,\boldsymbol{\delta}_n).$$

Under some regularity conditions it can be shown that for an arbitrary subsequence n_N of the sample size N we obtain

$$p\lim_{N\to\infty}||\widehat{\theta}_{n_N}-\theta_0||=0 \quad\text{and}\quad p\lim_{N\to\infty}d\left(\widehat{\boldsymbol{\delta}}_{n_N},\boldsymbol{\delta}_\infty^0\right)=0.$$

Moreover, under further regularity conditions the subsequence n_N can be chosen such that

$$\sqrt{N}(\widehat{\theta}_{n_N}-\theta_0)\xrightarrow{d}N[0,\Sigma].$$

See Shen (1997), Chen (2007), and Bierens (2013). As shown in Bierens (2013), the asymptotic variance matrix Σ can be estimated consistently by treating n_N as constant and then estimating the asymptotic variance matrix involved in the standard parametric way.

Note that Bierens and Carvalho (2007) assume that $\boldsymbol{\delta}_\infty^0 \in \cup_{n=1}^\infty \Delta_n$, so that for some n we have $\boldsymbol{\delta}_\infty^0 = \boldsymbol{\delta}_n^0 \in \Delta_n$. This is quite common in empirical applications because then the model is fully parametric, albeit with unknown dimension of the parameter space. See, for example, Gabler et al. (1993). The minimal order n in this case can be estimated consistently via an information criterion, such as the Hannan–Quinn (1979) and Schwarz (1978) information criteria. Asymptotically, the estimated order $\widehat{n}_N$ may then be treated as the true order, so that the consistency and asymptotic normality of the parameter estimates can be established in the standard parametric way.

In the case $\boldsymbol{\delta}_\infty^0 \in \overline{\cup_{n=1}^\infty \Delta_n} \backslash \cup_{n=1}^\infty \Delta_n$ the estimated sieve order $\widehat{n}_N$ via these information criteria will converge to ∞. Nevertheless, using $\widehat{n}_N$ in this case may preserve consistency of the sieve estimators, as in Bierens and Song (2012, Theorem 4), but whether asymptotic normality is also preserved is an open question.

1.8. Concluding Remarks

Admittedly, this discussion of the sieve estimation approach is very brief and incomplete. However, the main focus of this chapter is on SNP modeling. A full review of the sieve estimation approach is beyond the scope and size limitation of this chapter. Besides, a recent complete review has already been done by Chen (2007).

This chapter is part of the much wider area of *approximation theory.* The reader may wish to consult some textbooks on the latter—for example, Cheney (1982), Lorentz (1986), Powell (1981), and Rivlin (1981).

Notes

1. Of course, there are many more examples of SNP models.
2. See, for example, Bierens (2004, Theorem 3.10, p. 77).
3. See (1.41) below.
4. Note that due to the presence of scale parameters in the Weibull baseline hazards (1.3), the condition $E[V] = 1$ is merely a normalization of the condition that $E[V] < \infty$.
5. That is, $\{v : f(v) > 0\}$ is an interval.
6. See, for example, Bierens (2004, Theorem 7.A.1, p. 200).
7. Here the bar denotes the closure.
8. See, for example, Bierens (2004, Theorem 7.A.5, p. 202) for a proof of the projection theorem.

9. See, for example, Bierens (2004, Theorem 7.A.2., p. 200). The latter result is confined to the Hilbert space of zero-mean random variables with finite second moments, but its proof can easily be adapted to $\mathcal{R}$.
10. The existence of such a complete orthonormal sequence will be shown in the next section.
11. See, for example, Bierens (2004, Theorem 6.B.3, p. 168).
12. See, for example, Bierens (2004, Theorem 2.B.2, p. 168).
13. Charles Hermite (1822–1901).
14. Edmund Nicolas Laguerre (1834–1886).
15. Adrien-Marie Legendre (1752–1833).
16. See, for example, Bierens (1994, Theorem 3.1.1, p. 50).
17. See Section 1.7.
18. See, for example, Young (1988, Chapter 5).
19. The latter is the case if we choose $\rho_m(x) = \overline{p}(x|w)$.

References

Abbring, Jaap H., and Gerard J. van den Berg. 2003. "The Identifiability of the Mixed Proportional Hazards Competing Risks Model." *Journal of the Royal Statistical Society B*, **65**, pp. 701–710.

Bierens, Herman J. 1994. *Topics in Advanced Econometrics*: *Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge, UK: Cambridge University Press.

Bierens, Herman J. 2004. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge, UK: Cambridge University Press.

Bierens, Herman J. 2008. "Semi-Nonparametric Interval-Censored Mixed Proportional Hazard Models: Identification and Consistency Results." *Econometric Theory*, **24**, pp. 749–794.

Bierens, Herman J. 2013. "Consistency and Asymptotic Normality of ML Sieve Estimators Under Low-Level Conditions." Forthcoming in *Econometric Theory*. (http://econ.la.psu.edu/hbierens/SNPMODELS.PDF).

Bierens, Herman J., and Hosin Song. 2012. "Semi-Nonparametric Estimation of Independently and Identically Repeated First-Price Auctions via an Integrated Simulated Moments Method." *Journal of Econometrics*, **168**, pp. 108–119.

Bierens, Herman J., and Hosin Song. 2013. "Semi-Nonparametric Modeling and Estimation of First-Price Auction Models with Auction-Specific Heterogeneity." Working paper (http://econ.la.psu.edu/hbierens/AUCTIONS_HETERO.PDF).

Bierens, Herman J., and Jose R. Carvalho. 2006. "Separate Appendix to: Semi-Nonparametric Competing Risks Analysis of Recidivism" (http://econ.la.psu.edu/~hbierens/RECIDIVISM_APP.PDF).

Bierens, Herman J., and Jose R. Carvalho. 2007. "Semi-Nonparametric Competing Risks Analysis of Recidivism." *Journal of Applied Econometrics*, **22**, pp. 971–993.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." Chapter 76 in *Handbook of Econometrics*, Vol. 6B, eds. James J. Heckman and Edward E. Leamer. Amsterdam: North-Holland.

Cheney, Elliott W. 1982. *Introduction to Approximation Theory*. Providence, RI: AMS Chelsea Publishers.

Eastwood, Brian J., and A. Ronald Gallant. 1991. "Adaptive Rules for Semi-Nonparametric Estimators that Achieve Asymptotic Normality." *Econometric Theory,* **7**, pp. 307–340.

Elbers, Chris, and Geert Ridder. 1982. "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model." *Review of Economic Studies,* **49**, pp. 403–409.

Gabler, Siegfried, Francois Laisney and Michael Lechner. 1993. "Seminonparametric Estimation of Binary-Choice Models with an Application to Labor-Force Participation." *Journal of Business & Economic Statistics,* **11**, pp. 61–80.

Gallant, A. Ronald. 1981. "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form: The Fourier Flexible Form." *Journal of Econometrics,* **15**, pp. 211–245.

Gallant, A. Ronald, and Douglas W. Nychka. 1987. "Semi-Nonparametric Maximum Likelihood Estimation." *Econometrica,* **55**, pp. 363–390.

Grenander, Ulf. 1981. *Abstract Inference.* New York: John Wiley.

Guerre, Emmanuel, Isabelle Perrigne, and Quang Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions." *Econometrica,* **68**, pp. 525–574.

Hahn, Jinyong. 1994. "The Efficiency Bound of the Mixed Proportional Hazard Model." *Review of Economic Studies,* **61**, pp. 607–629.

Hamming, Richard W. 1973. *Numerical Methods for Scientists and Engineers.* New York: Dover Publications.

Hannan, Edward J., and Barry G. Quinn. 1979. "The Determination of the Order of an Autoregression." *Journal of the Royal Statistical Society, Series B,* **41**, pp. 190–195.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica,* **47**, pp. 153–161.

Heckman, James J., and Bo E. Honore. 1989. "The Identifiability of the Competing Risks Model." *Biometrika,* **76**, pp. 325–330.

Heckman, James J., and Burton Singer. 1984. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica,* **52**, pp. 271–320.

Jennrich, Robert I. 1969. "Asymptotic Properties of Nonlinear Least Squares Estimators." *Annals of Mathematical Statistics,* **40**, pp. 633–643.

Krishna, Vijay. 2002. *Auction Theory.* San Diego: Academic Press.

Kronmal, R., and M. Tarter. 1968. "The Estimation of Densities and Cumulatives by Fourier Series Methods." *Journal of the American Statistical Association,* **63**, pp. 925–952.

Lancaster, Tony. 1979. "Econometric Methods for the Duration of Unemployment." *Econometrica,* **47**, pp. 939–956.

Lorentz, George G. 1986. *Approximation of Functions.* Providence, RI: AMS Chelsea Publishers.

Newey, Whitney K. 1990. "Semiparametric Efficiency Bounds." *Journal of Applied Econometrics,* **5**, pp. 99–135.

Newey, Whitney K. 1997. "Convergence Rates and Asymptotic Normality for Series Estimators." *Journal of Econometrics,* **79**, pp. 147–168.

Powell, Michael J. D. 1981. *Approximation Theory and Methods.* Cambridge, U.K.: Cambridge University Press.

Ridder, Geert, and Tiemen Woutersen. 2003. "The Singularity of the Efficiency Bound of the Mixed Proportional Hazard Model." *Econometrica,* **71**, pp. 1579–1589.

Riley, John G., and William F. Samuelson. 1981. "Optimal Auctions." *American Economic Review,* **71**, pp. 381–392.

Rivlin, Theodore J. 1981. *An Introduction to the Approximation of Functions*. New York: Dover Publications.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *Annals of Statistics,* **6**, pp. 461–464.

Shen, Xiaotong. 1997. "On the Method of Sieves and Penalization." *Annals of Statistics,* **25**, pp. 2555–2591.

Stewart, Mark B. 2004. "Semi-Nonparametric Estimation of Extended Ordered Probit Models." *The Stata Journal,* **4**, pp. 27–39.

van den Berg, Gerard. 2000. "Duration models: Specification, identification, and multiple duration." Chapter 55 in *Handbook of Econometrics*, Vol. 5, eds. James J. Heckman and Edward E. Leamer. Amsterdam: North-Holland.

van der Vaart, Aart W. 1998. *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wald, Abraham. 1949. "Note on the Consistency of the Maximum Likelihood Estimate." *Annals of Mathematical Statistics,* **20**, pp. 595–601.

Wold, Herman. 1938. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksell.

Young, Nicholas. 1988. *An Introduction to Hilbert Space*. Cambridge, UK: Cambridge University Press.

CHAPTER 2

AN OVERVIEW OF THE SPECIAL REGRESSOR METHOD

ARTHUR LEWBEL[†]

2.1. Introduction

The goal of this chapter is to provide some background for understanding how and why special regressor methods work, as well as provide their application to identification and estimation of latent variable moments and parameters. Other related surveys include that of Dong and Lewbel (2012), who describe the simplest estimators for applying special regressor methods to binary choice problems (particularly those involving endogenous regressors), and that of Lewbel, Dong, and Yang (2012), who provide a comparison of special regressor methods to other types of estimators, specifically, to control functions, maximum likelihood, and linear probability models for binary choice model estimation.

A special regressor is an observed covariate with properties that facilitate identification and estimation of a latent variable model. For example, suppose that an observed binary variable D satisfies $D = I(V + W^* \geq 0)$, where V is the observed special regressor and W^* is an unobserved latent variable. Such a W^* will exist as long as the probability that $D = 1$ is increasing in V. The goal is estimation of the distribution of W^*, or estimation of features of its distribution like a conditional or unconditional mean or median of W^*. Many standard models have this form; for example, a probit model has $W^* = X'\beta + \varepsilon$ with ε normal, and estimating β would correspond to estimating the conditional mean of W^* given X. A simple probit doesn't require a special regressor, but the special regressor would be useful here if ε is heteroskedastic with an unknown distribution, or if some or the regressors in X are endogenous. Special regressor methods work by exploiting the fact that if V is independent of W^* (either unconditionally or after conditioning on covariates), then variation in V changes the probability that $D = 1$ in a way that traces out the distribution of W^* (either the unconditional distribution or the distribution conditional on covariates).

The term special regressor was first used in Lewbel (1998), but the commonest class of applications for the special regressor is to binary choice models as described in Lewbel (2000). The special regressor method has been employed in a wide variety of limited dependent variable models, including binary, ordered, and multinomial choice as well as censored regression, selection, and treatment models (Lewbel, 1998, 2000, 2007a; Magnac and Maurin, 2007, 2008), truncated regression models (Khan and Lewbel, 2007), binary and other nonlinear panel models with fixed effects (Honore and Lewbel, 2002; Ai and Gan, 2010; Gayle, 2012), contingent valuation models (Lewbel, Linton, and McFadden, 2011), dynamic choice models (Heckman and Navarro, 2007; Abbring and Heckman, 2007), market equilibrium models of multinomial discrete choice (Berry and Haile, 2009a, 2009b), models of games, including entry games and matching games (Lewbel and Tang, 2011; Khan and Nekipelov, 2011; Fox and Yang, 2012), and a variety of models with (partly) nonseparable errors (Lewbel, 2007b; Matzkin, 2007; Briesch, Chintagunta, and Matzkin, 2010).

Additional empirical applications of special regressor methods include Anton, Fernandez Sainz, and Rodriguez-Poo (2002), Maurin (2002), Cogneau and Maurin (2002), Goux and Maurin (2005), Stewart (2005), Avelino (2006), Pistolesi (2006), Lewbel and Schennach (2007), and Tiwari, Mohnen, Palm, and van der Loeff (2007). Earlier results that can be reinterpreted as special cases of special regressor-based identification methods include Matzkin (1992, 1994) and Lewbel (1997). Vytlacil and Yildiz (2007) describe their estimator as a control function, but their identification of the endogenous regressor coefficient essentially treats the remainder of the latent index as a special regressor. Recent identification and limiting distribution theory involving special regressor models include Jacho-Chávez (2009), Khan and Tamer (2010), and Khan and Nekipelov (2010a, 2010b).

The remainder of this chapter lays out the basic ideas behind special regressor methods. The focus here is on identification and associated construction of estimators, not on limiting distribution theory. Most of the estimators provided here are multistep estimators, where each step takes the form of a standard parametric or nonparametric density or regression estimator.

The next section provides the basic idea of how a special regressor can identify the distribution of a latent variable, specifically the latent index W^* in a threshold crossing binary choice model $D = I(V + W^* \geq 0)$. This is followed by sections that look at estimation of unconditional moments of the latent index such as $E(W^*)$, as well as discuss estimation of conditional moments like $E(W^* \mid X)$, conditioning on covariates X. This will then lead to estimation of coefficients β in discrete choice models $D = I(V + X'\beta + \varepsilon \geq 0)$ when the latent error ε is heteroskedastic and has unknown distribution. Later sections will consider semiparametric instrumental variables estimators for these models when the regressors X are mismeasured or endogenous. The final sections consider some extensions, including allowing the special regressor V to be discrete, and also consider other latent variable models.

2.2. Identifying a Latent Marginal Distribution

To illustrate the main ideas behind special regressor identification and estimation, begin with an example from Lewbel, Linton, and McFadden (2011). Suppose we want to uncover the distribution of people's willingness to pay (wtp) W^* to preserve a wetland habitat. Denote this distribution function as $F_{W^*}(w)$. For simplicity in this exposition, assume that the distribution of W^* is continuously distributed.

A survey is performed in which, for each person in the sample, researchers draw a price P from some distribution function chosen by the researcher and ask the sampled individual if they would be willing to pay P dollars or more to preserve the wetland. Let D denote an individual's response ($D = 1$ for yes or $D = 0$ for no), so $D = I(W^* \geq P)$, where W^* is the individual's unobserved (latent) willingness to pay and I is the indicator function that equals one if its argument is true and zero otherwise. The experiment is designed so that prices are randomly assigned to individuals, meaning that the distribution that P is drawn from is independent of W^*. Let $E(D \mid P = p)$ denote the conditional expectation of D, conditioning on the event that the random variable P equals the value p. By construction of the experiment, P is drawn independently of each subject's willingness to pay, so P is distributed independently of W^*. It follows from this independence that $E(D \mid P = p) = \Pr(W^* \geq p) = 1 - \Pr(W^* < p) = 1 - F_{W^*}(p)$. For example, suppose that among a random sample of people, 70% said they would be *not* be willing to pay more than \$50 to preserve the wetland. In this example, $p = 50$ and so 0.7 would be an unbiased estimate of $1 - E(D \mid P = 50) = F_{W^*}(50)$.

In the data set examined by An (2000) and reconsidered by Lewbel, Linton, and McFadden (2011), P takes on one of 14 different values, so without additional assumptions, these data could identify the distribution function $F_{W^*}(w^*)$ only at $w^* = p$ for these 14 values of p. To identify the entire distribution function F_{W^*}, it would be desirable to design the experiment so that P can take on any value that the willingness-to-pay W^* could equal, meaning that each p should be drawn from a continuous distribution with support equal to an interval that is at least as large as the range of possible values of W^*.[1]

In this application, P is a special regressor that allows us to identify the distribution function $F_{W^*}(w^*)$ at points $w^* = p$ in the distribution of P. If the experiment were designed so that the distribution of P was continuous, then the entire distribution of W^* would be identified as the number of subjects goes to infinity and could be consistently estimated by a nonparametric regression of $1 - D$ on P, since $E(1 - D \mid P = p) = F_{W^*}(p)$.

To be consistent with later notation on latent variable models (e.g., the notation of a probit model), let $V = -P$, so $D = I(V + W^* \geq 0)$. Suppose P and hence V are continuously distributed, and let $F_V(v)$ denote the distribution function of the special regressor V. Let $H(v) = E(D \mid V = v)$ and let $\widehat{H}(v)$ be a uniformly consistent estimator of $H(v)$, such as a Nadaraya–Watson local constant or local linear kernel regression of D on V. Then, replacing p above with $-v$, we have $H(v) = 1 - F_{W^*}(-v)$, and therefore $1 - \widehat{H}(-w^*)$ is a consistent estimator of $F_{W^*}(w^*)$. If the support of V contains the support of $-W^*$, then the entire distribution function $F_{W^*}(w^*)$ would be identified and consistently estimated by $1 - \widehat{H}(-v)$ for values of v in the support of V.

In this example the special regressor V allows us to recover the distribution of the latent variable W^* in a limited dependent variable model. Suppose we wanted to estimate $E(W^*)$, the average willingness to pay in the population. Let w_L and w_U denote lower and upper bounds on the range of values that w^* can take on, and let f_{w^*} denote the probability density function of w^*. Then

$$\begin{aligned} E(W^*) &= \int_{w_L}^{w_U} w^* f_{w^*}(w^*)\, dw^* = \int_{w_L}^{w_U} w^* \frac{\partial F_{w^*}(w^*)}{\partial w^*}\, dw^* \\ &= \int_{w_L}^{w_U} w^* \frac{\partial [1 - H(-w^*)]}{\partial w^*}\, dw^*. \end{aligned} \tag{2.1}$$

This equation shows that $E(W^*)$ is identified and could be consistently estimated by replacing H with $\widehat{H}$ in the above integral. The next section will provide a much simpler estimator for $E(W^*)$.

The key features of the special regressor V that allow us to recover the distribution of the latent W^* are independence (later this will be relaxed to conditional independence), additivity, continuity, and large support. In this example, V is statistically independent of the latent W^*, V appears added to the latent W^* in the model $D = I(V + W^* \geq 0)$, V is continuously distributed, and the range of values that V takes on is at least as great as the range that $-W^*$ takes on. Having one regressor in a model that satisfies conditional independence, additivity, continuity, and large support is the typical set of requirements for special regressor methods, though these features can sometimes be relaxed in various directions.

Note in particular that large support is not required, but without it the distribution of W^* will only be identified at the values that $-V$ can take on. This may suffice for some applications. For example, if support for the wetland will be decided by a referendum, then by the median voter model the maximum amount of money that can be raised to protect the wetland will be obtained by having people vote on whether they are willing to pay the median of the distribution of W^*. So in this application we only need the range of $-V$ to include the median of W^*, which would be estimated as $-v$ for the value v that makes $\widehat{H}(v) = 1/2$. However, if we instead wish to estimate the mean of W^*, then large support will generally be required.[2]

These requirements for special regressors can be restrictive, but many discrete choice estimators make use of similar restrictions like these, though sometimes implicitly. For example, Manski's (1985) maximum score and Horowitz's (1992) smooth maximum score estimator papers assume the presence of a regressor with these properties. Except for large support, standard logit and probit models assume that all regressors satisfy these properties.

The above results can be generalized in a variety of directions. For example, suppose we had the more general model $Y = g(V + W^*)$ for some possibly unknown function g. Here V may now be thought of as just a regressor, not necessarily set by experimental design. As long as g is nonconstant and weakly monotonic, then the mean and all other features of the distribution of W^* can still be identified (up to an implicit location normalization) by first letting $D = I(Y \geq y_0)$ for any constant y_0 in the support of Y that makes D be nonconstant and then applying the same methods as above.

Even more generally, if Y depends on two or more latent variables, then their distribution can be generally identified if we have a separate special regressor for each latent variable.

To summarize, the key result derived in this section is that, in a latent index model of the form $D = I(V + W^* \geq 0)$, variation in the special regressor V can be used to identify and estimate the distribution function of the unobserved latent variable W^*, based on $F_{W^*}(w^*) = 1 - E(D \mid V = -w^*)$ for any constant w^* in the support of $-V$. This result more generally shows identification of $Y = g(V + W^*)$, where g is weakly monotonic.

2.3. Unconditional Moments

Identification and hence estimation of the mean willingness to pay $E(W^*)$ based on Eq. (2.1) is somewhat elaborate, since this equation involves integrating a function of the derivative of the conditional expectation function H. In this section we derive a simpler expression for $E(W^*)$. Let $f_v(v)$ denote the probability density function of V, which in the experimental design context is determined by, and hence known to, the researcher. Let c be any constant in the interior of the supports of $-W^*$. Define the variable T_1 by

$$T_1 = \frac{D - I(V \geq c)}{f_v(V)} - c. \tag{2.2}$$

Then it will be shown below that

$$E(T_1) = E(W^*). \tag{2.3}$$

Using Eq. (2.3), the mean of W^* can be consistently estimated just by constructing the variable $T_{1i} = [D_i - I(V_i \geq 0)] / f_v(V_i)$ for each individual i in a sample and then letting the estimate of $E(W^*)$ be the sample average $\sum_{i=1}^n T_{1i}/n$. Here we have made the simplest choice for the constant c of letting $c = 0$.

The derivation of $E(T_1) = E(W^*)$ has two main components. One component is to show that

$$E(W^*) = -c + \int_{w_L}^{w_U} [H(-w^*) - I(-w^* \geq c)]dw^*. \tag{2.4}$$

The crux of the argument behind Eq. (2.4) is that, by definition of an expectation $E(W^*) = \int w^* dF_{w^*}(w^*)$ and using integration by parts, this integral is equivalent to $-\int F_{w^*}(w^*)dw^*$ plus boundary terms. Since $F_{w^*}(w^*) = 1 - H(-w^*)$ we get $\int H(-w^*)dw^*$ plus boundary terms. The role of c and of $I(-w^* \geq c)$ in Eq. (2.4) is to handle these boundary terms.

The second component to proving Eq. (2.3) is to show that $E(T_1)$ equals the right-hand side of Eq. (2.4). This is shown by rewriting the integral in Eq. (2.4) as $\int_{-w_U}^{-w_L} [H(v) - I(v \geq c)]dv$ and plugging in $H(v) = E(D \mid V = v)$. The argument inside this integral can then be rewritten as $E(T_1 + c \mid V = v)f_v(v)$. Notice that the c inside the expectation will be canceled out with $-c$ before the integral in Eq. (2.4), so the right-hand side of Eq. (2.4) is equivalent to an expectation over V, giving $E(T_1)$.

The remainder of this section writes out these steps of the proof of Eq. (2.3) in detail. To ease exposition, technical conditions required for applying integration by parts, and for allowing the limits of integration w_L and w_U to be infinite, will be ignored here.

We begin by deriving Eq. (2.4), applying the following integration by parts argument.

$$\begin{aligned}
&-c + \int_{w_L}^{w_U} [H(-w^*) - I(-w^* \geq c)]dw^* \\
&\quad = -c - \int_{w_L}^{w_U} [1 - H(-w^*) - I(w^* > -c)]dw^* \\
&\quad = -c - \int_{w_L}^{w_U} [F_{w^*}(w^*) - I(w^* > -c)]\frac{\partial w^*}{\partial w^*}\, dw^* \\
&\quad = -c - (w^*[F_{w^*}(w^*) - I(w^* > -c)]|_{w_L}^{w_U}) + \int_{w_L}^{w_U} w^* \frac{\partial [F_{w^*}(w^*) - I(w^* > -c)]}{\partial w^*}\, dw^*.
\end{aligned}$$

To evaluate this expression, first consider the boundary related terms, noting that $F_{w^*}(w_L) = 0$ and $F_{w^*}(w_U) = 1$.

$$\begin{aligned}
&-c - (w^*[F_{w^*}(w^*) - I(w^* > -c)]|_{w_L}^{w_U}) = -c - (w^* F_{w^*}(w^*)|_{w_L}^{-c}) \\
&\qquad - (w^*[F_{w^*}(w^*) - 1]|_{-c}^{w_U}) \\
&\quad = -c - (-cF_{w^*}(-c) - w_L F_{w^*}(w_L)) - (w_U[F_{w^*}(w_U) - 1] - (-c)[F_{w^*}(-c) - 1]) \\
&\quad = -c - (-cF_{w^*}(-c)) + (-c)[F_{w^*}(-c) - 1] = c - c = 0.
\end{aligned}$$

Observe that the role of the term $I(w^* > -c)$ is to simplify these boundary terms, since $F_{w^*}(w^*) - I(w^* > -c)$ equals zero for $w^* = w_L$ and for $w^* = w_U$.

To finish deriving Eq. (2.2), now consider the remaining integral part.

$$\begin{aligned}
&\int_{w_L}^{w_U} w^* \frac{\partial [F_{w^*}(w^*) - I(w^* > -c)]}{\partial w^*}\, dw^* \\
&\quad = \int_{w_L}^{-c} w^* \frac{\partial F_{w^*}(w^*)}{\partial w^*}\, dw^* + \int_{-c}^{w_U} w^* \frac{\partial [F_{w^*}(w^*) - 1]}{\partial w^*}\, dw^*
\end{aligned}$$

$$= \int_{w_L}^{w_U} w^* \frac{\partial F_{w^*}(w^*)}{\partial w^*} dw^* = \int_{w_L}^{w_U} w^* dF_{w^*}(w^*) = E(W^*),$$

where the last equality just used the definition of an expectation. We have now shown that Eq. (2.4) holds.

Recall that H is defined by $H(v) = E(D \mid V = v)$. It would be simpler to base an estimator on Eq. (2.4) rather than Eq. (2.1), because now one would only need to replace H with a nonparametric regression estimator $\widehat{H}$ in an integral without taking a derivative of $\widehat{H}$. However, a further simplification is possible, by showing that Eq. (2.3) holds, as follows. Start from Eq. (2.4), making the change of variables $v = -w^*$ in the integral.

$$\begin{aligned} E(W^*) &= -c + \int_{w_L}^{w_U} [H(-w^*) - I(-w^* \geq c)] dw^* \\ &= -c + \int_{-w_U}^{-w_L} [H(v) - I(v \geq c)] dv. \end{aligned}$$

Let Ω_v be the support of V. Then

$$\begin{aligned} E(W^*) &= -c + \int_{-w_U}^{-w_L} [H(v) - I(v \geq c)]\, dv = -c + \int_{-w_U}^{-w_L} \left[\frac{H(v) - I(v \geq c)}{f_v(v)}\right] f_v(v) dv \\ &= -c + \int_{-w_U}^{-w_L} \left[\frac{H(v) - I(v \geq c)}{f_v(v)}\right] f_v(v) dv \\ &= -c + \int_{-w_U}^{-w_L} \left[\frac{E(D \mid V = v) - I(v \geq c)}{f_v(v)}\right] f_v(v) dv \\ &= -c + \int_{-w_U}^{-w_L} \left[E(\frac{D - I(V \geq c)}{f_v(V)} \mid V = v)\right] f_v(v) dv \\ &= -c + \int_{v \in \Omega_v} \left[E(\frac{D - I(V \geq c)}{f_v(V)} \mid V = v)\right] f_v(v) dv. \end{aligned}$$

The last equality above used the assumption that V can take on any value that $-W^*$ can take on, so the support of $-W^*$, which is the interval from $-w_U$ to $-w_L$, lies inside Ω_v, and also that $D - I(V \geq c) = I(V + W^* \geq 0) - I(V \geq c)$, which equals zero for any value of V that lies outside the interval $-w_U$ to $-w_L$.

Now substitute in the definition of T_1 to get

$$\begin{aligned} E(W^*) &= -c + \int_{v \in \Omega_v} E(T_1 + c \mid V = v) f_v(v) dv = -c + E[E(T_1 + c \mid V)] \\ &= E[E(T_1 \mid V)] = E(T_1), \end{aligned}$$

where we have used the definition of expectation over V and applied the law of iterated expectations. This completes the derivation of Eq. (2.3).

2.4. An Alternative Derivation

Let $c = 0$ for simplicity. The previous section proved $E(T_1) = E(W^*)$ by applying an integration by parts argument to get that $E(W^*) = \int [H(-w^*) - I(-w^* \geq 0)]dw^*$, or equivalently $E(W^*) = \int [H(v) - I(v \geq 0)]dv$. Multiplying and dividing the expression in this integral by $f_v(v)$ and using $H(v) = E(D \mid V = v)$ then shows that this integral is the same as $E(T_1)$.

The advantage of the derivation in the previous section is that it directly shows how T_1 is obtained from the definition of the mean of a random variable W^*, and hence how it directly follows from identification of H.

A more direct, but perhaps less insightful, derivation of the result follows from the proof in Lewbel (2000). Starting from $E(T_1)$ with $c = 0$, follow the steps of the previous section in reverse order to get

$$\begin{aligned}
E(T_1) &= \int_{-w_U}^{-w_L} \left[E\left(\frac{D - I(V \geq 0)}{f_v(V)} \mid V = v \right) \right] f_v(v) dv \\
&= \int_{-w_U}^{-w_L} E[D - I(V \geq 0) \mid V = v] dv \\
&= \int_{-w_U}^{-w_L} E[I(V + W^* \geq 0) - I(V \geq 0) \mid V = v] dv \\
&= \int_{-w_U}^{-w_L} E[I(V + W^* \geq 0) - I(V \geq 0) \mid V = v] dv \\
&= E[\int_{-w_U}^{-w_L} [I(v + W^* \geq 0) - I(v \geq 0)] dv],
\end{aligned}$$

where this last step used the independence of V and W^* to pass the integral into the expectation. Equivalently, the expectation is just over W^*, so we are just changing the order of integration over v and over W^*. When W^* is positive, this expression becomes

$$\begin{aligned}
E(T_1) &= E\left(\int_{-w_U}^{-w_L} I(0 \leq v \leq W^*) dv \right) \\
&= E\left(\int_0^{W^*} 1\, dv \right) = E(W^*)
\end{aligned}$$

and an analogous result holds when W^* is negative.

An advantage of this derivation is that it does not entail direct consideration of the boundary terms associated with the integration by parts. The technical assumptions required for dealing with integration by parts and possibly infinite boundaries w_L or w_U are replaced with the assumption that we can change the order of integration (e.g., Fubini's theorem).

2.5. Estimating Unconditional Moments

Given that $E(T_1) = E(W^*)$, we can construct the variable $T_{1i} = [D_i - I(V_i \geq 0)]/f_v(V_i)$ for each individual i in a sample, and let the estimate of $E(W^*)$ be the sample average $\overline{T}_1 = \sum_{i=1}^{n} T_{1i}/n$. Here we have made the simplest choice for the constant c, letting $c = 0$. In applications where the density of V is not already known to the researcher by experimental design, $\overline{T}$ can still be used to estimate $E(W^*)$ by replacing $f_v(V_i)$ in T_{1i} with a uniformly consistent density estimator $\widehat{f}_v(V_i)$. For example, an ordinary Rosenblatt–Parzen kernel density estimator could be used, or the simpler sorted data estimator described by Lewbel and Schennach (2007).

Although $\overline{T}_1$ is a sample average, it is possible that this estimator will not converge at rate root n. This is because the density $f_v(V_i)$ may have thin tails, and we are dividing by this density, which means that the distribution of T_1 can have tails that are too thick to satisfy the Lindeberg condition for the central limit theorem. It can be shown that obtaining parametric rates requires finite support for W^* and V, or that V has infinite variance, or that W^* satisfies a tail symmetry condition as defined by Magnac and Maurin (2007). See Khan and Tamer (2010), Khan and Nekipelov (2010a, 2010b), and Dong and Lewbel (2012) for more discussion of this point, and see Lewbel and Schennach (2007), Jacho-Chávez (2009), and Chu and Jacho-Chávez (2012) for more general limiting distribution theory regarding averages with a density in the denominator.

Based on Lewbel, Linton, and McFadden (2011), Eq. (2.3) readily extends to estimating other moments of W^*, using the fact that $D = I(V + W^* \geq 0) = I[-h(-V) + h(W^*) \geq 0]$ for any strictly monotonically increasing function h. Therefore, if we let $\widetilde{V} = -h(-V)$ and $h'(v) = \partial h(v)/\partial v$, we have

$$\begin{aligned} E[h(W^*)] &= E\left[c + \frac{D - I(\widetilde{V} \geq c)}{f_{\widetilde{v}}(\widetilde{V})}\right] \\ &= E\left(c + \frac{[D - I([h(-V) \leq -c])]h'(-V)}{f_v(V)}\right). \end{aligned}$$

It follows that, given a function h, to estimate $E[h(W^*)]$ we can construct

$$T_{hi} = c + \frac{[D_i - I([h(-V_i) \leq -c])]h'(-V_i)}{f_v(V_i)}$$

for each observation i and then take the sample average $\overline{T}_h = \sum_{i=1}^{n} T_{hi}/n$ as an estimator for $E[h(W^*)]$. For example, letting $h(W^*) = (W^*)^\lambda$ for integers λ provides direct estimators for second and higher moments of W^*, if W^* is everywhere non-negative or non-positive as in the willingness-to-pay example.

2.6. Identification with Covariates

Now consider how the previous section's results can be generalized by the inclusion of additional covariates. Continue with the willingness-to-pay application for now, so we still have $D = I(V + W^* \geq 0)$. Let X be a vector of observed covariates consisting of attributes of the sampled individuals, such as their age and income, which might affect or otherwise correlate with their willingness to pay. The distribution that V is drawn from could depend on X (e.g., quoting higher prices to richer individuals), but by construction it will be the case that $W^* \perp V \mid X$. The symbols $\perp$ and $\mid$ denote statistical independence and conditioning respectively, so $W^* \perp V \mid X$ says that the conditional distribution of W^* given X is independent of the conditional distribution of V given X.

Let $H(v,x) = E(D \mid V = v, X = x)$ so now H could be estimated as a nonparametric regression of D on both X and V. It follows from $W^* \perp V \mid X$ that

$$H(v,x) = 1 - \Pr[W^* \geq -v \mid X = x] = 1 - F_{W^*|X}(-v \mid x), \tag{2.5}$$

where $F_{W^*|X}$ denotes the conditional distribution of W^* given X. Therefore

$$F_{W^*|X}(w \mid x) = 1 - E(D \mid V = -w, X = x), \tag{2.6}$$

so one minus a nonparametric regression of D on both X and V, evaluated at $V = -w$, provides a consistent estimator of the conditional distribution of W^* given X, that is, the distribution of willingness to pay conditional on observed attributes of individuals like their age or income.

An analogous calculation to that of the previous sections can be applied to calculate the conditional mean willingness to pay $E(W^* \mid X)$. All that is required is to replace every expectation, density, and distribution function in the previous sections with conditional expectations, densities, or distributions, conditioning on X. In place of Eq. (2.2), define T_2 by

$$T_2 = \frac{D - I(V \geq 0)}{f_{v|x}(V \mid X)}, \tag{2.7}$$

where $f_{v|x}$ denotes the conditional probability density function of V given X (for simplicity, we will hereafter let $c = 0$). As before with the pdf f_v in T_1, when constructing T_2 the pdf $f_{v|x}$ is either known by experimental design or can be estimated using a Rosenblatt–Parzen kernel density estimator. Repeating the derivations of the previous sections, but now conditioning on $X = x$, shows that $E(W^* \mid X) = E(T_2 \mid X)$. Therefore the conditional mean willingness to pay $E(W^* \mid X)$ can be consistently estimated by a nonparametric regression (e.g., a kernel or local linear regression) of T_2 on X, where each observation of T_2 is defined by $T_{2i} = [D_i - I(V_i \geq 0)]/f_{v|x}(V_i \mid X_i)$ for each individual i in the sample.

The fact that the entire conditional distribution of W^* given X is identified means that any model for W^* that would have been identified if W^* were observable will now be identified via the special regressor even though W^* is latent. The next sections give some examples.

2.7. Latent Linear Index Models

Consider the standard consumer demand threshold crossing model $D = I(V + X'\beta + \varepsilon \geq 0)$, where D indicates whether a consumer buys a product or not, and $V + X'\beta + \varepsilon$ is the individual's utility from purchasing the good. This corresponds to the special case of the model in the previous section in which the model $W^* = X'\beta + \varepsilon$ is imposed. The vector X can (and typically would) include a constant term. If V is the negative of the price, then normalizing the coefficient of V to equal one puts the utility in a money metric form, that is, $V + X'\beta + \varepsilon$ is then the consumer's surplus, defined as their reservation price W^* minus their cost of purchasing, which is $-V$. The willingness-to-pay model is an example where the product is a public good like preserving a wetland while price is determined by experimental design. In the more general demand context, we may not want to take price to be the special regressor since, for example, real-world prices may be endogenous, or not vary sufficiently to give V the necessary large support. But whatever variable that we take to be V can have its coefficient normalized to equal one (by changing the sign of V if necessary).

We will not require ε to be independent of X. So, for example, the random coefficients model $D = I(V + X'e \geq 0)$ would be permitted, defining $\beta = E(e)$ and $\varepsilon = X'(e - \beta)$. More generally, the variance and higher moments of ε can depend on X in arbitrary ways, allowing for any unknown form of heteroskedasticity with respect to X.

This type of model would typically be estimated by parameterizing ε and doing maximum likelihood; for example, this would be a probit model if ε were normal and independent of X. However, the special regressor V allows this model to be identified and estimated even if the distribution of ε is unknown, and even if the second and higher moments of that distribution depends on X in unknown ways (such as having heteroskedasticity of unknown form).

If all we know about the latent error ε in the threshold crossing model $D = I(V + X'\beta + \varepsilon \geq 0)$ is that $E(X\varepsilon) = 0$, then it can be shown that β would not be identified. However, here we also have the special regressor conditional independence assumption that $W^* \perp V \mid X$, which with $W^* = X'\beta + \varepsilon$ implies that $\varepsilon \perp V \mid X$. This condition along with $E(X\varepsilon) = 0$ suffices to identify the entire model, as follows.

First note that in the usual linear regression way, having $W^* = X'\beta + \varepsilon$ and $E(X\varepsilon) = 0$ means that $\beta = E(XX')^{-1}E(XW^*)$, assuming that $E(XX')$ is nonsingular. We do not observe W^*; however,

$$E(XW^*) = E(XE(W^* \mid X)) = E\left(X \int w^* f_{w^*|X}(w^* \mid x)dw^*\right)$$

$$= E\left(X \int w^* \frac{dF_{w^*|X}(w^* \mid x)}{dw^*}\, dw^*\right),$$

where the integral is over the support of W^*. So, to identify β, recall that $H(v,x) = E(D \mid V = v, X = x)$ is identified (and could be estimated by an ordinary nonparametric regression). Then $W^* \perp V \mid X$ implies that $1 - H(v,x) = F_{W^*|X}(-v \mid x)$ as before. Plugging this into the above integral shows that

$$\beta = E(XX')^{-1}E(XW^*) = E(XX')^{-1}E\left(X \int v \frac{d[1 - H(-v,x)]}{dv}\, dv\right)$$

and therefore that β is identified, because all of the terms on the right are identified.

Following the logic of the previous sections, a much simpler estimator for β can be constructed using T_2. In particular, by the derivation of Eqs. (2.6) and (2.7) we have

$$\beta = E(XX')^{-1}E(XW^*) = E(XX')^{-1}E(XT_2),$$

so β is given by a linear ordinary least squares regression of T_2 on X, where T_2 is defined by Eq. (2.7). This is one of the estimators proposed in Lewbel (2000).

To implement this estimator, we first construct $T_{2i} = [D_i - I(V_i \geq 0)]/f_{v|x}(V_i \mid X_i)$ for each observation i in the data and then construct the usual ordinary least squares regression estimator $\widehat{\beta} = (\sum_{i=1}^{n} X_i X_i')^{-1}(\sum_{i=1}^{n} X_i T_{2i})$. As before, if the density $f_{v|x}(V_i \mid X_i)$ is unknown, it can be estimated using, for example, a kernel density estimator.

2.8. Latent Random Coefficients

A number of authors have considered binary choice models with random coefficients, including Berry, Levinsohn, and Pakes (1995), Ichimura and Thompson (1998), Gautier and Kitamura (2009), and Hoderlein (2009). Suppose the latent W^* equals $X'e$, where e is a vector of random coefficients. Then we have the random coefficients binary choice model $D = I(V + X'e \geq 0)$, where the scale normalization is imposed by setting the coefficient of V equal to one. It then follows immediately from Eq. (2.6) that $F_{X'e|X}(w \mid x) = 1 - E(D \mid V = -w, X = x)$. The conditional distribution $F_{X'e|X}$ is the same information that one could identify from a linear random coefficients model (i.e., a model where one observed W and X with $W = X'e$), so the known nonparametric identification of linear random coefficients models (see, e.g., Hoderlein, Klemelae, and Mammen (2010) and references therein) can be immediately applied to show nonparametric identification of binary choice random coefficients, by means of the special regressor.

2.9. Latent Partly Linear Regression

Suppose that $W^* = g(X) + \varepsilon$, where $E(\varepsilon \mid X) = 0$ for some unknown function g of the vector X. This corresponds to the partly linear latent variable regression model we have $D = I(V + g(X) + \varepsilon \geq 0)$. In this model we have $g(X) = E(T_2 \mid X)$, which could therefore be estimated by an ordinary nonparametric regression of the constructed variable T_2 on X. In this model, the required conditional independence assumption that $W^* \perp V \mid X$ will hold if $\varepsilon \perp V \mid X$, in addition to $E(\varepsilon \mid X) = 0$. More simply, though stronger than necessary, it could just be assumed that $\varepsilon \perp (V, X)$ to make all the required independence assumptions regarding V, X, and ε hold. The only other requirement would then be that the support of $g(X) + \varepsilon$ equals, or is contained in, the support of $-V$ (or that the tail symmetry condition of Magnac and Maurin (2007) holds).

These results immediately extend to the general partly latent variable model $I(V + X_1'B + g_2(X_2) + \varepsilon \geq 0)$, where X_1 and X_2 are subvectors comprising X. In this model we would have $E(T_2 \mid X) = X_1'B + g_2(X_2)$, so the vector B and the function g_2 could be estimated by applying Robinson (1988), using T_2 as the dependent variable.

2.10. Latent Nonparametric Instrumental Variables

In the previous section we had the model $D = I(V + g(X) + \varepsilon \geq 0)$, where the unobserved W^* satisfied $W^* = g(X) + \varepsilon$ with $E(\varepsilon \mid X) = 0$. Suppose now we have the same model, except instead of $E(\varepsilon \mid X) = 0$ we assume $E(\varepsilon \mid Z) = 0$, where Z is a vector of observed instruments. Some elements of X, corresponding to exogenous covariates, may also be in Z. Other elements of X are endogenous, in that they could be correlated with ε.

If W^* were observed, then $g(X)$ would need to be identified and estimated from the conditional moments

$$E[W^* - g(X) \mid Z] = 0. \tag{2.8}$$

If W^* were observed, this would be the instrumental variables nonparametric regression model of, for example, Newey and Powell (2003), Hall and Horowitz (2005), Darolles, Fan, Florens, and Renault (2011), and Chen and Reiss (2011). Assume, as those authors do, that $g(X)$ is identified from the conditional moments in Eq. (2.8), and consider how we might construct sample moment analogues to this equation in the case where W^* is not observed.

Replace the definition of T_2 in Eq. (2.7) with T_3 defined by

$$T_3 = \frac{D - I(V \geq 0)}{f_{v|Z}(V \mid Z)}, \tag{2.9}$$

where $f_{v|Z}$ denotes the conditional probability density function of V given Z. Let $\Omega_{v|z}$ be the support of $V \mid Z$. Then, by replacing X with Z everywhere in Section 2.6, we have that $E(W^* \mid Z) = E(T_3 \mid Z)$, assuming that $W^* \perp V \mid Z$, and assuming that $V \mid Z$ has sufficiently large support. It follows immediately that $E[T_3 - g(X) \mid Z] = 0$, so after constructing T_3, estimators like those of Newey and Powell (2003) or the other authors listed above could be directly applied to estimate $g(X)$ by instrumental variables nonparametric regression.

In a later section we will discuss implications of the conditional independence assumption $W^* \perp V \mid Z$ and describe how this restriction can be relaxed.

2.11. Latent Linear Index Models with Endogenous or Mismeasured Regressors

Return to the linear latent variable model $D = I(V + X'\beta + \varepsilon \geq 0)$. The earlier estimator for β in this model assumed that $V \mid X$ is continuous with large support, that $\varepsilon \perp V \mid X$, that $E(X\varepsilon) = 0$, and that $E(XX')$ is nonsingular. Apart from restrictions on V, the only assumptions regarding X and ε were the same as the assumptions required for linear ordinary least squares regression, that is, $E(XX')$ nonsingular and $E(X\varepsilon) = 0$.

We now extend this model to allow for endogeneous or mismeasured regressors, by replacing the linear ordinary least squares estimator with a linear two-stage least squares estimator. Let Z be a vector of observed instruments. The vector Z is assumed to include any elements of X that are exogenous (including a constant term), but does not include V. Any elements of X that are endogenous or mismeasured are not included in Z.

Suppose that we still have the model $D = I(V + X'\beta + \varepsilon \geq 0)$, but where some or all of the elements of X are endogenous or mismeasured and so may be correlated with ε. First make the exact same assumptions that would be required for linear model two-stage least squares with instruments Z. These are $E(Z\varepsilon) = 0$, the rank of $E(ZX')$ equals the dimension of X, and $E(ZZ')$ is nonsingular. Now add special regressor assumptions regarding V, by assuming that $V \mid Z$ is continuous with large support and $\varepsilon \perp V \mid Z$. Note that the special regressor assumptions involve V, Z, and ε, but not X. The only assumptions regarding the endogenous regressors in X that are required are the same as the minimal assumptions needed for linear model two-stage least squares regression.

Letting $W^* = X'\beta + \varepsilon$, it follows immediately from the results in the previous section that if we define T_3 by Eq. (2.9), then $E(T_3 \mid Z) = E(X'\beta + \varepsilon \mid Z)$ and therefore $E(ZT_3) = E(ZX')\beta$, so

$$\beta = [E(XZ')E(ZZ')^{-1}E(ZX')]^{-1}E(XZ')E(ZZ')^{-1}E\left(Z\frac{D - I(V \geq 0)}{f_{v|Z}(V \mid Z)}\right),$$

which is identical to a linear two-stage least squares regression of T_3 on X, using instruments Z. More generally, we could estimate β by applying standard GMM estimation to the moment conditions $E[Z(T_3 - X'\beta)] = 0$. This GMM could be more efficient than two-stage least squares if errors are heteroskedastic.

A particularly useful feature of this construction is that, apart from restrictions involving V, all that is required regarding the endogenous regressors and instruments is identical to what is required for linear two-stage least squares. In particular, the type of restrictions required for control function or maximum likelihood based estimation are not needed. This is particularly useful for cases where some of the endogenous regressors are themselves discrete or limited. See Lewbel, Dong, and Yang (2012) for details.

Finally, it should be noted that only one special regressor is required, regardless of how many other regressors are endogenous. If more than one regressor satisfies the assumptions required to be special, then, based on experiments in Lewbel (2000), the one with the largest spread (e.g., largest variance or interquartile range) should be chosen to minimize finite sample bias.

2.12. Relaxing the Conditional Independence Assumption

For instrumental variables estimation, in the previous two sections it was assumed that

$$W^* \mid V, Z = W^* \mid Z \tag{2.10}$$

to obtain either $E(W^* \mid Z) = E(T_3 \mid Z)$ or, for the linear index model, $E(ZW^*) = E(ZT_3)$. Then, given $W^* = g(X) + \varepsilon$, assuming either $E(\varepsilon \mid Z) = 0$ or just $E(Z\varepsilon) = 0$, the model could be estimated by nonparametric instrumental variables or, in the linear index case where $g(X) = X'\beta$, by ordinary linear two-stage least squares, treating T_3 as the dependent variable. All that was required for these estimators to work is that Eq. (2.10) hold, that V have sufficiently large support, and that Z have the standard properties of instruments for either nonparametric instrumental variables or for linear two-stage least squares.

Since $W^* = g(X) + \varepsilon$, a sufficient but stronger than necessary condition for Eq. (2.10) to hold is that

$$X, \varepsilon \mid V, Z = X, \varepsilon \mid Z, \tag{2.11}$$

meaning that V is exogenous in a standard sense of being conditionally independent of the latent model error ε, but in addition $X \mid V, Z = X \mid Z$, meaning that V would drop out of a model of endogenous regressors as a function of V and Z. This is a strong restriction on the special regressor V relative to the endogenous regressors, but fortunately it is stronger than necessary.

One way to relax Eq. (2.10) is to replace T_3 in Eq. (2.9) with T_4 defined by

$$T_4 = \frac{D - I(V \geq 0)}{f_{v|S}(V \mid S)}, \tag{2.12}$$

where S is a vector containing the union of all the elements of Z and of X. We then have that

$$E(W^* \mid S) = E(T_4 \mid S) \tag{2.13}$$

holds, assuming $W^* \perp V \mid S$ and that $V \mid S$ has sufficiently large support. It follows from applying the law of iterated expectations to Eq. (2.13) that $E(W^* \mid Z) = E(T_4 \mid Z)$, which is what we require to estimate $g(X)$ using $E[T_4 - g(X) \mid Z] = 0$. Similarly, by applying the law of iterated expectations to Eq. (2.13), T_4 can be used in place of T_1, T_2, or T_3 in all of the estimators described so far.

Requiring $W^* \perp V \mid S$ is equivalent to $W^* \perp V \mid X, Z$. With $W^* = g(X) + \varepsilon$, this is in turn equivalent to

$$\varepsilon \perp V \mid X, Z. \tag{2.14}$$

Equation (2.14) relaxes Eq. (2.11), and in particular does not impose the condition that $X \mid V, Z = X \mid Z$. Equation (2.14) will hold if we can write a model $V = M(U, X, Z)$, where M is invertible in U and U is an error term that is independent of X, Z, ε. For example, define $P = E(V \mid X, Z)$, define $U = V - P$, and suppose that the endogenous elements of X are functions of Z, P, and an error vector e that is independent of Z, P, and U. It is not necessary to actually specify or estimate this or any model for any elements of X. With this construction, X can depend on V by depending on P, and X can be endogenous by e correlating with ε, with Eq. (2.14) holding. This construction also does not impose any control function type restrictions on W^*, X, and Z—and so, for example, still allows X to be discrete or limited.

2.13. Constructing T_{ji}

Implementing the estimators discussed in the previous sections requires constructing an estimate of T_{ji} for each observation i and for $j = 1, 2, 3$, or 4. For each observation i, the variable T_{ji} is given by $T_{ji} = -c + [D_i - I(V_i \geq c)]/f_{v|S}(V_i \mid R_i)$, where R is either empty as in Eq. (2.2) where $j = 1$, or $R = X$ in Eq. (2.7) where $j = 2$, or $R = Z$ in Eq. (2.9) where $j = 3$, or $R = S$ in Eq. (2.12) where $j = 4$. The constant c can be any value inside the support of V. A natural choice for the constant c is the mean or median of V. More simply, we will just assume that V is centered (e.g., demeaned prior to the analysis), and just let $c = 0$.

Lewbel and Tang (2011) prove that the term $I(V_i \geq 0)$ in the definition of T_{ji} can be replaced with $M(V_i)$, where M is any mean zero probability distribution function that lies inside the support of V. Choosing M to be a simple differentiable function like

$$M(V) = I(-\sigma \leq V \leq \sigma)\frac{V+\sigma}{2\sigma},$$

where σ is the standard deviation of V (corresponding to a uniform distribution on $c-\sigma$ to $c+\sigma$) can simplify the calculation of limiting distributions and possibly improve the finite sample performance of the estimators. The free substitution of $M(V)$ for $I(V \geq 0)$ can be made for all special regressor estimators.

If V is determined by the researcher as in the willingness to pay examples, then the density function $f_{v|R}$ is known by experimental design. Otherwise, $f_{v|R}$ will need to be estimated. If $f_{v|R}$ is parameterized as $f_{v|R}(V \mid R, \theta)$ for some parameter vector θ, then efficient estimation could be accomplished by GMM, combining the moments based on the score function for maximum likelihood estimation of θ, that is, $E[\partial f_{v|R}(V \mid R, \theta)/\partial\theta] = 0$ with the moments used for estimation of β, such as $E[Z(T - X'\beta)] = 0$ from the previous section.

The function $f_{v|R}$ could also be estimated nonparametrically by, for example, a Nadayara–Watson kernel density estimator, but this may be very imprecise if the dimension of R is high. Dong and Lewbel (2012) propose some alternative semiparametric estimators for this density. For example, suppose $V = S'\gamma + U$, where S is as defined in Eq. (2.12), as the union of all the elements of X and Z. If $U \perp S$, then $f_{v|S}(V \mid S) = f_u(U)$. We may then define T by

$$T = \frac{D - I(V \geq 0)}{f_u(U)} \tag{2.15}$$

and correspondingly construct data $T_i = [D_i - I(V_i \geq 0)]/f_u(U_i)$. By the law of iterated expectations, this T_i can be used in place T_{ji} for all the special regressor estimators. Moreover, this is a special case of the model discussed at the end of previous section, so the required conditional independence assumptions involving the special regressor V will be satisfied if U is independent of X, Z, and W^*.

The advantage of this construction is that each U_i can be estimated as the residuals from an ordinary least squares linear regression of V on S, and the density function f_u of the scalar random variable U can be estimated by a one-dimensional kernel density estimator applied to the data $U_1, \ldots, U_n$. Even more simply, the ordered data estimator of Lewbel and Schennach (2007) can be applied to $U_1, \ldots, U_n$, which does not require any choice of kernel or bandwidth. See Dong and Lewbel (2012) for more details and for alternative simple estimators.

One final note regarding construction of T is that $f_{v|R}$ must have a large support, and so $f_{v|R}(V_i \mid R_i)$ may be very close to zero for very low and very high values of V_i. Similarly, $f_u(U_i)$ may be very close to zero for large values of $|U_i|$. The corresponding values of T_i may then be extremely large in magnitude. The special regressor estimators that involve estimating either moments of T or regression models using T as the dependent variable can be very sensitive to outlier observations of T. It may therefore be prudent, based on a mean squared error criterion, to Winsorize T, or to trim out data observations i where either T_i or the residuals in the regressions of T_i on X_i take on extreme values.

2.14. What if the Special Regressor is Discrete?

Suppose we have a model that fits the requirements for special regressor estimation, except that the special regressor V is discretely distributed and thereby violates the required support restrictions. In this case, it is still sometimes possible to apply the estimators discussed earlier.

One possibility is to assume that the number of values that V can take on grows with the sample size. For example, in the earlier willingness-to-pay application, this would mean assuming that the larger the size of the experiment (measured in terms of the number of experimental subjects), the larger would be the number of different proposed willingness-to-pay values that experimentors would select from to offer to a subject. Analogous to the type of infill asymptotics that is sometimes used in the time series literature, suppose that in the limit as $n \to \infty$, the number of values that V can take on grows to make the support of V become dense on a large interval (larger than the support of W^*). Then Lewbel, Linton, and McFadden (2011) show that, asymptotically, the distribution of W^* can be identified, and they supply associated limiting distribution theory for estimation based on that design.

One implication of the results in Lewbel, Linton, and McFadden (2011) is that if V is discretely distributed, but the number and range of values V can take on grows sufficiently quickly with n, then one can ignore the fact that V is discrete and do special regressor estimation as if V was continuous. For the resulting estimator to perform well in practice, the number of different values that V takes on in the observed data, and the range of those values, will need to be relatively large.[3] Li and Racine (2007) give other examples of treating discrete data as if it was continuous in, for example, nonparametric kernel regressions.

Another situation in which a discrete special regressor can be used is when the true V is continuous with a uniform distribution, and what one observes is a discretized version of V. For example, consider an application like that of Maurin (2002), in which the outcome D is whether a student will be held back in school, and the special regressor V is how old the student is at the date of enrollment in school. Suppose we do not observe the student's birthdates and times, but instead only know if a student is either five or six years old when school starts. So here the observed, discretized version of the special regressor age is just the binary variable indicating whether the student is five or six. By defining c appropriately, $I(V_i \geq c)$ will be one for students who are six when school starts, and zero for students who are five. Assuming that birthdates and times are close to uniform within a year,[4] $f_{V|R}(V_i \mid R_i)$ is a constant (equal to one if V is measured in years) and so is known despite not observing V_i. Since both $f_{V|R}(V_i \mid R_i)$ and $I(V_i \geq c)$ are known for each student i, the variable T_{ji} can be constructed for each student i, and so the special regressor estimators can be applied. Special regressor estimation only requires observing T_{ji}, not V_i, for each observation i.

If situations like those described above do not apply, then special regressor methods can still be used when V is discrete, but as described in Section 2.2, the distribution of the latent variable W^* will only be identified at the values that $-V$ can take on. This in turn only permits bounds on coefficients and moments to be identified and estimated. See Magnac and Maurin (2008) for details on partial identification with a discrete special regressor.

2.15. Extensions

This chapter has focused on binary choice model estimation, but the main idea of the special regressor method can be applied to a variety of models. For example, ordered choice models are also identified, including models with random thresholds and endogenous regressors. Suppose for $j = 1, \ldots, J$ that Y equals the integer j when $\alpha_{j-1} + \varepsilon_{j-1} < V + g(X) \leq \alpha_j + \varepsilon_j$ for some constants α_j and errors ε_j having unknown distributions. Then let $D_j = I(Y \leq j) = I(V + g(X) - \alpha_j - \varepsilon_j)$ and apply the special regressor estimator to each D_j to identify the conditional distribution of $g(X) - \alpha_j - \varepsilon_j$ given X for each j (or given instruments Z if endogenous regressors are present). See Lewbel (2000) for details.

The special regressor method is convenient for panel data latent variable models with latent fixed effects, because if $D_{it} = I(V_{it} + W^*_{it} \geq 0)$, then we can construct T_{jit} such that $E(T_{jit} \mid X) = E(W^*_{it} \mid X)$; and so, for example, $E(T_{jit} - T_{jit-1} \mid X) = E(W^*_{it} - W^*_{it-1} \mid X)$, meaning that we can difference out fixed effects in the latent W^*_{it}. This construction permits, but does not require, the special regressor V_{it} to vary over time. See, for example, Honore and Lewbel (2002), Ai and Gan (2010), and Gayle (2012). Using the special regressor in this way, it is sometimes possible to estimate panel binary choice models with fixed effects that converge at rate root n, even when the error distribution is not only not logit, but not known at all. Here the special regressor conditional independence assumption overcomes Chamberlain's (1993) result that logit errors are required for root n convergence of panel binary choice models with fixed effects.

As briefly noted earlier, if $Y = g(V + W^*)$ for some possibly unknown, weakly monotonic function g, then the conditional distribution of W^* given a vector of covariates X can be identified (up to a location normalization) by first letting $D = I(Y \geq y_0)$ for any constant y_0 in the support of Y that makes D be nonconstant, which reduces the problem to an equivalent binary choice problem. Identification is only up to a location normalization because for any constant a, one could replace W^* with $W^* + a$ and redefine g accordingly.

The particular location value for W^* that is obtained by applying previous estimators to $D = I(Y \geq y_0)$ will depend on g and on the choice of y_0. To increase the efficiency of terms other than location, one could combine estimates based on multiple choices of y_0. This can be also be combined with the above panel data generalization to permit estimation of many nonlinear or nonparametric panel data models with latent fixed effects.

Suppose $Y = g(V_1 + W_1^*, V_2 + W_2^*)$, where V_1 and V_2 are two special regressors and W_1^* and W_2^* are two latent variables. Then generally the joint distribution of W_1^* and W_2^* given X can be identified. Lewbel (2000) provides an example showing identification of general multinomial choice models. Similarly, Lewbel and Tang (2011), Khan and Nekipelov (2011), and Fox and Yang (2012) use multiple special regressors for identification of games, including entry games and matching games, with semiparametrically specified payoffs, while Berry and Haile (2009a, 2009b) use multiple special regressors to identify multinomial discrete choice market equilibrium models that are semiparametric generalizations of Berry, Levinsohn, and Pakes (1995).

2.16. Conclusions

The goal here has been to provide an understanding of how special regressor methods work and can be applied to estimate features of latent variable models. The focus was on identification and associated construction of estimators, not on limiting distribution theory. The estimators are multistep estimators, where each step takes the form of a standard parametric or nonparametric density or regression estimator. Yet despite being comprised of standard estimators, a number of technical issues can arise. In particular, the rates of convergence of these estimators can vary depending upon the thickness of the tails of the distribution of the special regressor. In general, converging at standard parametric root n rates requires either parametric specification of the density of V, or finite support of the model errors, or very thick tails for V (thick enough for the variance to be infinite), or conditions like the tail symmetry of Magnac and Maurin (2007). When special regressor estimators do converge at rate root n, the standard methods used to derive asymptotic distributions of multistep estimators can be applied, and in such cases the basic special regressor estimator has been shown to be semiparametrically efficient. Papers discussing limiting distribution theory for special regressor-based estimators include Lewbel (2000, 2007a), Magnac and Maurin (2007, 2008), Jacho-Chávez (2009), Khan and Tamer (2010), Khan and Nekipelov (2010a, 2010b), and Dong and Lewbel (2012).

Lewbel, Dong, and Yang (2012) provide a comparison of special regressor models versus maximum likelihood estimation, control function estimators, and linear probability models. They conclude that the greatest weakness of special regressor methods is the extent to which they rely on strong properties of just one regressor, some of which are difficult to verify, and the resulting sensitivity of estimates to this one regressor.

But the strength of special regressor methods is that they impose very mild conditions on the relationships between the remaining regressors and on the model errors. This is what makes special regressor methods particularly useful for proving identification of models that have some relatively intractable components, such as their use by Berry and Haile (2009a, 2009b) to deal with endogenous prices in semiparametric generalizations of Berry, Levinsohn, and Pakes (1995)-type market models, or their use by Lewbel and Tang (2011) to identify semiparametric payoff functions in discrete games.

The restrictions required for consistency of special regressor estimators are generally quite different from what is required for other estimators. This suggests that for empirical work, they may be particularly useful as robustness checks. If both special regressor estimators and more standard methods provide similar answers, one may have a greater degree of confidence that the findings are not due to violations of standard modeling assumptions. The simplicity of many special regressor estimators makes this an easy suggestion to follow.

Notes

[†] I would like to thank Thomas (Tao) Yang for helpful research assistance.

1. Past empirical practice has not been to design experiments sufficient to nonparametrically identify the willingness to pay distribution by drawing from a continuous distribution. Instead, the custom in the literature on valuing public goods is to make functional form assumptions that suffice to identify the desired features of the willingness to pay distribution from the few points on the F_{W^*} distribution that are revealed by standard experiments. Lewbel, Linton, and McFadden (2011) observe that identification could be obtained by increasing the number of mass points in the distribution of P in the experimental design as the sample size grows, to become asymptotically dense on the line. They supply associated limiting distribution theory for estimation based on that design. See Section 2.14, regarding when the special regressor is discrete.
2. While estimation of the mean formally requires knowing $F_{w^*}(w^*)$ on the entire support of w^*, if the tails of F_{w^*} are thin, then the bias of failing to estimate those tails will be small. Magnac and Maurin (2007) provide conditions, called tail symmetry, that make this bias not just small but actually zero, permitting consistent estimation of means without large support.
3. In Lewbel, Linton, and McFadden's (2011) empirical application, the sample size was $n = 518$ and the number of willingness to pay bid values that V took on was 14, ranging in value from 25 to 375. This range between the lowest and highest value of V in the data likely covered a large portion of the actual range of willingness to pay in the population.
4. There exists evidence of statistically significant seasonal departures from uniformity in the distribution of births within a year, but the magnitude of these departures from uniformity is quite small. See, for example, Beresford (1980). A related use of uniformity to account for discretization in observed age is Dong (2012).

References

Abbring, J. H., and J. J. Heckman. 2007. "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." Chapter 72 in *Handbook of Econometrics*, Vol. 6B, first edition, eds. J. J. Heckman & E. E. Leamer. New York: Elsevier.

Ai, C., and L. Gan. 2010. "An Alternative Root-*n* Consistent Estimator for Panel Data Binary Choice Models." *Journal of Econometrics*, **157**, pp. 93–100.

An, M. Y. 2000. "A Semiparametric Distribution for Willingness to Pay and Statistical Inference with Dichotomous Choice CV Data." *American Journal of Agricultural Economics*, **82**, pp. 487–500.

Avelino, R. R. G. 2006. "Estimation of Dynamic Discrete Choice Models with Flexible Correlation in the Unobservables with an Application to Migration within Brazil." Unpublished manuscript, University of Chicago.

Anton, A. A., A. Fernandez Sainz, and J. Rodriguez-Poo. 2002. "Semiparametric Estimation of a Duration Model." *Oxford Bulletin of Economics and Statistics*, **63**, pp. 517–533.

Beresford, G. C. 1980. "The Uniformity Assumption in the Birthday Problem." *Mathematics Magazine*, **53**, pp. 286–288.

Berry, S. T., and P. A. Haile. 2009a. "Identification in Differentiated Products Markets Using Market Level Data," Unpublished manuscript.

Berry, S. T., and P. A. Haile. 2009b. "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers." Unpublished manuscript.

Berry, S., J. Levinsohn, and A. Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica*, **63**, pp. 841–890.

Briesch, R., P. Chintagunta, and R. L. Matzkin. 2010. "Nonparametric Discrete Choice Models with Unobserved Heterogeneity." *Journal of Business and Economic Statistics*, **28**, pp. 291–307.

Chamberlain, G. 1993. "Feedback in Panel Data Models." Unpublished manuscript, Department of Economics, Harvard University.

Chen, X., and M. Reiss. 2011. "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics." *Econometric Theory*, **27**, pp. 497–521.

Cogneau, D., and E. Maurin. 2002. "Parental Income and School Attendance in a Low-Income Country: A Semiparametric Analysis." Unpublished manuscript.

Chu, B., and D. T. Jacho-Chávez. 2012. "*k*-Nearest Neighbor Estimation of Inverse-Density-Weighted Expectations with Dependent Data." *Econometric Theory*, **28**, pp. 769–803.

Darolles, S., Fan, Y., Florens, J. P., and Renault, E. 2011. "Nonparametric Instrumental Regression." *Econometrica*, **79**, pp. 1541–1565.

Dong, Y. 2012. "Regression Discontinuity Applications with Rounding Errors in the Running Variable." Unpublished manuscript, University of California, Irvine.

Dong, Y., and A. Lewbel. 2012. "A Simple Estimator for Binary Choice Models with Endogenous Regressors." *Econometrics Reviews*, forthcoming.

Fox, J., and C. Yang. 2012. "Unobserved Heterogeneity in Matching Games." Unpublished manuscript.

Gautier, E., and Y. Kitamura. 2009. "Nonparametric Estimation in Random Coefficients Binary Choice Models." Unpublished manuscript.

Gayle, W.-R. 2012. “Identification and $\sqrt{N}$-Consistent Estimation of a Nonlinear Panel Data Model with Correlated Unobserved Effects.” Unpublished manuscript, University of Virginia.

Goux, D., and E. Maurin. 2005. “The Effect of Overcrowded Housing on Children’s Performance at School.” *Journal of Public Economics*, **89**, pp. 797–819.

Hall, P., and J. L. Horowitz. 2005. “Nonparametric Methods for Inference in the Presence of Instrumental Variables.” *Annals of Statistics*, **33**, pp. 2904–2929.

Heckman, J. J., and S. Navarro. 2007. “Dynamic Discrete Choice and Dynamic Treatment Effects.” *Journal of Econometrics*, **136**, pp. 341–396.

Hoderlein, S. 2009. “Endogeneity in Semiparametric Binary Random Coefficient Models.” Unpublished manuscript.

Hoderlein, S., J. Klemelae, and E. Mammen. 2010. “Analyzing the Random Coefficient Model Nonparametrically.” *Econometric Theory*, **26**, pp. 804–837.

Honore, B., and A. Lewbel. 2002. “Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors.” *Econometrica*, **70**, pp. 2053–2063.

Horowitz, J. L. 1992. “A Smoothed Maximum Score Estimator for the Binary Response Model.” *Econometrica*, **60**, pp. 505–532.

Ichimura, H., and T. S. Thompson. 1998. “Maximum Likelihood Estimation of a Binary Choice Model with Random Coefficients of Unknown Distribution.” *Journal of Econometrics*, **86**, pp. 269–295.

Jacho-Chávez, D. T. 2009. “Efficiency Bounds for Semiparametric Estimation of Inverse Conditional-Density-Weighted Functions.” *Econometric Theory*, **25**, pp. 847–855.

Khan, S., and A. Lewbel. 2007. “Weighted and Two Stage Least Squares Estimation of Semiparametric Truncated Regression Models.” *Econometric Theory*, **23**, pp. 309–347.

Khan, S., and D. Nekipelov. 2010a. “Semiparametric Efficiency in Irregularly Identified Models.” Unpublished working paper.

Khan, S., and D. Nekipelov. 2010b. “Information Bounds for Discrete Triangular Systems.” Unpublished working paper.

Khan, S., and D. Nekipelov. 2011. “Information Structure and Statistical Information in Discrete Response Models.” Unpublished working paper.

Khan, S., and E. Tamer. 2010. “Irregular Identification, Support Conditions, and Inverse Weight Estimation.” *Econometrica*, **78**, pp. 2021–2042.

Lewbel, A. 1997. “Semiparametric Estimation of Location and Other Discrete Choice Moments.” *Econometric Theory*, **13**, pp. 32–51.

Lewbel, A. 1998. “Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors.” *Econometrica*, **66**, pp. 105–121.

Lewbel, A. 2000. “Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables.” *Journal of Econometrics*, **97**, pp. 145–177.

Lewbel, A. 2007a. “Endogenous Selection or Treatment Model Estimation.” *Journal of Econometrics*, **141**, pp. 777–806.

Lewbel, A. 2007b. “Modeling Heterogeneity.” Chapter 5 in *Advances in Economics and Econometrics: Theory and Applications*, Vol. III, Ninth World Congress (Econometric Society Monographs), eds. Richard Blundell, Whitney K. Newey, and Torsten Persson. Cambridge, UK: Cambridge University Press, pp. 111–121.

Lewbel, A., Dong, Y., and T. Yang. 2012. “Comparing Features of Convenient Estimators for Binary Choice Models with Endogenous Regressors.” *Canadian Journal of Economics*, forthcoming.

Lewbel, A., and S. Schennach. 2007. "A Simple Ordered Data Estimator for Inverse Density Weighted Functions." *Journal of Econometrics*, **186**, pp. 189–211.

Lewbel, A., and X. Tang. 2011. "Identification and Estimation of Games with Incomplete Information using Excluded Regressors." Unpublished manuscript.

Lewbel, A., O. Linton, and D. McFadden. 2011. "Estimating Features of a Distribution from Binomial Data." *Journal of Econometrics*, **162**, pp. 170–188.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*, Princeton. NJ: Princeton University Press.

Magnac, T., and E. Maurin. 2007. "Identification and Information in Monotone Binary Models." *Journal of Econometrics*, **139**, pp. 76–104.

Magnac, T., and E. Maurin. 2008. "Partial Identification in Monotone Binary Models: Discrete Regressors and Interval Data." *Review of Economic Studies*, **75**, pp. 835–864.

Manski, C. F. 1985. "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator." *Journal of Econometrics*, **27**, pp. 313–333.

Matzkin, R. L. 1992. "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models." *Econometrica*, **60**, pp. 239–270.

Matzkin, R. L. 1994. "Restrictions of Economic Theory in Nonparametric Methods." Chapter 42 in *Handbook of Econometrics*, Vol. IV, eds. R. F. Engel and D. L. McFadden. Amsterdam: Elsevier, pp. 2524–2554.

Matzkin, R. 2007. "Heterogeneous Choice." Chapter 4 in *Advances in Economics and Econometrics: Theory and Applications*, Vol. III, Ninth World Congress (Econometric Society Monographs), eds. Richard Blundell, Whitney K. Newey, and Torsten Persson. Cambridge, UK: Cambridge University Press, pp. 75–110.

Maurin, E. 2002. "The Impact of Parental Income on Early Schooling Transitions A Re-examination Using Data over Three Generations." *Journal of Public Economics*, **85**, pp. 301–332.

Newey, W. K., and J. L. Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica*, **71**, pp. 1565–1578.

Pistolesi, N. 2006. "The Performance at School of Young Americans, with Individual and Family Endowments." Unpublished manuscript.

Robinson, P. M. 1988. "Root-n-Consistent Semiparametric Regression." *Econometrica*, **56**, pp. 931–954.

Stewart, M. B. 2005. "A Comparison of Semiparametric Estimators for the Ordered Response Model." *Computational Statistics and Data Analysis*, **49**, pp. 555–573.

Tiwari, A. K., P. Mohnen, F. C. Palm, S. S. van der Loeff. 2007. "Financial Constraint and R&D Investment: Evidence from CIS." United Nations University, Maastricht Economic and Social Research and Training Centre on Innovation and Technology (UNU-MERIT) Working Paper 011.

Vytlacil, E., and N. Yildiz. 2007. "Dummy Endogenous Variables in Weakly Separable Models." *Econometrica*, **75**, pp. 757–779.

PART II

INVERSE PROBLEMS

CHAPTER 3

ASYMPTOTIC NORMAL INFERENCE IN LINEAR INVERSE PROBLEMS

MARINE CARRASCO, JEAN-PIERRE FLORENS, AND ERIC RENAULT[1]

3.1. Introduction

At least since Hansen's (1982) seminal paper on Generalized Method of Moments (GMM), econometricians have been used to make inference on an object of interest defined by a family of orthogonality conditions. While Hansen's GMM is focused on inference on a finite-dimensional vector θ of structural unknown parameters, our object of interest in this chapter will typically be a function φ element of some Hilbert space $\mathcal{E}$.

While Hansen (1982) acknowledged upfront that "identification requires at least as many orthogonality conditions as there are coordinates in the parameter vector to be estimated," we will be faced with two dimensions of infinity. First, the object of interest, the function φ, is of infinite dimension. Second, similarly to above, identification will require a set of orthogonality conditions at least as rich as the infinite dimension of φ.

Then, a convenient general framework is to describe the set of orthogonality conditions through a linear operator T from the Hilbert space $\mathcal{E}$ to some other Hilbert space $\mathcal{F}$ and a target vector r given in $\mathcal{F}$. More precisely, the testable implications of our structural model will always be summarized by a linear inverse problem:

$$T\varphi = r, \tag{3.1}$$

which will be used for inference about the unknown object φ based on a consistent estimator $\hat{r}$ of r. Similarly to the Method of Moments, the asymptotic normality of estimators $\hat{\varphi}$ of φ will be derived from asymptotic normality of the sample counterpart $\hat{r}$ of the population vector r.

[1] We thank the editors and one referee for helpful comments. Carrasco thanks SSHRC for partial financial support.

However, it is worth realizing that the functional feature of r introduces an additional degree of freedom that is not common for GMM with a finite number of unknown parameters, except in the recent literature on many weak instruments asymptotics. More precisely, the accuracy of estimators of r, namely the rate of convergence of $\hat{r}$ for asymptotic normality, heavily depends on the "choice of instruments"—namely, on the choice of the inverse problem (3.1) to solve. It must actually be kept in mind that this choice is to some extent arbitrary since (3.1) can be transformed by any operator K to be rewritten:

$$KT\varphi = Kr. \tag{3.2}$$

An important difference with (semi)parametric settings is that even the transformation by a one-to-one operator K may dramatically change the rate of convergence of the estimators of the right-hand side (r.h.s.) of the equation. Some operators (as integration or convolution) are noise-reducing whereas some others (as differentiation or deconvolution) actually magnify the estimation error.

A maintained assumption will be that some well-suited linear transformation Kr allows us to get a root-n asymptotically normal estimator $K\hat{r}$ of Kr. Then, the key issue to address is the degree of ill-posedness of the inverse problem (3.2)—that is, precisely to what extent the estimation error in $K\hat{r}$ is magnified by the (generalized) inverse operator of (KT).

Because of the ill-posedness of the inverse problem, we need a regularization of the estimation to recover consistency. Here, we consider a class of regularization techniques which includes Tikhonov, iterated Tikhonov, spectral cutoff, and Landweber–Fridman regularizations. For the statistical properties of these methods, see Engl, Hanke, and Neubauer (2000). For a review of the econometric aspects, we refer the reader to Florens (2003) and Carrasco, Florens, and Renault (2007).

The focus of this chapter is the asymptotic normality of the estimator $\hat{\varphi}$ of φ in the Hilbert space $\mathcal{E}$. With normality in the Hilbert space $\mathcal{E}$ being defined through all linear functionals $< \hat{\varphi}, \delta >$ (see, e.g., Chen and White (1998)), it is actually the rate of convergence of such functionals that really matters. In the same way as going from (3.1) to (3.2) may modify the rate of convergence of sample counterparts of the r.h.s, rates of convergence of linear functionals $< \hat{\varphi}, \delta >$ will depend on the direction δ we consider. There may exist in particular some Hilbert subspace of directions warranting root-n asymptotic normality of our estimator $\hat{\varphi}$. However, it is worth stressing that focusing only on such directions amounts to overlooking the information content of other test functions and, as such, yields to suboptimal inference. It is then worth characterizing the rate of convergence to normality of estimators $\hat{\varphi}$ of φ in any possible direction δ of interest. Since this rate actually depends on the direction, we do not get a functional asymptotic normality result as in other settings put forward in Chen and White (1998).

The chapter is organized as follows. Section 3.2 presents the model and examples. Section 3.3 describes the estimation method. Section 3.4 investigates the normality for fixed regularization parameter α. This result is used in the tests described in Section 3.5. Section 3.6 establishes asymptotic normality when α goes to zero. Section 3.7 discusses the practical selection of α, and Section 3.8 describes the implementation. Section 3.9 concludes.

In the sequel, $\mathcal{D}$ and $\mathcal{R}$ denote the domain and range of an operator. Moreover, $t \wedge s = \min(t,s)$ and $t \vee s = \max(t,s)$.

3.2. Model and Examples

A wide range of econometric problems are concerned with estimating a function φ from a structural model

$$r = T\varphi \tag{3.3}$$

where T is a linear operator from a Hilbert (L^2 or Sobolev) space $\mathcal{E}$ into a Hilbert space $\mathcal{F}$. The function r is estimated by $\hat{r}$, and the operator T is either known or estimated. We present four leading examples.

3.2.1. Density

We observe data $x_1, x_2, \ldots, x_n$ of unknown density f we wish to estimate. The density f is related to the distribution function F through

$$F(t) = \int_{-\infty}^{t} f(s)\,ds = \big(Tf\big)(t).$$

In this setting, $r = F$ and the operator T is a known integral operator. F can be estimated by $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq t)$, where $\widehat{F}(t)$ converges at a parametric rate to F.

3.2.2. Deconvolution

Assume we observe n i.i.d. realizations $y_1, y_2, \ldots, y_n$ of a random variable Y with unknown density h. Y is equal to a unobservable variable X plus an error ε, where X and ε are mutually independent with density functions f and g respectively so that $h = f * g$. The aim is to estimate f assuming g is known. The problem consists in solving for f the equation

$$h(y) = \int g(y-x)f(x)dx.$$

In this setting, the operator T is known and defined by $(Tf)(y) = \int g(y-x)f(x)dx$ whereas $r = h$ can be estimated but at a slower rate than the parametric rate.

Here, the choice of the spaces of reference is crucial. If T is considered as an operator from $L^2(\mathbb{R})$ into $L^2(\mathbb{R})$ provided with Lebesgue measure, then T has a continuous spectrum. Carrasco and Florens (2011) chose spaces of reference for which T is compact and hence has a discrete spectrum. Let π_X and π_Y be two non-negative weighting functions. Let $L^2_{\pi_Y}$ denote the space of square integrable real-valued functions with respect to π_Y:

$$L^2_{\pi_Y} = \left\{ \psi\left(y\right) \text{ such that } \int \psi\left(y\right)^2 \pi_Y\left(y\right) dy < \infty \right\}.$$

$L^2_{\pi_X}$ is defined similarly. We formally define T as the operator from $L^2_{\pi_X}$ into $L^2_{\pi_Y}$ which associates to any function $\phi(x)$ of $L^2_{\pi_X}$ a function of $L^2_{\pi_Y}$ as

$$\left(T\phi\right)\left(y\right) = \int g(y-x)\phi(x)dx. \tag{3.4}$$

Provided that π_X and π_Y are such that

$$\int\int \left(g(y-x)\right)^2 \frac{\pi_Y\left(y\right)}{\pi_X\left(x\right)} dxdy < \infty,$$

T is a Hilbert–Schmidt operator and hence has a discrete spectrum. We define the adjoint, T^*, of T, as the solution of $\langle T\varphi, \psi\rangle = \langle \varphi, T^*\psi\rangle$ for all $\varphi \in L^2_{\pi_X}$ and $\psi \in L^2_{\pi_Y}$. It associates to any function $\psi(y)$ of $L^2_{\pi_Y}$ a function of $L^2_{\pi_X}$:

$$\left(T^*\psi\right)(x) = \int \frac{g\left(y-x\right)\pi_Y\left(y\right)}{\pi_X(x)} \psi(y)dy.$$

For convenience, we denote its kernel as

$$\pi_{Y|X}(y|x) = \frac{g(y-x)\pi_Y(y)}{\pi_X(x)}.$$

3.2.3. Functional Linear Regression with Possibly Endogenous Regressors

We observe i.i.d. data (Y_i, Z_i, W_i) where each explanatory variable Z_i is a random function element of a Hilbert space $\mathcal{E}$ and W_i is a random function in a Hilbert space $\mathcal{F}$. Let $\langle .,. \rangle$ denote the inner product in $\mathcal{E}$. The response Y_i is generated by the model

$$Y_i = \langle Z_i, \varphi\rangle + u_i, \tag{3.5}$$

where $\varphi \in \mathcal{E}$ and u_i is i.i.d. with zero mean and finite variance. $Y_i \in \mathbb{R}$ and $u_i \in \mathbb{R}$. The regressors are endogenous, but we observe a function W_i that plays the role of instruments so that φ is identified from

$$E(u_i W_i) = 0$$

or equivalently

$$E(Y_i W_i) = E(\langle Z_i, \varphi \rangle W_i).$$

As X_i and W_i are functions, one can think of them as real random variables observed in continuous time. In this setting, $r = E(Y_i W_i)$ is unknown and needs to be estimated, the operator T, defined by $T\varphi = E(\langle Z_i, \varphi \rangle W_i)$, also needs to be estimated. Both estimators converge at a parametric rate to the true values.

This model is considered in Florens and Van Bellegem (2012). In the case where the regressors are exogenous and $W = Z$, this model has been studied by Ramsay and Silverman (1997), Ferraty and Vieu (2000), Cardot and Sarda (2006), and Hall and Horowitz (2007).

3.2.4. Nonparametric Instrumental Regression

We observe an i.i.d. sample $(Y_i, Z_i, W_i) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$ where the relationship between the response Y_i and the vector of explanatory variable Z_i is represented by the equation

$$Y_i = \varphi(Z_i) + u_i. \tag{3.6}$$

We wish to estimate the unknown function φ using as instruments the vector W_i. We assume that

$$E(u_i | W_i) = 0$$

or equivalently

$$E(Y_i | W_i) = E(\varphi(Z_i) | W_i). \tag{3.7}$$

In this setting, $r(w) = E(Y_i | W_i = w)$ is estimated at a slow nonparametric rate (even for a given w) and the operator T defined by $(T\varphi)(w) = E(\varphi(Z) | W = w)$ is also estimated at a slow rate. The identification and estimation of φ has been studied in many recent papers—for example, Newey and Powell (2003), Darolles, Fan, Florens, and Renault (2011), Hall and Horowitz (2005), Blundell, Chen, and Kristensen (2007), Chen and Reiss (2011), and references below. While some authors considered orthogonal series, Darolles et al. (2011) consider a kernel estimator of the conditional expectation. There, the spaces of reference are $\mathcal{E} = L^2_Z$, the space of functions that are square integrable with respect to the true density of Z, similarly $\mathcal{F} = L^2_W$. For these spaces, the adjoint T^* of T is a conditional expectation operator: $(T^*\psi)(z) = E(\psi(W) | Z = z)$, which can also be estimated by kernel. While Darolles et al. (2011) use a Tikhonov regularization where the penalty is on the L_2 norm of φ, Florens, Johannes, and Van Bellegem (2011) and Gagliardini and Scaillet (2012) consider a Tikhonov regularization where the penalty is on a Sobolev norm of φ, that is, the L^2 norm of its derivatives.

3.3. Assumptions and Estimation Method

3.3.1. Ill-Posedness

First we impose some identification conditions on Eq. (3.3).

Assumption 3.1. The solution φ of (3.3) exists and is unique.

The uniqueness condition is equivalent to the condition that T is one-to-one; that is, the null space of T is reduced to zero. As discussed in Newey and Powel (2003), this identification condition in the case of nonparametric IV is, $E(\varphi(Z)|W = w) = 0$ for all w implies $\varphi = 0$, which is equivalent to the completeness in w of the conditional distribution of Z given $W = w$. Interestingly, this condition is not testable, see Canay, Santos, and Shaikh (2013).

Assumption 3.2. T is a linear bounded operator from a Hilbert space $\mathcal{E}$ to a Hilbert space $\mathcal{F}$. Moreover, T is a Hilbert–Schmidt operator.

T is an Hilbert–Schmidt operator if for some (and then any) orthonormal basis $\{e_k\}$, we have $\sum \|Te_k\|^2 < \infty$. It means in particular that its singular values are square summable. It implies that T is compact. Because T is compact, its inverse is unbounded so that the solution φ does not depend continuously on the data. Indeed, if r is replaced by a noisy observation $r + \varepsilon$, then $T^{-1}(r + \varepsilon)$ may be very far from the true $\varphi = T^{-1}r$. Therefore, the solution needs to be stabilized by regularization.

First, we need to define certain spaces of reference to characterize the properties of T and φ.

3.3.2. Hilbert Scales

To obtain the rates of convergence, we need assumptions on φ and T in terms of Hilbert scales. For a review on Hilbert scales, see Krein and Petunin (1966) and Engl, Hanke, and Neubauer (2000). We define L as an unbounded self-adjoint strictly positive operator defined on a dense subset of the Hilbert space $\mathcal{E}$. Let $\mathcal{M}$ be the set of all elements ϕ for which the powers of L are defined; that is, $\mathcal{M} := \bigcap_{k=0}^{\infty} \mathcal{D}(L^k)$, where $\mathcal{D}$ denotes the domain. For all $s \in \mathbb{R}$, we introduce the inner product and norm:

$$\langle \phi, \psi \rangle_s = \left\langle L^s\phi, L^s\psi \right\rangle,$$
$$\|\phi\|_s = \left\| L^s\phi \right\|,$$

where ϕ, $\psi \in \mathcal{M}$. The Hilbert space $\mathcal{E}_s$ is defined as the completion of $\mathcal{M}$ with respect to the norm $\|.\|_s$. $(\mathcal{E}_s)_{s\in\mathbb{R}}$ is called the Hilbert scale induced by L. If $s \geq 0$, then $\mathcal{E}_s = \mathcal{D}(L^s)$. Moreover, for $s \leq s'$, we have $\mathcal{E}_{s'} \subset \mathcal{E}_s$.

A typical example is the case where L is a differential operator. Let $\mathcal{E}$ be the set of complex-valued functions ϕ such that $\int_0^1 |\phi(s)|^2 ds < \infty$. Define the operator I on $\mathcal{E}$ by

$$(I\phi)(t) = \int_0^t \phi(s)ds.$$

Let I^* be the adjoint of I, I^* is such that

$$\left(I^* g\right)(t) = \int_t^1 g(s)ds.$$

Let L be such that $L^{-2} = I^* I$. Then, for $b > 0$, $\phi \in \mathcal{D}(L^b)$ is equivalent to saying that ϕ is b differentiable and satisfies some boundary conditions (e.g., $\phi \in \mathcal{D}(L^2)$ means that ϕ is twice differentiable and $\phi(0) = \phi'(0) = 0$). Note that we could not define $L\phi = \phi'$ because the derivative is not self-adjoint. The construction above gives heuristically $L\phi = \sqrt{-\phi''}$. Indeed, since $L^{-2} = I^* I$, we have $L^2(I^* I)\phi = \phi$. This is satisfied for $L^2\phi = -\phi''$.

The degree of ill-posedness of T is measured by the number a in the following assumption.

Assumption 3.3. T satisfies

$$\underline{m}\, \|\phi\|_{-a} \leq \|T\phi\| \leq \overline{m}\, \|\phi\|_{-a}$$

for any $\phi \in \mathcal{E}$ and some $a > 0$, $0 < \underline{m} < \overline{m} < \infty$.

Assumption 3.4. $\varphi \in \mathcal{E}_b$, for some $b > 0$.

In our example of a differential operator, Assumption 3.4 is equivalent to φ is b differentiable.

Let $B = TL^{-s}$, $s \geq 0$. According to Corollary 8.22 of Engl et al. (2000), for $|\nu| < 1$, we have

$$\underline{c}(\nu)\, \|\phi\|_{-\nu(a+s)} \leq \left\| \left(B^* B\right)^{\nu/2} \phi \right\| \leq \bar{c}(\nu)\, \|\phi\|_{-\nu(a+s)} \tag{3.8}$$

for any $\phi \in \mathcal{D}((B^* B)^{\nu/2})$, with $\underline{c}(\nu) = \min(\underline{m}^\nu, \overline{m}^\nu)$ and $\bar{c}(\nu) = \max(\underline{m}^\nu, \overline{m}^\nu)$. Moreover,

$$\mathcal{R}((B^* B)^{\nu/2}) = \mathcal{E}_{\nu(a+s)}, \tag{3.9}$$

where $(B^* B)^{\nu/2}$ has to be replaced by its extension to $\mathcal{E}$ if $\nu < 0$.

It is useful to make the link between Assumptions 3.3 and 3.4 and the source condition given in Carrasco, Florens, and Renault (2007, Definition 3.4). This condition is written in terms of the singular system of T denoted $(\lambda_j, \phi_j, \psi_j)$:

$$\sum_{j=1}^{\infty} \frac{\left\langle \varphi, \phi_j \right\rangle^2}{\lambda_j^{2\beta}} < \infty. \tag{3.10}$$

This means that $\varphi \in \mathcal{R}((T^*T)^{\beta/2})$ or equivalently $\varphi \in \mathcal{D}((T^*T)^{-\beta/2})$. If we let $L = (T^*T)^{-1/2}$, we see that Assumption 3.3 holds with $a = 1$. Then Assumption 3.4 is equivalent to (3.10) with $\beta = b$. Another interpretation is the following. Using (3.9), we see that $\mathcal{R}((T^*T)^{\beta/2}) = \mathcal{E}_{\beta a}$. Hence, Assumptions 3.3 and 3.4 with $b = \beta a$ imply the source condition (3.10). While the condition (3.10) relates the properties of φ and T directly, Assumptions 3.3 and 3.4 characterize the properties of φ and T with respect to an auxiliary operator L.

3.3.3. Regularization and Estimation

Because the inverse of T is not continuous, some regularization is needed. The most common one is Tikhonov regularization, which consists in penalizing the norm of φ:

$$\min_{\varphi} \left\| T\varphi - \hat{r} \right\|^2 + \alpha \left\| \varphi \right\|^2 .$$

We will consider a more general case where we penalize the $\mathcal{E}_s$ norm of φ:

$$\min_{\varphi \in \mathcal{E}_s} \left\| T\varphi - \hat{r} \right\|^2 + \alpha \left\| \varphi \right\|_s^2 . \tag{3.11}$$

The reason to do this is twofold. Assuming that L is a differential operator and φ is known to be s times differentiable, we may want to dampen the oscillations of $\hat{\varphi}$ by penalizing its derivatives. Second, if we are interested in estimating $L^c\varphi$ for some $0 < c < s$, then we immediately obtain an estimator $\widehat{L^c\varphi} = L^c\hat{\varphi}$ and its rate of convergence.

The solution to (3.11) is given by

$$\begin{aligned}\hat{\varphi} &= \left(\alpha L^{2s} + T^*T\right)^{-1} T^*\hat{r} \\ &= L^{-s}(\alpha I + L^{-s}T^*TL^{-s})^{-1}L^{-s}T^*\hat{r} \\ &= L^{-s}(\alpha I + B^*B)^{-1}B^*\hat{r}, \end{aligned} \tag{3.12}$$

where $B = TL^{-s}$.

We also consider other regularization schemes. Let us define the regularized solution to (3.3) as

$$\hat{\varphi} = L^{-s} g_\alpha(B^*B)B^*\hat{r}, \tag{3.13}$$

where g_α:$[0, \|B\|^2] \to \mathbb{R}$, $\alpha > 0$, is a family of piecewise continuous functions and

$$\lim_{\alpha \to 0} g_\alpha(\lambda) = \frac{1}{\lambda}, \qquad \lambda \neq 0,$$

$$\left| g_\alpha(\lambda) \right| \leq \widehat{c}\alpha^{-1}, \tag{3.14}$$

$$\lambda^\mu \left| 1 - \lambda g_\alpha(\lambda) \right| \leq c_\mu \alpha^\mu, \qquad 0 \leq \mu \leq \mu_0, \tag{3.15}$$

with $\widehat{c}$ and $c_\mu > 0$ independent of α and $\mu_0 \geq 1$. The main examples of functions g_α are the following.

1. The Tikhonov regularization is given by $g_\alpha(\lambda) = 1/(\lambda + \alpha)$.
2. The iterated Tikhonov regularization of order m is given by $g_\alpha(\lambda) = (1 - (\alpha/(\lambda + \alpha))^m)/\lambda$. The solution is obtained after m iterative minimizations:

$$\hat{\varphi}_j = \arg\min_{\phi \in \mathcal{E}_s} \left\| T\phi - \hat{r} \right\|^2 + \alpha \left\| \phi - \hat{\varphi}_{j-1} \right\|_s^2, \qquad j = 1, \ldots, m,\ \hat{\varphi}_0 = 0.$$

3. The spectral cutoff considers $g_\alpha(\lambda) = 1/\lambda$ for $\lambda \geq \alpha$.
4. The Landweber–Fridman regularization takes $g_\alpha(\lambda) = (1 - (1 - \lambda)^{1/\alpha})/\lambda$.

When B is unknown, we replace B by a consistent estimator $\hat{B}$ and B^* by $(\hat{B})^*$. The convergence of $\hat{\varphi}$ is studied in Engl et al. (2000), Carrasco et al. (2007), Chen and Reiss (2011), and Johannes, Van Bellegem, and Vanhems (2011).

3.3.4. Rate of Convergence of MSE

Here we study the mean square error (MSE) of $\hat{\varphi}$ when B is known. When B is estimated, the error due to its estimation usually goes to zero faster than the other terms and does not affect the convergence rate of the bias (see Carrasco et al. (2007)).

To simplify the exposition, we first let $s = c = 0$ and consider Tikhonov regularization. The general case is discussed at the end. The difference $\hat{\varphi} - \varphi$ can be decomposed as the following sum:

$$\hat{\varphi} - \varphi = \hat{\varphi} - \varphi_\alpha + \varphi_\alpha - \varphi,$$

where

$$\varphi_\alpha = \left(\alpha I + T^* T\right)^{-1} T^* T \varphi.$$

The term $\hat{\varphi} - \varphi_\alpha$ corresponds to an estimation error whereas the term $\varphi_\alpha - \varphi$ corresponds to a regularization bias. We first examine the latter (see Groetsch (1993)).

$$\begin{aligned}\varphi_\alpha - \varphi &= \sum_j \frac{\lambda_j^2}{\lambda_j^2 + \alpha} \langle \varphi, \varphi_j \rangle \varphi_j - \sum_j \langle \varphi, \varphi_j \rangle \varphi_j \\ &= -\alpha \sum_j \frac{1}{\lambda_j^2 + \alpha} \langle \varphi, \varphi_j \rangle \varphi_j.\end{aligned}$$

Given

$$\begin{aligned}\|\varphi_\alpha - \varphi\|^2 &= \alpha^2 \sum_j \frac{1}{\left(\lambda_j^2 + \alpha\right)^2} \left\langle \varphi, \varphi_j \right\rangle^2 \\ &\leq \sum_j \left\langle \varphi, \varphi_j \right\rangle^2 < \infty, \end{aligned} \tag{3.16}$$

we may, in passing to the limit as α goes to zero in (3.16), interchange the limit and the summation yielding

$$\lim_{\alpha \to 0} \|\varphi_\alpha - \varphi\|^2 = 0.$$

From this result, we understand that we cannot obtain a rate of convergence for $\|\varphi_\alpha - \varphi\|^2$ unless we impose more restrictions on φ. Assume that φ satisfies the source condition (3.10) for some $\beta > 0$, then

$$\begin{aligned}\|\varphi_\alpha - \varphi\|^2 &\leq \sup_\lambda \frac{\alpha^2 \lambda^{2\beta}}{\left(\lambda^2 + \alpha\right)^2} \sum_j \frac{\left\langle \varphi, \varphi_j \right\rangle^2}{\lambda_j^{2\beta}} \\ &= O\left(\alpha^{\beta \wedge 2}\right)\end{aligned}$$

by Kress (1999) and Carrasco and al. (2007).

We now turn to the estimation error. There are two ways to characterize the rate of convergence of $\|\hat{\varphi} - \varphi_\alpha\|^2$, depending on whether we have an assumption on $\|r - \hat{r}\|^2$ or $\|T^*(r - \hat{r})\|^2$. First we consider the rate of $\|\hat{\varphi} - \varphi_\alpha\|^2$ in terms of $\|r - \hat{r}\|^2$. We have

$$\begin{aligned}\hat{\varphi} - \varphi_\alpha &= \left(\alpha I + T^* T\right)^{-1} T^* \left(T\varphi - \hat{r}\right) \\ &= T^* \left(\alpha I + TT^*\right)^{-1} \left(T\varphi - \hat{r}\right), \\ \left\|\hat{\varphi} - \varphi_\alpha\right\|^2 &= \left\langle T^* \left(\alpha I + TT^*\right)^{-1} \left(T\varphi - \hat{r}\right), T^* \left(\alpha I + TT^*\right)^{-1} \left(T\varphi - \hat{r}\right)\right\rangle \\ &= \left\langle \left(\alpha I + TT^*\right)^{-1} \left(T\varphi - \hat{r}\right), TT^* \left(\alpha I + TT^*\right)^{-1} \left(T\varphi - \hat{r}\right)\right\rangle.\end{aligned}$$

Moreover,

$$\begin{aligned}\left\|\left(\alpha I + TT^*\right)^{-1}\right\| &\leq \frac{1}{\alpha}, \\ \left\|TT^* \left(\alpha I + TT^*\right)^{-1}\right\| &\leq 1.\end{aligned}$$

Hence,

$$\left\|\hat{\varphi} - \varphi_\alpha\right\|^2 \leq \frac{1}{\alpha} \left\|r - \hat{r}\right\|^2.$$

In summary, the MSE of $\hat{\varphi}$ is bounded in the following way:

$$E\left(\left\|\hat{\varphi} - \varphi\right\|^2\right) \leq \frac{1}{\alpha} E(\|r - \hat{r}\|^2) + C\alpha^{\beta \wedge 2} \tag{3.17}$$

for some constant C.

Second, we consider the rate of $\|\hat{\varphi}-\varphi_\alpha\|^2$ in terms of $\|T^*(\hat{r}-r)\|^2$.

$$\begin{aligned}\left\|\hat{\varphi}-\varphi_\alpha\right\|^2 &\leq \left\|\left(\alpha I+T^*T\right)^{-1}\right\|^2 \left\|T^*\left(T\varphi-\hat{r}\right)\right\|^2\\ &\leq \frac{1}{\alpha^2}\left\|T^*\left(r-\hat{r}\right)\right\|^2.\end{aligned}$$

The MSE of $\hat{\varphi}$ is bounded in the following way:

$$E\left(\left\|\hat{\varphi}-\varphi\right\|^2\right) \leq \frac{1}{\alpha^2}E(\|T^*(r-\hat{r})\|^2)+C\alpha^{\beta\wedge 2}. \tag{3.18}$$

In both expressions (3.17) and (3.18), there is a tradeoff between the regularization bias that declines as α goes to zero and the variance that increases as α goes to zero. The optimal α is selected so that the rate of the regularization bias equals that of the variance.

These results generalize to the other three regularization techniques described earlier. In the case of Spectral cutoff, Landweber–Fridman, and iterated Tikhonov regularizations, the rate of $\|\varphi_\alpha-\varphi\|^2$ is $O(\alpha^\beta)$. In the case of Tikhonov with $\beta<2$, it is also $O(\alpha^\beta)$. So the rates given below apply to the four methods. The optimal α is chosen so that $\alpha^{\beta+1}=E(\|r-\hat{r}\|^2)$ or $\alpha^{\beta+2}=E(\|T^*(r-\hat{r})\|^2)$, hence

$$E\left(\left\|\hat{\varphi}-\varphi\right\|^2\right)=O\left(\min\left(E(\|r-\hat{r}\|^2)^{\beta/(\beta+1)},E(\|T^*(r-\hat{r})\|^2)^{\beta/(\beta+2)}\right)\right). \tag{3.19}$$

We can see that, for the optimal α, $\sqrt{n}\|\varphi_\alpha-\varphi\|$ diverges so that there is an asymptotic bias remaining when studying the asymptotic distribution of $\sqrt{n}(\hat{\varphi}-\varphi)$.

We can analyze the rate of (3.19) in different scenarios.

- If $r-\hat{r}$ converges at a parametric rate $\sqrt{n}$, then $T^*(r-\hat{r})$ also converges at a parametric rate and the first term of the r.h.s of (3.19) converges to 0 faster than the second term. Thus the rate of the MSE is given by $n^{-\beta/(\beta+1)}$.
- If $r-\hat{r}$ converges at a nonparametric rate so that $\|r-\hat{r}\|^2=O_p(n^{-2\nu})$ with $\nu<1/2$ and $\|T^*(T\varphi-\hat{r})\|^2=O_p(n^{-1})$ and, moreover, $2\nu<(\beta+1)/(\beta+2)$, then the second term in the r.h.s of (3.19) converges to 0 faster than the first term. Thus the rate of the MSE is given by $n^{-\beta/(\beta+2)}$. This is encountered in nonparametric IV; see, for example, Darolles et al. (2011). There, $r=E(Y|W)$ and $\nu=d/(2d+q)$, where q is the dimension of W and d is the number of derivatives of $E(Y|W)$. If $\beta=2$, $d=2$, and $q\geq 2$, then the condition $2\nu<(\beta+1)/(\beta+2)$ holds. See also Chen and Reiss (2011) and Johannes et al. (2011).

So far, we derived the rate of convergence of the MSE using a source condition (3.10). Now we establish the results using assumptions on the degree of ill-posedness of T. Suppose, moreover, that we are interested in estimating the derivative of φ, $L^c\varphi$.

Proposition 3.1. *Assume that T satisfies Assumption 3.3, φ satisfies Assumption 3.4 with $b \leq a+2s$, and $\hat{\varphi}$ is defined as in (3.13). Then, for the optimal α, we have*

$$E\Big(\left\|L^c\hat{\varphi}-L^c\varphi\right\|^2\Big)=O\Big(\min\Big(E\Big(\left\|r-\hat{r}\right\|^2\Big)^{(b-c)/(a+b)},\\ E\Big(\left\|B^*\left(r-\hat{r}\right)\right\|^2\Big)^{(b-c)/(b+2a+s)}\Big)\Big).$$

Setting $c=s=0$, we see that this result is the same as the rates (3.19) obtained with the source condition (3.10) and $\beta=b/a$.

Proof of Proposition 3.1. We follow the steps of the proof of Engl, Hanke, and Neubauer (2000, Theorem 8.23). Note that by (3.14) and (3.15) we obtain

$$\lambda^t\left|g_\alpha(\lambda)\right|\leq C\alpha^{t-1}, \tag{3.20}$$

where C denotes a generic constant. We have

$$\begin{aligned}\left\|L^c\left(\hat{\varphi}-\varphi_\alpha\right)\right\| &= \left\|L^{(c-s)}g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right)\right\|\\ &= \left\|g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right)\right\|_{c-s}\\ &\leq C\left\|\left(B^*B\right)^{\frac{s-c}{2(a+s)}}g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right)\right\|,\end{aligned} \tag{3.21}$$

where the inequality follows from inequality (3.8) with $\nu=(s-c)/(a+s)$ and $\phi=g_\alpha(B^*B)B^*(\hat{r}-r)$. Note that

$$\begin{aligned}&\left\|\left(B^*B\right)^{\frac{s-c}{2(a+s)}}g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right)\right\|^2\\ &=\left\langle\left(B^*B\right)^{\frac{s-c}{2(a+s)}}g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right),\left(B^*B\right)^{\frac{s-c}{2(a+s)}}g_\alpha\left(B^*B\right)B^*\left(\hat{r}-r\right)\right\rangle\\ &\leq\left\|BB^*g_\alpha\left(BB^*\right)\left(\hat{r}-r\right)\right\|\left\|\left(BB^*\right)^{\frac{s-c}{(a+s)}}g_\alpha\left(BB^*\right)\left(\hat{r}-r\right)\right\|\\ &\leq C\alpha^{-(a+c)/(a+s)}\left\|\hat{r}-r\right\|^2,\end{aligned}$$

where the last inequality follows from (3.20). Hence,

$$\left\|\hat{\varphi}-\varphi_\alpha\right\|\leq C\left\|T\varphi-\hat{r}\right\|\alpha^{-(a+c)/(2(a+s))}.$$

Another majoration follows from (3.21) and (3.20):

$$\begin{aligned}\left\|L^c\left(\hat{\varphi}-\varphi_\alpha\right)\right\| &\leq C\left\|\left(B^*B\right)^{\frac{s-c}{2(a+s)}}g_\alpha\left(B^*B\right)\right\|\left\|B^*\left(\hat{r}-r\right)\right\|\\ &\leq C\alpha^{-(c+2a+s)/(2(a+s))}\left\|B^*\left(\hat{r}-r\right)\right\|.\end{aligned}$$

We turn our attention to the bias term. Note that $L^s\varphi \in \mathcal{E}_{b-s}$. By Eq. (3.9), there is a function $\rho \in \mathcal{E}$ such that

$$L^s\varphi = \left(B^*B\right)^{(b-s)/(2(a+s))}\rho.$$

We have

$$\begin{aligned}\left\|L^c\left(\varphi_\alpha - \varphi\right)\right\| &= \left\|L^{(c-s)}\left(g_\alpha\left(B^*B\right)B^*B - I\right)L^s\varphi\right\| \\ &= \left\|\left(g_\alpha\left(B^*B\right)B^*B - I\right)\left(B^*B\right)^{(b-s)/(2(a+s))}\rho\right\|_{c-s} \\ &\le \left\|\left(B^*B\right)^{(s-c)/(2(a+s))}\left(g_\alpha\left(B^*B\right)B^*B - I\right)\left(B^*B\right)^{(b-s)/(2(a+s))}\rho\right\| \\ &= \left\|\left(B^*B\right)^{(b-c)/(2(a+s))}\left(g_\alpha\left(B^*B\right)B^*B - I\right)\rho\right\| \\ &\le C'\alpha^{(b-c)/(2(a+s))}\left\|\rho\right\|,\end{aligned}$$

for some constant C', where the first inequality follows from (3.8) with $\nu = (s-c)/(a+s)$ and $\phi = (g_\alpha(B^*B)B^*B - I)(B^*B)^{(b-s)/(2(a+s))}\rho$ and the second inequality follows from (3.15) with $\mu = (b-c)/(2(a+s))$. Then using the optimal α, we obtain the rates given in Proposition 3.1. ■

3.4. Asymptotic Normality for Fixed α

Let φ be the true value. As seen in Section 3.3, the estimator $\hat{\varphi}$ defined in (3.13) has a bias which does not vanish. For testing, it is useful to fix α and use $\hat{\varphi}$ minus a regularized version of φ:

$$\varphi_\alpha = L^{-s}g_\alpha\left(B^*B\right)B^*T\varphi = L^{-s}g_\alpha\left(B^*B\right)B^*r. \tag{3.22}$$

Then, we have

$$\hat{\varphi} - \varphi_\alpha = L^{-s}g_\alpha\left(B^*B\right)B^*\left(\hat{r} - r\right).$$

Depending on the examples, we will assume either Assumption 3.5a or Assumption 3.5b below.

Assumption 3.5a. $\sqrt{n}(\hat{r} - r) \Rightarrow \mathcal{N}(0,\Omega)$ in $\mathcal{F}$.

Under Assumption 3.5a, we have for a fixed α

$$\sqrt{n}\left(\hat{\varphi} - \varphi_\alpha\right) \Rightarrow \mathcal{N}(0,\Sigma) \tag{3.23}$$

with $\Sigma = L^{-s}g_\alpha(B^*B)B^*\Omega Bg_\alpha(B^*B)L^{-s}$.

Assumption 3.5b. $\sqrt{n}B^*(\hat{r}-r) \Rightarrow \mathcal{N}(0,\Omega)$ in $\mathcal{F}$.

Under Assumption 3.5b, we have for a fixed α

$$\sqrt{n}\left(\hat{\varphi}-\varphi_\alpha\right) \Rightarrow \mathcal{N}(0,\Sigma) \tag{3.24}$$

with $\Sigma = L^{-s}g_\alpha(B^*B)\Omega g_\alpha(B^*B)L^{-s}$.

The results (3.23) and (3.24) are the basis to construct the test statistics of the next section. If T is unknown, we have an extra term corresponding to $\hat{T}-T$ which is negligible provided $\hat{T}$ converges sufficiently fast. We can check that either Assumption 3.5a or 3.5b is satisfied and the asymptotic variance Ω (and hence Σ) is estimable in all the examples considered here.

Example 3.1. Density. We have

$$\hat{r}-r=\hat{F}-F=\frac{1}{n}\sum_{i=1}^{n}\left[I(x_i\leq t)-F(t)\right],$$

$$\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\left[I(x_i\leq t)-F(t)\right] \Rightarrow \mathcal{N}(0,F(t\wedge s)-F(t)\,F(s)).$$

This example satisfies Assumption 3.5a. Here the asymptotic variance of $\hat{r}-r$ can be estimated using the empirical cumulative distribution function.

Example 3.2. Deconvolution. Following Carrasco and Florens (2011), we have

$$\widehat{T^*r}-T^*r=\frac{1}{n}\sum_{i=1}^{n}\left(\pi_{Y|X}\left(y_i|x\right)-E\left(\pi_{Y|X}\left(Y|x\right)\right)\right).$$

Here a slight modification of Assumption 3.5b is satisfied. Since $\pi_{Y|X}$ is known, the variance of $\widehat{T^*r}-T^*r$ can be estimated using the empirical variance.

Example 3.3. Functional Linear Regression. We have

$$\hat{r}=\frac{1}{n}\sum_{i=1}^{n}Y_iW_i,$$

$$E(\widehat{r})=r.$$

Thus Assumption 3.5a holds and

$$V\left(\hat{r}-r\right)=\frac{1}{n}V(Y_iW_i)$$

can be estimated using the sample variance of Y_iW_i.

Example 3.4. Nonparametric Instrumental Regression. Following Darolles, Florens, and Renault (2002, Assumption A7), we assume that

$$\sqrt{n}\left(\widehat{T}^*\widehat{r}-\widehat{T}^*\widehat{T}\varphi\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_i-\varphi\left(Z_i\right)\right)\frac{f_{Z,W}\left(Z,W_i\right)}{f_Z\left(Z\right)f_W\left(W_i\right)}+h_n^{\rho}\Gamma, \tag{3.25}$$

where the term $h_n^{\rho}\Gamma$ is negligible provided that the bandwidth h_n is sufficiently small, which is consistent with Assumption 3.5b. We denote the leading term in (3.25) by $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i$. We have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\eta_i\Rightarrow N\left(0,\sigma^2T^*T\right),$$

where $\sigma^2=V(Y-\varphi(Z)|W)$. An estimate of σ^2 can be obtained using a first step estimator of φ.

3.5. Test Statistics

3.5.1. Case Where φ_0 is Fully Specified

We want to test $H_0:\varphi=\varphi_0$ where φ_0 is fully specified. A test can be based on the difference between $\hat{\varphi}$ and $\varphi_{0\alpha}$ defined in (3.22). We can construct a Kolmogorov–Smirnov test

$$\sup_z\sqrt{n}\left|\hat{\varphi}\left(z\right)-\varphi_{0\alpha}(z)\right|$$

or a Cramer–Von Mises test

$$\left\|\sqrt{n}\left(\hat{\varphi}-\varphi_{0\alpha}\right)\right\|^2.$$

Using (3.24), we have

$$\left\|\sqrt{n}\left(\hat{\varphi}-\varphi_{0\alpha}\right)\right\|^2\Rightarrow\sum_{j=1}^{\infty}\tilde{\lambda}_j\chi_j^2\left(1\right),$$

where χ_j^2 are independent chi-square random variables and $\tilde{\lambda}_j$ are the eigenvalues of Σ. As Σ is estimable, $\tilde{\lambda}_j$ can be estimated by the eigenvalues of the estimate of Σ; see, for instance, Blundell and Horowitz (2007).

Another testing strategy consists in using a test function δ and basing the test on a rescaled version of $\sqrt{n}\langle\hat{\varphi}-\varphi_{0\alpha},\delta\rangle$ to obtain a standard distribution.

$$\xi_n=\frac{\sqrt{n}\left\langle\hat{\varphi}-\varphi_{0\alpha},\delta\right\rangle}{\left\langle\widehat{\Sigma}\delta,\delta\right\rangle^{1/2}}\xrightarrow{d}\mathcal{N}(0,1). \tag{3.26}$$

A more powerful test can be obtained by considering a vector

$$\begin{pmatrix} \sqrt{n}\left\langle \hat{\varphi} - \varphi_{0\alpha}, \delta_1 \right\rangle \\ \vdots \\ \sqrt{n}\left\langle \hat{\varphi} - \varphi_{0\alpha}, \delta_q \right\rangle \end{pmatrix}$$

for a given family δ_l, $l = 1, 2, \ldots, q$, of linearly independent test functions of $\mathcal{E}$. This vector converges to a q-dimensional normal distribution. The covariance between the various components of the vector can be easily deduced from (3.26) since it holds for any linear combinations of test functions δ_l and δ_h, $l \neq h$, chosen in the same space. Then, the appropriately rescaled statistic asymptotically follows a chi-square distribution with q degrees of freedom.

3.5.2. Case Where φ_0 is Parametrically Specified

We want to test H_0: $\exists \theta \in \Theta$, $\varphi(.) = h(., \theta)$, where h is a known function. Assume that we have an estimator of θ, $\hat{\theta}$, such that $\sqrt{n}(\hat{\theta} - \theta_0) \overset{d}{\to} \mathcal{N}(0, V)$. Then, a test statistic can be based on $\sqrt{n}(\hat{\varphi} - h_\alpha(\hat{\theta}))$, where h_α is a regularized version of $h(\hat{\theta})$, namely

$$h_\alpha\left(\hat{\theta}\right) = L^{-s} g_\alpha\left(B^* B\right) B^* T h\left(\hat{\theta}\right).$$

This testing strategy permits us to eliminate the bias and is very similar in spirit to the twin-smoothing first proposed by Härdle and Mammen (1993) to test a parametric regression model against a nonparametric alternative (see also Fan (1994) and Altissimo and Mele (2009)).

Because θ is estimated, the asymptotic variance of $\sqrt{n}(\hat{\varphi} - h_\alpha(\hat{\theta}))$ will differ from that of $\sqrt{n}(\hat{\varphi} - \varphi_{0\alpha})$. We illustrate this point on two examples. In both examples, h is specified as $h(\theta) = \sum_{d=1}^{D} \theta_d e_d$, where e_d, $d = 1, \ldots, D$, are known functions and θ_d, $d = 1, \ldots, D$, are unknown scalars.

Functional Linear Regression

Consider the model (3.5) with exogenous regressors and homoskedastic error, $E(u_i Z_i) = 0$ and $\mathrm{V}(u_i|Z_i) = \sigma^2$. Replacing φ by $h(\theta)$ in (3.5), we obtain

$$Y_i = \sum_{d=1}^{D} \theta_d \langle Z_i, e_d \rangle + u_i.$$

Denote $x_{i,d} = \langle Z_i, e_d \rangle$, X the $n \times D$ matrix of $x_{i,d}$, e the $D \times 1$ vector of e_d, and Y the $n \times 1$ vector of Y_i. Then, θ can be estimated by the OLS estimator, $\hat{\theta} = (X'X)^{-1} X'Y$. The estimator of $h(\theta)$ is given by

$$h\left(\hat{\theta}\right) = e'\left(X'X\right)^{-1} X'Y.$$

Consider standard Tikhonov regularization

$$\hat{\varphi} - h_\alpha\left(\hat{\theta}\right) = \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\hat{r} \\ - \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\widehat{T}h\left(\hat{\theta}\right).$$

Replacing $\hat{r}$ by $\frac{1}{n}\sum_{i=1}^n Z_iY_i = \frac{1}{n}\sum_{i=1}^n Z_i\langle Z_i, \varphi\rangle + \frac{1}{n}\sum_{i=1}^n Z_iu_i = \widehat{T}\varphi + \frac{1}{n}\sum_{i=1}^n Z_iu_i$ and $h(\hat{\theta}) = e'\theta + e'(X'X)^{-1}X'u = \varphi + e'(X'X)^{-1}X'u$ (under H_0), we have

$$\hat{\varphi} - h_\alpha\left(\hat{\theta}\right) = \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\left(\frac{1}{n}\sum_{i=1}^n Z_iu_i\right) \\ - \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1}\widehat{T}^*\widehat{T}e'\left(\frac{X'X}{n}\right)^{-1}\frac{1}{n}X'u.$$

Let us denote $A_n = (\alpha I + \widehat{T}^*\widehat{T})^{-1}\widehat{T}^*$ and $B_n = -(\alpha I + \widehat{T}^*\widehat{T})^{-1}\widehat{T}^*\widehat{T}e'(\frac{X'X}{n})^{-1}$. We obtain

$$\hat{\varphi} - h_\alpha\left(\hat{\theta}\right) = [A_n\ B_n]\frac{1}{n}\sum_{i=1}^n \begin{pmatrix} Z_i \\ X_i \end{pmatrix} u_i.$$

Provided that $E\|\begin{pmatrix} Z \\ X \end{pmatrix} u\|^2 < \infty$, we know from van der Vaart and Wellner (1996) that a central limit theorem holds, so that

$$\frac{\sqrt{n}}{n}\sum_{i=1}^n \begin{pmatrix} Z_i \\ X_i \end{pmatrix} u_i \Rightarrow \mathcal{N}(0, \Gamma).$$

If, moreover, $\|A_n - A\| \overset{P}{\to} 0$ and $\|B_n - B\| \overset{P}{\to} 0$, we have

$$\sqrt{n}\left(\hat{\varphi} - h_\alpha\left(\hat{\theta}\right)\right) \Rightarrow \mathcal{N}\left(0, [A\ B]\,\Gamma\begin{bmatrix} A^* \\ B^* \end{bmatrix}\right).$$

Nonparametric Instrumental Regression

Consider the model (3.6). The null hypothesis of interest is again H_0: $\exists\theta \in \Theta$, $\varphi(.) = h(., \theta) = \sum_{d=1}^D \theta_d e_d$ for some known functions e_d. The finite-dimensional parameter θ can be estimated by two-stage least squares. Denote W the $n \times q$ matrix $(W_1', \ldots, W_n')'$, Y the $n \times 1$ matrix $(Y_1, \ldots, Y_n)'$, and E the $n \times d$ matrix with (i,j) elements $e_d(Z_i)$. Denote P_W the projection matrix on W, $P_W = W(W'W)^{-1}W'$. Then the two-stage least squares estimator of θ is

$$\hat{\theta} = \left(E'P_WE\right)^{-1}E'P_WY \equiv M\frac{W'Y}{n}.$$

Using the notation e for the $D \times 1$ vector of functions $(e_1, \ldots, e_d, \ldots, e_D)'$, $h(.,\hat{\theta})$ takes the simple form

$$h\left(\hat{\theta}\right) = e'\hat{\theta}.$$

Similarly to the previous example, we have

$$\begin{aligned}\sqrt{n}\left(\hat{\varphi} - h_\alpha\left(\hat{\theta}\right)\right) = \left(\alpha I + \widehat{T}^*\widehat{T}\right)^{-1} &\left[\frac{1}{n}\sum_{i=1}^{n} \frac{f_{Z,W}(Z,W_i)}{f_Z(Z) f_W(W_i)} u_i \right. \\ &\left. - \widehat{T}^*\widehat{T}e'M\left(\frac{1}{n}\sum_{i=1}^{n} W_i u_i\right)\right] + o_p(1).\end{aligned}$$

Under some mild conditions (see van der Vaart and Wellner, 1996),

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\begin{array}{c}\frac{f_{Z,W}(Z,W_i)}{f_Z(Z)f_W(W_i)} \\ W_i\end{array}\right] u_i$$

converges to a Gaussian process, which permits us to establish the asymptotic variance of $\sqrt{n}(\hat{\varphi} - h_\alpha(\hat{\theta}))$.

This test procedure can be related to that of Horowitz (2006). The test proposed by Horowitz (2006) is based on $\|\widehat{T}^*(\hat{r} - \widehat{T}h(\hat{\theta}))\|^2$, while our test is based on $\|(\alpha I + \widehat{T}^*\widehat{T})^{-1}\widehat{T}^*(\hat{r} - \widehat{T}h(\hat{\theta}))\|^2$ with a fixed α.

3.6. Asymptotic Normality for Vanishing α

In this section, we are looking at conditions under which $\hat{\varphi} - \varphi$ is asymptotically normal when α goes to zero. There are various ways to state the results.

Carrasco and Florens (2011) and Horowitz (2007) prove a pointwise convergence:

$$\frac{\hat{\varphi}(z) - \varphi(z)}{\sqrt{\hat{V}(z)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where typically the rate of convergence depends on z. Another possibility is to focus on the convergence of inner products:

$$\frac{\sqrt{n}\langle\hat{\varphi} - \varphi - b_n, \delta\rangle}{\langle\Sigma\delta,\delta\rangle^{1/2}} \xrightarrow{d} \mathcal{N}(0,1),$$

where b_n is the bias corresponding to $\varphi_\alpha - \varphi$ and $\langle\Sigma\delta,\delta\rangle$ may be finite or infinite depending on the regularity of φ and δ.

We are going to focus on the second case.

3.6.1. Asymptotic Normality with Known Operator

Here we consider the case of the Tikhonov regularization where T (hence B) is known. The case where B is estimated is studied in the next subsection.

We want to prove the asymptotic normality of $\sqrt{n}\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle$, where $c < s$ and $\hat{\varphi}$ is defined in (3.12):

$$\varphi_\alpha = L^{-s}\big(\alpha I + B^*B\big)^{-1}B^*r,$$

$$\hat{\varphi} - \varphi_\alpha = L^{-s}\big(\alpha I + B^*B\big)^{-1}L^{-s}T^*\big(\hat{r} - T\varphi\big).$$

The following assumption will be used to strengthen Assumption 3.5a.

Assumption 3.6. η_i, $i = 1, 2, \ldots, n$, are i.i.d. with mean 0 and variance Ω and satisfy a functional CLT:

$$\frac{\sum_{i=1}^n \eta_i}{\sqrt{n}} \Rightarrow \mathcal{N}(0,\Omega) \qquad \text{in } \mathcal{F}.$$

Define M such that $M = L^{c-s}(\alpha I + B^*B)^{-1}L^{-s}$.

Proposition 3.2. *Suppose that $\hat{\varphi}$ is as in (3.12). Assume that $T^*(\hat{r} - T\varphi) = \sum_{i=1}^n \eta_i/n$, where η_i satisfies Assumption 3.6. If $\delta \in \mathcal{E}$ satisfies*

$$\frac{E\big[|\langle M\eta_i, \delta\rangle|^{2+\varepsilon}\big]}{\big\|\Omega^{1/2}M^*\delta\big\|^{2+\varepsilon}} = O(1) \tag{3.27}$$

for some $\varepsilon > 0$, then

$$\frac{\sqrt{n}\big\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\big\rangle}{\big\|\Omega^{1/2}M^*\delta\big\|} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.28}$$

Proof of Proposition 3.2. We have

$$\sqrt{n}\big\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\big\rangle = \frac{\sqrt{n}}{n}\sum_{i=1}^n \langle M\eta_i, \delta\rangle.$$

It follows from Assumption 3.6 that $\langle M\eta_i, \delta\rangle$ are i.i.d. with $\mathrm{Var}(\langle M\eta_i, \delta\rangle) = \frac{1}{n}\langle M\Omega M^*\delta, \delta\rangle = \frac{1}{n}\|\Omega^{1/2}M^*\delta\|^2$. A sufficient condition for the asymptotic normality is the Lyapunov condition (Billingsley, 1995, (27.16)):

$$\lim_n \sum_{i=1}^n \frac{E\left[\left(\frac{\sqrt{n}}{n}|\langle M\eta_i, \delta\rangle|\right)^{2+\varepsilon}\right]}{\big\|\Omega^{1/2}M^*\delta\big\|^{2+\varepsilon}} = 0$$

for some $\varepsilon > 0$. By the stationarity, this relation simplifies to

$$\lim_n \frac{E\left[|\langle M\eta_i, \delta\rangle|^{2+\varepsilon}\right]}{n^{\varepsilon/2}\left\|\Omega^{1/2}M^*\delta\right\|^{2+\varepsilon}} = 0. \tag{3.29}$$

A sufficient condition for (3.29) is given by (3.27). The result follows. ■

The rate of convergence of $\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle$ will be slower than $\sqrt{n}$ if $\|\Omega^{1/2}M^*\delta\|$ diverges (which is the usual case). Moreover, the rate of convergence depends on the regularity of δ. The case of a $\sqrt{n}$ rate of convergence is discussed in Section 3.6.3. We see that condition (3.27) imposes in general restrictions on both η and δ.

First, we are going to investigate cases where condition (3.27) is satisfied for all δ. Assume that there exists μ_i such that

$$L^{-s}\eta_i = B^*B\mu_i. \tag{3.30}$$

This is equivalent to, say, that $L^{-s}\eta_i \in \mathcal{R}(B^*B) = \mathcal{D}(L^{2(a+s)})$ or equivalently $\eta_i \in \mathcal{D}(L^{2a+s})$. Under assumption (3.30), we have

$$\begin{aligned}|\langle M\eta_i, \delta\rangle| &= \left|\left\langle L^{c-s}\left(\alpha I + B^*B\right)^{-1}B^*B\mu_i, \delta\right\rangle\right| \\ &\le \left\|L^{c-s}\left(\alpha I + B^*B\right)^{-1}B^*B\mu_i\right\| \|\delta\| \\ &\le C\|\mu_i\|\,\|\delta\|\end{aligned}$$

for some constant C. If, moreover, $E(\|\mu_i\|^{2+\varepsilon}) < \infty$, Lyapunov condition (3.27) is satisfied for all δ such that $\|\delta\| < \infty$.

Now, we consider a more general case.

Assumption 3.7. $L^{-s}\eta_i \in \mathcal{R}((B^*B)^{\nu/2}) = \mathcal{D}(L^{\nu(a+s)})$ for some $0 \le \nu \le 2$.

Assumption 3.8. $L^{c-s}\delta \in \mathcal{R}((B^*B)^{\tilde{\nu}/2}) = \mathcal{D}(L^{\tilde{\nu}(a+s)})$ for some $\tilde{\nu} \ge 0$.

By a slight abuse of notation, we introduce the following μ_i and ρ:

$$\begin{aligned}L^{-s}\eta_i &= \left(B^*B\right)^{\nu/2}\mu_i, \\ L^{c-s}\delta &= \left(B^*B\right)^{\tilde{\nu}/2}\rho.\end{aligned}$$

We have

$$\begin{aligned}|\langle M\eta_i, \delta\rangle| &= \left|\left\langle\left(\alpha I + B^*B\right)^{-1}L^{-s}\eta_i, L^{c-s}\delta\right\rangle\right| \\ &= \left|\left\langle\left(\alpha I + B^*B\right)^{-1}\left(B^*B\right)^{\nu/2}\mu_i, \left(B^*B\right)^{\tilde{\nu}/2}\rho\right\rangle\right| \\ &= \left|\left\langle\left(\alpha I + B^*B\right)^{-1}\left(B^*B\right)^{(\nu+\tilde{\nu})/2}\mu_i, \rho\right\rangle\right|.\end{aligned}$$

If $\nu + \tilde{\nu} \geq 2$, this term is bounded by $\|\mu_i\|\|\rho\|$. If, moreover, $E(\|\mu_i\|^{2+\varepsilon}) < \infty$, the condition (3.27) is satisfied and the asymptotic normality (3.28) holds for this specific δ.

3.6.2. Case Where the Operator Is Estimated

Let

$$\hat{\varphi} = L^{-s}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\hat{r},$$

$$\tilde{\varphi}_\alpha = L^{-s}\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\widehat{T}\varphi.$$

We want to study the asymptotic normality of

$$\left\langle L^c\left(\hat{\varphi} - \tilde{\varphi}_\alpha\right), \delta\right\rangle.$$

Assumption 3.9. $L^{c-s}\delta = (B^*B)^{d/2}\rho$ for some ρ with $\|\rho\| < \infty$.

Proposition 3.3. *Suppose that $\hat{\varphi}$ is as in* (3.31). *Assume that $\widehat{T}^*(\hat{r} - \widehat{T}\varphi) = \sum_{i=1}^n \eta_i/n$ for some η_i satisfying Assumption 3.6 and for some δ satisfying Assumption 3.9 and (3.27). If*

$$\frac{\sqrt{n}}{\alpha^{\frac{(3-d)\vee 1}{2}}}\frac{\left\|\hat{B} - B\right\|}{\left\|\Omega^{1/2}M^*\delta\right\|} \to 0 \tag{3.31}$$

and

$$\frac{\sqrt{n}}{\alpha^{\frac{(3-d)\vee 2}{2}}}\frac{\left\|\hat{B}^* - B^*\right\|}{\left\|\Omega^{1/2}M^*\delta\right\|} \to 0, \tag{3.32}$$

then

$$\frac{\sqrt{n}\left\langle L^c\left(\hat{\varphi} - \tilde{\varphi}_\alpha\right), \delta\right\rangle}{\left\|\Omega^{1/2}M^*\delta\right\|} \xrightarrow{d} \mathcal{N}(0,1).$$

The notation $a \vee b$ means $\max(a,b)$. In the IV example, $\|\hat{B} - B\|$ depends on a bandwidth h_n. By choosing h_n in an appropriate way, Conditions (3.31) and (3.32) will be satisfied.

Proof of Proposition 3.3. We have

$$L^c\left(\hat{\varphi} - \tilde{\varphi}_\alpha\right) = L^{c-s}\left(\alpha I + B^*B\right)^{-1}L^{-s}\widehat{T}^*\left(\hat{r} - \widehat{T}\varphi\right) \tag{3.33}$$

$$+ L^{c-s}\left\{\left(\alpha I + \hat{B}^*\hat{B}\right)^{-1} - \left(\alpha I + B^*B\right)^{-1}\right\}L^{-s}\widehat{T}^*\left(\hat{r} - \widehat{T}\varphi\right). \tag{3.34}$$

Using the fact that $\widehat{T}^*(\hat{r}-\widehat{T}\varphi)=\sum_{i=1}^n \eta_i/n$ for some η_i satisfying Assumption 3.6, we can establish the asymptotic normality of $\sqrt{n}\langle L^{c-s}(\alpha I+B^*B)^{-1}L^{-s}\widehat{T}^*(\hat{r}-\widehat{T}\varphi),\delta\rangle$ using the same proof as in Proposition 3.2.

Now we show that the term (3.38) is negligible. By Assumption 3.9, we have

$$\begin{aligned}
&\left|\left\langle L^{c-s}\left\{\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}-\left(\alpha I+B^*B\right)^{-1}\right\}L^{-s}\widehat{T}^*\left(\hat{r}-\widehat{T}\varphi\right),\delta\right\rangle\right|\\
&\quad=\left|\left\langle L^{-s}\frac{\sum_{i=1}^n\eta_i}{\sqrt{n}},\left\{\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}-\left(\alpha I+B^*B\right)^{-1}\right\}L^{c-s}\delta\right\rangle\right|\\
&\quad\le\left\|\frac{\sum_{i=1}^n L^{-s}\eta_i}{\sqrt{n}}\right\|\left\|\left\{\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}-\left(\alpha I+B^*B\right)^{-1}\right\}\left(B^*B\right)^{d/2}\rho\right\|
\end{aligned}$$

The first term on the r.h.s is $O(1)$. We focus on the second term:

$$\begin{aligned}
&\left\|\left\{\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}-\left(\alpha I+B^*B\right)^{-1}\right\}\left(B^*B\right)^{d/2}\rho\right\|\\
&\quad=\left\|\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}\left(B^*B-\hat{B}^*\hat{B}\right)\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\rho\right\|\\
&\quad=\left\|\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}\left(\hat{B}^*\left(B-\hat{B}\right)+\left(B^*-\hat{B}^*\right)B\right)\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\rho\right\|\\
&\quad\le\left\|\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}\hat{B}^*\left(B-\hat{B}\right)\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\rho\right\| \qquad \text{(term 1)}\\
&\qquad+\left\|\left(\alpha I+\hat{B}^*\hat{B}\right)^{-1}\left(B^*-\hat{B}^*\right)B\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\rho\right\| \qquad \text{(term 2)}
\end{aligned}$$

Term 1: We have $\|(\alpha I+\hat{B}^*\hat{B})^{-1}\hat{B}^*\|^2\le 1/\alpha$ and

$$\left\|\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\right\|^2\le C\frac{1}{\alpha^{(2-d)}}$$

for $d\le 2$ (see Carrasco et al. (2007)). If $d>2$, this term is bounded. Thus

$$\begin{aligned}
(\text{term }1)^2&\le\frac{1}{\alpha}\left\|\hat{B}-B\right\|^2\frac{1}{\alpha^{(2-d)\vee 0}}\\
&=\left\|\hat{B}-B\right\|^2\frac{1}{\alpha^{(3-d)\vee 1}}.
\end{aligned}$$

Term 2:

$$\begin{aligned}
(\text{term }2)^2&\le\frac{1}{\alpha^2}\left\|\hat{B}^*-B^*\right\|^2\left\|B\left(\alpha I+B^*B\right)^{-1}\left(B^*B\right)^{d/2}\right\|\\
&=\frac{1}{\alpha^2}\left\|\hat{B}^*-B^*\right\|^2\frac{1}{\alpha^{(1-d)\vee 0}}
\end{aligned}$$

$$= \left\| \hat{B}^* - B^* \right\|^2 \frac{1}{\alpha^{(3-d)\vee 2}}.$$

Under the assumptions of Proposition 3.3, $\sqrt{n}(3.34)/\|\Omega^{1/2}M^*\delta\|$ is negligible. ■

3.6.3. Root *n* Rate of Convergence

The rate of convergence of $\langle L^c\hat{\varphi} - L^c\varphi_\alpha, \delta\rangle$ is $\sqrt{n}$ if $\|\Omega^{1/2}M^*\delta\|$ is bounded. A $\sqrt{n}$ rate of convergence may sound strange in a nonparametric setting. However, it should be noted that taking the inner product has a smoothing property. Moreover, a $\sqrt{n}$ rate will in general be obtained only for functions δ that are sufficiently smooth.

We can illustrate this point in the context of IV estimation where we set $s = c = 0$ to facilitate the exposition. In this case, $\Omega = T^*T$. Assuming that δ satisfies Assumption 3.8, we have

$$\|\Omega^{1/2}M^*\delta\| = \left\| \left(T^*T\right)^{1/2} \left(T^*T + \alpha I\right)^{-1} \left(T^*T\right)^{\tilde{\nu}/2} \rho \right\|,$$

which is finite if $\tilde{\nu} > 1$. Here it is always possible to choose ρ and then δ so that the inner product $\langle\hat{\varphi} - \varphi_\alpha, \delta\rangle$ converges at a $\sqrt{n}$ rate.

The root n rate of convergence of inner products has been discussed in various papers (e.g., Carrasco et al. (2007, p. 57) and Ai and Chen (2007, 2012) where an efficiency bound is derived). Severini and Tripathi (2012) derive the efficiency bound for estimating inner products of φ which remains valid when φ is not identified.

3.7. Selection of the Regularization Parameter

Engl et al. (2000) propose to select α using the criterion

$$\min_\alpha \frac{1}{\sqrt{\alpha}} \|\hat{r} - T\widehat{\varphi}_\alpha\|$$

and show that the resulting α has the optimal rate of convergence when T is known.

Darolles et al. (2011) suggest a slightly different rule. Let $\widehat{\varphi}_{\alpha(2)}$ be the iterated Tikhonov estimator of order 2. Then α is chosen to minimize

$$\frac{1}{\alpha} \left\| \hat{T}^*\hat{r} - \hat{T}^*\hat{T}\widehat{\varphi}_{\alpha(2)} \right\|.$$

They show that this selection rule delivers an α with optimal speed of convergence for the model (3.6). See Fève and Florens (2010, 2011) for the practical implementation of this method.

Other adaptive selection rules have been proposed for the IV model (3.6) but using different estimators than Darolles et al. Loubes and Marteau (2009) consider a spectral cutoff estimator and give a selection criterion of α such that the mean square error of the resulting estimator of φ achieves the optimal bound up to a $\ln(n)^2$ factor. They assume that the eigenfunctions are known but the eigenvalues are estimated. Johannes and Schwarz (2011) consider an estimator combining spectral cutoff and thresholding. They show that their data-driven estimator can attain the lower risk bound up to a constant, provided that the eigenfunctions are known trigonometric functions.

Recently, Horowitz (2011) proposed a selection rule that does not require the knowledge of the eigenfunctions and/or eigenvalues. The estimator considered in Horowitz (2011) is a modification of Horowitz's (2012) estimator. Let us briefly explain how to construct such an estimator. Multiply the left-hand and right-hand sides of Eq. (3.7) by $f_W(w)$ to obtain

$$E(Y_i|W_i = w)f_W(w) = E(\varphi(Z_i)|W_i = w)f_W(w).$$

Now define $r(w) = E(Y_i|W_i = w)f_W(w)$ and $(T\varphi)(z) = \int \varphi(z)f_{Z,W}(z,w)dz$. Assume that the support of Z and W is $[0,1]$. Let $\{\psi_j : j = 1,2,\ldots\}$ be a given complete orthonormal basis for $L^2[0,1]$. Contrary to Darolles et al., the ψ_j are not related to the eigenfunctions of T^*T. Then, T and r are approximated by a series expansion on this basis:

$$\hat{r}(w) = \sum_{k=1}^{J_n} \hat{r}_k \psi_k(w),$$

$$\widehat{f}_{Z,W}(z,w) = \sum_{j=1}^{J_n}\sum_{k=1}^{J_n} \hat{c}_{jk}\psi_j(z)\psi_k(w),$$

where J_n is a nonstochastic truncation point and $\hat{r}_k$ and $\hat{c}_{jk}$ are estimated Fourier coefficients:

$$\hat{r}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_k(w_i),$$

$$\hat{c}_{jk} = \frac{1}{n}\sum_{i=1}^{n} \psi_j(z_i)\,\psi_k(w_i).$$

For any function $\nu: [0,1] \to \mathbb{R}$, define $D_j\nu(z) = d^j\nu(z)/dz^j$. Let

$$\mathcal{H}_{Js} = \left\{ \nu = \sum_{j=1}^{J} \nu_j\psi_j : \sum_{0\leq j\leq s} \int_0^1 \left[D_j\nu(z)\right]^2 dz \leq C_0 \right\}$$

for some finite $C_0 > 0$. Then Horowitz's (2011) sieve estimator is defined as

$$\tilde{\varphi} = \arg\min_{\nu \in \mathcal{H}_{Jns}} \left\| \hat{T}\nu - \hat{r} \right\|.$$

For $j = 1, 2, \ldots, J_n$, define $\tilde{b}_j = \langle \tilde{\varphi}, \psi_j \rangle$. Let $J \leq J_n$ be a positive integer, the modified estimator of φ considered in Horowitz (2012) is

$$\hat{\varphi}_J = \sum_{j=1}^{J} \tilde{b}_j \psi_j.$$

The optimal J, denoted J_{opt}, is defined as the value that minimizes the asymptotic mean square error (AIMSE) of $\hat{\varphi}_J$. The AIMSE is $E_A \|\hat{\varphi}_J - \varphi\|^2$, where $E_A(.)$ denotes the expectation of the leading term of the asymptotic expansion of (.). The selection rule is the following:

$$\hat{J} = \arg\min_{1 \leq J \leq J_n} \widehat{T}_n(J)$$

with

$$\widehat{T}_n(J) = \frac{2}{3} \frac{\ln(n)}{n^2} \sum_{i=1}^{n} \left\{ (Y_i - \tilde{\varphi}(W_i))^2 \sum_{j=1}^{J} \left(\left(\hat{T}^{-1} \right)^* \psi_j(W_i) \right)^2 \right\} - \left\| \hat{\varphi}_J \right\|^2.$$

For this $\hat{J}$,

$$E_A \left\| \hat{\varphi}_{\hat{J}} - \varphi \right\|^2 \leq \left(2 + \frac{4}{3} \ln(n) \right) E_A \left\| \hat{\varphi}_{J_{opt}} - \varphi \right\|^2.$$

Thus, strictly speaking, $\hat{J}$ is not optimal, but the rate of convergence in probability of $\|\hat{\varphi}_{\hat{J}} - \varphi\|^2$ is within a factor of $\ln(n)$ of the asymptotically optimal rate.

3.8. Implementation

We discuss the implementation in the four examples studied in Section 3.2.

3.8.1. Case Where T Is Known

When T is known, the implementation is relatively simple.

Example 3.1. Density (Continued). The Tikhonov estimator of the density is given by the solution of

$$\min_f \int_{-\infty}^{\infty} \left(\int_{-\infty}^{t} f(u)\, du - \hat{F}(t) \right)^2 dt + \alpha \int_{-\infty}^{\infty} f^{(s)}(u)^2\, du$$

where f possesses s derivatives. This problem has a closed-form solution (Vapnik, 1998, pages 309–311):

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} G_\alpha(x - x_i)$$

which is a kernel estimator with kernel

$$G_\alpha(x) = \int_{-\infty}^{\infty} \frac{e^{ix\omega}}{1 + \alpha\omega^{2(s+1)}}\, d\omega.$$

This formula simplifies when $s = 0$ (the desired density belongs to L^2):

$$G_\alpha(x) = \frac{1}{2\sqrt{\alpha}} \exp\left\{-\frac{|x|}{\sqrt{\alpha}}\right\}.$$

Example 3.2 Deconvolution (Continued). We describe the estimator of Carrasco and Florens (2011). Given that T and T^* are known, their spectral decompositions are also known (or can be approximated arbitrarily well by simulations). The solution f of $(\alpha I + T^*T)f = T^*h$ is given by

$$f(x) = \sum_{j=0}^{\infty} \frac{1}{\alpha + \lambda_j^2} \left\langle T^*h, \phi_j \right\rangle \phi_j(x).$$

The only unknown is $(T^*h)(x) = \int h(y)\pi_{Y|X}(y|x)dy = E[\pi_{Y|X}(Y|x)]$. It can be estimated by

$$\left(\widehat{T^*h}\right)(x) = \frac{1}{n}\sum_{i=1}^{n} \pi_{Y|X}\left(y_i|x\right)$$

so that the Tikhonov estimator of f is given by

$$\hat{f}(x) = \sum_{j=0}^{\infty} \frac{1}{\alpha + \lambda_j^2} \left\langle \frac{1}{n}\sum_{i=1}^{n} \pi_{Y|X}\left(y_i|.\right), \phi_j(.) \right\rangle \phi_j(x).$$

3.8.2. Case Where T Is Estimated

Given that the number of observations is n, the estimated operators $\hat{T}$ and $\hat{T}^*$ are necessarily finite-dimensional operators of dimensions that cannot exceed n. Assume that the operators $\hat{T}$ and $\hat{T}^*$ take the following forms:

$$\hat{T}\varphi = \sum_{i=1}^{n} a_i(\varphi) f_i,$$

$$\hat{T}^* \psi = \sum_{i=1}^{n} b_i(\psi) e_i,$$

where f_i and e_i are elements of $\mathcal{F}$ and $\mathcal{E}$, respectively, and a_i and b_i are linear functions. Assume that r takes the form

$$\hat{r} = \sum_{i=1}^{n} c_i f_i.$$

Then, $\hat{T}^* \hat{T} \varphi + \alpha \varphi = \hat{T}^* \hat{r}$ can be rewritten as

$$\sum_{i=1}^{n} b_i \left(\sum_{j=1}^{n} a_j(\varphi) f_j \right) e_i + \alpha \varphi = \sum_{i=1}^{n} b_i \left(\sum_{j=1}^{n} c_j f_j \right) e_i \tag{3.35}$$

$$\sum_{i,j=1}^{n} b_i \left(f_j \right) a_j \left(\varphi \right) e_i + \alpha \varphi = \sum_{i,j=1}^{n} b_i \left(f_j \right) c_j e_i. \tag{3.36}$$

Now, applying a_l on the r.h.s. and l.h.s of (3.36) and using the linearity of the function a_l yields

$$\sum_{i,j=1}^{n} b_i \left(f_j \right) a_j(\varphi) a_l(e_i) + \alpha a_l(\varphi) = \sum_{i,j=1}^{n} b_i \left(f_j \right) c_j a_l(e_i). \tag{3.37}$$

We obtain n equations with n unknowns $a_j(\varphi), j = 1, 2, \ldots, n$. We can solve this system and then replace $a_j(\varphi)$ by its expression in (3.35) to obtain φ. We illustrate this method in two examples.

Example 3.3. Functional Linear Regression (Continued). To simplify, let $\mathcal{E} = \mathcal{F} = L^2[0,1]$. We have

$$\hat{T} \varphi = \frac{1}{n} \sum_{i=1}^{n} \langle Z_i, \varphi \rangle W_i,$$

$$\hat{T}^* \psi = \frac{1}{n} \sum_{i=1}^{n} \langle W_i, \psi \rangle Z_i,$$

$$\hat{r} = \frac{1}{n} \sum_{i=1}^{n} Y_i W_i.$$

Then $f_i = W_i/n$, $e_i = Z_i/n$, $a_i(\varphi) = \langle Z_i, \varphi \rangle$, $b_i(\psi) = \langle W_i, \psi \rangle$, $c_i = Y_i$. Equation (3.37) gives

$$\alpha \langle \varphi, Z_l \rangle + \frac{1}{n^2} \sum_{i,j=1}^{n} \langle Z_i, Z_l \rangle \left\langle W_i, W_j \right\rangle \left\langle \varphi, Z_j \right\rangle = \frac{1}{n^2} \sum_{i,j=1}^{n} \langle Z_i, Z_l \rangle \left\langle W_i, W_j \right\rangle Y_j, \qquad l = 1, \ldots, n.$$

To compute the inner products $\langle Z_i, Z_l \rangle$, Florens and Van Bellegem (2012) propose to discretize the integrals as follows:

$$\langle Z_i, Z_l \rangle = \frac{1}{T} \sum_{t=1}^{T} Z_i\left(\frac{t}{T}\right) Z_l\left(\frac{t}{T}\right)$$

and the same for $\langle W_i, W_j \rangle$. Let Z and W denote the $T \times n$ matrices with (t,i) elements $Z_i(t/T)$ and $W_i(t/T)$, respectively. Let ξ and Y be the $n \times 1$ vectors of $\langle \varphi, Z_i \rangle$ and Y_i. Then, closed-form expressions for ξ and φ are given by

$$\xi = \left(\alpha I + \frac{1}{n^2} \frac{Z'Z}{T} \frac{W'W}{T}\right)^{-1} \left(\frac{1}{n^2} \frac{Z'Z}{T} \frac{W'W}{T} Y\right),$$

$$\hat{\varphi} = \frac{1}{\alpha n^2} Z \frac{W'W}{T} (Y - \xi).$$

Example 3.4. Nonparametric Instrumental Regression (Continued). In Darolles et al. (2002), the conditional expectation operator is estimated by a kernel estimator with kernel k and bandwith h_n.

$$\widehat{T}\varphi = \frac{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right) \varphi(z_i)}{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)},$$

$$\hat{T}^* \psi = \frac{\sum_{i=1}^{n} k\left(\frac{z-z_i}{h_n}\right) \psi(w_i)}{\sum_{i=1}^{n} k\left(\frac{z-z_i}{h_n}\right)},$$

$$\hat{r} = \frac{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{w-w_i}{h_n}\right)}.$$

Thus $f_i = \frac{k(\frac{w-w_i}{h_n})}{\sum_{i=1}^{n} k(\frac{w-w_i}{h_n})}$, $e_i = \frac{k(\frac{z-z_i}{h_n})}{\sum_{i=1}^{n} k(\frac{z-z_i}{h_n})}$, $a_i(\varphi) = \varphi(z_i)$, $b_i(\psi) = \psi(w_i)$, $c_i = y_i$.

Note that in Darolles et al. (2011), Z and W are assumed to have bounded supports $[0,1]^p$ and $[0,1]^q$ and a generalized kernel is used to avoid having a larger bias at the boundaries of the support.

Now we illustrate the role of L^{-s}. Consider $\mathcal{F}$ the space of square integrable functions defined on $[0,1]$ that satisfy the conditions $\phi(0) = 0$ and $\phi'(1) = 0$. The inner product on this space is defined by $\langle \phi, \psi \rangle = \int_0^1 \phi(x)\psi(x)dx$. Let $L\phi = -\phi''$, which satisfies all the properties of Hilbert scale (L is self-adjoint, etc). The estimator is given by

$$\hat{\varphi} = L^{-1}\left(\alpha I + L^{-1} T^* T L^{-1}\right)^{-1} L^{-1} K^* \hat{r}.$$

This approach is particularly useful if one is interested in the second derivative of φ since we have

$$\widehat{\varphi''} = \left(\alpha I + L^{-1} T^* T L^{-1}\right)^{-1} L^{-1} K^* \hat{r}.$$

Note that even if φ does not satisfy the boundary conditions $\varphi(0)=0$ and $\varphi'(1)=0$, $\hat{\varphi}$ satisfies these properties. It has no impact on the second derivatives. Moreover, we know that

$$L^{-1}\varphi = \int_0^1 (s \wedge t)\varphi(s)\,ds.$$

Hence L^{-1} can be approximated by a numerical integral:

$$L^{-1}\varphi = \frac{1}{N}\sum_{i=1}^{N}(s_i \wedge t)\varphi(s_i).$$

Florens and Racine (2012) propose an estimation procedure of the first partial derivative of φ by Landweber–Fridman. Their paper derives the rate of convergence of the estimator, investigates the small-sample performance via Monte Carlo, and applies the method to the estimation of the Engel curve as in Blundell et al. (2007).

3.9. Concluding Remarks

In this chapter, we mainly focused on the asymptotic normality of $\sqrt{n}\langle\hat{\varphi}-\varphi_\alpha,\delta\rangle$ and omitted to study the regularization bias. However, the bias has the form

$$b_n = \varphi_\alpha - \varphi = -\alpha\sum_j \frac{1}{\lambda_j^2+\alpha}\left\langle \varphi,\varphi_j\right\rangle \varphi_j,$$

which is estimable. Given a consistent α, denoted $\tilde{\alpha}$, we can construct a first-step estimator of φ denoted $\hat{\varphi}_{\tilde{\alpha}}$. Then an estimator of the bias is given by

$$\hat{b}_n = -\alpha\sum_j \frac{1}{\hat{\lambda}_j^2+\alpha}\left\langle \hat{\varphi}_{\tilde{\alpha}},\hat{\varphi}_j\right\rangle \hat{\varphi}_j,$$

where $\hat{\varphi}_j$ and $\hat{\lambda}_j$ are consistent estimators of φ_j and λ_j as described in Carrasco et al. (2007). Given this estimator, we can construct a bias-corrected estimator of $\hat{\varphi}$, $\hat{\varphi}-\hat{b}_n$. Although this estimator will have a smaller bias than the original one, it may have a larger variance.

References

Ai, Chunrong, and Xiaohong Chen. 2007. "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models with Different Conditioning Variables." *Journal of Econometrics*, **141**, pp. 5–43.

Ai, Chunrong, and Xiahong Chen. 2012. "Semiparametric Efficiency Bound for Models of Sequential Moment Restrictions Containing Unknown Functions." *Journal of Econometrics*, **170**, pp. 442–457.

Altissimo, Filippo, and Antonio Mele. 2009. "Simulated Nonparametric Estimation of Dynamic Models." *Review of Economic Studies*, **76**, pp. 413–450.

Billingsley, Patrick. 1995. *Probability and Measure*. New York: John Wiley & Sons.

Blundell, Richard, Xiaohong Chen, and Dennis Kristensen. 2007. "Semi-nonparametric IV Estimation of Shape-Invariant Engle Curves." *Econometrica*, **75**, pp. 1613–1669.

Blundell, Richard, and Joel Horowitz. 2007. "A Nonparametric Test of Exogeneity." *Review of Economic Studies*, **74**, pp. 1035–1058.

Canay, Ivan, Andrés Santos, and Azeem Shaikh. 2013. "On the Testability of Identification in Some Nonparametric Models with Endogeneity." Forthcoming in Econometrica, Northwestern.

Cardot, Hervé, and Pascal Sarda. 2006. "Linear Regression Models for functional Data." In *The Art of Semiparametrics*, ed. W. Härdle. Heidelberg: Physica-Verlag, Springer, pp. 49–66.

Carrasco, Marine, and Jean-Pierre Florens. 2011. "Spectral Method for Deconvolving a Density." *Econometric Theory*, **27**(3), pp. 546–581.

Carrasco, Marine, Jean-Pierre Florens, and Eric Renault. 2007. "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization." In *Handbook of Econometrics*, Vol. 6B, ed. James Heckman and Edward Leamer. Amsterdam: North Holland, pp. 5633–5746.

Chen, Xiaohong, and Markus Reiss. 2011. "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics." *Econometric Theory*, **27**, pp. 497–521.

Chen, Xiaohong, and Halbert White. 1998. "Central Limit and Functional Central Limit Theorems for Hilbert Space-Valued Dependent Heterogeneous Arrays with Applications." *Econometric Theory*, **14**, pp. 260–284.

Darolles, Serge, Jean-Pierre Florens, and Eric Renault. 2002. "Nonparametric Instrumental Regression." CRDE working paper, cahier 05–2002.

Darolles, Serge, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. 2011. "Nonparametric Instrumental Regression." *Econometrica*, **79**, pp. 1541–1565.

Engl, Heinz W., Martin Hanke, and Andreas Neubauer. 2000. *Regularization of Inverse Problems*. Dordrecht: Kluwer Academic.

Fan, Yanqin. 1994. "Testing the Goodness-of-Fit of a Parameter Density Function by Kernel Method." *Econometric Theory*, **10**, pp. 316–356.

Ferraty, Frédéric, and Philippe Vieu. 2000. "Dimension fractale et estimation de la regression dans des espaces vectoriels semi-normes." *Comptes Rendus de l'Académie des Sciences de Paris, Série I, Mathématique*, **330**, pp. 139–142.

Fève, Frédérique, and Jean-Pierre Florens. 2010. "The Practice of Non Parametric Estimation by Solving Inverse Problems: The Example of Transformation Models." *The Econometrics Journal*, **13**, pp. S1–S27.

Fève, Frédérique, and Jean-Pierre Florens. 2011. "Non Parametric Analysis of Panel Data Models with Endogenous Variables." Mimeo, Toulouse School of Economics.

Florens, Jean-Pierre. 2003. "Inverse Problems and Structural Econometrics: The Example of Instrumental Variables." In *Advances in Economics and Econometrics: Theory and*

Applications, Volume 2, eds. Mathias Dewatripont, Lars Peter Hansen, and Stephen Turnovsky, New York: Cambridge University Press, pp. 284–311.

Florens, Jean-Pierre, and Sébastien Van Bellegem. 2012. "Functional Linear Instrumental Regression." Mimeo.

Florens, Jean-Pierre, Jan Johannes, and Sébastien Van Bellegem. 2011. "Identification and Estimation by Penalization in Nonparametric Instrumental Regression." *Econometric Theory*, **27**, pp. 472–496.

Florens, Jean-Pierre, and Jeffrey Racine. 2012. "Nonparametric Instrumental Derivatives." Mimeo, McMaster University.

Gagliardini, Patrick, and Olivier Scaillet. 2012. "Tikhonov Regularization for Nonparametric Instrumental Variable Estimators." *Journal of Econometrics*, **167**, pp. 61–75.

Groetsch, Charles. 1993. *Inverse Problems in the Mathematical Sciences*. Braunschweig, Wiesbaden: Vieweg & Sohn.

Hall, Peter, and Joel Horowitz. 2005. "Nonparametric Methods for Inference in the Presence of Instrumental Variables." *Annals of Statistics*, **33**, pp. 2904–2929.

Hall, Peter, and Joel Horowitz. 2007. "Methodology and Convergence Rates for Functional Linear Regression." *Annals of Statistics*, **35**, pp. 70–91.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, **50**, pp. 1029–1054.

Härdle, Wolfgang, and Enno Mammen. 1993. "Comparing Nonparametric Versus Parametric Regression Fits." *Annals of Statistics*, **21**, pp. 1926–1947.

Horowitz, Joel. 2006. "Testing a Parametric Model Against a Nonparametric Alternative with Identification Through Instrumental Variables." *Econometrica*, **74**, pp. 521–538.

Horowitz, Joel. 2007. "Asymptotic Normality of a Nonparametric Instrumental Variables Estimator." *International Economic Review*, **48**, pp. 1329–1349.

Horowitz, Joel. 2011. "Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter." Mimeo, Northwestern University.

Horowitz, Joel. 2012. "Specification Testing in Nonparametric Instrumental Variable Estimation." *Journal of Econometrics*, **167**, pp. 383–396.

Johannes, Jan, and Maik Schwarz. 2011. "Partially Adaptive Nonparametric Instrumental Regression by Model Selection." arxiv:1003.3128v1, Journal of the Indian Statistical Association, Vol 49, pp. 149–175.

Johannes, Jan, Sébastien Van Bellegem, and Anne Vanhems. 2011. "Convergence Rates for Ill-Posed Inverse Problems with an Unknown Operator." *Econometric Theory*, **27**, 522–545.

Kress, Rainer. 1999. *Linear Integral Equations*, Applied Mathematical Sciences, New York: Springer-Verlag.

Krein, S. G., and Yu I. Petunin. 1966. "Scales of Banach Spaces." *Russian Mathematical Surveys*, **21**, pp. 85–160.

Loubes, Jean-Michel, and Clément Marteau. 2009. "Oracle Inequality for Instrumental Variable Regression." arxiv:0901.4321v1, University of Toulouse 3, France.

Newey, Whitney, and James Powell. 2003. "Instrumental Variable Estimation of Nonparametric Models." *Econometrica*, **71**, pp. 1565–1578.

Ramsay, James O., and Bernard W. Silverman. 1997. *Functional Data Analysis*. New York: Springer.

Severini Thomas, and Gautam Tripathi. 2012. “Efficiency Bounds for Estimating Linear Functionals of Nonparametric Regression Models with Endogenous Regressors.” *Journal of Econometrics*, **170**, pp. 491–498.

van der Vaart, Aad, and Jon Wellner. 1996. *Weak Convergence and Empirical Processes*. New York: Springer.

Vapnik, Vladimir. 1998. *Statistical Learning Theory*, New York: John Wiley & Sons.

CHAPTER 4

IDENTIFICATION AND WELL-POSEDNESS IN NONPARAMETRIC MODELS WITH INDEPENDENCE CONDITIONS

VICTORIA ZINDE-WALSH[†]

4.1. Introduction

Many statistical and econometric models involve independence (or conditional independence) conditions that can be expressed via convolution. Examples are independent errors, classical measurement error and Berkson error, regressions involving data measured with these types of errors, common factor models, and models that conditionally on some variables can be represented in similar forms, such as a nonparametric panel data model with errors conditionally on observables independent of the idiosyncratic component.

Although the convolution operator is well known, this chapter explicitly provides convolution equations for a wide list of models for the first time. In many cases the analysis in the literature takes Fourier transforms as the starting point—for example, characteristic functions for distributions of random vectors, as in the famous Kotlyarski lemma (Kotlyarski, 1967). The emphasis here on convolution equations for the models provides the opportunity to explicitly state nonparametric classes of functions defined by the model for which such equations hold, in particular, for densities, conditional densities, and regression functions. The statistical model may give rise to different systems of convolution equations and may be overidentified in terms of convolution equations; some choices may be better suited to different situations; for example, here in Section 4.2 two sets of convolution equations (4 and 4a in Table 4.1) are provided for the same classical measurement error model with two measurements; it turns out that one of those allows us to relax some independence conditions, while the other makes it possible to relax a support assumption in identification. Many of the convolution equations derived here are based on density-weighted conditional averages of the observables.

The main distinguishing feature is that here all the functions defined by the model are considered within the space of generalized functions S^*, the space of so-called tempered distributions (they will be referred to as generalized functions). This is the dual space, the space of linear continuous functionals, on the space S of well-behaved functions: The functions in S are infinitely differentiable, and all the derivatives go to zero at infinity faster than any power. An important advantage of assuming that the functions are in the space of generalized functions is that in that space any distribution function has a density (generalized function) that continuously depends on the distribution function, so that distributions with mass points and fractal measures have well-defined generalized densities.

Any regular function majorized by a polynomial belongs to S^*; this includes polynomially growing regression functions and binary choice regression as well as many conditional density functions. Another advantage is that Fourier transform is an isomorphism of this space, and thus the usual approaches in the literature that employ characteristic functions are also included. Details about the space S^* are in Schwartz (1966) and are summarized in Zinde-Walsh (2013).

The model classes examined here lead to convolution equations that are similar to each other in form; the main focus of this chapter is on existence, identification, partial identification, and well-posedness conditions. Existence and uniqueness of solutions to some systems of convolution equations in the space S^* were established in Zinde-Walsh (2013). Those results are used here to state identification in each of the models. Identification requires examining support of the functions and generalized functions that enter into the models; if support excludes an open set, then identification at least for some unknown functions in the model fails; however, some isolated points or lower-dimensional manifolds where, for example, the characteristic function takes zero values (an example is the uniform distribution) does not preclude identification in some of the models. This point was made, for example, in Carrasco and Florens (2010) and in Evdokimov and White (2012) and is expressed here in the context of operating in S^*. Support restriction for the solution may imply that only partial identification will be provided. However, even in partially identified models, some features of interest (see, e.g., Matzkin (2007)) could be identified; thus some questions could be addressed, even in the absence of full identification. A common example of incomplete identification which nevertheless provides important information is Gaussian deconvolution of a blurred image of a car obtained from a traffic camera; the filtered image is still not very good, but the license plate number is visible for forensics.

Well-posedness conditions are emphasized here. The well-known definition by Hadamard (1923) defines well-posedness via three conditions: existence of a solution, uniqueness of the solution, and continuity in some suitable topology. The first two are essentially identification. Since here we shall be defining the functions in subclasses of S^*, we shall consider continuity in the topology of this generalized functions space. This topology is weaker than the topologies in functions spaces, such as the uniform or L_p topologies; thus differentiating the distribution function to obtain a density is a well-posed problem in S^*, by contrast, even in the class of absolutely continuous distributions with uniform metric where identification for density in the space L_1 holds; well-posedness, however, does not obtain (see discussion in Zinde-Walsh (2011)). But even though in the weaker topology of S^* well-posedness obtains more widely, for the problems considered here some additional restrictions may be required for well-posedness.

Well-posedness is important for plug-in estimation since if the estimators are in a class where the problem is well-posed, they are consistent; and conversely, if well-posedness does not hold, consistency will fail for some cases. Lack of well-posedness can be remedied by regularization, but the price is often more extensive requirements on the model and slower convergence. For example, in deconvolution (see, e.g., Fan (1991) and most other papers cited here) spectral cutoff regularization is utilized; it crucially depends on knowing the rate of the decay at infinity of the density.

Often nonparametric identification is used to justify parametric or semi-parametric estimation; the claim here is that well-posedness should be an important part of this justification. The reason for that is that in estimating a possibly misspecified parametric model, the misspecified functions of the observables belong in a nonparametric neighborhood of the true functions; if the model is nonparametrically identified, the unique solution to the true model exists, but without well-posedness the solution to the parametric model and to the true one may be far apart.

For deconvolution, An and Hu (2012) demonstrate well-posedness in spaces of integrable density functions when the measurement error has a mass point; this may happen in surveys when probability of truthful reporting is non-zero. The conditions for well-posedness here are provided in S^*; this then additionally does not exclude mass points in the distribution of the mismeasured variable itself; there is some empirical evidence of mass points in earnings and income. The results here show that in S^*, well-posedness holds more generally—as long as the error distribution is not supersmooth.

The solutions for the systems of convolution equations can be used in plug-in estimation. Properties of nonparametric plug-in estimators are based on results on stochastic convergence in S^* for the solutions that are stochastic functions expressed via the estimators of the known functions of the observables.

Section 4.2 enumerates the classes of models considered here. They are divided into three groups: (1) measurement error models with classical and Berkson errors and possibly an additional measurement, along with common factor models that transform into those models; (2) nonparametric regression models with classical measurement and Berkson errors in variables; (3) measurement error and regression models with conditional independence. The corresponding convolution equations and systems of equations are provided and discussed. Section 4.3 is devoted to describing the solutions to the convolution equations of the models. The main mathematical aspect of the different models is that they require solving equations of a similar form. Section

4.4 provides a table of identified solutions and discusses partial identification and well-posedness. Section 4.5 examines plug-in estimation. A brief conclusion follows.

4.2. Convolution Equations in Classes of Models with Independence or Conditional Independence

This section derives systems of convolution equations for some important classes of models. The first class of model is measurement error models with some independence (classical or Berkson error) and possibly a second measurement; the second class is regression models with classical or Berkson-type error; the third is models with conditional independence. For the first two classes the distributional assumptions for each model and the corresponding convolution equations are summarized in tables; we indicate which of the functions are known and which are unknown; a brief discussion of each model and derivation of the convolution equations follows. The last part of this section discusses convolution equations for two specific models with conditional independence; one is a panel data model studied by Evdokimov (2011), the other a regression model where independence of measurement error of some regressors obtains conditionally on a covariate.

The general assumption made here is that all the functions in the convolution equations belong to the space of generalized functions S^*.

Assumption 4.1. *All the functions defined by the statistical model are in the space of generalized functions S^*.*

This space of generalized function includes functions from most of the function classes that are usually considered, but allows for some useful generalizations. The next subsection provides the necessary definitions and some of the implications of working in the space S^*.

4.2.1. The Space of Generalized Functions S^*

The space S^* is the dual space, that is, the space of continuous linear functionals on the space S of functions. The theory of generalized functions is in Schwartz (1966); relevant details are summarized in Zinde-Walsh (2013). In this subsection the main definitions and properties are reproduced.

Recall the definition of S.

For any vector of non-negative integers $m = (m_1, \ldots, m_d)$ and vector $t \in R^d$ denote by t^m the product $t_1^{m_1} \ldots t_d^{m_d}$ and by ∂^m the differentiation operator $\frac{\partial^{m_1}}{\partial x_1^{m_1}} \ldots \frac{\partial^{m_d}}{\partial x_d^{m_d}}$; C_∞

is the space of infinitely differentiable (real or complex-valued) functions on R^d. The space $S \subset C_\infty$ of test functions is defined as

$$S = \left\{ \psi \in C_\infty(R^d) : |t^l \partial^k \psi(t)| = o(1) \text{ as } t \to \infty \right\},$$

for any $k = (k_1, \dots, k_d), l = (l_1, \dots, l_d)$, where $k = (0, \dots, 0)$ corresponds to the function itself, $t \to \infty$ coordinatewise; thus the functions in S go to zero at infinity faster than any power as do their derivatives; they are rapidly decreasing functions. A sequence in S converges if in every bounded region each $\left|t^l \partial^k \psi(t)\right|$ converges uniformly.

Then in the dual space S^* any $b \in S^*$ represents a linear functional on S; the value of this functional for $\psi \in S$ is denoted by (b, ψ). When b is an ordinary (pointwise defined) real-valued function, such as a density of an absolutely continuous distribution or a regression function, the value of the functional on real-valued ψ defines it and is given by

$$(b, \psi) = \int b(x) \psi(x)\, dx.$$

If b is a characteristic function that may be complex-valued, then the value of the functional b applied to $\psi \in S$, where S is the space of complex-valued functions, is

$$(b, \psi) = \int b(x) \overline{\psi(x)}\, dx,$$

where the overbar denotes complex conjugate. The integrals are taken over the whole space R^d.

The generalized functions in the space S^* are continuously differentiable and the differentiation operator is continuous; Fourier transforms and their inverses are defined for all $b \in S^*$, the operator is a (continuous) isomorphism of the space S^*. However, convolutions and products are not defined for all pairs of elements of S^*, unlike, say, the space L_1; on the other hand, in L_1, differentiation is not defined and not every distribution has a density that is an element of L_1.

Assumption 4.1 places no restrictions on the distributions, since in S^* any distribution function is differentiable and the differentiation operator is continuous. The advantage of not restricting distributions to be absolutely continuous is that mass points need not be excluded; distributions representing fractal measures such as the Cantor distribution are also allowed. This means that mixtures of discrete and continuous distributions are included—such as those examined by An and Hu (2012) for measurement error in survey responses where some are error—contaminated, some truthful leading to a mixture with a mass point distribution. Moreover, in S^* the case of mass points in the distribution of the mismeasured variable is also easily handled; in the literature such mass points are documented for income or work hours distributions in the presence of rigidities such as unemployment compensation rules (e.g., Green and Riddell, 1997). Fractal distributions may arise in some situations—for example, Karlin's (1959) example of the equilibrium price distribution in an oligopolistic game.

For regression functions the assumption $g \in S^*$ implies that growth at infinity is allowed but is somewhat restricted. In particular, for any ordinary pointwise defined function $b \in S^*$ the condition

$$\int \dots \int \Pi_{i=1}^{d} \left((1+t_i^2)^{-1}\right)^{m_i} |b(t)|\, dt_1 \dots dt_d < \infty, \tag{4.1}$$

needs to be satisfied for some non-negative valued $m_1, \dots, m_d$. If a locally integrable function g is such that its growth at infinity is majorized by a polynomial, then $b \equiv g$ satisfies this condition. While restrictive, this still widens the applicability of many currently available approaches. For example, in Berkson regression the common assumption is that the regression function be absolutely integrable (Meister, 2009); this excludes binary choice, linear and polynomial regression functions that belong to S^* and satisfy Assumption 4.1. Also, it is advantageous to allow for functions that may not belong to any ordinary function classes, such as sums of δ-functions ("sum of peaks") or (mixture) cases with sparse parts of support, such as isolated points; such functions are in S^*. Distributions with mass points can arise when the response to a survey question may be only partially contaminated; regression "sum of peaks" functions arise, for example, in spectroscopy and astrophysics, where isolated point supports are common.

4.2.2. Measurement Error and Related Models

Current reviews for measurement error models are in Carrol et al. (2006), Chen et al. (2011), and Meister (2009).

Here and everywhere below, the variables x, z, x^*, u, u_x are assumed to be in R^d; y, v are in R^1; all the integrals are over the corresponding space; density of v for any v is denoted by f_v; independence is denoted by $\perp$; expectation of x conditional on z is denoted by $E(x|z)$.

4.2.2.1. List of Models and Corresponding Equations

Table 4.1 lists various models and corresponding convolution equations. Many of the equations are derived from density-weighted conditional expectations of the observables.

Recall that for two functions, f and g, convolution $f * g$ is defined by

$$(f * g)(x) = \int f(w) g(x-w)\, dw;$$

this expression is not always defined. A similar expression (with some abuse of notation since generalized functions are not defined pointwise) may hold for generalized functions in S^*; similarly, it is not always defined. With Assumption 4.1 for the models considered here, we show that convolution equations given in Table 4.1 hold in S^*.

Table 4.1 Measurement Error Models: 1. Classical Measurement Error; 2. Berkson Measurement Error; 3. Classical Measurement Error with Additional Observation (with Zero Conditional Mean Error); 4, 4a. Classical Error with Additional Observation (Full Independence)

Model	Distributional Assumptions	Convolution Equations	Known Functions	Unknown Functions
1	$z = x^* + u$, $x^* \perp u$	$f_{x^*} * f_u = f_z$	f_z, f_u	f_{x^*}
2	$z = x^* + u$, $z \perp u$	$f_z * f_{-u} = f_{x^*}$	f_z, f_u	f_{x^*}
3	$z = x^* + u$, $x = x^* + u_x$, $x^* \perp u$, $E(u_x \| x^*, u) = 0$, $E\\|z\\| < \infty; E\\|u\\| < \infty$	$f_{x^*} * f_u = f_z$, $h_k * f_u = w_k$, with $h_k(x) \equiv x_k f_{x^*}(x)$, $k = 1, 2, \ldots, d$	f_z, w_k, $k = 1, 2, \ldots, d$	$f_{x^*}; f_u$
4	$z = x^* + u$, $x = x^* + u_x; x^* \perp u$, $x^* \perp u_x; E(u_x) = 0$, $u \perp u_x$, $E\\|z\\| < \infty; E\\|u\\| < \infty$	$f_{x^*} * f_u = f_z$, $h_k * f_u = w_k$, $f_{x^*} * f_{u_x} = f_x$, with $h_k(x) \equiv x_k f_{x^*}(x)$, $k = 1, 2, \ldots, d$	f_z, $f_x; w; w_k$ $k = 1, 2, \ldots, d$	$f_{x^*}; f_u$, f_{u_x}
4a	Same model as 4, alternative equations	$f_{x^*} * f_u = f_z$, $f_{u_x} * f_{-u} = w$, $h_k * f_{-u} = w_k$, with $h_k(x) \equiv x_k f_{u_x}(x)$, $k = 1, 2, \ldots, d$	Same as for 4	Same as for 4

Notation: $k = 1, 2, \ldots, d$; in 3 and 4, $w_k = E(x_k f_z(z) | z)$; in 4a, $w = f_{z-x}$; $w_k = E(x_k w(z - x) | \, (z - x))$.

Theorem 4.1. *Under Assumption 4.1 for each of the models 1–4 the corresponding convolution equations of Table 4.1 hold in the generalized functions space S^*.*

The proof is in the derivations of the following subsection.

Assumption 4.1 requires considering all the functions defined by the model as elements of the space S^*; but if the functions (e.g., densities, the conditional moments) exist as regular functions, the convolutions are just the usual convolutions of functions, on the other hand, the assumption allows us to consider convolutions for cases where distributions are not absolutely continuous.

4.2.2.2. Measurement Error Models and Derivation of the Corresponding Equations

Model 1. The Classical Measurement Error Model. The case of the classical measurement error is well known in the literature. Independence between error and the variable of interest is applicable to problems in many fields as long as it may be assumed that the source of the error is unrelated to the signal. For example, in Cunha et al. (2010) it is assumed that some constructed measurement of ability of a child derived from test scores fits into this framework. As is well known in regression, a measurement error in the regressor can result in a biased estimator (attenuation bias).

Typically the convolution equation

$$f_{x^*} * f_u = f_z$$

is written for density functions when the distribution function is absolutely continuous. The usual approach to possible nonexistence of density avoids considering the convolution and focuses on the characteristic functions. Since density always exists as a generalized function and convolution for such generalized functions is always defined, it is possible to write convolution equations in S^* for any distributions in model 1. The error distribution (and thus generalized density f_u) is assumed known; thus the solution can be obtained by "deconvolution" (Carrol et al. (2006), Meister (2009), the review of Chen et al. (2011), and papers by Fan (1991) and by Carrasco and Florens (2010), among others).

Model 2. The Berkson Error Model. For Berkson error the convolution equation is also well known. Berkson error of measurement arises when the measurement is somehow controlled and the error is caused by independent factors; for example, amount of fertilizer applied is given but the absorption into soil is partially determined by factors independent of that, or students' grade distribution in a course is given in advance, or distribution of categories for evaluation of grant proposals is determined by the granting agency. The properties of Berkson error are very different from that of classical error of measurement; for example, it does not lead to attenuation bias in regression; also in the convolution equation the unknown function is directly expressed via the known ones when the distribution of Berkson error is known. For discussion see Carrol et al. (2006), Meister (2009), and Wang (2004).

Models 3 and 4. The Classical Measurement Error with Another Observation. In 3 and 4 in the classical measurement error model the error distribution is not known, but another observation for the mismeasured variable is available; this case has been treated in the literature and is reviewed in Carrol et al. (2006) and Chen et al. (2011). In econometrics, such models were examined by Li and Vuong (1998), Li (2002), Schennach (2004), and subsequently others (see, for example, the review by Chen et al. (2011)). In case 3 the additional observation contains an error that is not necessarily independent, but just has conditional mean zero.

Note that here the multivariate case is treated where arbitrary dependence for the components of vectors is allowed. For example, it may be of interest to consider the vector of not necessarily independent latent abilities or skills as measured by different sections of an IQ test, or the GRE scores.

Extra measurements provide additional equations. Consider for any $k = 1, \dots, d$ the function of observables w_k defined by density-weighted expectation $E(x_k f_z(z)|z)$ as a generalized function; it is then determined by the values of the functional (w_k, ψ) for every $\psi \in S$. Note that by assumption $E(x_k f_z(z)|z) = E(x_k^* f_z(z)|z)$; then for any $\psi \in S$ the value of the functional is given by

$$\begin{aligned}(E(x_k^* f_z(z)|z), \psi) &= \int [\int x_k^* f_{x^*,z}(x^*, z)\, dx^*] \psi(z)\, dz \\ &= \int\int x_k^* f_{x^*,z}(x^*, z)\psi(z)\, dx^*\, dz \\ &= \int\int x_k^* \psi(x^* + u) f_{x^*,u}(x^*, u)\, dx^*\, du \\ &= \int\int x_k^* f_{x^*}(x^*) f_u(u)\psi(x^* + u)\, dx^*\, du = (h_k * f_u, \psi).\end{aligned}$$

The third expression is a double integral that always exists if $E\|x^*\| < \infty$; this is a consequence of boundedness of the expectations of z and u. The fourth is a result of change of variables (x^*, z) into (x^*, u), the fifth uses independence of x^* and u, and the sixth expression follows from the corresponding expression for the convolution of generalized functions (Schwartz, 1966, p. 246). The conditions of model 3 are not sufficient to identify the distribution of u_x; this is treated as a nuisance part in model 3.

Model 4 with all the errors and mismeasured variables independent of each other was investigated by Kotlyarski (1967), who worked with the joint characteristic function. In model 4 consider in addition to the equations written for model 3 another that uses the independence between x^* and u_x and involves f_{u_x}.

In representation 4a the convolution equations involving the density f_{u_x} are obtained by applying the derivations that were used here for model 3,

$$\begin{aligned}z &= x^* + u, \\ x &= x^* + u_x,\end{aligned}$$

to model 4 with $x-z$ playing the role of z, u_x playing the role of x^*, $-u$ playing the role of u, and x^* playing the role of u_x. The additional convolution equations arising from the extra independence conditions provide extra equations and involve the unknown density f_{u_x}. This representation leads to a generalization of Kotlyarski's identification result similar to that obtained by Evdokimov (2011), who used the joint characteristic function. The equations in model 4a make it possible to identify f_u, f_{u_x} ahead of f_{x^*}; for identification this will require less restrictive conditions on the support of the characteristic function for x^*.

4.2.2.3. *Some Extensions*

A. Common Factor Models

Consider a model $\tilde{z} = AU$, with A a matrix of known constants, $\tilde{z}$ an $m \times 1$ vector of observables, and U a vector of unobservable variables. Usually, A is a block matrix and AU can be represented via a combination of mutually independent vectors. Then without loss of generality consider the model

$$\tilde{z} = \tilde{A}x^* + \tilde{u}, \tag{4.2}$$

where $\tilde{A}$ is an $m \times d$ known matrix of constants, $\tilde{z}$ is an $m \times 1$ vector of observables, unobserved x^* is $d \times 1$, and unobserved $\tilde{u}$ is $m \times 1$. If the model (4.2) can be transformed to model 3 considered above, then x^* will be identified whenever identification holds for model 3. Once some components are identified, identification of other factors could be considered sequentially.

Lemma 4.1. *If in* (4.2) *the vectors x^* and $\tilde{u}$ are independent and all the components of the vector $\tilde{u}$ are mean independent of each other and are mean zero and the matrix A can be partitioned after possibly some permutation of rows as $\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ with rank $A_1 =$ rank $A_2 = d$, then the model* (4.2) *implies model 3.*

Proof. Define $z = T_1\tilde{z}$, where conformably to the partition of A we have the partitioned $T_1 = \begin{pmatrix} \tilde{T}_1 \\ 0 \end{pmatrix}$, with $\tilde{T}_1 A_1 x^* = x^*$ (such a $\tilde{T}_1$ always exists by the rank condition); then $z = x^* + u$, where $u = T_1\tilde{u}$ is independent of x^*. Next define $T_2 = \begin{pmatrix} 0 \\ \tilde{T}_2 \end{pmatrix}$ similarly with $\tilde{T}_2 A_2 x^* = x^*$.

Then $x = T_2\tilde{z}$ is such that $x = x^* + u_x$, where $u_x = T_2\tilde{u}$ and does not include any components from u. This implies $Eu_x|(x^*, u) = 0$. Model 3 holds. ■

Here dependence in components of x^* is arbitrary. A general structure with subvectors of U independent of each other but with components that may be only mean independent (as $\tilde{u}$ here) or arbitrarily dependent (as in x^*) is examined by Ben-Moshe

(2012). Models of linear systems with full independence were examined by, for example, Li and Vuong (1998). These models lead to systems of first-order differential equations for the characteristic functions.

It may be that there are no independent components x^* and $\tilde{u}$ for which the conditions of Lemma 4.1 are satisfied. Bonhomme and Robin (2010) proposed to consider products of the observables to increase the number of equations in the system and analyzed conditions for identification; Ben-Moshe (2012) provided the necessary and sufficient conditions under which this strategy leads to identification when there may be some dependence.

B. Error Correlations with More Observables

The extension to nonzero $E(u_x|z)$ in model 3 is trivial if this expectation is a known function. A more interesting case results if the errors u_x and u are related—for example,

$$u_x = \rho u + \eta; \eta \perp z.$$

With an unknown parameter (or function of observables) ρ, if more observations are available, then more convolution equations can be written to identify all the unknown functions. Suppose that additionally an observation y is available with

$$\begin{aligned} y &= x^* + u_y, \\ u_y &= \rho u_x + \eta_1; \eta_1 \perp, \eta, z. \end{aligned}$$

Without loss of generality, consider the univariate case and define $w_x = E(xf(z)|z); w_y = E(yf(z)|z)$. Then the system of convolution equations expands to

$$\begin{aligned} f_{x^*} * f_u &= w, \\ (1-\rho)h_{x^*} * f_u + \rho z f(z) &= w_x, \\ (1-\rho^2)h_{x^*} * f_u + \rho^2 z f(z) &= w_y. \end{aligned} \tag{4.3}$$

The three equations have three unknown functions, f_{x^*}, f_u, and ρ. Assuming that support of ρ does not include the point 1, ρ can be expressed as a solution to a linear algebraic equation derived from the two equations in (4.3) that include ρ:

$$\rho = (w_x - zf(z))^{-1}(w_y - w_x).$$

4.2.3. Regression Models with Classical and Berkson Errors and the Convolution Equations

4.2.3.1. The List of Models

Table 4.2 provides several regression models and the corresponding convolution equations involving density-weighted conditional expectations.

Table 4.2 Regression Models: 5. Regression with Classical Measurement Error and an Additional Observation; 6. Regression with Berkson Error (x, y, z Are Observable); 7. Regression with Zero Mean Measurement Error and Berkson Instruments

Model	Distributional Assumptions	Convolution Equations	Known Functions	Unknown Functions
5	$y = g(x^*) + v,$ $z = x^* + u,$ $x = x^* + u_x,$ $x^* \perp u; E(u) = 0,$ $E(u_x \vert x^*, u) = 0,$ $E(v \vert x^*, u, u_x) = 0$	$f_{x^*} * f_u = f_z,$ $(g f_{x^*}) * f_u = w,$ $h_k * f_u = w_k,$ with $h_k(x) \equiv x_k g(x) f_{x^*}(x),$ $k = 1, 2, \ldots, d$	f_z; w; w_k	f_{x^*}; f_u; g
6	$y = g(x) + v,$ $z = x + u; E(v \vert z) = 0,$ $z \perp u; E(u) = 0$	$f_x = f_{-u} * f_z,$ $g * f_{-u} = w$	$f_z; f_x, w$	f_u; g
7	$y = g(x^*) + v,$ $x = x^* + u_x,$ $z = x^* + u; z \perp u,$ $E(v \vert z, u, u_x) = 0,$ $E(u_x \vert z, v) = 0$	$g * f_u = w,$ $h_k * f_u = w_k,$ with $h_k(x) \equiv x_k g(x),$ $k = 1, 2, \ldots, d$	w, w_k	f_u; g

Notation: $k = 1, 2, \ldots, d$; in model 5, $w = E(y f_z(z) \vert z), w_k = E(x_k f_z(z) \vert z)$; in model 6, $w = E(y \vert z)$; in model 7, $w = E(y \vert z), w_k = E(x_k y \vert z)$.

Theorem 4.2. *Under Assumption 4.1 for each of the models 5–7 the corresponding convolution equations hold.*

The proof is in the derivations of the next subsection.

4.2.3.2. Discussion of the Regression Models and Derivation of the Convolution Equations

Model 5. The Nonparametric Regression Model with Classical Measurement Error and an Additional Observation. This type of model was examined by Li (2002) and Li and Hsiao (2004); the convolution equations derived here provide a convenient representation. Often models of this type were considered in semiparametric settings. Butucea and Taupin (2008) (extending the earlier approach by Taupin, 2001) consider a regression function known up to a finite-dimensional parameter with the mismeasured variable observed with independent error where the error distribution is known. Under the latter condition, model 5 here would reduce to the two first equations

$$f_{x^*} * f_u = f_z, \qquad \left(g f_{x^*}\right) * f_u = w,$$

where f_u is known and two unknown functions are g (here nonparametric) and f_{x^*}.

Model 5 incorporates model 3 for the regressor, and thus the convolution equations from that model apply. An additional convolution equation is derived here; it is obtained from considering the value of the density-weighted conditional expectation in the dual space of generalized functions, S^*, applied to arbitrary $\psi \in S$,

$$(w, \psi) = (E(f(z)y|z), \psi) = (E(f(z)g(x^*)|z), \psi);$$

this equals

$$\begin{aligned}
&\int\int g(x^*) f_{x^*,z}(x^*, z)\psi(z)\, dx^*\, dz \\
&\quad = \int\int g(x^*) f_{x^*,u}(x^*, u)\psi(x^* + u)\, dx^*\, du \\
&\quad = \int g(x^*) f_{x^*}(x^*) f_u(u)\, dx^*\, \psi(x^* + u)\, dx^*\, du = ((g f_{x^*}) * f_u, \psi).
\end{aligned}$$

Conditional moments for the regression function need not be integrable or bounded functions of z; we require them to be in the space of generalized functions S^*.

Model 6. Regression with Berkson Error. This model may represent the situation when the regressor (observed) x is correlated with the error v, but z is a vector possibly representing an instrument uncorrelated with the regression error.

Then as is known in addition to the Berkson error convolution equation, the equation,

$$w = E(y|z) = E(g(x)|z) = \int g(x) \frac{f_{x,z}(x, z)}{f_z(z)}\, dx = \int g(z - u) f_u(u)\, dx = g * f_u$$

holds. This is stated in Meister (2009); however, the approach there is to consider g to be absolutely integrable so that convolution can be defined in the L_1 space. Here by working in the space of generalized functions S^* a much wider nonparametric class of functions that includes regression functions with polynomial growth is allowed.

Model 7. Nonparametric regression with error in the regressor, where Berkson type instruments are assumed available. This model was proposed by Newey (2001), examined in the univarite case by Schennach (2007) and Zinde-Walsh (2009), and studied in the multivariate case in Zinde-Walsh (2013), where the convolution equations given here in Table 4.2 were derived.

4.2.4. Convolution Equations in Models with Conditional Independence Conditions

Models 1–7 can be extended to include some additional variables where conditionally on those variables, the functions in the model (e.g., conditional distributions) are defined and all the model assumptions hold conditionally.

Evdokimov (2011) derived the conditional version of model 4 from a very general nonparametric panel data model. Model 8 below describes the panel data setup and how it transforms to conditional model 4 and 4a and possibly model 3 with relaxed independence condition (if the focus is on identifying the regression function).

Model 8. Panel Data Model with Conditional Independence. Consider a two-period panel data model with an unknown regression function m and an idiosyncratic (unobserved) α:

$$Y_{i1} = m(X_{i1}, \alpha_i) + U_{i1},$$
$$Y_{i2} = m(X_{i2}, \alpha_i) + U_{i2}.$$

To be able to work with various conditional characteristic functions, corresponding assumptions ensuring existence of the conditional distributions need to be made, and in what follows we assume that all the conditional density functions and moments exist as generalized functions in S^*.

In Evdokimov (2011), independence (conditional on the corresponding period $X's$) of the regression error from α, as well as from the $X's$ and error of the other period, is assumed:

$$f_t = f_{Uit|X_{it},\alpha_i,X_{i(-t)},U_{i(-t)}}(u_t|x,\ldots) = f_{Uit|X_{it}}(u_t|x), \qquad t = 1,2,$$

with $f_{\cdot|\cdot}$ denoting corresponding conditional densities. Conditionally on $X_{i2} = X_{i1} = x$ the model takes the form 4,

$$z = x^* + u,$$
$$x = x^* + u_x,$$

with z representing Y_1, x representing Y_2, and x^* standing in for $m(x,\alpha)$, u for U_1, and u_x for U_2. The convolution equations derived here for 4 or 4a now apply to conditional densities.

The convolution equations in 4a are similar to Evdokimov; they allow for equations for f_u, f_{u_x} that do not rely on f_{x^*}. The advantage of those lies in the possibility of identifying the conditional error distributions without placing the usual nonzero restrictions on the characteristic function of x^* (that represents the function m for the panel model).

The panel model can be considered with relaxed independence assumptions. Here in the two-period model we look at forms of dependence that assume zero conditional mean of the second period error, rather than full independence of the first period error:

$$f_{Ui1|X_{i1},\alpha_i,X_{i2},Ui2}(u_t|x,\dots) = f_{Ui1|Xi1}(u_t|x),$$
$$E(U_{i2}|X_{i1},\alpha_i,X_{i2},U_{i1}) = 0,$$
$$f_{Ui2|\alpha_i,X_{i2}=X_{i1}=x}(u_t|x,\dots) = f_{Ui2|Xi2}(u_t|x).$$

Then the model maps into model 3, with the functions in the convolution equations representing conditional densities, and allows us to identify distribution of x^* (function m in the model). But the conditional distribution of the second-period error in this setup is not identified.

Evdokimov introduced parametric AR(1) or MA(1) dependence in the errors U, and to accommodate this he extended the model to three periods. Here this would lead in the AR case to the Eq. (4.3).

Model 9. Errors in Variables Regression with Classical Measurement Error Conditionally on Covariates. Consider the regression model

$$y = g(x^*, t) + v,$$

with a measurement of unobserved x^* given by $\tilde{z} = x^* + \tilde{u}$, with $x^* \perp \tilde{u}$ conditionally on t. Assume that $E(\tilde{u}|t) = 0$ and that $E(v|x^*, t) = 0$. Then redefining all the densities and conditional expectations to be conditional on t, we get the same system of convolution equations as in Table 4.2 for model 5 with the unknown functions now being conditional densities and the regression function, g.

Conditioning requires assumptions that provide for existence of conditional distribution functions in S^*.

4.3. Solutions for the Models

4.3.1. Existence of Solutions

To state results for nonparametric models, it is important first to clearly indicate the classes of functions where the solution is sought. Assumption 4.1 requires that all the (generalized) functions considered are elements in the space of generalized functions S^*. This implies that in the equations the operation of convolution applied to the two functions from S^* provides an element in the space S^*. This subsection gives high-level assumptions on the nonparametric classes of the unknown functions where the solutions can be sought: Any functions from these classes that enter into the convolution provide a result in S^*.

No assumptions are needed for existence of convolution and full generality of identification conditions in models 1,2 where the model assumptions imply that the functions represent generalized densities. For the other models including regression models, convolution is not always defined in S^*. Zinde-Walsh (2013) defines the concept of convolution pairs of classes of functions in S^* where convolution can be applied.

To solve the convolution equations, a Fourier transform is usually employed; thus, for example, one transforms generalized density functions into characteristic functions. Fourier transform is an isomorphism of the space S^*. The Fourier transform of a generalized function $a \in S^*$, $Ft(a)$, is defined as follows. For any $\psi \in S$, as usual $Ft(\psi)(s) = \int \psi(x)e^{isx}dx$; then the functional $Ft(a)$ is defined by

$$(Ft(a), \psi) \equiv (a, Ft(\psi)).$$

The advantage of applying Fourier transform is that integral convolution equations transform into algebraic equations when the "exchange formula" applies:

$$a*b = c \Longleftrightarrow Ft(a) \cdot Ft(b) = Ft(c). \tag{4.4}$$

In the space of generalized functions S^*, the Fourier transform and inverse Fourier transform always exist. As shown in Zinde-Walsh (2013), there is a dichotomy between convolution pairs of subspaces in S^* and the corresponding product pairs of subspaces of their Fourier transforms.

The classical pairs of spaces (Schwartz, 1966) are the convolution pair $\left(S^*, O_C^*\right)$ and the corresponding product pair (S^*, O_M), where O_C^* is the subspace of S^* that contains rapidly decreasing (faster than any polynomial) generalized functions and O_M is the space of infinitely differentiable functions with every derivative growing no faster than a polynomial at infinity. These pairs are important in that no restriction is placed on one of the generalized functions that could be any element of space S^*; the other belongs to a space that needs to be correspondingly restricted. A disadvantage of the classical pairs is that the restriction is fairly severe; for example, the requirement that a characteristic function be in O_M implies existence of all moments for the random variable. Relaxing this restriction would require placing constraints on the other space in the pair; Zinde-Walsh (2013) introduces some pairs that incorporate such tradeoffs.

In some models the product of a function with a component of the vector of arguments is involved, such as $d(x) = x_k a(x)$; then for Fourier transforms $Ft(d)(s) = -i\frac{\partial}{\partial s_k}Ft(a)(s)$; the multiplication by a variable is transformed into $(-i)$ times the corresponding partial derivative. Since the differentiation operators are continuous in S^*, this transformation does not present a problem.

Assumption 4.2. *The functions $a \in A, b \in B$, are such that (A, B) form a convolution pair in S^*.*

Equivalently, $Ft(a)$, $Ft(b)$ are in the corresponding product pair of spaces.

Assumption 4.2 is applied to model 1 for $a = f_{x^*}, b = f_u$; to model 2 with $a = f_z, b = f_u$; to model 3 with $a = f_{x^*}, b = f_u$ and with $a = h_k, b = f_u$, for all $k = 1, \ldots, d$; to model 4a for $a = f_{x^*}$, or f_{u_x}, or h_k for all k and $b = f_u$; to model 5 with $a = f_{x^*}$, or gf_{x^*}, or $h_k f_{x^*}$ and $b = f_u$; to model 6 with $a = f_z$, or g and $b = f_u$; to model 7 with $a = g$ or h_k and $b = f_u$.

Assumption 4.2 is a high-level assumption that is a sufficient condition for a solution to the models 1–4 and 6–7 to exist. Some additional conditions are needed for model 5 and are provided below.

Assumption 4.2 is automatically satisfied for generalized density functions, so is not needed for models 1 and 2. Denote by $\bar{D} \subset S^*$ the subset of generalized derivatives of distribution functions (corresponding to Borel probability measures in R^d), then in models 1 and 2 $A = B = \bar{D}$; and for the characteristic functions there are correspondingly no restrictions; denote the set of all characteristic functions, $Ft\left(\bar{D}\right) \subset S^*$, by $\bar{C}$.

Below a (non-exhaustive) list of nonparametric classes of generalized functions that provide sufficient conditions for existence of solutions to the models here is given. The classes are such that they provide minimal or often no restrictions on one of the functions and restrict the class of the other in order that the assumptions be satisfied.

In models 3 and 4 the functions h_k are transformed into derivatives of continuous characteristic functions. An assumption that either the characteristic function of x^* or the characteristic function of u be continuously differentiable is sufficient, without any restrictions on the other to ensure that Assumption 4.2 holds. Define the subset of all continuously differentiable characteristic functions by $\bar{C}^{(1)}$.

In model 5, equations involve a product of the regression function g with f_{x^*}. Products of generalized functions in S^* do not always exist, and so additional restrictions are needed in that model. If g is an arbitrary element of S^*, then for the product to exist, f_{x^*} should be in O_M. On the other hand, if f_{x^*} is an arbitrary generalized density, it is sufficient that g and h_k belong to the space of d times continuously differentiable functions with derivatives that are majorized by polynomial functions for $gf_{x^*}, h_k f_{x^*}$ to be elements of S^*. Indeed, the value of the functional $h_k f_{x^*}$ for an arbitrary $\psi \in S$ is defined by

$$(h_k f_{x^*}, \psi) = (-1)^d \int F_{x^*}(x)\partial^{(1,\ldots,1)}(h_k(x)\psi(x))\,dx;$$

here F is the distribution (ordinary bounded) function and this integral exists because ψ and all its derivatives go to zero at infinity faster than any polynomial function. Denote by $\bar{S}^{B,1}$ the space of continuously differentiable functions $g \in S^*$ such that the functions $h_k(x) = x_k g(x)$ are also continuously differentiable with all derivatives majorized by polynomial functions. Since the products are in S^*, then the Fourier transforms of the products are defined in S^*. Further restrictions requiring the Fourier transforms of the products gf_{x^*} and $h_k f_{x^*}$ to be continuously differentiable functions in S^* would remove any restrictions on f_u for the convolution to exist. Denote the space of all continuously differentiable functions in S^* by $\bar{S}^{(1)}$.

If g is an ordinary function that represents a regular element in S^*, the infinite differentiability condition on f_{x^*} can be relaxed to simply requiring continuous first derivatives.

In models 6 and 7, if the generalized density function for the error, f_u, decreases faster than any polynomial (all moments need to exist for that), so that $f_u \in O_C^*$, then g could be any generalized function in S^*; this will of course hold if f_u has bounded support. Generally, the more moments the error is assumed to have, the fewer the number of restrictions on the regression function g that are needed to satisfy the convolution equations of the model and the exchange formula. Models 6 and 7 satisfy the assumptions for any error u when support of generalized function g is compact (as for the "sum of peaks"); then $g \in E^* \subset S^*$, where E^* is the space of generalized functions with compact support. More generally the functions g and all the h_k could belong to the space O_C^* of generalized functions that decrease at infinity faster than any polynomial, and still no restrictions need to be placed on u.

Denote for any generalized density function f the corresponding characteristic function, $Ft(f_.)$, by $\phi_.$. Denote the Fourier transform of the (generalized) regression function g, $Ft(g)$, by γ.

Table 4.3 summarizes some fairly general sufficient conditions on the models that place restrictions on the functions themselves or on the characteristic functions of distributions in the models that will ensure that Assumption 4.2 is satisfied and a solution exists. The nature of these assumptions is to provide restrictions on some of the functions that allow the others to be completely unrestricted for the corresponding model.

Table 4.4 states the equations and systems of equations for Fourier transforms that follow from the convolution equations.

Assumption 4.2 (that is fulfilled, for example, by generalized functions classes of Table 4.3) ensures existence of solutions to the convolution equations for models 1–7; this does not exclude multiple solutions, and the next section provides a discussion of solutions for equations in Table 4.4.

Table 4.3 Some Nonparametric Classes of Generalized Functions for Which the Convolution Equations of the Models are Defined in S^*

Model	Sufficient assumptions	
1	No restrictions: $\phi_{x^*} \in \bar{C}; \phi_u \in \bar{C}$	
2	No restrictions: $\phi_{x^*} \in \bar{C}; \phi_u \in \bar{C}$	
	Assumptions A	Assumptions B
3	Any $\phi_{x^*} \in \bar{C}; \phi_u \in \bar{C}^{(1)}$	Any $\phi_u \in \bar{C}; \phi_{x^*} \in \bar{C}^{(1)}$
4	Any $\phi_{u_x}, \phi_{x^*} \in \bar{C}; \phi_u \in \bar{C}^{(1)}$	Any $\phi_u, \phi_{x^*} \in \bar{C}; \phi_{u_x} \in \bar{C}^{(1)}$
4a	Any $\phi_{u_x}, \phi_{x^*} \in \bar{C}; \phi_u \in \bar{C}^{(1)}$	Any $\phi_u, \phi_{u_x} \in \bar{C}; \phi_{x^*} \in \bar{C}^{(1)}$
5	Any $g \in S^*; f_{x^*} \in O_M; f_u \in O_C^*$	Any $f_{x^*} \in \bar{D};\ g, h_k \in \bar{S}^{B,1}; f_u \in O_C^*$
6	Any $g \in S^*; f_u \in O_C^*$	$g \in O_C^*$; any f_u: $\phi_u \in \bar{C}$
7	Any $g \in S^*; f_u \in O_C^*$	$g \in O_C^*$; any f_u: $\phi_u \in \bar{C}$

4.3.2. Classes of Solutions; Support and Multiplicity of Solutions

Typically, support assumptions are required to restrict multiplicity of solutions; here we examine the dependence of solutions on the support of the functions. The results here also give conditions under which some zeros—for example, in the characteristic functions—are allowed. Thus in common with, for example, Carrasco and Florens (2010) and Evdokimov and White (2012), distributions such as the uniform or triangular for which the characteristic function has isolated zeros are not excluded. The difference here is the extension of the consideration of the solutions to S^* and to models such as the regression model where this approach to relaxing support assumptions was not previously considered.

Recall that for a continuous function $\psi(x)$ on R^d, support is closure of the set $W = \text{supp}(\psi)$, such that

$$\psi(x) = \begin{cases} a \neq 0 & \text{for } x \in W, \\ 0 & \text{for } x \in R^d \backslash W. \end{cases}$$

The set W is an open set.

Table 4.4 The Form of the Equations for the Fourier Transforms

Model	Equations for Fourier Transforms	Unknown Functions
1	$\phi_{x^*}\phi_u = \phi_z$	ϕ_{x^*}
2	$\phi_{x^*} = \phi_z\phi_{-u}$	ϕ_{x^*}
3	$\phi_{x^*}\phi_u = \phi_z,$ $(\phi_{x^*})'_k\phi_u = \varepsilon_k, k = 1,\ldots,d$	ϕ_{x^*},ϕ_u
4	$\phi_{x^*}\phi_u = \phi_z,$ $(\phi_{x^*})'_k\phi_u = \varepsilon_k, k = 1,\ldots,d,$ $\phi_{x^*}\phi_{u_x} = \phi_x$	$\phi_{x^*},\phi_u,\phi_{u_x}$
4a	$\phi_{u_x}\phi_u = \phi_{z-x},$ $\left(\phi_{u_x}\right)'_k\phi_u = \varepsilon_k, k = 1,\ldots,d,$ $\phi_{x^*}\phi_{u_x} = \phi_x$	–"–
5	$\phi_{x^*}\phi_u = \phi_z,$ $Ft(gf_{x^*})\phi_u = \varepsilon,$ $(Ft\,(gf_{x^*}))'_k\phi_u = \varepsilon_k, k = 1,\ldots,d.$	ϕ_{x^*},ϕ_u,g
6	$\phi_x = \phi_{-u}\phi_z,$ $Ft(g)\phi_{-u} = \varepsilon$	ϕ_u,g
7	$Ft(g)\phi_u = \varepsilon,$ $(Ft\,(g))'_k\phi_u = \varepsilon_k, k = 1,\ldots,d$	ϕ_u,g

Notation: $(\cdot)'_k$ denotes the kth partial derivative of the function. The functions ε are Fourier transforms of the corresponding w, and $\varepsilon_k = -iFt(w_k)$ is defined for the models in Tables 4.1 and 4.2.

Generalized functions are functionals on the space S, and support of a generalized function $b \in S^*$ is defined as follows (Schwartz, 1966, p. 28). Denote by (b, ψ) the value of the functional b for $\psi \in S$. Define a null set for $b \in S^*$ as the union of supports of all functions in S for which the value of the functional is zero: $\Omega = \{\cup \text{supp}(\psi), \psi \in S,$ such that $(b, \psi) = 0\}$. Then $\text{supp}(b) = R^d \backslash \Omega$. Note that a generalized function has support in a closed set, for example, support of the δ-*function* is just one point 0.

Note that for model 2, Table 4.4 gives the solution for ϕ_{x^*} directly and the inverse Fourier transform can provide the (generalized) density function, f_{x^*}.

In Zinde-Walsh (2013), identification conditions in S^* were given for models 1 and 7 under assumptions that include the ones in Table 4.3 but could also be more flexible.

The equations in Table 4.3 for models 1, 3, 4, 4a, 5, 6, and 7 are of two types, similar to those solved in Zinde-Walsh (2013). One is a convolution with one unknown function; the other is a system of equations with two unknown functions, each leading to the corresponding equations for their Fourier transforms.

4.3.2.1. Solutions to the Equation $\alpha\beta = \gamma$

Consider the equation

$$\alpha\beta = \gamma, \tag{4.5}$$

with one unknown function α; β is a given continuous function. By Assumption 4.2 the nonparametric class for α is such that the equation holds in S^* on R^d; it is also possible to consider a nonparametric class for α with restricted support, $\bar{W}$. Of course without any restrictions $\bar{W} = R^d$. Recall the differentiation operator, ∂^m, for $m = (m_1, \ldots, m_d)$ and denote by $\text{supp}(\beta, \partial)$ the set $\cup_{\Sigma m_i = 0}^{\infty} \text{supp}(\partial^m \beta)$; where $\text{supp}(\partial^m \beta)$ is an open set where a continuous non-zero derivative $\partial^m \beta$ exists. Any point where β is zero belongs to this set if some finite-order partial continuous derivative of β is not zero at that point (and in some open neighborhood); for β itself $\text{supp}(\beta) \equiv \text{supp}(\beta, 0)$.

Define the functions

$$\alpha_1 = \beta^{-1}\gamma I\big(\text{supp}(\beta, \partial)\big); \qquad \alpha_2(x) = \begin{cases} 1 & \text{for } x \in \text{supp}(\beta, \partial), \\ \tilde{\alpha} & \text{for } x \in \bar{W} \backslash (\text{supp}(\beta, \partial)) \end{cases} \tag{4.6}$$

with any $\tilde{\alpha}$ in S^* such that $\alpha_1\alpha_2 \in Ft(A)$.

Consider the case when α, β, and thus γ are continuous. For any point x_0 if $\beta(x_0) \neq 0$, there is a neighborhood $N(x_0)$ where $\beta \neq 0$, and division by β is possible. If $\beta(x_0)$ has a zero, it could only be of finite order, and in some neighborhood, $N(x_0) \in \text{supp}(\partial^m \beta)$, a representation

$$\beta = \eta(x) \Pi_{i=1}^{d} (x_i - x_{0i})^{m_i} \tag{4.7}$$

holds for some continuous function η in S^*, such that $\eta > 0$ on $\text{supp}(\eta)$. Then $\eta^{-1}\gamma$ in $N(x_0)$ is a nonzero continuous function; division of such a function by $\Pi_{i=1}^{d} (x_i - x_{0i})^{m_i}$ in S^* is defined (Schwartz, 1966, pp. 125–126), thus division by β is defined in this neighborhood $N(x_0)$. For the set $\text{supp}(\beta, \partial)$ consider a covering of every point by such neighborhoods, the possibility of division in each neighborhood leads to the possibility of division globally on the whole $\text{supp}(\beta, \partial)$. Then a_1 as defined in (4.6) exists in S^*.

In the case where γ is an arbitrary generalized function, if β is infinitely differentiable, then by Schwartz (1966, pp. 126–127) division by β is defined on $\text{supp}(\beta, \partial)$ and the solution is given by (4.6) and is unique up to a function supported in isolated zeros of β.

For the cases where γ is not continuous and β is not infinitely differentiable, the solution is provided by

$$\alpha_1 = \beta^{-1}\gamma I(\text{supp}(\beta,0)); \qquad \alpha_2(x) = \begin{cases} 1 & \text{for } x \in \text{supp}(\beta,0), \\ \tilde{\alpha} & \text{for } x \in \bar{W}\backslash(\text{supp}(\beta,0)) \end{cases}$$

with any $\tilde{\alpha}$ such that $\alpha_1\alpha_2 \in Ft\,(A)$.

Theorem 2 in Zinde-Walsh (2013) implies that the solution to (4.5) is $a = Ft^{-1}(\alpha_1\alpha_2)$; the sufficient condition for the solution to be unique is $\text{supp}(\beta,0) \supset \bar{W}$; if additionally either γ is a continuous function or β is an infinitely continuously differentiable function, it is sufficient for uniqueness that $\text{supp}(\beta,\partial) \supset \bar{W}$.

This provides solutions for models 1 and 6, where only equations of this type appear.

4.3.2.2. Solutions to the System of Equations

For models 3, 4, 5, and 7 a system of equations of the form

$$\begin{aligned} &\alpha\beta = \gamma, \\ &\alpha\beta_k' = \gamma_k, \qquad k = 1,\ldots,d. \end{aligned} \tag{4.8}$$

(with β continuously differentiable) arises. Theorem 3 in Zinde-Walsh (2013) provides the solution and uniqueness conditions for this system of equations. It is first established that a set of continuous functions $\varkappa_k, k = 1,\ldots,d$, that solves the equation

$$\varkappa_k\gamma - \gamma_k = 0 \tag{4.9}$$

in the space S^* exists and is unique on $W = \text{supp}(\gamma)$ as long as $\text{supp}(\beta) \supset W$. Then $\beta_k'\beta^{-1} = \varkappa_k$ and substitution into (4.9) leads to a system of first-order differential equations in β.

Case 4.1. Continuous functions; W is an open set.

For models 3 and 4 the system (4.8) involves continuous characteristic functions; thus W is an open set. In some cases W can be an open set under conditions of models 5 and 7—for example, if the regression function is integrable in model 7.

For this case represent the open set W as a union of (maximal) connected components $\cup_\nu W_\nu$.

Then by the same arguments as in the proof of Theorem 3 in Zinde-Walsh (2012) the solution can be given uniquely on W as long as at some point $\zeta_{0\nu} \in (W_\nu \cap W)$ the value $\beta(\zeta_{0\nu})$ is known for each of the connected components. Consider then $\beta_1(\zeta) = \Sigma_\nu [\beta\,(\zeta_{0\nu}) \exp \int_{\zeta_0}^{\zeta} \sum_{k=1}^{d} \varkappa_k(\xi)\, d\xi\,] I(W_\nu)$, where integration is along any arc within the component that connects ζ to $\zeta_{0\nu}$. Then $\alpha_1 = \beta_1^{-1}\gamma$, and α_2, β_2 are defined as above by being 1 on $\cup_\nu W_\nu$ and arbitrary outside of this set.

When $\beta(0) = 1$, as is the case for the characteristic function, the function is uniquely determined on the connected component that includes 0.

Evdokimov and White (2012) provide a construction that permits, in the univariate case, to extend the solution $\beta\,(\zeta_{0\nu})\,[\exp\int_{\zeta_0}^{\zeta}\sum_{k=1}^{d}\varkappa_k(\xi)d\xi\,]I(W_\nu)$ from a connected component of support where $\beta\,(\zeta_{0\nu})$ is known (e.g., at 0 for a characteristic function) to a contiguous connected component when on the border between the two where $\beta = 0$, at least some finite order derivative of β is not zero. In the multivariate case this approach can be extended to the same construction along a one-dimensional arc from one connected component to the other. Thus identification is possible on a connected component of $\text{supp}(\beta,\partial)$.

Case 4.2. W is a closed set.

Generally for models 5 and 7, W is the support of a generalized function and is a closed set. It may intersect with several connected components of support of β. Denote by W_ν here the intersection of a connected component of support of β and W. Then similarly $\beta_1(\zeta) = \sum_\nu\,[\beta(\zeta_{0\nu})\exp\int_{\zeta_0}^{\zeta}\sum_{k=1}^{d}\varkappa_k(\xi)\,d\xi\,]I(W_\nu)$, where integration is along any arc within the component that connects ζ to $\zeta_{0\nu}$. Then $\alpha_1 = \beta_1^{-1}\varepsilon$, and α_2, β_2 are defined as above by being 1 on $\cup_\nu W_\nu$ and arbitrary outside of this set. The issue of the value of β at some point within each connected component arises. In the case of β being a characteristic function, if there is only one connected component, W, and $0 \in W$ the solution is unique, since then $\beta(0) = 1$.

Note that for model 5 the solution to equations of the type (4.8) would only provide $Ft(gf_{x^*})$ and ϕ_u; then from the first equation for this model in Table 4.4, ϕ_{x^*} can be obtained; it is unique if $\text{supp}\phi_{x^*} = \text{supp}\phi_z$. To solve for g, find $g = Ft^{-1}(Ft(gf_{x^*}))\cdot(f_{x^*})^{-1}$.

4.4. Identification, Partial Identification and Well-Posedness

4.4.1. Identified Solutions for the Models 1–7

As follows from the discussion of the solutions, uniqueness in models 1, 2, 3, 4, 4a, 5, and 6 holds (in a few cases up to a value of a function at a point) if all the Fourier transforms are supported over the whole R^d; in many cases it is sufficient that $\text{supp}(\beta,\partial) = R^d$.

The classes of functions could be defined with Fourier transforms supported on some known subset $\bar{W}$ of R^d, rather than on the whole space; if all the functions considered have $\bar{W}$ as their support, and the support consists of one connected component that includes 0 as an interior point, then identification for the solutions holds. For Table 4.5, assume that $\bar{W}$ is a single connected component with 0 as an interior point; again $\bar{W}$ could coincide with $\text{supp}(\beta,\partial)$. For model 5 under Assumption B,

Table 4.5 The Solutions for Identified Models on $\bar{W}$

Model	Solution to Equations
1	$f_{x^*} = Ft^{-1}(\phi_u^{-1}\phi_z)$.
2	$f_{x^*} = Ft^{-1}(\phi_{-u}\phi_z)$.
3	Under Assumption A $f_{x^*} = Ft^{-1}(\exp \int_{\zeta_0}^{\zeta} \sum_{k=1}^{d} \varkappa_k(\xi)\, d\xi)$, where $\varkappa_k$ solves $\varkappa_k \phi_z - [(\phi_z)'_k - \varepsilon_k] = 0$; $f_u = Ft^{-1}(\phi_{x^*}^{-1}\varepsilon)$. Under Assumption B $f_u = Ft^{-1}(\exp \int_{\zeta_0}^{\zeta} \sum_{k=1}^{d} \varkappa_k(\xi)\, d\xi)$; $\varkappa_k$ solves $\varkappa_k \phi_z - \varepsilon_k = 0$; $f_{x^*} = Ft^{-1}(\phi_u^{-1}\varepsilon)$.
4	f_{x^*}, f_u obtained similarly to those in model 3; $\phi_{u_x} = \phi_{x^*}^{-1}\phi_x$.
4a	f_{u_x}, f_u obtained similarly to ϕ_{x^*}, ϕ_u in model 3; $\phi_{x^*} = \phi_{u_x}^{-1}\phi_x$.
5	Three steps: 1. (a) Get $Ft(gf_{x^*}), \phi_u$ similarly to ϕ_{x^*}, ϕ_u in model 3 (under Assumption A use $Ft(gf_{x^*})(0)$); 2. Obtain $\phi_{x^*} = \phi_u^{-1}\phi_z$; 3. Get $g = \left[Ft^{-1}(\phi_{x^*})\right]^{-1} Ft^{-1}(Ft(gf_{x^*}))$.
6	$\phi_{-u} = \phi_z^{-1}\phi_x$ and $g = Ft^{-1}(\phi_x^{-1}\phi_z\varepsilon)$.
7	$\phi_{x^*}, Ft(g)$ obtained similarly to ϕ_{x^*}, ϕ_u in model 3 (under Assumption A use $Ft(g)(0)$).

assume additionally that the value at zero, $Ft(gf_{x^*})(0)$, is known; similarly for model 7 under Assumption B, additionally assume that $Ft(g)(0)$ is known.

4.4.2. Implications of Partial Identification

Consider the case of model 1. Essentially lack of identification, say in the case when the error distribution has characteristic function supported on a convex domain W_u around zero, results in the solution for $\phi_{x^*} = \phi_1\phi_2$; here ϕ_1 is nonzero and unique on W_u and thus captures the lower-frequency components of x^* and where ϕ_2 is a characteristic function of a distribution with arbitrary high-frequency components.

Transforming back to densities provides a corresponding model with independent components

$$z = x_1^* + x_2^* + u,$$

where x_1^* uniquely extracts the lower-frequency part of observed z. The more important the contribution of x_1^* to x^*, the less important is lack of identification.

If the feature of interest as discussed, for example, by Matzkin (2007) involves only low-frequency components of x^*, it may still be fully identified even when the distribution for x^* itself is not. An example of that is a deconvolution applied to an image of a car captured by a traffic camera, although even after deconvolution the image may still appear blurry the license plate number may be clearly visible. In nonparametric regression the polynomial growth of the regression or the expectation of the response function may be identifiable even if the regression function is not fully identified.

Features that are identified include any functional, Φ, linear or nonlinear on a class of functions of interest, such that in the frequency domain Φ is supported on W_u.

4.4.3. Well-Posedness in S^*

Conditions for well-posedness in S^* for solutions of the equations entering in models 1–7 were established in Zinde-Walsh (2013). Well-posedness is needed to ensure that if a sequence of functions converges (in the topology of S^*) to the known functions of the equations characterizing the models 1–7 in Tables 4.1 and 4.2, then the corresponding sequence of solutions will converge to the solution for the limit functions. A feature of well-posedness in S^* is that the solutions are considered in a class of functions that is a bounded set in S^*.

The properties that differentiation is a continuous operation, and that the Fourier transform is an isomorphism of the topological space S^*, make conditions for convergence in this space much weaker than those in functions spaces, say, L_1, L_2. Thus for density that is given by the generalized derivative of the distribution function, well-posedness holds in spaces of generalized functions by the continuity of the differentiation operator.

For the problems here, however, well-posedness does not always obtain. The main sufficient condition is that the inverse of the characteristic function of the measurement error satisfy the condition (4.1) with $b = \phi_u^{-1}$ on the corresponding support. This holds if either the support is bounded or the distribution is not supersmooth. If ϕ_u has some zeros but satisfies the identification conditions so that it has local representation (4.7) where (4.1) is satisfied for $b = \eta^{-1}$, well-posedness will hold.

The example in Zinde-Walsh (2013) demonstrates that well-posedness of deconvolution will not hold even in the weak topology of S^* for supersmooth (e.g., Gaussian) distributions on unbounded support. On the other hand, well-posedness of deconvolution in S^* obtains for ordinary smooth distributions; and thus under less restrictive conditions than in function spaces, such as L_1 or L_2 usually considered.

In models 3–7 with several unknown functions, more conditions are required to ensure that all the operations by which the solutions are obtained are continuous in the topology of S^*. It may not be sufficient to assume (4.1) for the inverses of unknown functions where the solution requires division; for continuity of the solution the condition may need to apply uniformly.

Define a class of regular functions on R^d, $\Phi(m, V)$ (with m a vector of integers, V a positive constant) where $b \in \Phi(m, V)$ if

$$\int \Pi((1+t_i^2)^{-1})^{m_i} |b(t)| \, dt < V < \infty. \tag{4.10}$$

Then similarly to Zinde-Walsh (2013), well-posedness can be established for model 7 as long as in addition to Assumption A or B, for some $\Phi(m, V)$ both ϕ_u and ϕ_u^{-1} belong to the class $\Phi(m, V)$. This condition is fulfilled by non-supersmooth ϕ_u; this could be an ordinary smooth distribution or a mixture with some mass point.

A convenient way of imposing well-posedness is to restrict the support of functions considered to a bounded $\bar{W}$. If the features of interest are associated with low-frequency components only, then if the functions are restricted to a bounded space, the low-frequency part can be identified and is well-posed.

4.5. Implications for Estimation

4.5.1. Plug-in Nonparametric Estimation

Solutions in Table 4.5 for the equations that express the unknown functions via known functions of observables give scope for plug-in estimation. We see for example, that the equations in models 4 and 4a are different expressions that will provide different plug-in estimators for the same functions.

The functions of the observables here are characteristic functions and Fourier transforms of density-weighted conditional expectations—and in some cases their derivatives—that can be estimated by nonparametric methods. There are some direct estimators—for example, for characteristic functions. In the space S^* the Fourier transform and inverse Fourier transform are continuous operations; thus using standard estimators of density-weighted expectations and applying the Fourier transform would provide consistency in S^*; the details are provided in Zinde-Walsh (2013). Then the solutions can be expressed via those estimators by the operations from Table 4.5; and, as long as the problem is well-posed, the estimators will be consistent and the convergence will obtain at the appropriate rate. As in An and Hu (2012), the convergence rate may be even faster for well-posed problems in S^* than the usual nonparametric rate in (ordinary) function spaces. For example, as demonstrated in Zinde-Walsh (2008), kernel estimators of density that may diverge if the distribution function is not absolutely continuous are always (under the usual assumptions on kernel/bandwidth) consistent in the weak topology of the space of generalized functions, where the density problem is well-posed. Here, well-posedness holds for deconvolution as long as the error density is not supersmooth.

4.5.2. Regularization in Plug-In Estimation

When well-posedness cannot be ensured, plug-in estimation will not provide consistent results and some regularization is required; usually spectral cutoff is employed for the problems considered here. In the context of these non-parametric models, regularization requires extra information: the knowledge of the rate of decay of the Fourier transform of some of the functions.

For model 1 this is not a problem since ϕ_u is assumed known; the regularization uses the information about the decay of this characteristic function to construct a sequence of compactly supported solutions with support increasing at a corresponding rate. In S^* no regularization is required for plug-in estimation unless the error distribution is supersmooth. Exponential growth in ϕ_u^{-1} provides a logarithmic rate of convergence in function classes for the estimator (Fan, 1991). Below we examine spectral cutoff regularization for the deconvolution in S^* when the error density is supersmooth.

With supersmooth error in S^*, define a class of generalized functions $\Phi(\Lambda, m, V)$ for some non-negative-valued function Λ; we have a generalized function $b \in \Phi(\Lambda, m, V)$ if there exists a function $\bar{b}(\zeta) \in \Phi(m, V)$ such that also $\bar{b}(\zeta)^{-1} \in \Phi(m, V)$ and $b = \bar{b}(\zeta)\exp(-\Lambda(\zeta))$. Note that a linear combination of functions in $\Phi(\Lambda, m, V)$ belongs to the same class. Define convergence: A sequence of $b_n \in \Phi(\Lambda, m, V)$ converges to zero if the corresponding sequence $\bar{b}_n$ converges to zero in S^*.

Convergence in probability for a sequence of random functions, ε_n, in S^* is defined as follows: $(\varepsilon_n - \varepsilon) \to_p 0$ in S^* if for any set $\psi_1, \ldots, \psi_v \in S$ the random vector of the values of the functionals converges: $((\varepsilon_n - \varepsilon, \psi_1), \ldots, (\varepsilon_n - \varepsilon, \psi_v)) \to_p 0$.

Lemma 4.2. *If in model 1 $\phi_u = b \in \Phi(\Lambda, m, V)$, where Λ is a polynomial function of order no more than k, and ε_n is a sequence of estimators of ε that are consistent in S^*: $r_n(\varepsilon_n - \varepsilon) \to_p 0$ in S^* at some rate $r_n \to \infty$, then for any sequence of constants $\bar{B}_n$: $0 < \bar{B}_n < (\ln r_n)^{1/k}$ and the corresponding set $B_n = \{\zeta : \|\zeta\| < \bar{B}_n\}$ the sequence of regularized estimators $\phi_u^{-1}(\varepsilon_n - \varepsilon)I(B_n)$ converges to zero in probability in S^*.*

Proof. For n the value of the random functional

$$(\phi_u^{-1}(\varepsilon_n - \varepsilon)I(B_n), \psi) = \int \bar{b}^{-1}(\zeta) r_n(\varepsilon_n - \varepsilon) r_n^{-1} I(B_n) \exp(\Lambda(\zeta)) \psi(\zeta)\, d\zeta.$$

Multiplication by $\bar{b}^{-1} \in \Phi(m, V)$, which corresponds to $\phi_u = b$, does not affect convergence; thus $\bar{b}^{-1}(\zeta) r_n(\varepsilon_n - \varepsilon)$ converges to zero in probability in S^*. To show that $(\phi_u^{-1}(\varepsilon_n - \varepsilon)I(B_n), \psi)$ converges to zero, it is sufficient to show that the function $r_n^{-1} I(B_n) \exp(\Lambda(\zeta))\psi(\zeta)$ is bounded. It is then sufficient to find B_n such that $r_n^{-1} I(B_n) \exp(\Lambda(\zeta))$ is bounded (by possibly a polynomial), thus it is sufficient that $\sup_{B_n} |\exp(\Lambda(\zeta)) r_n^{-1}|$ be bounded. This will hold if $\exp(\bar{B}_n^k) < r_n, \bar{B}_n^k < \ln r_n$. ■

Of course an even slower growth for spectral cutoff would result from Λ that grows faster than a polynomial. The consequence of the slow growth of the support is usually a correspondingly slow rate of convergence for $\phi_u^{-1}\varepsilon_n I(B_n)$. Additional conditions (as in function spaces) are needed for the regularized estimators to converge to the true γ.

It may be advantageous to focus on lower-frequency components and ignore the contribution from high frequencies when the features of interest depend on the contribution at low frequency.

4.6. Concluding Remarks

Working in spaces of generalized functions extends the results on nonparametric identification and well-posedness for a wide class of models. Here identification in deconvolution is extended to generalized densities in the class of all distributions from the usually considered classes of integrable density functions. In regression with Berkson error, nonparametric identification in S^* holds for functions of polynomial growth, extending the usual results obtained in L_1; a similar extension applies to regression with measurement error and Berkson-type measurement; this allows us to consider binary choice and polynomial regression models. Also, identification in models with a sum-of-peaks regression function that cannot be represented in function spaces is included. Well-posedness results in S^* also extend the results in the literature provided in function spaces; well-posedness of deconvolution holds as long as the characteristic function of the error distribution does not go to zero at infinity too fast (as, e.g., supersmooth), and a similar condition provides well-posedness in the other models considered here.

Further investigation of the properties of estimators in spaces of generalized functions requires (a) deriving the generalized limit process for the function being estimated and (b) investigating when it can be described as a generalized Gaussian process. A generalized Gaussian limit process holds for kernel estimator of the generalized density function (Zinde-Walsh, 2008). Determining the properties of inference based on the limit process for generalized random functions requires both further theoretical development and simulations evidence.

Notes

[†] McGill University and CIREQ. The support of the Social Sciences and Humanities Research Council of Canada (SSHRC) and the *Fonds québecois de la recherche sur la société et la culture* (FRQSC) is gratefully acknowledged.

References

An, Y., and Y. Hu. 2012. "Well-Posedness of Measurement Error Models for Self-Reported Data." *Journal of Econometrics*, **168**, pp. 259–269.

Ben-Moshe, D. 2012. "Identification of Dependent Multidimensional Unobserved Variables in a System of Linear Equations." Working paper, UCLA.

Bonhomme, S., and J.-M. Robin. 2010. "Generalized Nonparametric Deconvolution with an Application to Earnings Dynamics." *Review of Economic Studies*, **77**, pp. 491–533.

Butucea, C., and M. L. Taupin. 2008. "New M-Estimators in Semi-parametric Regression with Errors in Variables." *Annales de l'Institut Henri Poincaré—Probabilités et Statistiques*, **44**, pp. 393–421.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and C. M. Crainiceanu. 2006. "*Measurement Error in Nonlinear Models: A Modern Perspective*. New York: Chapman & Hall.

Carrasco, M., and J.-P. Florens. 2010. "A Spectral Method for Deconvolving a Density." *Econometric Theory*, **27**, pp. 546–581.

Chen, X., H. Hong, and D. Nekipelov. 2011. "Nonlinear Models of Measurement Errors." *Journal of Economic Literature*, **49**, pp. 901–937.

Cunha, F., J. J. Heckman, and S. M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica*, **78**, 883–933.

Evdokimov, K. 2011. "Identification and Estimation of a Nonparametric Panel Data Model with Unobserved Heterogeneity." Working paper.

Evdokimov, K., and H. White. 2012. An Extension of a Lemma of Kotlyarski." *Econometric Theory*, **28**(4), 925–932.

Fan, J. Q. 1991. "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems." *Annals of Statistics*, **19**, pp. 1257–1272.

Green, D. A., and W. C. Riddell. 1997. "Qualifying for Unemployment Insurance: An Empirical Analysis." *Economic Journal*, **107**, pp. 67–84.

Hadamard, J. 1923. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. New Haven, CT: Yale University Press.

Karlin, S. 1959. *The Theory of Infinite Games*, Vol. II of Mathematical Methods and Theory in Games, Programming and Economics. Boston: Addison-Wesley.

Kotlyarski, I. 1967. "On Characterizing the Gamma and Normal Distribution." *Pacific Journal of Mathematics*, **20**, pp. 69–76.

Li, T. 2002. "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models." *Journal of Econometrics*, **110**, pp. 1–26.

Li, T., and Ch. Hsiao. 2004. “Robust Estimation of Generalized Models with Measurement Error.” *Journal of Econometrics*, **118**, pp. 51–65.

Li, T., and Vuong, Q. 1998. “Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators.” *Journal of Multivariate Analysis*, **65**, 139–165.

Matzkin, R. L. 2007. “Nonparametric Identification.” Chapter 73 in *Handbook of Econometrics*, Vol. 6b, eds. J. J. Heckman and E. E. Leamer. Amsterdam: Elsevier B. V., pp. 5307–5368.

Meister, A. 2009. *Deconvolution Problems in Nonparametric Statistics*, Lecture Notes in Statistics. Berlin: Springer-Verlag.

Newey, W. 2001. “Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models.” *Review of Economics and Statistics*, **83**, 616–627.

Schennach, S. 2004. “Nonparametric Regression in the Presence of Measurement Error.” *Econometric Theory*, **20**, pp. 1046–1093.

Schennach, S. 2007. “Instrumental Variable Estimation in Nonlinear Errors-in-Variables Models.” *Econometrica*, **75**, pp. 201–239.

Schwartz, L. 1966. *Théorie des distributions.* Paris: Hermann.

Taupin, M.-L. 2001. “Semi-parametric Estimation in the Nonlinear structural Errors-in-Variables Model.” *Annals of Statistics*, **29**, pp. 66–93.

Wang, L. “Estimation of Nonlinear Models with Berkson Measurement Errors.” *Annals of Statistics*, **32**, pp. 2559–2579.

Zinde-Walsh, V. 2008. “Kernel Estimation When Density May Not Exist.” *Econometric Theory*, **24**, 696–725.

Zinde-Walsh, V. 2009. “Errors-in-Variables Models: A Generalized Functions Approach.” Working paper, arXiv:0909.5390v1 [stat.ME], McGill University.

Zinde-Walsh, V. 2011. “Presidential Address: Mathematics in Economics and Econometrics.” *Canadian Journal of Economics*, **44**, pp. 1052–1068.

Zinde-Walsh, V. 2013. “Measurement Error and Deconvolution in Spaces of Generalized Functions.” Econometric Theory, accepted for publication, Working paper, arXiv:1009.4217v2 [math.ST]; earlier version, 2010: arXiv:1009.4217v1[MATH.ST].

PART III

ADDITIVE MODELS

CHAPTER 5

NONPARAMETRIC ADDITIVE MODELS

JOEL L. HOROWITZ

5.1. Introduction

Much applied research in statistics, economics, and other fields is concerned with estimation of a conditional mean or quantile function. Specifically, let (Y, X) be a random pair, where Y is a scalar random variable and X is a d-dimensional random vector that is continuously distributed. Suppose we have data consisting of the random sample $\{Y_i, X_i : i = 1, \ldots, n\}$. Then the problem is to use the data to estimate the conditional mean function $g(x) \equiv E(Y|X = x)$ or the conditional α-quantile function $Q_\alpha(x)$. The latter is defined by $P[Y \leq Q_\alpha(x)|X = x] = \alpha$ for some α satisfying $0 < \alpha < 1$. For example, the conditional median function is obtained if $\alpha = 0.50$.

One way to proceed is to assume that g or Q_α is known up to a finite-dimensional parameter θ, thereby obtaining a parametric model of the conditional mean or quantile function. For example, if g is assumed to be linear, then $g(x) = \theta_0 + \theta_1' x$, where θ_0 is a scalar constant and θ_1 is a vector that is conformable with x. Similarly, if Q_α is assumed to be linear, then $Q_\alpha(x) = \theta_0 + \theta_1' x$. Given a finite-dimensional parametric model, the parameter θ can be estimated consistently by least squares in the case of conditional mean function and by least absolute deviations in the case of the conditional median function $Q_{0.5}$. Similar methods are available for other quantiles. However, a parametric model is usually arbitrary. For example, economic theory rarely, if ever, provides one, and a misspecified parametric model can be seriously misleading. Therefore, it is useful to seek estimation methods that do not require assuming a parametric model for g or Q_α.

Many investigators attempt to minimize the risk of specification error by carrying out a specification search. In a specification search, several different parametric models are estimated, and conclusions are based on the one that appears to fit the data best. However, there is no guarantee that a specification search will include the correct model or a good approximation to it, and there is no guarantee that the correct model will be selected if it happens to be included in the search. Therefore, the use of specification searches should be minimized.

The possibility of specification error can be essentially eliminated through the use of nonparametric estimation methods. Nonparametric methods assume that g or Q_α satisfies certain smoothness conditions, but no assumptions are made about the shape or functional form of g or Q_α. See, for example, Fan and Gijbels (1996), Härdle (1990), Pagan and Ullah (1999), Li and Racine (2007), and Horowitz (2009), among many other references. However, the precision of a nonparametric estimator decreases rapidly as the dimension of X increases. This is called the curse of dimensionality. As a consequence of it, impracticably large samples are usually needed to obtain useful estimation precision if X is multidimensional.

The curse of dimensionality can be avoided through the use of dimension-reduction techniques. These reduce the effective dimension of the estimation problem by making assumptions about the form of g or Q_α that are stronger than those made by fully nonparametric estimation but weaker than those made in parametric modeling. Single-index and partially linear models (Härdle, Liang, and Gao, 2000, Horowitz, 2009) and nonparametric additive models, the subject of this chapter, are examples of ways of doing this. These models achieve greater estimation precision than do fully nonparametric models, and they reduce (but do not eliminate) the risk of specification error relative to parametric models.

In a nonparametric additive model, g or Q_α is assumed to have the form

$$\left.\begin{array}{l} g(x) \\ \text{or} \\ Q_\alpha(x) \end{array}\right\} = \mu + f_1(x^1) + f_2(x^2) + \cdots + f_d(x^d), \tag{5.1}$$

where μ is a constant, x^j $(j = 1,\ldots,d)$ is the jth component of the d-dimensional vector x, and $f_1,\ldots,f_d$ are functions that are assumed to be smooth but are otherwise unknown and are estimated nonparametrically. Model (5.1) can be extended to

$$\left.\begin{array}{l} g(x) \\ \text{or} \\ Q_\alpha(x) \end{array}\right\} = F[\mu + f_1(x^1) + f_2(x^2) + \cdots + f_d(x^d)], \tag{5.2}$$

where F is a strictly increasing function that may be known or unknown.

It turns out that under mild smoothness conditions, the additive components $f_1,\ldots,f_d$ can be estimated with the same precision that would be possible if X were a scalar. Indeed, each additive component can be estimated as well as it could be if all the other additive components were known. This chapter reviews methods for achieving these results. Section 5.2 describes methods for estimating model (5.1). Methods for estimating model (5.2) with a known or unknown link function F are described in Section 5.3. Section 5.4 discusses tests of additivity. Section 5.5 presents an empirical example that illustrates the use of model (5.1), and Section 5.6 presents conclusions.

Estimation of derivatives of the functions $f_1, \ldots, f_d$ is important in some applications. Estimation of derivatives is not discussed in this chapter but is discussed by Severance-Lossin and Sperlich (1999) and Yang, Sperlich, and Härdle (2003). The discussion in this chapter is informal. Regularity conditions and proofs of results are available in the references that are cited in the chapter. The details of the methods described here are lengthy, so most methods are presented in outline form. Details are available in the cited references.

5.2. Methods for Estimating Model (5.1)

We begin with the conditional mean version of model (5.1), which can be written as

$$E(Y|X=x) = \mu + f_1(x^1) + f_2(x^2) + \cdots + f_d(x^d). \tag{5.3}$$

The conditional quantile version of (5.1) is discussed in Section 5.2.1.

Equation (5.3) remains unchanged if a constant, say γ_j, is added to f_j $(j = 1, \ldots, d)$ and μ is replaced by $\mu - \sum_{j=1}^{d} \gamma_j$. Therefore, a location normalization is needed to identify μ and the additive components. Let X^j denote the jth component of the random vector X. Depending on the method that is used to estimate the f_j's, location normalization consists of assuming that $Ef_j(X^j) = 0$ or that

$$\int f_j(v)\, dv = 0 \tag{5.4}$$

for each $j = 1, \ldots, d$.

Stone (1985) was the first to give conditions under which the additive components can be estimated with a one-dimensional nonparametric rate of convergence and to propose an estimator that achieves this rate. Stone (1985) assumed that the support of X is $[0,1]^d$, that the probability density function of X is bounded away from 0 on $[0,1]^d$, and that $\mathrm{Var}(Y|X=x)$ is bounded on $[0,1]^d$. He proposed using least squares to obtain spline estimators of the f_j's under the location normalization $Ef_j(X^j) = 0$. Let $\hat{f}_j$ denote the resulting estimator of f_j. For any function h on $[0,1]$, define

$$\|h\|^2 = \int_0^1 h(v)^2 dv.$$

Stone (1985) showed that if each f_j is p times differentiable on $[0,1]$, then $E(\|\hat{f}_j - f_j\|^2 | X^1, \ldots, X^d) = O_p[n^{-2p/(2p+1)}]$. This is the fastest possible rate of convergence. However, Stone's result does not establish pointwise convergence of $\hat{f}_j$ to f_j or the asymptotic distribution of $n^{p/(2p+1)}[\hat{f}_j(x) - f_j(x)]$.

Since the work of Stone (1985), there have been many attempts to develop estimators of the f_j's that are pointwise consistent with the optimal rate of convergence and are asymptotically normally distributed. Oracle efficiency is another desirable property of such estimators. Oracle efficiency means that the asymptotic distribution of the estimator of any additive component f_j is the same as it would be if the other components were known.

Buja, Hastie, and Tibshirani (1989) and Hastie and Tibshirani (1990) proposed an estimation method called backfitting. This method is based on the observation that

$$f_k(x^k) = E[Y - \mu - \sum_{j \neq k} f_j(x^j) | X = (x^1, \ldots, x^d)].$$

If μ and the f_j's for $j \neq k$ were known, then f_k could be estimated by applying nonparametric regression to $Y - \mu - \sum_{j \neq k} f_j(X^j)$. Backfitting replaces the unknown quantities by preliminary estimates. Then each additive component is estimated by nonparametric regression, and the preliminary estimates are updated as each additive component is estimated. In principle, this process continues until convergence is achieved. Backfitting is implemented in many statistical software packages, but theoretical investigation of the statistical properties of backfitting estimators is difficult. This is because these estimators are outcomes of an iterative process, not the solutions to optimization problems or systems of equations. Opsomer and Ruppert (1997) and Opsomer (2000) investigated the properties of a version of backfitting and found, among other things, that strong restrictions on the distribution of X are necessary to achieve results and that the estimators are not oracle efficient. Other methods described below are oracle efficient and have additional desirable properties. Compared to these estimators, backfitting is not a desirable approach, despite its intuitive appeal and availability in statistical software packages.

The first estimator of the f_j's that was proved to be pointwise consistent and asymptotically normally distributed was developed by Linton and Nielsen (1995) and extended by Linton and Härdle (1996). Tjøstheim and Auestad (1994) and Newey (1994) present similar ideas. The method is called marginal integration and is based on the observation that under the location normalization $Ef_j(X^j) = 0$, we have $\mu = E(Y)$ and

$$f_j(x^j) = \int E(Y|X = x) p_{-j}(x^{(-j)}) dx^{(-j)} - \mu, \tag{5.5}$$

where $x^{(-j)}$ is the vector consisting of all components of x except x^j and p_{-j} is the probability density function of $X^{(-j)}$. The constant μ is estimated consistently by the sample analogue

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} Y_i.$$

To estimate, say, $f_1(x^1)$, let $\hat{g}(x^1, x^{(-1)})$ be the following kernel estimator of $E(Y|X^1 = x^1, X^{(-1)} = x^{(-1)})$:

$$\hat{g}(x^1, x^{(-1)}) = \hat{P}(x^1, x^{(-1)})^{-1} \sum_{i=1}^{n} Y_i K_1 \left(\frac{x^1 - X_i^1}{h_1} \right) K_2 \left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2} \right), \tag{5.6}$$

where

$$\hat{P}(x^1, x^{(-1)}) = \sum_{i=1}^{n} K_1 \left(\frac{x^1 - X_i^1}{h_1} \right) K_2 \left(\frac{x^{(-1)} - X_i^{(-1)}}{h_2} \right), \tag{5.7}$$

K_1 is a kernel function of a scalar argument, K_2 is a kernel function of a $(d-1)$-dimensional argument, $X_i^{(-1)}$ is the ith observation of $X^{(-1)}$, and h_1 and h_2 are bandwidths. The integral on the right-hand side of (5.5) is the average of $E(Y|X^1 = x^1, X^{(-1)} = x^{(-1)})$ over $X^{(-1)}$ and can be estimated by the sample average of $\hat{g}(x^1, X^{(-1)})$. The resulting marginal integration estimator of f_1 is

$$\hat{f}_1(x^1) = n^{-1} \sum_{i=1}^{n} \hat{g}(x^1, X_i^{(-1)}) - \hat{\mu}.$$

Linton and Härdle (1996) give conditions under which $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)] \to^d N[\beta_{1,MI}(x^1), V_{1,MI}(x^1)]$ for suitable functions $\beta_{1,MI}$ and $V_{1,MI}$. Similar results hold for the marginal integration estimators of the other additive components. The most important condition is that each additive component is at least d times continuously differentiable. This condition implies that the marginal integration estimator has a form of the curse of dimensionality, because maintaining an $n^{-2/5}$ rate of convergence in probability requires the smoothness of the additive components to increase as d increases. In addition, the marginal integration estimator is not oracle efficient and can be hard to compute.

There have been several refinements of the marginal integration estimator that attempt to overcome these difficulties. See, for example, Linton (1997), Kim, Linton, and Hengartner (1999), and Hengartner and Sperlich (2005). Some of these refinements overcome the curse of dimensionality, and others achieve oracle efficiency. However, none of the refinements is both free of the curse of dimensionality and oracle efficient.

The marginal integration estimator has a curse of dimensionality because, as can be seen from (5.6) and (5.7), it requires full-dimensional nonparametric estimation of $E(Y|X = x)$ and the probability density function of X. The curse of dimensionality can be avoided by imposing additivity at the outset of estimation, thereby avoiding the need for full-dimensional nonparametric estimation. This cannot be done with kernel-based estimators, such as those used in marginal integration, but it can be done easily with series estimators. However, it is hard to establish the asymptotic distributional properties of series estimators. Horowitz and Mammen (2004) proposed a two-step estimation procedure that overcomes this problem. The first step of the procedure is series estimation of the f_j's. This is followed by a backfitting step that turns the series estimates into kernel estimates that are both oracle efficient and free of the curse of dimensionality.

Horowitz and Mammen (2004) use the location normalization (5.4) and assume that the support of X is $[-1,1]^d$. Let $\{\psi_k\colon k = 1,2,\ldots\}$ be an orthonormal basis for smooth functions on $[-1, 1]$ that satisfies (5.4). The first step of the Horowitz–Mammen (2004) procedure consists of using least squares to estimate μ and the generalized Fourier coefficients $\{\theta_{jk}\}$ in the series approximation

$$E(Y|X=x) \approx \mu + \sum_{j=1}^{d}\sum_{k=1}^{\kappa}\theta_{jk}\psi_k(x^j), \tag{5.8}$$

where κ is the length of the series approximations to the additive components. In this approximation, f_j is approximated by

$$f_j(x^j) \approx \sum_{k=1}^{\kappa}\theta_{jk}\psi_k(x^j).$$

Thus, the estimators of μ and the θ_{jk}'s are given by

$$\{\tilde{\mu},\tilde{\theta}_{jk}\colon j=1,\ldots,d;\ k=1,\ldots,\kappa\} = \arg\min_{\mu,\theta_{jk}}\sum_{i=1}^{n}\left[Y_i - \mu - \sum_{j=1}^{d}\sum_{k=1}^{\kappa}\theta_{jk}\psi_k(X_i^j)\right]^2,$$

where X_i^j is the jth component of the vector X_i. Let $\tilde{f}_j$ denote the resulting estimator of μ and f_j $(j=1,\ldots,d)$. That is,

$$\tilde{f}_j(x^j) = \sum_{k=1}^{\kappa}\tilde{\theta}_{jk}\psi_k(x^j).$$

Now let K and h, respectively, denote a kernel function and a bandwidth. The second-step estimator of, say, f_1 is

$$\hat{f}_1(x^1) = \left[\sum_{i=1}^{n}K\left(\frac{x^1 - X_i^1}{h}\right)\right]^{-1}\sum_{i=1}^{n}[Y_i - \tilde{f}_{-1}(X_i^{(-1)})]K\left(\frac{x^1 - X_i^1}{h}\right), \tag{5.9}$$

where $X_i^{(-1)}$ is the vector consisting of the ith observations of all components of X except the first and $\tilde{f}_{-1} = \tilde{f}_2 + \cdots + \tilde{f}_d$. In other words, $\hat{f}_1$ is the kernel nonparametric regression of $Y - \tilde{f}_{-1}(X^{(-1)})$ on X^1. Horowitz and Mammen (2004) give conditions under which $n^{2/5}[\hat{f}_1(x^1) - f_1(x^1)] \to^d N[\beta_{1,HM}(x^1), V_{1,HM}(x^1)]$ for suitable functions $\beta_{1,HM}$ and $V_{1,HM}$. Horowitz and Mammen (2004) also show that the second-step estimator is free of the curse of dimensionality and is oracle efficient. Freedom from the curse of dimensionality means that the f_j's need to have only two continuous derivatives, regardless of d. Oracle efficiency means that the asymptotic distribution of

$n^{2/5}[\hat{f}_1(x^1)-f_1(x^1)]$ is the same as it would be if the estimator $\tilde{f}_{-1}$ in (5.9) were replaced with the true (but unknown) sum of additive components, f_{-1}. Similar results apply to the second-step estimators of the other additive components. Thus, asymptotically, each additive component f_j can be estimated as well as it could be if the other components were known. Intuitively, the method works because the bias due to truncating the series approximations to the f_j's in the first estimation step can be made negligibly small by making κ increase at a sufficiently rapid rate as n increases. This increases the variance of the $\tilde{f}_j$'s, but the variance is reduced in the second estimation step because this step includes averaging over the $\tilde{f}_j$'s. Averaging reduces the variance enough to enable the second-step estimates to have an $n^{-2/5}$ rate of convergence in probability.

There is also a local linear version of the second step estimator. For estimating f_1, this consists of choosing b_0 and b_1 to minimize

$$S_n(b_0,b_1)=(nh)^{-1}\sum_{i=1}^{n}[Y_i-\tilde{\mu}-b_0-b_1(X_i^1-x^1)-\tilde{f}_{-1}(X_i^{(-1)})]^2K\left(\frac{X_i^1-x^1}{h}\right).$$

Let $(\hat{b}_0,\hat{b}_1)$ denote the resulting value of (b_0,b_1). The local linear second-step estimator of $f_1(x^1)$ is $\hat{f}_1(x^1)=\hat{b}_0$. The local linear estimator is pointwise consistent, asymptotically normal, oracle efficient, and free of the curse of dimensionality. However, the mean and variance of the asymptotic distribution of the local linear estimator are different from those of the Nadaraya–Watson (or local constant) estimator (5.9). Fan and Gijbels (1996) discuss the relative merits of local linear and Nadaraya–Watson estimators.

Mammen, Linton, and Nielsen (1999) developed an asymptotically normal, oracle-efficient estimation procedure for model (5.1) that consists of solving a certain set of integral equations. Wang and Yang (2007) generalized the two-step method of Horowitz and Mammen (2004) to autoregressive time-series models. Their model is

$$Y_t=\mu+f_1(X_t^1)+\cdots+f_d(X_t^d)+\sigma(X_t^1,\ldots,X_t^d)\varepsilon_t; \qquad t=1,2,\ldots,$$

where X_t^j is the jth component of the d-vector X_t, $E(\varepsilon_t|X_t)=0$, and $E(\varepsilon_t^2|X_t)=1$. The explanatory variables $\{X_t^j\colon j=1,\ldots,d\}$ may include lagged values of the dependent variable Y_t. The random vector (X_t,ε_t) is required to satisfy a strong mixing condition, and the additive components have two derivatives. Wang and Yang (2007) propose an estimator that is like that of Horowitz and Mammen (2004), except the first step uses a spline basis that is not necessarily orthogonal. Wang and Yang (2007) show that their estimator of each additive component is pointwise asymptotically normal with an $n^{-2/5}$ rate of convergence in probability. Thus, the estimator is free of the curse of dimensionality. It is also oracle efficient. Nielsen and Sperlich (2005) and Wang and Yang (2007) discuss computation of some of the foregoing estimators.

Song and Yang (2010) describe a different two-step procedure for obtaining oracle efficient estimators with time-series data. Like Wang and Yang (2007), Song and Yang (2010) consider a nonparametric, additive, autoregressive model in which the covariates and random noise component satisfy a strong mixing condition.

The first estimation step consists of using least squares to make a constant-spline approximation to the additive components. The second step is like that of Horowitz and Mammen (2004) and Wang and Yang (2007), except a linear spline estimator replaces the kernel estimator of those papers. Most importantly, Song and Yang (2010) obtain asymptotic uniform confidence bands for the additive components. They also report that their two-stage spline estimator can be computed much more rapidly than procedures that use kernel-based estimation in the second step. Horowitz and Mammen (2004) and Wang and Yang (2007) obtained pointwise asymptotic normality for their estimators but did not obtain uniform confidence bands for the additive components. However, the estimators of Horowitz and Mammen (2004) and Wang and Yang (2007) are, essentially, kernel estimators. Therefore, these estimators are multivariate normally distributed over a grid of points that are sufficiently far apart. It is likely that uniform confidence bands based on the kernel-type estimators can be obtained by taking advantage of this multivariate normality and letting the spacing of the grid points decrease slowly as n increases.

5.2.1. Estimating a Conditional Quantile Function

This section describes estimation of the conditional quantile version of (5.1). The discussion concentrates on estimation of the conditional median function, but the methods and results also apply to other quantiles. Model (5.1) for the conditional median function can be estimated using series methods or backfitting, but the rates of convergence and other asymptotic distributional properties of these estimators are unknown. De Gooijer and Zerom (2003) proposed a marginal integration estimator. Like the marginal integration estimator for a conditional mean function, the marginal integration estimator for a conditional median or other conditional quantile function is asymptotically normally distributed but suffers from the curse of dimensionality.

Horowitz and Lee (2005) proposed a two-step estimation procedure that is similar to that of Horowitz and Mammen (2004) for conditional mean functions. The two-step method is oracle efficient and has no curse of dimensionality. The first step of the method of Horowitz and Lee (2005) consists of using least absolute deviations (LAD) to estimate μ and the θ_{jk}'s in the series approximation (5.8). That is,

$$\{\tilde{\mu}, \tilde{\theta}_{jk}: j=1,\ldots,d;\ k=1,\ldots,\kappa\} = \arg\min_{\mu,\theta_{jk}} \sum_{i=1}^{n} \left| Y_i - \mu - \sum_{j=1}^{d}\sum_{k=1}^{\kappa} \theta_{jk}\psi_k(X_i^j) \right|,$$

As before, $\tilde{f}_j$ denote the first-step estimator of f_j. The second step of the method of Horowitz and Lee (2005) is of a form local-linear LAD estimation that is analogous to the second step of the method of Horowitz and Mammen (2004). For estimating f_1, this step consists of choosing b_0 and b_1 to minimize

$$S_n(b_0,b_1)=(nh)^{-1}\sum_{i=1}^{n}|Y_i-\tilde{\mu}-b_0-b_1(X_i^1-x^1)-\tilde{f}_{-1}(X_i^{(-1)}|K\left(\frac{X_i^1-x^1}{h}\right),$$

where h is a bandwidth, K is a kernel function, and $\tilde{f}_{-1}=\tilde{f}_2+\cdots+f_d$. Let $(\hat{b}_0,\hat{b}_1)$ denote resulting value of (b_0,b_1). The estimator of $f_1(x^1)$ is $\hat{f}_1(x^1)=\hat{b}_0$. Thus, the second-step estimator of any additive component is a local-linear conditional median estimator. Horowitz and Lee (2005) give conditions under which $n^{2/5}[\hat{f}_1(x^1)-f_1(x^1)]\to^d N[\beta_{1,HL}(x^1),V_{1,HL}(x^1)]$ for suitable functions $\beta_{1,HL}$ and $V_{1,HL}$. Horowitz and Lee (2005) also show that $\hat{f}_1$ is free of the curse of dimensionality and is oracle efficient. Similar results apply to the estimators of the other f_j's.

5.3. Methods for Estimating Model (5.2)

This section describes methods for estimating model (5.2) when the link function F is not the identity function. Among other applications, this permits extension of methods for nonparametric additive modeling to settings in which Y is binary. For example, an additive binary probit model is obtained by setting

$$P(Y=1|X=x)=\Phi[\mu+f_1(x^1)+\cdots+f_d(X^d)], \tag{5.10}$$

where Φ is the standard normal distribution function. In this case, the link function is $F=\Phi$. A binary logit model is obtained by replacing Φ in (5.10) with the logistic distribution function.

Section 5.3.1 treats the case in which F is known. Section 5.3.2 treats bandwidth selection for one of the methods discussed in Section 5.3.1. Section 5.3.3 discusses estimation when F is unknown.

5.3.1. Estimation with a Known Link Function

In this section, it is assumed that the link function F is known. A necessary condition for point identification of μ and the f_j's is that F is strictly monotonic. Given this requirement, it can be assumed without loss of generality that F is strictly increasing. Consequently, $F^{-1}[Q_\alpha(x)]$ is the α conditional quantile of $F^{-1}(Y)$ and has a nonparametric additive form. Therefore, quantile estimation of the additive components of model (5.2) can be carried out by applying the methods of Section 5.2.1 to $F^{-1}(Y)$. Accordingly, the remainder of this section is concerned with estimating the conditional mean version of model (5.2).

Linton and Härdle (1996) describe a marginal integration estimator of the additive components in model (5.2). As in the case of model (5.1), the marginal integration estimator has a curse of dimensionality and is not oracle efficient. The two-step method of Horowitz and Mammen (2004) is also applicable to model (5.2). When F has a Lipschitz continuous second derivative and the additive components are twice continuously differentiable, it yields asymptotically normal, oracle efficient estimators of the additive components. The estimators have an $n^{-2/5}$ rate of convergence in probability and no curse of dimensionality.

The first step of the method of Horowitz and Mammen (2004) is nonlinear least squares estimation of truncated series approximations to the additive components. That is, the generalized Fourier coefficients of the approximations are estimated by solving

$$\begin{aligned}&\{\tilde{\mu}, \tilde{\theta}_{jk} : j = 1, \ldots, d;\ k = 1, \ldots, \kappa\}\\ &\quad = \arg\min_{\mu, \theta_{jk}} \sum_{i=1}^{n} \left\{ Y_i - F\left[\mu + \sum_{j=1}^{d} \sum_{k=1}^{\kappa} \theta_{jk} \psi_k(x^j) \right] \right\}^2 .\end{aligned}$$

Now set

$$\tilde{f}_j(x^j) = \sum_{k=1}^{\kappa} \tilde{\theta}_{jk} \psi_k(x^j).$$

A second-step estimator of $f_1(x^1)$, say, can be obtained by setting

$$\tilde{\tilde{f}}_1(x^1) = \arg\min_b \sum_{i=1}^{n} \left\{ Y_i - F\left[\tilde{\mu} + b + \sum_{j=2}^{d} \tilde{f}_j(X_i^j) \right] \right\}^2 K\left(\frac{x^1 - X_i^1}{h} \right),$$

where, as before, K is a kernel function and h is a bandwidth. However, this requires solving a difficult nonlinear optimization problem. An asymptotically equivalent estimator can be obtained by taking one Newton step from $b_0 = \tilde{f}_1(x^1)$ toward $\tilde{\tilde{f}}_1(x^1)$. To do this, define

$$\begin{aligned}S'_{n1}(x^1, f) = {} & -2 \sum_{i=1}^{n} \left\{ Y_i - F[\mu + f_1(x^1) + f_2(X_i^2) + \cdots + f_d(X_i^d)] \right\}\\ & \times F'[\mu + f_1(x^1) + f_2(X_i^2) + \cdots + f_d(X_i^d)] K\left(\frac{x^1 - X_i^1}{h} \right)\end{aligned}$$

and

$$\begin{aligned}S''_{n1}(x^1, f) = {} & 2 \sum_{i=1}^{n} F'[\mu + f_1(x^1) + f_2(X_i^2) + \cdots + f_d(X_i^d)]^2 K\left(\frac{x^1 - X_i^1}{h} \right)\\ & -2 \sum_{i=1}^{n} \{ Y_i - F[\mu + f_1(x^1) + f_2(X_i^2) + \cdots + f_d(X_i^d)] \}\\ & \times F''[\mu + f_1(x^1) + f_2(X_i^2) + \cdots + f_d(X_i^d)] K\left(\frac{x^1 - X_i^1}{h} \right).\end{aligned}$$

The second-step estimator is

$$\hat{f}_1(x^1) = \tilde{f}_1(x^1) - S'_{n1}(x^1\tilde{f})/S''_{n1}(x^1,\tilde{f}).$$

Horowitz and Mammen (2004) also describe a local-linear version of this estimator.

Liu, Yang, and Härdle (2011) describe a two-step estimation method for model (5.2) that is analogous to the method of Wang and Yang (2007) but uses a local pseudo-log-likelihood objective function based on the exponential family at each estimation stage instead of a local least squares objective function. As in Wang and Yang (2007), the method of Liu, Yang, and Härdle (2011) applies to an autoregressive model in which the covariates and random noise satisfy a strong mixing condition. Yu, Park, and Mammen (2008) proposed an estimation method for model (5.2) that is based on numerically solving a system of nonlinear integral equations. The method is more complicated than that of Horowitz and Mammen (2004), but the results of Monte Carlo experiments suggest that the estimator of Yu, Park, and Mammen (2008) has better finite-sample properties than that of Horowitz and Mammen (2004), especially when the covariates are highly correlated.

5.3.2. Bandwidth Selection for the Two-Step Estimator of Horowitz and Mammen (2004)

This section describes a penalized least squares (PLS) method for choosing the bandwidth h in the second step of the procedure of Horowitz and Mammen (2004). The method is described here for the local-linear version of the method, but similar results apply to the local constant version. The method described in this section can be used with model (5.1) by setting F equal to the identity function.

The PLS method simultaneously estimates the bandwidths for second-step estimation of all the additive components f_j $(j = 1,\ldots,d)$. Let $h_j = C_j n^{-1/5}$ be the bandwidth for $\hat{f}_j$. The PLS method selects the C_j's that minimize an estimate of the average squared error (ASE):

$$ASE(\bar{h}) = n^{-1}\sum_{i=1}^{n}\{F[\tilde{\mu}+\hat{f}(X_i)] - F[\mu + f(X_i)]\}^2,$$

where $\hat{f} = \hat{f}_1 + \cdots + \hat{f}_d$ and $\bar{h} = (C_1 n^{-1/5},\ldots,C_d n^{-1/5})$. Specifically, the PLS method selects the C_j's to

$$\begin{aligned}\text{minimize}_{C_1,\ldots,C_d}: \ PLS(\bar{h}) = n^{-1}\sum_{i=1}^{n}[Y_i - F[\tilde{\mu}+\hat{f}(X_i)]^2 + 2K(0)n^{-1} \\ \times \sum_{i=1}^{n}\{F'[\tilde{\mu}+\hat{f}(X_i)]^2\hat{V}(X_i)\}\sum_{j=1}^{d}[n^{4/5}C_j\hat{D}_j(X_i^j)]^{-1},\end{aligned} \tag{5.11}$$

where the C_j's are restricted to a compact, positive interval that excludes 0,

$$D_j(x^j) = (nh_j)^{-1} \sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) F'[\tilde{\mu} + \hat{f}(X_i)]^2$$

and

$$\begin{aligned}\hat{V}(x) = &\left[\sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right) \ldots K\left(\frac{X_i^d - x^d}{h_d}\right)\right]^{-1} \\ &\times \sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right) \ldots K\left(\frac{X_i^d - x^d}{h_d}\right) \{Y_i - F[\tilde{\mu} + \hat{f}(X_i)]^2.\end{aligned}$$

The bandwidths for $\hat{V}$ may be different from those used for $\hat{f}$, because $\hat{V}$ is a full-dimensional nonparametric estimator. Horowitz and Mammen (2004) present arguments showing that the solution to (5.11) estimates the bandwidths that minimize ASE.

5.3.3. Estimation with an Unknown Link Function

This section is concerned with estimating model (5.2) when the link function F is unknown. When F is unknown, model (5.2) contains semiparametric single-index models as a special case. This is important, because semiparametric single-index models and nonparametric additive models with known link functions are non-nested. In a semiparametric single-index model, $E(Y|X = x) = G(\theta' x)$ for some unknown function G and parameter vector θ. This model coincides with the nonparametric additive model with link function F only if the additive components are linear and $F = G$. An applied researcher must choose between the two models and may obtain highly misleading results if an incorrect choice is made. A nonparametric additive model with an unknown link function makes this choice unnecessary, because the model nests semiparametric single index models and nonparametric additive models with known link functions. A nonparametric additive model with an unknown link function also nests the multiplicative specification

$$E(Y|X = x) = F[f_1(x^1) f_2(x^2) \ldots f_d(x^d)].$$

A further attraction of model (5.2) with an unknown link function is that it provides an informal, graphical method for checking the additive and single-index specifications. One can plot the estimates of F and the f_j's. Approximate linearity of the estimate of F favors the additive specification (5.1), whereas approximate linearity of the f_j's favors the single-index specification. Linearity of F and the f_j's favors the linear model $E(Y|X) = \theta' X$.

Identification of the f_j's in model (5.2) requires more normalizations and restrictions when F is unknown than when F is known. First, observe that μ is not identified when f is unknown, because $F[\mu + f_1(x^1) + \cdots + f_d(x^d)] = F^*[f_1(x^1) + \cdots + f_d(x^d)]$, where the function F^* is defined by $F^*(v) = F(\mu + v)$ for any real v. Therefore, we can set $\mu = 0$ without loss of generality. Similarly, a location normalization is needed because model (5.2) remains unchanged if each f_j is replaced by $f_j + \gamma_j$, where γ_j is a constant, and $F(v)$ is replaced by $F^*(v) = F(v - \gamma_1 - \cdots - \gamma_d)$. In addition, a scale normalization is needed because model (5.2) is unchanged if each f_j is replaced by cf_j for any constant $c \neq 0$ and $F(v)$ is replaced by $F^*(v) = F(v/c)$. Under the additional assumption that F is monotonic, model (5.2) with F unknown is identified if at least two additive components are not constant. To see why this assumption is necessary, suppose that only f_1 is not constant. Then conditional mean function is of the form $F[f_1(x^1) + \text{constant}]$. It is clear that this function does not identify F and f_1. The methods presented in this discussion use a slightly stronger assumption for identification. We assume that the derivatives of two additive components are bounded away from 0. The indices j and k of these components do not need to be known. It can be assumed without loss of generality that $j = d$ and $k = d - 1$.

Under the foregoing identifying assumptions, oracle-efficient, pointwise asymptotically normal estimators of the f_j's can be obtained by replacing F in the procedure of Horowitz and Mammen (2004) for model (5.2) with a kernel estimator. As in the case of model (5.2) with F known, estimation takes place in two steps. In the first step, a modified version of Ichimura's (1993) estimator for a semiparametric single-index model is used to obtain a series approximation to each f_j and a kernel estimator of F. The first-step procedure imposes the additive structure of model (5.2), thereby avoiding the curse of dimensionality. The first-step estimates are inputs to the second step. The second-step estimator of, say, f_1 is obtained by taking one Newton step from the first-step estimate toward a local nonlinear least squares estimate. In large samples, the second-step estimator has a structure similar to that of a kernel nonparametric regression estimator, so deriving its pointwise rate of convergence and asymptotic distribution is relatively easy. The details of the two-step procedure are lengthy. They are presented in Horowitz and Mammen (2011). The oracle-efficiency property of the two-step estimator implies that asymptotically, there is no penalty for not knowing F in a nonparametric additive model. Each f_j can be estimated as well as it would be if F and the other f_j's were known.

Horowitz and Mammen (2007) present a penalized least squares (PLS) estimation procedure that applies to model (5.2) with an unknown F and also applies to a larger class of models that includes quantile regressions and neural networks. The procedure uses the location and scale normalizations $\mu = 0$, (5.4), and

$$\sum_{j=1}^{d} \int f_j^2(v)\, dv = 1. \tag{5.12}$$

The PLS estimator of Horowitz and Mammen (2007) chooses the estimators of F and the additive components to solve

$$\begin{aligned}
&\text{minimize}_{\breve{F},\breve{f}_1,\ldots,\breve{f}_d}: && \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \breve{F}[\breve{f}_1(X_i^1)+\cdots+\breve{f}_d(X_i^d)]\} + \lambda_n^2 J(\breve{F},\breve{f}_1,\ldots,\breve{f}_d) \\
&\text{subject to:} && (5.4) \text{ and } (5.12), \qquad (5.13)
\end{aligned}$$

where $\{\lambda_n\}$ is a sequence of constants and J is a penalty term that penalizes roughness of the estimated functions. If F and the f_j's are k times differentiable, the penalty term is

$$J(\breve{F},\breve{f}_1,\ldots,\breve{f}_d) = J_1^{\nu_1}(\breve{F},\breve{f}_1,\ldots,\breve{f}_d) + J_2^{\nu_2}(\breve{F},\breve{f}_1,\ldots,\breve{f}_d),$$

where ν_1 and ν_2 are constants satisfying $\nu_2 \geq \nu_1 > 0$,

$$J_1(\breve{F},\breve{f}_1,\ldots,\breve{f}_d) = T_k(\breve{F})\left\{\sum_{j=1}^{d}[T_1^2(\breve{f}_j)+T_k^2(\breve{f}_j)]\right\}^{(2k-1)/4},$$

$$J_2(\breve{F},\breve{f}_1,\ldots,\breve{f}_d) = T_1(\breve{F})\left\{\sum_{j=1}^{d}[T_1^2(\breve{f}_j)+T_k^2(\breve{f}_j)]\right\}^{1/4},$$

and

$$T_\ell^2(f) = \int f^{(\ell)}(v)^2 dv$$

for $0 \leq \ell \leq k$ and any function f whose ℓth derivative is square integrable. The PLS estimator can be computed by approximating $\breve{F}$ and the $\breve{f}_j$'s by B-splines and minimizing (5.13) over the coefficients of the spline approximation. Denote the estimator by $\hat{F},\hat{f}_1,\ldots,\hat{f}_d$. Assume without loss of generality that the X is supported on $[0,1]^d$. Horowitz and Mammen (2007) give conditions under which the following result holds:

$$\int_0^1 [\hat{f}_j(v) - f_j(v)]^2 dv = O_p(n^{-2k/(2k+1)})$$

for each $j = 1,\ldots,d$ and

$$\int\left\{\hat{F}\left[\sum_{j=1}^{d}\hat{f}_j(x^j)\right] - F\left[\sum_{j=1}^{d} f_j(x^j)\right]\right\}^2 dx^1\ldots\, dx^d = O_p(n^{-2k/(2k+1)}).$$

In other words, the integrated squared errors of the PLS estimates of the link function and additive components converge in probability to 0 at the fastest possible rate under the assumptions. There is no curse of dimensionality. The available results do not provide an asymptotic distribution for the PLS estimator. Therefore, it is not yet possible to carry out statistical inference with this estimator.

5.4. Tests of Additivity

Models (5.1) and (5.2) are misspecified and can give misleading results if the conditional mean or quantile of Y is not additive. Therefore, it is useful to be able to test additivity. Several tests of additivity have been proposed for models of conditional mean functions. These tests undoubtedly can be modified for use with conditional quantile functions, but this modification has not yet been carried out. Accordingly, the remainder of this section is concerned with testing additivity in the conditional mean versions of models (5.1) and (5.2). Bearing in mind that model (5.1) can be obtained from model (5.2) by letting F be the identity function, the null hypothesis to be tested is

$$H_0: \ E(Y|X=x) = F[\mu + f_1(x^1) + \cdots + f_d(x^d)].$$

The alternative hypothesis is

$$H_1: \ E(Y|X=x) = F[\mu + f(x)],$$

where there are no functions $f_1, \ldots, f_d$ such that

$$P[f(X) = f_1(X^1) + \cdots + f_d(X^d)] = 1.$$

Gozalo and Linton (2001) have proposed a general class of tests. Their tests are applicable regardless of whether F is the identity function. Wang and Carriere (2011) and Dette and von Lieres und Wilkau (2001) proposed similar tests for the case of an identity link function. These tests are based on comparing fully a fully nonparametric estimator of f with an estimator that imposes additivity. Eubank, Hart, Simpson, and Stefanski (1995) also proposed tests for the case in which F is the identity function. These tests look for interactions among the components of X and are based on Tukey's (1949) test for additivity in analysis of variance. Sperlich, Tjøstheim, and Yang (2002) also proposed a test for the presence of interactions among components of X. Other tests have been proposed by Abramovich, De Fesis, and Sapatinas (2009) and Derbort, Dette, and Munk (2002).

The remainder of this section outlines a test that Gozalo and Linton (2001) found through Monte Carlo simulation to have satisfactory finite sample performance. The test statistic has the form

$$\hat{\tau}_n = \sum_{i=1}^{n} \{F^{-1}[\hat{f}(X_i)] - [\hat{\mu} + \hat{f}_1(X_i^1) + \cdots + f_d(X_i^d)]\}^2 \pi(X_i),$$

where $\hat{f}(x)$ is a full-dimensional nonparametric estimator of $E(Y|X=x)$, $\hat{\mu}$ and the $\hat{f}_j$'s are estimators of μ and f_j under H_0, and π is a weight function. Gozalo and Linton (2001) use a Nadaraya–Watson kernel estimator for $\hat{f}$ and a marginal integration estimator for $\hat{\mu}$ and the $\hat{f}_j$'s. Dette and von Lieres und Wilkau (2001) also use these marginal integration estimators in their version of the test. However, other estimators can be used. Doing so might increase the power of the test or enable some of the regularity conditions of Gozalo and Linton (2001) to be relaxed. In addition, it is clear that $\hat{\tau}_n$ can be applied to conditional quantile models, though the details of the statistic's asymptotic distribution would be different from those with conditional mean models. If F is unknown, then $F^{-1}[f(x)]$ is not identified, but a test of additivity can be based on the following modified version of $\hat{\tau}_n$:

$$\hat{\tau}_n = \sum_{i=1}^{n} \{\hat{f}(X_i) - \hat{F}[\hat{\mu} + \hat{f}_1(X_i^1) + \cdots + f_d(X_i^d)]\}^2 \pi(X_i),$$

where $\hat{f}$ is a full-dimensional nonparametric estimator of the conditional mean function, $\hat{F}$ is a nonparametric estimator of F, and the $\hat{f}_j$'s are estimators of the additive components.

Gozalo and Linton (2001) give conditions under which a centered, scaled version of $\hat{\tau}_n$ is asymptotically normally distributed as $N(0,1)$. Dette and von Lieres und Wilkau (2001) provide similar results for the case in which F is the identity function. Gozalo and Linton (2001) and Dette and von Lieres und Wilkau (2001) also provide formulae for estimating the centering and scaling parameters. Simulation results reported by Gozalo and Linton (2001) indicate that using the wild bootstrap to find critical values produces smaller errors in rejection probabilities under H_0 than using critical values based on the asymptotic normal distribution. Dette and von Lieres und Wilkau (2001) also used the wild bootstrap to estimate critical values.

5.5. An Empirical Application

This section illustrates the application of the estimator of Horowitz and Mammen (2004) by using it to estimate a model of the rate of growth of gross domestic product (GDP) among countries. The model is

$$G = f_T(T) + f_S(S) + U,$$

where G is the average annual percentage rate of growth of a country's GDP from 1960 to 1965, T is the average share of trade in the country's economy from 1960 to 1965 measured as exports plus imports divided by GDP, and S is the average number of years of schooling of adult residents of the country in 1960. U is an unobserved random variable satisfying $E(U|T,S) = 0$. The functions f_T and f_S are unknown and are estimated by the method of Horowitz and Mammen (2004). The data are taken from the data set **Growth** in Stock and Watson (2011). They comprise values of G, T, and S for 60 countries.

Estimation was carried out using a cubic B-spline basis in the first step. The second step consisted of Nadaraya–Watson (local constant) kernel estimation with the

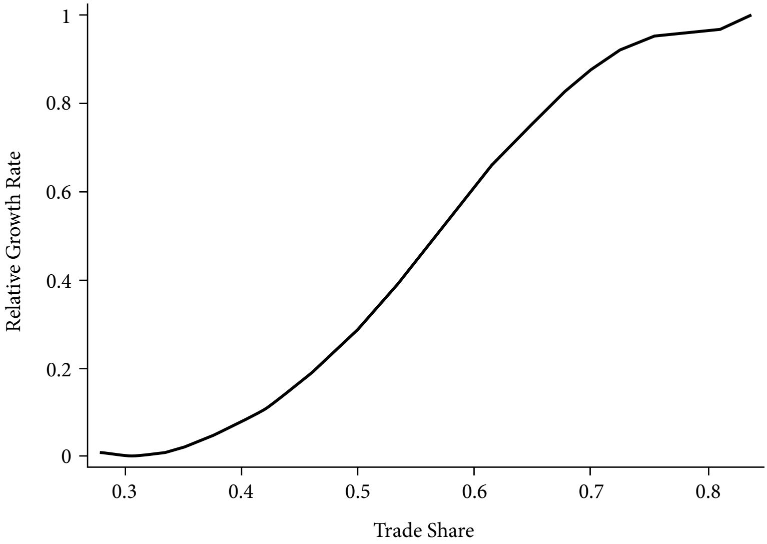


FIGURE 5.1 Additive component f_T in the growth model.

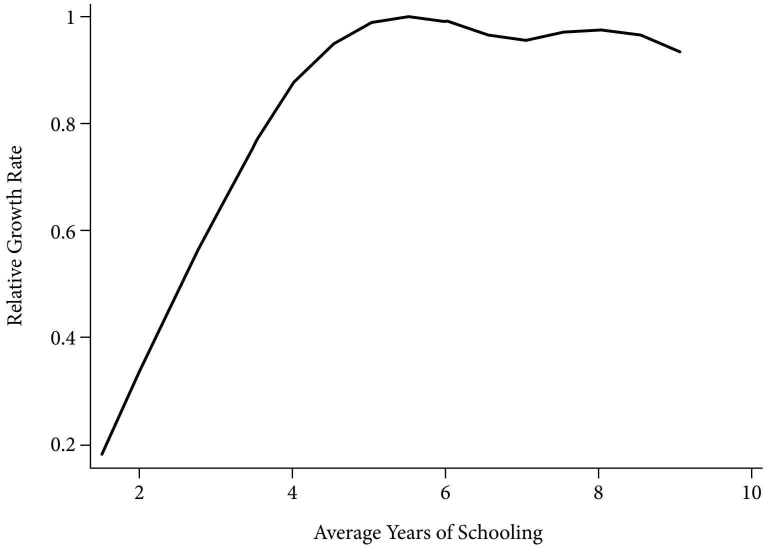


FIGURE 5.2 Additive component f_S in the growth model.

biweight kernel. Bandwidths of 0.5 and 0.8 were used for estimating f_T and f_S, respectively.

The estimation results are shown in Figures 5.1 and 5.2. The estimates of f_T and f_S are nonlinear and differently shaped. The dip in f_S near $S = 7$ is almost certainly an artifact of random sampling errors. The estimated additive components are not well-approximated by simple parametric functions such as quadratic or cubic functions. A lengthy specification search might be needed to find a parametric model that produces shapes like those in Figures 5.1 and 5.2. If such a search were successful, the resulting parametric models might provide useful compact representations of f_T and f_S but could not be used for valid inference.

5.6. Conclusions

Nonparametric additive modeling with a link function that may or may not be known is an attractive way to achieve dimension reduction in nonparametric models. It greatly eases the restrictions of parametric modeling without suffering from the lack of precision that the curse of dimensionality imposes on fully nonparametric modeling. This chapter has reviewed a variety of methods for estimating nonparametric additive models. An empirical example has illustrated the usefulness of the nonparametric additive approach. Several issues about the approach remain unresolved. One of these is to find ways to carry out inference about additive components based on the estimation method of Horowitz and Mammen (2007) that is described in Section 5.3.3. This is the most general and flexible method that has been developed to date. Another issue is the extension of the tests of additivity described in Section 5.5 to estimators other than partial integration and models of conditional quantiles. Finally, finding data-based methods for choosing tuning parameters for the various estimation and testing procedures remains an open issue.

References

Abramovich, F., I. De Fesis, and T. Sapatinas. 2009. "Optimal Testing for Additivity in Multiple Nonparametric Regression." *Annals of the Institute of Statistical Mathematics*, **61**, pp. 691–714.

Buja, A., T. Hastie, and R. Tibshirani. 1989. "Linear Smoothers and Additive Models." *Annals of Statistics*, **17**, pp. 453–555.

De Gooijer, J. G., and D. Zerom. 2003. "On Additive Conditional Quantiles with High Dimensional Covariates." *Journal of the American Statistical Association*, **98**, pp. 135–146.

Dette, H., and C. von Lieres und Wilkau. 2001. "Testing Additivity by Kernel-Based Methods—What Is a Reasonable Test?" *Bernoulli*, **7**, pp. 669–697.

Derbort, S., H. Dette, and A. Munk. 2002. "A Test for Additivity in Nonparametric Regression." *Annals of the Institute of Statistical Mathematics*, **54**, pp. 60–82.

Eubank, R. L., J. D. Hart, D. G. Simpson, and L. A. Stefanski. 1995. "Testing for Additivity in Nonparametric Regression." *Annals of Statistics*, **23**, pp. 1896–1920.

Fan, J., and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

Gozalo, P. L., and O. B. Linton. 2001. "Testing Additivity in Generalized Nonparametric Regression Models with Estimated Parameters." *Journal of Econometrics*, **104**, pp. 1–48.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge, UK: Cambridge University Press.

Härdle, W., H. Liang, and J. Gao. 2000. *Partially Linear Models.* New York: Springer.

Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized Additive Models.* London: Chapman and Hall.

Hengartner, N. W., and S. Sperlich. 2005. "Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates." *Journal of Multivariate Analysis*, **95**, pp. 246–272.

Horowitz, J. L. 2009. *Semiparametric and Nonparametric Methods in Econometrics.* New York: Springer.

Horowitz, J. L., and S. Lee. 2005. "Nonparametric Estimation of an Additive Quantile Regression Model." *Journal of the American Statistical Association*, **100**, pp. 1238–1249.

Horowitz, J. L., and E. Mammen. 2004. "Nonparametric Estimation of an Additive Model with a Link Function." *Annals of Statistics*, **32**, pp. 2412–2443.

Horowitz, J. L., and E. Mammen. 2007. "Rate-Optimal Estimation for a General Class of Nonparametric Regression Models with Unknown Link Functions." *Annals of Statistics*, **35**, pp. 2589–2619.

Horowitz, J. L., and E. Mammen. 2011. "Oracle-Efficient Nonparametric Estimation of an Additive Model with an Unknown Link Function." *Econometric Theory*, **27**, pp. 582–608.

Ichimura, H. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics*, **58**, pp. 71–120.

Kim, W., O. B. Linton, and N. W. Hengartner. 1999. "A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals." *Journal of Computational and Graphical Statistics*, **8**, pp. 278–297.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics.* Princeton, NJ: Princeton University Press.

Linton, O. B. (1997). "Efficient Estimation of Additive Nonparametric Regression Models." *Biometrika*, **84**, pp. 469–473.

Linton, O. B., and W. Härdle. 1996. "Estimating Additive Regression Models with Known Links." *Biometrika*, **83**, pp. 529–540.

Linton, O. B., and J. B. Nielsen. 1995. "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration." *Biometrika*, **82**, pp. 93–100.

Liu, R., L. Yang, and W. K. Härdle. 2011. "Oracally Efficient Two-Step Estimation of Generalized Additive Model." SFB 649 discussion paper 2011–016, Humboldt-Universität zu Berlin, Germany.

Mammen, E., O. Linton, and J. Nielsen. 1999. "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions." *Annals of Statistics*, **27**, pp. 1443–1490.

Newey, W. K. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory*, **10**, pp. 233–253.

Nielsen, J. P. and S. Sperlich. 2005. "Smooth Backfitting in Practice." *Journal of the Royal Statistical Society, Series B*, **67**, pp. 43–61.

Opsomer, J. D. 2000. "Asymptotic Properties of Backfitting Estimators." *Journal of Multivariate Analysis*, **73**, pp. 166–179.

Opsomer, J. D., and D. Ruppert. 1997. “Fitting a Bivariate Additive Model by Local Polynomial Regression.” *Annals of Statistics*, **25**, pp. 186–211.

Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics*. Cambridge, UK: Cambridge University Press.

Severance-Lossin, E., and S. Sperlich. 1999. “Estimation of Derivatives for Additive Separable Models.” *Statistics*, **33**, pp. 241–265.

Song. Q., and L. Yang. 2010. “Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Models with Simultaneous Confidence Band.” *Journal of Multivariate Analysis*, **101**, pp. 2008–2025.

Sperlich, S., D. Tjøstheim, and L. Yang. 2002. “Nonparametric Estimation and Testing of Interaction in Additive Models.” *Econometric Theory*, **18**, pp. 197–251.

Stone, C. J. 1985. “Additive Regression and Other Nonparametric Models.” *Annals of Statistics*, **13**, pp. 689–705.

Stock, J. H., and M. W. Watson. 2011. *Introduction to Econometrics*, 3rd edition. Boston: Pearson/Addison Wesley.

Tjøstheim, D., and Auestad, A. 1994. “Nonparametric Identification of Nonlinear Time Series Projections.” *Journal of the American Statistical Association*, **89**, 1398–1409.

Tukey, J. 1949. “One Degree of Freedom Test for Non-Additivity.” *Biometrics*, **5**, pp. 232–242.

Wang, L., and L. Yang. 2007. “Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model.” *Annals of Statistics*, **35**, pp. 2474–2503.

Wang, X., and K. C. Carriere. 2011. “Assessing Additivity in Nonparametric Models—a Kernel-Based Method.” *Canadian Journal of Statistics*, **39**, pp. 632–655.

Yang, L., S. Sperlich, and W. Härdle. 2003. “Derivative Estimation and Testing in Generalized Additive Models.” *Journal of Statistical Planning and Inference*, **115**, pp. 521–542.

Yu, K., B. U. Park, and E. Mammen. 2008. “Smooth Backfitting in Generalized Additive Models.” *Annals of Statistics*, **36**, pp. 228–260.

CHAPTER 6

ORACALLY EFFICIENT TWO-STEP ESTIMATION FOR ADDITIVE REGRESSION

SHUJIE MA AND LIJIAN YANG

6.1. Introduction and Overview of Additive Regression

Linear regression is one of the most widely used technique for studying the relationship between a scalar variable Y and a vector of independent variables $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$. Given a data set $(Y_i, \mathbf{X}_i^{\mathrm{T}})^{\mathrm{T}}$ of n subjects or experimental units, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^{\mathrm{T}}$, a linear model has the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{6.1}$$

where ε_i is an unobserved random variable that adds noise to this relationship. Linear regression has gained its popularity because of its simplicity and easy-to-interpret nature, but it has suffered from inflexibility in modeling possible complicated relationships between Y and $\mathbf{X}$. To avoid the strong linearity assumption and capture possible nonlinear relationships, nonparametric models were proposed and have gained much attention in the last three decades. In nonparametric models, the response Y depends on the explanatory variables $\mathbf{X}$ through a nonlinear function $m(\cdot)$ such that

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{6.2}$$

The functional form of $m(\cdot)$ is not predetermined, which is estimated from the data, so that we can let the data speak for themselves. Under smoothness condition, the unknown function can be estimated nonparametrically by such methods as kernel and spline smoothing.

Nonparametric modeling imposes no specific model structure and enables one to explore the data more flexibly, but it does not perform well when the dimension of the predictor vector in the model is high. The variances of the resulting estimates tend to be unacceptably large due to the sparseness of data, which is the so-called "curse of dimensionality." To overcome these difficulties, Stone (1985a) proposed additive models. In model (6.2), the unknown function $m(\cdot)$ is replaced by sum of univariate functions, so an additive model is given as

$$Y_i = m(X_{i1},\ldots,X_{id}) + \sigma(X_{i1},\ldots,X_{id})\varepsilon_i,\ m(x_1,\ldots,x_d) = c + \sum_{\alpha=1}^{d} m_\alpha(x_\alpha), \qquad (6.3)$$

where m and σ are the mean and standard deviation of the response Y_i conditional on the predictor vector $\mathbf{X}_i$, and each ε_i is white noise conditional on $\mathbf{X}_i$. By definition of conditional mean and variance, we have

$$m(\mathbf{X}_i) = E(Y_i|\mathbf{X}_i), \sigma^2(\mathbf{X}_i) = \mathrm{var}(Y_i|\mathbf{X}_i), \qquad i = 1,\ldots,n$$

and so the error term $\varepsilon_i = \{Y_i - m(\mathbf{X}_i)\}\sigma^{-1}(\mathbf{X}_i)$ accommodates the most general form of heteroskedasticity, because we do not assume independence of ε_i and $\mathbf{X}_i$ but only $E(\varepsilon_i|\mathbf{X}_i) \equiv 0, E(\varepsilon_i^2|\mathbf{X}_i) \equiv 1$. For identifiability, it is commonly assumed that $Em_\alpha(X_{i\alpha}) \equiv 0$, $\alpha = 1,\ldots,d$. Some other restrictions can also solve the identifiability problem such as by letting $m_\alpha(0) = 0$, for $\alpha = 1,\ldots,d$. Because the unknown functions $m_\alpha(\cdot), 1 \le \alpha \le d$, are one-dimensional, the problem associated with the so-called "curse of dimensionality" is solved.

In model (6.3), each predictor $X_\alpha, 1 \le \alpha \le d$, is required to be a continuous variable. In order to incorporate discrete variables, different forms of semiparametric models have been proposed, including partially linear additive models (PLAM) given as

$$Y_i = m(\mathbf{X}_i,\mathbf{T}_i) + \sigma(\mathbf{X}_i,\mathbf{T}_i)\varepsilon_i, m(\mathbf{x},\mathbf{t}) = c_{00} + \sum\nolimits_{l=1}^{d_1} c_{0l}t_l + \sum\nolimits_{\alpha=1}^{d_2} m_\alpha(x_\alpha) \qquad (6.4)$$

in which the sequence $\{Y_i,\mathbf{X}_i^T,\mathbf{T}_i^T\}_{i=1}^n = \{Y_i, X_{i1},\ldots,X_{id_2}, T_{i1},\ldots T_{id_1}\}_{i=1}^n$. For identifiability, we let both the additive and linear components be centered, that is, $Em_\alpha(X_{i\alpha}) \equiv 0, \alpha = 1,\ldots,d_2$, $ET_{il} = 0$, $l = 1,\ldots,d_1$. Model (6.4) is more parsimonious and easier to interpret than purely additive models (6.3) by allowing a subset of predictors $(T_{il})_{l=0}^{d_1}$ to be discrete, and it is more flexible than linear models (6.1) by allowing nonlinear relationships. To allow the coefficients of the linear predictors to change with some other variable, Xue and Yang (2006a,b) and Yang et al. (2006) proposed an additive coefficient model (ACM) that allows a response variable Y to depend linearly on some regressors, with coefficients as smooth additive functions of other predictors, called tuning variables. Specifically,

$$Y_i = \sum\nolimits_{l=1}^{d_1} m_l(\mathbf{X}_i)\, T_{il}, m_l(X_i) = m_{0l} + \sum\nolimits_{\alpha=1}^{d_2} m_{\alpha l}(X_{i\alpha}), \qquad 1 \le l \le d_1, \qquad (6.5)$$

in which the predictor vector $(\mathbf{X}_i^T, \mathbf{T}_i^T)^T$ consists of the tuning variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{id_2})^T$ and linear predictors $\mathbf{T}_i = (T_{i1}, \ldots, T_{id_1})^T$.

Model (6.5)'s versatility for econometric applications is illustrated by the following example: Consider the forecasting of the U.S. GDP annual growth rate, which is modeled as the total factor productivity (TFP) growth rate plus a linear function of the capital growth rate and the labor growth rate, according to the classic Cobb–Douglas model (Cobb and Douglas, 1928). As pointed out in Li and Racine (2007, p. 302), it is unrealistic to ignore the non-neutral effect of R&D spending on the TFP growth rate and on the complementary slopes of capital and labor growth rates. Thus, a smooth coefficient model should fit the production function better than the parametric Cobb–Douglas model. Indeed, Figure 6.1 (see Liu and Yang, 2010) shows that a smooth coefficient model has much smaller rolling forecast errors than the parametric Cobb–Douglas model, based on data from 1959 to 2002. In addition, Figure 6.2 (see Liu and Yang, 2010) shows that the TFP growth rate is a function of R&D spending, not a constant.

The additive model (6.3), the PLAM (6.4) and the ACM (6.5) achieve dimension reduction through representing the multivariate function of the predictors by sum of additive univariate functions. People have been making great efforts to develop statistical tools to estimate these additive functions. In review of literature, there are four types of kernel-based estimators: the classic backfitting estimators of Hastie and Tibshirani (1990) and Opsomer and Ruppert (1997); marginal integration estimators of Fan et al. (1997), Linton and Nielsen (1995), Linton and Härdle (1996), Kim et al.

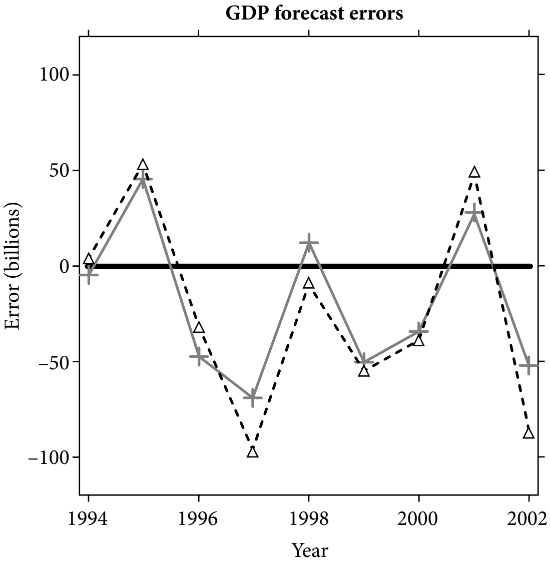


FIGURE 6.1 Errors of GDP forecasts for a smooth coefficient model (solid line); Cobb–Douglas model (dashed line).

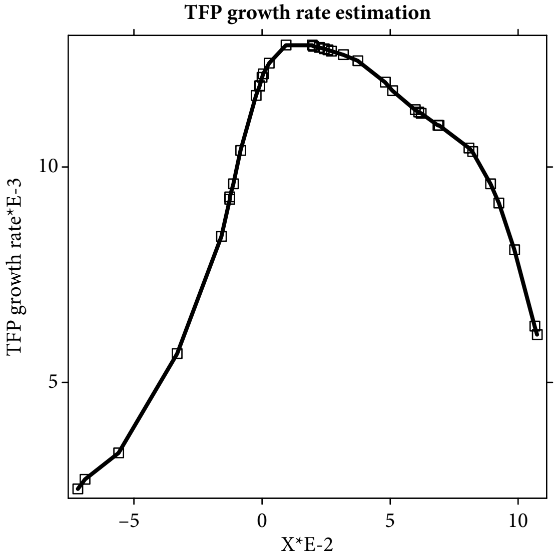


FIGURE 6.2 Estimation of TFP growth rate function.

(1999), Sperlich et al. (2002), and Yang et al. (2003) and a kernel-based method of estimating rate to optimality of Hengartiner and Sperlich (2005); the smoothing backfitting estimators of Mammen et al. (1999); and the two-stage estimators, such as one-step backfitting of the integration estimators of Linton (1997), one-step backfitting of the projection estimators of Horowitz et al. (2006) and one Newton step from the nonlinear LSE estimators of Horowitz and Mammen (2004). For the spline estimators, see Huang (1998), Stone (1985a,b), and Xue and Yang (2006b).

Satisfactory estimators of the additive functions should be (i) computationally expedient, (ii) theoretically reliable, and (iii) intuitively appealing. The kernel procedures mentioned above satisfy criterion (iii) and partly (ii) but not (i), since they are computationally intensive when sample size n is large, as illustrated in the Monte Carlo results of Xue and Yang (2006b) and Wang and Yang (2007). Kim et al. (1999) introduces a computationally efficient marginal integration estimator for the component functions in additive models, which provides a reduction in computation of order n. Spline approaches are fast to compute, thus satisfying (i), but they do not satisfy criterion (ii) because they lack limiting distribution. By combining the best features of both kernel and spline methods, Wang and Yang (2007), Ma and Yang (2011), and Liu and Yang (2010) proposed a “spline-backfitted kernel smoothing” (SBK) method for the additive autoregressive model (6.3), the PLAM (6.4), and the ACM (6.5), respectively.

The SBK estimator is essentially as fast and accurate as a univariate kernel smoothing, satisfying all three criteria (i)–(iii), and is oracle efficient such that it has the same limiting distribution as the univariate function estimator by assuming that other parametric and nonparametric components are known. The SBK method is proposed for time series data, which has a geometrically α-mixing distribution. The SBK estimation method has several advantages compared to most of the existing methods. First, as pointed out in Sperlich et al. (2002), the estimator of Linton (1997) mixed up different projections, making it uninterpretable if the real data generating process deviates from additivity, while the projections in both steps of the SBK estimator are with respect to the same measure. Second, the SBK method is computationally expedient, since the pilot spline estimator is thousands of times faster than the pilot kernel estimator in Linton (1997), as demonstrated in Table 2 of Wang and Yang (2007). Third, the SBK estimator is shown to be as efficient as the "oracle smoother" uniformly over any compact range, whereas Linton (1997) proved such "oracle efficiency" only at a single point. Moreover, the regularity conditions needed by the SBK estimation procedure are natural and appealing and close to being minimal. In contrast, higher order smoothness is needed with growing dimensionality of the regressors in Linton and Nielsen (1995). Stronger and more obscure conditions are assumed for the two-stage estimation proposed by Horowitz and Mammen (2004).

Wang and Yang (2011) applied the SBK method to survey data. As an extension, Song and Yang (2010) proposed a spline backfitted spline (SBS) approach in the framework of additive autoregressive models. The SBS achieves the oracle efficiency as the SBK method, and is more computationally efficient. Asymptotically simultaneous confidence bands can be constructed for each functional curve by the proposed SBK and SBS methods. In the following sections, we will discuss the SBK method with applications to the additive model (6.3), the PLAM (6.4) and the ACM (6.5), and the SBS method for the additive model (6.3).

6.2. SBK in Additive Models

In model (6.3), if the last $d-1$ component functions were known by "oracle," one could create $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$ with $Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}) = m_1(X_{i1}) + \sigma(X_{i1}, \ldots, X_{id})\varepsilon_i$, from which one could compute an "oracle smoother" to estimate the only unknown function $m_1(x_1)$, thus effectively bypassing the "curse of dimensionality." The idea of Linton (1997) was to obtain an approximation to the unobservable variables Y_{i1} by substituting $m_\alpha(X_{i\alpha})$, $i = 1, \ldots, n$, $\alpha = 2, \ldots, d$, with marginal integration kernel estimates and arguing that the error incurred by this "cheating" is of smaller magnitude than the rate $O(n^{-2/5})$ for estimating function $m_1(x_1)$ from the unobservable data. Wang and Yang (2007) modify the procedure of Linton (1997) by substituting $m_\alpha(X_{i\alpha})$, $i = 1, \ldots, n$, $\alpha = 2, \ldots, d$, with spline estimators; specifically, Wang and Yang (2007) propose a two-stage estimation procedure: First they pre-estimate $\{m_\alpha(x_\alpha)\}_{\alpha=2}^{d}$ by its pilot estimator through an undersmoothed centered standard spline procedure, and next they construct the pseudo-response $\hat{Y}_{i1}$ and approximate $m_1(x_1)$ by its Nadaraya–Watson estimator.

The SBK estimator achieves its seemingly surprising success by borrowing the strengths of both spline and kernel: Spline does a quick initial estimation of all additive components and removes them all except the one of interests; kernel smoothing is then applied to the cleaned univariate data to estimate with asymptotic distribution. The proposed estimators achieve uniform oracle efficiency. The two-step estimating procedure accomplishes the well-known "reducing bias by undersmoothing" in the first step using spline and "averaging out the variance" in the second step with kernel, both steps taking advantage of the joint asymptotics of kernel and spline functions.

6.2.1. The SBK Estimator

In this section, we describe the spline-backfitted kernel estimation procedure. Let $\{Y_i, \mathbf{X}_i^{\mathrm{T}}\}_{i=1}^n = \{Y_i, X_{i1}, \ldots, X_{id}\}_{i=1}^n$ be observations from a geometrically α-mixing process following model (6.3). We assume that the predictor X_α is distributed on a compact interval $[a_\alpha, b_\alpha], \alpha = 1, \ldots, d$. Without loss of generality, we take all intervals $[a_\alpha, b_\alpha] = [0,1], \alpha = 1, \ldots, d$. Denote by $\|\varphi_\alpha\|_2$ the theoretical L_2 norm of a function φ_α on $[0,1]$, $\|\varphi_\alpha\|_2^2 = \int_0^1 \varphi_\alpha^2(x_\alpha) f(x_\alpha) dx_\alpha$. We preselect an integer $N = N_n \sim n^{2/5} \log n$; see Assumption (A6) in Wang and Yang (2007). Next, we define for any $\alpha = 1, \ldots, d$, the first-order B spline function (De Boor, 2001, p. 89), or say the constant B spline function as the indicator function $I_J(x_\alpha)$ of the $(N+1)$ equally spaced subintervals of the finite interval $[0,1]$ with length $H = H_n = (N+1)^{-1}$, that is,

$$I_{J,\alpha}(x_\alpha) = \begin{cases} 1, & JH \le x_\alpha < (J+1)H, \\ 0, & \text{otherwise,} \end{cases} \qquad J = 0, 1, \ldots, N.$$

Define the following centered spline basis:

$$b_{J,\alpha}(x_\alpha) = I_{J+1,\alpha}(x_\alpha) - \frac{\left\| I_{J+1,\alpha} \right\|_2}{\left\| I_{J,\alpha} \right\|_2} I_{J,\alpha}(x_\alpha), \qquad \forall \alpha = 1, \ldots, d, J = 1, \ldots, N, \tag{6.6}$$

with the standardized version given for any $\alpha = 1, \ldots, d$,

$$B_{J,\alpha}(x_\alpha) = \frac{b_{J,\alpha}(x_\alpha)}{\left\| b_{J,\alpha} \right\|_2}, \qquad \forall J = 1, \ldots, N.$$

Define next the $(1 + dN)$-dimensional space $G = G[0,1]$ of additive spline functions as the linear space spanned by $\{1, B_{J,\alpha}(x_\alpha), \alpha = 1, \ldots, d, J = 1, \ldots, N\}$, while $G_n \subset R^n$ is spanned by $\{1, \{B_{J,\alpha}(X_{i\alpha})\}_{i=1}^n, \alpha = 1, \ldots, d, J = 1, \ldots, N\}$. As $n \to \infty$, the dimension of G_n becomes $1 + dN$ with probability approaching one. The spline estimator of additive function $m(\mathbf{x})$ is the unique element $\hat{m}(\mathbf{x}) = \hat{m}_n(\mathbf{x})$ from the space G so that the vector $\{\hat{m}(\mathbf{X}_1), \ldots, \hat{m}(\mathbf{X}_n)\}^T$ best approximates the response vector $\mathbf{Y}$. To be precise, we define

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0' + \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha}' I_{J,\alpha}(x_\alpha), \tag{6.7}$$

where the coefficients $(\hat{\lambda}'_0, \hat{\lambda}'_{1,1}, \ldots, \hat{\lambda}'_{N,d})$ are solutions of the least squares problem

$$\left\{\hat{\lambda}'_0, \hat{\lambda}'_{1,1}, \ldots, \hat{\lambda}'_{N,d}\right\}^T = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \lambda_{J,\alpha} I_{J,\alpha}(X_{i\alpha}) \right\}^2 .$$

Simple linear algebra shows that

$$\hat{m}(\mathbf{x}) = \hat{\lambda}_0 + \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha), \tag{6.8}$$

where $(\hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d})$ are solutions of the following least squares problem

$$\left\{\hat{\lambda}_0, \hat{\lambda}_{1,1}, \ldots, \hat{\lambda}_{N,d}\right\}^T = \underset{R^{dN+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d} \sum_{J=1}^{N} \lambda_{J,\alpha} B_{J,\alpha}(X_{i\alpha}) \right\}^2 , \tag{6.9}$$

while (6.7) is used for data analytic implementation, the mathematically equivalent expression (6.8) is convenient for asymptotic analysis.

The pilot estimators of the component functions and the constant are

$$\hat{m}_\alpha(x_\alpha) = \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(X_{i\alpha}),$$

$$\hat{m}_c = \hat{\lambda}_0 + n^{-1} \sum_{\alpha=1}^{d} \sum_{i=1}^{n} \sum_{J=1}^{N} \hat{\lambda}_{J,\alpha} B_{J,\alpha}(X_{i\alpha}). \tag{6.10}$$

These pilot estimators are then used to define new pseudo-responses $\hat{Y}_{i1}$, which are estimates of the unobservable "oracle" responses Y_{i1}. Specifically,

$$\hat{Y}_{i1} = Y_i - \hat{c} - \sum_{\alpha=2}^{d} \hat{m}_\alpha(X_{i\alpha}), \qquad Y_{i1} = Y_i - c - \sum_{\alpha=2}^{d} m_\alpha(X_{i\alpha}), \tag{6.11}$$

where $\hat{c} = \overline{Y}_n = n^{-1} \sum_{i=1}^{n} Y_i$, which is a $\sqrt{n}$-consistent estimator of c by the Central Limit Theorem. Next, we define the spline-backfitted kernel (SBK) estimator of $m_1(x_1)$ as $\hat{m}_{\mathrm{SBK},1}(x_1)$ based on $\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^{n}$, which attempts to mimic the would-be Nadaraya–Watson estimator $\tilde{m}_{\mathrm{K},1}(x_1)$ of $m_1(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$ if the unobservable "oracle" responses $\{Y_{i1}\}_{i=1}^{n}$ were available:

$$\hat{m}_{\mathrm{SBK},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1)\, \hat{Y}_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)},$$

$$\tilde{m}_{\mathrm{K},1}(x_1) = \frac{\sum_{i=1}^{n} K_h(X_{i1} - x_1)\, Y_{i1}}{\sum_{i=1}^{n} K_h(X_{i1} - x_1)}, \tag{6.12}$$

where $\hat{Y}_{i1}$ and Y_{i1} are defined in (6.11). Similarly, the spline-backfitted local linear (SBLL) estimator $\hat{m}_{\mathrm{SBLL},1}(x_1)$ based on $\{\hat{Y}_{i1}, X_{i1}\}_{i=1}^{n}$ mimics the would-be local linear estimator $\tilde{m}_{\mathrm{LL},1}(x_1)$ based on $\{Y_{i1}, X_{i1}\}_{i=1}^{n}$:

$$\left\{\hat{m}_{\mathrm{SBLL},1}(x_1), \tilde{m}_{\mathrm{LL},1}(x_1)\right\} = (1,0)\left(\mathbf{Z}^T\mathbf{W}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{W}\left(\hat{\mathbf{Y}}_1, \mathbf{Y}_1\right), \tag{6.13}$$

in which the oracle and pseudo-response vectors are $\mathbf{Y}_1 = (Y_{11,\ldots,}Y_{n1})^T$, $\hat{\mathbf{Y}}_1 = (\hat{Y}_{11,\ldots,}\hat{Y}_{n1})^T$, and the weight and design matrices are

$$\mathbf{W} = \operatorname{diag}\{K_h(X_{i1} - x_1)\}_{i=1}^{n}, \qquad \mathbf{Z}^T = \begin{pmatrix} 1 & \cdots & 1 \\ X_{11} - x_1 & \cdots & X_{n1} - x_1 \end{pmatrix}.$$

The asymptotic properties of the smoothers $\tilde{m}_{\mathrm{K},1}(x_1)$ and $\tilde{m}_{\mathrm{LL},1}(x_1)$ are well-developed. Under Assumptions (A1)–(A5) in Wang and Yang (2009), according to Theorem 4.2.1 of Härdle (1990), one has the following for any $x_1 \in [h, 1-h]$:

$$\sqrt{nh}\left\{\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1) - b_{\mathrm{K}}(x_1)\,h^2\right\} \xrightarrow{D} N\left\{0, v^2(x_1)\right\},$$

$$\sqrt{nh}\left\{\tilde{m}_{\mathrm{LL},1}(x_1) - m_1(x_1) - b_{\mathrm{LL}}(x_1)\,h^2\right\} \xrightarrow{D} N\left\{0, v^2(x_1)\right\},$$

where

$$\begin{array}{lcl} b_{\mathrm{K}}(x_1) & = & \int u^2K(u)\,du\left\{m_1''(x_1)f_1(x_1)/2 + m_1'(x_1)f_1'(x_1)\right\}f_1^{-1}(x_1), \\ b_{\mathrm{LL}}(x_1) & = & \int u^2K(u)\,du\left\{m_1''(x_1)/2\right\} \\ v^2(x_1) & = & \int K^2(u)\,du E\left[\sigma^2(X_1,\ldots,X_d)\,|X_1 = x_1\right]f_1^{-1}(x_1). \end{array} \tag{6.14}$$

The equation for $\tilde{m}_{\mathrm{K},1}(x_1)$ requires additional Assumption (A7) in Wang and Yang (2009). The next two theorems state that the asymptotic magnitude of difference between $\hat{m}_{\mathrm{SBK},1}(x_1)$ and $\tilde{m}_{\mathrm{K},1}(x_1)$ is of order $o_p(n^{-2/5})$, both pointwise and uniformly, which is dominated by the asymptotic size of $\tilde{m}_{\mathrm{K},1}(x_1) - m_1(x_1)$. Hence $\hat{m}_{\mathrm{SBK},1}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{\mathrm{K},1}(x_1)$. The same is true for $\hat{m}_{\mathrm{SBLL},1}(x_1)$ and $\tilde{m}_{\mathrm{LL},1}(x_1)$.

Theorem 6.1. *Under Assumptions (A1)–(A6) in Wang and Yang (2009), the estimators $\hat{m}_{SBK,1}(x_1)$ and $\hat{m}_{SBLL,1}(x_1)$ given in (6.12) and (6.13) satisfy*

$$\left|\hat{m}_{SBK,1}(x_1) - \tilde{m}_{K,1}(x_1)\right| + \left|\hat{m}_{SBLL,1}(x_1) - \tilde{m}_{LL,1}(x_1)\right| = o_p\left(n^{-2/5}\right).$$

Hence with $b_K(x_1)$, $b_{LL}(x_1)$ and $v^2(x_1)$ as defined in (6.19), for any $x_1 \in [h, 1-h]$, we obtain

$$\sqrt{nh}\left\{\hat{m}_{SBLL,1}(x_1) - m_1(x_1) - b_{LL}(x_1)\,h^2\right\} \xrightarrow{D} N\left\{0, v^2(x_1)\right\},$$

and with the additional Assumption (A7) in Wang and Yang (2009), we have

$$\sqrt{nh}\left\{\hat{m}_{SBK,1}(x_1) - m_1(x_1) - b_K(x_1)h^2\right\} \xrightarrow{D} N\left\{0, v^2(x_1)\right\}.$$

Theorem 6.2. *Under Assumptions (A1)–(A6) and (A2′) in Wang and Yang (2009), the estimators* $\hat{m}_{SBK,1}(x_1)$ *and* $\hat{m}_{SBLL,1}(x_1)$ *given in (6.12) and (6.13) satisfy*

$$\sup_{x_1 \in [0,1]} \left\{ \left| \hat{m}_{SBK,1}(x_1) - \tilde{m}_{K,1}(x_1) \right| + \left| \hat{m}_{SBLL,1}(x_1) - \tilde{m}_{LL,1}(x_1) \right| \right\} = o_p\left(n^{-2/5}\right).$$

Hence for any z,

$$\lim_{n\to\infty} P\left[\left\{ \log\left(h^{-2}\right) \right\}^{1/2} \left(\sup_{x_1 \in [h,1-h]} \frac{\sqrt{nh}}{v(x_1)} \left| \hat{m}_{SBLL,1}(x_1) - m_1(x_1) \right| - d_n \right) < z \right]$$
$$= \exp\left\{ -2\exp(-z) \right\},$$

in which $d_n = \{\log(h^{-2})\}^{1/2} + \{\log(h^{-2})\}^{-1/2} \log\{c(K')(2\pi)^{-1}c^{-1}(K)\}$*. With the additional Assumption (A7) in Wang and Yang (2009), it is also true that*

$$\lim_{n\to\infty} P\left[\left\{ \log\left(h^{-2}\right) \right\}^{1/2} \left(\sup_{x_1 \in [h,1-h]} \frac{\sqrt{nh}}{v(x_1)} \left| \hat{m}_{SBK,1}(x_1) - m_1(x_1) \right| - d_n \right) < z \right]$$
$$= \exp\left\{ -2\exp(-z) \right\}.$$

For any $\alpha \in (0,1)$*, an asymptotic* $100(1-\alpha)\%$ *confidence band for* $m_1(x_1)$ *over interval* $[h, 1-h]$ *is*

$$\hat{m}_{SBLL,1}(x_1) \pm v(x_1)\,(nh)^{-1/2} \left[d_n - \log^{-1/2}\left(h^{-2}\right) \log\left\{ -\frac{\log(1-\alpha)}{2} \right\} \right].$$

Remark 6.1. Similar estimators $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ and $\hat{m}_{\mathrm{SBLL},\alpha}(x_\alpha)$ can be constructed for $m_\alpha(x_\alpha)$, $2 \le \alpha \le d$ with same oracle properties.

6.2.2. Application to Boston Housing Data

Wang and Yang (2009) applied the proposed method to the well-known Boston housing data, which contains 506 different houses from a variety of locations in Boston Standard Metropolitan Statistical Area in 1970. The median value and 13 sociodemographic statistics values of the Boston houses were first studied by Harrison and Rubinfeld (1978) to estimate the housing price index model. Breiman and Friedman (1985) did further analysis to deal with the multicollinearity for overfitting by using a stepwise method. The response and explanatory variables of interest are:

MEDV: Median value of owner-occupied homes in $1000's

RM: Average number of rooms per dwelling

TAX: Full-value property-tax rate per $10,000

PTRATIO: Pupil–teacher ratio by town school district

LSTAT: Proportion of population that is of "lower status" in %.

In order to ease off the trouble caused by big gaps in the domain of variables TAX and LSTAT, logarithmic transformation is done for both variables before fitting the model. Wang and Yang (2009) fitted an additive model as follows:

$$\text{MEDV} = \mu + m_1(\text{RM}) + m_2\big(\log(\text{TAX})\big) + m_3(\text{PTRATIO}) + m_4\big(\log(\text{LSTAT})\big) + \varepsilon.$$

In Figure 6.3 (see Wang and Yang, 2009), the univariate function estimates and corresponding confidence bands are displayed together with the "pseudo-data points" with

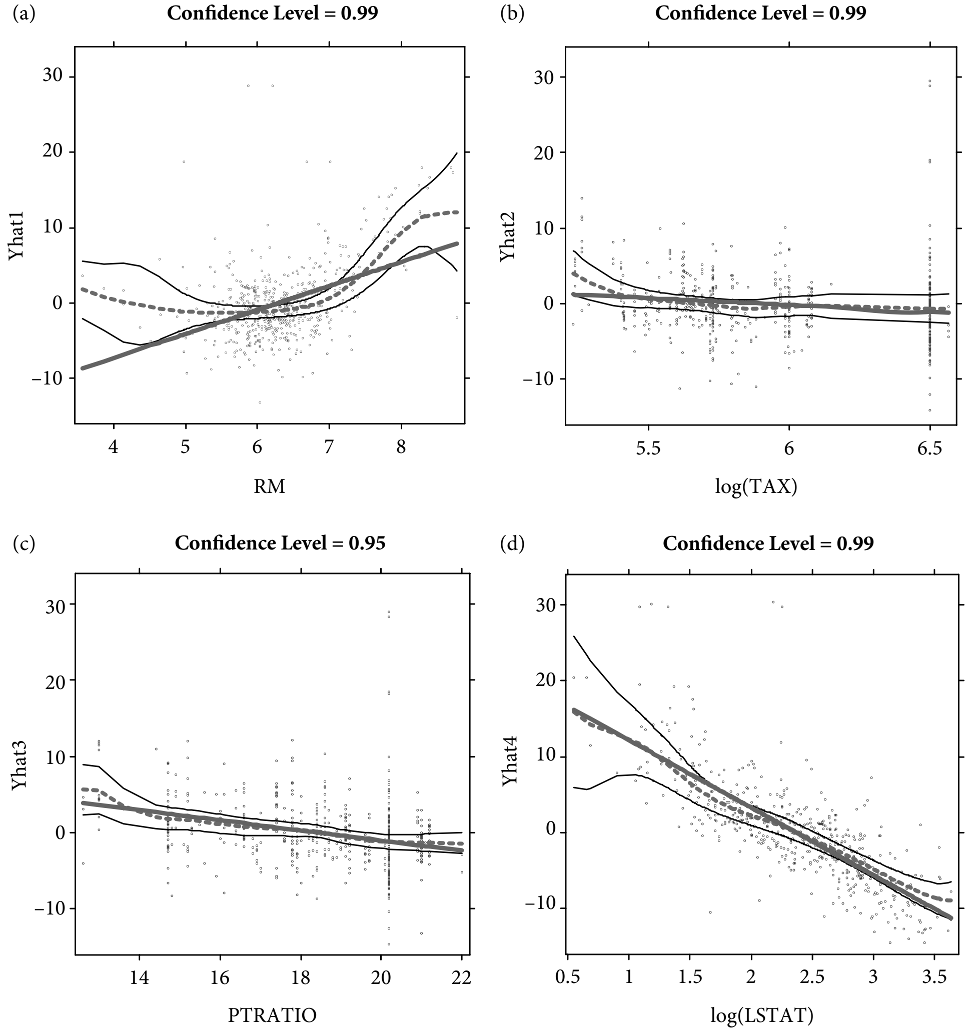


FIGURE 6.3 Linearity test for the Boston housing data. Plots of null hypothesis curves of $\mathcal{H}_0$: $m(x_\alpha) = a_\alpha + b_\alpha \cdot x_\alpha$, $\alpha = 1, 2, 3, 4$ (solid line), linear confidence bands (upper and lower thin lines), the linear spline estimator (dotted line), and the data (circle).

pseudo-response as the backfitted response after subtracting the sum function of the remaining three covariates. All the function estimates are represented by the dotted lines, "data points" by circles, and confidence bands by upper and lower thin lines. The kernel used in SBLL estimator is quartic kernel, $K(u) = \frac{15}{16}(1-u^2)^2$ for $-1 < u < 1$.

The proposed confidence bands are used to test the linearity of the components. In Figure 6.3 the straight solid lines are the least squares regression lines. The first figure shows that the null hypothesis $\mathcal{H}_0$: $m_1(\mathrm{RM}) = a_1 + b_1\mathrm{RM}$ will be rejected since the confidence bands with 0.99 confidence couldn't totally cover the straight regression line; that is, the p-value is less than 0.01. Similarly, the linearity of the component functions for log(TAX) and log(LSTAT) are not accepted at the significance level 0.01. While the least squares straight line of variable PTRATIO in the upper right figure totally falls between the upper and lower 95% confidence bands, the linearity null hypothesis $\mathcal{H}_0$: $m_3(\mathrm{PTRATIO}) = a_3 + b_3\mathrm{PTRATIO}$ is accepted at the significance level 0.05.

6.3. SBK in Partially Linear Additive Models (PLAM)

Wang and Yang (2009) fitted an additive model using RM, log(TAX), PTRSATIO and log(LSTAT) as predictors to test the linearity of the components and found that only PTRATIO is accepted at the significance level 0.05 for the linearity hypothesis test. Based on the conclusion drawn from Wang and Yang (2009), a PLAM can be fitted as

$$\begin{aligned}\mathrm{MEDV} = c_{00} &+ c_{01} \times \mathrm{PTRATIO} + m_1(\mathrm{RM}) \\ &+ m_2\big(\log(\mathrm{TAX})\big) + m_3\big(\log(\mathrm{LSTAT})\big) + \varepsilon. \qquad (6.15)\end{aligned}$$

PLAMs contain both linear and nonlinear additive components, and they are more flexible compared to linear models and more efficient compared to general nonparametric regression models. A general form of PLAMs is given in (6.4). In the PLAM (6.4), if the regression coefficients $\{c_{0l}\}_{l=0}^{d_1}$ and the component functions $\{m_\beta(x_\beta)\}_{\beta=1,\beta\neq\alpha}^{d_2}$ were known by "oracle", one could create $\{Y_{i\alpha}, X_{i\alpha}\}_{i=1}^n$ with $Y_{i\alpha} = Y_i - c_{00} - \sum_{l=1}^{d_1} c_{0l}T_{il} - \sum_{\beta=1,\beta\neq\alpha}^{d_2} m_\beta(X_{i\beta}) = m_\alpha(X_{i\alpha}) + \sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i$, from which one could compute an "oracle smoother" to estimate the only unknown function $m_\alpha(x_\alpha)$, bypassing the "curse of dimensionality." A major theoretical innovation is to resolve the dependence between $\mathbf{T}$ and $\mathbf{X}$, making use of Assumption (A5) in Ma and Yang (2011), which is not needed in Wang and Yang (2007). Another significant innovation is the $\sqrt{n}$-consistency and asymptotic distribution of estimators for parameters $\{c_{0l}\}_{l=0}^{d_1}$, which is trivial for the additive model of Wang and Yang (2007).

We denote by $\mathbf{I}_r$ the $r \times r$ identity matrix, $\mathbf{0}_{r\times s}$ the zero matrix of dimension $r \times s$, and $\mathrm{diag}(a, b)$ the 2×2 diagonal matrix with diagonal entries a, b. Let $\{Y_i, \mathbf{X}_i^{\mathrm{T}}, \mathbf{T}_i^{\mathrm{T}}\}_{i=1}^n$ be a sequence of strictly stationary observations from a geometrically α-mixing process following model (6.4), where Y_i and $(\mathbf{X}_i, \mathbf{T}_i) = \{(X_{i1}, \ldots, X_{id_2})^T, (T_{i1}, \ldots T_{id_1})^T\}$ are the ith response and predictor vector. Define next the space G of partially linear additive spline functions as the linear space spanned by $\{1, t_l, b_{J,\alpha}(x_\alpha), 1 \le l \le d_1, 1 \le \alpha \le d_2, 1 \le J \le N+1\}$. Let $\{1, \{T_l, b_{J,\alpha}(X_{i\alpha})\}_{i=1}^n, 1 \le l \le d_1, 1 \le \alpha \le d_2, 1 \le J \le N+1\}$ span the space $G_n \subset R^n$, where $b_{J,\alpha}$ is defined in (6.6). As $n \to \infty$, with probability approaching 1, the dimension of G_n becomes $\{1 + d_1 + d_2(N+1)\}$. The spline estimator of $m(\mathbf{x}, \mathbf{t})$ is the unique element $\hat{m}(\mathbf{x}, \mathbf{t}) = \hat{m}_n(\mathbf{x}, \mathbf{t})$ from G so that $\{\hat{m}(\mathbf{X}_i, \mathbf{T}_i)\}_{1 \le i \le n}^T$ best approximates the response vector $\mathbf{Y}$. To be precise, we define

$$\hat{m}(\mathbf{x}, \mathbf{t}) = \hat{c}_{00} + \sum\nolimits_{l=1}^{d_1} \hat{c}_{0l} t_l + \sum\nolimits_{\alpha=1}^{d_2} \sum\nolimits_{J=1}^{N+1} \hat{c}_{J,\alpha} b_{J,\alpha}(x_\alpha),$$

where the coefficients $(\hat{c}_{00}, \hat{c}_{0l}, \hat{c}_{J,\alpha})_{1 \le l \le d_1, 1 \le J \le N+1, 1 \le \alpha \le d_2}$ minimize

$$\sum\nolimits_{i=1}^{n} \left\{ Y_i - c_0 - \sum\nolimits_{l=1}^{d_1} c_l T_{il} - \sum\nolimits_{\alpha=1}^{d_2} \sum\nolimits_{J=1}^{N+1} c_{J,\alpha} b_{J,\alpha}(X_{i\alpha}) \right\}^2 .$$

Pilot estimators of $\mathbf{c}^T = \{c_{0l}\}_{l=0}^{d_1}$ and $m_\alpha(x_\alpha)$ are $\hat{\mathbf{c}}^T = \{\hat{c}_{0l}\}_{l=0}^{d_1}$ and $\hat{m}_\alpha(x_\alpha) = \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} b_{J,\alpha}(x_\alpha) - n^{-1} \sum_{i=1}^n \sum_{J=1}^{N+1} \hat{c}_{J,\alpha} b_{J,\alpha}(X_{i\alpha})$, which are used to define pseudo-responses $\hat{Y}_{i\alpha}$, estimates of the unobservable "oracle" responses $Y_{i\alpha}$:

$$\begin{aligned} \hat{Y}_{i\alpha} &= Y_i - \hat{c}_{00} - \sum\nolimits_{l=1}^{d_1} \hat{c}_{0l} T_{il} - \sum\nolimits_{\beta=1, \beta \ne \alpha}^{d_2} \hat{m}_\beta \left(X_{i\beta}\right), \\ Y_{i\alpha} &= Y_i - c_{00} - \sum\nolimits_{l=1}^{d_1} c_{0l} T_{il} - \sum\nolimits_{\beta=1, \beta \ne \alpha}^{d_2} m_\beta \left(X_{i\beta}\right). \end{aligned} \tag{6.16}$$

Based on $\{\hat{Y}_{i\alpha}, X_{i\alpha}\}_{i=1}^n$, the SBK estimator $\hat{m}_{\mathrm{SBK},\alpha}(x_\alpha)$ of $m_\alpha(x_\alpha)$ mimics the would-be Nadaraya–Watson estimator $\tilde{m}_{\mathrm{K},\alpha}(x_\alpha)$ of $m_\alpha(x_\alpha)$ based on $\{Y_{i\alpha}, X_{i\alpha}\}_{i=1}^n$, if the unobservable responses $\{Y_{i\alpha}\}_{i=1}^n$ were available:

$$\begin{aligned} \hat{m}_{\mathrm{SBK},\alpha}(x_\alpha) &= \left\{ n^{-1} \sum\nolimits_{i=1}^n K_h(X_{i\alpha} - x_\alpha) \hat{Y}_{i\alpha} \right\} / \hat{f}_\alpha(x_\alpha), \\ \tilde{m}_{\mathrm{K},\alpha}(x_\alpha) &= \left\{ n^{-1} \sum\nolimits_{i=1}^n K_h(X_{i\alpha} - x_\alpha) Y_{i\alpha} \right\} / \hat{f}_\alpha(x_\alpha), \end{aligned} \tag{6.17}$$

with $\hat{Y}_{i\alpha}, Y_{i\alpha}$ in (6.16), $\hat{f}_\alpha(x_\alpha) = n^{-1} \sum_{i=1}^n K_h(X_{i\alpha} - x_\alpha)$ an estimator of $f_\alpha(x_\alpha)$.

Define the Hilbert space

$$\mathcal{H} = \left\{ p(\mathbf{x}) = \sum\nolimits_{\alpha=1}^{d_2} p_\alpha(x_\alpha), E p_\alpha(X_\alpha) = 0, E^2 p_\alpha(X_\alpha) < \infty \right\}$$

of theoretically centered L_2 additive functions on $[0,1]^{d_2}$, while denote by $\mathcal{H}_n$ its subspace spanned by $\{B_{J,\alpha}(x_\alpha), 1 \le \alpha \le d_2, 1 \le J \le N+1\}$. Denote

$$\mathrm{Proj}_{\mathcal{H}} T_l = p_l(\mathrm{X}) = \mathrm{argmin}_{p \in \mathcal{H}} E\left\{T_l - p(\mathrm{X})\right\}^2, \tilde{T}_l = T_l - \mathrm{Proj}_{\mathcal{H}} T_l,$$

$$\mathrm{Proj}_{\mathcal{H}_n} T_l = \mathrm{argmin}_{p \in \mathcal{H}_n} E\left\{T_l - p(\mathrm{X})\right\}^2, \tilde{T}_{l,n} = T_l - \mathrm{Proj}_{\mathcal{H}_n} T_l,$$

for $1 \leq l \leq d_1$, where $\text{Proj}_{\mathcal{H}}\text{T}_l$ and $\text{Proj}_{\mathcal{H}_n} T_l$ are orthogonal projections of T_l unto subspaces $\mathcal{H}$ and $\mathcal{H}_n$, respectively. Denote next in vector form

$$\tilde{\text{T}}_n = \left\{\tilde{T}_{l,n}\right\}_{1 \leq l \leq d_1}, \qquad \tilde{\text{T}} = \left\{\tilde{T}_l\right\}_{1 \leq l \leq d_1}. \tag{6.18}$$

Without loss of generality, let $\alpha = 1$. Under Assumptions (A1)–(A5) and (A7) in Ma and Yang (2011), it is straightforward to verify (as in Bosq (1998)) that as $n \to \infty$, we obtain

$$\begin{gathered} \sup_{x_1 \in [h, 1-h]} \left|\tilde{m}_{\text{K},1}(x_1) - m_1(x_1)\right| = o_p\left(n^{-2/5} \log n\right), \\ \sqrt{nh}\left\{\tilde{m}_{\text{K},1}(x_1) - m_1(x_1) - b_1(x_1)\, h^2\right\} \xrightarrow{D} N\left\{0, v_1^2(x_1)\right\}, \end{gathered} \tag{6.19}$$

where,

$$b_1(x_1) = \int u^2 K(u)\, du\left\{m_1''(x_1) f_1(x_1)/2 + m_1'(x_1) f_1'(x_1)\right\} f_1^{-1}(x_1),$$

$$v_1^2(x_1) = \int K^2(u)\, du E\left[\sigma^2(\text{X}, \text{T})\,|X_1 = x_1\right] f_1^{-1}(x_1).$$

It is shown in Li (2000) and Schimek (2000) that the spline estimator $\hat{m}_1(x_1)$ in the first step uniformly converges to $m_1(x_1)$ with certain convergence rate, but lacks asymptotic distribution. Theorem 6.3 below states that the difference between $\hat{m}_{\text{SBK},1}(x_1)$ and $\tilde{m}_{\text{K},1}(x_1)$ is $o_p(n^{-2/5})$ uniformly, dominated by the asymptotic uniform size of $\tilde{m}_{\text{K},1}(x_1) - m_1(x_1)$. So $\hat{m}_{\text{SBK},1}(x_1)$ has identical asymptotic distribution as $\tilde{m}_{\text{K},1}(x_1)$.

Theorem 6.3. *Under Assumptions (A1)–(A7) in Ma and Yang (2011), as $n \to \infty$, the SBK estimator $\hat{m}_{\text{SBK},1}(x_1)$ given in (6.17) satisfies*

$$\sup_{x_1 \in [0,1]} \left|\hat{m}_{\text{SBK},1}(x_1) - \tilde{m}_{\text{K},1}(x_1)\right| = o_p\left(n^{-2/5}\right).$$

Hence with $b_1(x_1)$ and $v_1^2(x_1)$ as defined in (6.19), for any $x_1 \in [h, 1-h]$, $\sqrt{nh}\{\hat{m}_{\text{SBK},1}(x_1) - m_1(x_1) - b_1(x_1)h^2\} \xrightarrow{D} N\{0, v_1^2(x_1)\}$.

Instead of using a Nadaraya–Watson estimator, one can use a local polynomial estimator; see Fan and Gijbels (1996). Under Assumptions (A1)–(A7), for any $\alpha \in (0,1)$, an asymptotic $100(1-\alpha)\%$ confidence intervals for $m_1(x_1)$ is

$$\hat{m}_{\text{SBK},1}(x_1) - \hat{b}_1(x_1)h^2 \pm z_{\alpha/2}\hat{v}_1(x_1)\,(nh)^{-1/2},$$

where $\hat{b}_1(x_1)$ and $\hat{v}_1^2(x_1)$ are estimators of $b_1(x_1)$ and $v_1^2(x_1)$, respectively.

The following corollary provides the asymptotic distribution of $\hat{m}_{\text{SBK}}(\mathbf{x})$.

Corollary 6.1. *Under Assumptions (A1)–(A7) in Ma and Yang (2011) and $m_\alpha \in C^{(2)}[0,1]$, $2 \leq \alpha \leq d_2$. Let $\hat{m}_{\text{SBK}}(\mathbf{x}) = \sum_{\alpha=1}^{d_2} \hat{m}_{\text{SBK},\alpha}(x_\alpha)$, $b(\mathbf{x}) = \sum_{\alpha=1}^{d_2} b_\alpha(x_\alpha)$, $v^2(\mathbf{x}) = \sum_{\alpha=1}^{d_2} v_\alpha^2(x_\alpha)$, for any $\mathbf{x} \in [0,1]^{d_2}$, with SBK estimators $\hat{m}_{\text{SBK},\alpha}(x_\alpha)$, $1 \leq \alpha \leq d_2$, defined* *in (6.17), and $b_\alpha(x_\alpha)$, $v_\alpha^2(x_\alpha)$ similarly defined as in (6.19), as $n \to \infty$,*

$$\sqrt{nh}\left\{\hat{m}_{\mathrm{SBK}}(\mathbf{x}) - \sum\nolimits_{\alpha=1}^{d_2} m_\alpha(x_\alpha) - b(\mathbf{x})\,h^2\right\} \xrightarrow{D} N\{0, v^2(\mathbf{x})\}.$$

The next theorem describes the asymptotic behavior of estimator $\hat{\mathbf{c}}$ for $\mathbf{c}$.

Theorem 6.4. *Under Assumptions (A1)–(A6) in Ma and Yang (2011), as $n \to \infty$, $\|\hat{\mathbf{c}} - \mathbf{c}\| = O_p(n^{-1/2})$. With the additional Assumption A8 in Ma and Yang (2011),*

$$\sqrt{n}\left(\hat{\mathbf{c}} - \mathbf{c}\right) \to_d N\left(0, \sigma_0^2 \left\{ \begin{array}{cc} 1 & 0_{d_1}^T \\ 0_{d_1} & \Sigma^{-1} \end{array} \right\}\right),$$

for $\Sigma = \mathrm{cov}(\tilde{\mathrm{T}})$ with random vector $\tilde{\mathrm{T}}$ defined in (6.18).

6.3.1. Application to Boston Housing Data

The Boston housing data were studied by Ma and Yang (2011) by fitting model (6.15). Figure 6.4 (see Ma and Yang (2011)) shows the univariate nonlinear function estimates (dashed lines) and corresponding simultaneous confidence bands (thin lines) together with the "pseudo-data points" (dots) with pseudo-response as the backfitted response after subtracting the sum function of the remaining covariates. The confidence bands are used to test the linearity of the nonparametric components. In Figure 6.4 the straight solid lines are the least squares regression lines through the pseudo-data points. The first figure confidence band with 0.999999 confidence level does not totally cover the straight regression line; that is, the p-value is less than 0.000001. Similarly, the linearity of the component functions for log(TAX) and log(LSTAT) are rejected at the significance levels 0.017 and 0.007, respectively. The estimators $\hat{c}_{00}$ and $\hat{c}_{01}$ of c_{00} and c_{01} are 33.393 and -0.58845 and both are significant with p-values close to 0. The correlation between the estimated and observed values of MEDV is 0.89944, much higher than 0.80112 obtained by Wang and Yang (2009). This improvement is due to fitting the variable PTRATIO directly as linear with the higher accuracy of parametric model instead of treating it unnecessarily as a nonparametric variable. In other words, the PLAM fits the housing data much better than the additive model of Wang and Yang (2009).

6.4. SBK in Additive Coefficient Models (ACM)

To estimate the additive function components in model (6.5), we introduce the similar idea as in the previous two sections for additive models (6.3) and PLAMs (6.4).

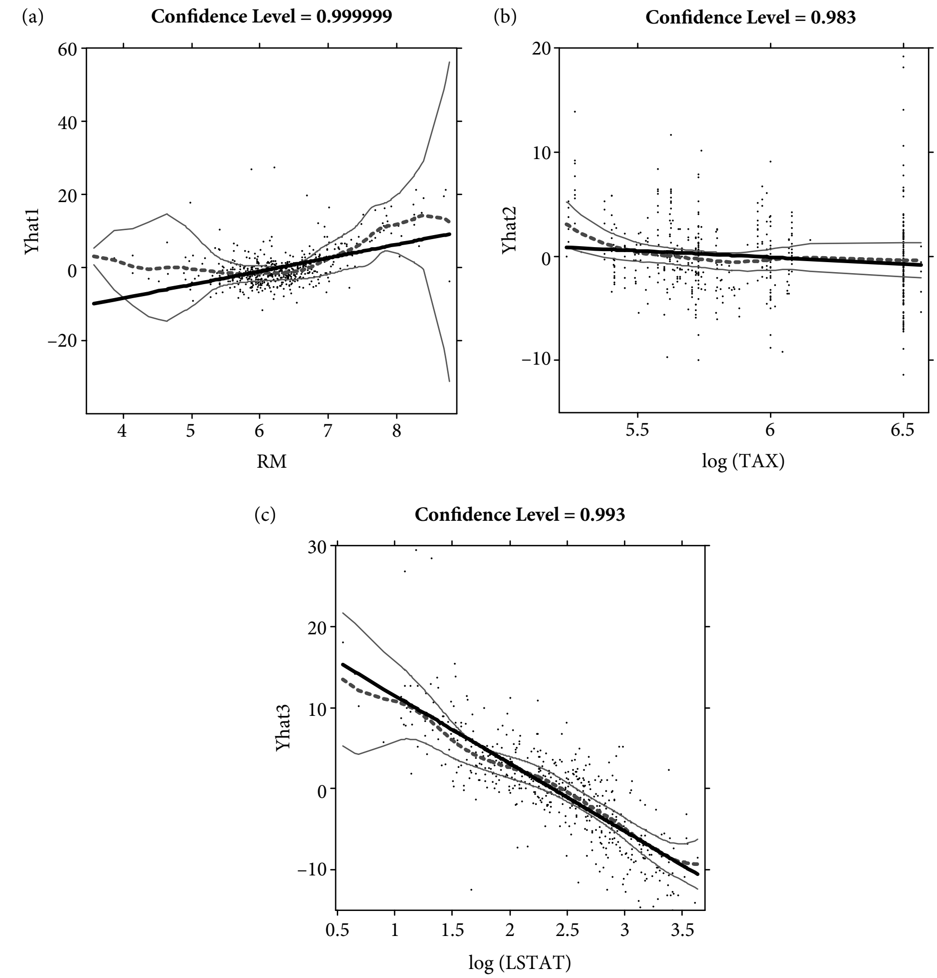


FIGURE 6.4 Plots of the least squares regression estimator (solid line), confidence bands (upper and lower thin lines), the spline estimator (dashed line), and the data (dot).

If all the nonparametric functions of the last $d_2 - 1$ variables, $\{m_{\alpha l}(x_\alpha)\}_{l=1,\alpha=2}^{d_1,d_2}$ and all the constants $\{m_{0l}\}_{l=1}^{d_1}$ were known by "oracle," one could define a new variable $Y_{,1} = \sum_{l=1}^{d_1} m_{1l}(X_1)T_l + \sigma(\mathbf{X},\mathbf{T})\varepsilon = Y - \sum_{l=1}^{d_1}\{m_{0l} + \sum_{\alpha=2}^{d_2} m_{\alpha l}(X_\alpha)\}T_l$ and estimate all functions $\{m_{1l}(x_1)\}_{l=1}^{d_1}$ by linear regression of $Y_{,1}$ on $T_1, \ldots, T_{d_1}$ with kernel weights computed from variable X_1. Instead of using the Nadaraya–Watson estimating method in the second step, Liu and Yang (2010) proposed to pre-estimate the functions $\{m_{\alpha l}(x_\alpha)\}_{l=1,\alpha=2}^{d_1,d_2}$ and constants $\{m_{0l}\}_{l=1}^{d_1}$ by linear spline and then use these estimates as substitutes to obtain an approximation $\hat{Y}_{,1}$ to the variable $Y_{,1}$, and construct "oracle" estimators based on $\hat{Y}_{,1}$.

Following Stone (1985a, p. 693), the space of α-centered square integrable functions on $[0,1]$ is

$$\mathcal{H}_\alpha^0 = \left\{g\colon E\left\{g\left(X_\alpha\right)\right\} = 0, E\left\{g^2(X_\alpha)\right\} < +\infty\right\}, \qquad 1 \le \alpha \le d_2.$$

Next define the model space $\mathcal{M}$, a collection of functions on $\chi \times R^{d_1}$, as

$$\mathcal{M} = \left\{ g(\mathbf{x},\mathbf{t}) = \sum_{l=1}^{d_1} g_l(\mathbf{x})\, t_l; \qquad g_l\,(\mathbf{x}) = g_{0l} + \sum_{\alpha=1}^{d_2} g_{\alpha l}(x_\alpha);\ g_{\alpha l} \in \mathcal{H}_\alpha^0 \right\},$$

in which $\{g_{0l}\}_{l=1}^{d_1}$ are finite constants. The constraints that $E\{g_{\alpha l}(X_\alpha)\} = 0$, $1 \le \alpha \le d_2$, ensure unique additive representation of m_l as expressed in (6.5) but are not necessary for the definition of space $\mathcal{M}$.

For any vector $\mathbf{x} = (x_1, x_2, \ldots, x_{d_2})$, denote the deleted vector as $\mathbf{x}_{\_1} = (x_2, \ldots, x_{d_2})$; for the random vector $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id_2})$, denote the deleted vector as $\mathbf{X}_{i,\_1} = (X_{i2}, \ldots, X_{id_2})$, $1 \le i \le n$. For any $1 \le l \le d_1$, write $m_{\_1,l}(\mathbf{x}_{\_1}) = m_{0l} + \sum_{\alpha=2}^{d_2} m_{\alpha l}(x_\alpha)$. Denote the vector of pseudo-responses $\mathbf{Y}_1 = (Y_{1,1}, \ldots, Y_{n,1})^T$ in which

$$Y_{i,1} = Y_i - \sum_{l=1}^{d_1} \left\{m_{0l} + m_{\_1,l}\left(\mathbf{X}_{i,\_1}\right)\right\} T_{il} = \sum_{l=1}^{d_1} m_{1l}(X_{i1})\, T_{il} + \sigma(\mathbf{X}_i, \mathbf{T}_i)\varepsilon_i.$$

These would be the "responses" if the unknown functions $\{m_{\_1,l}(\mathbf{x}_{\_1})\}_{1 \le l \le d_1}$ had been given. In that case, one could "estimate" all the coefficient functions in x_1, the vector function $m_{1,\cdot}(x_1) = \{m_{11}(x_1), \ldots, m_{1d_1}(x_1)\}^T$ by solving a kernel weighted least squares problem

$$\tilde{m}_{\mathrm{K},1,\cdot}\,(x_1) = \left\{\tilde{m}_{\mathrm{K},11}(x_1), \ldots, \tilde{m}_{\mathrm{K},1d_1}(x_1)\right\}^T = \underset{\boldsymbol{\lambda}=(\lambda_l)_{1\le l\le d_1}}{\operatorname{argmin}}\ L\left(\boldsymbol{\lambda}, m_{\_1,\cdot}, x_1\right)$$

in which

$$L\left(\boldsymbol{\lambda}, m_{\_1,\cdot}, x_1\right) = \sum_{i=1}^{n} \left(Y_{i,1} - \sum_{l=1}^{d_1} \lambda_l T_{il} \right)^2 K_h(X_{i1} - x_1).$$

Alternatively, one could rewrite the above kernel oracle smoother in matrix form

$$\tilde{m}_{\mathrm{K},1,\cdot}(x_1) = \left(\mathbf{C}_\mathrm{K}^T \mathbf{W}_1 \mathbf{C}_\mathrm{K}\right)^{-1} \mathbf{C}_\mathrm{K}^T \mathbf{W}_1 \mathbf{Y}_1 = \left(\frac{1}{n}\mathbf{C}_\mathrm{K}^T \mathbf{W}_1 \mathbf{C}_\mathrm{K}\right)^{-1} \frac{1}{n}\mathbf{C}_\mathrm{K}^T \mathbf{W}_1 \mathbf{Y}_1 \qquad (6.20)$$

in which

$$\mathbf{T}_i = \left(T_{i1}, \ldots, T_{id_1}\right)^T, \qquad \mathbf{C}_\mathrm{K} = \{\mathbf{T}_1, \ldots, \mathbf{T}_n\}^T,$$
$$\mathbf{W}_1 = \operatorname{diag}\{K_h\,(X_{11} - x_1), \ldots, K_h(X_{n1} - x_1)\}.$$

Likewise, one can define the local linear oracle smoother of $m_{1,\cdot}(x_1)$ as

$$\tilde{m}_{\mathrm{LL},1,\cdot}(x_1) = \left(\mathbf{I}_{d_1\times d_1}, \mathbf{0}_{d_1\times d_1}\right)\left(\frac{1}{n}\mathbf{C}_{\mathrm{LL},1}^T\mathbf{W}_1\mathbf{C}_{\mathrm{LL},1}\right)^{-1}\frac{1}{n}\mathbf{C}_{\mathrm{LL},1}^T\mathbf{W}_1\mathbf{Y}_1, \tag{6.21}$$

in which

$$\mathbf{C}_{\mathrm{LL},1} = \left\{\begin{array}{ccc} \mathbf{T}_1 & ,\ldots, & \mathbf{T}_n \\ \mathbf{T}_1(X_{11}-x_1), & \ldots, & \mathbf{T}_n(X_{n1}-x_1) \end{array}\right\}^T.$$

Denote $\mu_2(K) = \int u^2K(u)du, \|K\|_2^2 = \int K(u)^2du$, $\mathbf{Q}_1(x_1) = \{q_1(x_1)\}_{l,l'=1}^{d_1} = E(\mathbf{T}\mathbf{T}^T|X_1 = x_1)$, and define the following bias and variance coefficients:

$$\begin{aligned} b_{\mathrm{LL},l,l',1}(x_1) &= \frac{1}{2}\mu_2(K)\, m_{1l}''(x_1) f_1(x_1)\, q_{ll',1}(x_1), \\ b_{\mathrm{K},l,l',1}(x_1) &= \frac{1}{2}\mu_2(K)\left[2m_{1l}'(x_1)\frac{\partial}{\partial x_1}\left\{f_1(x_1)\, q_{ll',1}(x_1)\right\}\right. \\ &\quad \left. + \, m_{1l}''(x_1) f_1(x_1)\, q_{ll',1}(x_1)\right], \\ \Sigma_1(x_1) &= \|K\|_2^2 f_1(x_1)\, E\left\{\mathbf{T}\mathbf{T}^T\sigma^2(\mathbf{X},\mathbf{T})\,|X_1 = x_1\right\}, \\ \left\{v_{l,l',1}(x_1)\right\}_{l,l'=1}^{d_1} &= \mathbf{Q}_1(x_1)^{-1}\,\Sigma_1(x_1)\,\mathbf{Q}_1(x_1)^{-1}. \end{aligned} \tag{6.22}$$

Theorem 6.5. *Under Assumptions (A1)–(A5) and (A7) in Liu and Yang (2010), for any $x_1 \in [h, 1-h]$, as $n \to \infty$, the oracle local linear smoother $\tilde{m}_{\mathrm{LL},1,\cdot}(x_1)$ given in (6.21) satisfies*

$$\sqrt{nh}\left[\tilde{m}_{\mathrm{LL},1,\cdot}(x_1) - m_{1,\cdot}(x_1) - \left\{\sum\nolimits_{l=1}^{d_1} b_{\mathrm{LL},l,l',1}(x_1)\right\}_{l'=1}^{d_1} h^2\right]$$
$$\to N\left(0, \left\{v_{l,l',1}(x_1)\right\}_{l,l'=1}^{d_1}\right).$$

With Assumption (A6) in addition, the oracle kernel smoother $\tilde{m}_{\mathrm{K},1,\cdot}(x_1)$ in (6.20) satisfies

$$\sqrt{nh}\left[\tilde{m}_{\mathrm{K},1,\cdot}(x_1) - m_{1,\cdot}(x_1) - \left\{\sum\nolimits_{l=1}^{d_1} b_{\mathrm{K},l,l',1}(x_1)\right\}_{l'=1}^{d_1} h^2\right]$$
$$\to N\left(0, \left\{v_{l,l',1}(x_1)\right\}_{l,l'=1}^{d_1}\right).$$

Theorem 6.6. *Under Assumptions (A1)–(A5) and (A7) in Liu and Yang (2010), as $n \to \infty$, the oracle local linear smoother $\tilde{m}_{\mathrm{LL},1,\cdot}(x_1)$ given in (6.21) satisfies*

$$\sup_{x_1\in[h,1-h]}\left|\tilde{m}_{\mathrm{LL},1,\cdot}(x_1) - m_{1,\cdot}(x_1)\right| = O_p\left(\log n/\sqrt{nh}\right).$$

With Assumption (A6) in addition, the oracle kernel smoother $\tilde{m}_{K,1,\cdot}(x_1)$ in (6.20) satisfies

$$\sup_{x_1 \in [h,1-h]} \left| \tilde{m}_{K,1,\cdot}(x_1) - m_{1,\cdot}(x_1) \right| = O_p\left(\log n/\sqrt{nh}\right).$$

Remark 6.2. *The above theorems hold for $\tilde{m}_{LL,\alpha,\cdot}(x_\alpha)$ and $\tilde{m}_{K,\alpha,\cdot}(x_\alpha)$ similarly constructed as $\tilde{m}_{LL,1,\cdot}(x_1)$and $\tilde{m}_{K,1,\cdot}(x_1)$, for any $\alpha = 2, \ldots, d_2$.*

The same oracle idea applies to the constants as well. Define the would-be "estimators" of constants $(m_{0l})^T_{1 \le l \le d_1}$ as the least squares solution

$$\tilde{m}_0 = (\tilde{m}_{0l})^T_{1 \le l \le d_1} = \arg\min \sum_{i=1}^{n} \left\{ Y_{ic} - \sum_{l=1}^{d_1} m_{0l} T_{il} \right\}^2 ,$$

in which the oracle responses are

$$Y_{ic} = Y_i - \sum_{l=1}^{d_1} \sum_{\alpha=1}^{d_2} m_{\alpha l}(X_{i\alpha})\, T_{il} = \sum_{l=1}^{d_1} m_{0l} T_{il} + \sigma(\mathbf{X}_i, \mathbf{T}_i)\, \varepsilon_i. \tag{6.23}$$

The following result provides optimal convergence rate of $\tilde{m}_0$ to m_0, which are needed for removing the effects of m_0 for estimating the functions $\{m_{1l}(x_1)\}_{l=1}^{d_1}$.

Proposition 6.1. *Under Assumptions (A1)–(A5) and (A8) in Liu and Yang (2010), as $n \to \infty$, $\sup_{1 \le l \le d_1} |\tilde{m}_{0l} - m_{0l}| = O_p(n^{-1/2})$.*

Although the oracle smoothers $\tilde{m}_{LL,\alpha,\cdot}(x_\alpha)$, $\tilde{m}_{K,\alpha,\cdot}(x_\alpha)$ possess the theoretical properties in Theorems 6.5 and 6.6, they are not useful statistics because they are computed based on the knowledge of unavailable functions $\{m_{\alpha l}(x_\alpha)\}_{l=1,\alpha=2}^{d_1,d_2}$ and constants $\{m_{0l}\}_{l=1}^{d_1}$. They do, however, motivate the spline-backfitted estimators that we introduce next.

In the following, we describe how the unknown functions $\{m_{\alpha l}(x_\alpha)\}_{l=1,\alpha=2}^{d_1,d_2}$ and constants $\{m_{0l}\}_{l=1}^{d_1}$ can be pre-estimated by linear spline and how the estimates are used to construct the "oracle estimators." Define the space of α-empirically centered linear spline functions on $[0,1]$ as

$$G^0_{n,\alpha} = \left\{ g_\alpha : g_\alpha(x_\alpha) \equiv \sum_{J=0}^{N+1} \lambda_J b_J(x_\alpha), E_n\left\{g_\alpha(X_\alpha)\right\} = 0 \right\}, \qquad 1 \le \alpha \le d_2,$$

and the space of additive spline coefficient functions on $\chi \times R^{d_1}$ as

$$G^0_n = \left\{ g(\mathbf{x}, \mathbf{t}) = \sum_{l=1}^{d_1} g_l(\mathbf{x})\, t_l; \qquad g_l(\mathbf{x}) = g_{0l} + \sum_{\alpha=1}^{d_2} g_{\alpha l}(x_\alpha); g_{0l} \in R, g_{\alpha l} \in G^0_{n,\alpha} \right\},$$

which is equipped with the empirical inner product $\langle \cdot, \cdot \rangle_{2,n}$.

The multivariate function $m(\mathbf{x},\mathbf{t})$ is estimated by an additive spline coefficient function

$$\hat{m}(\mathbf{x},\mathbf{t}) = \sum_{l=1}^{d_1} \hat{m}_l(\mathbf{x})t_l = \underset{g\in G_n^0}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{Y_i - g(\mathbf{X}_i,\mathbf{T}_i)\right\}^2. \tag{6.24}$$

Since $\hat{m}(\mathbf{x},\mathbf{t}) \in G_n^0$, one can write $\hat{m}_l(\mathbf{x}) = \widehat{m}_{0l} + \sum_{\alpha=1}^{d_2} \hat{m}_{\alpha l}(x_\alpha)$; for $\widehat{m}_{0l} \in R$ and $\hat{m}_{\alpha l}(x_\alpha) \in G_{n,\alpha}^0$. Simple algebra shows that the following oracle estimators of the constants m_{0l} are exactly equal to $\widehat{m}_{0l}$, in which the oracle pseudo-responses $\hat{Y}_{ic} = Y_i - \sum_{l=1}^{d_1}\sum_{\alpha=1}^{d_2} \hat{m}_{\alpha l}(X_{i\alpha})T_{il}$ which mimic the oracle responses Y_{ic} in (6.23)

$$\hat{m}_0 = \left(\hat{m}_{0l}\right)_{1\le l\le d_1}^T = \arg \min_{(\lambda_{01},\ldots,\lambda_{0d_1})} \sum_{i=1}^{n} \left\{\hat{Y}_{ic} - \sum_{l=1}^{d_1} \lambda_{0l}T_{il}\right\}^2. \tag{6.25}$$

Proposition 6.2. *Under Assumptions (A1)–(A5) and (A8) in Liu and Yang (2010), as* $n \to \infty$, $\sup_{1\le l\le d_1} |\hat{m}_{0l} - \tilde{m}_{0l}| = O_p(n^{-1/2})$, *hence* $\sup_{1\le l\le d_1} |\hat{m}_{0l} - m_{0l}| = O_p(n^{-1/2})$ *following Proposition 6.1.*

Define next the oracle pseudo-responses

$$\hat{Y}_{i1} = Y_i - \sum\nolimits_{l=1}^{d_1} \left(\hat{m}_{0l} + \sum\nolimits_{\alpha=2}^{d_2} \hat{m}_{\alpha l}(X_{i\alpha})\right) T_{il}$$

and $\hat{\mathbf{Y}}_1 = (\hat{Y}_{11},\ldots,\hat{Y}_{n1})^T$, with $\hat{m}_{0l}$ and $\hat{m}_{\alpha l}$ defined in (6.25) and (6.24), respectively. The spline-backfitted kernel (SBK) and spline-backfitted local linear (SBLL) estimators are

$$\hat{m}_{\mathrm{SBK},1,\cdot}(x_1) = \left(\mathbf{C}_{\mathrm{K}}^T\mathbf{W}_1\mathbf{C}_{\mathrm{K}}\right)^{-1}\mathbf{C}^T\mathbf{W}_1\hat{\mathbf{Y}}_1 = \left(\frac{1}{n}\mathbf{C}_{\mathrm{K}}^T\mathbf{W}_1\mathbf{C}_{\mathrm{K}}\right)^{-1}\frac{1}{n}\mathbf{C}^T\mathbf{W}_1\hat{\mathbf{Y}}_1, \tag{6.26}$$

$$\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1) = \left(\mathbf{I}_{d_1\times d_1},\mathbf{0}_{d_1\times d_1}\right)\left(\frac{1}{n}\mathbf{C}_{\mathrm{LL},1}^T\mathbf{W}_1\mathbf{C}_{\mathrm{LL},1}\right)^{-1}\frac{1}{n}\mathbf{C}_{\mathrm{LL},1}^T\mathbf{W}_1\hat{\mathbf{Y}}_1. \tag{6.27}$$

The following theorem states that the asymptotic uniform magnitude of difference between $\hat{m}_{\mathrm{SBK},1,\cdot}(x_1)$ and $\tilde{m}_{\mathrm{K},1,\cdot}(x_1)$ is of order $o_p(n^{-2/5})$, which is dominated by the asymptotic size of $\tilde{m}_{\mathrm{K},1,\cdot}(x_1) - m_{1,\cdot}(x_1)$. As a result, $\hat{m}_{\mathrm{SBK},1,\cdot}(x_1)$ will have the same asymptotic distribution as $\tilde{m}_{\mathrm{K},1,\cdot}(x_1)$. The same is true for $\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1)$ and $\tilde{m}_{\mathrm{LL},1,\cdot}(x_1)$.

Theorem 6.7. *Under Assumptions (A1)–(A5), (A7), and (A8) in Liu and Yang (2010), as* $n\to\infty$, *the SBK estimator* $\hat{m}_{\mathrm{SBK},1,\cdot}(x_1)$ *in (6.26) and the SBLL estimator* $\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1)$ *in (6.27) satisfy*

$$\sup_{x_1\in[0,1]} \left\{\left|\hat{m}_{\mathrm{SBK},1,\cdot}(x_1) - \tilde{m}_{\mathrm{K},1,\cdot}(x_1)\right| + \left|\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1) - \tilde{m}_{\mathrm{LL},1,\cdot}(x_1)\right|\right\} = o_p\left(n^{-2/5}\right).$$

The following corollary provides the asymptotic distributions of $\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1)$ and $\tilde{m}_{\mathrm{K},1,\cdot}(x_1)$. The proof of this corollary is straightforward from Theorems 6.5 and 6.7.

Corollary 6.2. *Under Assumptions (A1)–(A5), (A7), and (A8) in Liu and Yang (2010), for any $x_1 \in [h, 1-h]$, as $n \to \infty$, the SBLL estimator $\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1)$ in (6.27) satisfies*

$$\sqrt{nh}\left[\hat{m}_{\mathrm{SBLL},1,\cdot}(x_1) - m_{1,\cdot}(x_1) - \left\{\sum_{l=1}^{d_1} b_{\mathrm{LL},l,l',1}(x_1)\,(x_1)\right\}_{l'=1}^{d_1} h^2\right]$$
$$\to N\left(0, \left\{v_{l,l',1}(x_1)\right\}_{l,l'=1}^{d_1}\right)$$

and with the additional Assumption (A6), the SBK estimator $\hat{m}_{\mathrm{SBK},1,\cdot}(x_1)$ in (6.26) satisfies

$$\sqrt{nh}\left[\tilde{m}_{\mathrm{K},1,\cdot}(x_1) - m_{1,\cdot}(x_1) - \left\{\sum_{l=1}^{d_1} b_{\mathrm{K},l,l',1}(x_1)\right\}_{l'=1}^{d_1} h^2\right]$$
$$\to N\left(0, \left\{v_{l,l',1}\,(x_1)\right\}_{l,l'=1}^{d_1}\right),$$

where $b_{\mathrm{LL},l,l',1}(x_1)$, $b_{\mathrm{K},l,l',1}(x_1)$ and $v_{l,l',1}(x_1)$ are defined as (6.22).

Remark 6.3. *For any $\alpha = 2, \ldots, d$, the above theorem and corollary hold for $\hat{m}_{\mathrm{SBK},\alpha,\cdot}(x_\alpha)$ and $\hat{m}_{\mathrm{SBLL},\alpha,\cdot}(x_\alpha)$ similarly constructed, that is,*

$$\hat{m}_{\mathrm{SBK},\alpha,\cdot}(x_\alpha) = \left(\frac{1}{n}\mathbf{C}_{\mathrm{K}}^{T}\mathbf{W}_{\alpha}\mathbf{C}_{\mathrm{K}}\right)^{-1}\frac{1}{n}\mathbf{C}_{\mathrm{K}}^{T}\mathbf{W}_{\alpha}\hat{\mathbf{Y}}_{\alpha},$$

where $\hat{Y}_{i\alpha} = Y_i - \sum_{l=1}^{d_1}\{\hat{m}_{0l} + \sum_{1\le\alpha'\le d_2,\alpha'\ne\alpha}\hat{m}_{\alpha l}(X_{i\alpha})\}$.

6.4.1. Application to Cobb–Douglas Model

Liu and Yang (2010) applied the SBLL procedure to a varying coefficient extension of the Cobb–Douglas model for the US GDP that allows non-neutral effects of the R&D on capital and labor as well as in total factor productivity (TFP). Denoted by Q_t the US GDP at year t, and K_t the US capital at year t, L_t the US labor at year t, and X_t the growth rate of ratio of R&D expenditure to GDP at year t, all data have been downloaded from the Bureau of Economic Analysis (BEA) website for years $t = 1959, \ldots, 2002$ ($n = 44$). The standard Cobb–Douglas production function (Cobb and Douglas (1928)) is $Q_t = A_t K_t^{\beta_1} L_t^{1-\beta_1}$, where A_t is the total factor productivity (TFP) of year t, β_1 is a parameter determined by technology. Define the following stationary time series variables

$$Y_t = \log Q_t - \log Q_{t-1}, T_{1t} = \log K_t - \log K_{t-1}, T_{2t} = \log L_t - \log L_{t-1},$$

then the Cobb–Douglas equation implies the following simple regression model

$$Y_t = \left(\log A_t - \log A_{t-1}\right) + \beta_1 T_{1t} + (1 - \beta_1) T_{2t}.$$

According to Solow (1957), the total factor productivity A_t has an almost constant rate of change, thus one might replace $\log A_t - \log A_{t-1}$ with an unknown constant and arrive at the following model:

$$Y_t - T_{2t} = \beta_0 + \beta_1 (T_{1t} - T_{2t}). \tag{6.28}$$

Because technology growth is one of the biggest subsections of TFP, it is reasonable to examine the dependence of both β_0 and β_1 on technology rather than treating them as fixed constants. We use exogenous variables X_t (growth rate of ratio of R&D expenditure to GDP at year t) to represent technology level and model $Y_t - T_{2t} = m_1(\mathbf{X}_t) + m_2(\mathbf{X}_t)(T_{1t} - T_{2t})$, where $m_l(\mathbf{X}_t) = m_{0l} + \sum_{\alpha=1}^{d_2} m_{\alpha l}(X_{t-\alpha+1})$, $l = 1,2$, $\mathbf{X}_t = (X_t, \ldots, X_{t-d_2+1})$. Using the BIC of Xue and Yang (2006b) for additive coefficient model with $d_2 = 5$, the following reduced model is considered optimal:

$$Y_t - T_{2t} = c_1 + m_{41}(X_{t-3}) + \{c_2 + m_{52}(X_{t-4})\}(T_{1t} - T_{2t}). \tag{6.29}$$

The rolling forecast errors of GDP by SBLL fitting of model (6.29) and linear fitting of (6.28) are show in Figure 6.1. The averaged squared prediction error (ASPE) for model (6.29) is

$$\frac{1}{9} \sum_{t=1994}^{2002} \left[Y_t - T_{2t} - \hat{c}_1 - \hat{m}_{\mathrm{SBLL},41}(X_{t-3}) - \left\{\hat{c}_2 + \hat{m}_{\mathrm{SBLL},52}(X_{t-4})\right\}(T_{1t} - T_{2t})\right]^2$$
$$= 0.001818,$$

which is about 60% of the corresponding ASPE (0.003097) for model (6.28). The in-sample averaged squared estimation error (ASE) for model (6.29) is 5.2399×10^{-5}, which is about 68% of the in sample ASE (7.6959×10^{-5}) for model (6.28).

In model (6.29), $\hat{c}_1 + \hat{m}_{\mathrm{SBLL},41}(X_{t-3})$ estimates the TFP growth rate, which is shown as a function of X_{t-3} in Figure 6.2. It is obvious that the effect of X_{t-3} is positive when $X_{t-3} \leq 0.02$, but negative when $X_{t-3} > 0.02$. On average, the higher R&D investment spending causes faster GDP growing. However, overspending on R&D often leads to high losses (Tokic, 2003).

Liu and Yang (2010) also computed the average contribution of R&D to GDP growth for 1964–2001, which is about 40%. The GDP and estimated TFP growth rates is shown in Figure 6.5 (see Liu and Yang, 2010), it is obvious that TFP growth is highly correlated to the GDP growth.

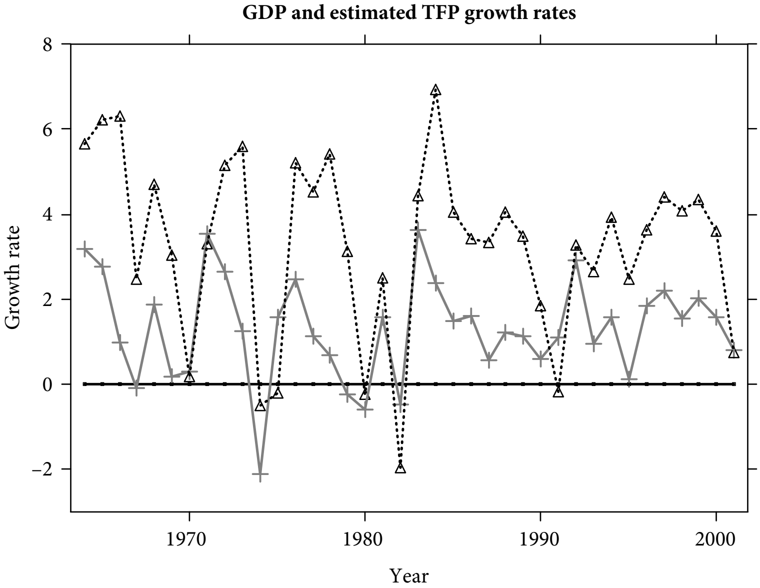


FIGURE 6.5 Estimation of function $\hat{c}_1 + \hat{m}_{\text{SBLL},41}(X_{t-3})$: GDP growth rate—dotted line; $\hat{c}_1 + \hat{m}_{\text{SBLL},41}(X_{t-3})$—solid line.

6.5. SBS in Additive Models

In this section, we describe the spline-backfitted spline estimation procedure for model (6.3). Let $0 = t_0 < t_1 < \cdots < t_{N+1} = 1$ be a sequence of equally spaced knots, dividing $[0,1]$ into $(N+1)$ subintervals of length $h = h_n = 1/(N+1)$ with a preselected integer $N \sim n^{1/5}$ given in Assumption (A5) of Song and Yang (2010), and let $0 = t_0^* < t_1^* < \cdots < t_{N^*+1}^* = 1$ be another sequence of equally spaced knots, dividing $[0,1]$ into N^* subintervals of length $H = H_n = N^{*-1}$, where $N^* \sim n^{2/5} \log n$ is another preselected integer; see Assumption (A5) in Song and Yang (2010). Denote by G_α the linear space spanned by $\{1, b_J(x_\alpha)\}_{J=1}^{N+1}$, whose elements are called linear splines, piecewise linear functions of x_α which are continuous on $[0,1]$ and linear on each subinterval $[t_J, t_{J+1}], 0 \leq J \leq N$. Denote by $G_{n,\alpha} \subset R^n$ the corresponding subspace of R^n spanned by $\{1, \{b_J(X_{i\alpha})\}_{i=1}^n\}_{J=1}^{N+1}$. Similarly, define the $\{1 + dN^*\}$-dimensional space $G^* = G^*([0,1]^d)$ of additive constant spline functions as the linear space spanned by $\{1, I_{J^*}(x_\alpha)\}_{\alpha=1,J^*=1}^{d,N^*}$, while denote by $G_n^* \subset R^n$ the corresponding subspace spanned by $\{1, \{I_{J^*}(X_{i\alpha})\}_{i=1}^n\}_{\alpha=1,J^*=1}^{d,N^*}$. As $n \to \infty$, with probability approaching one, the dimension of $G_{n,\alpha}$ becomes $N+2$, and the dimension of G_n^* becomes $1 + dN^*$.

The additive function $m(\mathbf{x})$ has a multivariate additive regression spline (MARS) estimator $\hat{m}(\mathbf{x}) = \hat{m}_n(\mathbf{x})$, the unique element of G^* so that the vector

$\{\hat{m}(\mathbf{X}_1),\ldots,\hat{m}(\mathbf{X}_n)\}^T \in G_n^*$ best approximates the response vector $\mathbf{Y}$. Precisely

$$\hat{m}(\mathbf{x}) = \underset{g\in G^*}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{Y_i - g\left(\mathbf{X}_i\right)\right\}^2 = \hat{\lambda}_0' + \sum_{\alpha=1}^{d}\sum_{J^*=1}^{N^*} \hat{\lambda}_{J^*,\alpha}' I_{J^*}\left(x_\alpha\right),$$

where $(\hat{\lambda}_0', \hat{\lambda}_{1,1}', \ldots, \hat{\lambda}_{N^*,d}')$ is the solution of the least squares problem

$$\left\{\hat{\lambda}_0', \hat{\lambda}_{1,1}', \ldots, \hat{\lambda}_{N^*,d}'\right\}^T = \underset{R^{d(N^*)+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ Y_i - \lambda_0 - \sum_{\alpha=1}^{d}\sum_{J^*=1}^{N^*} \lambda_{J^*,\alpha} I_{J^*}\left(X_{i\alpha}\right) \right\}^2.$$

Estimators of each component function and the constant are derived as

$$\hat{m}_\alpha\left(x_\alpha\right) = \sum_{J^*=1}^{N^*} \hat{\lambda}_{J^*,\alpha}' \left\{ I_{J^*}\left(x_\alpha\right) - n^{-1}\sum_{i=1}^{n} I_{J^*}\left(X_{i\alpha}\right) \right\},$$

$$\hat{m}_c = \hat{\lambda}_0' + n^{-1}\sum_{i=1}^{n}\sum_{\alpha=1}^{d}\sum_{J^*=1}^{N^*} \hat{\lambda}_{J^*,\alpha}' I_{J^*}(X_{i\alpha}) = \hat{c} = \overline{Y}.$$

These pilot estimators are used to define pseudo-responses $\hat{Y}_{i\alpha}$, $\forall 1 \le \alpha \le d$, which approximate the "oracle" responses $Y_{i\alpha}$. Specifically, we define

$$\hat{Y}_{i\alpha} = Y_i - \hat{c} - \sum_{\beta=1,\beta\neq\alpha}^{d} \hat{m}_\beta\left(X_{i\beta}\right),$$

where $\hat{c} = \overline{Y}_n = n^{-1}\sum_{i=1}^{n} Y_i$, which is a $\sqrt{n}$-consistent estimator of c by the Central Limit Theorem for strongly mixing sequences. Correspondingly, we denote vectors

$$\hat{\mathbf{Y}}_\alpha = \left\{\hat{Y}_{1\alpha}, \ldots, \hat{Y}_{n\alpha}\right\}^T, \qquad \mathbf{Y}_\alpha = \left\{Y_{1\alpha}, \ldots, Y_{n\alpha}\right\}^T. \tag{6.30}$$

We define the spline-backfitted spline (SBS) estimator of $m_\alpha(x_\alpha)$ as $\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha)$ based on $\{\hat{Y}_{i\alpha}, X_{i\alpha}\}_{i=1}^n$, which attempts to mimic the would-be spline estimator $\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)$ of $m_\alpha(x_\alpha)$ based on $\{Y_{i\alpha}, X_{i\alpha}\}_{i=1}^n$ if the unobservable "oracle" responses $\{Y_{i\alpha}\}_{i=1}^n$ were available. Then,

$$\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha) = \underset{g_\alpha\in G_\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{\hat{Y}_{i\alpha} - g_\alpha\left(X_{i\alpha}\right)\right\}^2,$$

$$\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha) = \underset{g_\alpha\in G_\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{Y_{i\alpha} - g_\alpha\left(X_{i\alpha}\right)\right\}^2. \tag{6.31}$$

Theorem 6.8. *Under Assumptions (A1)–(A5) in Song and Yang (2010), as $n \to \infty$, the SBS estimator $\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha)$ and the oracle smoother $\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)$ given in (6.31) satisfy*

$$\sup_{x_\alpha \in [0,1]} \left|\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha) - \tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)\right| = o_p\left(n^{-2/5}\right).$$

Theorem 6.8 provides that the maximal deviation of $\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha)$ from $\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)$ over $[0,1]$ is of the order $O_p(n^{-2/5}(\log n)^{-1}) = o_p(n^{-2/5}(\log n)^{1/2})$, which is needed for the maximal deviation of $\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha)$ from $m_\alpha(x_\alpha)$ over $[0,1]$ and the maximal deviation of $\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)$ from $m_\alpha(x_\alpha)$ to have the same asymptotic distribution, of order $n^{-2/5}(\log n)^{1/2}$. The estimator $\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha)$ is therefore asymptotically oracally efficient; that is, it is asymptotically equivalent to the oracle smoother $\tilde{m}_{\alpha,\mathrm{S}}(x_\alpha)$ and, in particular, the next theorem follows. The simultaneous confidence band given in (6.32) has width of order $n^{-2/5}(\log n)^{1/2}$ at any point $x_\alpha \in [0,1]$, consistent with published works on nonparametric simultaneous confidence bands such as Bosq (1998) and Claeskens and Van Keilegom (2003).

Theorem 6.9. *Under Assumptions (A1)–(A5) in Song and Yang (2010), for any $p \in (0,1)$, as $n \to \infty$, an asymptotic $100(1-p)\%$ simultaneous confidence band for $m_\alpha(x_\alpha)$ is*

$$\hat{m}_{\alpha,\mathrm{SBS}}(x_\alpha) \pm 2\hat{\sigma}_\alpha(x_\alpha)\left\{3\boldsymbol{\Delta}^T(x_\alpha)\,\Xi_{j(x_\alpha)}\boldsymbol{\Delta}(x_\alpha)\log(N+1)/2\hat{f}_\alpha(x_\alpha)\,nh\right\}^{1/2}$$
$$\times\left[1-\{2\log(N+1)\}^{-1}\left[\log\left(p/4\right)+\frac{1}{2}\{\log\log(N+1)+\log 4\pi\}\right]\right], \qquad (6.32)$$

where $\hat{\sigma}_\alpha(x_\alpha)$ and $\hat{f}_\alpha(x_\alpha)$ are some consistent estimators of $\sigma_\alpha(x_\alpha)$ and $f_\alpha(x_\alpha)$, $j(x_\alpha) =$ $\min\{[x_\alpha/h], N\}, \delta(x_\alpha) = \{x_\alpha - t_{j(x_\alpha)}\}/h$, and

$$\boldsymbol{\Delta}(x_\alpha) = \begin{pmatrix} c_{j(x_\alpha)-1}\{1-\delta(x_\alpha)\} \\ c_{j(x_\alpha)}\delta(x_\alpha) \end{pmatrix}, \qquad c_j = \begin{cases} \sqrt{2}, & j=-1,N, \\ 1, & 0 \le j \le N-1, \end{cases}$$

$$\Xi_j = \begin{pmatrix} l_{j+1,j+1} & l_{j+1,j+2} \\ l_{j+2,j+1} & l_{j+2,j+2} \end{pmatrix}, \qquad 0 \le j \le N,$$

where terms $\{l_{ik}\}_{|i-k|\le 1}$ are the entries of the inverse of the $(N+2)\times(N+2)$ matrix $\mathbf{M}_{N+2}$:

$$\mathbf{M}_{N+2} = \begin{pmatrix} 1 & \sqrt{2}/4 & & & & 0 \\ \sqrt{2}/4 & 1 & 1/4 & & & \\ & 1/4 & 1 & \ddots & & \\ & & \ddots & \ddots & 1/4 & \\ & & & 1/4 & 1 & \sqrt{2}/4 \\ 0 & & & & \sqrt{2}/4 & 1 \end{pmatrix}.$$

6.6. Future Research

Fan and Jiang (2005) provides generalized likelihood ratio (GLR) tests for additive models using the backfitting estimator. Similar GLR test based on the two-step estimator is feasible for future research. The SBS method can also be applied to the PLAMs (6.4) and the ACMs (6.5). The two-step estimating procedure can be extended to generalized additive, partially linear additive, and additive coefficient models. Ma et al. (2013) proposed a one-step penalized spline estimation and variable selection procedure in PLAMs with clustered/longitudinal data. The procedure is fast to compute, but lacks asymptotic distributions for the additive function components. Thus no confidence measures can be established. As another future work, our target is to (a) apply the two-step estimation to the analysis of clustered/longitudinal data and (b) establish the oracle efficiency of the estimators. The confidence bands of each additive function can be constructed based on the same idea in Section 6.5.

References

Bosq, D. 1998. *Nonparametric Statistics for Stochastic Processes.* New York: Springer-Verlag.

Breiman, X., and Y. Friedman. 1985. Estimating Optimal Transformation for Multiple Regression and Correlation. Journal of the American Statistical Association, **80**, pp. 580–598.

Claeskens, G., and I. Van Keilegom. 2003. "Bootstrap Confidence Bands for Regression Curves and Their Derivatives." *Annals of Statistics*, **31**, pp. 1852–1884.

Cobb, C. W., and P. H. Douglas. 1928. "A Theory of Production." *American Economic Review*, **18**, pp. 139–165.

De Boor, C. 2001. *A Practical Guide to Splines.* New York: Springer.

Fan, J., and Gijbels, I. 1996. *Local Polynomial Modelling and Its Applications.* London: Chapman and Hall.

Fan, J., and Jiang, J. 2005. "Nonparametric Inferences for Additive Models." *Journal of the American Statistical Association*, **100**, pp. 890–907.

Fan, J., W. Härdle, and E. Mammen. 1997. "Direct Estimation of Low-Dimensional Components in Additive Models." *Annals of Statistics*, **26**, pp. 943–971.

Härdle, W. 1990. *Applied Nonparametric Regression.* Cambridge, UK: Cambridge University Press.

Harrison, X., and Rubinfeld, Y. 1978. Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management, **5**, pp. 81–102.

Hastie, T. J. and R. J. Tibshirani. 1990. *Generalized Additive Models.* London: Chapman and Hall.

Hengartiner, N. W., and S. Sperlich. 2005. "Rate Optimal Estimation with the Integration Method in the Presence of Many Covariates." *Journal of Multivariate Analysis*, **95**, pp. 246–272.

Horowitz, J., and E. Mammen. 2004. Nonparametric Estimation of an Additive Model with a Link Function." *Annals of Statistics*, **32**, pp. 2412–2443.

Horowitz, J., J. Klemelä, and E. Mammen. 2006. “Optimal Estimation in Additive Regression. *Bernoulli*, **12**, pp. 271–298.

Huang, J. Z. 1998. “Projection Estimation in Multiple Regression with Application to Functional ANOVA Models.” *Annals of Statistics*, **26**, pp. 242–272.

Kim, W., O. B. Linton, and N. W. Hengartner. 1999. “A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals.” *Journal of Computational and Graphical Statistics*, **8**, pp. 278–297.

Li, Q. 2000. “Efficient Estimation of Additive Partially Linear Models.” *International Economic Review*, **41**, pp. 1073–1092.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Linton, O. B. 1997. “Efficient Estimation of Additive Nonparametric Regression Models.” *Biometrika*, **84**, pp. 469–473.

Linton, O. B., and W. Härdle. 1996. “Estimation of Additive Regression Models with Known Links.” *Biometrika*, **83**, pp. 529–540.

Linton, O. B., and J. P. Nielsen. 1995. “A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration.” *Biometrika*, **82**, pp. 93–100.

Liu, R., and L. Yang. 2010. “Spline-Backfitted Kernel Smoothing of Additive Coefficient Model.” *Econometric Theory*, **26**, pp. 29–59.

Ma, S., and L. Yang. 2011. “Spline-Backfitted Kernel Smoothing of Partially Linear Additive Model.” *Journal of Statistical Planning and Inference*, **141**, pp. 204–219.

Ma, S., Q. Song, and L. Wang. 2013. “Variable Selection and Estimation in Marginal Partially Linear Additive Models for Longitudinal Data.” *Bernoulli*, **19**, pp. 252–274.

Mammen, E., O. Linton, and J. Nielsen. 1999. “The Existence and Asymptotic Properties of a Backfitting Projection Algorithm Under Weak Conditions.” *Annals of Statistics*, **27**, pp. 1443–1490.

Opsomer, J. D., and D. Ruppert. 1997. “Fitting a Bivariate Additive Model by Local Polynomial Regression.” *Annals of Statistics*, **25**, pp. 186–211.

Schimek, M. 2000. “Estimation and Inference in Partially Linear Models with Smoothing Splines.” *Journal of Statistical Planning and Inference*, **91**, pp. 525–540.

Solow, R. M. 1957. “Technical Change and the Aggregate Production Function.” *The Review of Economics and Statistics*, **39**, pp. 312–320.

Song, Q., and L. Yang. 2010. “Oracally Efficient Spline Smoothing of Nonlinear Additive Autoregression Model with Simultaneous Confidence Band.” *Journal of Multivariate Analysis*, **101**, pp. 2008–2025.

Sperlich, S., D. Tjøstheim, and L. Yang. 2002. Nonparametric Estimation and Testing of Interaction in Additive Models.” *Econometric Theory*, **18**, pp. 197–251.

Stone, C. J. 1985a. “Additive Regression and Other Nonparametric Models.” *Annals of Statistics*, **13**, pp. 242–272.

Stone, C. J. 1985b. “The Use of Polynomial Splines and Their Tensor Products in Multivariate Function Estimation (with Discussion).” *Annals of Statistics*, **22**, pp. 118–184.

Tokic, D. 2003. “How Efficient Were r&d and Advertising Investments for Internet Firms Before the Bubble Burst? A Dea Approach.” *Credit and Financial anagement Review*, **9**, pp. 39–51.

Wang, J., and L. Yang. 2009. “Efficient and Fast Spline-Backfitted Kernel Smoothing of Additive Regression Model.” *Annals of the Institute of Statistical Mathematics*, **61**, pp. 663–690.

Wang, L., and S. Wang. 2011. "Nonparametric Additive Model-Assisted Estimation for Survey Data." *Journal of Multivariate Analysis*, **102**, pp. 1126–1140.

Wang, L., and L. Yang. 2007. "Spline-Backfitted Kernel Smoothing of Nonlinear Additive Autoregression Model." *Annals of Statistics*, **35**, pp. 2474–2503.

Xue, L., and L. Yang. 2006a. "Estimation of Semiparametric Additive Coefficient Model." *Journal of Statistical Planning and Inference*, **136**, pp. 2506–2534.

Xue, L., and L. Yang. 2006b. "Additive Coefficient Modelling via Polynomial Spline." *Statistica Sinica*, **16**, pp. 1423–1446.

Yang, L., B. U. Park, L. Xue, and W. Härdle. 2006. "Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration." *Journal of the American Statistical Association*, **101**, pp. 1212–1227.

Yang, L., S. Sperlich, and W. Härdle. 2003. "Derivative Estimation and Testing in Generalized Additive Models." *Journal of Statistical Planning and Inference*, **115**, pp. 521–542.

CHAPTER 7

ADDITIVE MODELS: EXTENSIONS AND RELATED MODELS

ENNO MAMMEN,[†] BYEONG U. PARK,[‡]
AND MELANIE SCHIENLE[§]

7.1. Introduction

In this chapter we continue the discussions on additive models of chapters 5 and 6. We come back to the smooth backfitting approach that was already mentioned there. The basic idea of the smooth backfitting is to replace the least squares criterion by a smoothed version. We now explain its definition in an additive model

$$E(Y|X) = \mu + f_1(X^1) + \cdots + f_d(X^d). \tag{7.1}$$

We assume that $(X_i^1,\ldots,X_i^d,Y_i)$, $1 \le i \le n$, are n observed i.i.d. copies of $(X^1,\ldots,X^d,Y)$, or more generally, n stationary copies. Below, in Section 7.4., we will also weaken the stationarity assumption.

In an additive model (7.1) the smooth backfitting estimators $\widehat{\mu},\widehat{f}_1,\ldots,\widehat{f}_d$ are defined as the minimizers of the smoothed least squares criterion

$$\begin{aligned}&\int \sum_{i=1}^{n}\Big[Y_i - \mu - f_1(x^1) - \cdots - f_d(x^d)\Big]^2 K\left(\frac{X_i^1 - x^1}{h_1}\right)\\ &\quad \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) dx^1 \cdots\, dx^d\end{aligned} \tag{7.2}$$

under the constraint

$$\int f_1(x^1)\widehat{p}_{X^1}(x^1)\,dx^1 = \cdots = \int f_d(x^d)\widehat{p}_{X^d}(x^d)\,dx^d = 0. \tag{7.3}$$

Here K is a kernel function; that is, a positive probability density function and $h_1, \dots, h_d$ are bandwidths. Furthermore, $\widehat{p}_{X^j}$ is the kernel density estimator of the density p_{X^j} of X^j defined by

$$\widehat{p}_{X^j}(x^j) = \frac{1}{nh_j} \sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j} \right).$$

Below, we will outline that the smooth backfitting estimator can be calculated by an iterative backfitting algorithm. While the estimator got its name from the corresponding algorithm, it could, however, better be described as *smooth least squares estimator* highlighting its statistical motivation.

If there is only one additive component—that is, if we have $d = 1$—we get a kernel estimator $\widetilde{f}_1(x^1) = \widehat{\mu} + \widehat{f}_1(x^1)$ as the minimizer of

$$f_1 \rightsquigarrow \int \sum_{i=1}^{n} \left[Y_i - f_1(x^1) \right]^2 K\left(\frac{X_i^1 - x^1}{h_1} \right) dx^1. \qquad (7.4)$$

The minimizer of this criterion is given as

$$\widetilde{f}_1(x^1) = \left[\sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1} \right) \right]^{-1} \sum_{i=1}^{n} Y_i K\left(\frac{X_i^1 - x^1}{h_1} \right).$$

Thus, $\widetilde{f}_1(x^1)$ is just the classical Nadaraya–Watson estimator. We get the smooth backfitting estimator as a natural generalization of Nadaraya–Watson smoothing to additive models.

In this chapter we present a broad discussion of estimators based on minimizing a smoothed least squares criterion. We do this for two reasons. First, we argue that, even for additive models, this method is a powerful alternative to the two-step procedures that were extensively discussed in Chapters 5 and 6. Furthermore, smooth least squares estimators also work in models that are closely related to the additive model but are not of the form that is directly suitable for two-step estimation. We illustrate this with an example. Suppose that one observes (X_i, Y_i) with $Y_i = f(X_i) + \varepsilon_i$, where ε_i is a random walk: that is, $\eta_i = \varepsilon_{i+1} - \varepsilon_i$ are zero mean i.i.d. variables that are independent of $X_1, \dots, X_n$. In this model the Nadaraya–Watson estimator (7.4) is not consistent. Consistent estimators can be based on considering $Z_i = Y_{i+1} - Y_i$. For these variables we get the regression model

$$Z_i = f(X_{i+1}) - f(X_i) + \eta_i.$$

The smooth least squares estimator in this model is based on the minimization of

$$f \rightsquigarrow \int \sum_{i=1}^{n} [Z_i - f(x^1) + f(x^2)]^2 K\left(\frac{X_{i+1} - x^1}{h_1} \right) K\left(\frac{X_i - x^2}{h_2} \right) dx^1 dx^2.$$

Clearly, an alternative approach would be to calculate estimators $\widehat{f_1}$ and $\widehat{f_2}$ in the model $Z_i = f_1(X_{i+1}) - f_2(X_i) + \eta_i$ and to use $[\widehat{f_1}(x) - \widehat{f_2}(x)]/2$ as an estimator of f. We will come back to related models below.

The additive model is important for two reasons:

1. It is the simplest nonparametric regression model with several nonparametric components. The theoretical analysis is quite simple because the nonparametric components enter linearly into the model. Furthermore, the mathematical analysis can be built on localization arguments from classical smoothing theory. The simple structure allows for completely understanding of how the presence of additional terms influences estimation of each one of the nonparametric curves. This question is related to semiparametric efficiency in models with a parametric component and nonparametric nuisance components. We will come back to a short discussion of *nonparametric efficiency* below.
2. The additive model is also important for practical reasons. It efficiently avoids the curse of dimensionality of a full-dimensional nonparametric estimator. Nevertheless, it is a powerful and flexible model for high-dimensional data. Higher-dimensional structures can be well approximated by additive functions. As lower-dimensional curves they are also easier to visualize and hence to interpret than a higher-dimensional function.

Early references that highlight the advantages of additive modeling are Stone (1985, 1986), Buja, Hastie, and Tibshirani (1989), and Hastie and Tibshirani (1990). In this chapter we concentrate on the discussion of smooth backfitting estimators for such additive structures. For a discussion of two-step estimators we refer to Chapters 5 and 6. For sieve estimators in additive models, see Chen (2006) and the references therein. For the discussion of penalized splines we refer to Eilers and Marx (2002).

In this chapter we only discuss estimation of nonparametric components. Estimation of parametric components such as $\theta = \theta(f_1) = \int f_1(u)w(u)\ du$ for some given function w requires another type of analysis. In the latter estimation problem, natural questions are, for example, whether the plug-in estimator $\widehat{\theta} = \theta(\widehat{f_1}) = \int \widehat{f_1}(u)w(u)\ du$ for a nonparametric estimator $\widehat{f_1}$ of f_1 converges to θ at a parametric $\sqrt{n}$ rate, and whether this estimator achieves the semiparametric efficiency bound. Similar questions arise in related semiparametric models. An example is the partially linear additive model: $Y_i = \theta^\top Z_i + \mu + f_1(X_i^1) + \cdots + f_d(X_i^d) + \varepsilon_i$. Here, Z is an additional covariate vector. A semiparametric estimation problem arises when $\mu, f_1, \ldots, f_d$ are nuisance components and θ is the only parameter of interest. Then naturally the same questions as above arise when estimating θ. As said, such semiparametric considerations will not be in the focus of this chapter. For a detailed discussion of the specific example, we refer to Schick (1996) and Yu, Mammen, and Park (2011).

In this chapter, we concentrate on the description of estimation procedures. Smooth backfitting has also been used in testing problems by Haag (2006, 2008) and Lundervold, Tjøstheim, and Yao (2007). For related tests based on kernel smoothing, see also the overview article Fan and Jiang (2007). In Lundervold, Tjøstheim, and Yao (2007) additive models are used to approximate the distribution of spatial Markov random fields. The conditional expectation of the outcome of the random field at a point, given the outcomes in the neighborhood of the point, are modeled as sum of functions of the neighbored outcomes. They propose tests for testing this additive structure. They also discuss the behavior of smooth backfitting if the additive model is not correct. Their findings are also interesting for other applications where the additive model is not valid but can be used as a powerful approximation.

Another approach that will not be pursued here is parametrically guided nonparametrics. The idea is to fit a parametric model in a first step and then apply nonparametric smoothing in a second step, see Fan, Wu, and Feng (2009) for a description of the general idea. The original idea was suggested by Hjort and Glad (1995) in density estimation. See also Park, Kim, and Jones (2002) for a similar idea.

The next section discusses the smooth backfitting estimator in additive models. In Section 7.3 we discuss some models that are related to additive models. The examples include nonparametric regression with dependent error variables where the errors can be transformed to white noise by a linear transformation, nonparametric regression with repeatedly measured data, nonparametric panels with fixed effects, simultaneous nonparametric equation models, and non- and semiparametric autoregression and GARCH models. Other extensions that we will shortly mention are varying coefficient models and additive models with missing observations. In Section 7.4 we discuss the case of nonstationary covariates. Throughout the chapter we will see that many of the discussed models can be put in a form of noisy Fredholm integral equation of the second kind. We come back to this representation in the final section of this chapter. We show that this representation can be used as an alternative starting point for the calculation and also for an asymptotic understanding of smooth least squares estimators.

7.2. Smooth Least Squares Estimator in Additive Models

7.2.1. The Backfitting Algorithm

In the additive model (7.1) the smooth backfitting estimator can be calculated by an iterative algorithm. To see this, fix a value of x^1 and define $\widehat{\mu}_1 = \widehat{\mu} + \widehat{f}_1(x^1)$. One can easily see that $\widehat{\mu}_1$ minimizes

$$\mu_1 \rightsquigarrow \int \sum_{i=1}^{n} K\left(\frac{X_i^1 - x^1}{h_1}\right)\left[Y_i - \mu_1 - f_2(x^2) - \cdots - f_d(x^d)\right]^2$$
$$\times K\left(\frac{X_i^2 - x^2}{h_2}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) dx^2 \cdots dx^d. \tag{7.5}$$

This holds because we have no constraint on the function $x^1 \rightsquigarrow \widehat{\mu} + \widehat{f}_1(x^1)$. Thus we can minimize the criterion pointwise in this function and we do not integrate over the argument x^1 in (7.5). Thus, we get

$$\widehat{\mu}_1 = \left[\int \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_i^j - x^j}{h_j}\right) dx^2 \cdots dx^d\right]^{-1}$$
$$\times \int \sum_{i=1}^{n} \left[Y_i - f_2(x^2) - \cdots - f_d(x^d)\right] \prod_{j=1}^{d} K\left(\frac{X_i^j - x^j}{h_j}\right) dx^2 \cdots dx^d.$$

The expression on the right-hand side of this equation can be simplified by noting that $\int \frac{1}{h_j} K\left(\frac{X_i^j - x^j}{h_j}\right) dx^j = 1$ for $i = 1, \ldots, n; j = 1, \ldots, d$. We get

$$\widehat{\mu}_1 = \widehat{\mu} + \widehat{f}_1(x^1) = \widehat{f}_1^*(x^1) - \sum_{k=2}^{d} \int \frac{\widehat{p}_{X^1,X^k}(x^1, x^k)}{\widehat{p}_{X^1}(x^1)} \widehat{f}_k(x^k)\, dx^k. \tag{7.6}$$

Here, for $1 \le j \le d$ we define

$$\widehat{f}_j^*(x^j) = \left[\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i$$
$$= \widehat{p}_{X^j}(x^j)^{-1} \frac{1}{nh_j} \sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i.$$

This is the marginal Nadaraya–Watson estimator, based on smoothing the response Y_i versus one covariate X_i^j. Furthermore, $\widehat{p}_{X^j,X^k}$ is the two-dimensional kernel density estimator of the joint density p_{X^j,X^k} of two covariates X^j and X^k, defined for $1 \le j \ne k \le d$ by

$$\widehat{p}_{X^j,X^k}(x^j, x^k) = \frac{1}{nh_j h_k} \sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) K\left(\frac{X_i^k - x^k}{h_k}\right).$$

Similarly to Eq. (7.6), we get for all $j = 1, \ldots, d$ that

$$\widehat{f}_j(x^j) = \widehat{f}_j^*(x^j) - \widehat{\mu} - \sum_{k \ne j} \int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k(x^k)\, dx^k. \tag{7.7}$$

One can show that

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{7.8}$$

A proof of this equation is postponed to the end of this subsection.

We are now in the position to define the smooth backfitting algorithm. Our main ingredients are Eq. (7.7) and the formula for $\widehat{\mu}$. After an initialization step, the backfitting algorithm proceeds in cycles of d steps:

- **Initialization step:** Put $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\widehat{f}_j^{[0]}(x^j) \equiv 0$ for $j = 1,\ldots,d$.
- ***l*th iteration cycle:**
  - ***j*th step of the *l*th iteration cycle:** In the jth step of the lth iteration cycle, one updates the estimator $\widehat{f}_j$ of the jth additive component f_j

$$\widehat{f}_j^{[l]}(x^j) = \widehat{f}_j^{*}(x^j) - \widehat{\mu} - \sum_{k=1}^{j-1}\int \frac{\widehat{p}_{X^j,X^k}(x^j,x^k)}{\widehat{p}_{X^j}(x^j)}\widehat{f}_k^{[l]}(x^k)\,dx^k$$
$$-\sum_{k=j+1}^{d}\int \frac{\widehat{p}_{X^j,X^k}(x^j,x^k)}{\widehat{p}_{X^j}(x^j)}\widehat{f}_k^{[l-1]}(x^k)\,dx^k. \tag{7.9}$$

We now discuss some computational aspects of the smooth backfitting algorithm. One can show that there exist constants $C > 0$ and $0 < \gamma < 1$ that do not depend on n such that with probability tending to one

$$\int [\widehat{f}_j^{[l]}(x^j) - \widehat{f}_j(x^j)]^2 p_{X^j}(x^j)\,dx^j \le C\gamma^{2l}. \tag{7.10}$$

For a detailed statement, see Theorem 7.1 in Mammen, Linton, and Nielsen (1999) where a proof of (7.10) can be also found. The essential argument of the proof is that the approximation error $\sum_{j=1}^{d}[\widehat{f}_j^{[l]}(x^j) - \widehat{f}_j(x^j)]$ behaves like a function that is cyclically and iteratively projected onto d linear subspaces of a function space. Each cycle of projections reduces the norm of this function by a factor γ, for some fixed $\gamma < 1$, with probability tending to one.

The bound (7.10) allows for two important conclusions.

1. For a fixed accuracy, the number of iterations of the algorithm can be chosen as constant in n; in particular, it does not need to increase with n.
2. Furthermore, for an accuracy of order $n^{-\alpha}$ it suffices that the number of iterations increases with a logarithmic order. This implies, in particular, that the complexity of the algorithm does not explode but increases only slowly in n. We will see in the next subsection that for an optimal choice of bandwidth the rate of $\widehat{f}_j(x^j) - f_j(x^j)$ is of order $O_p(n^{-2/5})$. In that case, a choice of α with $\alpha > 2/5$ guarantees that the numerical error is of smaller order than the statistical error.

When numerically implementing smooth backfitting, estimators $\widehat{f}_j^{[l]}(x^j)$ are only calculated on a finite grid of points and integrals in (7.10) are replaced by discrete approximations. Suppose that the number of grid points is of order n^{β} for some $\beta > 0$. Then in the initialization step we have to calculate $n^{2\beta}$ two-dimensional kernel density estimators. This results in $O(n^{1+2\beta})$ calculations. Let us briefly discuss this for the case where all functions $f_j(x^j)$ have bounded support and all bandwidths are chosen so that $\widehat{f}_j(x^j) - f_j(x^j)$ is of order $O_p(n^{-2/5})$. It can be shown that one has to choose $\beta > 4/19$ to obtain a numerical error of smaller order than the statistical error. Then the computational complexity of the algorithm is of order $O(n\log(n) + n^{1+2\beta}) = O(n^{1+2\beta}) = O(n^{(27/19)+2\delta})$ with $\delta = \beta - \frac{4}{19}$. This amount of calculations can still be carried out even for large values of n in reasonable time.

Proof of (7.8). To get Eq. (7.8) we multiply both sides of Eq. (7.7) with $\widehat{p}_{X^j}(x^j)$ and integrate both sides of the resulting equation over x^j. Because of the norming (7.3), this yields

$$\begin{aligned}
0 &= \int \widehat{f}_j(x^j)\widehat{p}_{X^j}(x^j)\, dx^j \\
&= \int \widehat{f}_j^{*}(x^j)\widehat{p}_{X^j}(x^j)\, dx^j - \widehat{\mu}\int \widehat{p}_{X^j}(x^j)\, dx^j - \sum_{k\neq j}\int \widehat{p}_{X^j,X^k}(x^j,x^k)\widehat{f}_k(x^k)\, dx^k\, dx^j \\
&= \int \frac{1}{nh_j}\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i\, dx^j - \widehat{\mu} - \sum_{k\neq j}\int \widehat{p}_{X^k}(x^k)\widehat{f}_k(x^k)\, dx^k \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\mu},
\end{aligned}$$

where we use the facts that $\int \frac{1}{h_j}K\left(\frac{X_i^j - x^j}{h_j}\right)\, dx^j = 1$ and that $\int \widehat{p}_{X^j,X^k}(x^j,x^k)\, dx^j = \widehat{p}_{X^k}(x^k)$. This completes the proof. ■

7.2.2. Asymptotics of the Smooth Backfitting Estimator

Under appropriate conditions, the following result holds for the asymptotic distribution of each component function $\widehat{f}_j(x^j)$, $j = 1, \ldots, d$:

$$\sqrt{nh_j}\Big(\widehat{f}_j(x^j) - f_j(x^j) - \beta_j(x^j)\Big) \xrightarrow{d} N\left(0, \int K^2(u)\, du\, \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right). \qquad (7.11)$$

Here the asymptotic bias terms $\beta_j(x^j)$ are defined as minimizers of

$$(\beta_1, \ldots, \beta_d) \rightsquigarrow \int [\beta(x) - \beta_1(x^1) - \cdots - \beta_d(x^d)]^2 p_X(x)\, dx$$

under the constraint that

$$\int \beta_j(x^j) p_{X^j}(x^j)\, dx^j = \frac{1}{2} h_j^2 \int [2 f_j'(x^j) p_{X^j}'(x^j) + f_j''(x^j) p_{X^j}(x^j)]\, dx^j \times \int u^2 K(u)\, du, \tag{7.12}$$

where p_X is the joint density of $X = (X^1, \ldots, X^d)$ and

$$\beta(x) = \frac{1}{2} \sum_{j=1}^{d} h_j^2 \left[2 f_j'(x^j) \frac{\partial \log p_X}{\partial x^j}(x) + f_j''(x^j) \right] \int u^2 K(u)\, du.$$

In Mammen, Linton, and Nielsen (1999) and Mammen and Park (2005) this asymptotic statement has been proved for the case that f_j is estimated on a compact interval I_j. The conditions include a boundary modification of the kernel. Specifically, the convolution kernel $h_j^{-1} K(h_j^{-1}(X_i^j - x^j))$ is replaced by $K_{h_j}(X_i^j, x^j) = h_j^{-1} K(h_j^{-1}(X_i^j - x^j)) / \int_{I_j} h_j^{-1} K(h_j^{-1}(X_i^j - u^j))\, du^j$. Then it holds that $\int_{I_j} K_{h_j}(X_i^j, x^j)\, dx^j = 1$. In particular, this implies $\int_{I_j} \widehat{p}_{X^j, X^k}(x^j, x^k)\, dx^j = \widehat{p}_{X^k}(x^k)$ and $\int_{I_j} \widehat{p}_{X^j}(x^j)\, dx^j = 1$ if one replaces $h_j^{-1} K(h_j^{-1}(X_i^j - x^j))$ by $K_{h_j}(X_i^j, x^j)$ in the definitions of the kernel density estimators. In fact, we have already made use of these properties of kernel density estimators in the previous subsection.

Before illustrating how the asymptotic result (7.11) is obtained, we discuss its interpretations. In particular, it is illustrative to compare $\widehat{f}_j$ with the Nadaraya–Watson estimator $\widetilde{f}_j$ in the classical nonparametric regression model $Y_i = f_j(X_i^j) + \varepsilon_i$. Under standard smoothness assumptions, it holds that

$$\sqrt{nh_j}\left(\widetilde{f}_j(x^j) - f_j(x^j) - \beta_j^*(x^j)\right) \xrightarrow{d} N\left(0, \int K^2(u)\, du\, \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right) \tag{7.13}$$

with the asymptotic bias $\beta_j^*(x^j) = \frac{1}{2} h_j^2 [2 f_j'(x^j)(\partial \log p_{X^j}(x^j))/(\partial x^j) + f_j''(x^j)] \int u^2 K(u)\, du$. We see that $\widetilde{f}_j(x^j)$ has the same asymptotic variance as $\widehat{f}_j(x^j)$ but that the two estimators differ in their asymptotic bias. Thus, as long as one only considers the asymptotic variance, one does not have to pay any price for not knowing the other additive components f_k $(k \neq j)$. One gets the same asymptotic variance in the additive model as in the simplified model $Y_i = f_j(X_i^j) + \varepsilon_i$ where all other additive components f_k $(k \neq j)$ are set equal to 0. As said, the bias terms differ. The asymptotic bias of $\widehat{f}_j(x^j)$ may be larger or smaller than that of $\widetilde{f}_j(x^j)$. This depends on the local characteristics of the function f_j at the point x^j and also on the global shape of the other functions f_k $(k \neq j)$. It is a disadvantage of the Nadaraya–Watson smooth backfitting estimator. There may be structures in $\widehat{f}_j(x^j)$ that are caused by other functions. We will argue below that this is not the case for the local linear smooth backfitting estimator. For the local linear smooth backfitting estimator, one gets the same asymptotic bias and variance as for the local linear estimator in the classical model $Y_i = f_j(X_i^j) + \varepsilon_i$. In particular, both estimators have the same asymptotic normal distribution. In Chapter 6 this was called oracle efficiency. This notion of efficiency is appropriate for nonparametric models. Typically in nonparametric models there exists no asymptotically optimal estimator, in contrast to parametric models and to the case of estimating the parametric parts of semiparametric models.

We now come to a heuristic explanation of the asymptotic result (7.11). For a detailed proof we refer to Mammen, Linton, and Nielsen (1999) and Mammen and Park (2005). The main argument is based on a decomposition of the estimator into a *mean part* and a *variance part*. For this purpose, one applies smooth backfitting to the "data" $(X^1,\ldots,X^d,f_1(X^1)+\cdots+f_d(X^d))$ and to $(X^1,\ldots,X^d,\varepsilon)$. We will argue below that $\widehat{f}_j(x^j)$ is the sum of these two estimators.

Justification of (7.11). We start with a heuristic derivation of the asymptotic bias and variance of the smooth backfitting estimator $\widehat{f}_j(x^j)$. For this purpose note first that the smooth backfitting estimators $\widehat{\mu},\widehat{f}_1,\ldots,\widehat{f}_d$ are the minimizers of

$$(\mu,f_1,\ldots,f_d) \rightsquigarrow \int [\widehat{f}(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\,dx \qquad (7.14)$$

under the constraint (7.3), where $\widehat{p}_X$ is the kernel density estimator of p_X and $\widehat{f}$ is the Nadaraya–Watson estimator of the regression function $f(x) = E(Y|X=x)$:

$$\widehat{p}_X(x) = \frac{1}{nh_1\cdots h_d}\sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right),$$

$$\widehat{f}(x) = \widehat{p}_X(x)^{-1}\frac{1}{nh_1\cdots h_d}\sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) Y_i.$$

One may show that this minimization problem leads to (7.7) and (7.8). We omit the details. For a geometric argument see also Mammen, Marron, Turlach, and Wand (2001).

For heuristics on the asymptotics of $\widehat{f}_j$, $1 \le j \le d$, we now decompose $\widehat{f}$ into its bias and variance component $\widehat{f}(x) = \widehat{f}^A(x) + \widehat{f}^B(x)$, where

$$\widehat{f}^A(x) = \widehat{p}_X(x)^{-1}\frac{1}{nh_1\cdots h_d}\sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right)\varepsilon^i,$$

$$\widehat{f}^B(x) = \widehat{p}_X(x)^{-1}\frac{1}{nh_1\cdots h_d}\sum_{i=1}^n K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right)$$
$$\times[\mu + f_1(x^1) + \cdots + f_d(x^d)].$$

Denote by $(\widehat{\mu}^A, \widehat{f}_1^A, \ldots, \widehat{f}_d^A)$ the minimizer of

$$(\mu, f_1, \ldots, f_d) \rightsquigarrow \int [\widehat{f}^A(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\, dx$$

under the constraint (7.3), and by $(\widehat{\mu}^B, \widehat{f}_1^B, \ldots, \widehat{f}_d^B)$ the minimizer of

$$(\mu, f_1, \ldots, f_d) \rightsquigarrow \int [\widehat{f}^B(x) - \mu - f_1(x^1) - \cdots - f_d(x^d)]^2 \widehat{p}_X(x)\, dx$$

under the constraint (7.3). Then, we obtain $\widehat{\mu} = \widehat{\mu}^A + \widehat{\mu}^B, \widehat{f}_1 = \widehat{f}_1^A + \widehat{f}_1^B, \ldots, \widehat{f}_d = \widehat{f}_d^A + \widehat{f}_d^B$. By standard smoothing theory, $\widehat{f}^B(x) \approx \mu + f_1(x^1) + \cdots + f_d(x^d) + \beta(x)$. This immediately implies that $\widehat{f}_j^B(x^j) \approx c_j + f_j(x^j) + \beta_j(x^j)$ with a random constant c_j. Our constraint (7.12) implies that c_j can be chosen equal to zero. This follows by some more lengthy arguments that we omit.

For an understanding of the asymptotic result (7.11), it remains to show that

$$\sqrt{nh_j}\Big(\widehat{f}_j^A(x^j) - f_j(x^j)\Big) \xrightarrow{d} N\left(0, \int K^2(u)\, du\, \frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right). \tag{7.15}$$

To see this claim, we proceed similarly as in the derivation of (7.7). Using essentially the same arguments as there, one can show that

$$\widehat{f}_j^A(x^j) = \widehat{f}_j^{A,*}(x^j) - \widehat{\mu}^A - \sum_{k \neq j} \int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^A(x^k)\, dx^k, \tag{7.16}$$

where

$$\widehat{f}_j^{A,*}(x^j) = \left[\sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i$$

is the stochastic part of the marginal Nadaraya–Watson estimator $\widehat{f}_j^*(x^j)$. We now argue that

$$\int \frac{\widehat{p}_{X^j,X^k}(x^j, x^k)}{\widehat{p}_{X^j}(x^j)} \widehat{f}_k^A(x^k)\, dx^k \approx \int \frac{p_{X^j,X^k}(x^j, x^k)}{p_{X^j}(x^j)} \widehat{f}_k^A(x^k)\, dx^k \approx 0.$$

The basic argument for the second approximation is that a global average of a local average behaves like a global average; or, more explicitly, consider, for example, the local average $\widehat{r}_j(x^j) = (nh_j)^{-1} \sum_{i=1}^n K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i$. This local average is of order $O_p(n^{-1/2} h_j^{-1/2})$. For a smooth weight function w we now consider the global average $\widehat{\rho}_j = \int_{I_j} w(x^j)\widehat{r}_j(x^j)\, dx^j$ of the local average $\widehat{r}_j(x^j)$. This average is of order

$O_p(n^{-1/2}) = o_p(n^{-1/2}h_j^{-1/2})$ because of

$$\begin{aligned}\widehat{\rho}_j &= \int_{I_j} w(x^j)\widehat{r}_j(x^j)\,dx^j \\ &= \int_{I_j} w(x^j)(nh_j)^{-1}\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right)\varepsilon_i\,dx^j \\ &= n^{-1}\sum_{i=1}^{n} w_{h_j}(X_i^j)\varepsilon_i\end{aligned}$$

with $w_{h_j}(X_i^j) = \int_{I_j} w(x^j)h_j^{-1}K\left(\frac{X_i^j - x^j}{h_j}\right)dx^j$.

7.2.3. Smooth Backfitting Local Linear Estimator

In the additive model (7.1) the smooth backfitting local linear estimators $\widehat{\mu}, \widehat{f}_1, \widehat{f}_1^\dagger, \ldots, \widehat{f}_d, \widehat{f}_d^\dagger$ are defined as the minimizers of the smoothed least squares criterion

$$\int \sum_{i=1}^{n}\Big[Y_i - \mu - f_1(x^1) - f_1^\dagger(x^1)(X_i^1 - x^1) - \cdots - f_d(x^d) - f_d^\dagger(x^d)(X_i^d - x^d)\Big]^2 \\ \times K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) dx^1 \cdots dx^d \tag{7.17}$$

under the constraint (7.3). This is a natural generalization of the local linear estimator. For the case $d = 1$ the minimization gives the classical local linear estimator as the minimization of (7.4) leads to the classical Nadaraya–Watson estimator. The estimators, $\widehat{f}_j^\dagger$, $1 \le j \le d$, are estimators of the derivatives of the additive components f_j.

The smooth backfitting local linear estimator is given as the solution of a random integral equation. Similarly to Eq. (7.7), the tuples $(\widehat{f}_j, \widehat{f}_j^\dagger)$ fulfill now a two-dimensional integral equation. This integral equation can be used for the iterative calculation of the estimators. For details we refer to Mammen, Linton, and Nielsen (1999). We only mention the following asymptotic result from Mammen, Linton, and Nielsen (1999) for the smooth backfitting local linear estimator. Under appropriate conditions it holds that for $1 \le j \le d$

$$\sqrt{nh_j}\Big(\widehat{f}_j(x^j) - f_j(x^j) - \beta_j(x^j)\Big) \xrightarrow{d} N\left(0, \int K^2(u)\,du\,\frac{\sigma_j^2(x^j)}{p_{X^j}(x^j)}\right), \tag{7.18}$$

where now the asymptotic bias terms $\beta_j(x^j)$ are defined as

$$\beta_j(x^j) = \frac{1}{2}h_j^2\left[f_j''(x^j) - \int f_j''(u^j)p_{X^j}(u^j)\,du^j\right]\int u^2K(u)\,du.$$

Up to an additive norming term, the asymptotic bias of $\widehat{f}_j(x^j)$ coincides with the asymptotic bias of local linear estimator $\widetilde{f}_j$ in the classical nonparametric regression model $Y_i = f_j(X_i^j) + \varepsilon_i$. Moreover, we get the same asymptotic distribution for both estimators (up to an additive norming term). Asymptotically, one does not lose any efficiency by not knowing the additive components $f_k : k \neq j$ compared to the *oracle model* where these components are known. This is an asymptotic optimality result for the local linear smooth backfitting. It achieves the same asymptotic bias and variance as in the oracle model. As discussed above, the Nadaraya–Watson smooth backfitting estimator achieves only the asymptotic variance of the oracle model. For an alternative implementation of local linear smooth backfitting, see Lee, Mammen, and Park (2012b).

7.2.4. Smooth Backfitting as Solution of a Noisy Integral Equation

We write the smooth backfitting estimators as solutions of an integral equation. We discuss this briefly for Nadaraya–Watson smoothing. Put $\widehat{\mathbf{f}}(x^1,\ldots,x^d) = (\widehat{f}_1(x^1),\ldots,\widehat{f}_d(x^d))^\top$ and $\widehat{\mathbf{f}}^*(x^1,\ldots,x^d) = (\widehat{f}_1^*(x^1),\ldots,\widehat{f}_d^*(x^d))^\top$. With this notation and taking, $\widehat{\mu} = 0$, we can rewrite (7.7) as

$$\widehat{\mathbf{f}}(x) = \widehat{\mathbf{f}}^*(x) - \int \widehat{\mathcal{H}}(x,z)\widehat{\mathbf{f}}(z)\,dz, \tag{7.19}$$

where for each value of $x, z \in \mathbb{R}$ the integral kernel $\widehat{\mathcal{H}}(x,z)$ is a matrix with element (j,k) equal to $\widehat{p}_{X^j,X^k}(x^j,x^k)/\widehat{p}_{X^j}(x^j)$. This representation motivates an alternative algorithm. One can use a discrete approximation of the integral equation and approximate the integral equation (7.19) by a finite linear equation. This can be solved by standard methods of linear algebra. Equation (7.19) can also be used as an alternative starting point for an asymptotic analysis of the estimator $\widehat{\mathbf{f}}$. We will come back to this in Section 7.5 after having discussed further those models in Section 7.3 whose estimation can be formulated as solving an integral equation.

7.2.5. Relations to Classical Backfitting and Two-Stage Estimation

Smooth backfitting (7.9) is related to classical backfitting and to two-stage estimation. In the classical backfitting, the jth step of the lth iteration cycle (7.9) of the smooth backfitting is replaced by

$$\widehat{f}_j^{[l]}(X_i^j) = \widehat{p}_{X^j}(x^j)^{-1}\frac{1}{nh_j}\sum_{i=1}^{n} K\left(\frac{X_i^j - x^j}{h_j}\right)\left[Y_i - \widehat{\mu} - \sum_{k=1}^{j-1}\widehat{f}_k^{[l]}(X_i^k) - \sum_{k=j+1}^{d}\widehat{f}_k^{[l-1]}(X_i^k)\right] \tag{7.20}$$

for $1 \le j \le d$ and $1 \le i \le n$. This iteration equation can be interpreted as a limiting case of (7.9) where one lets the second bandwidth h_k in the definition of the kernel density estimator $\widehat{p}_{X^j,X^k}(x^j, x^k)$ converge to zero.

If the backfitting algorithm runs through $O(\log n)$ cycles, the algorithm needs $O(n \log n)$ calculation steps. This is slightly faster than the smooth backfitting. In contrast to the smooth backfitting, the backfitting estimator is only defined as the limit of the iterative algorithm (7.20). Note that the smooth backfitting is explicitly defined as minimizer of the smoothed least squares criterion (7.2). The fact that backfitting estimators are only implicitly defined as limit of an iterative algorithm complicates the asymptotic mathematical analysis. Note also that the algorithm runs in $\mathbb{R}^n$—that is, in spaces with increasing dimension. An asymptotic treatment of the classical backfitting can be found in Opsomer (2000) and Opsomer and Ruppert (1997). Nielsen and Sperlich (2005) illustrated by simulation that smooth backfitting, in comparison with the classical backfitting, is more robust against degenerated designs and a large number of additive components. The reason behind this is that the iteration equation (7.9) is a smoothed version of (7.20). The smoothing stabilizes the “degenerated integral equation” (7.20). In Opsomer (2000) and Opsomer and Ruppert (1997), stronger assumptions are made on the joint density of the covariates than are needed for the study of the smooth backfitting. This may be caused by the same reasons, but there has been made no direct theoretical argument that supports the empirical finding that the classical backfitting is more sensitive to degenerate designs than smooth backfitting. For another modification of the classical backfitting that takes care of correlated covariates, see Jiang, Fan, and Fan (2010).

Two-stage estimation differs from smooth backfitting in several respects. First of all, only two steps are used instead of an iterative algorithm that runs until a convergence criterion is fulfilled. Furthermore, different bandwidths are used in different steps: Undersmoothing is done in the first step, but an optimal bandwidth is chosen in the second step. The algorithm of two-step estimation is as simple as that of backfitting. On the other hand, choice of the bandwidth in the first-step is rather complex. Asymptotically, optimal choices will not affect the first-order properties of the outcomes of the second step. But for finite samples the influence of the first-step bandwidth is not clear. The calculation of theoretically optimal values would require a second-order optimal theory that is not available and, as with other higher-order theory, may not be accurate for small to moderate sample sizes. In particular, in models with many nonparametric components, backfitting may be preferable because it does not require an undersmoothing step.

Another kernel smoothing method that can be applied to additive models is marginal integration. It was discussed in Chapter 5 that marginal integration only achieves optimal rates for low-dimensional additive models but that it does not work in higher-dimensional models. This drawback is not shared by backfitting, smooth backfitting, and two-stage estimation. There is also another aspect in which smooth backfitting and marginal integration differ. If the additive model is not correct, smooth backfitting as a weighted least squares estimator estimates the best additive fit to the non-additive model. On the other side, marginal integration estimates a weighted average effect for each covariate. This follows because marginal integration is based on a weighted average of the full-dimensional regression function. Thus, the methods estimate quite different quantities if the model is not additive.

7.2.6. Bandwidth Choice and Model Selection

Bandwidth selection for additive models has been discussed in Mammen and Park (2005). There, consistency has been shown for bandwidth selectors based on plug-in and penalized least squares criteria. Nielsen and Sperlich (2005) discusses practical implementations of cross-validation methods. Because an additive model contains several nonparametric functions, there exist two types of optimal bandwidths: bandwidths that are optimal for the estimation of the sum of the additive components and bandwidths that optimize estimation of a single additive component. While the former criterion in particular may be appropriate in prediction, the latter is more motivated in data analytic-oriented inference. Whereas all three-bandwidth selectors (cross-validation, penalized least squares, and plug-in) can be designed for the former criterion, only plug-in based approaches can be used. For a further discussion we refer to the two papers cited above. For the models that will be discussed in the next section, bandwidth selection has been only partially studied. The asymptotic results for the estimators that will be discussed can be used to design plug-in methods. For cross-validation it is questionable if for all models algorithms can be found that run in reasonable time.

In very-high-dimensional additive models, backfitting methods will suffer from the complexity of the models, in statistical performance and in computational costs. For this reason, component selection is an important step to control the size of the model. Recently, some proposals have been made that are influenced by the study of high-dimensional models with sparsity constraints. We refer to Lin and Zhang (2006), Meier, van de Geer, and Bühlmann (2009), and Huang, Horowitz, and Wei (2010).

7.2.7. Generalized Additive Models

We now discuss nonlinear extensions of the additive models. In a generalized additive model a link function g is introduced and it is assumed that the following equation holds for the regression function $E(Y|X_1,\ldots,X_d)$:

$$E(Y|X^1,\ldots,X^d) = g^{-1}\{\mu + f_1(X^1) + \cdots + f_d(X^d)\}.$$

It has been considered that the link function is known or that it is unknown and has to be estimated. An important example where generalized additive models make sense is the case of binary responses Y. If Y is $\{0,1\}$-valued, the function g^{-1} maps the additive function onto the interval $[0,1]$. In the generalized additive model, the additive functions $f_1,\ldots,f_d$ can be estimated by smoothed least squares. An alternative approach for heterogeneous errors is a smoothed quasi-likelihood criterion. Quasi-likelihood is motivated for regression models where the conditional variance of the errors is equal to $V(\mu)$ with μ equal to the conditional expectation of Y. Here, V is a specified variance function. Quasi-likelihood coincides with classical likelihood if the conditional error distribution is an exponential family. It also leads to consistent estimators if the conditional variances have another form. The quasi-likelihood criterion $Q(\mu,y)$ is defined as

$$\frac{\partial}{\partial\mu}Q(\mu,y) = \frac{y-\mu}{V(\mu)}.$$

An early reference to quasi-likelihood approaches in additive models is Hastie and Tibshirani (1990). For the discussion of local linear smoothing in generalized partially linear models see also Fan, Heckman, and Wand (1995). For a discussion of the asymptotics of classical backfitting in the generalized additive model, see Kauermann and Opsomer (2003). The Smoothed Quasi-Likelihood criterion is defined as follows: Minimize for $\mathbf{f} = (\mu,f_1,\ldots,f_d)^\top$

$$SQ(\mathbf{f}) = \int \sum_{i=1}^{n} Q(g^{-1}(f^+(x)), Y_i) K\left(\frac{X_i^1 - x^1}{h_1}\right) \times \cdots \times K\left(\frac{X_i^d - x^d}{h_d}\right) dx^1 \cdots dx^d.$$

where $f^+(x) = \mu + f_1(x^1) + \cdots + f_d(x^d)$. Minimization of the smoothed quasi-likelihood criterion over $\mathbf{f}$ results in the smoothed maximum quasi-liklihood estimator. Algorithms for the calculation of this estimator were discussed in Yu, Park, and Mammen (2008). In that paper an asymptotic theory for this estimator was also developed. In other applications the quasi-likelihood criterion may be replaced by other M-functionals. We do not discuss this here. An example is quantile regression. For a discussion of backfitting and smooth backfitting in additive quantile models, see Lee, Mammen, and Park (2010).

7.3. Some Models That are Related to Additive Models

In linear regression, the standard least squares method produces consistent estimators when the errors are uncorrelated. When the errors are correlated, the method may not give consistent or efficient estimators of the regression parameters. In the latter case it is often appropriate to take a linear transformation of the response variable in such a way that it corrects for the correlations between the errors. Linear transformations may be also used to remove some unobserved effects in a regression model that are correlated with the regressors or errors. Taking a linear transformation in parametric linear models does not alter the linear structure of the model, so that conventional methods still work with the transformed data. In nonparametric regression models, however, it often yields an additive model where classical smoothing methods cannot be applied, as we illustrate on several cases in this section. Some of the models of this section were also discussed in the overview papers Linton and Mammen (2003) and Mammen and Yu (2009). A general discussion of smooth least squares in a general class of nonparametric models can also be found in Mammen and Nielsen (2003).

7.3.1. Nonparametric Regression with Time Series Errors

Suppose we observe (X_t, Y_t) for $1 \le t \le T$ such that $Y_t = f(X_t) + u_t$, where the errors u_t have an AR(1) time series structure so that $\varepsilon_t = u_t - \rho u_{t-1}$ is a sequence of uncorrelated errors. The transformed model $Z_t(\rho) \equiv Y_t - \rho Y_{t-1} = f(X_t) - \rho f(X_{t-1}) + \varepsilon_t$ has uncorrelated errors, but has an additive structure in the mean function. For simplicity, assume that the errors u_t are independent of the covariates X_t. Then, the target function f minimizes

$$Q_T(m) = \frac{1}{T}\sum_{t=1}^{T} E[Z_t(\rho) - m(X_t) + \rho m(X_{t-1})]^2$$

over m, so that it satisfies

$$\int \big[E(Z_t(\rho)|X_t = x, X_{t-1} = y) - f(x) + \rho f(y)\big]\big[g(x) - \rho g(y)\big] f_{0,1}(x,y)\, dx\, dy = 0 \tag{7.21}$$

for all square integrable functions g. Here $f_{0,1}$ denotes the joint density of (X_t, X_{t-1}) and f_0 is the density of X_t. Equation (7.21) holds for all square integrable functions g if and only if

$$f(x) = f_\rho^*(x) - \int \mathcal{H}_\rho(x,y) f(y)\, dy \tag{7.22}$$

where

$$f_\rho^*(x) = \frac{1}{1+\rho^2}[E(Z_t(\rho)|X_t = x) - \rho E(Z_t(\rho)|X_{t-1} = x)],$$

$$\mathcal{H}_\rho(x,y) = -\frac{\rho}{1+\rho^2}\left[\frac{f_{0,1}(x,y)}{f_0(x)} + \frac{f_{0,1}(y,x)}{f_0(x)}\right].$$

An empirical version of the integral equation (7.22) may be obtained by estimating f_0, $f_{0,1}$, $E(Z_t(\rho)|X_t = \cdot)$ and $E(Z_t(\rho)|X_{t-1} = \cdot)$. Let $\widehat{f}(\cdot,\rho)$ denote the solution of the latter integral equation. In case ρ is known, $\widehat{f}(\cdot,\rho)$ can be used as an estimator of f. Otherwise, the parameter ρ can be estimated by $\widehat{\rho}$ that minimizes

$$\frac{1}{T}\sum_{t=1}^{T}\big[Z_t(\rho) - \widehat{f}(X_t,\rho) + \rho\widehat{f}(X_{t-1},\rho)\big]^2,$$

and then f by $\widehat{f} = \widehat{f}(\cdot,\widehat{\rho})$. We note that the estimator $\widehat{f}(\cdot,\rho)$ is consistent even if the autoregressive coefficient ρ is 1. In contrast, smoothing of the original untransformed data (Y_t, X_t) leads to an inconsistent estimator. We mentioned this example already in the introduction.

The above discussion may be extended to a general setting where the errors u_t admit a time series structure such that $\varepsilon_t = \sum_{j=0}^{\infty} a_j u_{t-j}$ is a sequence of uncorrelated errors. In this general case, if we take the transformation $Z_t(a_0, a_1, \ldots) = \sum_{j=0}^{\infty} a_j Y_{t-j}$, then the transformed model $Z_t(a_0, a_1, \ldots) = \sum_{j=0}^{\infty} a_j f(X_{t-j}) + \varepsilon_t$ has an additive structure with uncorrelated errors. For a discussion of this general case, see Linton and Mammen (2008). There weaker assumptions are made on the errors u_t. In particular, it is not assumed that the errors u_t are independent of the covariates X_t.

7.3.2. Nonparametric Regression with Repeated Measurements

Suppose that one has J repeated measurements on each of n subjects. Let (X_{ij}, Y_{ij}) be the jth observation on the ith subject. Write $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})^\top$ and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^\top$. Assume that $(\mathbf{X}_i, \mathbf{Y}_i), i = 1\ldots, n$, are i.i.d. copies of $(\mathbf{X}, \mathbf{Y})$. Consider the simple nonparametric regression model

$$Y_{ij} = f(X_{ij}) + \epsilon_{ij}, \tag{7.23}$$

where the errors ϵ_{ij} have zero conditional mean, but are allowed to be correlated within each subject. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iJ})^\top$ and $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{\epsilon}_i)$. The kernel regression estimator based on the ordinary least squares criterion is consistent even in this case where $\boldsymbol{\Sigma}$ is not the identity matrix. However, we may find a better estimator that is based on a weighted least squares criterion. This is in line with parametric linear regression with repeated measurements, where a weighted least squares estimator outperforms the ordinary least squares estimator. A weighted least squares estimation is equivalent to taking a linear transformation of the response and then applying the ordinary least squares criterion to the transformed model. In contrast to the parametric case, introducing weights in the nonparametric model (7.23) leads to a more complicated estimation problem, as is demonstrated below.

Let $\mathrm{f}(x_1, \ldots, x_J) = (f(x_1), \ldots, f(x_J))^\top$. The regression function f at (7.23) minimizes

$$E[\{\mathbf{Y} - \mathbf{m}(X_1, \ldots, X_J)\}^\top \boldsymbol{\Sigma}^{-1}\{\mathbf{Y} - \mathbf{m}(X_1, \ldots, X_J)\}] \tag{7.24}$$

over all square integrable functions m, where $\mathbf{m}(x_1,\ldots,x_J) = (m(x_1),\ldots,m(x_J))^\top$. Note that the transformed response vector $\mathbf{\Sigma}^{-1/2}\mathbf{Y}$ admits an additive model and the variance of the transformed error vector $\mathbf{\Sigma}^{-1/2}\boldsymbol{\epsilon}$ equals the identity matrix. The minimizer f satisfies

$$\sum_{j=1}^{J}\sum_{k=1}^{J}\sigma^{jk}E\{[Y_j - f(X_j)]g(X_k)\} = 0$$

for all square integrable functions g, where σ^{jk} denotes the (j,k)th entry of the matrix $\mathbf{\Sigma}^{-1}$. This gives the following integral equation for f;

$$f(x) = f^*(x) - \int \mathcal{H}(x,z)f(z)\,dz, \tag{7.25}$$

where

$$f^*(x) = \left[\sum_{j=1}^{J}\sigma^{jj}p_j(x)\right]^{-1}\sum_{j=1}^{J}\sum_{k=1}^{J}\sigma^{jk}E(Y_k|X_j = x)p_j(x),$$

$$\mathcal{H}(x,z) = \left[\sum_{j=1}^{J}\sigma^{jj}p_j(x)\right]^{-1}\sum_{j=1}^{J}\sum_{k\neq j}^{J}\sigma^{jk}p_{jk}(x,z).$$

Here, p_j and p_{jk} denote the densities of X_j and (X_j,X_k), respectively. The quantities f^*, p_j, and p_{jk} can be estimated by the standard kernel smoothing techniques. Plugging these into (7.25) gives an integral equation for estimating f.

One may apply other weighting schemes replacing $\mathbf{\Sigma}^{-1}$ at (7.24) by a weight matrix $\mathbf{W}$. It can be shown the choice $\mathbf{W} = \mathbf{\Sigma}^{-1}$ leads to an estimator with the minimal variance, see Carroll, Maity, Mammen, and Yu (2009) for details. The foregoing weighted least squares regression may be extended to the additive regression model $Y_{ij} = \sum_{d=1}^{D} f_d(X_{ij}^d) + \epsilon_{ij}$ with covariates $\mathbf{X}_{ij} = (X_{ij}^1,\ldots,X_{ij}^D)^\top$. Details are also given in Carroll, Maity, Mammen, and Yu (2009).

7.3.3. Panels with Individual Effects

Suppose we have panel data (X_{ij}, Y_{ij}) for $i = 1,\ldots,n$ and $j = 1,\ldots,J$. We assume that

$$Y_{ij} = f(X_{ij}) + \alpha_i + \epsilon_{ij}, \tag{7.26}$$

where α_i are the unobserved random or nonrandom individual effects that are invariant over time j, and ϵ_{ij} are errors such that $E(\epsilon_{ij}|X_{i1},\ldots,X_{iJ}) = 0$. The individual effect α_i can be uncorrelated or correlated with the regressors $X_{i1},\ldots,X_{iJ}$ and the error variables ϵ_{ij}. If $E(\alpha_i|X_{i1},\ldots,X_{iJ}) = 0$, then the model reduces to the model considered in

Subsection 7.3.2. An interesting case is when the individual effect is correlated with the regressors so that $E(\alpha_i|X_{i1},\dots,X_{iJ}) \neq 0$. In this case, the ordinary nonparametric kernel regression fails to obtain a consistent estimator. Recall that the latter is also the case with parametric linear regression.

Here again, we may use a simple linear transformation to remove the unobserved individual effect from the regression model. Let $Z_i = \sum_{j=1}^J a_j Y_{ij}$ for some constants a_j such that $\sum_{j=1}^J a_j = 0$. Examples include

(i) $a_1 = \cdots = a_{k-2} = 0$, $a_{k-1} = -1$, $a_k = 1$, $a_{k+1} = \cdots = a_J = 0$ for some $1 \leq k \leq J$;
(ii) $a_1 = \cdots = a_{k-1} = -J^{-1}$, $a_k = 1 - J^{-1}$, $a_{k+1} = \cdots = a_J = -J^{-1}$ for some $1 \leq k \leq J$.

For the transformed response variables Z_i, we obtain

$$Z_i = \sum_{j=1}^J a_j f(X_{ij}) + u_i, \tag{7.27}$$

where $u_i = \sum_{j=1}^J a_j \epsilon_{ij}$ has zero conditional mean given $X_{i1},\dots,X_{iJ}$. Let Z and X_j denote the generics of Z_i and X_{ij}, respectively. Since f minimizes the squared error risk $E[Z - \sum_{j=1}^J a_j m(X_j)]^2$ over m, it satisfies

$$E\left\{\left[Z - \sum_{j=1}^J a_j f(X_j)\right] \sum_{j=1}^J a_j g(X_j)\right\} = 0 \tag{7.28}$$

for all square integrable functions g. Equation (7.28) is equivalent to

$$\int \left[\sum_{j=1}^J a_j E(Z|X_j = x) p_j(x) - \sum_{j=1}^J \sum_{k \neq j}^J a_j a_k E[f(X_k)|X_j = x] p_j(x) - f(x) \sum_{j=1}^J a_j^2 p_j(x)\right]$$
$$\times\, g(x)\, dx = 0,$$

where p_j and p_{jk} denote the density of X_j and (X_j, X_k), respectively. This gives the following integral equation

$$f(x) = f^*(x) - \int \mathcal{H}(x,z) f(z)\, dz, \tag{7.29}$$

where

$$f^*(x) = \left[\sum_{j=1}^J a_j^2 p_j(x)\right]^{-1} \sum_{j=1}^J a_j E(Z|X_j = x) p_j(x),$$

$$\mathcal{H}(x,z) = \left[\sum_{j=1}^J a_j^2 p_j(x)\right]^{-1} \sum_{j=1}^J \sum_{k \neq j}^J a_j a_k p_{jk}(x,z).$$

As in the additive regression model, we need a norming condition for identification of f in the transformed model (7.27). The reason is that in the transformed model we have $\sum_{j=1}^{J} a_j f(X_{ij}) = \sum_{j=1}^{J} a_j [c + f(X_{ij})]$ for any constant c since $\sum_{j=1}^{J} a_j = 0$. We may also see this from the integral equation (7.29) since $\int \mathcal{H}(x,z)\,dz = -1$. For a norming condition, we may define α_i such that $E(Y_{ij}) = Ef(X_{ij})$. This motivates the normalizing constraint

$$J^{-1}\sum_{j=1}^{J}\int \widehat{f}(x)\widehat{p}_j(x)\,dx = n^{-1}J^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}$$

for an estimator $\widehat{f}$ of f. For a kernel estimator based on differencing, see also Henderson, Carroll, and Li (2008).

The differencing technique we have discussed above may also be applied to a more general setting that allows for discrete response variables. For example, consider a binary response model where each of the n subjects has matched observations (X_{ij}, Y_{ij}) such that the responses Y_{ij}, conditionally on the regressors $X_{i1},\ldots,X_{iJ}$ and the individual effect α_i, are independent across j and have Bernoulli distributions with success probabilities $p(X_{ij},\alpha_i)$, respectively. Assume that

$$\log\left[\frac{p(X_{ij},\alpha_i)}{1-p(X_{ij},\alpha_i)}\right] = f(X_{ij}) + \alpha_i$$

and consider the case where $J=2$ for simplicity. Let $Z_i = I(Y_{i1}=1)$ and $N_i = Y_{i1}+Y_{i2}$, where I denotes the indicator function. Then, it can be shown that

$$\log\left[\frac{E(Z_i|X_{i1},X_{i2},N_i=1)}{1-E(Z_i|X_{i1},X_{i2},N_i=1)}\right] = f(X_{i1}) - f(X_{i2}). \tag{7.30}$$

This follows from the equation

$$E(Z_i|X_{i1},X_{i2},N_i=1) = \frac{E\big[p(X_{i1},\alpha_i)(1-p(X_{i2},\alpha_i))|X_{i1},X_{i2}\big]}{E\big[p(X_{i1},\alpha_i)(1-p(X_{i2},\alpha_i)) + p(X_{i2},\alpha_i)(1-p(X_{i1},\alpha_i))|X_{i1},X_{i2}\big]}$$

and the fact that

$$\frac{p(X_{i1},\alpha_i)[1-p(X_{i2},\alpha_i)]}{p(X_{i1},\alpha_i)[1-p(X_{i2},\alpha_i)] + p(X_{i2},\alpha_i)[1-p(X_{i1},\alpha_i)]} = \frac{\exp[f(X_{i1})-f(X_{i2})]}{1+\exp[f(X_{i1})-f(X_{i2})]}$$

does not involve α_i. This generalizes an observation that has been made for parametric conditional maximum likelihood estimation in panel logit models; see Rasch (1960), Rasch (1961), Andersen (1970), and Chamberlain (1994). For extensions of the conditional logit approach see Magnac (2004).

Let Z, X_j, Y_j denote the generics for Z_i, X_{ij}, Y_{ij}, respectively. The function f in the transformed model (7.30) maximizes the expected log-likelihood, so that it satisfies

$$EI(N=1)\big[Z-\eta(X_1,X_2;f)\big][g(X_1)-g(X_2)] = 0$$

for all square integrable function g, where

$$\eta(x,y;m)=\frac{\exp[m(x)-m(y)]}{1+\exp[m(x)-m(y)]}.$$

It can be shown that f satisfies $F(f)=0$, where F is a nonlinear operator defined by

$$\begin{aligned}F(m)(x)&=E[I(N=1)(Z-\eta(X_1,X_2;m))|\,X_1=x]\\&\quad\times p_1(x)-E[I(N=1)(Z-\eta(X_1,X_2;m))|\,X_2=x]p_2(x)\end{aligned}$$

and p_j denotes the density of X_j, $j=1,2$. Here, we also need a norming condition for identifiability of f. The integral equation $F(m)=0$ is nonlinear, but it can be linearized in the same way as the nonlinear equation in Section 7.2. The linear approximation basically puts the problem back to the framework for the model (7.26). To detail this, define $\eta_1(x,y;m)=[1+\exp(m(x)-m(y))]^{-2}$ and let $f^{[0]}$ be a function close to f. Note that $F(m)\simeq F(f^{[0]})+F_1(f^{[0]})(m-f^{[0]})$, where $F_1(f^{[0]})$ is a linear operator and $F_1(f^{[0]})(g)$ denotes the Fréchet differential of F at $f^{[0]}$ with increment g. Put $\delta=f-f^{[0]}$ and

$$\begin{aligned}\mathcal{H}_0(x,y)&=E\big[I(N=1)|X_1=x,X_2=y\big]\eta_1(x,y;f^{[0]})p_{12}(x,y)\\&\quad+E\big[I(N=1)|X_1=y,X_2=x\big]\eta_1(y,x;f^{[0]})p_{12}(y,x),\end{aligned}$$

where p_{12} denotes the density of (X_1,X_2). Then, the approximating linear integral equation $F(f^{[0]})+F_1(f^{[0]})(\delta)=0$ is equivalent to

$$\delta(x)=\delta^*(x)-\int\mathcal{H}(x,y)\delta(y)\,dy, \tag{7.31}$$

where

$$\begin{aligned}\delta^*(x)&=\left[\int\mathcal{H}_0(x,y)\,dy\right]^{-1}F(f^{[0]})(x),\\ \mathcal{H}(x,y)&=-\left[\int\mathcal{H}_0(x,z)\,dz\right]^{-1}\mathcal{H}_0(x,y).\end{aligned}$$

We may estimate F and $\mathcal{H}_0$ by kernel methods. Plugging the estimators $\widehat{F}$ and $\widehat{\mathcal{H}}_0$ into (7.31) gives an integral equation for the update $\widehat{f}^{[1]}$ of the starting estimator $\widehat{f}^{[0]}$. The statistical properties of the resulting backfitting algorithm and the limit of the algorithm $\widehat{f}$ which satisfies $\widehat{F}(\widehat{f})=0$ have been studied by Hoderlein, Mammen, and Yu (2011).

7.3.4. Additive Models for Panels of Time Series and Factor Models

Similar to (7.26), one can consider models with an unobserved time effect η_t instead of an individual effect. We now denote time by t. Suppose that we have panel data $(X_{it}^1, \dots, X_{it}^d, Y_{it})$ for individuals $1 \leq i \leq n$ and time points $1 \leq T$. We assume that

$$Y_{it} = \sum_{j=1}^{d} m_j(X_{it}^j) + \eta_t + \varepsilon_{it}. \tag{7.32}$$

This model naturally generalizes linear panel data models. It has been studied in Mammen, Støve, and Tjøstheim (2009) for two asymptotic frameworks: $n \to \infty, T$ fixed and $n, T \to \infty$. Their asymptotic analysis includes the case where $\{X_{it}^j\}$, $j = 1, \dots, p$, are time lagged values of Y_{it}. No assumptions are made on the unobserved temporary effects η_t. They may be deterministic or random, and they may be correlated with covariates or error terms. The basic idea of Mammen, Støve, and Tjøstheim (2009) is to use difference schemes that cancel out the time effects η_t, simliar to the approaches in the last subsection that cancel out individual effects. Here, the values η_t are nuissance parameters.

In Linton and Nielsen (2009) also the model (7.32) is considered, but the statistical aim there is inference on the structure of η_t. It is assumed that η_t is a random process following a parametric specification. A two-step procedure is proposed where the process η_t is fitted in the first step. In their mathematics they compare parametric inference based on the fitted values of η_t with an infeasible statistical inference that is based on the unobserved η_t. The main result is that these two approaches are asymptotically equivalent. This can be interpreted as an oracle property and it can be used to construct efficient estimators of the parameters.

Another modification of model (7.32) is the factor model

$$Y_{tl} = m_0(X_{tl}^0) + \sum_{j=1}^{d} Z_t^j m_j(X_{tl}^j) + \varepsilon_{tl} \tag{7.33}$$

for $l = 1, \dots, L$. Here, the dynamics of the L-dimensional process Y_t is approximated by the unobserved d-dimensional time series Z_t. The basic idea is that elements Y_{tl} of Y_t with similar characteristics $(X_{tl}^j : 1 \leq j \leq d)$ show similar dynamics and that the dynamics of Y_t can be accurately modeled by choices of d that are much smaller than L. This model has been applied in Connor, Hagmann, and Linton (2012) to the analysis of stock returns Y_{tl} with characteristics $(X_{tl}^j : 1 \leq j \leq d)$. Again, a two-step procedure is proposed where in the first-step the unobserved process Z_t is fitted. Also, an oracle property applies: Inference based on estimates $\widehat{Z}_t$ of Z_t is asymptotically equivalent to infeasible inference based on the unobserved Z_t.

In Fengler, Härdle, and Mammen (2007) and Park, Mammen, Härdle, and Borak (2009) the following model has been considered:

$$Y_{tl} = m_0(X_{tl}) + \sum_{j=1}^{d} Z_t^j m_j(X_{tl}) + \varepsilon_{tl}.$$

This model differs from (7.33) because now the nonparametric components m_j are functions of a single characteristic X_{tl}. As a result, the multivariate time series Z_t is only identified up to linear transformations. Again, an oracle property for parametric inference based on fitted values was shown in Park, Mammen, Härdle, and Borak (2009). The model has been used in functional principal component analysis. One application in Fengler, Härdle, and Mammen (2007) and Park, Mammen, Härdle, and Borak (2009) is for implied volatility surfaces that develop over time. The surfaces are approximated by a finite-dimensional process and the random movement of the surfaces is then analyzed by a VAR representation of the finite-dimensional process.

7.3.5. Semiparametric GARCH Models

Another example that leads to an additive model is a semiparametric GARCH model. In this model we observe a process Y_t such that $E(Y_t|\mathcal{F}_{t-1}) = 0$, where $\mathcal{F}_{t-1}$ denotes the sigma field generated by the entire past history of the Y process, and $\sigma_t^2 \equiv E(Y_t^2|\mathcal{F}_{t-1})$ assumes a semiparametric model

$$\sigma_t^2 = \theta\sigma_{t-1}^2 + f(Y_{t-1}). \tag{7.34}$$

This model is a natural generalization of the GARCH(1,1) model of Bollerslev (1986), where a parametric assumption is made on f such that $f(x) = \alpha + \beta x$. The generalization was introduced by Engle and Ng (1993) to allow for more flexibility in the "news impact curve"—that is, the function f, which measures the effect of news onto volatilities in financial markets.

The parameters θ and the function f in the semiparametric model (7.34) are unknown. Since $E(Y_t^2|\mathcal{F}_{t-1}) = \sum_{j=1}^{\infty}\theta^{j-1}f(Y_{t-j})$, the parameter θ and the function $f(\cdot,\theta)$ together minimize $E[Y_0^2 - \sum_{j=1}^{\infty}\theta^{j-1}f(X_{-j})]^2$. For each θ, let f_θ denote the minimizer of the criterion. Then, it satisfies

$$\sum_{j=1}^{\infty}\sum_{k=1}^{\infty}\theta^{j+k-2}f_\theta(Y_{-k})g(Y_{-j}) = \sum_{j=1}^{\infty}E[Y_0^2\theta^{j-1}g(Y_{-j})]$$

for all square integrable functions g. This gives the following integral equation.

$$f_\theta(x) = f_\theta^*(x) - \int \mathcal{H}_\theta(x,y)f_\theta(y)\,dy, \tag{7.35}$$

where

$$f_\theta^*(x) = (1-\theta^2)\sum_{j=1}^{\infty}\theta^{j-1}E(Y_0^2|Y_{-j}=x),$$

$$\mathcal{H}_\theta(x,y) = \sum_{j=1}^{\infty}\theta^j\left[\frac{p_{0,-j}(x,y)+p_{0,j}(x,y)}{p_0(x)}\right],$$

p_0 and $p_{0,j}$ are the densities of Y_0 and (Y_0, Y_j), respectively. For an asymptotic and empirical analysis of the estimators based on the integral equation (7.35), we refer to Linton and Mammen (2005). For a recent extension of the model, see also Chen and Ghysels (2011).

7.3.6. Varying Coefficient Models

Suppose we are given a group of covariates $X^1,\ldots,X^d$ and a response Y. The most general form of varying coefficient model was introduced and studied by Lee, Mammen, and Park (2012a). It is given by

$$E(Y|X^1,\ldots,X^d) = g^{-1}\left(\sum_{k\in I_1}X^k f_{k1}(X^1)+\cdots+\sum_{k\in I_p}X^k f_{kp}(X^p)\right), \tag{7.36}$$

where g is a link function and $p \le d$. The index sets I_j may intersect with each other, but each I_j does not include j. It is also allowed that the two groups of covariates, $\{X^j : 1 \le j \le p\}$ and $\{X^k : k \in \cup_{j=1}^p I_j\}$, may have common variables. The coefficient functions are identifiable if we set the following constraints: for non-negative weight functions w_j, (i) $\int f_{kj}(x^j)w_j(x^j)\,dx^j = 0$ for all $k \in \cup_{j=1}^p I_j$ and $1 \le j \le p$; (ii) $\int x^j f_{kj}(x^j)w_j(x^j)\,dx^j = 0$ for all $j,k \in \{1,\ldots,p\} \cap (\cup_{j=1}^p I_j)$. In this model, the effect of the covariate X^k for $k \in \cup_{j=1}^p I_j$ is set in a nonparametric way as $\sum_{j:I_j \ni k} f_{kj}(X^j)$. The model is flexible enough to include various types of varying coefficient models as special cases. For example, it is specialized to the generalized additive model discussed in Section 7.2.7 if one takes $I_1 = \cdots = I_p = \{p+1\}$ and sets $X^{p+1} \equiv 1$. The model also reduces to the varying coefficient model studied by Lee, Mammen, and Park (2012b) and Yang, Park, Xue, and Härdle (2006) if the two groups, $\{X^j\!: 1 \le j \le p\}$ and $\{X^k\!: k \in \cup_{j=1}^p I_j\}$, are disjoint and the sets I_j contain only one element $(1 \le j \le p)$. In this case, one can rewrite model (7.36) as

$$Y_i = g^{-1}\left(\sum_{j=1}^{p} Z_i^j f_j(X_i^j)\right) + \varepsilon_i.$$

With an identity link g and with the additional constraint $f_j \equiv f$, this model has been used in Linton, Mammen, Nielsen, and Tanggaard (2001) for nonparametric estimation of yield curves by smoothed least squares. There, Y_i was the trading price of a coupon bond, Z_i^j denotes the payment returned to the owner of bond i at date X_i^j, and f is the discount function. In case $p = 1$ and $I_1 = \{2, \ldots, d\}$, the approach with disjoint sets of covariates results in the model studied, for example, by Fan and Zhang (1999).

For simplicity, suppose that the link g is the identity function. In this case, the coefficient functions f_{kj} minimize $E\Big[Y - \sum_{k \in I_1} X^k f_{k1}(X^1) - \cdots - \sum_{k \in I_p} X^k f_{kp}(X^p)\Big]^2$. This gives the following system of integral equations for f_{kj}: for $1 \leq j \leq p$, we have

$$\begin{aligned}\mathbf{f}_j(x^j) =& E(\mathbf{X}_j \mathbf{X}_j^\top | X^j = x^j)^{-1} E(\mathbf{X}_j Y | X^j = x^j) - E(\mathbf{X}_j \mathbf{X}_j^\top | X^j = x^j)^{-1} \\ & \times \sum_{l=1,\neq j}^{p} \int E\Big[\mathbf{X}_j \mathbf{X}_l^\top | X^j = x^j, X^l = x^l\Big] \mathbf{f}_l(x^l) \frac{p_{jl}(x^j, x^l)}{p_j(x^j)}\, dx^l,\end{aligned}$$

where $\mathbf{X}_j = (X^k\colon k \in I_j)$ and $\mathbf{f}_j = (f_{kj}\colon k \in I_j)$. Note that $\mathbf{X}_j$ does not contain X^j as its entry. To get an empirical version of the above integral equations, one may replace the conditional expectations, the joint density p_{jl} of (X^j, X^l), and the marginal density p_j of X^j by kernel estimators. Lee, Mammen, and Park (2012a) presented complete theory for the estimation of the general model (7.36). Their theory includes sieve and penalized quasi-likelihood estimation as well as the smooth backfitting method described above.

7.3.7. Missing Observations

Additive models can also be consistently estimated if the tuples $(Y_i, X_i^1, \ldots, X_i^d)$ are only partially observed. We will discuss this for a simple scheme of missing observations. Let N_{jk} denote the set of indices i where X_i^j and X_i^k are observed; N_j the set of indices i where X_i^j is observed; N_{0j} the set of indices i where X_i^j and Y_i are observed; and N_0 the set of indices i where Y_i is observed.

- Denote by $\mathcal{N}_{jk}$ the set of indices i where X_i^j and X_i^k are observed.
- Denote by $\mathcal{N}_j$ the set of indices i where X_i^j is observed.
- Denote by $\mathcal{N}_{0j}$ the set of indices i where X_i^j and Y_i are observed.
- Denote by $\mathcal{N}_0$ the set of indices i where Y_i is observed.

These sets may be random or nonrandom. We denote the number of elements of these sets by N_{jk}, N_j, N_{0j} or N_0, respectively. We assume that the observations $\{(X_i^j, X_i^k) : i \in \mathcal{N}_{jk}\}$, $\{X_i^j : i \in \mathcal{N}_j\}$, $\{(X_i^j, Y_i) : i \in \mathcal{N}_{0j}\}$, and $\{Y_i\colon i \in \mathcal{N}_0\}$ are i.i.d. This assumption holds under simple random missingness schemes and also in the case of pooling samples where different subsets of covariates were observed.

Then, under the assumption that $N_{jk} \to \infty$, $N_j \to \infty$, $N_{0j} \to \infty$ and $N_0 \to \infty$, the estimators of p_{X^j,X^k}, p_{X^j}, f_j^*, and μ that are based on the subsamples $\mathcal{N}_{jk}$, $\mathcal{N}_j$, $\mathcal{N}_{0j}$, and $\mathcal{N}_0$, respectively, are consistent. More precisely, for $1 \le j \ne k \le d$, put

$$\widetilde{p}_{X^j,X^k}(x^j,x^k) = \frac{1}{N_{jk}h_jh_k}\sum_{i\in\mathcal{N}_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right)K\left(\frac{X_i^k - x^k}{h_k}\right),$$

$$\widetilde{p}_{X^j}(x^j) = \frac{1}{N_jh_j}\sum_{i\in\mathcal{N}_j} K\left(\frac{X_i^j - x^j}{h_j}\right),$$

$$\widetilde{f}_j^*(x^j) = \widetilde{p}_{X^j}(x^j)^{-1}\frac{1}{N_{0j}h_j}\sum_{i\in\mathcal{N}_{0j}} K\left(\frac{X_i^j - x^j}{h_j}\right)Y_i,$$

$$\widetilde{\mu} = \frac{1}{N_0}\sum_{i\in\mathcal{N}_0}^{n} Y_i.$$

Under appropriate conditions on the bandwidths h_j, these estimators converge to $p_{X^j,X^k}(x^j,x^k)$, $p_{X^j}(x^j)$, $f_j^*(x^j)$, and μ, respectively, in probability. Similarly as in Eq. (7.6), we consider the solutions $\widetilde{f}_1,\dots,\widetilde{f}_d$ of the equations

$$\widetilde{f}_j(x^j) = \widetilde{f}_j^*(x^j) - \widetilde{\mu} - \sum_{k\ne j}\int \frac{\widetilde{p}_{X^j,X^k}(x^j,x^k)}{\widetilde{p}_{X^j}(x^j)}\widetilde{f}_k(x^k)\,dx^k.$$

Using the stochastic convergence of $\widetilde{p}_{X^j,X^k}(x^j,x^k)$, $\widetilde{p}_{X^j}(x^j)$, $\widetilde{f}_j^*(x^j)$, and $\widetilde{\mu}$, one can show that $\widetilde{f}_j(x^j)$ converges in probability to $f_j(x^j)$ for $1 \le j \le d$. These consistency proofs can be generalized to more complex missingness schemes. Furthermore, under appropriate conditions, one can study normal distribution limits of these estimators. We remark that these identification, consistency, and asymptotic normality results are not available for the full-dimensional model specification: $Y = f(X^1,\dots,X^d) + \varepsilon$.

7.3.8. Additive Diffusion Models

Some multivariate diffusion models are based on additive parametric specifications of the mean. Nonparametric generalizations of such models were considered in Haag (2006). There also nonparametric specifications of the volatility term were considered.

7.3.9. Simultaneous Nonparametric Equation Models

Additive models also naturally occur in economic models, where some covariates are correlated with the disturbance. Despite these so-called endogenous regressors, such models can be identified via a control function approach. In particular, Newey, Powell, and Vella (1999) proposed the following model with additive error terms:

$$Y = f(X^1, Z^1) + e,$$

where X^1 and Z^1 are observed covariates and Y is a one-dimensional response. While Z^1 is independent of the error variable e, no assumptions are made on the dependence between X^1 and e at this stage. For identification, however, assume that the following control equation holds for the endogenous variable X^1:

$$X^1 = h(Z^1, Z^2) + V,$$

where Z^2 is an observed covariate not contained in the original equation and (Z^1, Z^2) is independent of the joint vector of errors (e, V).

Under the stated independence conditions, it follows that

$$E(Y|X^1, Z^1, Z^2) = f(X^1, Z^1) + \lambda(V) = E[Y|X^1, Z^1, V] \tag{7.37}$$

with $\lambda(V) = E(e|V)$. Thus, we get an additive model where the regressor in the second additive component is not observed but can be estimated as residual of the control equation. This additive model can be also obtained under slightly weaker conditions than the above independence conditions, namely under the assumption that $E(e|Z^1, Z^2, V) = E(e|V)$ and $E(V|Z^1, Z^2) = 0$. The corresponding system of integral equations to be solved for (7.37) is

$$f(x^1, z^2) = f^*(x^1, z^2) - \int \frac{p_{X^1,Z^2,V}(x^1, z^2, v)}{p_{X^1,Z^2}(x^1, z^2)} \lambda(v)\, dv$$

$$\lambda(v) = \lambda^*(v) - \int \frac{p_{X^1,Z^2,V}(x^1, z^2, v)}{p_V(v)} f(x^1, z^2)\, d(x^1, z^2),$$

where $f^*(z^1, z^2) = E[Y|(X^1, Z^1) = (x^1, z^2)]$ and $\lambda^*(v) = E(Y|V = v)$. Note that some ingredients of the smooth backfitting iteration algorithm thus require nonparametric pre-estimates of marginal objects with the nonparametrically generated regressor $\widehat{V} = X^1 - \widehat{h}(Z^1, Z^2)$. The paper by Mammen, Rothe, and Schienle (2012) studies how asymptotic theory in nonparametric models has to be adjusted to take care of nonparametrically generated regressors.

7.4. Nonstationary Observations

Additive models are a powerful tool in case of stochastically nonstationary covariates. For this data generality, consistent estimation of a fully nonparametric model requires that the whole compound vector fulfills a specific recurrence condition; that is, it has to be guaranteed that the full-dimensional process X returns infinitely often to local neighborhoods (see, for example, Karlsen and Tjøstheim (2001), Wang and Phillips (2009), and Karlsen, Myklebust, and Tjøstheim (2007)). For an additive model, however, recurrence conditions are only needed for two-dimensional subvectors of X. An illustrative example is a multivariate random walk. A fully nonparametric model cannot be consistently estimated for dimensions greater than two, since beyond dimension two random walks become transient and do not fulfill the above recurrence property. For an additive model, however, there is no dimension restriction, because any pair of bivariate random walks is recurrent. Here we briefly outline the main ideas. The detailed theory of additive models for nonstationary covariates is developed in Schienle (2008).

The setting is as follows: Suppose we want to estimate a standard additive model (7.1) where covariates and response are potentially nonstationary Markov chains but satisfy a pairwise recurrence condition, and the residual is stationary mixing. Instead of a stationary data generating process density function, a nonstationary pairwise recurrent Markov chain can be characterized by the densities of pairwise bivariate invariant measures π_{jk} with $j, k \in \{1, \ldots, d\}$. For the specific kind of recurrence imposed, it is guaranteed that such a bivariate invariant measure exists for each pair and is unique up to a multiplicative constant; but it is generally only finite on so-called small sets and only σ-finite on the full support. Note, for example, that for random walks any compact set is small.

Furthermore, under the type of pairwise recurrence imposed, bivariate component Markov chains $(X^j, X^k) = (X^{jk})$ can be decomposed into i.i.d. parts of random length depending on the recurrence times of the chain. In particular, the stochastic number of recurrence times $T^{jk}(n)$ characterizes the amount of i.i.d. block observations and thus corresponds to the effective sample size available for inference with the particular pair of components. Thus for different components and pairs of components available, effective sample sizes are path-dependent and generally vary depending on the recurrence frequency being smaller for more nonstationary processes and closer to the stationary deterministic full sample size n for more stationary processes. Correspondingly, consistent kernel-type estimators are weighted averages of $T^{jk}(n)$ i.i.d. block elements

$$\begin{aligned}\widehat{\pi}_{jk}(x^{jk}) &= \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} K\left(\frac{X_i^{jk} - x^{jk}}{h_{jk}}\right), \\ \widehat{f}_j(x^j) &= \left[\sum_{i \in I_j} K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i \in I_j} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i,\end{aligned} \tag{7.38}$$

$$\begin{aligned}\widehat{\pi}_j^{(k)}(x^j) &= \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right),\\ \widehat{f}_j^{(k)}(x^j) &= \left[\sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right)\right]^{-1} \sum_{i \in I_{jk}} K\left(\frac{X_i^j - x^j}{h_j}\right) Y_i.\end{aligned} \tag{7.39}$$

The estimators in (7.38) provide pointwise consistent estimates of the corresponding bivariate invariant measure density π_{jk} and a general nonparametric link function f_j, respectively (see Karlsen, Myklebust, and Tjøstheim (2007)). Their rates of convergence are driven by respective recurrence frequencies and occupation times $\widehat{L}_{jk}(x^{jk}) = \sum_{i \in I_{jk}} K_{x^{jk},h_{jk}}(X_i^{jk})$ and $\widehat{L}_j$, respectively, which are generally of different order on average over all sample paths. Asymptotically in both cases, they are on average of size $(n^{\beta^{jk}} h)^{-1/2}$ and $(n^{\beta^j} h)^{-1/2}$, respectively, where the global β^{jk}-parameter $\in [0,1]$ characterizes the underlying type of nonstationarity of the corresponding recurrent chain as the tail index on the distribution of recurrence times. For a bivariate random walk we have $\beta^{jk} = 0$, for a stationary process we have $\beta^{jk} = 1$ recovering standard rates, and generally $\beta^{jk} \leq \beta^j$. The kernel estimators in (7.39) artificially "downgrade" their univariate speed of convergence to the respective bivariate one. Note that the index sets I_{jk} ensure that only $T^{jk}(n)$ i.i.d. sub-blocks are considered of the $T^j(n)$ original ones.

For balancing terms in the empirical version of the smooth backfitting integral equations, such potentially slower-than-standard estimators $\widehat{\pi}_j^{(k)}$, $\widehat{\pi}_{jl}^{(k)}$, and $\widehat{f}_j^{(k)}$ of bivariate nonstationary type β^{jk} are necessary. Also in the backfitting operator for component j, the impact of other directions on any pair of components containing X^j might now differ depending on respective occupation times of component pairs. Both aspects are reflected by a respectively generalized procedure ensuring consistent estimates. The generalized smooth backfitting estimates $(\widehat{f_j})_{j=1}^d$ are defined as

$$\widehat{f}_j(x^j) = \frac{1}{d-1}\left[\sum_{k \neq j}\left(\widehat{f}_j^{(k)*}(x^j) - \widehat{f}_{0,j}^{(k)*}\right) - \sum_{k \neq j} \frac{1}{\widehat{\lambda}_{jk}} \sum_{l \neq j} \int_{\mathcal{G}_l} \widehat{f}_l(x^l) \frac{\widehat{\pi}_{jl}^{(k)}(x^{jl})}{\widehat{\pi}_j^{(k)}(x^j)}\, dx^l\right], \tag{7.40}$$

where $\widehat{f}_j^{(k)*}(x^j)$ are the marginal local constant estimates with bivariate speed of convergence as defined above and constants

$$\widehat{f}_{0,j}^{(k)*} = \frac{\int_{\mathcal{G}_j} \widehat{f}_j^{(k)*}(x^j)\widehat{\pi}_j^{(k)}(x^j)\, dx^j}{\int_{\mathcal{G}_j} \widehat{\pi}_j^{(k)}(x^j)\, dx^j} = \frac{1}{T^{jk}(n)} \sum_{i \in I_{jk}} Y_i, \tag{7.41}$$

which follow from appropriate analogues of the standard norming constraints

$$\sum_{k \neq j} \int_{\mathcal{G}_j} f_j(x^j) \pi_j^{(k)}(x^j)\, dx^j = 0. \tag{7.42}$$

Note that asymptotically in the projection part of (7.40), only those elements $\widehat{\pi}_{jl}$ prevail, where $\beta^{jl} = \beta^{jk}$ while all others vanish. The projection property of standard backfitting only prevails in a generalized sense, since in general an invariant measure for the full-dimensional compound process does not exist for pairwise recurrent X. For each j and k, $\widehat{\lambda}_{jk}$ counts the number of such elements in the sample. In a nonstationary setting, also the regions of integration $\mathcal{G}_j$ must be chosen with some care to ensure that integrals exist. Related to small sets—for example, in a random walk case—compact areas are appropriate. If all pairs of components of X have the same type of nonstationarity, the backfitting equations reduce to

$$\widehat{f}_j(x^j) = \frac{1}{d-1} \sum_{k \neq j} \left(\widehat{f}_j^{(k)*}(x^j) - \widehat{f}_{0,j}^{(k)*} \right) - \sum_{k \neq j} \int_{\mathcal{G}_k} \widehat{f}_k(x^k) \frac{\widehat{\pi}_{jk}(x^{jk})}{\widehat{\pi}_j^{(k)}(x^j)} \, dx^k \, ,$$

since $\lambda_{jk} = d - 1$ and $\widehat{\pi}_{jl}^{(k)} = \widehat{\pi}_{jl}$ in this case. In particular, for the special case of identical one- and two-dimensional scales, generalized smooth backfitting reduces to the standard case. This usually occurs for sufficiently stationary data.

Asymptotic results for the generalized backfitting are univariate in form; that is, the standard curse of dimensionality can be circumvented. However, they are driven by the worst-case bivariate type of nonstationarity in the data. In particular, the difference between the true component function f_j and the backfitting estimate $\widehat{f}_j$ is asymptotically normal when inflated with the stochastic occupation time factor $\sqrt{\min_{k \neq j} \widehat{L}_j^{(k)}(x^j) h}$. Because $\widehat{L}_j^{(k)}$ is asymptotically of the same order as $T^{jk}(n)$, the rate of convergence is, on average, of size $\sqrt{n^{\beta^{j+}+\epsilon} h}$, where β^{j+} is the highest degree of nonstationarity, and thus the smallest number among the β^{jk}, and $\epsilon > 0$ is very small. That means, if all components are random walks—that is, $\beta^{jk} = 0$—estimation of each component is possible, but with logarithmic rate. This should be compared to the fact that a fully nonparametric model cannot be estimated in this case where the compound vector is transient. If one component X^{j_0} follows a random walk and all others are stationary, all components are estimated at rate $\sqrt{n^{\beta^{j_0}} h} = \sqrt{n^{1/2} h}$.

7.5. Noisy Fredholm Integral Equations of Second Kind

As outlined in Subsection 7.2.4, we can define the smooth backfitting estimators in the additive models as solutions of an integral equation $\widehat{\mathbf{f}}(x) = \widehat{\mathbf{f}}^*(x) - \int \widehat{\mathcal{H}}(x,z) \widehat{\mathbf{f}}(z) \, dz$, where $\widehat{\mathbf{f}}(x^1, \ldots, x^d) = (\widehat{f}_1(x^1), \ldots, \widehat{f}_d(x^d))^\top$, $\widehat{\mathbf{f}}^*(x^1, \ldots, x^d) = (\widehat{f}_1^*(x^1), \ldots, \widehat{f}_d^*(x^d))^\top$, and the integral kernel $\widehat{\mathcal{H}}(x,z)$ equals a matrix with elements $\widehat{p}_{X^j,X^k}(x^j, x^k)/\widehat{p}_{X^j}(x^j)$. We also rewrite this noisy integral equation as

$$\widehat{\mathbf{f}} = \widehat{\mathbf{f}}^* - \widehat{\mathcal{H}}\widehat{\mathbf{f}}.$$

In Section 7.3 we have also seen that smooth least squares for various models leads to estimators that are given as solutions of such noisy integral equations. There are several approaches to the numerical solution of the integral equation. As already mentioned in Subsection 7.2.4, one can use a discrete approximation of the integral equation for the numerical solution. This results in a finite system of linear equations that can be solved by standard methods. One approach would be based on a iterative scheme that uses a discrete approximation of the iteration steps:

$$\widehat{\mathbf{f}}^{NEW} = \widehat{\mathbf{f}}^* - \widehat{\mathcal{H}}\widehat{\mathbf{f}}^{OLD}.$$

If $\widehat{\mathbf{f}}$ is a d-dimensional vector of functions with $d \geq 2$, one can also use an iteration scheme that runs cyclically through componentwise updates

$$\widehat{f}_j^{NEW} = \widehat{f}_j^* - \widehat{\mathcal{H}}_j\widehat{\mathbf{f}}^{OLD}, \qquad 1 \leq j \leq d,$$

with an obvious definition of $\widehat{\mathcal{H}}_j$. This was the algorithm we discussed in Subsection 7.2.1. Compare also the Gauss–Seidel method and the Jacobi method in numerical linear algebra.

We now use the definition of the estimators by a noisy integral equation for an asymptotic understanding of the distributional properties of the estimators. We consider the case of one-dimensional $\widehat{\mathbf{f}}$ and $\widehat{\mathbf{f}}^*$ and we rewrite the equation as $\widehat{f} = \widehat{f}^* - \widehat{\mathcal{H}}\widehat{f}$. We now suppose that $\widehat{f}^*$ is a smoothing estimator with

$$\widehat{f}^* \approx \widehat{f}_A^* + f^* + f_B^*,$$

where $\widehat{f}_A^*$ is the *stochastic part* of $\widehat{f}^*$ that is of order $(nh)^{-1/2}$. The function f^* is the stochastic limit of $\widehat{f}^*$ and f_B^* is a bias term that we suppose to be of the standard order h^2. Here, h is a bandwidth that is chosen of order $n^{-1/5}$ so that the stochastic term and the bias term are of order $n^{-2/5}$. A similar discussion applies to $\widehat{\mathcal{H}}f$. This variable has stochastic limit $\mathcal{H}f$, where $\mathcal{H}$ is the stochastic limit of $\widehat{\mathcal{H}}$. We now get

$$\widehat{\mathcal{H}}f \approx (\widehat{\mathcal{H}}f)_A + \mathcal{H}f + (\mathcal{H}f)_B,$$

where $(\widehat{\mathcal{H}}f)_A$ is the *stochastic part* of $\widehat{\mathcal{H}}f$. Again this term is of order $(nh)^{-1/2}$. Although $\widehat{\mathcal{H}}$ is a higher-dimensional smoother, all variables up to one are integrated out in $\widehat{\mathcal{H}}f$. Furthermore, $(\mathcal{H}f)_B$ is a bias term that is of order h^2. By subtracting $f = f^* - \mathcal{H}f$ from $\widehat{f} = \widehat{f}^* - \widehat{\mathcal{H}}\widehat{f}$, we get

$$\begin{aligned}
\widehat{f} - f &= \widehat{f}^* - f^* - \widehat{\mathcal{H}}\widehat{f} + \mathcal{H}f \\
&= \widehat{f}^* - f^* - \mathcal{H}(\widehat{f} - f) - (\widehat{\mathcal{H}} - \mathcal{H})f - (\widehat{\mathcal{H}} - \mathcal{H})(\widehat{f} - f) \\
&\approx \widehat{f}^* - f^* - \mathcal{H}(\widehat{f} - f) - (\widehat{\mathcal{H}} - \mathcal{H})f.
\end{aligned}$$

Now, simple algebra gives

$$\begin{aligned}\widehat{f}-f &\approx (I+\mathcal{H})^{-1}(\widehat{f}^{*}-f^{*}-(\widehat{\mathcal{H}}-\mathcal{H})f)\\ &\approx (I+\mathcal{H})^{-1}(\widehat{f_A^{*}}+f_B^{*}-(\widehat{\mathcal{H}f})_A-(\mathcal{H}f)_B).\end{aligned}$$

We now argue that $(I+\mathcal{H})^{-1}\widehat{f_A^{*}}\approx\widehat{f_A^{*}}$ and $(I+\mathcal{H})^{-1}(\widehat{\mathcal{H}f})_A\approx(\widehat{\mathcal{H}f})_A$. These claims follow immediately from $(I+\mathcal{H})^{-1}=I-(I+\mathcal{H})^{-1}\mathcal{H}$, $\mathcal{H}\widehat{f_A^{*}}\approx 0$ and $\mathcal{H}(\widehat{\mathcal{H}f})_A\approx 0$. Here, the first equality can be easily seen by multiplying both sides of the equation with $(I+\mathcal{H})$. For the two approximations, one notes that the integral, over an interval, of the stochastic part of a kernel smoother is typically of order $n^{-1/2}$. For example, one has $\int w(x)n^{-1}\sum_{i=1}^{n}K_h(x-X_i)\varepsilon_i\,dx=n^{-1}\sum_{i=1}^{n}w_h(X_i)\varepsilon_i$ with $w_h(u)=\int w(x)K_h(x-u)\,dx$, which is of order $n^{-1/2}$. Using the above approximations, we get that

$$\begin{aligned}\widehat{f}-f &\approx (I+\mathcal{H})^{-1}(\widehat{f_A^{*}}+f_B^{*}-(\widehat{\mathcal{H}f})_A-(\mathcal{H}f)_B)\\ &=\widehat{f_A^{*}}-(\widehat{\mathcal{H}f})_A-(I+\mathcal{H})^{-1}\mathcal{H}(\widehat{f_A^{*}}-(\widehat{\mathcal{H}f})_A)+(I+\mathcal{H})^{-1}(f_B^{*}-(\mathcal{H}f)_B)\\ &\approx\widehat{f_A^{*}}-(\widehat{\mathcal{H}f})_A+(I+\mathcal{H})^{-1}(f_B^{*}-(\mathcal{H}f)_B).\end{aligned}$$

The expressions on the right-hand side of this expansion can be easily interpreted. The first term $\widehat{f_A^{*}}-(\widehat{\mathcal{H}f})_A$ is of order $(nh)^{-1/2}$ and asymptotically normal with mean zero. This can be shown as in classical kernel smoothing theory. The second term $(I+\mathcal{H})^{-1}(f_B^{*}-(\mathcal{H}f)_B)$ is purely deterministic, and it is of order h^2 because already $f_B^{*}-(\mathcal{H}f)_B$ is of this order. For a more detailed discussion of the above arguments, we refer to Mammen, Støve, and Tjøstheim (2009) and Mammen and Yu (2009).

We conclude this section by noting that the above noisy integral equations are quite different from integral equations of the form

$$0=\widehat{\mathrm{f}^{*}}-\widehat{\mathcal{H}\mathrm{f}}.$$

This is called an ill-posed inverse problem because, typically, the eigenvalues of the operator $\widehat{\mathcal{H}}$ accumulate at 0. For this reason the inverse of the operator $\widehat{\mathcal{H}}$ is not continuous. The integral equation studied in this chapter leads to the inversion of the operator $(I+\widehat{\mathcal{H}})$. The eigenvalues of this operator accumulate around 1 and allow for a continuous inverse of $(I+\widehat{\mathcal{H}})$. Thus our setup is quite different from ill-posed problems. For a discussion of ill-posed problems, we refer to Carrasco, Florens, and Renault (2006), Chen and Reiss (2011), Darolles, Florens, and Renault (2011), Donoho (1995), Engl and Neubauer (1996), Johnstone and Silverman (1990), and Newey and Powell (2003).

Notes

† Enno Mammen gratefully acknowledges research support of the German Science Foundation through the Collaborative Research Center 884 "Political Economy of Reforms."

[‡] Byeong U. Park's research was supported by the NRF Grant funded by the Korea government (MEST) (No. 2010-0017437).

[§] Melanie Schienle gratefully ackowledges research support of the German Science Foundation through the Collaborative Research Center 649.

References

Andersen, E. 1970. "Asymptotic Properties of Conditional Maximum Likelihood Estimators." *Journal of the Royal Statistical Society Series B*, **32**, pp. 283–301.

Bollerslev, T. 1986. "Generalized Autoregressive Conditional Heteroscedasticity," *Journal of Econometrics*, **31**, pp. 307–327.

Buja, A., T. Hastie, and R. Tibshirani. 1989. "Linear Smoothers and Additive Models (with Discussion)." *Annals of Statistics*, **17**, pp. 453–510.

Carrasco, M., J. Florens, and E. Renault. 2006. "Linear Inverse Problems in Structural Econometrics: Estimation Based on Spectral Decomposition and Regularization." In *Handbook of Econometrics*, Vol. 6, eds. J. Heckman and E. Leamer. Amsterdam: Elsevier.

Carroll, R. J., A. Maity, E. Mammen, and K. Yu. 2009. "Nonparametric Additive Regression for Repeatedly Measured Data." *Biometrika*, **96**, pp. 383–398.

Chamberlain, G. 1994. "Panel Data." In *Handbook of Econometrics*, eds. Z. Griliches and M. Intriligator, Amsterdam: Elsevier, pp. 1247–1318.

Chen, X. 2006. "Large Sample Sieve Estimation of Semi-nonparametric Models." In *Handbook of Econometrics*, Vol. 6, eds. J. Heckman and E. Leamer, Amsterdam: Elsevier, pp. 5549–5632.

Chen, X., and E. Ghysels. 2011. "News—Good or Bad—and Its Impact on Volatility Predictions over Multiple Horizons." *Review of Financial Studies*, **24**, pp. 46–81.

Chen, X., and M. Reiss. 2011. "On Rate Optimality for Ill-Posed Inverse Problems in Econometrics." *Econometric Theory*, **27**, pp. 497–52.

Connor, G., M. Hagmann, and Linton. 2012. "Efficient Estimation of a Semiparametric Characteristic-Based Factor Model of Security Returns." *Econometrica*, **18**, pp. 730–754.

Darolles, S., J. P. Florens, and E. Renault. 2011. "Nonparametric Instrumental Regression." *Econometrica*, **79**, pp. 1541–1565.

Donoho, D. L. 1995. "Nonlinear Solutions of Linear Inverse Problems by Wavelet-Vaguelette Decomposition." *Journal of Applied and Computational Harmonic Analysis*, **2**, pp. 101–126.

Eilers, P. H. C., and B. D. Marx. 2002. "Generalized Linear Additive Smooth Structures." *Journal of Computational Graphical Statistics*, **11**, pp. 758–783.

Engl, H., M. Hanke, and A. Neubauer. 1996. *Regularization of Inverse Problems*. London: Kluwer Academic Publishers.

Engle, R. F., and V. K. Ng. 1993. "Measuring and Testing the Impact of News on Volatility." *Journal of Finance*, **48**, pp. 987–1008.

Fan, J., N. Heckman, and M. P. Wand. 1995. "Local Polynomial Kernel Regression for Generalized Linear Models and Quasi-likelihood Functions." *Journal of the American Statistical Association*, **90**, pp. 141–150.

Fan, J., and J. Jiang. 2007. "Nonparametric Inference with Generalized Likelihood Ratio Tests (with Discussion)." *Test*, **16**, pp. 409–478.

Fan, J., Y. Wu, and Y. Feng. 2009. "Local Quasi-likelihood with a Parametric Guide." *Annals of Statistics*, **27**, pp. 4153–4183.

Fan, J., and W. Zhang. 1999. “Statistical Estimation in Varying Coefficient Models.” *Annals of Statistics*, **27**, pp. 1491–1518.

Fengler, M., W. Härdle, and E. Mammen. 2007. “A Semiparametric factor model for implied volatility surface dynamics.” *Journal of Financial Econometrics*, **5**, pp. 189–218.

Haag, B. 2006. “Model Choice in Structured Nonparametric Regression and Diffusion Models.” Ph.D. thesis, Mannheim University.

Haag, B. 2008. “Non-parametric Regression Tests Using Dimension Reduction Techniques.” *Scandinavian Journal of Statistics*, **2008**, pp. 719–738.

Hastie, T. J., and R. J. Tibshirani. 1990. *Generalized Additive Models*. London: Chapman and Hall.

Henderson, D. J., Carroll, R. J. and Li, Q. 2008. “Nonparametric Estimation and Testing of Fixed Effects Panel Data Models,” *Journal of Econometrics*, **144**, pp. 257–275.

Hjort, N. L., and I. K. Glad. 1995. “Nonparametric Density Estimation with a Parametric Start.” *Annals of Statistics*, **23**, pp. 882–904.

Hoderlein, S., E. Mammen, and K. Yu. 2011. “Nonparametric Models in Binary Choice Fixed Effects Panel Data.” *Econometrics Journal*, **14**, pp. 351–367.

Huang, J., J. L. Horowitz, and F. Wei. 2010. “Variable Selection in Nonparametric Additive Models.” *Annals of Statistics*, **38**, pp. 2282–2313.

Jiang, J., Y. Fan, and J. Fan. 2010. “Estimation of Additive Models with Highly or Nonhighly Correlated Covariates.” *Annals of Statistics*, **38**, pp. 1403–1432.

Johnstone, I. M., and B. W. Silverman. 1990. “Speed of Estimation in Positron Emission Tomography and Related Inverse Problems.” *Annals of Statistics*, **18**, pp. 251–280.

Karlsen, H. A., T. Myklebust, and D. Tjøstheim. 2007. “Nonparametric Estimation in a Nonlinear Cointegration Type Model.” *Annals of Statistics*, **35**, pp. 1–57.

Karlsen, H. A., and D. Tjøstheim. 2001. “Nonparametric Estimation in Null–Recurrent Time Series.” *Annals of Statistics*, **29**, pp. 372–416.

Kauermann, G., and J. D. Opsomer. 2003. “Local Likelihood Estimation in Generalized Additive Models.” *Scandinavian Journal of Statistics*, **30**, pp. 317–337.

Lee, Y. K., E. Mammen, and B. U. Park. 2010. “Backfitting and Smooth Backfitting for Additive Quantile Models.” *Annals of Statistics*, **38**, pp. 2857–2883.

———. 2012a. “Flexible Generalized Varying Coefficient Regression Models.” *Annals of Statistics*, **40**, pp. 1906–1933.

———. 2012b. “Projection-Type Estimation for Varying Coefficient Regression Models.” *Bernoulli*, **18**, pp. 177–205.

Lin, Y., and H. Zhang. 2006. “Component Selection and Smoothing in Multivariate Nonparametric Regression.” *Annals of Statistics*, **34**, pp. 2272–2297.

Linton, O., and E. Mammen. 2003. “Nonparametric Smoothing Methods for a Class of Non-standard Curve Estimation Problems.” In *Recent Advances and Trends in Nonparametric Statistics*, eds. M. Akritas, and D. N. Politis. Amsterdam: Elsevier.

———. 2005. “Estimating Semiparametric ARCH(∞) Models by Kernel Smoothing Methods.” *Econometrica*, **73**, pp. 771–836.

———. 2008. “Nonparametric Transformation to White Noise.” *Journal of Econometrics*, **142**, pp. 241–264.

Linton, O., E. Mammen, J. Nielsen, and C. Tanggaard. 2001. “Estimating Yield Curves by Kernel Smoothing Methods.” *Journal of Econometrics*, **105**, pp. 185–223.

Linton, O., and J. Nielsen. 2009. “Nonparametric Regression with a Latent Time Series.” *Econometrics Journal*, **12**, pp. 187–207.

Lundervold, L., D. Tjøstheim, and Q. Yao. 2007. "Exploring Spatial Nonlinearity Using Additive Approximation." *Bernoulli*, **13**, pp. 447–472.

Magnac, T. 2004. "Panel Binary Variables and Sufficiency: Generalizing Conditional Logit." *Econometrica*, **72**, pp. 1859–1876.

Mammen, E., O. Linton, and J. Nielsen. 1999. "The Existence and Asymptotic Properties of a Backfitting Projection Algorithm under Weak Conditions." *Annals of Statistics*, **27**, pp. 1443–1490.

Mammen, E., J. S. Marron, B. A. Turlach, and M. P. Wand. 2001. "A General Framework for Constrained Smoothing." *Statistical Science*, **16**, pp. 232–248.

Mammen, E., and J. Nielsen. 2003. "Generalised Structured Models." *Biometrika*, **90**, pp. 551–566.

Mammen, E., and B. U. Park. 2005. "Bandwidth Selection for Smooth Backfitting in Additive Models." *Annals of Statistics*, **33**, pp. 1260–1294.

———. 2006. "A Simple Smooth Backfitting Method for Additive Models." *Annals of Statistics*, **34**, pp. 2252–2271.

Mammen, E., C. Rothe, and M. Schienle. 2012. "Nonparametric Regression with Nonparametrically Generated Covariates." *Annals of Statistics*, **40**, pp. 1132–1170.

Mammen, E., B. Støve, and D. Tjøstheim. 2009. "Nonparametric Additive Models for Panels of Time Series." *Econometric Theory*, **25**, pp. 442–481.

Mammen, E., and K. Yu. 2009. "Nonparametric Estimation of Noisy Integral Equations of the Second Kind." *Journal of the Korean Statistical Soceity*, **38**, pp. 99–110.

Meier, L., S. van de Geer, and P. Bühlmann. 2009. "High-Dimensional Additive Modeling." *Annals of Statistics*, **37**, pp. 3779–3821.

Newey, W., J. Powell, and F. Vella. 1999. "Nonparametric Estimation of Triangular Simultaneous Equations Models." *Econometrica*, **67**(3), pp. 565–603.

Newey, W. K., and J. L. Powell. 2003. "Instrumental Variables Estimation for Nonparametric Models." *Econometrica*, **71**, pp. 1565–1578.

Nielsen, J., and S. Sperlich. 2005. "Smooth Backfitting in Practice." *Journal of the Royal Statistical Society B*, **67**, pp. 43–61.

Opsomer, J. D.. 2000. "Asymptotic Properties of Backfitting Estimators." *Journal of Multinomial Analysis*, **73**, pp. 166–179.

Opsomer, J. D., and D. Ruppert. 1997. "Fitting a Bivariate Additive Model by Local Polynomial Regression." *Annals of Statistics*, **25**, pp. 186–211.

Park, B. U., W. C. Kim, and M. Jones. 2002. "On Local Likelihood Density Estimation." *Annals of Statistics*, **30**, pp. 1480–1495.

Park, B. U., E. Mammen, W. Härdle, and S. Borak. 2009. "Time Series Modelling with Semiparametric Factor Dynamics." *Journal of the American Statistical Association*, **104**, pp. 284–298.

Rasch, G. 1960. *Probabilistic Models for some Intelligence and Attainment Tests.* Chicago: University of Chicago Press.

———. 1961. "On General Law and the Meaning of Measurement in Psychology." In *Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4. Berkeley: UC Press and Los Angeles.

Schick, A. 1996. "Root-*n*-Consistent and Efficient Estimation in Semiparametric Additive Regression Models." *Statistics and Probability Letters*, **30**, pp. 45–51.

Schienle, M. 2008. "Nonparametric Nonstationary Regression." Ph.D. thesis, Universität Mannheim.

Stone, C. J. 1985. "Additive Regression and Other Nonparametric Models." *Annals of Statistics*, **13**, pp. 689–705.

———. 1986. "The Dimensionality Reduction Principle for Generalized Additive Models." *Annals of Statistics*, **14**, pp. 590–606.

Wang, Q., and P. C. B. Phillips. 2009. "Structural Nonparametric Cointegrating Regression." *Econometrica*, **77**, pp. 1901–1948.

Yang, L., B. U. Park, L. Xue, and W. Härdle. 2006. "Estimation and Testing for Varying Coefficients in Additive Models with Marginal Integration." *Journal of the American Statistical Association*, **101**, pp. 1212–1227.

Yu, K., E. Mammen, and B. U. Park. 2011. "Semiparametric Regression: Efficiency Gains from Modeling the Nonparametric Part." *Bernoulli*, **17**, pp. 736–748.

Yu, K., B. U. Park, and E. Mammen. 2008. "Smooth Backfitting in Generalized Additive Models." *Annals of Statistics*, **36**, pp. 228–260.

PART IV

MODEL SELECTION AND AVERAGING

CHAPTER 8

NONPARAMETRIC SIEVE REGRESSION: LEAST SQUARES, AVERAGING LEAST SQUARES, AND CROSS-VALIDATION

BRUCE E. HANSEN[†]

8.1. Introduction

One of the most popular nonparametric techniques in applied econometric analysis is sieve regression. A sieve is sequence of finite-dimensional models of increasing complexity. The most common examples of sieve regression are polynomials and splines. For a fixed order of complexity, the model can be estimated by classical (parametric) methods. An important difference with parametric regression is that the order of the sieve (the number of regressors) must be selected. This fundamentally changes both the distributional theory and applied practice.

In this chapter we consider selection and combination of nonparametric sieve regression estimators. We review the concepts of series and sieve approximations, introduce least squares estimates of sieve approximations, and measure the accuracy of the estimators by integrated mean-squared error (IMSE). We show that a critical issue in applications is the order of the sieve, because the IMSE greatly varies across the choice.

We develop the relationship between IMSE and mean-squared forecast error (MSFE), and we introduce the cross-validation criterion as an estimator of MSFE and IMSE. A major theoretical contribution is that we show that selection based on cross-validation is asymptotically equivalent (with respect to IMSE) to estimation based on the infeasible best sieve approximation. This is an important extension of the theory of cross-validation, which currently has only established optimality with respect to conditional squared error.

We also introduce averaging estimators, which are weighted averages of sieve regression estimators. Averaging estimators have lower IMSE than selection estimators. The critical applied issue is the selection of the averaging weights. Following Hansen and Racine (2012) we introduce a cross-validation criterion for the weight vector, and recommend selection of the weights by minimizing this criterion. The resulting estimator—jackknife model averaging (JMA)—is a feasible averaging sieve estimator. We show that the JMA estimator is asymptotically optimal in the sense that it is asymptotically equivalent (with respect to IMSE) to the infeasible optimal weighted average sieve estimator. Computation of the JMA weights is a simple application of quadratic programming. We also introduce a simple algorithm that closely approximates the JMA solution without the need for quadratic programming.

Sieve approximation has a long history in numerical analysis, statistics, and econometrics. See Chui (1992) and de Boor (2001) for numerical properties of splines, Grenander (1981) for the development of the theory of sieves, Li and Racine (2006) for a useful introduction for econometricians, and Chen (2006) for a review of advanced econometric theory.

Nonparametric sieve regression has been studied by Andrews (1991a) and Newey (1995, 1997), including asymptotic bounds for the IMSE of the series estimators.

Selection by cross-validation was introduced by Stone (1974), Allen (1974), Wahba and Wold (1975), and Craven and Wahba (1979). The optimality of cross-validation selection was investigated by Li (1987) for homoskedastic regression and Andrews (1991b) for heteroskedastic regression. These authors established that the selected estimated is asymptotically equivalent to the infeasible best estimator, where "best" is defined in terms of conditional squared error.

Averaging estimators for regression models was introduced by Hansen (2007). A cross-validation (jacknife) method for selecting the averaging weights was introduced by Hansen and Racine (2012).

The organization of this chapter is as follows. Section 8.2 introduces nonparametric sieve regression, and Section 8.3 discusses sieve approximations. Section 8.4 introduces the sieve regression model and least squares estimation. Section 8.5 derives the IMSE of the sieve estimators. Section 8.6 is a numerical illustration of how the sieve order is of critical practical importance. Section 8.7 develops the connection between IMSE and MSFE. Section 8.8 introduces cross-validation for sieve selection. Section 8.9 presents the theory of optimal cross-validation selection. Section 8.10 is a brief discussion of how to preselect the number of models, and Section 8.11 discusses alternative selection criteria. Section 8.12 is a continuation of the numerical example. Section 8.13 introduces averaging regression estimators, and Section 8.14 introduces the JMA averaging weights and estimator. Section 8.15 introduces methods for numerical computation of the JMA weights. Section 8.16 presents an optimality result for JMA weight selection. Section 8.17 is a further continuation of the numerical example. Section 8.18 concludes. Regularity conditions for the theorems are listed in Section 8.19, and the proofs of the theoretical results are presented in Section 8.20. Computer programs that create the numerical work is available on my webpage www.ssc.wisc.edu/~bhansen.

8.2. NonParametric Sieve Regression

Suppose that we observe a random sample (y_i, x_i), $i = 1, \ldots, n$, with y_i real-valued and $x_i \in \mathcal{X}$ possibly vector-valued with $\mathcal{X}$ compact and density $f(x)$. We are interested in estimating the regression of y_i on x_i, that is, the conditional mean $g(x) = \mathbb{E}(y \mid x)$, which is identified almost surely if $\mathbb{E}|y| < \infty$. We can write the regression equation as

$$y_i = g(x_i) + e_i, \tag{8.1}$$

$$\mathbb{E}(e_i \mid x_i) = 0. \tag{8.2}$$

The regression problem is nonparametric when $g(x)$ cannot be summarized by a finite set of parameters.

Note that Eqs. (8.1) and (8.2) do not impose any restrictions on the regression function $g(x)$ or on the regression error e_i (such as conditional homoskedasticity). This is because in a nonparametric context the goal is to be minimalistic regarding parametric assumptions. To develop distributional approximations for estimators, it will be necessary to impose some smoothness and moment restrictions. But these restrictions are technical regularity conditions, not fundamental features of the nonparametric model.

A sieve expansion for $g(x)$ is a sequence of finite-dimensional models $g_m(x)$, $m = 1, 2, \ldots$, with increasing complexity. Particularly convenient are linear sieves, which take the form

$$\begin{aligned} g_m(x) &= \sum_{j=1}^{K_m} z_{jm}(x)\beta_{jm} \\ &= Z_m(x)'\beta_m, \end{aligned}$$

where $z_{jm}(x)$ are (nonlinear) functions of x. The number of terms K_m indexes the complexity of the approximation, and it plays an important role in the theory. Given a sieve expansion $Z_m(x)$, we define the $K_m \times 1$ regressor vector $z_{mi} = Z_m(x_i)$.

An important special case of a sieve is a series expansion, where the terms $z_{jm}(x)$ are not a function of the sieve order m. For example, a polynomial series expansion is obtained by setting $z_j(x) = x^{j-1}$. When the sieve is a series expansion, then the models are nested in the sense that $m_2 > m_1$ implies that $g_{m_2}(x)$ contains $g_{m_1}(x)$ as a special case.

While polynomial series expansions are quite well known, better approximations can be typically achieved by a spline. A spline is a piecewise continuous polynomial, constrained to be smooth up to the order of the polynomial. There is more than one way to write out the basis of a regression spline. One convenient choice takes the form

$$g_m(x) = \sum_{j=0}^{p} x^j \beta_{jm} + \sum_{j=1}^{m} \beta_{p+j}(x - t_j)^p 1\big(x \geq t_j\big). \tag{8.3}$$

Here, p is the order of the polynomial. There are m constants $t_1,\dots,t_m$ called knots which are the join points between the piecewise polynomials. Splines thus have $K_m = p+1+m$ coefficients, and a spline has similar flexibility to a $(p+m)$th-order polynomial. Splines require a rule to determine the location of the knots t_j. A common choice is to set the knots to evenly partition the support of x_i. An alternative is to set the knots to evenly partition the percentiles of the distribution of x_i (that is, if $m = 3$, then set t_1, t_2, and t_3 equal the 25th, 50th, and 75th percentile, respectively).

Typically, the order p of the spline is preselected based on desired smoothness (linear, quadratic, and cubic are typical choices), and the number of knots m are then selected to determine the complexity of the approximation.

If the knots are set evenly, then the sequence of spline sieves with $m = 1,2,3,\dots,$ are non-nested in the sense that $m_2 > m_1$ does not imply that $g_{m_2}(x)$ contains $g_{m_1}(x)$. However, a sequence of splines can be nested if the knots are set sequentially, or if they are set to partition evenly but the number of knots doubled with each sequential sieve, that is, if we consider the sequence $m = 1,2,4,8,\dots$.

In a given sample with n observations, we consider a set of sieves $g_m(x)$ for $m = 1,..,M_n$, where M_n can depend on sample size. For example, the set of sieve expansions could be the set of pth-order polynomials for $p = 1,\dots,M$. Or alternatively, the sieve could be the set of pth-order splines with m knots, for $m = 0,1,\dots,M-1$.

8.3. Sieve Approximation

We have been using the notation β_m to denote the coefficients of the mth sieve approximation, but how are they defined? There is not a unique definition, but a convenient choice is the best linear predictor

$$\begin{aligned}\beta_m &= \underset{\beta}{\operatorname{argmin}}\, \mathbb{E}\big(y_i - z_{mi}'\beta\big)^2 \\ &= \big(\mathbb{E}\big(z_{mi}z_{mi}'\big)\big)^{-1}\mathbb{E}\big(z_{mi}y_i\big). \qquad (8.4)\end{aligned}$$

Given β_m, define the approximation error

$$r_m(x) = g(x) - Z_m(x)'\beta_m,$$

set $r_{mi} = r_m(x_i)$, and define the expected squared approximation error

$$\phi_m^2 = \mathbb{E}r_{mi}^2 = \int r_m(x)^2 f(x)\,dx.$$

ϕ_m^2 measures the quality of $g_m(x)$ as an approximation to $g(x)$ in the sense that a smaller ϕ_m^2 means a better approximation. Note that ϕ_m^2 is the variance of the projection error from the population regression of the true regression function $g(x_i)$ on the sieve regressors z_{mi}:

$$\phi_m^2 = \int g(x)^2 f(x)\,dx - \int g(x) Z_m(x)' f(x)\,dx \left(\int Z_m(x) Z_m(x)' f(x)\,dx \right)^{-1}$$
$$\int Z_m(x) g(x) f(x)\,dx.$$

It therefore follows that for nested series approximations, ϕ_m^2 is monotonically decreasing as K_m increases. That is, larger models mean smaller approximation error.

Furthermore, we can describe the rate at which ϕ_m^2 decreases to zero. As discussed on page 150 of Newey (1997), if $\dim(x) = q$ and $g(x)$ has s continuous derivatives, then for splines and power series there exists an approximation $\beta' z_m(x)$ such that $\left|g(x) - \beta' z_m(x)\right| = O\left(K_m^{-s/q}\right)$, uniformly in x. Thus

$$\phi_m^2 = \inf_\beta \mathbb{E}\left(g(x_i) - \beta' z_m(x_i)\right)^2 \leq \inf_\beta \sup_x \left|g(x) - \beta' z_m(x)\right|^2 \leq O\left(K_m^{-2s/q}\right).$$

This shows that the magnitude of the approximation error depends on the dimensionality and smoothness of $g(x)$. Smoother functions $g(x)$ can be approximated by a smaller number of series terms K_m, so the rate of convergence is increasing in the degree of smoothness.

8.4. Sieve Regression Model and Estimation

As we have described, for each sieve approximation there are a set of regressors z_{mi} and best linear projection coefficient β_m. The sieve regression model is then

$$y_i = z_{mi}' \beta_m + e_{mi}, \tag{8.5}$$

where e_{mi} is a projection error and satisfies

$$\mathbb{E}(z_{mi} e_{mi}) = 0.$$

It is important to recognize that e_{mi} is defined by this construction, and it is therefore inappropriate to *assume* properties for e_{mi}. Rather they should be derived.

Recall that the approximation error is $r_{mi} = r_m(x_i) = g(x_i) - z_{mi}'\beta_m$. Since the true regression (8.1) is $y_i = g(x_i) + e_i$, it follows that the projection error is $e_{mi} = e_i + r_{mi}$, the sum of the true regression error e_i and the sieve approximation error r_{mi}.

The least squares (LS) estimator of Eq. (8.5) is

$$\widehat{\beta}_m = \left(\sum_{i=1}^{n} z_{mi} z'_{mi}\right)^{-1} \sum_{i=1}^{n} z_{mi} y_i,$$
$$\widehat{g}_m(x) = Z_m(x)'\widehat{\beta}_m.$$

Least squares is an appropriate estimator because β_m is defined as the best linear predictor. The least squares estimator is a natural moment estimator of the projection coefficient β_m.

8.5. Integrated Mean Squared Error

As a practical matter, the most critical choice in a series regression is the number of series terms. The choice matters greatly and can have a huge impact on the empirical results.

Statements such as "the number of series terms should increase with the sample size" do not provide any useful guidance for practical selection. Applied nonparametric analysis needs practical, data-based rules. Fortunately, there are sound theoretical methods for data-dependent choices.

The foundation for a data-dependent choice is a (theoretical) criterion that measures the performance of an estimator. The second step is to constuct an estimator of this criterion. Armed with such an estimate, we can select the number of series terms or weights to minimize the empirical criterion.

Thus to start, we need a criterion to measure the performance of a nonparametric regression estimator. There are multiple possible criteria, but one particularly convenient choice is integrated mean-squared error (IMSE). For a sieve estimator $\widehat{g}_m(x)$ the IMSE equals

$$IMSE_n(m) = \int \mathbb{E}\big(\widehat{g}_m(x) - g(x)\big)^2 f(x)\,dx.$$

Using the fact that $\widehat{g}_m(x) - g(x) = z_m(x)'\big(\widehat{\beta}_m - \beta_m\big) - r_m(x)$, we can calculate that

$$\begin{aligned}&\int \big(\widehat{g}_m(x) - g(x)\big)^2 f(x)\,dx \\ &= \int r_m(x)^2 f(x)\,dx - 2\big(\widehat{\beta}_m - \beta_m\big)' \int x_m(x) r_m(x) f(x)\,dx \\ &\quad + \big(\widehat{\beta}_m - \beta_m\big)' \int z_m(x) z_m(x)' f(x)\,dx \big(\widehat{\beta}_m - \beta_m\big).\end{aligned}$$

Note the the first term equals the expected squared approximation error ϕ_m^2. The second term is zero because $\int x_m(z) r_m(z) f(z)\,dz = \mathbb{E}(z_{mi} r_{mi}) = 0$. Defining

$Q_m = \mathbb{E}\big(z_{mi} z'_{mi}\big)$, we can write

$$IMSE_n(m) = \phi_m^2 + \mathrm{tr}\Big[Q_m \mathbb{E}\Big(\big(\widehat{\beta}_m - \beta_m\big)\big(\widehat{\beta}_m - \beta_m\big)'\Big)\Big].$$

Asymptotically, $\mathbb{E}\big(\big(\widehat{\beta}_m - \beta_m\big)\big(\widehat{\beta}_m - \beta_m\big)'\big) \simeq \frac{1}{n} Q_m^{-1} \Omega_m Q_m^{-1}$ where $\Omega_m = \mathbb{E}\big(z_{mi} z'_{mi} \sigma_i^2\big)$ and $\sigma_i^2 = \mathbb{E}\big(e_i^2 \mid x_i\big)$. Making these substitutions, we expect that $IMSE_n(m)$ should be close to

$$IMSE_n^*(m) = \phi_m^2 + \frac{1}{n}\,\mathrm{tr}\big(Q_m^{-1} \Omega_m\big). \tag{8.6}$$

The second term in (8.6) is the integrated asymptotic variance. Under conditional homoskedasticity $\mathbb{E}\big(e_i^2 \mid x_i\big) = \sigma^2$, we have the simplification $\Omega_m = \mathbb{E}\big(z_{mi} z'_{mi}\big)\sigma^2 = Q_m \sigma^2$. Thus in this case $\frac{1}{n}\,\mathrm{tr}\big(Q_m^{-1} \Omega_m\big) = \sigma^2 K_m / n$, a simple function of the number of coefficients and sample size. That is, homoskedasticity implies the following simplification of (8.6):

$$IMSE_n^*(m) = \phi_m^2 + \sigma^2 \frac{K_m}{n}.$$

However, in the general case of conditional heteroskedasticity, (8.6) is the appropriate expression.

Hansen (2012) showed that $IMSE_n(m)$ and $IMSE_n^*(m)$ are uniformly close under quite general regularity conditions, listed in Section 8.19.

Theorem 8.1. *Under Assumption 8.1, uniformly across $m \le M_n$,*

$$IMSE_n(m) = IMSE_n^*(m)(1 + o(1)).$$

This shows that $IMSE_n^*(m)$ is a good approximation to $IMSE_n(m)$.

8.6. The Order of the Approximation Matters

The way that nonparametric methods are often presented, some users may have received the false impression that the user is free to select the order of the approximation m. So long as m increases with n, the method works, right? Unfortunately it is not so simple in practice. Instead, the actual choice of m in a given application can have large and substantive influence on the results.

To illustrate this point, we take a simple numerical example. We consider the following data-generating process.

$$
\begin{aligned}
y_i &= g(x_i) + e_i, \\
g(x) &= a \sin\left(2\pi x + \frac{\pi}{4}\right), \\
x_i &\sim U[0,1], \\
e_i &\sim N(0, \sigma_i^2), \\
\sigma_i^2 &= \sqrt{5} x_i^2,
\end{aligned}
$$

This is a simple normal regression with conditional heteroskedasticity. The parameter a is selected to control the population $R^2 = a^2/(2 + a^2)$, and we vary $R^2 = 0.25$, 0.5, 0.75, and 0.9. We vary the sample size n from 50 to 1000.

We consider estimation of $g(x)$ using quadratic splines, ranging the number of knots m from 1 to 5. For each R^2, n, and m, the integrated mean-squared error (IMSE) is calculated and displayed in Figure 8.1 as a function of sample size using a logarithmic scale. The four displays are for the four values of R^2, and each line corresponds to a different number of knots. Thus each line corresponds to a distinct sieve approximation m. To render the plots easy to read, the IMSE has been normalized by the IMSE of the infeasible optimal averaging estimator. Thus the reported IMSEs are multiples of the infeasible best.

One striking feature of Figure 8.1 is the strong variation with m. That is, for a given R^2 and n, the IMSE varies considerably across estimators. For example, take $n = 200$ and $R^2 = 0.25$. The relative IMSE ranges from about 1.4 (2 knots) to 2.1 (5 knots). Thus the choice really matters. Another striking feature is that the IMSE rankings strongly depend on unknowns. For example, again if $n = 200$ but we consider $R^2 = 0.9$, then the sieve with two knots performs quite poorly with IMSE = 2.9, while the sieve with 5 knots has a relative IMSE of about 1.5.

A third striking feature is that the IMSE curves are U-shaped functions of the sample size n. When they reach bottom, they tend to be the sieve with the lowest IMSE. Thus if we fix R^2 and vary n from small to large, we see how the best sieve is increasing. For example, take $R^2 = 0.25$. For $n = 50$, the lowest IMSE is obtained by the spline with one knot. The one-knot spline has the lowest IMSE until $n = 150$, at which point the two-knot spline has the lowest IMSE. The three-knot spline has lower IMSE for $n \geq 800$. Or, consider the case $R^2 = 0.75$. In this case, the two-knot spline has the lowest IMSE for $n < 100$, while the three-knot spline is best for $100 \leq n \leq 400$, with the four-knot spline for $n \leq 600$.

The overall message is that the order of the series approximation matters, and it depends on features that we know (such as the sample size n) but also features that we do not know. Data-dependent methods for selection of m are essential, otherwise the selection between the estimators is arbitrary.

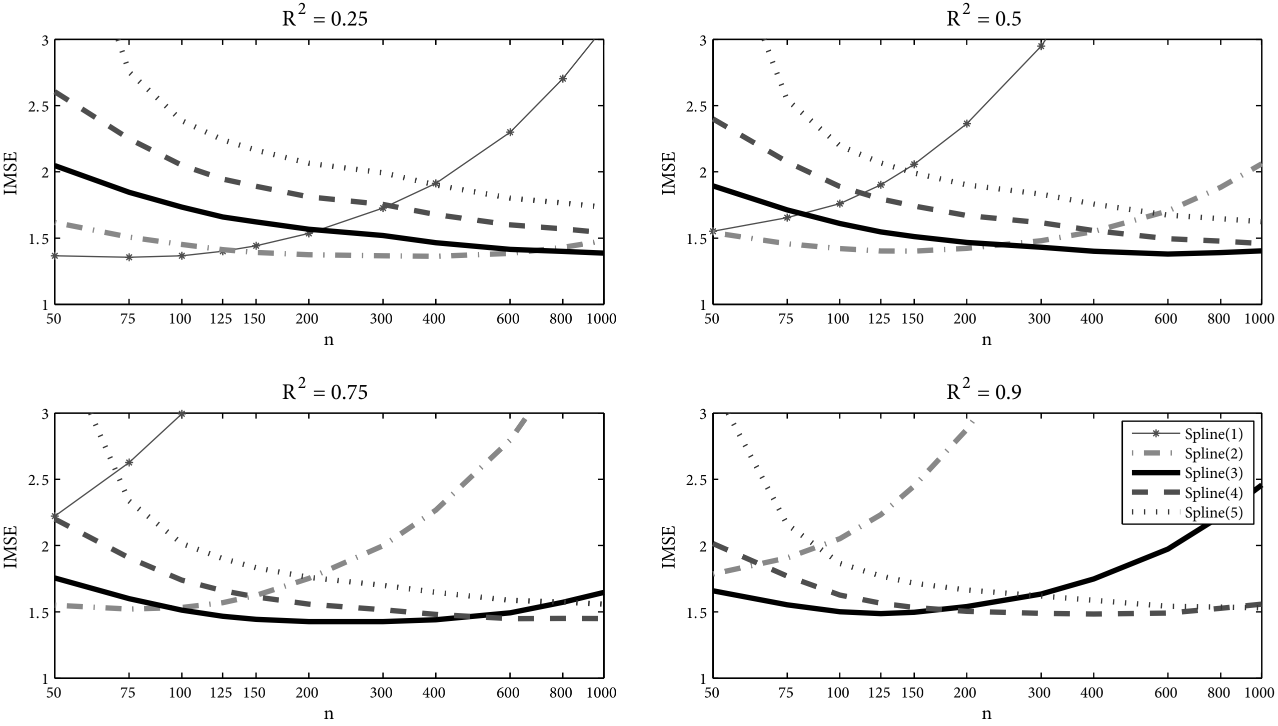


FIGURE 8.1 Integrated Mean-Squared Error of Spline Regression Estimators

8.7. Mean-Squared Forecast Error

A concept related to IMSE is the mean-squared forecast error (MSFE). This is the expected squared error from the prediction of an out-of-sample observation. Specifically, let (y_{n+1}, x_{n+1}) be an out-of-sample observation drawn from the same distribution as the in-sample observations. The forecast of y_{n+1} given x_{n+1} is $\widehat{g}_m(x_{n+1})$. The MSFE is the expected squared forecast error

$$MSFE_n(m) = \mathbb{E}\big(y_{n+1} - \widehat{g}_m(x_{n+1})\big)^2,$$

which depends on the sample size n as well as the estimator $\widehat{g}_m$.

Making the substitution $y_{n+1} = g(x_{n+1}) + e_{n+1}$ and using the fact that e_{n+1} is independent of $g(x_{n+1}) - \widehat{g}_m(x_{n+1})$, we can calculate that the MSFE equals

$$MSFE_n(m) = \mathbb{E}\big(e_{n+1}^2\big) + \mathbb{E}\big(g(x_{n+1}) - \widehat{g}_m(x_{n+1})\big)^2.$$

The second term on the right is an expectation over the random vector x_{n+1} and the estimator $\widehat{g}_m(x)$, which are independent since the estimator is a function only of the in-sample observations. We can write the expectation over x_{n+1} as an integral with respect to its marginal density $f(x)$, thus

$$\begin{aligned} MSFE_n(m) &= \mathbb{E}\big(e_{n+1}^2\big) + \int \mathbb{E}\big(\widehat{g}_m(x) - g(x)\big)^2 f(x)\, dx \\ &= \mathbb{E}\big(e_{n+1}^2\big) + IMSE_n(m). \end{aligned}$$

Thus $MSFE_n(m)$ equals $IMSE_n(m)$ plus $\mathbb{E}\big(e_{n+1}^2\big)$. Note that $\mathbb{E}\big(e_{n+1}^2\big)$ does not depend on the estimator $\widehat{g}_m(x)$. Thus ranking estimators by MSFE and IMSE are equivalent.

8.8. Cross-Validation

Ideally, we want to select the estimator m that minimizes $IMSE_n(m)$ or equivalently $MSFE_n(m)$. However, the true MSFE is unknown. In this section we show how to estimate the MSFE.

Observe that

$$MSFE_n(m) = \mathbb{E}\big(\tilde{e}_{m,n+1}^2\big),$$

where $\tilde{e}_{m,n+1} = y_{n+1} - \widehat{g}_m(x_{n+1})$. This is a prediction error. Estimation is based on the sample (y_i, x_i): $i = 1, \ldots, n$, and the error calculated is based on the out-of-sample observation $n+1$. Thus $MSFE_n(m)$ is the expectation of a squared leave-one-out prediction error from a sample of length $n+1$.

For each observation i, we can create a similar leave-one-out prediction error. For each i we can create a pseudo-prediction error by estimating the coefficients using the observations excluding i. That is, define the leave-one-out estimator

$$\widehat{\beta}_{m,-i} = \left(\sum_{j \neq i} z_{mj} z'_{mj} \right)^{-1} \sum_{j \neq i} z_{mj} y_j \tag{8.7}$$

and prediction error

$$\tilde{e}_{mi} = y_i - z'_{mi} \widehat{\beta}_{m,-i}. \tag{8.8}$$

The only difference between $\tilde{e}_{m,n+1}$ and $\tilde{e}_{mi}$ is that the former is based on the extended sample of length $n+1$ while the latter are based on a sample of length n. Otherwise, they have the same construction. It follows that for each i, $\mathbb{E}\tilde{e}^2_{mi} = MSFE_{n-1}(m)$. Similarly, the sample average, known as the *cross-validation criterion*

$$CV_n(m) = \frac{1}{n} \sum_{i=1}^{n} \tilde{e}^2_{mi}$$

also has mean $MSFE_{n-1}(m)$. This is a natural moment estimator of $MSFE_{n-1}(m)$.

We have established the following result.

Theorem 8.2. $\mathbb{E}CV_n(m) = MSFE_{n-1}(m)$.

As $MSFE_{n-1}(m)$ should be very close to $MSFE_n(m)$, we can view $CV_n(m)$ as a nearly unbiased estimator of $MSFE_n(m)$.

Computationally, the following algebraic relationship is convenient.

Proposition 8.1. $\tilde{e}_{mi} = \hat{e}_{mi}(1 - h_{mi})^{-1}$, *where* $\hat{e}_{mi} = y_i - z'_{mi}\widehat{\beta}_m$ *are the least squares residuals and* $h_{mi} = z'_{mi}\left(\sum_{i=1}^{n} z_{mi} z'_{mi}\right)^{-1} z_{mi}$ *are known as the leverage values.*

While Proposition 8.1 is well known, we include a complete proof in Section 8.20 for completeness.

Proposition 8.1 directly implies the simple algebraic expression

$$CV_n(m) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{e}^2_{mi}}{(1 - h_{mi})^2}. \tag{8.9}$$

This shows that for least squares estimation, cross-validation is a quite simple calculation and does not require the explicit leave-one-out operations suggested by (8.7).

The estimator $\widehat{m}$ that is selected by cross-validation is the one with the smallest value of $CV(m)$. We can write this as

$$\widehat{m} = \underset{1 \leq m \leq M_n}{\operatorname{argmin}}\ CV_n(m).$$

Computationally, we estimate each series regression $m = 1, \ldots, M_n$, compute the residuals $\hat{e}_{mi}$ for each, determine the CV criterion $CV_n(m)$ using (8.9), and then find $\widehat{m}$ as the value that yields the smallest value of $CV_n(m)$.

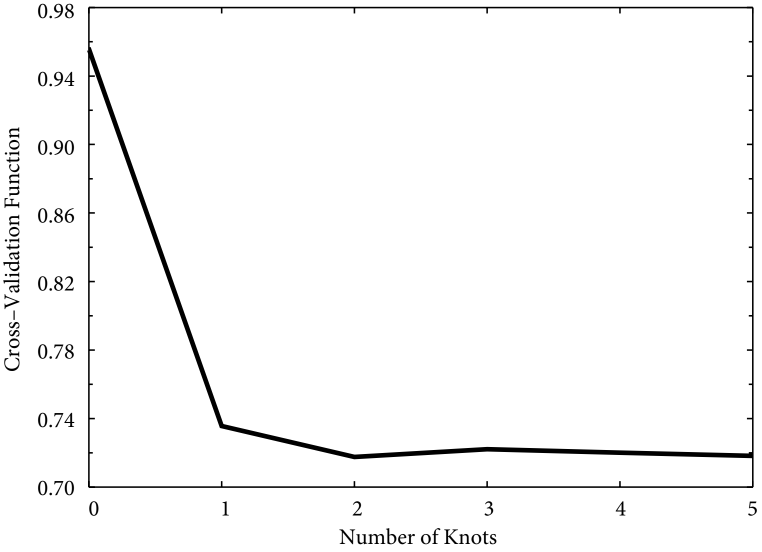


FIGURE 8.2 Typical Cross-Validation Function, $n = 200$

It is useful to plot $CV_n(m)$ against m to visually check if there are multiple local minima or flat regions. In these cases some statisticians have argued that it is reasonable to select the most parsimonious local minima or the most parsimonious estimator among near-equivalent values of the CV function. The reasons are diverse, but essentially the cross-validation function can be quite a noisy estimate of the IMSE, especially for high-dimensional models. The general recommendation is to augment automatic model-selection with visual checks and judgment.

To illustrate, Figure 8.2 plots the cross-validation function for one of the samples from Section 8.6. The cross-validation function is sharply decreasing until 2 knots, then flattens out, with the minimum $m = 2$ knots. In this particular example, the sample was drawn from the DGP of Section 8.6 with $n = 200$ and $R^2 = 0.5$. From Figure 8.1 we can see that the lowest IMSE is obtained by $m = 2$, so indeed the CV function is a constructive guide for selection.

8.9. Asymptotic Optimality of Cross-Validation Selection

Li (1987), Andrews (1991b) and Hansen and Racine (2012) have established conditions under which the CV-selected estimator is asymptotically optimal, in the sense that the selected model is asymptotically equivalent to the infeasible optimum. The criterion they used to assess optimality is the conditional squared error fit

$$R_n(m) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\Big(\big(\widehat{g}_m(x_i) - g(x_i)\big)^2 \mid X\Big), \tag{8.10}$$

where $X = \{x_1, \ldots, x_n\}$. This is similar to IMSE, but only assesses fit on the support points of the data. In contrast, the literature on sieve approximations focuses on IMSE. We now extend the asymptotic optimality theory and show that the CV-selected estimator is asymptotically optimal with respect to IMSE.

Theorem 8.3. *Under Assumptions 8.1, 8.2, and 8.3, as $n \to \infty$,*

$$\left|\frac{IMSE_n(\widehat{m})}{\inf\limits_{1\le m\le M_n} IMSE_n(m)}\right| \xrightarrow{p} 1.$$

The assumptions and proof are presented in Sections 8.19 and 8.20, respectively.

Theorem 8.3 shows that in large samples, the IMSE of the CV-selected estimator $\widehat{g}_{\widehat{m}}(x)$ is equivalent to the IMSE of the infeasible best estimator in the class $\widehat{g}_m(x)$ for $1 \le m \le M_n$. This is an oracle property for cross-validation selection.

A critical assumption for Theorem 8.3 is that $\phi_m^2 > 0$ for all $m < \infty$ (Assumption 8.2 in Section 8.20). Equivalently, the approximation error is nonzero for all finite-dimensional models; that is, all models are approximations. If, instead, one of the finite-dimensional models is the true conditional mean (so that $\phi_m^2 = 0$ for some model m), then cross-validation asymptotically over-selects the model order with positive probability and is thus asymptotically suboptimal. In this context consistent model selection methods (such as BIC) are optimal. This classification was carefully articulated in the review paper by Shao (1997). Some researchers refer to cross-validation as a *conservative* selection procedure (it is optimal for the broad class of nonparametric models) and to BIC as a consistent selection procedure (it selects the correct model when it is truly finite-dimensional).

8.10. PreSelection of the Number of Models

To implement cross-validation selection, a user first has to select the set of models $m = 1, \ldots, M_n$ over which to search. For example, if using a power series approximation, a user has to first determine the highest power, or if using a spline, a user has to determine the order of the spline and the maximum number of knots. This choice affects the results, but unfortunately there is no theory about how to select these choices. What we know is that the assumptions restrict both the number of estimated parameters in each model K_m and the number of models M_n relative to sample size.

Specifically, Assumption 8.1.5 specifies that for a power series $K_m^4/n = O(1)$ and for a spline sieve $K_m^3/n = O(1)$, uniformly for $m \leq M_n$. These conditions may be stronger than necessary, but they restrict the number of estimated parameters to be increasing at a rate much slower than sample size. Furthermore, Assumption 8.3.2 allows non-nested models, but controls the number of models. While these conditions do not give us precise rules for selecting the initial set of models, they do suggest that we should be reasonably parsimonious and not too aggressive in including highly parameterized models.

Unfortunately, these comments still do not give precise guidance on how to determine the number of models M_n. It may be a useful subject for future research to construct and justify data-dependent rules for determining M_n.

8.11. Alternative Selection Criteria

We have discussed the merits of cross-validation to select the sieve approximation, but many other selection methods have been proposed. In this section we briefly describe the motivation and properties of some of these alternative criteria.

The Mallows criterion (Mallows, 1973)

$$Mallows(m) = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_{mi}^2 + 2\tilde{\sigma}^2 K_m$$

with $\tilde{\sigma}^2$ a preliminary estimate of $\mathbb{E}\left(e_i^2\right)$ is an alternative estimator of the IMSE under the additional assumption of conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid x_i\right) = \sigma^2$. Li (1987) provided conditions under which Mallows selection is asymptotically optimal, but Andrews (1991b) shows that its optimality fails under heteroskedasticity.

The Akaike information criterion (Akaike, 1973)

$$AIC(m) = n\log\left(\frac{1}{n}\sum_{i=1}^{n} \hat{e}_{mi}^2\right) + 2K_m$$

is an estimate of the Kullback-Leibler divergence between the estimated Gaussian model and the true model density. AIC selection has asymptotic properties that are similar to those of Mallows selection, in that it is asymptotically optimal under conditional homoskedasticity but not under heteroskedasticity.

The corrected AIC (Hurvich and Tsai, 1989)

$$AIC_c(m) = AIC(m) + \frac{2K_m(K_m+1)}{n - K_m - 1}$$

is a finite-sample unbiased estimate of the Kullback–Leibler divergence under the auxiliary assumption that the errors e_i are independent and Gaussian. Its asymptotic properties are the same as AIC, but has improved finite-sample performance, especially when the model dimension K_m is high relative to sample size.

The Bayesian information criterion (Schwarz, 1978)

$$BIC(m) = n\log\left(\frac{1}{n}\sum_{i=1}^{n}\hat{e}_{mi}^2\right) + \log(n)K_m$$

is an approximation to the log posterior probability that model m is the true model, under the auxiliary assumption that the errors are independent and Gaussian, the true model is finite-dimension, the models have equal prior probability, and priors for each model m are diffuse. BIC selection has the property of *consistent model selection*: When the true model is a finite-dimensional series, BIC will select that model with probability approaching one as the sample size increases. However, when there is no finite-dimensional true model, then BIC tends to select overly parsimonious models (based on IMSE).

The above methods are all information criteria, similar in form to cross-validation. A different approach to selection is the class of penalized least squares estimators. Let z_i denote the $K_n \times 1$ vector of all potential regressors in all models, let $\beta = (\beta_1, \ldots, \beta_{K_n})$ denote its projection coefficient, and define the penalized least squares criteria

$$P_n(\beta,\lambda) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - z_i'\beta\right)^2 + \sum_{j=1}^{K_n} p_\lambda\left(\beta_j\right)$$

and the PLS estimator

$$\widehat{\beta}_\lambda = \underset{\beta}{\operatorname{argmin}}\, P_n(\beta,\lambda),$$

where $p_\lambda(u)$ is a non-negative symmetric penalty function and λ is a tuning parameter.

The choice of $p_\lambda(u)$ determines the estimator. In the recent literature a popular choice is $p_\lambda(|u|) = \lambda|u|$ which yields the LASSO (least absolute shrinkage and selection operator) estimator, proposed by Tibshirani (1996). Different variants of LASSO have been proposed, including SCAD (smoothly clipped absolute deviation) (Fan and Li, 2001) and the adaptive LASSO (Zou, 2006).

PLS estimators are generally appropriate when the dimension of z_i is high (some estimators such as the LASSO are defined even when K_n exceeds n). The LASSO family are selection methods as the estimators $\widehat{\beta}_\lambda$ typically set most individual coefficients to zero. The nonzero coefficient estimates are the selected variables, and the zero coefficient estimates are the excluded variables. SCAD and the adaptive LASSO have optimality (oracle) properties when the true regression function is *sparse*, meaning that the true regression function is a finite-dimensional series. When the true regression function is not sparse, the properties of LASSO selection are unclear.

Among these methods, selection by cross-validation is uniquely the only method that is asymptotically optimal for general nonparametric regression functions and unknown conditional heteroskedasticity. Most of the other selection methods explicitly or implicity rely on conditional homoskedasticity, and some of the methods rely on sparsity (finite-dimensionality), neither of which are generally appropriate for nonparametric estimation.

8.12. Numerical Simulation

We return to the simulation experiment introduced in Section 8.6. Recall that we reported the integrated mean-squared error of a set of least squares estimates of a quadratic spline with given knots. Now we compare the IMSE of estimators that select the number of knots. We consider CV selection and compare its performance with selection based on the Akaike information criterion (Akaike, 1973) and the Hurvich–Tsai (Hurvich and Tsai, 1989) corrected AIC.

For all methods, we estimate nonparametric quadratic splines with knots $m = 0, 1, \ldots, M_n$ with $M_n = 4n^{0.15}$. The selection criteria were calculated for each set of knots, and the model was selected with the lowest value of the criteria.

We report the IMSE of the three methods in Figure 8.3 (along with the IMSE of the JMA method, to be discussed below). Again, the IMSE is normalized by the IMSE of the infeasible best averaging estimator, so all results are relative to this infeasible optimum.

One striking feature of this figure is that the three methods (CV, AIC, and AIC$_c$) have similar performance for $n \geq 100$, though CV has slightly lower IMSE, especially for small n.

Another striking feature is that for $n \geq 100$, the IMSE of the selection methods is relatively unaffected by sample size n and the value of R^2. This is especially important when contrasted with Figure 1, where we found that the IMSE of individual sieve estimators depend greatly upon n and R^2. This is good news, it shows that the selection methods are adapting to the unknown features of the sampling distribution.

8.13. Averaging Regression

Let $w = (w_1, w_2, \ldots, w_M)$ be a set of non-negative weights that sum to one, $\sum_{m=1}^{M} w_m = 1$. An averaging LS estimator is

$$\widehat{g}_w(x) = \sum_{m=1}^{M} w_m \widehat{g}_m(x). \tag{8.11}$$

The averaging estimator includes the mth least squares estimator as a special case by setting w to equal the unit vector with a weight of 1 in the mth place.

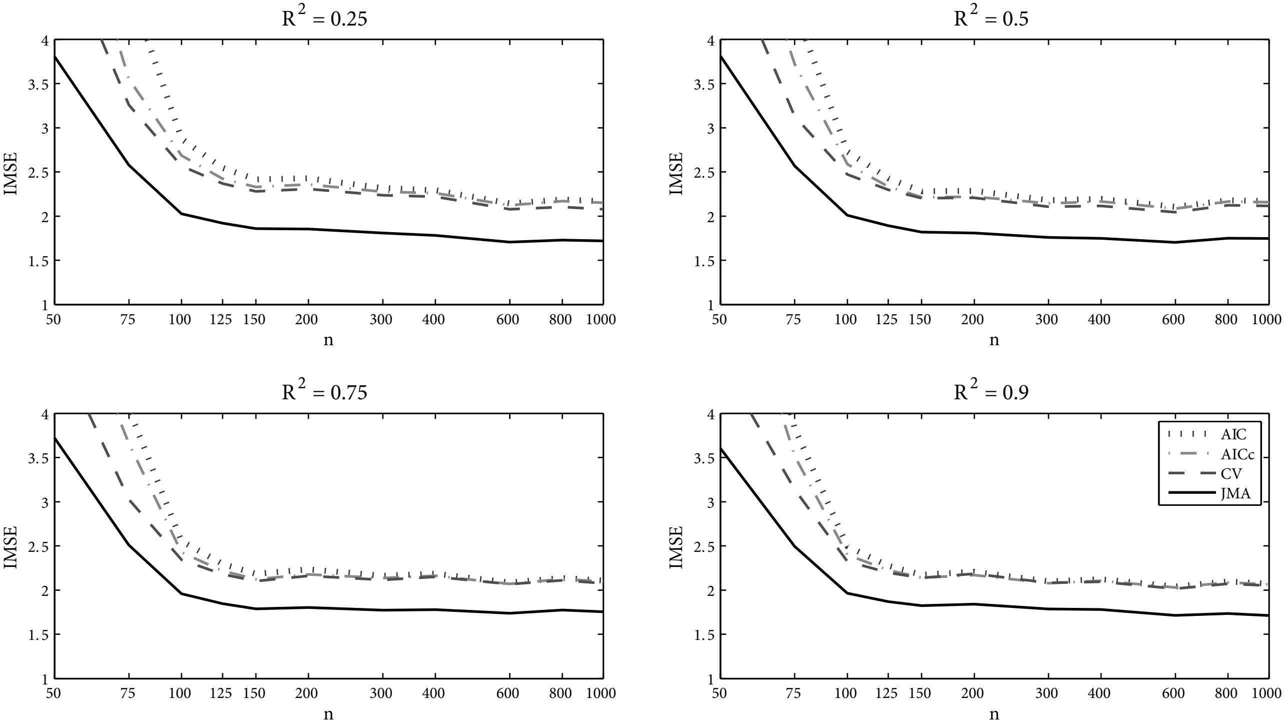


FIGURE 8.3 Integrated Mean-Squared Error, Selection and Averaging Estimators

For example, consider a set of spline estimators with $m = 0, 1, 2$, and 3 knots. The averaging estimator takes an average of these four estimators. In general, averaging is a smoother function of the data than selection, and smoothing generally reduces variance. The reduction in variance can result in estimators with lower IMSE.

We define the IMSE of the averaging estimator as

$$IMSE_n(w) = \int \mathbb{E}\big(\widehat{g}_w(x) - g(x)\big)^2 f(x)\, dx,$$

which is a function of the weight vector.

It is recommended to constrain the weights w_m to be non-negative, that is, $w_m \geq 0$. In this case the weight vector w lies on $\mathcal{H}$, the unit simplex in $\mathbb{R}^{M_n}$. This restriction may not be necessary, but some bounds on the weights are required. Hansen and Racine (2012) suggested that in the case of nested models, non-negativity is a necessary condition for admissibility, but they made a technical error. The actual condition is that $0 \leq \sum_{j=m}^{M} w_j \leq 1$, which is somewhat broader. (I thank Guido Kuersteiner and Ryo Okui for pointing out this error to me.) It is unclear if this broader condition is compatible with the optimality theory, or what restrictions are permissible in the case of non-nested models.

Hansen (2012) provides an approximation to the IMSE of an averaging estimator.

Theorem 8.4. *Under Assumptions 8.1, 8.2, and 8.4, uniformly across $w \in \mathcal{H}$,*

$$IMSE_n(w) = IMSE_n^*(w)(1 + o(1)),$$

where

$$\begin{aligned} IMSE_n^*(w) &= \sum_{m=1}^{M_n} w_m^2 \left(\phi_m^2 + \frac{1}{n} \operatorname{tr}\big(Q_m^{-1} \Omega_m\big) \right) \\ &\quad + 2 \sum_{\ell=1}^{M_n} \sum_{m=1}^{\ell-1} w_\ell w_m \left(\phi_\ell^2 + \frac{1}{n} \operatorname{tr}\big(Q_m^{-1} \Omega_m\big) \right) \\ &= \sum_{m=1}^{M_n} w_m^* n \phi_m^2 + \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\big(Q_m^{-1} \Omega_m\big) \qquad (8.12) \end{aligned}$$

and

$$w_m^* = w_m^2 + 2 w_m \sum_{\ell=1}^{m-1} w_\ell, \qquad (8.13)$$

$$w_m^{**} = w_m^2 + 2 w_m \sum_{\ell=m+1}^{M_n} w_\ell. \qquad (8.14)$$

8.14. JMA for Averaging Regression

The method of cross-validation for averaging regressions is much the same as for selection. First, note that the discussion about the equivalence of mean-square forecast error (MSFE) and IMSE from Section 8.7 is not specific to the estimation method. Thus it equally applies to averaging estimators—namely the averaging forecast of y_{n+1} given x_{n+1} is $\widehat{g}_w(x_{n+1})$, with MSFE

$$\begin{aligned} MSFE_n(w) &= \mathbb{E}\left(y_{n+1} - \widehat{g}_w(x_{n+1})\right)^2 \\ &= \mathbb{E}\left(e_{n+1}^2\right) + IMSE_n(w), \end{aligned}$$

where the second equality follows by the same discussion as in Section 8.7.

Furthermore, the discussion in Section 8.8 about estimation of MSFE by cross-validation is also largely independent of the estimation method, and thus applies to averaging regression. There are some differences, however, in the algebraic implementation. The leave-one-out averaging prediction errors are

$$\tilde{e}_{wi} = \sum_{m=1}^{M} w_m \tilde{e}_{mi}$$

where, as before, $\tilde{e}_{mi}$ is defined in (8.8) and Proposition 8.1. The cross-validation function for averaging regression is then

$$\begin{aligned} CV_n(w) &= \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{wi}^2 \\ &= \sum_{m=1}^{M}\sum_{\ell=1}^{M} w_m w_\ell \left(\frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{mi}\tilde{e}_{\ell i}\right) \\ &= w' S w \end{aligned}$$

where S is an $M \times M$ matrix with $m\ell$th entry

$$\begin{aligned} S_{m\ell} &= \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_{mi}\tilde{e}_{\ell i} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{e}_{mi}\hat{e}_{\ell i}}{(1-h_{mi})(1-h_{\ell i})}. \end{aligned}$$

with $\hat{e}_{mi}$ the least squares residuals for the mth estimator, and the second equality uses Proposition 8.1.

$CV_n(w)$ is also the jackknife estimator of the expected squared error, and thus Hansen and Racine (2012) call $CV_n(w)$ the jackknife model averaging (JMA) criterion.

8.15. Computation

The cross-validation or jackknife choice of weight vector w is the one that minimizes the cross-validation criterion $CV_n(w) = w'Sw$. Since the weights w_m are restricted to be non-negative and sum to one, the vector w lies on the M-dimension unit simplex $\mathcal{H}$, so we can write this problem as

$$\widehat{w} = \underset{w \in \mathcal{H}}{\operatorname{argmin}}\, w'Sw.$$

The weights $\widehat{w}$ are called the JMA weights, and when plugged into the estimator (8.11) they yield the JMA nonparametric estimator

$$\widehat{g}_w(x) = \sum_{m=1}^{M} \widehat{w}_m \widehat{g}_m(x). \tag{8.15}$$

Since the criterion is quadratic in w and the weight space $\mathcal{H}$ is defined by a set of linear equality and inequality restrictions, this minimization problem is known as a quadratic programming problem. In matrix programming languages, solution algorithms are available. For example, $\widehat{w}$ can be easily solved using the qprog command in GAUSS, the quadprog command in MATLAB, or the quadprog command in R.

In other packages, quadratic programming may not be available. However, it is often possible to call the calculation through an external call to a compatible language (for example, calling R from within STATA). This, however, may be rather cumbersome.

However, it turns out that $\widehat{w}$ can be found using a relatively simple set of linear regressions. First, let $\tilde{g}_i = (\tilde{g}_{1i}, \ldots, \tilde{g}_{Mi})'$ be the $M \times 1$ vector of leave-one-out predicted values for the ith observation. Then note that $\tilde{e}_{wi} = y_i - \tilde{g}_i' w$, so the CV criterion can be written as

$$\begin{aligned} CV_n(w) &= \frac{1}{n} \sum_{i=1}^{n} \hat{e}_{wi}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \tilde{g}_i' w \right)^2. \end{aligned}$$

This is the sum-of-squared error function from a regression of y_i on the vector $\tilde{g}_i$, with coefficients w. Thus the problem of solving for $\hat{w}$ is algebraically equivalent to a constrained least squares regression of y_i on $\tilde{g}_i$. We can write the least squares regression as

$$y_i = \tilde{g}_i' w + \tilde{e}_{wi}$$

or in vector notation

$$y = \widetilde{G} w + \tilde{e}_w,$$

where $\widetilde{G}$ is an $n \times M$ matrix whose m'th columnn are the leave-one-out predicted values from the mth series approximation.

The simple unconstrained least squares estimator of w

$$\widetilde{w} = \left(\widetilde{G}'\widetilde{G}\right)^{-1}\widetilde{G}'y \tag{8.16}$$

will satisfy neither the summing up nor non-negativity constraints. To impose the constraint that the coefficients sum to one, letting $\mathbf{1}$ denote an $M \times 1$ vector of ones, then the least squares estimator subject to the constraint $\mathbf{1}'w = 1$ is

$$\bar{w} = \widetilde{w} - \left(\widetilde{G}'\widetilde{G}\right)^{-1}\mathbf{1}\left(\mathbf{1}'\left(\widetilde{G}'\widetilde{G}\right)^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}'\widetilde{w} - 1\right). \tag{8.17}$$

Alternatively, subtract $\tilde{g}_{Mi}$ from y_i and $\tilde{g}_{1i}, \ldots, \tilde{g}_{M-1,i}$ and run the regression

$$y_i - \tilde{g}_{Mi} = \bar{w}_1\left(\tilde{g}_{1i} - \tilde{g}_{Mi}\right) + \bar{w}_2\left(\tilde{g}_{2i} - \tilde{g}_{Mi}\right) + \cdots + \bar{w}_{M-1}\left(\tilde{g}_{M-1,i} - \tilde{g}_{Mi}\right) + \tilde{e}_{wi} \tag{8.18}$$

and then set $\bar{w}_M = 1 - \sum_{m=1}^{M-1} \bar{w}_m$. Equations (8.17) and (8.18) are algebraically equivalent methods to compute $\bar{w}$.

While the weights $\bar{w}$ will sum to one, they will typically violate the non-negativity constraints and thus are not a good estimator. However, a simple iterative algorithm will convert $\bar{w}$ to the desired $\widehat{w}$. Here are the steps.

1. If $\bar{w}_m \geq 0$ for all m, then $\widehat{w} = \bar{w}$ and stop.
2. If $\min_m \bar{w}_m < 0$, find the index $\bar{m}$ with the most negative weight $\bar{w}_{\bar{m}}$ (e.g., $\bar{m} = \operatorname{argmin} \bar{w}_m$).
3. Remove the estimator $\bar{m}$ from the set of M estimators. We are left with a set of $M - 1$ estimators, with $\widetilde{G}$ an $n \times (M - 1)$ matrix.
4. Recompute $\widetilde{w}$ and $\bar{w}$ in (8.16) and (8.17) using this new $\widetilde{G}$.
5. Go back to step 1 and iterate until all weights are non-negative.

This is a simple algorithm and has at most M iteration steps, where M is the number of initial estimators and is thus quite efficient. It is simple enough that it can be computed using simple least squares methods and thus can be used in many packages.

8.16. Asymptotic Optimality of JMA Averaging

Hansen and Racine (2012) have established conditions under which the JMA weights are asymptotically optimal, in the sense that the selected averaging estimator is asymptotically equivalent to the infeasible optimal weights. They established optimality with respect to the conditional squared error fit (8.10). We now show that this can be extended to optimality with respect to IMSE.

As in Hansen (2007) and Hansen and Racine (2012), we only establish optimality with respect to a discrete set of weights. For some integer $N \geq 1$, let the weights w_j take values from the set $\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\}$, and let $\mathcal{H}_n$ denote the subset of the unit simplex $\mathcal{H}$ restricted to these points. If N is large, then this is not restrictive. This restriction is for technical reasons and does not affect how the method is implemented in practical applications.

Theorem 8.5. *Under Assumptions 8.1–8.4, as $n \to \infty$, we have*

$$\left| \frac{IMSE_n(\widehat{w})}{\inf\limits_{w \in \mathcal{H}_n} IMSE_n(w)} \right| \xrightarrow{p} 1.$$

The assumptions and proof are presented in Sections 8.19 and 8.20, respectively.

Theorem 8.5 shows that in large samples, the IMSE of the JMA estimator $\widehat{g}_{\widehat{w}}(x)$ is equivalent to the IMSE of the infeasible best estimator in the class $\widehat{g}_w(x)$ for $w \in \mathcal{H}_n$. This is an oracle property for weight selection by cross-validation.

8.17. Numerical Simulation

We return to the simulation experiment introduced in Sections 8.6 and 8.12. Now we add the JMA estimator (8.15). The IMSE of the estimator is plotted in Figure 8.3 along with the other estimators. The IMSE of the JMA estimator is uniformly better than the other estimators, with the difference quite striking.

The plots display the IMSE relative to the IMSE of the infeasible optimal averaging estimator. The optimality theory (Theorem 8.5) suggests that the relative IMSE of the JMA estimator should approach one as the sample size n diverges. Examining the figures, we can see that the IMSE of the estimator is converging extremely slowly to this asymptotic limit. This suggests that while the JMA is “asymptotically” optimal, there is considerable room for improvement in finite samples.

We illustrate implementation with the simulated sample ($n = 200$) from Section 8.12. We report the cross-validation function and JMA weights in Table 8.1. As we saw in Figure 8.2, the CV function is minimized at $m = 2$. However, the value of the CV function is quite flat for $m \geq 2$, and in particular its value at $m = 5$ is nearly identical to $m = 2$. This means that cross-validation ranks $m = 2$ and $m = 5$ quite similarly. The JMA weights account for this. Note that JMA divides the weight between $m = 1$, $m = 2$, and $m = 5$, rather than putting all the weight on a single estimator. The estimators are plotted (along with the true conditional mean $g(x)$) in Figure 8.4. Both estimators are close to the true $g(x)$.

Table 8.1 Cross-Validation Function and JMA Weights

	$m=0$	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$CV_n(m)$	0.955	0.735	0.717	0.722	0.720	0.718
$\widehat{w}_m$	0.02	0.17	0.35	0.00	0.00	0.46

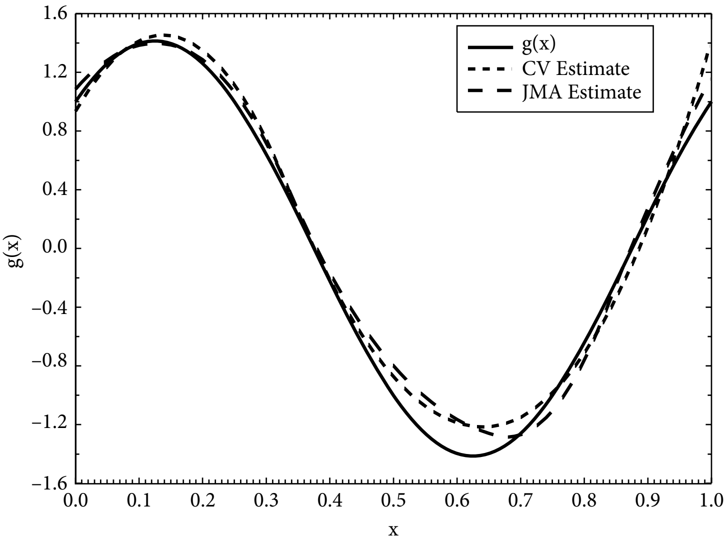


FIGURE 8.4 CV and JMA Feasible Series Estimators

8.18. Summary

Sieves are routinely used in applied econometrics to approximate unknown functions. Power series and splines are particularly popular and convenient choices. In all applications, the critical issue is selecting the order of the sieve. The choice greatly affects the results and the accuracy of the estimates. Rules of thumb are insufficient because the ideal choice depends on the unknown function to be estimated.

In regression estimation, a simple, straightforward and computationally easy method for selecting the sieve approximation is cross-validation. The method is also asymptotically optimal, in the sense that the CV-selected estimator is asymptotically equivalent to the infeasible best-fitting estimator, when we evaluate estimators based on IMSE (integrated mean-squared error).

Further improvements can be obtained by averaging. Averaging estimators reduce estimation variance and thereby IMSE. Selection of the averaging weights is analogous to the problem of selection of the order of a sieve approximation, and a feasible method is again cross-validation. Numerical computation of the averaging weights is simple using quadratic programming. Good approximations can be obtained by a simple iterative algorithm. The JMA weights selected by cross-validation are asymptotically optimal in the sense that the fitted averaging estimator is asymptotically equivalent (with respect to IMSE) to the infeasible best weighted average.

8.19. Regularity Conditions

In this section we list the regularity conditions for the theoretical results.

Assumption 8.1.

1. *The support $\mathcal{X}$ of x_i is a Cartesian product of compact connected intervals on which the density $f(x)$ is bounded away from zero.*
2. *$g(x)$ is continuously differentiable on $x \in \mathcal{X}$.*
3. *For some $\alpha > 0$, $\eta > 0$, and $\psi < \infty$, for all $\ell' Q_m \ell = 1$ and $0 \le u \le \eta$, $\sup_m \mathbb{P}\left(\left|\ell' z_{mi}\right| \le u\right) \le \psi u^{\alpha}$.*
4. $0 < \underline{\sigma}^2 \le \sigma_i^2 \le \overline{\sigma}^2 < \infty$.
5. $\max_{m \le M_n} K_m^4/n = O(1)$ *for a power series, or* $\max_{m \le M_n} K_m^3/n = O(1)$ *for a spline sieve.*

Assumption 8.2. $\phi_m^2 > 0$ *for all* $m < \infty$.

The role of Assumption 8.1.1 is to ensure that the expected design matrix Q_m is uniformly invertible. Assumption 8.1.2 is used to ensure that $r_m(x)$ is uniformly bounded. Assumption 8.1.3 is unusual, but is used to ensure that moments of the inverse sample design matrix $\left(n^{-1}\sum_{i=1}^n z_{mi} z_{mi}'\right)^{-1}$ exist. Assumption 8.1.4 bounds the extent of conditional heteroskedasticity, and Assumption 8.1.5 restricts the complexity of the fitted models.

Assumption 8.2 is quite important. It states that the approximation error is nonzero for all finite-dimensional models; thus all models are approximations. This is standard in the nonparametrics optimality literature. One implication is that $\xi_n = \inf_m nIMSE_n^*(m) \longrightarrow \infty$ as $n \to \infty$.

Let $q_{jn} = \#\{m : K_m = j\}$ be the number of models which have exactly j coefficients, and set $\overline{q}_n = \max_{j \le M_n} q_{jn}.$ This is the largest number of models of any given dimension. For nested models, then $\overline{q}_n = 1$, but when the models are non-nested then $\overline{q}_n$ can exceed one.

Assumption 8.3. *For some $N \geq 1$*

1. $\sup_i \mathbb{E}\left(e_i^{4(N+1)} \mid x_i\right) < \infty$.
2. $\overline{q}_n = o(\xi_n^{1/N})$ *where* $\xi_n = \inf_m nIMSE_n^*(m)$.
3. $\max_{m \leq M_n} \max_{i \leq n} h_{mi} \longrightarrow 0$ *almost surely.*

Assumption 8.3.1 is a strengthing of Assumption 8.1.4. Assumption 8.3.2 allows for non-nested models, but bounds the number of models. Assumption 8.3.3 states that the design matrix cannot be too unbalanced. Under our conditions it is easy to show that $\max_{m \leq M_n} \max_{i \leq n} h_{mi} = o_p(1)$. The technical strengthening here is to almost sure convergence.

Assumption 8.4.

1. $z_m(x)$ *is either a spline or power series and is nested.*
2. $g(x)$ *has s continuous derivatives on* $x \in \mathcal{X}$ *with* $s \geq q/2$ *for a spline and* $s \geq q$ *for a power series.*

8.20. Technical Proofs

Proof of Proposition 8.1. The key is the Sherman–Morrison formula (Sherman and Morrison, 1950), which states that for nonsingular $\boldsymbol{A}$ and vector $\boldsymbol{b}$ we have

$$\left(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1 - \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

This can be verified by premultiplying the expression by $\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'$ and simplifying.

Let $\boldsymbol{Z}_m$ and $\boldsymbol{y}$ denote the matrices of stacked regressors and dependent variable so that the LS estimator is $\widehat{\beta}_m = \left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1}\boldsymbol{Z}_m'\boldsymbol{y}$. An application of the Sherman–Morrison formula yields

$$\begin{aligned}\left(\sum_{j \neq i} z_{mj} z_{mj}'\right)^{-1} &= \left(\boldsymbol{Z}_m'\boldsymbol{Z}_m - z_{mi}z_{mi}'\right)^{-1} \\ &= \left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1} + (1 - h_{mi})^{-1}\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1} z_{mi}z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1}.\end{aligned}$$

Thus

$$\begin{aligned}\tilde{e}_{mi} &= y_i - z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m - z_{mi}z_{mi}'\right)^{-1}\left(\boldsymbol{Z}_m'\boldsymbol{y} - z_{mi}y_i\right) \\ &= y_i - z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1}\boldsymbol{Z}_m'\boldsymbol{y} + z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1} z_{mi}y_i \\ &\quad -(1 - h_{mi})^{-1} z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1} z_{mi}z_{mi}'\left(\boldsymbol{Z}_m'\boldsymbol{Z}_m\right)^{-1}\boldsymbol{Z}_m'\boldsymbol{y}\end{aligned}$$

$$\begin{aligned}
&+(1-h_{mi})^{-1}z'_{mi}\left(Z'_mZ_m\right)^{-1}z_{mi}z'_{mi}\left(Z'_mZ_m\right)^{-1}z_{mi}y_i \\
&= \hat{e}_{mi}+h_{mi}y_i-(1-h_{mi})^{-1}h_{mi}z'_{mi}\widehat{\beta}_m+(1-h_{mi})^{-1}h^2_{mi}y_i \\
&= \hat{e}_{mi}+(1-h_{mi})^{-1}h_{mi}\hat{e}_{mi} \\
&= (1-h_{mi})^{-1}\hat{e}_{mi},
\end{aligned}$$

where the third equality makes the substitutions $\widehat{\beta}_m = \left(Z'_mZ_m\right)^{-1}Z'_my$ and $h_{mi} = z'_{mi}\left(Z'_mZ_m\right)^{-1}z_{mi}$, with the remainder collecting terms. ■

Define

$$\zeta_m = \sup_{x\in\mathcal{X}}\left(z_m(x)'Q_m^{-1}z_m(x)\right)^{1/2}, \tag{8.19}$$

the largest normalized Euclidean length of the regressor vector. Under Assumption 8.1, if z_{mi} is a power series, then $\zeta_m^2 = O(k_m^2)$ (see Andrews (1991a)), and when z_{mi} is a regression spline, then $\zeta_m^2 = O(k_m)$ (see Newey (1995)). For further discussion see Newey (1997) and Li and Racine (2006).

Without loss of generality, assume $Q_m = I_{K_m}$ throughout this section.

Proof of Theorem 3. Assumptions (A.1), (A.2), (A.7), (A.9), and (A.10) of Hansen and Racine (2012) are satisfied under our Assumptions 8.1–8.3. Thus by their Theorem 2 with $N=1$, CV selection is optimal with respect to the criterion $R_n(m)$, that is,

$$\left|\frac{R_n(\widehat{m})}{\inf\limits_{1\le m\le M_n} R_n(m)}\right| \xrightarrow{p} 1.$$

Furthermore, Theorem 8.1 shows that $IMSE_n^*(m)$ and $IMSE_n(m)$ are asymptotically equivalent. Thus for Theorem 8.3 it is thus sufficient to show that $R_n(m)$ and $IMSE_n^*(m)$ are asymptotically equivalent. To reduce the notation, we will write $I_n(m) = IMSE_n^*(m) = \phi_m^2 + n^{-1}\operatorname{tr}(\Omega_m)$. Thus what we need to show is

$$\sup_{1\le m\le M_n}\left|\frac{R_n(m)-I_n(m)}{I_n(m)}\right| \xrightarrow{p} 0. \tag{8.20}$$

It is helpful to note the following inequalities:

$$n\phi_m^2 \le nI_n(m), \tag{8.21}$$

$$\operatorname{tr}(\Omega_m) \le nI_n(m), \tag{8.22}$$

$$1 \le nI_n(m), \tag{8.23}$$

$$\frac{\zeta_m^2}{n} \le \frac{\zeta_m^2K_m}{n} \le \frac{\zeta_m^2K_m^2}{n} \le \Psi < \infty. \tag{8.24}$$

Equations (8.21) and (8.22) follow from the formula $nI_n(m) = n\phi_m^2 + \mathrm{tr}(\Omega_m)$. Equation (8.23) holds for n sufficiently large since $\xi_n = \inf_m nI_n(m) \to \infty$. The first two inequalities in (8.24) hold since either $K_m \geq 1$ or $\zeta_m^2 = 0$; the third inequality holds for n sufficiently large under Assumption 8.1.5.

Set

$$\widehat{Q}_m = \frac{1}{n}\sum_{i=1}^{n} z_{mi}z_{mi}',$$

$$\widehat{\gamma}_m = \frac{1}{n}\sum_{i=1}^{n} z_{mi}r_{mi},$$

$$\widehat{\Omega}_m = \frac{1}{n}\sum_{i=1}^{n} z_{mi}z_{mi}'\sigma_i^2.$$

As shown in Andrews (1991a) and Hansen and Racine (2012),

$$nR_n(m) = \sum_{i=1}^{n} r_{mi}^2 - n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m + \mathrm{tr}\big(\widehat{Q}_m^{-1}\widehat{\Omega}_m\big).$$

Then

$$\begin{aligned} n(R_n(m) - I_n(m)) = \sum_{i=1}^{n}\big(r_{mi}^2 - \phi_m^2\big) - n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m + \mathrm{tr}\big(\big(\widehat{Q}_m^{-1} - I_{K_m}\big)\Omega_m\big) \\ + \mathrm{tr}\big(\widehat{\Omega}_m - \Omega_m\big) + \mathrm{tr}\big(\big(\widehat{Q}_m^{-1} - I_{K_m}\big)\big(\widehat{\Omega}_m - \Omega_m\big)\big). \end{aligned}$$

and for any $J \geq 2$

$$\begin{aligned} \Big(\mathbb{E}\big|n(R_n(m) - I_n(m))\big|^J\Big)^{1/J} \leq \left(\mathbb{E}\left|\sum_{i=1}^{n}\big(r_{mi}^2 - \phi_m^2\big)\right|^J\right)^{1/J} \quad (8.25) \\ + \Big(\mathbb{E}\big|n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m\big|^J\Big)^{1/J} \quad (8.26) \\ + \Big(\mathbb{E}\big|\mathrm{tr}\big(\big(\widehat{Q}_m^{-1} - I_{K_m}\big)\Omega_m\big)\big|^J\Big)^{1/J} \quad (8.27) \\ + \Big(\mathbb{E}\big|\mathrm{tr}\big(\widehat{\Omega}_m - \Omega_m\big)\big|^J\Big)^{1/J} \quad (8.28) \\ + \Big(\mathbb{E}\big|\mathrm{tr}\big(\big(\widehat{Q}_m^{-1} - I_{K_m}\big)\big(\widehat{\Omega}_m - \Omega_m\big)\big)\big|^J\Big)^{1/J}. \quad (8.29) \end{aligned}$$

We use some bounds developed in Hansen (2013) for the moment matrices that appear on the right-hand side of (8.25)–(8.29). A key bound is the matrix Rosenthal inequality (Theorem 1 of Hansen (2013)), which states that for any $J \geq 2$ there is a constant $A_J < \infty$ such that for any i.i.d. random matrix X_i we have

$$\left(\mathbb{E}\left\|\sum_{i=1}^{n}(X_i-\mathbb{E}X_i)\right\|_2^J\right)^{1/J} \leq \left[A_J\left(\left(n\mathbb{E}\|X_i\|_2^2\right)^{J/2}+\left(n\mathbb{E}\|X_i\|_2^J\right)\right)\right]^{1/J}$$

$$\leq A_J^{1/J}\left(n\mathbb{E}\|X_i\|_2^2\right)^{1/2}+A_J^{1/J}\left(n\mathbb{E}\|X_i\|_2^J\right)^{1/J}. \quad (8.30)$$

where the second inequality is the c_r inequality. Using this bound, Hansen (2013, Lemmas 2 and 3) established that for n sufficiently large we have

$$\mathbb{E}\left\|\widehat{Q}_m^{-1}\right\|^J \leq 2 \quad (8.31)$$

$$\left(\mathbb{E}\left\|\widehat{Q}_m^{-1}-I_{K_m}\right\|^J\right)^{1/J} \leq A_{2J}^{1/J}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2}. \quad (8.32)$$

We use (8.30)–(8.32) to bound the terms (8.25)–(8.29).

We start with (8.25). Define $\overline{r} = \sup_m \sup_{x\in\mathcal{X}}|r_m(x)|$, which is bounded under Assumption 8.1.2. WLOG assume that $\overline{r} \geq 1$. Note that $|r_{mi}| \leq \overline{r}$. Applying (8.30) to (8.25), and then $\mathbb{E}r_{mi}^N \leq \overline{r}^{N-2}\mathbb{E}r_{mi}^2 \leq \overline{r}^{N-2}\phi_m^2$,

$$\left(\mathbb{E}\left|\sum_{i=1}^{n}\left(r_{mi}^2-\phi_m^2\right)\right|^J\right)^{1/J} \leq A_J^{1/J}\left(n\mathbb{E}r_{mi}^4\right)^{1/2}+A_J^{1/J}\left(n\mathbb{E}r_{mi}^{2J}\right)^{1/J}$$

$$\leq A_J^{1/J}\overline{r}\left(n\phi_m^2\right)^{1/2}+A_J^{1/J}\overline{r}^{2-2/J}\left(n\phi_m^2\right)^{1/J}. \quad (8.33)$$

We next take (8.26). Note that (8.19) implies $\|z_{mi}\| \leq \zeta_m$. Then $\mathbb{E}\|z_{mi}r_{mi}\|^2 \leq \overline{r}^2\mathbb{E}\|z_{mi}\|^2 = \overline{r}^2\operatorname{tr}(Q_m) = \overline{r}^2 K_m$ and $\mathbb{E}\|z_{mi}r_{mi}\|^2 \leq \zeta_m^2\phi_m^2$. Together,

$$\mathbb{E}\|z_{mi}r_{mi}\|^2 \leq \overline{r}\left(\frac{\zeta_m^2 K_m}{n}\right)^{1/2}\left(n\phi_m^2\right)^{1/2} \leq \overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2} \quad (8.34)$$

where the second inequality uses (8.24). Similarly,

$$\frac{\mathbb{E}\|z_{mi}r_{mi}\|^{4J}}{n^{2J-1}} \leq \frac{\overline{r}^{4J-2}\zeta_m^{4J}\mathbb{E}r_{mi}^2}{n^{2J-1}} = \overline{r}^{4J-2}\left(\frac{\zeta_m^2}{n}\right)^{2J} n\phi_m^2 \leq \overline{r}^{4J-2}\Psi^{2J}n\phi_m^2. \quad (8.35)$$

Applying (8.30) to (8.26), and then (8.34) and (8.35), we find

$$\left(\mathbb{E}\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{4J}\right)^{1/2J} \leq A_{4J}^{1/2J}\mathbb{E}\|z_{mi}r_{mi}\|^2+A_{4J}^{1/2J}\left(\frac{\mathbb{E}\|z_{mi}r_{mi}\|^{4J}}{n^{2J-1}}\right)^{1/2J}$$

$$\leq A_{4J}^{1/2J}\overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2}+A_{4J}^{1/2J}\overline{r}^{2-1/J}\Psi\left(n\phi_m^2\right)^{1/2J}. \quad (8.36)$$

Using the trace and Cauchy–Schwarz inequalities, (8.31), and (8.36), we obtain

$$\begin{aligned}\left(\mathbb{E}\left|n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m\right|^J\right)^{1/J} &\le \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1}\right\|^J\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{2J}\right)\right)^{1/J}\\ &\le \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1}\right\|^{2J}\right)\mathbb{E}\left(\left\|n^{1/2}\widehat{\gamma}_m\right\|_2^{4J}\right)\right)^{1/2J}\\ &\le \left(2A_{4J}\right)^{1/2J}\overline{r}\Psi^{1/2}\left(n\phi_m^2\right)^{1/2}\\ &\quad+\left(2A_{4J}\right)^{1/2J}\overline{r}^{2-1/J}\Psi\left(n\phi_m^2\right)^{1/2J}. \qquad (8.37)\end{aligned}$$

Now we take (8.27). Using the trace inequality, (8.32), we obtain $\mathrm{tr}(\Omega_m) = \mathbb{E}\left|z_{mi}'z_{mi}\sigma_i^2\right| \le \overline{\sigma}^2\mathbb{E}\left|z_{mi}'z_{mi}\right| = \overline{\sigma}^2 K_m$; and using (8.24), we get

$$\begin{aligned}\left(\mathbb{E}\left|\mathrm{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J} &\le \left(\mathbb{E}\left\|\widehat{Q}_m^{-1}-I_{K_m}\right\|^J\right)^{1/J}\mathrm{tr}(\Omega_m)\\ &\le A_J^{1/J}\left(\frac{\zeta_m^2K_m}{n}\right)^{1/2}\overline{\sigma}K_m^{1/2}\,\mathrm{tr}(\Omega_m)^{1/2}\\ &\le \overline{\sigma}A_J^{1/J}\;\Psi^{1/2}\,\mathrm{tr}(\Omega_m)^{1/2}. \qquad (8.38)\end{aligned}$$

Next, take (8.28). Applying (8.30) to (8.28) and using $\left|z_{mi}'z_{mi}\sigma_i^2\right| \le \zeta_m^2\overline{\sigma}^2$ and (8.24), we obtain

$$\begin{aligned}&\left(\mathbb{E}\left|\mathrm{tr}\left(\widehat{\Omega}_m-\Omega_m\right)\right|^J\right)^{1/J}\\ &\le A_J^{1/J}\left(\frac{\mathbb{E}\left|z_{mi}'z_{mi}\sigma_i^2\right|^2}{n}\right)^{1/2}+A_J^{1/J}\left(\frac{\mathbb{E}\left|z_{mi}'z_{mi}\sigma_i^2\right|^J}{n^{J-1}}\right)^{1/J}\\ &\le \overline{\sigma}A_J^{1/J}\left(\frac{\zeta_m^2}{n}\right)^{1/2}\mathrm{tr}(\Omega_m)^{1/2}+\overline{\sigma}^{2(1-1/J)}A_J^{1/J}\left(\frac{\zeta_m^2}{n}\right)^{1-1/J}\mathrm{tr}(\Omega_m)^{1/J} \qquad (8.39)\\ &\le \overline{\sigma}A_J^{1/J}\Psi^{1/2}\,\mathrm{tr}(\Omega_m)^{1/2}+\overline{\sigma}^{2(1-1/J)}A_J^{1/J}\Psi^{1-1/J}\,\mathrm{tr}(\Omega_m)^{1/J}. \qquad (8.40)\end{aligned}$$

Finally, take (8.29). Using the trace inequality, Cauchy–Schwarz, (8.32), and (8.39), we get

$$\begin{aligned}&\left(\mathbb{E}\left|\mathrm{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\left(\widehat{\Omega}_m-\Omega_m\right)\right)\right|^J\right)^{1/J}\\ &\le \left(\mathbb{E}\left(\left\|\widehat{Q}_m^{-1}-I_{K_m}\right\|^J\left\|\widehat{\Omega}_m-\Omega_m\right\|^J\right)\right)^{1/J}K_m\\ &\le \left(\mathbb{E}\left\|\widehat{Q}_m^{-1}-I_{K_m}\right\|^{2J}\right)^{1/2J}\left(\mathbb{E}\left\|\widehat{\Omega}_m-\Omega_m\right\|^{2J}\right)^{1/2J}K_m\\ &\le A_{4J}^{1/2J}\left(\frac{\zeta_m^2K_m}{n}\right)^{1/2}\left(\overline{\sigma}A_{2J}^{1/2J}\left(\frac{\zeta_m^2}{n}\right)^{1/2}\mathrm{tr}(\Omega_m)^{1/2}\right.\end{aligned}$$

$$+\overline{\sigma}^{2(1-1/2J)} A_{2J}^{1/2J} \left(\frac{\zeta_m^2}{n}\right)^{1-1/2J} \operatorname{tr}(\Omega_m)^{1/2J} \Bigg) K_m$$

$$\leq \overline{\sigma} A_{4J}^{1/2J} A_J^{1/J} \Psi \operatorname{tr}(\Omega_m)^{1/2} + \overline{\sigma}^{2(1-1/2J)} A_{4J}^{1/2J} A_{2J}^{1/2J} \Psi^{3/2-1/2J} \operatorname{tr}(\Omega_m)^{1/2J}. \tag{8.41}$$

Combining (8.33) and (8.37) and then applying (8.21) and (8.23), we find that

$$\left(\mathbb{E}\left|\sum_{i=1}^{n}\left(r_{mi}^2 - \phi_m^2\right)\right|^J\right)^{1/J} + \left(\mathbb{E}\left|n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m\right|^J\right)^{1/J}$$

$$\leq C_1\left(n\phi_m^2\right)^{1/2} C_2\left(n\phi_m^2\right)^{1/J} + C_3\left(n\phi_m^2\right)^{1/2J} \tag{8.42}$$

$$\leq C_1(nI_n(m))^{1/2} + C_2(nI_n(m))^{1/J} + C_3(nI_n(m))^{1/2J}$$

$$\leq (C_1 + C_2 + C_3)(nI_n(m))^{1/2}, \tag{8.43}$$

where $C_1 = A_J^{1/J}\overline{r} + \left(2A_{4J}\right)^{1/2J}\overline{r}\Psi^{1/2}$, $C_2 = A_J^{1/J}\overline{r}^{2-2/J}$, and $C_3 = \left(2A_{4J}\right)^{1/2J}\overline{r}^{2-1/J}\Psi$.

Similarly, combining (8.38), (8.40), and (8.41) and then applying (8.22) and (8.23), we obtain

$$\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J} + \left(\mathbb{E}\left|\operatorname{tr}\left(\widehat{\Omega}_m - \Omega_m\right)\right|^J\right)^{1/J}$$

$$+\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1} - I_{K_m}\right)\left(\widehat{\Omega}_m - \Omega_m\right)\right)\right|^J\right)^{1/J}$$

$$\leq C_4 \operatorname{tr}(\Omega_m)^{1/2} C_5 \operatorname{tr}(\Omega_m)^{1/J} + C_6 \operatorname{tr}(\Omega_m)^{1/2J} \tag{8.44}$$

$$\leq C_4(nI_n(m))^{1/2} + C_5(nI_n(m))^{1/J} + C_6(nI_n(m))^{1/2J}$$

$$\leq (C_4 + C_5 + C_6)(nI_n(m))^{1/2}. \tag{8.45}$$

where $C_4 = \overline{\sigma} A_J^{1/J}\left(2\Psi^{1/2} + A_{4J}^{1/2J}\Psi\right)$, $C_5 = \overline{\sigma}^{2(1-1/J)} A_J^{1/J}\Psi^{1-1/J}$, and $C_6 = \overline{\sigma}^{2(1-1/2J)} A_{4J}^{1/2J} A_{2J}^{1/2J}\Psi^{3/2-1/2J}$.

Setting $J = 4$, (8.25)–(8.29), (8.43), and (8.45) imply that

$$\left(\mathbb{E}|n(R_n(m) - I_n(m))|^4\right)^{1/4} \leq C(nI_n(m))^{1/2}. \tag{8.46}$$

where $C = C_1 + C_2 + C_3 + C_4 + C_5 + C_6$.

Applying Boole's inequality, Markov's inequality, and (8.46), we obtain

$$\mathbb{P}\left(\sup_{1\leq m\leq M_n}\left|\frac{R_n(m) - I_n(m)}{I_n(m)}\right| > \eta\right) = \mathbb{P}\left(\bigcup_{m=1}^{M_n}\left\{\left|\frac{R_n(m) - I_n(m)}{I_n(m)}\right| > \eta\right\}\right)$$

$$\leq \sum_{m=1}^{M_n}\mathbb{P}\left(\left\{\left|\frac{n(R_n(m) - I_n(m))}{nI_n(m)}\right| > \eta\right\}\right)$$

$$\begin{aligned}
&\leq \eta^{-4} \sum_{m=1}^{M_n} \frac{\mathbb{E}|n(R_n(m) - I_n(m))|^4}{(nI_n(m))^4} \\
&\leq C^4 \eta^{-4} \sum_{m=1}^{M_n} \frac{1}{(nI_n(m))^2}.
\end{aligned}$$

Recall the definitions of $\overline{q}_n$ and ξ_n. Pick a sequence $m_n \to \infty$ such that $m_n \xi_n^{-2} \to 0$ yet $\overline{q}_n^2 = o(m_n)$, which is possible since $\xi_n \to \infty$ and $\overline{q}_n^2 = o(\xi_n^2)$ under Assumption 8.3.2. Then since $nI_n(m) \geq \xi_n$ and $nI_n(m) \geq \operatorname{tr}(\Omega_m) \geq \underline{\sigma}^2 K_m \geq \underline{\sigma}^2 m/\overline{q}_n$, the sum on the right-hand side is bounded by

$$m_n \xi_n^{-2} + \sum_{m=m_n+1}^{\infty} \frac{\overline{q}_n^2}{\underline{\sigma}^4 m^2} \leq m_n \xi_n^{-2} + \frac{\overline{q}_n^2}{\underline{\sigma}^4 m_n} \longrightarrow 0$$

as $n \to \infty$. This establishes (8.20) as desired, completing the proof. ■

Proof of Theorem 8.5. As in the proof of Theorem 8.2, it is sufficient to show that

$$\sup_{w \in \mathcal{H}_n} \left| \frac{R_n(w) - I_n(w)}{I_n(w)} \right| \xrightarrow{p} 0. \tag{8.47}$$

where we have written $I_n(w) = IMSE_n^*(w)$. WLOG assumes $Q_m = I_{K_m}$. For w_m^* and w_m^{**} defined in (8.13) and (8.14), observe that $\sum_{m=1}^{M_n} w_m^* = \sum_{m=1}^{M_n} w_m^{**} = 1$. Since w_m^* are non-negative and sum to one, they define a probability distribution. Thus by Liapunov's inequality, for any $s \geq 1$ and any constants $a_m \geq 0$ we get

$$\sum_{m=1}^{M_n} w_m^* a_m^{1/s} \leq \left(\sum_{m=1}^{M_n} w_m^* a_m \right)^{1/s} \tag{8.48}$$

and similarly

$$\sum_{m=1}^{M_n} w_m^{**} a_m^{1/s} \leq \left(\sum_{m=1}^{M_n} w_m^{**} a_m \right)^{1/s}. \tag{8.49}$$

As shown in Andrews (1991a) and Hansen and Racine (2012), we have

$$nR_n(w) = \sum_{m=1}^{M_n} w_m^* n\phi_m^2 - \sum_{m=1}^{M_n} w_m^* n\widehat{\gamma}_m' \widehat{Q}_m^{-1} \widehat{\gamma}_m + \sum_{m=1}^{M_n} w_m^{**} \operatorname{tr}\big(\widehat{Q}_m^{-1} \widehat{\Omega}_m\big).$$

Then applying Minkowski's inequality, (8.42), (8.44), and then (8.48) and (8.49), we obtain

$$
\begin{aligned}
&\left(\mathbb{E}|n(R_n(m)-I_n(m))|^J\right)^{1/J}\\
&\le \sum_{m=1}^{M_n} w_m^*\left[\left(\mathbb{E}\left|\sum_{i=1}^{n}\left(r_{mi}^2-\phi_m^2\right)\right|^J\right)^{1/J}+\left(\mathbb{E}\left|n\widehat{\gamma}_m'\widehat{Q}_m^{-1}\widehat{\gamma}_m\right|^J\right)^{1/J}\right]\\
&+\sum_{m=1}^{M_n} w_m^{**}\left[\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\Omega_m\right)\right|^J\right)^{1/J}+\left(\mathbb{E}\left|\operatorname{tr}\left(\widehat{\Omega}_m-\Omega_m\right)\right|^J\right)^{1/J}\right.\\
&\left.+\left(\mathbb{E}\left|\operatorname{tr}\left(\left(\widehat{Q}_m^{-1}-I_{K_m}\right)\left(\widehat{\Omega}_m-\Omega_m\right)\right)\right|^J\right)^{1/J}\right]\\
&\le C_1\sum_{m=1}^{M_n} w_m^*\left(n\phi_m^2\right)^{1/2}+C_2\sum_{m=1}^{M_n} w_m^*\left(n\phi_m^2\right)^{1/J}+C_3\sum_{m=1}^{M_n} w_m^*\left(n\phi_m^2\right)^{1/2J}\\
&+C_4\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)^{1/2}+C_5\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)^{1/J}+C_6\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)^{1/2J}\\
&\le C_1\left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2\right)^{1/2}+C_2\left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2\right)^{1/J}+C_3\left(\sum_{m=1}^{M_n} w_m^* n\phi_m^2\right)^{1/2J}\\
&+C_4\left(\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)\right)^{1/2}+C_5\left(\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)\right)^{1/J}+C_6\left(\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m)\right)^{1/2J}\\
&\le C_1(nI_n(w))^{1/2}+C_2(nI_n(w))^{1/J}+C_3(nI_n(w))^{1/2J}\\
&+C_4(nI_n(w))^{1/2}+C_5(nI_n(w))^{1/J}+C_6(nI_n(w))^{1/2J}\\
&\le C(nI_n(m))^{1/2}
\end{aligned}
$$

where the final two inequalities use

$$
\begin{aligned}
\sum_{m=1}^{M_n} w_m^* n\phi_m^2 &\le nI_n(w),\\
\sum_{m=1}^{M_n} w_m^{**}\operatorname{tr}(\Omega_m) &\le nI_n(w),\\
1 &\le nI_n(w),
\end{aligned}
$$

where the first two follow from the formula (8.12) for $nI_n(w)$, and the third holds for n sufficiently large since $\inf_w nI_n(w)\to\infty$.

Setting $J = 2(N+1)$, we have shown that

$$\mathbb{E}|n(R_n(w) - I_n(w))|^{2(N+1)} \leq C^{1+N}(nI_n(w))^{N+1}.$$

Then

$$\begin{aligned}
\mathbb{P}\left(\sup_{w\in\mathcal{H}_n}\left|\frac{R_n(w)-I_n(w)}{I_n(w)}\right| > \eta\right) &= \mathbb{P}\left(\bigcup_{w\in\mathcal{H}_n}\left\{\left|\frac{R_n(w)-I_n(w)}{I_n(w)}\right| > \eta\right\}\right)\\
&\leq \sum_{w\in\mathcal{H}_n}\mathbb{P}\left(\left\{\left|\frac{n(R_n(w)-I_n(w))}{nI_n(w)}\right| > \eta\right\}\right)\\
&\leq \eta^{-2(N+1)}\sum_{w\in\mathcal{H}_n}\frac{\mathbb{E}|n(R_n(w)-I_n(w))|^{2(N+1)}}{(nI_n(w))^{2(N+1)}}\\
&\leq C^{1+N}\eta^{-2(N+1)}\sum_{w\in\mathcal{H}_n}\frac{1}{(nI_n(w))^{N+1}}.
\end{aligned}$$

As shown in Hansen and Racine (2012, Eqs. (23), (25), and (28)), the right-hand side is $o(1)$. ■

By Markov's inequality, we have established (8.47), as desired.

Notes

[†] Research supported by the National Science Foundation. I thank Chu-An Liu for excellent research assistance, Guido Kuersteiner and Ryo Okui for pointing out an error in Hansen and Racine (2012), and the referee and editor for helpful comments and suggestions.

References

Akaike, H. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." In *Second International Symposium on Information Theory*, eds. B. Petroc and F. Csake, Akademiai Kiado, Budapest, Hungary.

Allen, David M. 1974. "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction." *Technometrics*, **16**, pp. 125–127.

Andrews, Donald W. K. 1991a. "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Models." *Econometrica,* **59**, pp. 307–345.

Andrews, Donald W. K. 1991b. "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors." *Journal of Econometrics*, **47**, pp. 359–377.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-nonparametric Models." Chapter 76 In *Handbook of Econometrics*, Vol. 6B, eds. James J. Heckman and Edward E. Leamer. Amsterdam: North-Holland.

Chui, Charles K. 1992. *An Introduction to Wavelets.* New York: Academic Press.

Craven P., and Grace Wahba. 1979. "Smoothing Noisy Data with Spline Functions." *Numerische Mathematik*, **31**, pp. 377–403.

de Boor, Carl. 2001. *A Practical Guide to Splines.* Berlin: Springer.

Fan, Jianging, and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association*, **96**, pp. 1348–1360.

Grenander, U. 1981. *Abstract Inference.* New York: Wiley.

Hansen, Bruce E. 2007 "Least Squares Model Averaging." *Econometrica*, **75**, pp. 1175–1189.

Hansen, Bruce E. 2013. "The Integrated Squared Error of Series Regression and a Rosenthal Hilbert-Space Inequality." Working paper, University of Wisconsin.

Hansen, Bruce E., and Jeffrey S. Racine. 2012. "Jackknife Model Averaging." *Journal of Econometrics*, **167**, pp. 38–46.

Hurvich, Clifford M., and Chih-Ling Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika*, **76**, pp. 297–307.

Li, Ker-Chau. 1987. "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set." *Annals of Statistics*, **15**, pp. 958–975.

Li, Qi, and Jeffrey S. Racine. 2006. *Nonparametric Econometrics: Theory and Practice.* Princeton, NJ: Princeton University Press.

Mallows, C. L. 1973. "Some Comments on C_p." *Technometrics*, **15**, pp. 661–675.

Newey, Whitney K. 1995. "Convergence Rates for Series Estimators." In *Statistical Methods of Economics and Quantitative Economics: Essays in Honor of C.R. Rao*, eds. G. S. Maddalla, P. C. B. Phillips, T. N. Srinavasan. Blackwell, Cambridge, pp. 254–275.

Newey, Whitney K. 1997. "Convergence Rates and Asymptotic Normality for Series Estimators." *Journal of Econometrics*, **79**, pp. 147–168.

Schwarz, G. 1978. "Estimating the Dimension of a Model." *Annals of Statistics*, **6**, pp. 461–464.

Shao, Jun. 1997. "An Asymptotic Theory for Linear Model Selection." *Statistica Sinica*, **7**, pp. 221–264.

Sherman, Jack, and Winifred J. Morrison. 1950. "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix." *Annals of Mathematical Statistics*, **21**, pp. 124–127.

Stone, M. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions" (with Discussion). *Journal of the Royal Statistical Society, Series B*, **36**, pp. 111–147.

Tibshirani, R. J. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society*, Series B, **58**, pp. 267–288.

Wahba, Grace, and S. Wold. 1975. "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation." *Communications in Statistics*, **4**, pp. 1–17.

Zou, Hui. 2006. "The Adaptive LASSO and Its Oracle Properties." *Journal of the American Statistical Association*, **101**, pp. 1418–1429.

CHAPTER 9

VARIABLE SELECTION IN NONPARAMETRIC AND SEMIPARAMETRIC REGRESSION MODELS

LIANGJUN SU AND YONGHUI ZHANG

9.1. Introduction

Over the last 15 years or so, high-dimensional models have become increasingly popular and gained considerable attention in diverse fields of scientific research. Examples in economics include wage regression with more than 100 regressors (e.g., Belloni and Chernozhukov, 2011b), portfolio allocation among hundreds or thousands of stocks (e.g., Jagannathan and Ma, 2003; Fan, Zhang, and Yu, 2011), VAR models with hundreds of data series (e.g., Bernanke, Boivin, and Eliasz, 2005), and large-dimensional panel data models of home price (e.g., Fan, Lv, and Qi, 2011), among others. A common feature of high-dimensional models is that the number of regressors is very large, which may grow as the sample size increases. This poses a series of challenges for statistical modeling and inference. Penalized least squares or likelihood has become a standard unified framework for variable selection and feature extraction in such scenarios. For a comprehensive overview of high-dimensional modeling, see Fan and Li (2006) and Fan and Lv (2010).

In high-dimensional modeling, one of the most important problems is the choice of an optimal model from a set of candidate models. In many cases, this reduces to the choice of which subset of variables should be included into the model. Subset selection has attracted a lot of interest, leading to a variety of procedures. The majority of these procedures do variable selection by minimizing a penalized objective function with the following form:

$$\text{Loss function} + \text{Penalization.}$$

The most popular choices of loss functions are least squares, negative log-likelihood, and their variants (e.g., profiled least squares or profiled negative log-likelihood). There are many choices of penalization. The traditional subset selection criterion such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) uses the l_0-norm for the parameters entering the model so that the penalization term is proportional to the number of nonzero parameters. The bridge estimator (see, e.g., Frank and Friedman, 1993; Fu, 1998; Knight and Fu, 2000) uses the l_q-norm ($q > 0$). It boils down the commonly used ridge estimator (Horel and Kennard, 1970) when $q = 2$ and the Lasso estimator (Tibshirani, 1996) when $q = 1$. When $0 < q \leq 1$, some components of the estimator can be exactly zero with some correctly chosen tuning parameters. Thus, the bridge estimator with $0 < q \leq 1$ provides a way to combine variable selection and parameter estimation simultaneously. Among the class of bridge estimators, Lasso becomes most popular due to its computational and theoretical advantages compared with other bridge estimators and traditional variable selection methods. Allowing an adaptive amount of shrinkage for each regression coefficient results in an estimator called the adaptive Lasso, which was first proposed by Zou (2006) and can be as efficient as the oracle one in the sense that the method works asymptotically equivalent to the case as if the correct model were exactly known. Other variants of Lasso include the group Lasso, adaptive group Lasso, graphic Lasso, elastic net, and so on. Of course, the penalization term can take other forms; examples include the SCAD penalty of Fan and Li (2001) and the MC penalty of Zhang (2010).

Given the huge literature on variable selection that has developed over the last 15 years, it is impossible to review all of the works. Fan and Lv (2010) offer a selective overview of variable selection in high-dimensional feature space. By contrast, in this chapter we focus on variable selection in semiparametric and nonparametric regression models with fixed or large dimensions because semiparametric and nonparametric regressions have gained considerable importance over the last three decades due to their flexibility in modeling and robustness to model misspecification. In particular, we consider variable selection in the following models:

- Additive models
- Partially linear models
- Functional/varying coefficient models
- Single index models
- General nonparametric regression models
- Semiparametric/nonparametric quantile regression models

The first four areas are limited to semiparametric and nonparametric regression models that impose certain structure to alleviate the "curse of dimensionality" in the nonparametric literature. The fifth part focuses on variable or component selection in general nonparametric models. In the last part we review variable selection in semiparametric and nonparametric quantile regression models. Below we first briefly introduce variable selection via Lasso or SCAD type of penalties in general parametric regression models and then review its development in the above fields in turn. In the last section we highlight some issues that require further study.

9.2. Variable Selection via Lasso or SCAD Type of Penalties in Parametric Models

In this section we introduce the background of variable selection via Lasso or SCAD type of penalties.

9.2.1. The Lasso Estimator

The Lasso (*least absolute shrinkage and selection operator*) proposed by Tibshirani (1996) is a popular model building technique that can simultaneously produce accurate and parsimonious models. For $i = 1,\ldots,n$, let Y_i denote a response variable and let $X_i = \left(X_{i1},\ldots,X_{ip}\right)'$ denote a $p \times 1$ vector of covariates/predictors. For simplicity, one could assume that (X_i, Y_i) are independent and identically distributed (i.i.d.), or assume that $\{X_i, Y_i\}_{i=1}^n$ are standardized so that $\overline{Y} = n^{-1}\sum_{i=1}^n Y_i = 0$, $n^{-1}\sum_{i=1}^n X_{ij} = 0$, and $n^{-1}\sum_i X_{ij}^2 = 1$ for $j = 1,\ldots,p$. But these are not necessary. The Lasso estimates of the slope coefficients in a linear regression model solve the l_1-penalized least regression problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\left|\beta_j\right| \le s, \tag{9.1}$$

or, equivalently,

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|, \tag{9.2}$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)'$, and s and λ are tuning parameters. The restricted optimization in (9.1) suggests that the Lasso uses a constraint in the form of l_1-norm: $\sum_{j=1}^p\left|\beta_j\right| \le s$. It is similar to the ridge regression with the constraint of l_2-norm: $\sum_{j=1}^p \beta_j^2 \le s$. By using the l_1-penalty, the Lasso achieves variable selection and shrinkage simultaneously. However, ridge regression only does shrinkage. More generally, a *penalized least squares* (PLS) can have a generic l_q-penalty of the form $(\sum_{j=1}^p\left|\beta_j\right|^q)^{1/q}$, $0 \le q \le 2$. When $q \le 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients. In particular, when $q = 0$, the l_0-penalty term becomes $\sum_{j=1}^p \mathbf{1}\left(\beta_j \neq 0\right)$ with $\mathbf{1}(\cdot)$ being the usual indicator function, which counts the number of nonzero elements in the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. The Lasso uses the smallest value of q, that is, $q = 1$, which leads to a convex problem.

The tuning parameter λ is the shrinkage parameter that controls the amount of regularization. If $\lambda = 0$, there is no penalty put on the coefficients and hence we obtain the ordinary least squares solution; if $\lambda \to \infty$, the penalty is infinitely large and thus forces all of the coefficients to be zero. These are necessary but insufficient for the Lasso to produce sparse solutions. Large enough λ will set some coefficients exactly equal to zero and is thus able to perform variable selection. In contrast, a ridge penalty never sets coefficients exactly equal to 0.

Efron, Hastie, Johnstone, and Tibshirani (2004) propose the *least angel regression selection* (LARS) and show that the entire solution path of the Lasso can be computed by the LARS algorithm. Their procedure includes two steps. First, a solution path that is indexed by a tuning parameter is constructed. Then the final model is chosen on the solution path by cross-validation or using some criterion such as C_p. The solution path is piecewise linear and can be computed very efficiently. These nice properties make the Lasso very popular in variable selection.

9.2.2. Some Generalizations and Variants of the Lasso

In this subsection, we review some variants of the Lasso: Bridge, the adaptive Lasso, the group Lasso, and the elastic-net. For other work generalizing the Lasso, we give a partial list for reference.

Bridge. Knight and Fu (2000) study the asymptotics for the Bridge estimator $\hat{\boldsymbol{\beta}}_{\text{Bridge}}$ which is obtained via the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|^{\gamma}, \tag{9.3}$$

where λ_n is a tuning parameter and $\gamma > 0$. For $\gamma \leq 1$, the Bridge estimator has the potentially attractive feature of being exactly 0 if λ_n is sufficiently large, thus combining parametric estimation and variable selection in a single step.

Adaptive Lasso. Fan and Li (2001) and Fan and Peng (2004) argue that a good variable selection procedure should have the following oracle properties: (i) *Selection consistency.* This can identify the right subset models in the sense that it selects the correct subset of predictors with probability tending to one. (ii) *Asymptotic optimality.* This achieves the optimal asymptotic distribution as the oracle estimator in the sense that it estimates the nonzero parameters as efficiently as would be possible if we knew which variables were uninformative ahead of time. It has been shown that the Lasso of Tibshirani (1996) lacks such oracle properties whereas the Bridge estimator with $0 < \gamma < 1$

can possess them with a well-chosen tuning parameter. Fan and Li (2001) point out that asymptotically the Lasso has non-ignorable bias for estimating the nonzero coefficients. For more discussions on the consistency of Lasso, see Zhao and Yu (2006). Zou (2006) first shows that the Lasso could be inconsistent for model selection unless the predictor matrix satisfies a rather strong condition, and then propose a new version of the Lasso, called the *adaptive Lasso.* Adaptive Lasso assigns different weights to penalize different coefficients in the l_1-penalty. That is, the adaptive Lasso estimate $\hat{\boldsymbol{\beta}}_{\text{ALasso}}$ of $\boldsymbol{\beta}$ solves the following minimization problem:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{9.4}$$

where the weights $\hat{w}_j$'s are data-dependent and typically chosen as $\hat{w}_j = |\hat{\beta}_j|^{-\gamma}$ for some $\gamma > 0$, and $\hat{\beta}_j$ is a preliminary consistent estimator of β_j that typically has $\sqrt{n}$-rate of convergence. Intuitively, in adaptive Lasso the zero parameter is penalized more severely than a nonzero parameter as the weight of the zero parameter goes to infinity, while that of a nonzero parameter goes to a positive constant. Zou shows that the adaptive Lasso enjoys the oracle properties so that it performs as well as if the underlying true model were given in advance. Similar to the Lasso, the adaptive Lasso is also near-minimax optimal in the sense of Donoho and Johnstone (1994).

Group Lasso. Observing that the Lasso is designed for selecting individual regressor, Yuan and Lin (2006) consider extensions of the Lasso and the LARS to the case with "grouped variables." The group Lasso estimate is defined as the solution to

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \left\| \mathrm{Y} - \sum_{j=1}^{J} \mathrm{X}_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n \sum_{j=1}^{J} \left\| \boldsymbol{\beta}_j \right\|_{K_j}, \tag{9.5}$$

where $\mathrm{Y} = (Y_1, \ldots, Y_n)'$, X_j is an $n \times p_j$ regressor matrix, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_J)'$, $\boldsymbol{\beta}_j$ is a $p_j \times 1$ vector, K_j is a $p_j \times p_j$ positive definite matrix, $\|\cdot\|_2$ is a Euclidean norm, $\left\|\boldsymbol{\beta}_j\right\|_{K_j} = (\boldsymbol{\beta}'_j K_j \boldsymbol{\beta}_j)^{1/2}$, and $\lambda \geq 0$ is a tuning parameter. Two typical choices of K_j are I_{p_j} and $p_j I_{p_j}$, where I_{p_j} is a $p_j \times p_j$ identity matrix and the coefficient p_j in the second choice is used to adjust for the group size. Obviously, the penalty function in the group Lasso is intermediate between the l_1-penalty that is used in the Lasso and the l_2-penalty that is used in ridge regression. It can be viewed as an l_1-penalty used for coefficients from different groups and an l_2-penalty used for coefficients in the same group. Yuan and Lin propose a group version of the LARS algorithm to solve the minimization problem. See Huang, Breheny, and Ma (2012) for an overview on group selection in high-dimensional models.

Elastic-net. As shown in Zou and Hastie (2005), the Lasso solution paths are unstable when the predictors are highly correlated. Zou and Hastie (2005) propose the elastic-net as an improved version of the Lasso for analyzing high-dimensional data. The elastic-net estimator is defined as follows:

$$\hat{\boldsymbol{\beta}}_{\mathrm{Enet}} = \left(1+\frac{\lambda_2}{n}\right)\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\|\mathrm{Y}-X\boldsymbol{\beta}\|_2^2+\lambda_2\|\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1\right\}, \tag{9.6}$$

where $\mathrm{Y}=(Y_1,\ldots,Y_n)'$, X is an $n\times p$ regressor matrix, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)'$, $\|\boldsymbol{\beta}\|_q=\{\sum_{j=1}^p|\beta_j|^q\}^{1/q}$, and λ_1 and λ_2 are tuning parameters. When the predictors are standardized, $1+\lambda_2/n$ should be replaced by $1+\lambda_2$. The l_1-part of the elastic-net performs automatic variable selection, while the l_2-part stabilizes the solution paths to improve the prediction. Donoho, Johnstone, Kerkyacharian, and Picard (1995) show that in the case of orthogonal design the elastic-net automatically reduces to the Lasso. Zou and Hastie also propose the adaptive elastic-net estimates:

$$\hat{\boldsymbol{\beta}}_{\mathrm{AdaEnet}} = \left(1+\frac{\lambda_2}{n}\right)\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\|\mathrm{Y}-\mathrm{X}\boldsymbol{\beta}\|_2^2+\lambda_2\|\boldsymbol{\beta}\|_2^2+\lambda_1\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right\}, \tag{9.7}$$

where $\hat{w}_j=|\hat{\beta}_{\mathrm{Enet},j}|^{-\gamma}$ and $\hat{\beta}_{\mathrm{Enet},j}$ denotes the jth element of $\hat{\boldsymbol{\beta}}_{\mathrm{Enet}}$ for $j=1,\ldots,p$. Under some weak regularity conditions, they establish the oracle property of the adaptive elastic-net.

There has been a large amount of work in recent years, applying and generalizing the Lasso and l_1-like penalties to a variety of problems. This includes the adaptive group Lasso (e.g., Wang and Leng, 2008; Wei and Huang, 2010), fused Lasso (e.g., Tibshirani et al., 2005; Rinaldao, 2009), the graphical Lasso (e.g., Yuan and Lin, 2007; Friedman et al., 2008), the Dantzig selector (e.g., Candès and Tao, 2007), and near isotonic regularization (e.g., Tibshirani et al., 2010), among others. See Table 1 in Tibshirani (2011) for a partial list of generalizations of the Lasso.

9.2.3. Other Penalty Functions

Many non-Lasso-type penalization approaches have also been proposed, including the SCAD and MC penalties. In the linear regression framework, the estimates are given by

$$\hat{\boldsymbol{\beta}}_n(\lambda_n)=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\|\mathrm{Y}-\mathrm{X}\boldsymbol{\beta}\|^2+n\sum_{j=1}^{p}p_{\lambda_n}\left(|\beta_j|\right),$$

where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a penalty parameter. Different penalty functions yield different variable selection procedures, which have different asymptomatic properties. Note that the Bridge penalty function takes the form $p_\lambda(v)=\lambda|v|^\gamma$, where $\gamma>0$ is a constant. The ordinary Lasso penalty function corresponds to $p_\lambda(v)=\lambda|v|$.

SCAD. The SCAD (*smoothly clipped absolute deviation*) penalty function proposed by Fan and Li (2001) takes the form

$$p_\lambda(v) = \begin{cases} \lambda v & \text{if } 0 \le v \le \lambda, \\ -\frac{(v^2 - 2a\lambda v + \lambda^2)}{2(a-1)} & \text{if } \lambda < v < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } a\lambda \le v, \end{cases}$$

and its derivative satisfies

$$p'_\lambda(v) = \lambda \left[1(v \le \lambda) + \frac{(a\lambda - v)_+}{(a-1)\lambda} 1(v > \lambda) \right],$$

where $(b)_+ = \max(b,0)$ and $a > 2$ is a constant ($a = 3.7$ is recommended and used frequently). Fan and Li (2001) and Fan and Peng (2004) investigate the properties of penalized least squares and likelihood estimator with the SCAD penalty. In particular, they show that the SCAD penalty can yield estimators with the oracle properties. Hunter and Li (2004) suggest using MM algorithms to improve the performance of SCAD-penalized estimators.

MC. The MC (*minimax concave*) penalty function proposed by Zhang (2010) is given by

$$p_\lambda(v) = \lambda \int_0^v \left(1 - \frac{x}{\gamma\lambda}\right)_+ dx$$

where $\gamma > 0$ is a tuning parameter. Zhang proposes and studies the MC+ methodology that has two components: a MC penalty and a penalized linear unbiased selection (PLUS) algorithm. It provides a fast algorithm for nearly unbiased concave penalized selection in the linear regression model and achieves selection consistency and minimax convergence rates.

For other penalization methods, see Fan, Huang, and Peng (2005) and Zou and Zhang (2009).

9.3. Variable Selection in Additive Models

In this section, we consider the problem of variable selection in the following nonparametric additive model

$$Y_i = \mu + \sum_{j=1}^{p} f_j\left(X_{ij}\right) + \varepsilon_i, \tag{9.8}$$

where Y_i is a response variable and $X_i = \left(X_{i1}, \ldots, X_{ip}\right)'$ is a $p \times 1$ vector of covariates, μ is the intercept term, the f_j's are unknown smooth functions with zero means, and

ε_i is the unobserved random error term with mean zero and finite variance σ^2. It is typically assumed that $(Y_i, X_i), i = 1, \ldots, n$, are i.i.d. and some additive components $f_j(\cdot)$ are zero. The main problem is to distinguish the nonzero components from the zero components and estimate the nonzero components consistently. The number of additive components p can be either fixed or divergent as the sample size n increases. In the latter case, we frequently write $p = p_n$. In some scenarios, p_n is allowed to be much larger than n.

Many penalized methods have been proposed to select the significant nonzero components for model (9.8). Huang, Horowitz, and Wei (2010) apply the *adaptive group Lasso* to select nonzero component after using the group Lasso to obtain an initial estimator and to achieve an initial reduction of the dimension. They assume that the number of nonzero f_j's is fixed and give conditions under which the group Lasso selects a model whose number of components is comparable with the true model. They show that the adaptive group Lasso can select the nonzero components correctly with probability approaching one as n increases and achieves the optimal rates of convergence in the "large p, small n" setting. Meier, van de Geer, and Bühlmann (2009) consider an additive model that allows for both the numbers of zero and nonzero f_j's passing to infinity and being larger than n. They propose a sparsity–smoothness penalty for model selection and estimation. Under some conditions, they show that the nonzero components can be selected with probability approaching 1. However, the model selection consistency is not established. Ravikumar et al. (2007) propose a penalized approach for variable selection in nonparametric additive models. They impose penalty on the l_2-norm of the nonparametric components. Under some strong conditions on the design matrix and with special basis functions, they establish the model selection consistency. In all the above three approaches, the penalties are in the form of group/adaptive Lasso, or variants of the group Lasso. In addition, Xue (2009) proposes a penalized polynomial spline method for simultaneous variable selection and model estimation in additive models by using the SCAD penalty. We will review these methods in turn.

Several other papers have also considered variable selection in nonparametric additive models. Bach (2008) applies a method similar to the group Lasso to select variables in additive models with a fixed number of covariates and establishes the model selection consistency under a set of complicated conditions. Avalos, Grandvalet, and Ambroise (2007) propose a method for function estimation and variable selection for additive models fitted by cubic splines, but they don't give any theoretic analysis for their method. Lin and Zhang (2006) propose the component selection and smoothing operator (COSSO) method for model selection and estimation in multivariate nonparametric regression models, in the framework of smoothing spline ANOVA. The COSSO is a method of regularization with the penalty functional being the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. They show that in the special case of a tensor product design, the COSSO correctly selects the nonzero additive component with high probability. More recently, Fan, Feng, and Song (2011) propose several closely related variable screening procedures in sparse ultrahigh-dimensional additive models.

9.3.1. Huang, Horowitz, and Wei's (2010) Adaptive Group Lasso

Huang, Horowitz, and Wei (2010) propose a two-step approach to select and estimate the nonzero components simultaneously in (9.8) when p is fixed. It uses the group Lasso in the first stage and the adaptive group Lasso in the second stage.

Suppose that X_{ij} takes values in $[a,b]$ where $a<b$ are finite numbers. Suppose $E\left[f_j\left(X_{ij}\right)\right]=0$ for $j=1,\ldots,p$ to ensure unique identification of f_j's. Under some suitable smoothness assumptions, f_j's can be well approximated by functions in $\mathcal{S}_n$, a space of polynomial splines defined on $[a,b]$ with some restrictions. There exists a normalized B-spline basis $\{\phi_k, 1\le k\le m_n\}$ for $\mathcal{S}_n$ such that for any $f_{nj}\in\mathcal{S}_n$, we have

$$f_{nj}(x)=\sum_{k=1}^{m_n}\beta_{jk}\phi_k(x), \qquad 1\le j\le p. \tag{9.9}$$

Let $\boldsymbol{\beta}_{nj}=\left(\beta_{j1},\ldots,\beta_{jm_n}\right)'$ and $\boldsymbol{\beta}_n=(\boldsymbol{\beta}'_{n1},\ldots,\boldsymbol{\beta}'_{np})'$. Let $w_n=\left(w_{n1},\ldots,w_{np}\right)'$ be a weight vector and $0\le w_{nj}\le\infty$ for $j=1,\ldots,p$. Huang, Horowitz, and Wei (2010) consider the following penalized least squares (PLS) criterion

$$L_n\left(\mu,\boldsymbol{\beta}_n\right)=\sum_{i=1}^{n}\left[Y_i-\mu-\sum_{j=1}^{p}\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(X_{ij}\right)\right]^2+\lambda_n\sum_{j=1}^{p}w_{nj}\left\|\boldsymbol{\beta}_{nj}\right\|_2 \tag{9.10}$$

subject to

$$\sum_{i=1}^{n}\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(X_{ij}\right)=0, \qquad j=1,\ldots,p, \tag{9.11}$$

where λ_n is a tuning parameter.

Note that (9.10) and (9.11) define a constrained minimization problem. To convert it to an unconstrained optimization problem, one can center the response and the basis functions. Let $\bar{\phi}_{jk}=n^{-1}\sum_{i=1}^{n}\phi_k\left(X_{ij}\right)$, $\psi_{jk}(x)=\phi_k(x)-\bar{\phi}_{jk}$, $Z_{ij}=\left(\psi_{j1}\left(X_{ij}\right),\ldots,\psi_{jm_n}\left(X_{ij}\right)\right)'$, $\mathrm{Z}_j=\left(Z_{1j},\ldots,Z_{nj}\right)'$, $\mathrm{Z}=(\mathrm{Z}_1,\ldots,\mathrm{Z}_p)$, and $\mathrm{Y}=\left(Y_1-\bar{Y},\ldots,Y_n-\bar{Y}\right)'$, where $\bar{Y}=n^{-1}\sum_{i=1}^{n}Y_i$. Note that Z_j is an $n\times m_n$ "design" matrix for the jth covariate. It is easy to verify that minimizing (9.10) subject to (9.11) is equivalent to minimizing

$$L_n\left(\boldsymbol{\beta}_n;\lambda_n\right)=\left\|\mathrm{Y}-\mathrm{Z}\boldsymbol{\beta}_n\right\|_2^2+\lambda_n\sum_{j=1}^{p}w_{nj}\left\|\boldsymbol{\beta}_{nj}\right\|_2. \tag{9.12}$$

In the first step, Huang, Horowitz, and Wei (2010) compute the group Lasso estimator by minimizing $L_{n1}\left(\boldsymbol{\beta}_n;\lambda_{n1}\right) = \left\|\mathrm{Y} - \mathbf{Z}\boldsymbol{\beta}_n\right\|_2^2 + \lambda_{n1}\sum_{j=1}^{p}\left\|\boldsymbol{\beta}_{nj}\right\|_2$, which is a special case of (9.12) by setting $w_{nj} = 1$ for $j = 1,\ldots,p$ and $\lambda_n = \lambda_{n1}$. Denote the resulting group Lasso estimator as $\tilde{\boldsymbol{\beta}}_n \equiv \tilde{\boldsymbol{\beta}}_n(\lambda_{n1})$. In the second step they minimize the adaptive group Lasso objective function $L_n\left(\boldsymbol{\beta}_n;\lambda_{n2}\right)$ by choosing

$$w_{nj} = \begin{cases} ||\tilde{\boldsymbol{\beta}}_{nj}||_2^{-1} & \text{if } ||\tilde{\boldsymbol{\beta}}_{nj}||_2 > 0, \\ \infty & \text{if } ||\tilde{\boldsymbol{\beta}}_{nj}||_2 = 0. \end{cases}$$

Denote the adaptive group Lasso estimator of $\boldsymbol{\beta}_n$ as $\hat{\boldsymbol{\beta}}_n \equiv \hat{\boldsymbol{\beta}}_n(\lambda_{n2}) = (\hat{\boldsymbol{\beta}}'_{n1},\ldots,\hat{\boldsymbol{\beta}}'_{np})'$. The adaptive group Lasso estimators of μ and f_j are then given by

$$\hat{\mu}_n = \bar{Y} \quad \text{and} \quad \hat{f}_{nj}(x) = \sum_{k=1}^{m_n}\hat{\boldsymbol{\beta}}_{nj}\psi_k(x), \qquad 1 \le j \le p.$$

Assume $f_j(x) \neq 0$ for $j \in A_1 = \left\{1,\ldots,p^*\right\} = 0$ for $j \in A_0 = \left\{p^*+1,\ldots,p\right\}$. Let $\hat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n$ denote $\mathrm{sgn}_0(||\hat{\boldsymbol{\beta}}_{nj}||) = \mathrm{sgn}_0(||\boldsymbol{\beta}_{nj}||)$, $1 \le j \le p$, where $\mathrm{sgn}_0(|x|) = 1$ if $|x| > 0$ and $= 0$ if $|x| = 0$. Under some regularity conditions, Huang, Horowitz, and Wei (2010) show that $P(\hat{\boldsymbol{\beta}}_n =_0 \boldsymbol{\beta}_n) \to 1$ as $n \to \infty$, and $\sum_{j=1}^{p^*}||\hat{\boldsymbol{\beta}}_{nj} - \boldsymbol{\beta}_{nj}||_2^2 = O_P(m_n^2/n + 1/m_n^{2d-1} + 4m_n^2\lambda_{n2}^2/n^2)$, where d denotes the smoothness parameter of f_j's (e.g., $d = 2$ if each f_j has continuous second-order derivative). In terms of the estimators of the nonparametric components, they show that

$$P\left(\left\|\hat{f}_{nj}\right\|_2 > 0, j \in A_1 \text{ and } \left\|\hat{f}_{nj}\right\|_2 = 0, j \in A_0\right) \to 1 \qquad \text{as } n \to \infty$$

and

$$\sum_{j=1}^{p^*}\left\|\hat{f}_{nj} - f_j\right\|_2^2 = O_P\left(\frac{m_n}{n} + \frac{1}{m_n^{2d}} + \frac{4m_n\lambda_{n2}^2}{n^2}\right).$$

The above result states that the adaptive group Lasso can consistently distinguish nonzero components from zero components, and gives an upper bound on the rate of convergence of the estimators.

9.3.2. Meier, Geer, and Bühlmann's (2009) Sparsity–Smoothness Penalty

Meier, Geer, and Bühlmann (2009) consider the problem of estimating a high-dimensional additive model in (9.8), where $p = p_n \gg n$. For identification purpose, they assume that all f_j's are centered, that is, $\sum_{i=1}^{n} f_j\left(X_{ij}\right) = 0$ for $j = 1,\ldots,p$. For any vector $a = (a_1, \ldots, a_n)' \in \mathbb{R}^n$, define $\|a\|_n^2 \equiv n^{-1}\sum_{i=1}^n a_i^2$. Let $f_j = \big(f_j(X_{j1}), \ldots, f_j(X_{jn})\big)'$. Meier, Geer, and Bühlmann define the sparsity–smoothness penalty as

$$J(f_j) = \lambda_{1n}\sqrt{\|f_j\|_n^2 + \lambda_{2n} I^2(f_j)},$$

where $I^2(f_j) = \int [f_j''(x)]^2\,dx$ measures the smoothness of f_j with $f_j''(x)$ denoting the second-order derivative of $f_j(x)$, and λ_{1n} and λ_{2n} are two tuning parameters controlling the amount of penalization. The estimator is given by the following PLS problem:

$$\left(\hat{f}_1, \ldots, \hat{f}_p\right) = \underset{f_1,\ldots,f_p \in \mathcal{F}}{\operatorname{argmin}} \left\| Y - \sum_{j=1}^p f_j \right\|_n^2 + \sum_{j=1}^p J(f_j), \tag{9.13}$$

where $\mathcal{F}$ is a suitable class of functions and $Y = (Y_1, \ldots, Y_n)'$. They choose B-splines to approximate each function f_j: $f_j(x) = \sum_{k=1}^{m_n} \beta_{jk}\phi_{jk}(x)$, where $\phi_{jk}(x) : \mathbb{R} \to \mathbb{R}$ are the B-spline basis functions and $\boldsymbol{\beta}_j = \left(\beta_{j1}, \ldots, \beta_{jm_n}\right)' \in \mathbb{R}^{m_n}$ is the parameter vector corresponding to f_j. Let B_j denote the $n \times m_n$ design matrix of B-spline basis of the jth predictor. Denote the $n \times pm_n$ design matrix as $B = \left[B_1, B_2, \ldots, B_p\right]$. By assuming that all f_j's are second-order continuously differentiable, one can reformulate (9.13) as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \| Y - B\boldsymbol{\beta} \|_n^2 + \lambda_{1n} \sum_{j=1}^p \sqrt{\frac{1}{n}\boldsymbol{\beta}_j' B_j B_j' \boldsymbol{\beta}_j + \lambda_{2n}\boldsymbol{\beta}_j'\Omega_j\boldsymbol{\beta}_j}, \tag{9.14}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')'$, and Ω_j is an $m_n \times m_n$ matrix with (k, l)th element given by $\Omega_{j,kl} = \int \phi_{jk}''(x)\phi_{jl}''(x)\,dx$ for $k, l \in \{1, \ldots, m_n\}$. Let $M_j = \frac{1}{n}B_jB_j' + \lambda_{2n}\Omega_j$. Then (9.14) can be written as a general group Lasso problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \| Y - B\boldsymbol{\beta} \|_n^2 + \lambda_{1n} \sum_{j=1}^p \sqrt{\boldsymbol{\beta}_j' M_j \boldsymbol{\beta}_j}. \tag{9.15}$$

By Cholesky decomposition, $M_j = R_j'R_j$ for some $m_n \times m_n$ matrix R_j. Define $\tilde{\boldsymbol{\beta}}_j = R_j\boldsymbol{\beta}_j$ and $\tilde{B}_j = B_jR_j^{-1}$. (9.15) reduces to

$$\widehat{\tilde{\boldsymbol{\beta}}} = \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} \left\| Y - \tilde{B}\tilde{\boldsymbol{\beta}} \right\|_n^2 + \lambda_{1n} \sum_{j=1}^p \left\| \tilde{\boldsymbol{\beta}}_j \right\|, \tag{9.16}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \ldots, \tilde{\boldsymbol{\beta}}_p')'$, and $||\tilde{\boldsymbol{\beta}}_j||$ denotes the Euclidean norm in $\mathbb{R}^{m_n}$. For fixed λ_{2n}, (9.16) is an ordinary group Lasso problem. For large enough λ_{1n} some of the coefficient group $\boldsymbol{\beta}_j \in \mathbb{R}^{m_n}$ will be estimated as exactly zero.

Meier, Geer, and Bühlmann (2009) argue empirically that the inclusion of a smoothness part ($I^2(f_j)$) into the penalty functions yields much better results than having the sparsity term ($\|f_j\|_n$) only. Under some conditions, the procedure can select a set of

f_j's containing all the additive nonzero components. However, the model selection consistency of their procedure is not established. The selected set may include zero components and then be larger than the set of nonzero f_j's.

9.3.3. Ravikumar, Liu, Lafferty, and Wasserman's (2007) Sparse Additive Models

Ravikumar, Liu, Lafferty, and Wasserman (2007) propose a new class of methods for high-dimensional nonparametric regression and classification called *SParse Additive Models* (SPAM). The models combine ideas from sparse linear modeling and additive nonparametric regression. The models they consider take the form (9.8), but they restrict $X_i = \left(X_{i1}, \ldots, X_{ip}\right)' \in [0,1]^p$ and $p = p_n$ can diverge with n. They consider a modification of standard additive model optimization problem as follows

$$\min_{\beta \in \mathbb{R}^p, g_j \in \mathcal{H}_j} E\left[Y_i - \sum_{j=1}^{p} \beta_j g_j\left(X_{ij}\right)\right]^2$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \le L \quad \text{and} \quad E\left[g_j\left(X_{ij}\right)^2\right] = 1, \qquad j = 1, \ldots, p,$$

where $\mathcal{H}_j$ denotes the Hilbert subspace $L_2\left(\mu_j\right)$ of measurable function $f_j\left(X_{ij}\right)$ with $E\left[f_j\left(X_{ij}\right)\right] = 0$ and μ_j being the marginal distribution of X_{ij}, $\beta = (\beta_1, \ldots, \beta_p)'$, and β_j is the rescaling parameter such that $f_j = \beta_j g_j$ and $E[g_j\left(X_{ij}\right)^2] = 1$. The constraint that $\boldsymbol{\beta}$ lies in the l_1-ball $\left\{\beta : \|\beta\|_1 \le L\right\}$ induces sparsity for the estimate $\hat{\beta}$ of β as for the Lasso estimate. Absorbing β_j in f_j, we can re-express the minimization problem in the equivalent Lagrangian form:

$$L(f, \lambda) = \frac{1}{2} E\left[Y_i - \sum_{j=1}^{p} f_j\left(X_{ij}\right)\right]^2 + \lambda_n \sum_{j=1}^{p} \sqrt{E\left[f_j^2\left(X_{ij}\right)\right]}, \tag{9.17}$$

where λ_n is a tuning parameter. The minimizers for (9.17) satisfy

$$f_j\left(X_{ij}\right) = \left[1 - \frac{\lambda_n}{\sqrt{E\left[P_j\left(X_{ij}\right)^2\right]}}\right]_+ P_j\left(X_{ij}\right) \qquad \text{almost surely (a.s.),}$$

where $[\cdot]_+$ denotes the positive part of $\cdot$, and $P_j\left(X_{ij}\right) = E\left(R_{ij} | X_{ij}\right)$ denotes the projection of the residuals $R_{ij} = Y_i - \sum_{k \neq j} f_k(X_{ik})$ onto $\mathcal{H}_j$.

To get a sample version of the above solution, Ravikumar et al. insert the sample estimates into the population algorithm as in standard backfitting. The projection $P_j = (P_j(X_{1j}), \ldots, P_j(X_{nj}))'$ can be estimated by smoothing the residuals:

$$\hat{P}_j = \mathcal{S}_j R_j$$

where $\mathcal{S}_j$ is a linear smoother (e.g., an $n \times n$ matrix for the local linear or sieve smoother) and $R_j = (R_{1j}, \ldots, R_{nj})'$. Let $\hat{s}_j = \sqrt{n^{-1}\sum_{i=1}^{n} \hat{P}_{ij}^2}$ be an estimator of $s_j = \sqrt{E[P_j(X_{ij})^2]}$, where $\hat{P}_{ij}$ denotes the ith element of $\hat{P}_j$. They propose the SPAM backfitting algorithm to solve f_j's as follows: Given regularization parameter λ, initialize $\hat{f}_j(X_{ij}) = 0$ for $j = 1, \ldots, p$, and then iterate the following steps until convergence, for each $j = 1, \ldots, p$:

1. Compute the residual, $R_{ij} = Y_i - \sum_{k \neq j}^{p} \hat{f}_k(X_{ik})$.
2. Estimate P_j by $\hat{P}_j = \mathcal{S}_j R_j$.
3. Estimate s_j by $\hat{s}_j$.
4. Obtain the soft thresholding estimate $\hat{f}_j(X_{ij}) = [1 - \lambda/\hat{s}_i]_+ \hat{P}_{ij}$.
5. Center $\hat{f}_j$ to obtain $\hat{f}_j(X_{ij}) - n^{-1}\sum_{i=1}^{n} \hat{f}_j(X_{ij})$ and use this as an updated estimate of $f_j(X_{ij})$.

The outputs are $\hat{f}_j$, based on which one can also obtain $\sum_{i=1}^{p} \hat{f}_j(X_{ij})$.

If $f_j(x)$ can be written in terms of orthonormal basis functions $\{\psi_{jk} : k = 1, 2, \ldots\}$: $f_j(x) = \sum_{k=1}^{\infty} \beta_{jk}\psi_{jk}(x)$ with $\beta_{jk} = \int f_j(x)\psi_{jk}(x)dx$, we can approximate it by $\tilde{f}_j(x) = \sum_{k=1}^{m_n} \beta_{jk}\psi_{jk}(x)$ where $m_n \to \infty$ as $n \to \infty$. In this case, the smoother $\mathcal{S}_j$ can be taken to be the least squares projection onto the truncated set of basis $\{\psi_{j1}, \ldots, \psi_{jm_n}\}$. The orthogonal series smoother is $\mathcal{S}_j = \Psi_j(\Psi_j'\Psi_j)^{-1}\Psi_j'$, where Ψ_j denotes the $n \times m_n$ design matrix with (i, k)th element given by $\psi_{jk}(X_{ij})$. Then the backfitting algorithm reduces to choosing $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p')'$ to minimize

$$\frac{1}{2n}\left\| Y - \sum_{j=1}^{p} \Psi_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n \sum_{j=1}^{p} \sqrt{\frac{1}{n}\boldsymbol{\beta}_j'\left(\Psi_j'\Psi_j\right)\boldsymbol{\beta}_j}, \tag{9.18}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})'$. Note that (9.18) is a sample version of (9.17), and it can be regarded as a functional version of the group Lasso by using the similar transformation form as used to obtain (9.16). Combined with the soft thresholding step, the update of f_j in the above algorithm can be thought as to minimize

$$\frac{1}{2n}\left\| R_j - \Psi_j \boldsymbol{\beta}_j \right\|_2^2 + \lambda_n \sqrt{\frac{1}{n}\boldsymbol{\beta}_j'\left(\Psi_j'\Psi_j\right)\boldsymbol{\beta}_j}.$$

Under some strong conditions, they show that with truncated orthogonal basis the SPAM backfitting algorithm can recover the correct sparsity pattern asymptotically if the number of relevant variables p^* is bounded [p^* denotes the cardinality of the set $\{1 \leq j \leq p : f_j \neq 0\}$]. That is, their estimator can achieve the selection consistency.

9.3.4. Xue's (2009) SCAD Procedure

Xue (2009) considers a penalized polynomial spline method for simultaneous model estimation and variable selection in the additive model (9.8) where p is fixed. Let $\mathcal{M}_n = \{m_n(x) = \sum_{l=1}^{p} g_l(x_l) : g_l \in \varphi_l^{0,n}\}$ be the approximation space, where $\varphi_l^{0,n} = \{g_l : n^{-1} \sum_{i=1}^{n} g_l(X_{il}) = 0,\ g_l \in \varphi_l\}$ and φ_l is the space of empirically centered polynomial splines of degree $q \geq 1$ on the lth intervals constructed by interior knots on [0,1]. The penalized least squares estimator is given by

$$\hat{m} = \underset{m_n = \sum_{l=1}^{p} f_l \in \mathcal{M}_n}{\operatorname{argmin}} \left[\frac{1}{2} \|Y - m_n\|_n^2 + \sum_{l=1}^{p} p_{\lambda_n}\left(\|f_l\|_n\right) \right],$$

where $Y = (Y_1, \ldots, Y_n)'$, $m_n = (m_n(X_1), \ldots, m_n(X_n))'$, and $p_{\lambda_n}(\cdot)$ is a given penalty function depending on a tuning parameter λ_n. Different penalty functions lead to different variable selection procedures. Noting the desirable properties of the SCAD, they use the spline SCAD penalty.

The proposed polynomial splines have polynomial spline basis representation. Let $J_l = N_l + q$ and $B_l = \{B_{l1}, \ldots, B_{lJ_l}\}$ be a basis for $\varphi_l^{0,n}$. For any fixed $x = (x_1, \ldots, x_p)'$, let $B_l(x_l) = [B_{l1}(x_l), \ldots, B_{lJ_l}(x_l)]'$. One can express $\hat{m}$ as

$$\hat{m}(x) = \sum_{l=1}^{p} \hat{f}_l(x_l) \quad \text{and} \quad \hat{f}_l(x_l) = \hat{\boldsymbol{\beta}}_l' B_l(x_l) \qquad \text{for } l = 1, \ldots, p,$$

where $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \ldots, \hat{\boldsymbol{\beta}}_p')'$ minimizes the PLS criterion:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'}{\operatorname{argmin}} \left[\frac{1}{2} \left\| Y - \sum_{l=1}^{p} \boldsymbol{\beta}_l' B_l \right\|_n^2 + \sum_{l=1}^{p} p_{\lambda_n}\left(\|\boldsymbol{\beta}_l\|_{K_l} \right) \right],$$

where $\|\boldsymbol{\beta}_l\|_{K_l} = \sqrt{\boldsymbol{\beta}_l' K_l \boldsymbol{\beta}_l}$ with $K_l = n^{-1} \sum_{i=1}^{n} B_l(X_{il}) B_l(X_{il})'$.

Under some mild conditions, Xue (2009) shows that the SCAD penalized procedure estimates the nonzero function components with the same optimal mean square convergence rate as the standard polynomial spline estimators, and she correctly sets the zero function components to zero with probability approaching one as the sample size n goes to infinity.

9.4. Variable Selection in Partially Linear Models

In this section, we consider the problem of variable selection in the following *partially linear model* (PLM)

$$Y_i = \beta' X_i + g(Z_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{9.19}$$

where $X_i = \left(X_{i1}, \ldots, X_{ip}\right)'$ is a $p \times 1$ vector of regressors that enter the model linearly, Z_i is a $q \times 1$ vector of regressors that enter the model with an unknown functional form g, and ε_i is an error term such that

$$E(\varepsilon_i | X_i, Z_i) = 0 \qquad \text{a.s.} \tag{9.20}$$

To allow p to increase as n increases, we sometimes write p as p_n below.

Xie and Huang (2009) consider the problem of simultaneous variable selection and estimation (9.19) with a divergent number of covariates in the linear part. Ni, Zhang, and Zhang (2009) propose a *double-penalized least squares* (DPLS) approach to simultaneously achieve the estimation of the nonparametric component g and the selection of important variables in X_i in (9.19). Kato and Shiohama (2009) consider variable selection in (9.19) in the time series framework. In the large p framework, Chen, Yu, Zou, and Liang (2012) propose to use the adaptive elastic-net for variable selection for parametric components by using the profile least squares approach to convert the partially linear model to a classical linear regression model. Liang and Li (2009) consider variable selection in the PLM (9.19), where Z_i is a scalar random variable but X_i is measured with error and is not observable. In addition, Liu, Wang, and Liang (2011) consider the additive PLM where $g(Z_i) = \sum_{k=1}^{q} g_k(Z_{ik})$ in (9.19). Besides, it is worth mentioning that Bunea (2004) considers covariate selection in X_i when the dimension of X_i is fixed and Z_i is a scalar variable in (9.19) based on a BIC type of information criterion and a sieve approximation for $g(\cdot)$. He shows that one can consistently estimate the subset of nonzero coefficients of the linear part and establish its oracle property. But we will not review this paper in detail because the procedure is not a simultaneous variable selection and estimation procedure.

9.4.1. Xie and Huang's (2009) SCAD-Penalized Regression in High-Dimension PLM

Xie and Huang (2009) consider the problem of simultaneous variable selection and estimation in (9.19) when $p = p_n$ is divergent with n and q is fixed. To make it explicit that the coefficients depend on n, one can write $\beta = \beta^{(n)}$. They allow the number p_n^* of nonzero components in $\beta^{(n)}$ diverge with n too.

Since g is unknown, Xie and Huang use the polynomial splines to approximate it. Let $\{B_{nw}(z) : 1 \leq w \leq m_n\}$ be a sequence of basis functions. Let $B(z) = \big(B_{n1}(z), \ldots, B_{nm_n}(z)\big)'$, $\mathbf{B}^{(n)}$ be the $n \times m_n$ matrix whose ith row is $B(Z_i)'$. Under some smoothness conditions, $g(z)$ can be well approximated by $\alpha^{(n)\prime}B(z)$ for some $\alpha^{(n)} \in \mathbb{R}^{m_n}$. Then the problem of estimating g becomes that of estimating $\alpha^{(n)}$. Let $Y = (Y_1, \ldots, Y_n)'$ and $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)'$. Then one can consider the following PLS objective function for estimating $\beta^{(n)}$ and $\alpha^{(n)}$ with the SCAD penalty

$$Q_n\Big(\beta^{(n)}, \alpha^{(n)}\Big) = \Big\| Y - \mathbf{X}^{(n)}\beta^{(n)} - \mathbf{B}^{(n)}\alpha^{(n)} \Big\|_2^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}\Big(\Big|\beta_j^{(n)}\Big|\Big),$$

where $\beta_j^{(n)}$ denotes the jth element of $\beta^{(n)}$, and $p_{\lambda_n}(\cdot)$ is the SCAD penalty function with λ_n as a tuning parameter. Let $(\hat{\beta}^{(n)}, \hat{\alpha}^{(n)})$ denote the solution to the above minimization problem, and $\hat{g}_n(z) = \hat{\alpha}^{(n)\prime}B(z)$.

For any $\beta^{(n)}$, $\alpha^{(n)}$ minimizing Q_n satisfies

$$\mathbf{B}^{(n)\prime}\mathbf{B}^{(n)}\alpha^{(n)} = \mathbf{B}^{(n)\prime}(Y - \mathbf{B}^{(n)}\beta^{(n)}).$$

Let $P_{\mathbf{B}^{(n)}} = \mathbf{B}^{(n)}(\mathbf{B}^{(n)\prime}\mathbf{B}^{(n)})^{-1}\mathbf{B}^{(n)\prime}$ be the projection matrix. It is easy to verify that the profile least squares objective function of the parametric part becomes

$$\tilde{Q}_n\Big(\beta^{(n)}\Big) = \Big\|(I - P_{\mathbf{B}^{(n)}})\Big(Y - \mathbf{X}^{(n)}\beta^{(n)}\Big)\Big\|^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}\Big(\Big|\beta_j^{(n)}\Big|\Big),$$

and $\hat{\beta}^{(n)} = \operatorname{argmin}_{\beta^{(n)}} \tilde{Q}\big(\beta^{(n)}\big)$.

Under some regularity conditions that allow divergent p_n^*, Xie and Huang show that variable selection is consistent, the SCAD penalized estimators of the nonzero coefficients possess the oracle properties, and the estimator of the nonparametric estimate can achieve the optimal convergence rate.

9.4.2. Ni, Zhang, and Zhang's (2009) Double-Penalized Least Squares Regression in PLM

Ni, Zhang, and Zhang (2009) consider a unified procedure for variable selection in the PLM in (9.19) when Z_i is restricted to be a scalar variable on $[0,1]$. To simultaneously achieve the estimation of the nonparametric component g and the selection of important variables in X_i, they propose a *double-penalized least squares* (DPLS) approach by minimizing

$$Q(\beta, g) = \frac{1}{2}\sum_{i=1}^{n}\big[Y_i - \beta' X_i - g(Z_i)\big]^2 + \frac{n\lambda_{1n}}{2}\int_0^1 \big[g''(z)\big]^2 dz + n\sum_{j=1}^{p} p_{\lambda_{2n}}\big(|\beta_j|\big). \quad (9.21)$$

The first penalty term in (9.21) penalizes the roughness of the nonparametric fit $g(z)$, and the second penalty term imposes the usual SCAD penalty on the finite-dimensional parameter β. Let $\mathbf{X} = (X_1,\ldots,X_n)'$, $Y = (Y_1,\ldots,Y_n)'$, and $\mathbf{g} = \left(g(Z_1),\ldots,g(Z_n)\right)'$. It can be shown that given λ_{1n} and λ_{2n}, minimizing (9.21) leads to a smoothing spline estimate for g and one can rewrite the DPLS (9.21) as

$$Q_{dp}(\beta,g) = \frac{1}{2}\left(Y - \mathbf{X}\beta - \mathbf{g}\right)'\left(Y - \mathbf{X}\beta - \mathbf{g}\right) + \frac{n\lambda_{1n}}{2}\mathbf{g}'\mathbf{K}\mathbf{g} + n\sum_{j=1}^{p} p_{\lambda_{2n}}\left(|\beta_j|\right), \quad (9.22)$$

where $\mathbf{K}$ is the non-negative definite smoothing matrix defined by Green and Silverman (1994). Given β, one can obtain the minimizer of $\mathbf{g}$ as $\hat{\mathbf{g}}(\beta) = (I_n + n\lambda_{1n}\mathbf{K})^{-1}(Y - \mathbf{X}\beta)$, where I_n is an $n \times n$ identity matrix. With this, one can obtain readily the profile PLS objective function of β as follows:

$$Q(\beta) = \frac{1}{2}(Y - \mathbf{X}\beta)'\left[I_n - (I_n + n\lambda_{1n}\mathbf{K})^{-1}\right](Y - \mathbf{X}\beta) + n\sum_{j=1}^{p} p_{\lambda_{2n}}\left(|\beta_j|\right).$$

Let $\hat{\beta}$ denote the minimizing solution to the above problem. Ni, Zhang, and Zhang (2009) show that $\hat{\beta}$ has the oracle properties in the case of fixed p under some regularity conditions. In the case where $p = p_n$ is divergent with $p_n \ll n$, they also establish the selection consistency by allowing the number p_n^* of nonzero components in β to be divergent at a slow rate.

9.4.3. Kato and Shiohama's (2009) Partially Linear Models

Kato and Shiohama (2009) consider the PLM in (9.19) in the time series framework by restricting $Z_i = t_i = i/n$ and allowing ε_i to be a linear process. They assume that $g(t_i)$ is an unknown time trend function that can be *exactly* expressed as

$$g(t_i) = \sum_{k=1}^{m} \alpha_k \phi_k(t_i) = \alpha' \boldsymbol{\phi}_i,$$

where $\boldsymbol{\phi}_i = (\phi_1(t_i),\ldots,\phi_m(t_i))'$ is an m-dimensional vector constructed from basis functions $\{\phi_k(t_i)\colon k = 1,\ldots,m\}$, and $\alpha = (\alpha_1,\ldots,\alpha_m)'$ is an unknown parameter vector to be estimated. They propose variable selection via the PLS method:

$$\|Y - X\beta - \boldsymbol{\phi}\alpha\|_2^2 + n\lambda_{0n}\boldsymbol{\alpha}'\mathbf{K}\boldsymbol{\alpha} + n\left(\sum_{j=1}^{p} p_{\lambda_{1n}}\left(|\beta_j|\right) + \sum_{k=1}^{m} p_{\lambda_{2n}}(|\alpha_k|)\right),$$

where $\boldsymbol{\phi} = (\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_n)'$, λ_{0n} in the second term is used to control the tradeoff between the goodness-of-fit and the roughness of the estimated function, K is an appropriate positive semidefinite symmetric matrix, $p_{\lambda_{in}}(\cdot)$ are penalty functions, and $\lambda_{in}, i = 1,2$, are regularization parameters, which control the model complexity. They consider several different penalty functions: the hard thresholding penalty, l_2-penalty in ridge regression, the Lasso penalty, and the SCAD penalty. Under some conditions, they establish the convergence rates for the PLS estimator, and show its oracle property.

9.4.4. Chen, Yu, Zou, and Liang's (2012) Adaptive Elastic-Net Estimator

Chen, Yu, Zou, and Liang (2012) propose to use the adaptive elastic-net (Zou and Zhang, 2009) for variable selection for parametric components when the dimension p is large, using profile least squares approach to convert the PLM to a classical linear regression model.

Noting that $E(Y_i|Z_i) = \beta' E(X_i|Z_i) + g(Z_i)$ under (9.19)–(9.20), we have

$$Y_i - E(Y_i|Z_i) = \beta'[X_i - E(X_i|Z_i)] + \varepsilon_i, \tag{9.23}$$

which is a standard linear model if $E(Y_i|Z_i)$ and $E(X_i|Z_i)$ were known. Let $\hat{E}(Y_i|Z_i)$ and $\hat{E}(X_i|Z_i)$ be the local linear estimators for $E(Y_i|Z_i)$ and $E(X_i|Z_i)$, respectively. Let $\hat{X}_i = X_i - \hat{E}(X_i|Z_i)$, $\hat{Y}_i = Y_i - \hat{E}(Y_i|Z_i)$, $\hat{X} = (\hat{X}_1, \ldots, \hat{X}_n)'$, and $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)'$. Chen, Yu, Zou, and Liang's adaptive Elastic-Net procedure is composed of the following two steps:

1. Construct the Elastic-Net estimator of β given by

$$\hat{\beta}_{\text{Enet}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\beta}{\text{argmin}} \left\{ \left\| \hat{Y} - \hat{X}\beta \right\|_2^2 + \lambda_{2n}\|\beta\|_2^2 + \lambda_{1n}\|\beta\|_1 \right\}.$$

2. Let $\hat{w}_j = |\hat{\beta}_{\text{Enet},j}|^{-\gamma}$ for $j = 1, \ldots, p$ and some $\gamma > 0$, where $\hat{\beta}_{\text{Enet},j}$ denotes the jth element of $\hat{\beta}_{\text{Enet}}$. The adaptive elastic-net estimator of β is given by

$$\hat{\beta}_{\text{AdaEnet}} = \left(1 + \frac{\lambda_2}{n}\right) \underset{\beta}{\text{argmin}} \left\{ \left\| \hat{Y} - \hat{X}\beta \right\|_2^2 + \lambda_{2n}\|\beta\|_2^2 + \lambda_{1n}^* \sum_{j=1}^{p} \hat{w}_j |\beta_j| \right\}.$$

Here, the l_1 regularization parameters λ_{1n} and λ_{1n}^* control the sparsity of the Elastic-Net and adaptive Elastic-Net estimators, respectively. The same λ_{2n} for the l_2-penalty is used in both steps. Under some regular conditions that allows diverging p, they show that profiled adaptive Elastic-Net procedure has the oracle property. In particular, $P(\{j : \hat{\beta}_{\text{AdaEnet},j} \neq 0\} = \mathcal{A}) \to 1$ as $n \to \infty$, where $\mathcal{A} = \{j : \beta_j \neq 0, j = 1, \ldots, p\}$.

9.4.5. Liang and Li's (2009) Variable Selection with Measurement Errors

Liang and Li (2009) also consider the model in (9.19)–(9.20) where Z_i is a scalar random variable, and X_i is measured with error and is not observable. Let W_i denote the observed surrogate of X_i, that is,

$$W_i = X_i + U_i, \tag{9.24}$$

where U_i is the measurement error with mean zero and unknown covariance Σ_{uu}. Assume U_i is independent of (X_i, Z_i, Y_i). Since $E(U_i|Z_i) = 0$, we have $g(Z_i) = E(Y_i|Z_i) - E(W_i|Z_i)'\beta$. The PLS function based on partial residuals is defined as

$$\begin{aligned} L_p(\Sigma_{uu}, \beta) = \frac{1}{2}\sum_{i=1}^{n}\Big\{\big[Y_i - \hat{m}_y(Z_i)\big] - \big[W_i - \hat{m}_w(Z_i)\big]'\beta\Big\} \\ - \frac{n}{2}\beta'\Sigma_{uu}\beta + n\sum_{j=1}^{p} q_{\lambda_j}\big(|\beta_j|\big), \end{aligned} \tag{9.25}$$

where $\hat{m}_y(Z_i)$ and $\hat{m}_w(Z_i)$ are estimators of $E(Y_i|Z_i)$ and $E(W_i|Z_i)$, respectively, and $q_{\lambda_j}(\cdot)$ is a penalty function with a tuning parameter λ_j. The second term is included to correct the bias in the squared loss function due to the presence of measurement error.

The PLS function (9.25) provides a general framework of variable selection in PLMs with measurement errors. In principle, one can use all kinds of penalty functions. But Liang and Li focus on the case of SCAD penalty. Under some conditions, they show that with probability approaching one, there exists a $\sqrt{n}$-consistent PLS estimator $\hat{\beta}$ when Σ_{uu} is estimated from partially replicated observations, and the estimator $\hat{\beta}$ possesses the oracle properties. In addition, they also consider the penalized quantile regression for PLMs with measurement error. See Section 9.8.1.

9.4.6. Liu, Wang, and Liang's (2011) Additive PLMs

Liu, Wang, and Liang (2011) consider the additive PLM of the form

$$Y_i = \beta' X_i + \sum_{k=1}^{q} g_k(Z_{ik}) + \varepsilon_i$$

where $X_i = \big(X_{i1}, \ldots, X_{ip}\big)'$ and $Z_i = \big(Z_{i1}, \ldots, Z_{iq}\big)'$ enter the linear and nonparametric components, respectively, $g_1, \ldots, g_q$ are unknown smooth functions, and $E(\varepsilon_i|X_i, Z_i) = 0$ a.s. They are interested in the variable selection in the parametric component. For identification, assume that $E\big[g_k(Z_{ik})\big] = 0$ a.s. for $k = 1, \ldots, q$. In addition, they assume that both p and q are fixed.

Since g_k's are unknown, Liu, Wang, and Liang propose to use spline approximation. Let $\mathcal{S}_n$ be the space of polynomial functions on $[0,1]$ of degree $\varrho \geq 1$. Let $\mathcal{G}_n$ be the collection of functions g with additive form $g(z) = g_1(z_1) + \cdots + g_K(z_K)$. For the kth covariate z_k, let $b_{jk}(z_k)$ be the B-spline basis functions of degree ϱ. For any $g \in \mathcal{G}_n$, one can write

$$g(z) = \gamma' b(z),$$

where $b(z) = \left\{b_{jk}(z_k), j = -\varrho, \ldots, m_n,\ k = 1, \ldots, q\right\}' \in \mathbb{R}^{m_n q}$, and γ is the corresponding vector of coefficients and its elements are arranged in the same order as $b(z)$. The PLS objective function is given by

$$L_P(\beta, \gamma) = \frac{1}{2} \sum_{i=1}^{n} \left[Y_i - \gamma' b(Z_i) - \beta' X_i\right]^2 + n \sum_{j=1}^{p} q_{\lambda_j}\left(|\beta_j|\right),$$

where $q_{\lambda_j}(\cdot)$ is a penalty function with a tuning parameter λ_j. Let $\hat{\beta}$ be the PLS estimator. Liu, Wang, and Liang consider the SCAD penalty function and show that the SCAD variable selection procedure can effectively identify the significant components with the associated parametric estimators satisfying the oracle properties. But they do not study the asymptotic properties of the estimators of nonparametric components.

9.5. Variable Selection in Functional/Varying Coefficients Models

Nonparametric varying or functional coefficient models (VCMs or FCMs) are useful for studying the time-dependent or the variable-dependent effects of variables. Many methods have been proposed for estimation of these models. See, for example, Fan and Zhang (1998) for the local polynomial smoothing method and see Huang, Wu, and Zhou (2002) and Qu and Li (2006) for the basis expansion and spline method. Several procedures have been developed for variable selection and estimation simultaneously for these models. Wang and Xia (2009) propose adaptive group Lasso for variable selections in VCMs with fixed p based on kernel estimation; Lian (2010) extends their approach by using double adaptive lasso and allowing p to be divergent with n. Zhao and Xue (2011) consider SCAD variable selection for VCMs with measurement error. Li and Liang (2008) consider variable selection in generalized varying-coefficient partially linear models by using the SCAD penalty. In addition Wang, Chen, and Li (2007), Wang, Li, and Huang (2008) and Wei, Huang, and Li (2011) consider sieve-estimation-based variable selection in VCMs with longitudinal data where the penalty takes either the SCAD or group Lasso form.

9.5.1. Wang and Xia's (2009) Kernel Estimation with Adaptive Group Lasso Penalty

Wang and Xia (2009) consider the following varying coefficient model (VCM):

$$Y_i = X_i'\beta(Z_i) + \varepsilon_i, \tag{9.26}$$

where $X_i = \left(X_{i1},\ldots,X_{ip}\right)'$ is a $p \times 1$ vector of covariates, Z_i is a scalar variable that takes values on [0,1], and ε_i is the error term satisfying $E(\varepsilon_i|X_i,Z_i) = 0$ a.s. The coefficient vector $\beta(z) = (\beta_1(z),\ldots,\beta_p(z))' \in \mathbb{R}^p$ is an unknown but smooth function in z, whose true value is given by $\beta_0(z) = (\beta_{01}(z),\ldots,\beta_{0d}(z))'$. Without loss of generality, assume that there exists an integer $p^* \leq p$ such that $0 < E[\beta_{0j}^2(Z_i)] < \infty$ for any $j \leq p^*$ but $E[\beta_{0j}^2(Z_i)] = 0$ for $j > p^*$. The main objection is to select the variables in X_i with nonzero coefficients when p is fixed.

Let $B = (\beta(Z_1),\ldots,\beta(Z_n))'$ and $B_0 = (\beta_0(Z_1),\ldots,\beta_0(Z_n))'$. Note that the last $\left(p-p^*\right)$ columns for B_0 should be 0. The selection of variables becomes identifying sparse columns in the matrix B_0. Following the group Lasso of Yuan and Lin (2006), Wang and Xia propose the following PLS estimate

$$\hat{B}_\lambda = (\hat{\beta}_\lambda(Z_1),\ldots,\hat{\beta}_\lambda(Z_n))' = \operatorname{argmin}_{B\in\mathbb{R}^{n\times p}} Q_\lambda(B),$$

where

$$Q_\lambda(B) = \sum_{i=1}^{n}\sum_{t=1}^{n}[Y_i - X_i\beta(Z_t)]^2 K_h(Z_t - Z_i) + \sum_{j=1}^{p}\lambda_j\left\|b_j\right\|, \tag{9.27}$$

$\lambda = \left(\lambda_1,\ldots,\lambda_p\right)'$ is a vector of tuning parameters, $K_h(z) = h^{-1}K(z/h)$, $K(\cdot)$ is a kernel function, h is a bandwidth parameter, b_j denotes the jth column of B for $j = 1,\ldots,p$, and $\|\cdot\|$ denotes the usual Euclidean norm. Let $\hat{b}_{\lambda,k}$ denote the kth column of $\hat{B}_\lambda$ for $k = 1,\ldots,p$ so that we can also write $\hat{B}_\lambda = (\hat{b}_{\lambda,1},\ldots,\hat{b}_{\lambda,p})$. They propose an iterated algorithm based on the idea of the local quadratic approximation of Fan and Li (2001). Let $\tilde{B}$ be an initial estimate of B_0 and $\hat{B}_\lambda^{(m)} = (\hat{b}_{\lambda,1}^{(m)},\ldots,\hat{b}_{\lambda,p}^{(m)}) = (\hat{\beta}_\lambda^{(m)}(Z_1),\ldots,\hat{\beta}_\lambda^{(m)}(Z_n))'$ be the Lasso estimate in the mth iteration. The objective function in (9.27) can be locally approximated by

$$\sum_{i=1}^{n}\sum_{t=1}^{n}[Y_i - X_i\beta(Z_t)]^2 K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j\frac{\left\|b_j\right\|^2}{\left\|\hat{b}_{\lambda,j}^{(m)}\right\|}.$$

Let $\hat{B}_\lambda^{(m+1)}$ denote the minimizer. Its ith row is given by

$$\hat{\beta}_\lambda^{(m+1)}(Z_t) = \left[\sum_{i=1}^{n}X_iX_i'K_h(Z_t - Z_i) + D^{(m)}\right]^{-1}\sum_{i=1}^{n}X_iY_iK_h(Z_t - Z_i),$$

where $D^{(m)}$ is a $p \times p$ diagonal matrix with its jth diagonal component given by $\lambda_j/||\hat{b}^{(m)}_{\lambda,j}||, j=1,\ldots,p$. The estimate for $\beta(z)$ is

$$\hat{\beta}_\lambda(z) = \left[\sum_{i=1}^{n} X_i X_i' K_h(z-Z_i) + D^{(m)}\right]^{-1} \sum_{i=1}^{n} X_i Y_i K_h(z-Z_i).$$

Let $\hat{\beta}_{a,\lambda}(z) = (\hat{\beta}_{\lambda,1}(z),\ldots,\hat{\beta}_{\lambda,p^*}(z))'$ and $\hat{\beta}_{b,\lambda}(z) = (\hat{\beta}_{\lambda,p^*+1}(z),\ldots,\hat{\beta}_{\lambda,p}(z))'$. Under some regular conditions, Wang and Xia show that (i) $P(\sup_{z\in[0,1]}||\hat{\beta}_{\lambda,b}(z)||=0) \to 1$; and (ii) $\sup_{z\in[0,1]}||\hat{\beta}_{a,\lambda}(z) - \hat{\beta}_{ora}(z)|| = o_P\left(n^{-2/5}\right)$, where

$$\hat{\beta}_{ora}(z) = \left[\sum_{i=1}^{n} X_{i(a)} X_{i(a)}' K_h(Z_t - Z_i)\right]^{-1} \sum_{i=1}^{n} X_{i(a)} Y_i K_h(z-Z_i)$$

stands for the oracle estimator, and $X_{i(a)} = \left(X_{i1},\ldots,X_{ip_0}\right)'$. The second part implies that $\hat{\beta}_{a,\lambda}(z)$ has the oracle property.

They also propose to choose the tuning parameters λ by

$$\lambda_j = \frac{\lambda_0}{n^{-1/2}||\tilde{\beta}_j||},$$

where $\tilde{\beta}_j$ is the jth column of the unpenalized estimate $\tilde{B}$. λ_0 can be selected according to the following BIC-type criterion:

$$\text{BIC}_\lambda = \log(RRS_\lambda) + df_\lambda \times \frac{\log(nh)}{nh}$$

where $0 \leq df_\lambda \leq p$ is the number of nonzero coefficients identified by $\hat{B}_\lambda$ and RRS_λ is defined as

$$RRS_\lambda = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{t=1}^{n}\left[Y_i - X_i'\hat{\beta}_\lambda(Z_t)\right]^2 K_h(Z_i - Z_t).$$

Under some conditions, they show that the tuning parameter $\hat{\lambda}$ selected by the BIC-type criterion can identify the true model consistently.

9.5.2. Lian's (2010) Double Adaptive Group Lasso in High-Dimensional VCMs

Lian (2010) studies the problem of simultaneous variable selection and constant coefficient identification in high-dimensional VCMs based on B-spline basis expansion. He considers the VCM in (9.26) but allows for $p = p_n \gg n$. In addition, he explicitly allows some nonzero coefficients in $\beta(Z_t)$ to be constant a.s.

Using spline expansions, $\beta(z)$ can be approximated by $\sum_{k=1}^{m_n} b_{jk}B_k(z)$, where $\{B_k(z)\}_{k=1}^{m_n}$ is a normalized B-spline basis. Lian proposes the following PLS estimate:

$$\hat{b} = \operatorname{argmin}_b \frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\sum_{k=1}^{m_n} X_{ij}b_{jk}B_k(Z_i)\right]^2 + n\lambda_1\sum_{j=1}^{p} w_{1j}\|b_j\| + n\lambda_2\sum_{j=1}^{p} w_{2j}\|b_j\|_c,$$

where λ_1, λ_2 are regularization parameters, $w_1 = \left(w_{11},\ldots,w_{1p}\right)'$ and $w_2 = \left(w_{21},\ldots,w_{2p}\right)'$ are two given vectors of weights, $b_j = \left(b_{j1},\ldots,b_{jm_n}\right)'$, $\|b_j\| = \sqrt{\sum_{k=1}^{m_n} b_{jk}^2}$, and $\|b_j\|_c = \sqrt{\sum_{k=1}^{m_n}\left[b_{jk} - \bar{b}_j\right]^2}$ with $\bar{b}_j = m_n^{-1}\sum_{k=1}^{m_n} b_{jk}$. The first penalty is used for identifying the zero coefficients, while the second is used for identifying the nonzero constant coefficients.

The minimization problem can be solved by the locally quadratic approximation as Fan and Li (2001) and Wang and Xia (2009). He also proposes a BIC-type criterion to select λ_1 and λ_2. Under some suitable conditions, he shows that consistency in terms of both variable selection and constant coefficients identification can be achieved, and the oracle property of the constant coefficients can be established.

9.5.3. Zhao and Xue's (2011) SCAD Variable Selection for VCMs with Measurement Errors

Zhao and Xue (2011) consider variable selection for the VCM in (9.26) when the covariate X_i is measured with errors and Z_i is error-free. That is, X_i is not observed but measured with additive error:

$$\xi_i = X_i + V_i,$$

where V_i is the measurement error that is assumed to be independent of $(X_i, Z_i, \varepsilon_i)$ and have zero mean and variance–covariance matrix Σ_{VV}.

Like Lian (2010), Zhao and Xue propose to approximate $\beta(z)$ by a linear combination of B-spline basis functions. Let $B(z) = (B_1(z),\ldots,B_{m_n}(z))'$, $W_i = (X_{i1}B(Z_i)',\ldots,X_{ip}B(Z_i)') = [I_p \otimes B(Z_i)]X_i$, and $b = (b_1',\ldots,b_p')'$. Let $\tilde{W}_i = (\xi_{i1}B(Z_i)',\ldots,\xi_{ip}B(Z_i)') = [I_p \otimes B(Z_i)]\xi_i$, where ξ_{ij} denotes the jth element in ξ_i. Observing that

$$E\left[\tilde{W}_i\tilde{W}_i'|X_i, Z_i\right] = W_iW_i' + \Omega(Z_i),$$

where $\Omega(Z_i) = [I_p \otimes B(Z_i)]\Sigma_{VV}[I_p \otimes B(Z_i)]'$, they propose a bias-corrected PLS objective function

$$Q(b) = \sum_{i=1}^{n}\left[Y_i - b'\tilde{W}_i\right]^2 - \sum_{i=1}^{n}\sum_{j=1}^{p} b'\Omega(Z_i)b + n\sum_{j=1}^{p} q_\lambda\left(\|b_j\|_H\right),$$

where $q_\lambda(\cdot)$ is the SCAD penalty function with λ as a tuning parameter and $\left\| b_j \right\|_H = \left(b'Hb\right)^{1/2}$ with $H = \int_0^1 B(z)B(z)'dz$. They establish the consistency of the variable selection procedure and derive the optimal convergence rate of the regularized estimators.

9.5.4. Li and Liang's (2008) Variable Selection in Generalized Varying-Coefficient Partially Linear Model

Li and Liang (2008) consider the generalized varying-coefficient partially linear model (GVCPLM)

$$g\{\mu(u,x,z)\} = x'\alpha(u) + z'\beta,$$

where $\mu(u,x,z) = E(Y_i|U_i = u, X_i = x, Z_i = z)$, X_i is p-dimensional, Z_i is q-dimensional, and U_i is a scalar random variable. They assume that p and q are fixed and focus on the selection of significant variables in the parametric component based on i.i.d. observations (Y_i, U_i, X_i, Z_i), $i = 1, \ldots, n$. The conditional quasi-likelihood of Y_i is $Q(\mu(U_i, X_i, Z_i))$, where

$$Q(\mu, y) = \int_\mu^y \frac{s-y}{V(s)} ds$$

and $V(s)$ is a specific variance function. Then the penalized likelihood function is defined by

$$L(\alpha,\beta) = \sum_{i=1}^{n} Q\left[g^{-1}\left(X_i'\alpha(U_i) + Z_i'\beta\right), Y_i\right] - n\sum_{j=1}^{q} p_{\lambda_j}\left(|\beta_j|\right), \tag{9.28}$$

where β_j denotes the jth element of β, $p_{\lambda_j}(\cdot)$ is a specific penalty function with a tuning parameter λ_j, and g^{-1} denotes the inverse function of g. They propose to use the SCAD penalty. Since $\alpha(u)$ is unknown, they first use the local likelihood technique to estimate $\alpha(u)$ and then substitute the resulting estimate into the above penalized likelihood function and finally maximize (9.28) with respect to β. Under some conditions, they establish the rate of convergence for the resulting PLS estimator $\hat{\beta}$ of β. With proper choices of penalty functions and tuning parameters, they show the asymptotic normality of $\hat{\beta}$ and demonstrate that the proposed procedure performs as well as an oracle procedure. To select variables in X_i that are associated with the nonparametric component, they propose a *generalized likelihood ratio* (GLR) test statistic to test the null hypothesis of some selected components being zero.

9.5.5. Sieve-estimation-Based Variable Selection in VCMs with Longitudinal Data

Several papers address the issue of variable selection in VCMs with longitudinal data. They base the variable selection on sieve estimation with either SCAD or Lasso penalty with balanced or unbalanced data.

Let $Y_i(t_j)$ be the expression level of the ith individual at time t_j, where $i = 1, \ldots, n$ and $j = 1, \ldots, T$. Wang, Chen, and Li (2007) consider the following VCM:

$$Y_i(t) = \mu(t) + \sum_{k=1}^{p} \beta_k(t) X_{ik} + \varepsilon_i(t), \tag{9.29}$$

where $\mu(t)$ indicates the overall mean effect, $\varepsilon_i(t)$ is the error term, and other objects are defined as above. They approximate $\beta_k(t)$ by using the natural cubic B-spline basis: $\beta_k(t) \approx \sum_{l=1}^{m_n} \beta_{kl} B_l(t)$, where $B_l(t)$ is the natural cubic B-spline basis function, for $l = 1, \ldots, m_n$, and the number of interior knots is given by $m_n - 4$. They propose a general *group SCAD* (gSCAD) procedure for selecting the groups of variables in a linear regression setting. Specifically, to select nonzero $\beta_k(t)$, they minimize the following PLS loss function:

$$L(\beta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{T} \left\{ y_{it} - \mu\left(t_j\right) - \sum_{k=1}^{p} \sum_{l=1}^{m_n} \beta_{kl} \left[B_l\left(t_j\right) X_{ik} \right] \right\}^2 + nT \sum_{k=1}^{p} p_\lambda\left(\left\| \boldsymbol{\beta}_k \right\|\right),$$

where $y_{it} = Y_i\left(t_j\right)$, $\mu = (\mu(t_1), \ldots, \mu(t_T))'$, $p_\lambda(\cdot)$ is the SCAD penalty with λ as a tuning parameter, and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{km_n})'$. An iterative algorithm based on local quadratic approximation of the non-convex penalty $p_\lambda\left(\left\| \boldsymbol{\beta}_k \right\|\right)$ as in Fan and Li (2001) is proposed. Under some overly restrictive conditions such as the knot, locations are held *fixed* as the sample size increases, they generalize the arguments in Fan and Li (2001) to the group selection settings and establish the oracle property of gSCAD group selection procedure.

Wang, Li, and Huang (2008) consider a model similar to (9.29) but allow for unbalanced data:

$$Y_i\left(t_{ij}\right) = \sum_{k=1}^{p} \beta_k\left(t_{ij}\right) X_{ik}\left(t_{ij}\right) + \varepsilon_i\left(t_{ij}\right), \tag{9.30}$$

where $i = 1, \ldots, n$, $j = 1, \ldots, n_i$, $X_{ik}(t)$ is the covariate with time-varying effects, and the number of covariates p is fixed. They propose a PLS estimator using the SCAD penalty and basis expansion. The coefficients $\beta_k(t)$ can be approximated by a basis expansion $\beta_k(t) \approx \sum_{l=1}^{m_{nk}} \beta_{kl} B_{kl}(t)$ where various basis systems including Fourier bases, polynomial bases, and B-spline bases can be used in the basis expansion to obtain $B_{kl}(t)$

for $l = 1, \ldots, m_{nk}$. Their objective function is given by

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{T_i} \left\{ Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{m_{nk}} \beta_{kl} [B_{kl}(t_j) X_{ik}(t_{ij})] \right\}^2 + \sum_{k=1}^{p} p_{\lambda} \left(\|\boldsymbol{\beta}_k\|_{\mathrm{R}_k} \right),$$

where the w_i's are weights taking value 1 if we treat all observations equally or $1/T_i$ if we treat each individual subject equally, $\|\boldsymbol{\beta}_k\|_{\mathrm{R}_k}^2 = \boldsymbol{\beta}_k' \mathrm{R}_k \boldsymbol{\beta}_k$, R_k is an $m_{nk} \times m_{nk}$ kernel matrix whose (i,j)th element is given by

$$r_{k,ij} = \int_0^1 B_{ki}(t) B_{kj}(t)\, dt.$$

Under suitable conditions, they establish the theoretical properties of their procedure, including consistency in variable selection and the oracle property in estimation.

More recently, Wei, Huang, and Li (2011) also consider the model in (9.30) but allow the number of variables $p(=p_n)$ to be larger than n. They apply the group Lasso and basis expansion to simultaneously select the important variables and estimate the coefficient functions. The objective function of the group Lasso is given by

$$\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{T_i} \left\{ Y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{m_{nk}} \beta_{kl} [B_{kl}(t_j) X_{ik}(t_{ij})] \right\}^2 + \sum_{k=1}^{p} \lambda \|\boldsymbol{\beta}_k\|_{\mathrm{R}_k},$$

where $\|\boldsymbol{\beta}_k\|_{\mathrm{R}_k}$ is defined as above. Under some conditions, they establish the estimation consistency of group Lasso and the selection consistency of adaptive group Lasso.

9.6. Variable Selection in Single Index Models

As a natural extension of linear regression model, the single index model (SIM) provides a specification that is more flexible than parametric models while retaining the desired properties of parametric models. It also avoids the curse of dimensionality through the index structure. Many methods have been proposed to estimate the coefficients in SIMs. Most of them can be classified into three categories. The first category includes the average derivative estimation method (Härdle and Stoker, 1989), the structure adaptive method (Hristache et al., 2001) and the outer product of gradients method (Xia et al., 2002), which only focus on the estimation of unknown coefficients. The second category consists of methods that estimate the unknown link function and coefficients simultaneously, including Ichimura's (1993) semiparametric least square estimation and the minimum average conditional variance estimation (MAVE) by Xia et al. (2002). The third one is related to the inverse regression and is developed for sufficient dimension reduction (SDR) (see, e.g., Li (1991)).

Variable selection is a crucial problem in SIMs. Many classical variable selection procedures for linear regressions have been extended to SIMs. See, for example, Naik and Tsai (2001) and Kong and Xia (2007) for AIC and cross-validation. Based on the comparison of all the subsets of predictor variables, these methods are computationally intensive. Recently, Peng and Huang (2011) use the penalized least squares method to estimate the model and select the significant variables simultaneously. Zeng, He, and Zhu (2011) consider a Lasso-type approach called sim-Lasso for estimation and variable selection. Liang, Liu, Li, and Tsai (2010) consider variable selection in partial linear single-indexed models using the SCAD penalty. Yang (2012) considers variable selection for functional index coefficient models. Some variable selection procedures are also proposed for generalized SIMs (see Zhu and Zhu (2009), Zhu, Qian, and Lin (2011), and Wang, Xu, and Zhu (2012)).

9.6.1. Peng and Huang's (2011) Penalized Least Squares for SIM

Peng and Huang (2011) consider the SIM

$$Y_i = g\big(X_i'\beta\big) + \varepsilon_i, \qquad i = 1, 2, \ldots, n, \tag{9.31}$$

where $g(\cdot)$ is a smooth unknown function, X_i is a $p \times 1$ vector of covariates, $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of parameters, and ε_i is a white noise with unknown variance σ^2. For identification, let $\|\beta\| = 1$ and $\text{sign}(\beta_1) = 1$ where $\text{sign}(a) = 1$ if $a > 0, = -1$ otherwise. They follow the idea of Carroll et al. (1997) and use an iterative algorithm to estimate β and the link function g simultaneously.

The unknown function $g(\cdot)$ can be approximated locally by a linear function

$$g(v) \approx g(u) + g'(u)(v - u)$$

when v lies in the neighborhood of u, and $g'(u) = dg(u)/du$. Given β, one can estimate $g(u)$ and $g'(u)$ by choosing (a, b) to minimize

$$\sum_{i=1}^{n} \big[Y_i - a - b\big(X_i'\beta - u\big)\big]^2 k_h\big(X_i'\beta - u\big), \tag{9.32}$$

where $k_h(\cdot) = k(\cdot/h)/h$ and $k(\cdot)$ is a symmetric kernel function. Let $\hat{g}(\cdot, \beta)$ denote the estimate of $g(u)$ given β. Given $\hat{g}(\cdot, \beta)$, one can estimate β by minimizing the following PLS function

$$\sum_{i=1}^{n} \big[Y_i - \hat{g}\big(X_i'\beta, \beta\big)\big]^2 + n \sum_{j=1}^{p} p_\lambda\big(|\beta_j|\big), \tag{9.33}$$

where $p_\lambda(\cdot)$ is the SCAD penalty function with a tuning parameter λ. To solve the above nonlinear optimization problem, Peng and Huang propose to use the local approximation idea and update the estimate of β given the current estimates $\beta^{(0)}$ and $\hat{g}$ by minimizing the following penalized least squares function:

$$\sum_{i=1}^{n}\Big[Y_i-\hat{g}\Big(X_i'\hat{\beta}^{(0)},\hat{\beta}^{(0)}\Big)-\hat{g}'\Big(X_i'\hat{\beta}^{(0)},\hat{\beta}^{(0)}\Big)X_i'\Big(\beta-\hat{\beta}^{(0)}\Big)\Big]^2+n\sum_{j=1}^{p}p_\lambda\big(|\beta_j|\big). \quad (9.34)$$

The estimation procedure for β and $g(\cdot)$ is summarized as follows:

1. Obtain an initial estimate of β, say $\hat{\beta}^{(0)}$, by the least squares regression of Y_i on X_i. Let $\hat{\beta}=\hat{\beta}^{(0)}/\|\hat{\beta}^{(0)}\|\cdot\text{sign}(\hat{\beta}_1^{(0)})$, where $\hat{\beta}_1^{(0)}$ is the first element of $\hat{\beta}^{(0)}$.
2. Given $\hat{\beta}$, find $\hat{g}(u,\hat{\beta})=\hat{a}$ and $\hat{g}'(u,\hat{\beta})=\hat{b}$ by minimizing (9.32).
3. Update the estimate of β by minimizing (9.34) with $\hat{\beta}^{(0)}$ being replaced by $\hat{\beta}$.
4. Continue steps 2–3 until convergence.
5. Given the final estimate $\hat{\beta}$ from step 4, refine the estimate $\hat{g}(u,\hat{\beta})$ of $g(\cdot)$ by minimizing (9.32).

Peng and Huang argue that the above iterative algorithm can be regarded as an EM algorithm and different bandwidth sequence should be used in steps 2–3 and 5. In steps 2–3, one should assure the accuracy of the estimate of β and thus an undersmoothing bandwidth should be used to obtain the estimate of g; in step 5, one can obtain the final estimate of g by using the optimal bandwidth as if β were known. They discuss the choice of these two bandwidths and the tuning parameter λ as well. Under some conditions, they derive the convergence rate for $\hat{\beta}$ and show its oracle property.

9.6.2. Zeng, He, and Zhu's (2011) Lasso-Type Approach for SIMs

Zeng, He, and Zhu (2011) consider a Lasso-type approach called sim-Lasso for estimation and variable selection for the SIM in (9.32). The sim-Lasso method penalizes the derivative of the link function and thus can be considered as an extension of the usual Lasso. They propose the following PLS minimization problem:

$$\min_{a,b,\beta,\|\beta\|=1}\sum_{j=1}^{n}\sum_{i=1}^{n}\big[Y_i-a_j-b_j\beta'\big(X_i-X_j\big)\big]^2w_{ij}+\lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k|, \quad (9.35)$$

where $a=(a_1,\ldots,a_n)'$, $b=(b_1,\ldots,b_n)'$, $w_{ij}=K_h\big(X_i-X_j\big)/\sum_{l=1}^{n}K_h\big(X_l-X_j\big)$, $K_h(\cdot)=K(\cdot/h)/h^p$, $K(\cdot)$ is a kernel function and h is the bandwidth, and λ is the tuning parameter. Denote the objective function in (9.35) as $LM_\lambda(a,b,\beta)$ and denote its minimizer as $\hat{a}(\lambda)=\big(\hat{a}_1(\lambda),\ldots,\hat{a}_n(\lambda)\big)'$, $\hat{b}(\lambda)=(\hat{b}_1(\lambda),\ldots,\hat{b}_n(\lambda))'$, and $\hat{\beta}(\lambda)$.

Note that the first part of $LM_\lambda(a,b,\beta)$ is the objective function for the MAVE estimation of β. Its inner summation is

$$\sum_{i=1}^{n}\big[Y_i - a_j - b_j\beta'(X_i - X_j)\big]^2 w_{ij}, \tag{9.36}$$

which is the least squares loss function for the local smoothing of g at $\beta' X_j$. A natural way to penalize (9.36) is to penalize the vector of linear coefficient $b_j\beta$ via the lasso type of penalty, yielding

$$\sum_{i=1}^{n}\big[Y_i - a_j - b_j\beta'(X_i - X_j)\big]^2 w_{ij} + \lambda|b_j|\sum_{k=1}^{p}|\beta_k|. \tag{9.37}$$

Summing (9.37) over i leads to the objective function in (9.35). The penalty term $\lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k|$ has twofold impact on the estimation of β. First, as the usual Lasso, it makes $\hat{\beta}(\lambda)$ sparse and thus performs variable selection. Second, it also enforces shrinkage in $\hat{b}(\lambda)$ and may shrink some $\hat{b}_i(\lambda)$'s to zero. The second point is important because when g is relatively flat, its derivative is close to zero and does not contain much information about β.

Given β and b, the target function $LM_\lambda(a,b,\beta)$ can be minimized by $a_j = \tilde{Y}_i - b_j\beta'\left(\tilde{X}_i - X_j\right)$, where $\tilde{Y}_i = \sum_{i=1}^{n} w_{ij}Y_j$ and $\tilde{X}_i = \sum_{i=1}^{n} w_{ij}X_j$. Then $LM_\lambda(a,b,\beta)$ can be simplified to

$$\begin{aligned} L_\lambda(b,\beta) &= \min_a LM_\lambda(a,b,\beta) \\ &= \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda\sum_{j=1}^{n}|b_j|\sum_{k=1}^{p}|\beta_k|. \end{aligned}$$

When β is fixed, the target function L_λ is decoupled into n separate target functions, that is, $L_\lambda = \sum_{i=1}^{n} L_{\lambda,i}$, where

$$L_{\lambda,i} = \sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda_\beta^*|b_j|$$

and $\lambda_\beta^* = \lambda\sum_{k=1}^{p}|\beta_k|$. The solution is

$$\hat{b}_j = \operatorname{sgn}(\beta' R_j)\left(\frac{|\beta' R_j| - \lambda\sum_{k=1}^{p}|\beta_k|/2}{\beta' S_j\beta}\right)^{+}, \tag{9.38}$$

where

$$R_j = \sum_{i=1}^{n}\left(Y_i - \tilde{Y}_j\right)\left(X_i - \tilde{X}_j\right)w_{ij} \quad \text{and} \quad S_j = \sum_{i=1}^{n}\left(X_i - \tilde{X}_j\right)\left(X_i - \tilde{X}_j\right)' w_{ij}.$$

For fixed b, minimizing $L_\lambda(b,\beta)$ becomes

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{n}\sum_{i=1}^{n}\left[Y_i - \tilde{Y}_i - b_j\beta'\left(X_i - \tilde{X}_j\right)\right]^2 w_{ij} + \lambda_b^* \sum_{k=1}^{p}|\beta_k|, \tag{9.39}$$

where $\lambda_b^* = \lambda\sum_{j=1}^{n}\left|b_j\right|$. It can be solved by the LARS–Lasso algorithm. The algorithm is summarized as follows:

1. Get an initial estimate $\hat{\beta}$ of β.
2. Given $\hat{\beta}$, calculate $\hat{b}_j$ as (9.38).
3. Given $\hat{b} = (\hat{b}_1,\ldots,\hat{b}_n)'$, use the LARS–Lasso algorithm to solve (9.39).
4. Renormalize $\hat{b}$ to $||\hat{\beta}||\hat{b}$ and $\hat{\beta}$ to $\hat{\beta}/||\hat{\beta}||$ and use them as $\hat{b}$ and $\hat{\beta}$ below.
5. Repeat steps 2–4 until $(\hat{\beta},\hat{b})$ converges.

Zeng, He, and Zhu (2011) propose to use 10-fold cross-validation procedure to choose the penalty parameter λ and use the rule of thumb for bandwidth. They focus on the computational aspect of sim-Lasso but have not established its theoretical properties. They conjecture that by choosing the penalty parameter λ properly, the sim-Lasso possesses the usual consistency and convergence rate, but admit that the proof is nontrivial due to the interaction between the bandwidth h and the penalty parameter λ.

9.6.3. Liang, Liu, Li, and Tsai's (2010) Partially Linear Single-Index Models

Liang, Liu, Li, and Tsai (2010) consider the following partially linear single-index model (PLSIM):

$$Y_i = \eta\left(Z_i'\boldsymbol{\alpha}\right) + X_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1,\ldots,n, \tag{9.40}$$

where Z_i and X_i are q-dimensional and p-dimensional covariate vectors, respectively, $\boldsymbol{\alpha} = \left(\alpha_1,\ldots,\alpha_q\right)'$, $\boldsymbol{\beta} = \left(\beta_1,\ldots,\beta_p\right)'$, $\eta(\cdot)$ is an unknown differentiable function, ε_i is random error with zero mean and finite variance σ^2, and (X_i, Z_i) and ε_i are independent. They assume that $\|\boldsymbol{\alpha}\| = 1$ and α_1 is positive for identification. They propose a profile least squares procedure to estimate the model and the SCAD penalty to select the significant variables.

Let $Y_i^* = Y_i - X_i'\boldsymbol{\beta}$ and $\Lambda_i = Z_i'\boldsymbol{\alpha}$. For given $\boldsymbol{\xi} = \left(\boldsymbol{\alpha}',\boldsymbol{\beta}'\right)'$, $\eta(\cdot)$ can be estimated by the local linear regression to minimize

$$\sum_{i=1}^{n}\left[Y_i - a - b(\Lambda_i - u) - X_i'\boldsymbol{\beta}\right]^2 k_h(\Lambda_i - u),$$

with respect to a and b, where k_h is defined as before. Let $(\hat{a}, \hat{b})$ denote the minimizer. Then the profile estimator of $\eta(\cdot)$ is given by

$$\hat{\eta}(u;\boldsymbol{\xi}) = \hat{a} = \frac{K_{20}(u,\boldsymbol{\xi})K_{01}(u,\boldsymbol{\xi}) - K_{10}(u,\boldsymbol{\xi})K_{11}(u,\boldsymbol{\xi})}{K_{00}(u,\boldsymbol{\xi})K_{20}(u,\boldsymbol{\xi}) - K_{10}^2(u,\boldsymbol{\xi})},$$

where $K_{lj}(u,\boldsymbol{\xi}) = \sum_{i=1}^{n} k_h(\Lambda_i - u)(\Lambda_i - u)^l \big(X_i'\boldsymbol{\beta} - Y_i\big)^j$ for $l = 0,1,2$ and $j = 0,1$. They consider a PLS function

$$L_p(\boldsymbol{\xi}) = \frac{1}{2}\sum_{i=1}^{n}\big[Y_i - \hat{\eta}\big(Z_i'\boldsymbol{\alpha};\boldsymbol{\xi}\big) - X_i'\boldsymbol{\beta}\big]^2 + n\sum_{j=1}^{q} p_{\lambda_{1j}}\big(|\alpha_j|\big) + n\sum_{k=1}^{p} p_{\lambda_{2k}}(|\beta_k|),$$

where $p_\lambda(\cdot)$ is a penalty function with a regularization parameter λ. Different penalty functions for different elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are allowed. If one wants to select variables in X_i only, one can set $p_{\lambda_{2k}}(\cdot) = 0$ for $k = 1,\ldots,p$. Similarly, if one wants to select variables in Z_i only, one can set $p_{\lambda_{1j}}(\cdot) = 0$ for $j = 1,\ldots,q$.

Because it is computationally expensive to minimize a criterion function with respect to $\big(p+q\big)$-dimensional regularization parameters, Liang, Liu, Li, and Tsai follow the approach of Fan and Li (2004) and set $\lambda_{1j} = \lambda SE(\hat{\alpha}_j^u)$ and $\lambda_{2k} = \lambda SE(\hat{\beta}_k^u)$, where λ is the tuning parameter, and $SE(\hat{\alpha}_j^u)$ and $SE(\hat{\beta}_k^u)$ are the standard errors of the unpenalized profile least squares estimators of α_j and β_k, respectively, for $j = 1,\ldots,q$ and $k = 1,\ldots,p$. Then they propose to use the SCAD penalty and select λ by minimizing the BIC-like criterion function given by

$$BIC(\lambda) = \log\{MSE(\lambda)\} + \frac{\log n}{n} df_\lambda,$$

where $MSE(\lambda) = n^{-1}\sum_{i=1}^{n}\Big[Y_i - \hat{\eta}(Z_i'\hat{\boldsymbol{\alpha}}_\lambda;\hat{\boldsymbol{\xi}}_\lambda) - X_i'\hat{\boldsymbol{\beta}}_\lambda\Big]^2$, $\hat{\boldsymbol{\xi}}_\lambda = (\hat{\boldsymbol{\alpha}}_\lambda', \hat{\boldsymbol{\beta}}_\lambda')'$ is the SCAD estimator of $\boldsymbol{\xi}$ by using the tuning parameter λ, and df_λ is the number of nonzero coefficients in $\hat{\boldsymbol{\xi}}_\lambda$. They show that the BIC tuning parameter selector enables us to select the true model consistently and their estimate enjoys the oracle properties under some mild conditions.

In addition, it is worth mentioning that Liang and Wang (2005) consider the PLSIM in (9.40) when X_i is measured with additive error: $W_i = X_i + U_i$, where U_i is independent of (Y_i, X_i, Z_i). They propose two kernel estimation methods for this model but do not discuss the variable selection issue.

9.6.4. Yang's (2012) Variable Selection for Functional Index Coefficient Models

Yang (2012) considers the following functional index coefficient model (FICM) of Fan, Yao, and Cai (2003):

$$Y_i = g\big(\beta' Z_i\big)' X_i + \varepsilon_i, \qquad (9.41)$$

where $i = 1, \ldots, n$, $X_i = \left(X_{i1}, \ldots, X_{ip}\right)'$ is a $p \times 1$ vector of covariates, Z_i is a $q \times 1$ vector of covariate, ε_i is an error term with mean zero and variance σ^2, $\beta = (\beta_1, \ldots, \beta_q)'$ is a $q \times 1$ vector of unknown parameters, and $g(\cdot) = (g_1(\cdot), \ldots, g_p(\cdot))'$ is a vector of p-dimensional unknown functional coefficients. Assume that $\|\beta\| = 1$ and the first element β_1 of β is positive for identification. The sparsity of the model may come from two aspects: Some of the functional index coefficients, $g_j(\cdot)$, $j = 1, \ldots, p$, are identically zero, and some elements of β are zero. Yang proposes a two-step approach to select the significant covariates with functional coefficients, and then variable selection is applied to choose local significant variables with parametric coefficients. The procedure goes as follows:

1. Given a $\sqrt{n}$-consistent initial estimator $\hat{\beta}^{(0)}$ of β (e.g., that of Fan, Yao, and Cai (2003)), we minimize the penalized local least squares to obtain the estimator $\hat{g}$ of $g : \hat{g} = \arg\min_{g} Q_h(g, \hat{\beta}^{(0)})$, where

$$Q_h(g, \beta) = \sum_{j=1}^{n} \sum_{i=1}^{n} \left[Y_i - g\left(\beta' Z_i\right)' X_i\right]^2 k_h\left(\beta' Z_i - \beta' Z_j\right) + n \sum_{l=1}^{p} p_{\lambda_l}\left(\|g_{l,\beta}\|\right),$$

   $k_h(z) = k(z/h)/h$, k is a kernel function, h is a bandwidth parameter, $g_{l,\beta} = (g_l\left(\beta' Z_1\right), \ldots, g_l\left(\beta' Z_n\right))'$, and $p_{\lambda_l}(\cdot)$ is the SCAD penalty function with tuning parameter λ_l.
2. Given the estimator $\hat{g}$ of g, we minimize the penalized global least squares objective function $Q\left(\beta, \hat{g}\right)$ to obtain an updated estimator $\hat{\beta}$ of β, where

$$Q\left(\beta, \hat{g}\right) = \frac{1}{2} \sum_{i=1}^{n} \left[Y_i - \hat{g}\left(\beta' Z_i\right)' X_i\right]^2 + n \sum_{k=1}^{q} p_{\lambda_n}(|\beta_k|).$$

   and $p_{\lambda_n}(\cdot)$ is the SCAD penalty function with tuning parameter λ_n.

Note that if one uses the Lasso penalty ($p_{\lambda_k}(a) = \lambda_k |a|$), the objective function in the first step becomes the penalized least squares criterion function used by Wang and Xia (2009). Yang (2012) proposes to choose the tuning parameters to minimize a BIC-type criterion function. Assuming that both p and q are fixed, he studies the consistency, sparsity, and the oracle property of the resulting functional index coefficient estimators $\hat{g}(\hat{\beta}' z)$ and $\hat{\beta}$. He applies his methodology to both financial and engineering data sets.

9.6.5. Generalized Single Index Models

Zhu and Zhu (2009) consider estimating the direction of β and selecting important variables simultaneously in the following generalized single-index model (GSIM) proposed by Li and Duan (1989) and Li (1991):

$$Y_i = G\big(X_i'\beta, \varepsilon_i\big), \tag{9.42}$$

where $G(\cdot)$ is an unknown link function, X_i is a $p \times 1$ vector of covariates, and ε_i is an error term that is independent of X_i. They allow $p = p_n$ to diverge as the sample size $n \to \infty$. The model in (9.42) is very general and covers the usual SIM and the heteroskedastic SIM (e.g., $Y = g_1\big(X'\beta\big) + g_2\big(X'\beta\big)\varepsilon$) as two special cases. Assume that $E(X_i) = 0$ and let $\Sigma = \text{Cov}(X_i)$.

Let $F\big(y\big)$ denote the cumulative distribution function (CDF) of the continuous response variable Y_i. Define

$$\beta^* = \underset{b}{\text{argmin}} E\big[l\big(b'X, F(Y)\big)\big],$$

where $l\big(b'X, F(Y)\big) = -F(Y)b'X + \psi\big(b'X\big)$ is a loss function and $\psi(\cdot)$ is a convex function. They show that under the sufficient recovery condition (which intuitively requires $E\big(X|\beta'X\big)$ to be linear in $\beta'X : :\ E\big(X|\beta'X\big) = \Sigma\beta\big(\beta'\Sigma\beta\big)^{-1}\beta'X$), β^* identifies β in model (9.42) up to a multiplicative scalar. The main requirement for such an identification is that $E\big[l\big(b'X, F(Y)\big)\big]$ has a proper minimizer. This condition relates to the unknown link function $G(\cdot)$ and is typically regarded as mild and thus widely assumed in the literature on SDR. To exclude the irrelevant regressors in the regression, Zhu and Zhu (2009) propose to estimate β as follows:

$$\hat{\beta} = \underset{b}{\text{argmin}} \frac{1}{2} \sum_{i=1}^{n} l\big(b'X_i, F_n(Y_i)\big) + n \sum_{j=1}^{p} p_{\lambda_n}\big(|b_j|\big)$$

where $F_n\big(y\big) = n^{-1} \sum_{i=1}^{n} 1\big(y_i \leq y\big)$ is the empirical distribution function (EDF) of Y_i, b_j is the jth coordinate of b, and $p_{\lambda_n}(\cdot)$ is a penalty function with tuning parameter λ_n.

The loss function $l\big(b'X, F(Y)\big)$ covers the least squares measure as a special case, that is, $l\big(b'X_i, F(Y_i)\big) = \big[b'X_i - F(Y_i)\big]^2/2$ by letting $\psi\big(b'X_i\big) = \big[b'X_iX_i'b + F^2(Y_i)\big]/2$. Then the least square estimation is

$$\beta_{LS}^* = \underset{b}{\text{argmin}} E\big[l\big(b'X_i, F(Y_i)\big)\big] = \underset{b}{\text{argmin}} E\big[F(Y_i) - X_i'b\big]^2 = \Sigma^{-1}\text{Cov}(X_i, F(Y_i)).$$

The sample version of the least squares estimate is given by

$$\hat{\beta}_{LS} = \underset{b}{\text{argmin}} \frac{1}{2} \sum_{i=1}^{n} \big[b'X_i - F_n(Y_i)\big]^2/2 + n \sum_{j=1}^{p} p_{\lambda_n}\big(|b_j|\big).$$

Zhu and Zhu (2009) suggest using the SCAD penalty. Under some regular conditions, they show that $\hat{\beta}_{LS}$ enjoys the oracle properties.

Zhu, Qian, and Lin (2011) follow the idea of Zhu and Zhu (2009) and propose a kernel-based method that automatically and simultaneously selects important predictors and estimates the direction of β in (9.42). As in Zhu and Zhu (2009), they also assume that $E(X_i) = 0$ and use Σ to denote $\mathrm{Cov}(X_i)$. The definition of the model in (9.42) is equivalent to saying that conditional on $\beta' X_i$, Y_i and X_i are independent. This implies the existence of a function $\tilde{f}$ such that

$$f(y|x) = \tilde{f}(y|\beta' x), \tag{9.43}$$

where $f(y|x)$ is the conditional probability density function (PDF) of Y_i given X_i and $\tilde{f}$ can be regarded as the conditional PDF of Y_i given $\beta' X_i$. Equation (9.43), together with the chain rule, implies that

$$\frac{\partial f(y|x)}{\partial x} = \beta \frac{\partial \tilde{f}(y|\beta' x)}{\partial(\beta' x)}. \tag{9.44}$$

That is, the first derivative of $f(y|x)$ with respect to x is proportional to β. This motivates Zhu, Qian, and Lin (2011) to identify the direction of β through the derivative of the conditional PDF $f(y|x)$.

Let $k_h(u) = k(u/h)/h$, where $k(\cdot)$ is a univariate kernel function and h is a bandwidth parameter. Note that $f(y|x) \approx E[k_h(Y-y)|X=x] = E[k_h(Y-y)|\beta' X = \beta' x] \approx \tilde{f}(y|\beta' x)$ as $h \to 0$. When X is Gaussian with mean zero and covariance matrix Σ, a direct application of Stein's lemma yields

$$H(y) = \Sigma^{-1} E[k_h(Y-y)X] \approx E\left[\frac{\partial f(y|X)}{\partial X}\right] \qquad \text{as } h \to 0.$$

When the normality assumption does not hold, Zhu, Qian, and Lin (2011) relax it to the widely assumed linearity condition as in the sufficient recovery condition and show that $H(y)$ and thus $E[H(Y)]$ are proportional to β for any fixed bandwidth h. Let

$$f^c(Y) = E\left[k_h\left(\tilde{Y} - Y\right)|Y\right] - E\left[k_h\left(\tilde{Y} - Y\right)\right], \tag{9.45}$$

where $\tilde{Y}$ is an independent copy of Y. They find that $E[H(Y)]$ is in spirit the solution to the following least squares minimization problem:

$$\beta_0 = E[H(Y_i)] = \underset{b}{\mathrm{argmin}} E\left[f^c(Y_i) - X_i' b\right]^2. \tag{9.46}$$

Note that the sample analogue of $f^c(Y)$ is given by

$$\hat{f}^c(Y) = \frac{1}{n}\sum_{i=1}^{n} k_h(Y_i - Y) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k_h(Y_i - Y_j).$$

Then one can obtain the unpenalized estimate of β_0 by

$$\hat{\beta}_0 = \underset{b}{\mathrm{argmin}} \sum_{i=1}^{n}\left[\hat{f}^c(Y_i) - X_i' b\right]^2.$$

The adaptive Lasso estimate of β_0 is given by

$$\hat{\beta}_{0,ALasso} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\hat{f}^c(Y_i) - X_i'b\right]^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j \left|b_j\right|,$$

where $\hat{w}_j = \left|\hat{\beta}_{0,j}\right|^{-\gamma}$ and $\hat{\beta}_{0,j}$ is the jth element of $\hat{\beta}_0$ for $j = 1, \ldots, p$. Assuming that p is fixed, Zhu, Qian, and Lin (2011) establish the oracle properties of $\hat{\beta}_{0,ALasso}$ under some regularity conditions.

Wang, Xu, and Zhu (2012) consider the variable selection and shrinkage estimation for several parametric and semiparametric models with the single-index structure by allowing $p = p_n$ to be divergent with n. Let $\delta = \mathrm{Cov}\big(X_i, g(Y_i)\big)$ for any function g. Define $\beta_g = \Sigma^{-1}\delta$. Under the assumption, $E\big(X|\beta'X\big)$ is linear in $\beta'X$, Theorem 2.1 in Li (1991) immediately implies that β_g is proportional to β; that is, $\beta_g = \kappa_g \beta$ for some constant κ_g. The least squares index estimate of β_g is given by

$$\hat{\beta}_g = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \left[g(Y_i) - X_i'b\right]^2.$$

They propose a response-distribution transformation by replacing g by the CDF $F\big(y\big)$ of Y minus 1/2. Since F is unknown in practice, they suggest using its EDF F_n and define the distribution-transformation least squares estimator as

$$\hat{\beta}_{F_n} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \left[F_n(Y_i) - \frac{1}{2} - X_i'b\right]^2.$$

The penalized version is given by

$$\hat{\beta}_{F_n} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \left[F_n(Y_i) - \frac{1}{2} - X_i'b\right]^2 + \sum_{j=1}^{p} p_{\lambda_n}\big(|\beta_j|\big),$$

where $p_{\lambda_n}(\cdot)$ can be the SCAD penalty or the MC penalty of Zhang (2010). They establish the selection consistency by allowing $p = p_n$ to grow at any polynomial rate under some moment conditions for X_i. If X_i's are normally distributed, it also allows p_n to grow exponentially fast.

9.7. Variable/Component Selection in General Nonparametric Models

In the previous four sections we reviewed variable selection in semiparametric and nonparametric regression models that impose certain structures to alleviate the notorious “curse of dimensionality” problem in the literature. In this section we review variable selection in general nonparametric models that do not assume these structures. Even so, we remark that it is frequently assumed that certain decomposition of the general nonparametric regression functions exists, in which case the latter also exhibits a specific additive structure.

The literature on variable or component selection in general nonparametric models can be classified into two categories. The first category is carried out in the framework of *Smoothing Spline ANalysis Of VAriance* (SS-ANOVA) or *global* function approximation (see, e.g., Lin and Zhang (2006), Bunea (2008), Storlie, Bondell, Reich, and Zhang (2011), and Comminges and Dalayan (2011)). Lin and Zhang (2006) propose a new method called *COmponent Selection and Smoothing Operator* (COSSO) for model selection and model fitting in multivariate nonparametric regression models, in the framework of *smoothing spline ANOVA* (SS-ANOVA). As Huang, Breheny, and Ma (2012) remark, the COSSO can be viewed as a group Lasso procedure in a reproducing kernel Hilbert space. Storlie, Bondell, Reich, and Zhang (2011) propose the *adaptive COSSO* (ACOSSO) to improve the performance of COSSO. Bunea (2008) investigates the consistency of selection via the Lasso method in regression models, where the regression function is approximated by a given dictionary of M functions. Comminges and Dalayan (2011) consider consistent variable selection in high-dimensional nonparametric regression based on an orthogonal Fourier expansion of the regression function. The second category focuses on *local* selection of significant variables. Bertin and Lecué (2008) implement a two-step procedure to reduce the dimensionality of a local estimate. Lafferty and Wasserman (2008) introduce the Rodeo procedure, which attempts to assign adaptive bandwidths based on the derivative of kernel estimate with respect to the bandwidth for each dimension. Miller and Hall (2010) propose a method called LABAVS in local polynomial regression to select the variables and estimate the model.

9.7.1. Lin and Zhang's (2006) COSSO

Lin and Zhang (2006) consider the nonparametric regression

$$Y_i = f(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{9.47}$$

where f is the regression function to be estimated, $X_i = (X_{i1}, \ldots, X_{ip})' \in \mathcal{X} = [0,1]^p$ are p-dimensional vectors of covariates, and ε_i is independent noise with mean zero and finite variance σ^2. In the framework of SS-ANOVA, f exhibits the decomposition

$$f(x) = b + \sum_{j=1}^{p} f_j(x_j) + \sum_{1 \le j < k \le p} f_{jk}(x_j, x_k) + \cdots, \tag{9.48}$$

where $x = (x_1, \ldots, x_p)'$, b is a constant, f_j's are the main effects, f_{jk} are the two-way interactions, and so on. The sequence is usually truncated somewhere to enhance interpretability. One can assure the identifiability of the terms in (9.48) by some side conditions through averaging operators.

Let $\mathcal{F}$ be the *reproducing kernel Hilbert space* (RKHS) corresponding to the decomposition in (9.48). For the definition of RKHS, see Wahba (1990). Frequently, $\mathcal{F}$ is a space of functions with a certain degree of smoothness—for example, the second-order Sobolev space, $\mathcal{S}^2 = \{g : g, g' \text{ are absolutely continuous and } g'' \in L^2[0,1]\}$. Let $\mathcal{H}_j$ be a function space of functions of x_j over $[0,1]$ such that $\mathcal{H}_j = \{1\} \oplus \bar{\mathcal{H}}_j$. Then $\mathcal{F}$ is the tensor product space of $\mathcal{H}_j$,

$$\mathcal{F} = \otimes_{j=1}^{p} \mathcal{H}_j = \{1\} \oplus \sum_{j=1}^{p} \bar{\mathcal{H}}_j \oplus \sum_{j<k} \left(\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k \right) \oplus \cdots . \tag{9.49}$$

Each functional component in the SS-ANOVA decomposition (9.48) lies in a subspace in the orthogonal decomposition (9.49) of $\otimes_{j=1}^{p} \mathcal{H}_j$. But in practice the higher-order interactions are usually truncated for convenience to avoid the curse of dimensionality. In the simplest case where $f(x) = b + \sum_{j=1}^{p} f_j(x_j)$ with $f_j \in \bar{\mathcal{H}}_j$, the selection of functional components is equivalent to variable selection. In the general SS-ANOVA models, model selection amounts to the selection of main effects and interaction terms in the SS-ANOVA decomposition. A general expression for the truncated space can be written as

$$\mathcal{F} = \{1\} \otimes \left\{ \otimes_{j=1}^{q} \mathcal{F}^j \right\}, \tag{9.50}$$

where $\mathcal{F}^1, \ldots, \mathcal{F}^q$ are q orthogonal subspaces of $\mathcal{F}$. $q = p$ gives the special case of additive models. When only main effects and two-way interaction effects are retained, the truncated space has $q = p(p+1)/2$, which includes p main effect spaces and $p(p-1)/2$ two-way interaction spaces.

Denote the norm in the RKHS $\mathcal{F}$ by $\|\cdot\|$. A traditional smoothing spline-type method finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \left[Y_i - f(X_i) \right]^2 + \lambda \sum_{j=1}^{q} \theta_j^{-1} \left\| P^j f \right\|^2, \tag{9.51}$$

where $P^j f$ is the orthogonal projection of f onto $\mathcal{F}^j$ and $\theta_j \geq 0$. If $\theta_j = 0$, the minimizer is taken to satisfy $\left\| P^j f \right\|^2 = 0$. The COSSO procedure finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \left[Y_i - f(X_i) \right]^2 + \tau_n^2 J(f) \qquad \text{with } J(f) = \sum_{j=1}^{q} \left\| P^j f \right\|, \tag{9.52}$$

where τ_n is a smoothing parameter. The penalty term $J(f)$ is a sum of RKHS norms, instead of the squared RKHS norm penalty employed in the smoothing spline. $J(f)$ is a convex functional, which ensures the existence of the COSSO estimate. Let $\hat{f} = \hat{b} + \sum_{j=1}^{q} \hat{f}_j$ be a minimizer of (9.52).

Lin and Zhang (2006) form $\mathcal{F}$ using $\mathcal{S}^2$ with squared norm

$$\|g\|^2 = \left(\int_0^1 g(u)du\right)^2 + \left(\int_0^1 g'(u)du\right)^2 + \left(\int_0^1 g''(u)du\right)^2 \tag{9.53}$$

for each of the $\mathcal{H}_j$ in (9.49). They show that an equivalent expression of (9.52) is

$$\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - f(X_i)\right]^2 + \lambda_0 \sum_{j=1}^{q} \theta_j^{-1}\left\|P^j f\right\|^2 + \lambda \sum_{j=1}^{q}\theta_j \qquad \text{subject to } \theta_j \geq 0, \tag{9.54}$$

for $j = 1, \ldots, p$, where λ_0 is a constant and λ is a smoothing parameter. The constant λ_0 can be fixed at any positive value. For fixed θ, the COSSO (9.54) is equivalent to the smoothing spline (9.51). From the smoothing spline literature, it is well known that the solution f has the form

$$f(x) = \sum_{i=1}^{n} c_i R_\theta(X_i, x) + b,$$

where $c = (c_1, \ldots, c_n)'$ and $R_\theta = \sum_{j=1}^{q} \theta_j R_j$ with R_j being the reproducing kernel of $\mathcal{F}_j$ (the $n \times n$ matrix $\left\{R_j(X_i, X_k)\right\}_{i,k=1}^{n}$). Then $f(x) = R_\theta c + b\mathbf{1}_n$, where $\mathbf{1}_n$ is an $n \times 1$ vector of ones. The problem (9.52) can be written as

$$\frac{1}{n}\left(Y - \sum_{j=1}^{q}\theta_j R_j c - b\mathbf{1}_n\right)'\left(Y - \sum_{j=1}^{q}\theta_j R_j c - b\mathbf{1}_n\right) + \lambda_0 \sum_{j=1}^{q}\theta_j c' R_j c + \lambda \sum_{j=1}^{q}\theta_j, \tag{9.55}$$

where $Y = (Y_1, \ldots, Y_n)'$, and $\theta_j \geq 0$ for $j = 1, \ldots, q$. For fixed θ, (9.55) can be written as

$$\min_{c,b}\left(y - R_\theta c - b\mathbf{1}_n\right)'\left(y - R_\theta c - b\mathbf{1}_n\right) + n\lambda_0 c' R_\theta c.$$

Then c and b can be solved as in Wahba (1990). On the other hand, if c and b are fixed, let $g_j = R_j c$ and G be the matrix with the jth column being g_j. θ that minimizes (9.55) is the solution to

$$\min_\theta (z - G\theta)'(z - G\theta) + n\lambda \sum_{j=1}^{q}\theta_j \qquad \text{subject to } \theta_j \geq 0 \text{ for } j = 1, .., q$$

or

$$\min_\theta (z - G\theta)'(z - G\theta), \qquad \text{subject to } \theta_j \geq 0, j = 1, \ldots, q, \text{ and } \sum_{j=1}^{q}\theta_j \leq M,$$

where $z = y - (1/2)n\lambda_0 c - b\mathbf{1}_n$ and M is a positive constant. The tuning parameter can be chosen by 5-fold or 10-fold cross-validation. Lin and Zhang (2006) study the theoretical properties such as the existence and rate of convergence of the COSSO estimator.

In the framework of SS-ANOVA, Zhang and Lin (2006) study the component selection and smoothing for nonparametric regression in the more general setting of exponential family regression, and Leng and Zhang (2006) study the same issue for a nonparametric extension of the Cox proportional hazard model. The former allows the treatment of non-normal responses, binary and polychotomous responses, and event counts data. The latter demonstrates great flexibility and easy interpretability in modeling relative risk functions for censored data.

9.7.2. Storlie, Bondell, Reich, and Zhang's (2011) ACOSSO

The oracle properties used before are mainly defined for the finite-dimensional parameter in parametric or semiparametric models. In the context of nonparametric regression, Storlie, Bondell, Reich, and Zhang (2011) extend this notion by saying that a nonparametric regression estimator has the nonparametric *weak (np)-oracle property* if it (a) selects the correct subset of predictors with probability tending to one and (b) estimates the regression function at the optimal nonparametric rate. Note that the *strong* version of the oracle property requires that the estimator should have the asymptotic distribution as the oracle one. The SS-ANOVA-based COSSO procedure has not been demonstrated to possess the weak np-oracle property. Instead, it has a tendency to oversmooth the nonzero functional components in order to set the unimportant functional components to zero. Storlie, Bondell, Reich, and Zhang (2011) propose the adaptive COSSO (ACOSSO) which possesses the weak np-oracle properties.

Like Lin and Zhang (2006), Storlie, Bondell, Reich, and Zhang (2011) consider the nonparametric regression model in (9.47), where $X_i = (X_{i1}, \ldots, X_{ip})' \in \mathcal{X} = [0,1]^p$ and ε_i's are independent of X_i and are uniformly sub-Gaussian with zero mean. They obtain their estimate of the function $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - f(x_i)\right]^2 + \lambda \sum_{j=1}^{p} w_j \left\| P^j f \right\|, \tag{9.56}$$

where $0 < w_j \leq \infty$ are weights that can depend on an initial estimate of f, for example, the COSSO estimate $\tilde{f}$. They suggest the choice

$$w_j = \left\| P^j \tilde{f} \right\|_{L_2}^{-\gamma} \qquad \text{for } j = 1, \ldots, p,$$

where $\left\| P^j \tilde{f} \right\|_{L_2} = \{\int_{\mathcal{X}} [P^j \tilde{f}(x)]^2 dx\}^{1/2}$ and $\gamma > 0$. The tuning parameter is also chosen via 5-fold or 10-fold cross-validation. Under some regular conditions, they show that their estimator possesses the weak np-oracle property when $f \in \mathcal{F}$ is additive in the predictors so that $\mathcal{F} = \{1\} \oplus \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_p$, where each $\mathcal{F}_j$ is a space of functions corresponding to x_j.

9.7.3. Bunea's (2008) Consistent Selection via the Lasso

Bunea (2008) considers the approximation of the regression function f in (9.47) with elements of a given dictionary of M functions. Let

$$\Lambda = \left\{ \lambda \in \mathbb{R}^M : \left\| f - \sum_{j=1}^{M} \lambda_j f_j \right\|^2 \leq C_f r_{n,M}^2 \right\},$$

where $C_f > 0$ is a constant depending only on f and $r_{n,M}$ is a positive sequence that converges to zero. For any $\lambda = (\lambda_1, \ldots, \lambda_M)' \in \mathbb{R}^M$, let $J(\lambda)$ denote the index set corresponding to the nonzero components of λ and $M(\lambda)$ its cardinality. Let $p^* = \min\{M(\lambda) : \lambda \in \Lambda\}$. Define

$$\lambda^* = \underset{\lambda \in \mathbb{R}^M}{\arg\min} \left\{ \left\| f - \sum_{j=1}^{M} \lambda_j f_j \right\|^2 : M(\lambda) = p^* \right\}.$$

Let $I^* = J(\lambda^*)$ denote the index set corresponding to the nonzero elements of λ^*. Note that the cardinality of I^* is given by p^* and thus $f^* = \sum_{j \in I^*} \lambda_j^* f_j$ provides the sparest approximation to f that can be realized with $\lambda \in \Lambda$ and $\left\| f^* - f \right\|^2 \leq C_f r_{n,M}^2$. This motivates Bunea to treat I^* as the target index set.

Bunea considers estimating the set I^* via the l_1-penalized least squares. First, he computes

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}^M}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \lambda_j f_j(X_i) \right]^2 + 2 \sum_{j=1}^{M} w_{nj} |\lambda_j| \right\},$$

where $w_{nj} = r_{n,M} \left\| f_j \right\|_n$ and $\left\| f_j \right\|_n^2 = n^{-1} \sum_{i=1}^{n} \left[f_j(X_i) \right]^2$. Let $\hat{I}$ denote the index set corresponding to the nonzero components of $\hat{\lambda}$. He shows that $P\left(\hat{I} = I^* \right) \to 1$ as $n \to \infty$ under some conditions in conjunction with the requirement that $p^* r_{n,M} \to 0$.

9.7.4. Comminges and Dalayan's (2011) Consistent Variable Selection in High-Dimensional Nonparametric Regression

Comminges and Dalayan (2011) consider the general nonparametric regression model (9.47) where X_i's are assumed to take values in $[0,1]^p$, $E(\varepsilon_i | X_i) = 0$, $E\left(\varepsilon_i^2 | X_i \right) = \sigma^2$, and $p = p_n$ may diverge to the infinity with n. They assume that f is differentiable with a squared integrable gradient and that the density function $g(x)$ of X_i exists and is bounded away from 0 from below. Define the Fourier basis

$$\varphi_{\mathbf{k}}(x) = \begin{cases} 1 & \text{if } \mathbf{k} = \mathbf{0}, \\ \sqrt{2}\cos\left(2\pi \mathbf{k}'x\right) & \text{if } \mathbf{k} \in \left(\mathbb{Z}^p\right)_+, \\ \sqrt{2}\sin\left(2\pi \mathbf{k}'x\right) & \text{if } -\mathbf{k} \in \left(\mathbb{Z}^p\right)_+, \end{cases}$$

where $\left(\mathbb{Z}^p\right)_+$ denotes the set of all $\mathbf{k} = \left(k_1, \ldots, k_p\right)' \in \mathbb{Z}^p \backslash \{0\}$ such that the first nonzero element of $\mathbf{k}$ is positive. Let

$$\Sigma_L = \left\{ f : \sum_{\mathbf{k} \in \mathbb{Z}^p} k_j < f, \varphi_{\mathbf{k}} >^2 \leq L \, \forall j \in \left\{1, \ldots, p\right\} \right\},$$

where $<\cdot,\cdot>$ stands for the scalar product in $L^2\left([0,1]^p; \mathbb{R}\right)$, that is, $<a,b>$ $\int_{[0,1]^p} a(x)b(x)\,dx$ for any $a, b \in L^2\left([0,1]^p; \mathbb{R}\right)$. Comminges and Dalayan (2011) assume that the regression function f belongs to Σ_L and for some $J \subset \left\{1, \ldots, p\right\}$ of cardinality $p^* \leq p$, $f(x) = \bar{f}\left(x_J\right)$ for some $\bar{f} : \mathbb{R}^{|J|} \to \mathbb{R}$, and it holds that

$$Q_j\left[f\right] \equiv \sum_{\mathbf{k}: k_j \neq 0} \theta_{\mathbf{k}}\left[f\right]^2 \geq \kappa, \qquad \forall j \in J,$$

where $\theta_{\mathbf{k}}\left[f\right] = <f, \varphi_{\mathbf{k}}>$. Clearly, J refers to the sparsity pattern of f and $Q_j\left[f\right] = 0$ if $j \notin J$.

The Fourier coefficients $\theta_{\mathbf{k}}\left[f\right]$ can be estimated by their empirical counterparts

$$\hat{\theta}_{\mathbf{k}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\varphi_{\mathbf{k}}(X_i)}{g(X_i)} Y_i, \qquad \mathbf{k} \in \mathbb{Z}^p.$$

Let $S_{m,l} = \{\mathbf{k} \in \mathbb{Z}^p : \|\mathbf{k}\|_2 \leq m, \|\mathbf{k}\|_0 \leq l\}$ and $N\left(p^*, \gamma\right) = \text{Card}\{\mathbf{k} \in \mathbb{Z}^{p^*} : \|\mathbf{k}\|_2^2 \leq \gamma p^*, k_1 \neq 0\}$, where $l \in \mathbb{N}$ and $\gamma > 0$. Note that if $j \notin J$, then $\theta_{\mathbf{k}}\left[f\right] = 0$ for every $\mathbf{k}$ such that $k_j \neq 0$; and if $j \in J$, then there exists $\mathbf{k} \in \mathbb{Z}^p$ with $k_j \neq 0$ such that $\left|\theta_{\mathbf{k}}\left[f\right]\right| > 0$. Comminges and Dalayan define their estimator of J by

$$\hat{J}_n(m, \lambda) = \left\{ j \in \left\{1, \ldots, p\right\} : \max_{\mathbf{k} \in S_{m,p^*} : k_j \neq 0} \left|\hat{\theta}_{\mathbf{k}}\right| > \lambda \right\}.$$

They show that $P\left(\hat{J}_n(m, \lambda) \neq J\right) \leq 3\left(6mp\right)^{-p^*}$ under some regularity conditions related to $N\left(p^*, \gamma\right)$ and $\left(p, p^*, n\right)$. It is possible for p^* to either be fixed or tend to the infinity as $n \to \infty$. Unfortunately, Comminges and Dalayan (2011) deliberately avoid any discussion on the computational aspects of the variable selection and focus exclusively on the consistency of variable selection without paying any attention to the consistency of regression function estimation. Two problems have to be addressed in order to implement their procedure, namely, the estimate of the typically unknown density function g and the determination of p^*.

9.7.5. Bertin and Lecué (2008)

Bertin and Lecué (2008) consider the nonparametric regression model (9.47) where ε_i's are i.i.d. Gaussian random variables with variance σ^2 and independent of X_i's, and f is the unknown regression function. Suppose the nonparametric regression function f satisfies a sparseness condition:

$$f(x) = \bar{f}(x_{\mathbf{R}}), \tag{9.57}$$

where $x_{\mathbf{R}} = \left(x_j : j \in \mathbf{R}\right)$, $\mathbf{R} \subset \left\{1,\ldots,p\right\}$ is a subset of p covariates, of size $p^* = |\mathbf{R}| < p$. Obviously, $x_{\mathbf{R}}$ denotes the set of relevant variables. They are interested in the pointwise estimation of f at a fixed point $x = \left(x_1,\ldots,x_p\right)'$ and the construction of some estimate $\hat{f}_n$ having the smallest pointwise integrated quadratic risk $E[\hat{f}_n(x) - f(x)]^2$.

Assume f to be β-Holderian around x with $\beta > 0$, denoted by $f \in \Sigma(\beta, x)$. A function $f : \mathbb{R}^p \to \mathbb{R}$ is β-Holderian at point x with $\beta > 0$ if (i) f is l-times differentiable in x ($l = \lfloor \beta \rfloor$) and (ii) there exists $L > 0$ such that for any $t = (t_1,\ldots,t_n) \in B_\infty(x,1)$ (the unit l_∞-ball of center x and radius 1), $\left|f(t) - P_l\left(f\right)(t,x)\right| \le L\|t - x\|_1^\beta$, where $P_l\left(f\right)(\cdot,x)$ is the Taylor polynomial of order l associated with f at point x. Assume that there exists a subset of $J = \left\{i_1,\ldots,i_{p^*}\right\} \subset \left\{1,\ldots,p\right\}$ such that

$$f\left(x_1,\ldots,x_p\right) = \bar{f}(x_{i_1},\ldots,x_{i_{p^*}}).$$

That is, the "real" dimension of the model is not p but p^*. Bertin and Lecué's goal is twofold: (i) Determine the set of indices $J = \left\{i_1,\ldots,i_{p^*}\right\}$, and (ii) construct an estimator of $f(x)$ that converges at rate $n^{-2\beta/(2\beta+p^*)}$, which is the fastest convergence rate when $f \in \Sigma(\beta, x)$ and the above sparsity condition is satisfied.

To determine the set of indices, based on the principle of local linear regression, they consider the following set of vectors:

$$\bar{\Theta}_1(\lambda) = \underset{\theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \frac{1}{nh^p} \sum_{i=1}^{n} \left[Y_i - U\left(\frac{X_i - x}{h}\right)' \theta \right]^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right\},$$

where $U(v) = \left(1, v_1,\ldots,v_p\right)'$ for any $v = \left(v_1,\ldots,v_p\right)'$, $\theta = \left(\theta_0,\theta_1,\ldots,\theta_p\right)$, $\|\theta\|_1 = \sum_{j=0}^{p}\left|\theta_j\right|$, h is a bandwidth, and $K(\cdot)$ is a symmetric kernel function. The l_1 penalty makes the solution vector $\bar{\Theta}(\lambda)$ sparse and then selects the variables locally. Another selection procedure, which is close to the previous one but requires less assumption on the regression function, is given by

$$\bar{\Theta}_2(\lambda) = \underset{\theta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \times \left\{ \frac{1}{nh^p} \sum_{i=1}^{n} \left[Y_i + f_{\max} + Ch - U\left(\frac{X_i - x}{h}\right)' \theta \right]^2 K\left(\frac{X_i - x}{h}\right) + 2\lambda \|\theta\|_1 \right\},$$

where $C > 0$ is a constant, $f_{\max} > 0$, and $|f(x)| \leq f_{\max}$. Here, the response variable Y_i is translated by $f_{\max} + Ch$.

Let $\hat{J}_1$ and $\hat{J}_2$ be the subset of indices selected by the above procedures for a given λ. Based on these sets of indices, Bertin and Lecué (2008) consider a local polynomial regression of degree $l = \lfloor \beta \rfloor$ by regressing Y_i on the selected variables. Under different conditions on function f, they show that $\hat{J}_1 = J$ or $\hat{J}_2 = J$ with a probability approaching 1, and for $\hat{J}_2$ the local polynomial estimate in the second step can achieve the fastest convergence rate under some conditions. The selection is proved to be consistent when $p^* = O(1)$, but p is allowed to be as large as $\log n$, up to a multiplicative constant.

9.7.6. Lafferty and Wasserman's (2008) Rodeo Procedure

Lafferty and Wasserman (2008) presented a greedy method for simultaneously preforming local bandwidth selection and variable selection in the nonparametric regression model (9.47), where ε_i's are i.i.d. Gaussian random variables with zero mean and variance σ^2. Suppose the nonparametric regression function f satisfies a sparseness condition in (9.57) with $p^* = |\mathbf{R}| \ll p$. Without loss of generality, we assume that $x_{\mathbf{R}} = \left(x_1, \ldots, x_{p^*}\right)'$ so that the last $p - p^*$ elements in x are irrelevant. Based on the idea that bandwidth and variable selection can be simultaneously performed by computing the infinitesimal change in a nonparametric estimator as a function of smoothing parameters, Lafferty and Wasserman (2008) propose the general framework for the *regularization of derivative expectation operator* (Rodeo).

The key idea is as follows. Fix a point x and let $\hat{f}(x)$ be an estimator of $f(x)$ based on a vector of smoothing parameters $h = (h_1, \ldots, h_p)'$. Let $F(h) = E[\hat{f}_h(x)]$. Assume that $x = X_i$ is one of the observed data points and $\hat{f}_0(x) = Y_i$. In this case, $f(x) = F(0) = E(Y_i)$. If $P = (h(t) : 0 \leq t \leq 1)$ is a smooth path through the set of smoothing parameters with $h(0) = 0$ and $h(1) = 1$ (or any other fixed large bandwidth), then

$$\begin{aligned} f(x) &= F(0) = F(1) + F(0) - F(1) \\ &= F(1) - \int_0^1 \frac{dF(h(s))}{ds} ds \\ &= F(1) - \int_0^1 D(h(s))' \dot{h}(s) ds, \end{aligned}$$

where $D(h(s)) = \nabla F(h) = \left(\frac{\partial F}{\partial h_1}, \ldots, \frac{\partial F}{\partial h_p}\right)'$ and $\dot{h}(s) = \frac{dh(s)}{ds}$. Noting that an unbiased estimator of $F(1)$ is $\hat{f}_1(x)$, an unbiased estimator of $D(h)$ is

$$Z(h) = \left(\frac{\partial \hat{f}_h(x)}{\partial h_1}, \ldots, \frac{\partial \hat{f}_h(x)}{\partial h_p}\right)'.$$

The naive estimator

$$\hat{f}(x) = \hat{f}_1(x) - \int_0^1 Z(h(s))'\dot{h}(s)ds$$

is equal to $\hat{f}_0(x) = Y_i$, which is a poor estimator because of the large variance of $Z(h)$ for small h. Nevertheless, the sparsity assumption on f suggests that $D(h)$ is also sparse for some paths. Then using an estimator $\hat{D}(h)$ which uses the sparsity assumption yields the following estimate of $f(x)$:

$$\tilde{f}(x) = \hat{f}_1(x) - \int_0^1 \hat{D}(h(s))'\dot{h}(s)ds.$$

The implementation of such an estimator requires us to find a path for which the derivative $D(h)$ is also sparse, and then take advantage of this sparseness when estimating $D(h)$ along that path.

A key observation is that if x_j is irrelevant in x, then changing the bandwidth h_j should cause only a small change in $\hat{f}_h(x)$. Conversely, if x_j is relevant in x, then changing h_j should cause a large change in $\hat{f}_h(x)$. Thus $Z_j(h) = \partial\hat{f}_h(x)/\partial h_j$ should discriminate between relevant and irrelevant covariates. Let $h_j \in \mathcal{H} = \{h_0, \beta h_0, \beta^2 h_0, \ldots\}$ for some $\beta \in (0,1)$. A greedy version of estimator of $D_j(h)$, the jth element of $D(h)$, would set $\hat{D}_j(h) = 0$ when $h_j < \hat{h}_j$, where $\hat{h}_j$ is the first h such that $|Z_j(h)| < \lambda_j(h)$ for some threshold λ_j where $h = a$ for a scalar a means $h = (a, \ldots, a)'$, a $p \times 1$ vector. That is,

$$\hat{D}_j(h) = Z_j(h)1\left(|Z_j(h)| > \lambda_j(h)\right).$$

This greedy version, coupled with the hard threshold estimator, yields $\tilde{f}(x) = \hat{f}_{\hat{h}}(x)$, where $\hat{h} = (\hat{h}_1, \ldots, \hat{h}_p)'$. This is a bandwidth selection procedure based on testing.

For local linear regression, Lafferty and Wasserman (2008) give explicit expressions for $Z(h)$. The local linear estimator of $f(x)$ by using kernel K and bandwidth $h = (h_1, \ldots, h_p)'$ is given by

$$\hat{f}_h(x) = \sum_{i=1}^{n} G(X_i, x, h)Y_i,$$

where $G(u,x,h) = e_1'\left(X_x'W_xX_x\right)^{-1}\begin{pmatrix} 1 \\ u-x \end{pmatrix}K_h(u-x)$ is the effective kernel, $e_1 = (1,0,\ldots,0)'$, $K_h(u) = \left(h_1 \ldots h_p\right)^{-1}K\left(u_1/h_1, \ldots, u_p/h_p\right)$, X_x is an $n \times \left(p+1\right)$ matrix whose ith row is given by $(1, (X_i - x)')$, and W_x is a diagonal matrix with (i,i)-element $K_h(X_i - x)$. In this case,

$$Z_j(h) = \frac{\partial\hat{f}_h(x)}{\partial h_j} = \sum_{i=1}^{n} \frac{\partial G(X_i, x, h)}{\partial h_j} Y_i.$$

Lafferty and Wasserman derive the explicit expression for $\partial G(X_i, x, h)/\partial h_j$ and $Z_j(h)$. Let

$$s_j = \mathrm{Var}\big(Z_j(h)|X_1,\ldots,X_n\big) = \sigma^2 \sum_{i=1}^{n} \left(\frac{\partial G(X_i,x,h)}{\partial h_j}\right)^2.$$

They illustrate how to perform the Rodeo via the hard thresholding as follows:

1. Select a constant $\beta \in (0,1)$ and the initial bandwidth $h_0 = c/\log\log n$.
2. Initialize the bandwidths, and activate all covariates: (a) $h_j = h_0$ for $j = 1,\ldots,p$ and (b) $\mathcal{A} = \{1,2,\ldots,p\}$.
3. While $\mathcal{A}$ is nonempty, for each $j \in \mathcal{A}$: (a) Compute Z_j and s_j; (b) compute threshold value $\lambda_j = s_j\sqrt{2\log n}$; and (c) if $|Z_j| > \lambda_j$, reassign select βh_j to h_j; otherwise remove j from $\mathcal{A}$.
4. Output the bandwidth $h^* = (h_1,\ldots,h_p)$ and estimator $\tilde{f}(x) = \hat{f}_{h^*}(x)$.

Under some conditions, the Rodeo outputs bandwidths h^* that satisfies $P(h_j^* = h_0$ for all $j > p^*) \to 1$ where recall X_{ij}'s are irrelevant variables for all $j > p^*$. In particular, the Rodeo selection is consistent when $p = O(\log n/\log\log n)$ and its estimator achieves the near-optimal minimax rate of convergence while p^* does not increase with n. Lafferty and Wasserman explain how to (a) estimate σ^2 used in the definition of s_j, and (b) obtain other estimators of $D(h)$ based on the soft thresholding.

9.7.7. Miller and Hall's (2010) LABAVS in Local Polynomial Regression

Miller and Hall (2010) propose a flexible and adaptive approach to local variable selection using local polynomial regression. The key technique is careful adjustment of the local regression bandwidths to allow for variable redundancy. They refer to their method as LABAVS, standing for "*locally adaptive bandwidth and variable selection.*" The model is as given in (9.47). They consider the local polynomial estimation of f at a fixed point x. Let $H = \mathrm{diag}\left(h_1^2,\ldots,h_p^2\right)$. Let $K(x) = \Pi_{j=1}^{p} k(x_j)$ be the p-dimensional rectangular kernel formed from a univariate kernel k with support on $[-1,1]$ such as the tricubic kernel: $k(v) = (35/32)(1-v^2)^3 1(|v| \le 1)$. Let $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$. We write $H(x)$ when H varies as a function of x. Asymmetric bandwidths are defined as having a lower and an upper diagonal bandwidth matrix, H^L and H^U, respectively, for a given estimation point x. The kernel weight of an observation X_i at an estimation point x with asymmetrical local bandwidth matrices $H^L(x)$ and $H^U(x)$ is given by

$$K_{H^L(x),H^U(x)}(X_i - x) = \prod_{j:X_{ij}<x_j} h_j^L(x)^{-1} k\left(\frac{X_{ij}-x_j}{h_j^L(x)}\right) \times \prod_{j:X_{ij}\ge x_j} h_j^U(x)^{-1} k\left(\frac{X_{ij}-x_j}{h_j^U(x)}\right),$$

which amounts to having possibly different window sizes above and below x in each direction.

Miller and Hall's LABAVS algorithm works as follows:

1. Find an initial bandwidth matrix $H = \operatorname{diag}\left(h^2, \ldots, h^2\right)$.
2. For each point x of a representative grid in the data support, perform local variable selection to determine disjoint index sets $\hat{A}^+(x)$ and $\hat{A}^-(x)$ for variables that are considered relevant and redundant, respectively. Note that $\hat{A}^+(x) \cup \hat{A}^-(x) = \left\{1, \ldots, p\right\}$.
3. For any given x, derive new local bandwidth matrices $H^L(x)$ and $H^U(x)$ by extending the bandwidth in each dimension indexed in $\hat{A}^-(x)$. The resulting space given nonzero weight by the kernel $K_{H^L(x), H^U(x)}(u - x)$ is the rectangle of maximal area with all grid points x_0 inside the rectangle satisfying $\hat{A}^+(x_0) \subset \hat{A}^+(x)$, where $\hat{A}^+(x)$ is calculated explicitly as in step 2, or is taken as the set corresponding the closet grid point to x.
4. Shrink the bandwidth slightly for those variables in $\hat{A}^+(x)$ according to the amount that bandwidths have increased in the other variables.
5. Compute the local polynomial estimate at x, excluding variables in $\hat{A}^-(x)$ and using adjusted asymmetrical bandwidths $H^L(x)$ and $H^U(x)$. For example, in the local linear regression case, one chooses a and b to minimize

$$\sum_{i=1}^{n} \left[Y_i - a - b'(X_i - x)\right]^2 K_{H^L(x), H^U(x)}(X_i - x).$$

Steps 2 and 4 of the above algorithm are referred to as the variable selection step and variable shrinkage step, respectively. Miller and Hall suggest three possible ways to select variables at x in step 2, namely, hard thresholding, backwards stepwise approach, and local Lasso. Let

$$\bar{X}_{j,x} = \frac{\sum_{i=1}^{n} X_{ij} K_{H(x)}(X_i - x)}{\sum_{i=1}^{n} K_{H(x)}(X_i - x)} \quad \text{and} \quad \bar{Y}_x = \frac{\sum_{i=1}^{n} Y_i K_{H(x)}(X_i - x)}{\sum_{i=1}^{n} K_{H(x)}(X_i - x)},$$

which are the local standardization of the data at point x. Let $\tilde{Y}_i = \left(Y_i - \bar{Y}_x\right) \left[K_{H(x)}(X_i - x)\right]^{1/2}$ and $\tilde{X}_{ij} = \frac{(X_{ij} - \bar{X}_{j,x})\left[K_{H(x)}(X_i - x)\right]^{1/2}}{\left[\sum_{i=1}^{n}(X_{ij} - \bar{X}_{j,x})^2 K_{H(x)}(X_i - x)\right]^{1/2}}$. In the local linear regression case, the hard thresholding method chooses parameters to minimize the weighted least squares

$$\sum_{i=1}^{n} \left[\tilde{Y}_i - \beta_0 - \sum_{j=1}^{p} \beta_j \tilde{X}_{ij}\right]^2$$

and classifies as *redundant* the variables for which $\left|\hat{\beta}_j\right| < \lambda$ for some tuning parameter λ, where $(\hat{\beta}_0, \ldots, \hat{\beta}_p)$ is the solution to the above minimization problem. The variable shrinkage step and step 3 are fairly complicated and computationally demanding. We refer the readers directly to Miller and Hall (2010), who also compare their approach with other local variable selection approaches in Bertin and Lecué (2008) and Lafferty and Wasserman (2008). They establish the strong oracle property for their estimator.

9.8. Variable Selection in Semiparametric/Nonparametric Quantile Regression

As a generalization of least absolute deviation regression (LADR), quantile regression (QR) has attracted huge interest in the literature and has been widely used in economics and finance; see Koenker (2005) for an overview. To select the significant variables is an important problem for QR. Many procedures have been proposed. Koenker (2004) applies the Lasso penalty to the mixed-effects linear QR model for longitudinal data to shrink the estimator of random effects. Wang, Li, and Jiang (2007) consider linear LADR with the adaptive Lasso penalty. Zou and Yuan (2008) propose a model selection procedure based on composite linear QRs. Wu and Liu (2009) consider the SCAD and adaptive Lasso in linear QR models. Belloni and Chernozhukov (2011a) consider l_1-penalized or post-l_1-penalized QR in high-dimensional linear sparse models. Liang and Li (2009) propose penalized QR (PQR) for PLMs with measurement error by using orthogonal regression to correct the bias in the loss function due to measurement error. Koenker (2011) considers the additive models for QR which include both parametric and nonparametric components. Kai, Li, and Zou (2011) consider efficient estimation and variable selection for semiparametric varying-coefficient PLM using composite QR. Lin, Zhang, Bondell, and Zou (2012) consider variable selection for nonparametric QR via SS-ANOVA. In this section, we focus on reviewing variable selection in semiparametric/nonparametric QR models.

9.8.1. Liang and Li's (2009) Penalized Quantile Regression for PLMs with Measurement Error

Liang and Li (2009) consider the PLM in (9.19) when X_i is measured with additive error:

$$W_i = X_i + U_i, \tag{9.58}$$

where U_i is the measurement error with mean zero and unknown covariance Σ_{uu} and U_i is independent of (X_i, Z_i, Y_i). They propose a penalized quantile regression (PQR) based on the orthogonal regression. That is, the objective function is defined as the sum of squares of the orthogonal distances from the data points to the straight line of regression function, instead of residuals from the classical regression. He and Liang (2000) apply the idea of orthogonal regression for QR for both linear and partially linear models with measurement error, but do not consider the variable selection problem. Liang and Li further use the orthogonal regression method to develop a PQR procedure to select significant variables in the PLMs.

To define orthogonal regression for QR with measurement error, it is assumed that the random vector $(\varepsilon_i, U_i')'$ follows an elliptical distribution with mean zero and covariance matrix $\sigma^2\Sigma$ where σ^2 is unknown, and Σ is a block diagonal matrix with $(1,1)$-element being 1 and the last $p \times p$ diagonal block matrix being C_{uu}. Liang and Li assume that C_{uu} is known but discuss that it can be estimated with partially replicated observations in practice.

Let $\rho_\tau(v) = v(\tau - 1(v < 0))$. Note that the solution to minimizing $\rho_\tau(\varepsilon_i - v)$ over $v \in \mathbb{R}$ is the τth quantile of ε_i. Liang and Li define the PQR objective function to be of the form

$$L_\tau(\beta) = \sum_{i=1}^{n} \rho_\tau\left(\frac{\hat{Y}_i - \hat{W}_i'\beta}{\sqrt{1+\beta' C_{uu}\beta}}\right) + n\sum_{j=1}^{p} p_{\lambda_j}\big(|\beta_j|\big), \tag{9.59}$$

where $\hat{Y}_i = Y_i - \hat{m}_y(Z_i)$ and $\hat{W}_i = W_i - \hat{m}_w(Z_i)$ using the notation defined in Section 9.4.5, and $p_{\lambda_j}(\cdot)$ is a penalty function with tuning parameter λ_j. He and Liang (2000) propose the QR estimate of β by minimizing the first term in (9.59) and also provide insights for this. Compared with the PLS in Section 9.4.5, the PQR uses the factor $\sqrt{1+\beta' C_{uu}\beta}$ to correct the bias in the loss function due to the presence of measurement error in X_i. Liang and Li establish the oracle property for the PQR estimator by assuming that p is fixed.

9.8.2. Koenker's (2011) Additive Models for Quantile Regression

Koenker (2011) considers models for conditional quantiles indexed by $\tau \in (0,1)$ of the general form

$$Q_{Y_i|X_i,Z_i}(\tau|X_i, Z_i) = X_i'\theta_0 + \sum_{j=1}^{q} g_j\big(Z_{ij}\big), \tag{9.60}$$

where X_i is a $p \times 1$ vector of regressors that enter the conditional quantile function linearly, and the nonparametric component g_i's are continuous functions, either univariate or bivariate. Let $g = \big(g_1, \ldots, g_q\big)'$ be a vector of functions. Koenker proposes to estimate these unknown functions and θ_0 by solving

$$\min_{(\theta,g)} \sum_{i=1}^{n} \rho_\tau\left(Y_i - X_i'\theta - \sum_{j=1}^{q} g_j\big(Z_{ij}\big)\right) + \lambda_0\|\theta\|_1 + \sum_{j=1}^{q} \lambda_j V\big(\nabla g_j\big), \tag{9.61}$$

where $\rho_\tau(u)$ is defined as above, $\|\theta\|_1 = \sum_{k=1}^{p}|\theta_k|$, and $V(\nabla g_j)$ denotes the total variation of the derivative or gradient of the function g_j. For g with absolutely continuous derivative g', the total variation of $g' : \mathbb{R} \to \mathbb{R}$ is given by $V(g'(z)) = \int |g''(z)|\,dz$, while for $g : \mathbb{R}^2 \to \mathbb{R}$, $V(\nabla g) = \int \|\nabla^2 g(z)\|\,dz$, where $\|\cdot\|$ is the usual Hilbert–Schmidt norm for matrices and $\nabla^2 g(z)$ denotes the Hessian of $g(z)$. The Lasso penalty $\|\theta\|_1$ leads to a sparse solution for parametric components and then selects the nonzero parametric components.

To select the tuning parameter λ, Koenker proposes an SIC-like criterion

$$SIC(\lambda) = n \log \hat{\sigma}(\lambda) + \frac{1}{2} p(\lambda) \log n,$$

where $\hat{\sigma}(\lambda) = n^{-1} \sum_{i=1}^{n} \rho_\tau(Y_i - \hat{g}(X_i, Z_i))$, and $p(\lambda)$ is the effective dimension of the fitted model

$$g(X_i, Z_i) = X_i'\hat{\theta} + \sum_{j=1}^{q} \hat{g}_j(Z_{ij}),$$

where Z_i is a collection of Z_{ij}'s. For a linear estimator, $p(\lambda)$ is defined as the trace of a pseudo-projection matrix, which maps observed response into fitted values. In general form,

$$p(\lambda) = \operatorname{div}(\hat{g}) = \sum_{i=1}^{n} \frac{\partial \hat{g}(X_i, Z_i)}{\partial Y_i}.$$

He proposes some methods to obtain the pointwise and uniform confidence bands for the estimate of nonparametric components but does not study the theoretical properties of the above variable selection procedure.

9.8.3. Kai, Li, and Zou's (2011) Composite Quantile Regression

Kai, Li, and Zou (2011) consider the following varying coefficient partial linear models

$$Y_i = \alpha_0(U_i) + X_i'\boldsymbol{\alpha}(U_i) + Z_i'\boldsymbol{\beta} + \varepsilon_i, \tag{9.62}$$

where (Y_i, U_i, X_i, Z_i), $i = 1, \ldots, n$, are i.i.d., $\alpha_0(\cdot)$ is a baseline function of scalar random variable U_i, $\boldsymbol{\alpha}(\cdot) = \{\alpha_1(\cdot), \ldots, \alpha_{d_1}(\cdot)\}'$ consists of d_1 unknown varying coefficient functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_2})'$ is a d_2-dimensional coefficient vector, and ε_i is random error with zero mean and CDF $F(\cdot)$. They assume that ε_i is independent of U_i, X_i, and Z_i.

Note that the τth conditional quantile function of Y_i given $(U_i, X_i, Z_i) = (u, x, z)$ is

$$Q_\tau(u, x, z) = \alpha_0(u) + x'\boldsymbol{\alpha}(u) + z'\boldsymbol{\beta} + c_\tau,$$

where $c_\tau = F^{-1}(\tau)$. All quantile regression estimates ($\hat{\boldsymbol{\alpha}}_\tau(u)$ and $\hat{\boldsymbol{\beta}}_\tau$) estimate the same target quantities ($\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}$) with the optimal rate of convergence. Therefore, they consider combining the information across multiple quantile estimates to obtain improved estimates of $\boldsymbol{\alpha}(u)$ and $\boldsymbol{\beta}$, which leads to the *composite quantile regression* (CQR) proposed by Zou and Yuan (2008). Let $\tau_k = k/(q+1)$, $k = 1,\ldots,q$ for a given q. The CQR estimates of $\alpha_0(\cdot)$, $\boldsymbol{\alpha}(\cdot)$, and $\boldsymbol{\beta}$ are obtained by minimizing the following CQR loss function:

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\big(Y_i - \alpha_0(U_i) - X_i'\boldsymbol{\alpha}(U_i) - Z_i'\boldsymbol{\beta}\big).$$

Note that $\alpha_j(\cdot)$ are unknown for $j = 0,1,\ldots,d_1$, but they can be approximated locally by linear functions: $\alpha_j(U) \approx \alpha_j(u) + \alpha_j'(u)(U-u) = a_j + b_j(U-u)$ when U lies in the neighborhood of u. Then let $\{\tilde{\mathrm{a}}_0, \tilde{b}_0, \tilde{\mathrm{a}}, \tilde{\mathrm{b}}, \tilde{\boldsymbol{\beta}}\}$ be the minimizer of the local CQR function defined by

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\big\{Y_i - a_{0k} - b_0(U_i - u) - X_i'[\mathrm{a} + \mathrm{b}(U_i - u)] - Z_i'\beta\big\}K_h(U_i - u),$$

where $K_h(u) = K(u/h)/h$ with K and h being the kernel and bandwidth, respectively, $\mathrm{a}_0 = \big(a_{01},\ldots,a_{0q}\big)'$, $\mathrm{a} = \big(a_1,\ldots,a_{d_1}\big)'$, $\mathrm{b} = \big(b_1,\ldots,b_{d_1}\big)'$, and $\tilde{\mathrm{a}}_0 = \big(\tilde{a}_{01},\ldots,\tilde{a}_{0q}\big)'$, and we have suppressed the dependence of these estimates on u. Initial estimates of $\alpha_0(u)$ and $\boldsymbol{\alpha}(u)$ are then given by

$$\tilde{\alpha}_0(u) = \frac{1}{q}\sum_{k=1}^{q}\tilde{a}_{0k} \quad \text{and} \quad \tilde{\boldsymbol{\alpha}}(u) = \tilde{\mathrm{a}}.$$

Given these initial estimates, the estimate of $\boldsymbol{\beta}$ can be refined by

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}}\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\big[Y_i - \tilde{a}_{0k}(U_i) - X_i'\tilde{\mathrm{a}}(U_i) - Z_i'\boldsymbol{\beta}\big],$$

which is called the semi-CQR estimator of $\boldsymbol{\beta}$. Given $\hat{\boldsymbol{\beta}}$, the estimates of the nonparametric parts can be improved by the following minimization problem:

$$\min_{\mathrm{a}_0, b_0, \mathrm{a}, \mathrm{b}}\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\Big[Y_i - Z_i'\hat{\boldsymbol{\beta}} - a_{0k} - b_0(U_i - u) - X_i'[\mathrm{a} + \mathrm{b}(U_i - u)]\Big]K_h(U_i - u).$$

In view of the fact that variable selection is a crucial step in high-dimensional modeling, Kai, Li, and Zou focus on the selection of nonzero components in the vector $\boldsymbol{\beta}$ of parametric coefficients. Let $p_{\lambda_n}(\cdot)$ be a penalty function with tuning parameter λ_n. The penalized loss function is

$$\sum_{k=1}^{q}\sum_{i=1}^{n}\rho_\tau\big(Y_i - \tilde{a}_{0k}(U_i) - X_i'\tilde{\boldsymbol{\alpha}}(U_i) - Z_i'\boldsymbol{\beta}\big) + nq\sum_{j=1}^{d_2}p_{\lambda_n}\big(|\beta_j|\big).$$

Note that the objective function is nonconvex, and both loss function and penalty parts are nondifferentiable. They propose to follow the one-step sparse estimate scheme in Zou and Li (2008) to derive a one-step sparse semi-CQR estimator. First, they obtain the unpenalized semi-CQR estimator $\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_1^{(0)}, \ldots, \hat{\beta}_{d_2}^{(0)})'$. Then they define

$$G_{n,\lambda_n}(\boldsymbol{\beta}) = \sum_{k=1}^{q}\sum_{i=1}^{n} \rho_\tau\big(Y_i - \tilde{\alpha}_{0k}(U_i) - X_i'\tilde{\boldsymbol{\alpha}}(U_i) - Z_i'\boldsymbol{\beta}\big) + nq\sum_{j=1}^{d_2} p'_{\lambda_n}\Big(\Big|\hat{\beta}_j^{(0)}\Big|\Big)|\beta_j|.$$

They refer to $\hat{\boldsymbol{\beta}}^{OSE}(\lambda_n) = \operatorname{argmin}_{\boldsymbol{\beta}} G_{n,\lambda_n}(\boldsymbol{\beta})$ as a *one-step sparse semi-CQR estimator.*

Under some conditions, they show that $\hat{\boldsymbol{\beta}}^{OSE}$ enjoys the oracle property and that the property holds for a class of concave penalties. To choose the tuning parameter λ, a BIC-like criterion is proposed as follows

$$BIC(\lambda) = \log\left[\sum_{k=1}^{q}\sum_{i=1}^{n} \rho_\tau\Big(Y_i - \hat{\alpha}_{0k}(U_i) - X_i'\hat{\boldsymbol{\alpha}}(U_i) - Z_i'\hat{\boldsymbol{\beta}}^{OSE}(\lambda)\Big)\right] + \frac{\log n}{n}\, df_\lambda,$$

where df_λ is the number of nonzero coefficients in the parametric part of the fitted models. They propose to use $\hat{\lambda}_{BIC} = \operatorname{argmin}_\lambda BIC(\lambda)$ as the tuning parameter.

9.8.4. Lin, Zhang, Bondell, and Zou's (2012) Sparse Nonparametric Quantile Regression

Lin, Zhang, Bondell, and Zou (2012) adopt the COSSO-type penalty to develop a new penalized framework for joint quantile estimation and variable selection. In the framework of SS-ANOVA, a function $f(x) = f(x^{(1)}, \ldots, x^{(p)})$ has the ANOVA decomposition in (9.48). The entire tensor-product space for estimating $f(x)$ is given in (9.49). But in practice the higher-order interactions are usually truncated for convenience to avoid the curse of dimensionality. Equation (9.50) gives a general expression for truncated space. Using the notation defined in Section 9.7.1, the regularization problem of joint variable selection and estimation is defined by

$$\min_f \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\big(Y_i - f(X_i)\big) + \lambda\sum_{j=1}^{p} w_j\left\|P^j f\right\|_{\mathcal{F}},$$

where $P^j f$ is the projection of f on $\mathcal{F}^j$, the penalty function penalizes the sum of component norms, and $w_j \in (0,\infty)$ is weight. In principle, smaller weights are assigned to important function components while larger weights are assigned to less important components. This is in the same spirit of the adaptive Lasso and adaptive COSSO. They also propose to construct the weight w_j's from the data adaptively:

$$w_j^{-1} = \left\| P^j \tilde{f} \right\|_{n,L_2} = \left\{ n^{-1} \sum_{i=1}^{n} \left[P^j \tilde{f}(X_i) \right]^2 \right\}^{1/2} \qquad \text{for } j = 1, \ldots, p,$$

where $\tilde{f}$ is a reasonable initial estimator of f, say the kernel quantile regression (KQR) estimator of Li, Liu, and Zhu (2007) which is obtained by penalizing the roughness of the function estimator using its squared functional norm in a RKHS. That is, the KQR solves the regularization problem

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \big(Y_i - f(X_i) \big) + \frac{\lambda}{2} \|f\|_{\mathcal{H}_K}^2$$

where $\mathcal{H}_K$ is an RKHS and $\|\cdot\|_{\mathcal{H}_K}$ is the corresponding function norm.

An equivalent expression of the above optimization problem is

$$\min_{f,\theta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \big(Y_i - f(X_i) \big) + \lambda_0 \sum_{j=1}^{p} w_j^2 \theta_j^{-1} \left\| P^j f \right\|_{\mathcal{F}}^2 \qquad \text{s.t. } \sum_{j=1}^{p} \theta_j \le M, \theta_j \ge 0,$$

where both λ_0 and M are smoothing parameters. Lin, Zhang, Bondell, and Zou (2012) show that the solution has the following structure:

$$\hat{f}(x) = \hat{b} + \sum_{i=1}^{n} \hat{c}_i \sum_{j=1}^{p} \frac{\hat{\theta}_j}{w_j^2} R_j(X_i, x)$$

where $\hat{\mathbf{c}} = \big(\hat{c}_1, \ldots, \hat{c}_n\big)' \in \mathbb{R}^n$, $\hat{b} \in \mathbb{R}$, and $R_j(X_i, x)$ is the reproducing kernel of subspace $\mathcal{F}^j$. Let $\mathbf{R}^\theta$ be the $n \times n$ matrix with (k, l)th element $R_{kl}^\theta = \sum_{j=1}^{p} w_j^2 \theta_j^{-1} R_j(X_k, X_l)$. Let $\mathbf{c} = (c_1, \ldots, c_n)'$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$. The objective function becomes

$$\min_{b,\mathbf{c},\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - b - \sum_{k=1}^{n} c_k R_{kl}^\theta \right) + \lambda_0 \mathbf{c}' \mathbf{R}^\theta \mathbf{c} \qquad \text{s.t. } \sum_{j=1}^{p} \theta_j \le M, \ \theta_j \ge 0.$$

An iterative optimization algorithm is proposed to solve the above problem.

1. Fix θ, solve $(b, \mathbf{c})$ by

$$\min_{b,\mathbf{c}} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - b - \sum_{k=1}^{n} c_k R_{kl}^\theta \right) + \lambda_0 \mathbf{c}' \mathbf{R}^\theta \mathbf{c}.$$

2. Fix $(b, \mathbf{c})$, solve θ by

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i^* - \sum_{j=1}^{p} \theta_j G_{ij} \right) + \lambda_0 \mathbf{c}' \mathbf{G} \boldsymbol{\theta} \text{ s.t. } \sum_{j=1}^{p} \theta_j \le M, \ \theta_j \ge 0,$$

where $Y_i^* = Y_i - b$, and G_{ij} is the (i,j)th element of $n \times p$ matrix $\mathbf{G} = \left(w_1^{-2} R_1 \mathbf{c}, \ldots, w_p^{-2} R_p \mathbf{c}\right)$.

The optimization problems in steps 1–2 can be cast into quadratic programming and linear programming problems, respectively. So both can be solved using standard optimization softwares. A SIC-like criterion is proposed to select the tuning parameter. However, the theoretical properties of the new variable selection procedure are not discussed.

9.9. Concluding Remarks

In this chapter we survey some of the recent developments on variable selections in nonparametric and semiparametric regression models. We focus on the use of Lasso, SCAD, or COSSO-type penalty for variable or component selections because of the oracle property of the SCAD and the adaptive versions of Lasso and COSSO. The oracle property has been demonstrated for some of the variable selection procedures but not for others (e.g., variable selection in nonparametric/semiparametric QR). It is interesting to develop variable selection procedures with the oracle property for some of the models reviewed in this chapter. In addition, the i.i.d. assumption has been imposed in almost all papers in the literature. It is important to relax this assumption to allow for either heterogeneity or serial/spatial dependence in the data. More generally, one can study variable selection for more complicated semiparametric/nonparametric models via shrinkage.

Despite the huge literature on Lasso- or SCAD-type techniques in statistics, we have seen very few developments of them in econometrics until 2009. Almost all of the works on variable selection in statistics are based on the assumption that the regressors are uncorrelated with or independent of the error terms; that is they are exogenous. However, in economic applications there are many examples in which some covariates are endogenous due to measurement error, omitted variables, sample selection, or simultaneity. The endogeneity causes an inconsistent estimate by the PLS method, along with misleading statistical inference, and one has to resort to instrumental variables (IVs) to handle this problem. Caner (2009) seems to be the first published paper to address this issue through shrinkage GMM estimation. Since then we have observed a large literature on the use of Lasso- or SCAD-type techniques in econometrics to cope with endogeneity in parametric models. They fall into three categories. The first category focuses on selection of covariates or parameters in the structural equation (see Caner (2009), Caner and Zhang (2009), and Fan and Liao (2011, 2012)). Caner (2009) considers covariate selection in GMM with Bridge penalty when the number of parameters is fixed; Caner and Zhang (2009) study covariate selection in GMM via adaptive elastic-net estimation by allowing the number of parameters to diverge to infinity; Fan and Liao (2011) consider variable selection with endogenous covariates in ultra-high-dimensional regressions via penalized GMM and penalized empirical likelihood (EL); Fan and Liao (2012) propose a penalized focused GMM (FGMM) criterion function to select covariates. The second category focuses on the selection relevant IVs (or deletion of irrelevant/weak IVs) (see Belloni, Chernozhukov, and Hansen (2010), Caner and Fan (2011), and García (2011)). Belloni, Chernozhukov, and Hansen (2010) introduce a heteroskedasticity-consistent Lasso-type estimator to pick optimal instruments among many of them. Caner and Fan (2011) use the adaptive Lasso to distinguish relevant and irrelevant/weak instruments in heteroskedastic linear regression models with fixed numbers of covariates and IVs. García (2011) proposes a two stage least squares (2SLS) estimator in the presence of many weak and irrelevant instruments and heteroskedasticity. The third category focuses on the selection both covariates and valid IVs (see Liao (2011) and Gautier and Tsybakov (2011)). Liao (2011) considers the selection of valid moment restrictions via adaptive Lasso, Bridge, and SCAD, and the selection of group variables and group valid moment restrictions via adaptive group Lasso when the number of parameters is fixed. Gautier and Tsybakov (2011) extend the Dantzig selector of Candès and Tao (2007) to the linear GMM framework and propose a new procedure called *self-tuning instrumental variable* (STIV) estimator for the selection of covariates and valid IVs when the number of covariates/parameters can be larger than the sample size. Nevertheless, none of these works address the issue of flexible functional form. It is interesting to consider variable selection for semiparametric or nonparametric models with endogeneity. Sieve estimation or local GMM estimation via shrinkage seems to be a very promising field to delve into.

References

Avalos, M., Y., Grandvalet, and C. Ambroise, 2007. "Parsimonious Additive Models." *Computational Statistics & Data Analysis*, **51**, pp. 2851–2870.

Bach, F. R. 2008. "Consistency of the Group Lasso and Multiple Kernel Learning." *Journal of Machine Learning Research*, **9**, pp. 1179–1225.

Belloni, A., V. Chernozhukov, and C. Hansen. 2010. "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain." Working paper, MIT.

Belloni, A., and V. Chernozhukov. 2011a. "l_1-Penalized Quantile Regression in High-Dimensional Sparse Models." *Annals of Statistics*, **39**, pp. 82–130.

Belloni, A., and V. Chernozhukov. 2011b. "High-Dimensional Sparse Econometric Models: An Introduction." In *Inverse Problems and High Dimensional Estimation*, eds. P. Alquier, E. Gautier, and G. Stoltz, *Lectures in Statistics* **203**, Berlin: Springer, pp. 127–162.

Bernanke, B. S., J. Bovin, and P. Eliasz. 2005. "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach." *Quarterly Journal of Economics*, **120**, pp. 387–422.

Bertin, K., and G. Lecué, 2008. "Selection of Variable and Dimension Reduction in High-Dimensional Non-parametric Regression." *Electronic Journal of Statistics*, **2**, pp. 1223–1241.

Bunea, F. 2004. "Consistent Covariate Selection and Post Model Selection Inference in Semiparametric Regression." *Annals of Statistics*, **32**, pp. 898–927.

Bunea, F. 2008. “Consistent Selection via the Lasso for High Dimensional Approximating Regression Models.” *Institute of Mathematical Statistics Collection*, **3**, pp. 122–137.

Candès, E. J., and T. Tao. 2007. “The Dantzig Selector: Statistical Estimation when p Is Much Larger than n.” *Annals of Statistics*, **35**, pp. 2313–2351.

Caner, M. 2009. “Lasso-type GMM Estimator.” *Econometric Theory*, **25**, pp. 270–290.

Caner, M., and M. Fan. 2011. “A Near Minimax Risk Bound: Adaptive Lasso with Heteroskedastic Data in Instrumental Variable Selection.” Working paper, North Carolina State University.

Caner, M., and H. H. Zhang. 2009. “General Estimating Equations: Model Selection and Estimation with Diverging Number of Parameters.” Working paper, North Carolina State University.

Carroll, R. J., J. Fan, I. Gijbels, and M. P. Wand. 1997. “Generalized Partially Linear Single-Index Models.” *Journal of American Statistical Association*, **92**, pp. 477–489.

Chen, B., Y. Yu, H. Zou, and H. Liang. 2012. “Profiled Adaptive Elastic-Net Procedure for Partially Linear Models.” *Journal of Statistical Planning and Inference*, **142**, pp. 1773–1745.

Comminges, L., and A. S. Dalayan. 2011. “Tight Condition for Consistent Variable Selection in High Dimensional Nonparametric Regression.” *JMLR: Workshop and Conference Proceedings*, **19**, pp. 187–205.

Donoho, D. L., and I. M. Johnstone. 1994. “Ideal Spatial Adaptation via Wavelet Shrinkages.” *Biometrika*, **81**, pp. 425–455.

Donoho, D. L., I. M. Johnstone, G. Kerkyacharian, and D. Picard. 1995. “Wavelet Shrinkage: Asymptopia (with discussion)?” *Journal of the Royal Statistical Society, Series B*, **57**, pp. 301–369.

Efron B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. “Least Angle Regression.” *Annals of Statistics*, **32**, pp. 407–499.

Fan, J., Feng, Y., and R. Song. 2011. “Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models.” *Journal of American Statistical Association*, **116**, pp. 544–557.

Fan, J., T. Huang, and H. Peng. 2005. “Semilinear High-Dimensional Model for Normalization of Microarray Data: A Theoretical Analysis and Partial Consistency (with discussion).” *Journal of American Statistical Association*, **100**, pp. 781–813.

Fan, J., and R. Li. 2001. “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties.” *Journal of the American Statistical Association*, **96**, pp. 1348–1360.

Fan, J., and R. Li. 2006. “Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Recovery.” *Proceedings of the International Congress of Mathematicians, Madrid, Spain*, pp. 595–622.

Fan, J., and Y. Liao, 2011. “Ultra High Dimensional Variable Selection with Endogenous Covariates.” Working paper, Princeton University.

Fan, J., and Y. Liao. 2012. “Endogeneity in Ultrahigh Dimension.” Working paper, Princeton University.

Fan, J., and J. Lv. 2010. “A Selection Overview of Variable Selection in High Dimensional Feature Space.” *Statistica Sinica*, **20**, pp. 101–148.

Fan, J., J. Lv, and L. Qi. 2011. “Sparse High-Dimensional Models in Economics.” *Annual Review of Economics*, **3**, pp. 291–317.

Fan, J., and H. Peng. 2004. “On Non-Concave Penalized Likelihood with Diverging Number of Parameters.” *Annals of Statistics*, **32**, pp. 928–961.

Fan, J., Q. Yao, and Z. Cai. 2003. “Adaptive Varying-Coefficient Linear Models.” *Journal of the Royal Statistical Society, Series B*, **65**, pp. 57–80.

Fan, J., and J. T. Zhang. 1998. “Functional Linear Models for Longitudinal Data.” *Journal of the Royal Statistical Society, Series B*, **39**, pp. 254–261.

Fan J., J. Zhang, and K. Yu. 2011. “Asset Allocation and Risk Assessment with Gross Exposure Constraints for Vast Portfolios.” Working paper, Princeton University.

Frank, I. E., and J. H. Friedman. 1993. “A Statistical View of Some Chemometrics Regression Tools (with discussion).” *Technometrics*, **35**, pp. 109–148.

Friedman, J., T. Hastie, and R. Tibshirani. 2008. “Sparse Inverse Covariance Estimation with the Graphical Lasso.” *Biostatistics*, **9**, pp. 432–441.

Fu, W. 1998. “Penalized Regressions: The Bridge Versus the Lasso.” *Journal of Computational and Graphical Statistics*, **7**, pp. 397–416.

García, P. E. 2011. “Instrumental Variable Estimation and Selection with Many Weak and Irrelevant Instruments.” Working paper, University of Wisconsin, Madison.

Gautier, E., and A. Tsybakov. 2011. “High-dimensional Instrumental Variables Regression and Confidence Sets.” Working paper, CREST.

Green, P. J., and B. W. Silverman. 1994. *Nonparametric Regression and Generalized Linear Models*. London: Chapman and Hall.

Härdle, W., and T. M. Stoker. 1989. “Investigating Smooth Multiple Regression by the Method of Average Derivatives.” *Journal of American Statistical Association*, **84**, pp. 986–995.

He, X., and H. Liang. 2000. “Quantile Regression Estimates for a Class of Linear and Partially Linear Errors-in-Variables Models.” *Statistica Sinica*, **10**, pp. 129–140.

Horel, A. E., and R. W. Kennard. 1970. “Ridge Regression: Biased Estimation for Nonorthogonal Problems.” *Technometrics*, **12**, pp. 55–67.

Hristache, M., A. Juditsky, J. Polzehl, and V. Spokoiny. 2001. “Structure Adaptive Approach for Dimension Reduction.” *Annals of Statistics*, **29**, pp. 1537–1566.

Huang, J. P. Breheny, and S. Ma. 2012. “A Selective Review of Group Selection in High Dimensional Models.”*Statistical Science*, **27**, pp. 481–499.

Huang, J., J. L. Horowitz, and F. Wei. 2010. “Variable Selection in Nonparametric Additive Models.” *Annals of Statistics*, **38**, pp. 2282–2313.

Huang, J. Z., C. O. Wu, and L. Zhou. 2002. “Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements.” *Biometrika*, **89**, pp. 111–128.

Hunter, D. R., and R. Li. 2004. “Variable Selection Using MM Algorithms.” *Annals of Statistics*, **33**, pp. 1617–1642.

Ichimura, H. 1993. “Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models.” *Journal of Econometrics*, **58**, pp. 71–120.

Jagannathan, R., and T. Ma. 2003. “Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps.” *Journal of Finance*, **58**, pp. 1651–1683.

Kai, B., R. Li, and H. Zou. 2011. “New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-Coefficient Partially Linear Models.” *Annals of Statistics*, **39**, pp. 305–332.

Kato R., and T. Shiohama. 2009. “Model and Variable Selection Procedures for Semiparametric Time Series Regression.” *Journal of Probability and Statistics,* Article ID 487194, 37 pages. http://www.hindawi.com/journals/jps/2009/487194/.

Knight, K., and W. Fu. 2000. “Asymptotics for Lasso-Type Estimators.” *Annals of Statistics*, **28**, pp. 1356–1378.

Koenker, R. 2004. “Quantile Regression for Longitudinal Data.” *Journal of Multivariate Analysis*, **91**, pp. 74–89.

Koenker, R. 2005. *Quantile Regression.* Cambridge, UK: Cambridge University Press.

Koenker, R. 2011. “Additive Models for Quantile Regression: Model Selection and Confidence Bands.” *Brazilian Journal of Probability and Statistics*, **25**, pp. 239–262.

Kong, E., and Y. Xia. 2007. “Variable Selection for the Single-Index Model.” *Biometrika*, **94**, pp. 217–229.

Lafferty, J., and L. Wasserman. 2008. “Redeo: Sparse, Greedy Nonparametric Regression.” *Annals of Statistics*, **36**, pp. 18–63.

Leng, C., and H. H. Zhang. 2006. “Model Selection in Nonparametric Hazard Regression.” *Nonparametric Statistics*, **18**, pp. 316–342.

Li, K. C. 1991. “Sliced Inverse Regression for Dimensional Reduction (with discussion).” *Journal of the American Statistical Association*, **86**, pp. 417–429.

Li, K. C., N. H. Duan. 1989. “Regression Analysis under Link Violation.” *Annals of Statistics*, **17**, pp. 1009–1052.

Li, R., and H. Liang. 2008. “Variable Selection in Semiparametric Regression Modeling.” *Annals of Statistics*, **36**, pp. 261–286.

Li, Y., Liu, Y., and J. Zhu. 2007. “Quantile Regression in Reproducing Kernel Hilbert Spaces.” *Journal of the American Statistical Association*, **102**, pp. 255–267.

Lian, H. 2010. “Flexible Shrinkage Estimation in High-Dimensional Varying Coefficient Models.” Working paper, NTU.

Liang, H., and R. Li. 2009. “Variable Selection for Partially Linear Models with Measurement Errors.” *Journal of the American Statistical Association*, **104**, pp. 234–248.

Liang, H., X. Liu, R. Li, and C.-L. Tsai. 2010. “Estimation and Testing for Partially Linear Single-Index Models.” *Annals of Statistics*, **38**, pp. 3811–3836.

Liang, H., and N. Wang. 2005. “Partially Linear Single-Index Measurement Error Models.” *Statistica Sinica*, **15**, pp. 99–116.

Liao, Z. 2011. “Adaptive GMM Shrinkage Estimation with Consistent Moment Selection.” Working paper, UCLA.

Lin C., H. H. Zhang, H. D. Bondell, and H. Zou. 2012. “Variable Selection for Nonparametric Quantile Regression via Smoothing Spline ANOVA”. Working paper, North Carolina State University.

Lin, Y., and H. H. Zhang. 2006. “Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models–COSSO.” *Annals of Statistics*, **34**, pp. 2272–2297.

Liu, X., L. Wang, and H. Liang. 2011. “Estimation and Variable Selection for Semiparametric Additive Partially Linear Models.” *Statistica Sinica*, **21**, pp. 1225–1248.

Meier, L., S. A. Van De Geer, and P. Bühlmann. 2009. “High-Dimensional Additive Modeling.” *Annals of Statistics*, **37**, pp. 3379–3821.

Miller, H., and P. Hall. 2010. “Local Polynomial Regression and Variable Selection.” Borrowing Strength: Theory Powering Applications—A Festschrift for Lawrence D. Brown. *IMS Collections*, **6**, pp. 216–233.

Naik, P. A., and C.-L. Tsai. 2001. “Single-index Model Selections.” *Biometrika*, **88**, pp. 821–832.

Ni, X., H. H. Zhang, and D. Zhang. 2009. “Automatic Model Selection for Partially Linear Models.” *Journal of Multivariate Analysis*, **100**, pp. 2100–2111.

Peng, H., and T. Huang. 2011. “Penalized Least Squares for Single Index Models.” *Journal of Statistical Planning and Inference*, **141**, pp. 1362–1379.

Qu, A., and R. Li. 2006. "Quadratic Inference Functions for Varying Coefficient Models with Longitudinal Data." *Biometrics*, **62**, pp. 379–391.

Ravikumar, P., H. Liu, J. Lafferty, and L. Wasserman. 2007. "Spam: Sparse Additive Models." *Advances in Neural Information Processing Systems*, **20**, pp. 1202–1208.

Rinaldo, A., 2009. "Properties and Refinement of the Fused Lasso." *Annals of Statistics*, **37**, pp. 2922–2952.

Storlie, C. B., H. D. Bondell, B. J. Reich, and H. H. Zhang. 2011. "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property." *Statistica Sinica*, **21**, pp. 679–705.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B*, **58**, pp. 267–288.

Tibshirani, R. 2011. "Regression Shrinkage and Selection via the Lasso: A Retrospective." *Journal of the Royal Statistical Society, Series B*, **73**, pp. 273–282.

Tibshirani, R. J., H. Hoefling, and R. Tibshirani. 2010. "Nearly-Isotonic Regression." Working paper, Stanford University.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. "Sparsity and Smoothness via the Fused Lasso." *Journal of the Royal Statistical Society, Series B*, **67**, pp. 91–108.

Wahba, G., 1990. *Spline Models for Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics.

Wang, H., and C. Leng. 2008 "A Note of Adaptive Group Lasso." *Computational Statistics and Data Analysis*, **52**, pp. 5277–5286.

Wang, H., G. Li, and G. Jiang. 2007. "Robust Regression Shrinkage and Consistent Variable Selection through the LAD—Lasso." *Journal of Business & Economic Statistics*, **25**, pp. 347–355.

Wang, H., and Y. Xia. 2009. "Shrinkage Estimation of the Varying Coefficient Model." *Journal of the American Statistical Association*, **104**, pp. 747–757.

Wang, L., G. Chen, and H. Li. 2007. "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data." *Bioinformatics*, **23**, pp. 1486–1494.

Wang, L., H. Li, and J. H. Huang. 2008. "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements." *Journal of the American Statistical Association*, **103**, pp. 1556–1569.

Wang, T., P.-R. Xu, and L.-X. Zhu. 2012. "Non-convex Penalized Estimation in High-dimensional Models with Single-index Structure." *Journal of Multivariate Analysis*, **109**, pp. 221–235.

Wei, F., and J. Huang. 2010. "Consistent Group Selection in High-dimensional Linear Regression." *Bernoulli*, **16**, pp. 1369–1384.

Wei, X., J. Huang, and H. Li. 2011. "Variable Selection and Estimation in High-dimensional Varying-coefficient Models." *Statistica Sinica*, **21**, pp. 1515–1540.

Wu, Y., and Y. Liu. 2009. "Variable Selection in Quantile Regression." *Statistic Sinica*, **19**, pp. 801–817.

Xia, Y., H. Tong, W. K. Li, and L. Zhu. 2002. "An Adaptive Estimation of Dimension Reduction Space." *Journal of the Royal Statistical Society, Series B*, **64**, pp. 363–410.

Xie, H., and J. Huang. 2009. "SCAD-Penalized Regression in High-Dimensional Partially Linear Models." *Annals of Statistics*, **37**, pp. 673–696.

Xue, L. 2009. "Consistent Variable Selection in Additive Models." *Statistica Sinica*, **19**, pp. 1281–1296.

Yang, B. 2012. *Variable Selection for Functional Index Coefficient Models and Its Application in Finance and Engineering.* Ph.D. thesis, University of North Carolina at Charlotte.

Yuan, M., and Y. Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society, Series B*, **68**, pp. 49–67.

Yuan, M., and Y. Lin. 2007. "Model Selection and Estimation in the Gaussian Graphical Model." *Biometrika*, **94**, pp. 19–35.

Zeng, P., T. He, and Y. Zhu. 2011. "A Lasso-Type Approach for Estimation and Variable Selection in Single Index Models." *Journal of Computational and Graphical Statistical*, **21**, pp. 92–109.

Zhang, C.-H. 2010. "Nearly Unbiased Variable Selection under Minimax Concave Penalty." *Annals of Statistics*, **38**, pp. 894–932.

Zhang, H., and Y. Lin. 2006. "Component Selection and Smoothing for Nonparametric Regression in Exponential Families." *Statistica Sinica*, **16**, pp. 1021–1042.

Zhao, P., and L. Xue. 2011. "Variable Selection for Varying Coefficient Models with Measurement Errors." *Metrika*, **74**, pp. 231–245.

Zhao, P., and B. Yu. 2006. "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research*, **7**, pp. 2541–2563.

Zhu, L.-P., L. Qian, and J. Lin. 2011. "Variable Selection in a Class of Single-index Models." *Annals of the Institute of Statistical Mathematics*, **63**, pp. 1277–1293.

Zhu, L.-P., and L.-X. Zhu. 2009. "Nonconcave Penalized Inverse Regression in Single-index Models with High Dimensional Predictors." *Journal of Multivariate Analysis*, **100**, pp. 862–875.

Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101, pp. 1418–1429.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B*, **67**, pp. 301–320.

Zou, H., and R. Li. 2008. "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models (with discussion)." *Annals of Statistics*, **36**, pp. 1509–1533.

Zou, H., and M. Yuan. 2008. "Composite Quantile Regression and the Oracle Model Selection Theory." *Annals of Statistics*, **36**, pp. 1509–1533.

Zou, H., and H. H. Zhang. 2009. "On the Adaptive Elastic-Net with a Diverging Number of Parameters." *Annals of Statistics*, **37**, pp. 1733–1751.

CHAPTER 10

DATA-DRIVEN MODEL EVALUATION: A TEST FOR REVEALED PERFORMANCE

JEFFREY S. RACINE AND CHRISTOPHER F. PARMETER[†]

10.1. Introduction

Having estimated a parametric model in the course of applied data analysis, one ought naturally test for model adequacy (i.e., for correct specification). When the parametric model is rejected by the data, practitioners often turn to more flexible methods—for example, nonparametric models. But there is no guarantee that the nonparametric model that one has adopted will perform any better than the parametric model that has been deemed inadequate, even though the nonparametric model may indeed exhibit an apparent marked improvement in (within-sample) fit according to a variety of metrics.[1]

This is widely appreciated in the time-series literature where out-of-sample predictive performance is an overriding concern.[2] By way of example, Medeiros, Teräsvirta, and Rech (2006) consider using autoregressive neural network models (AR-NN) to model financial time series. However, having rejected linearity, fitted an AR-NN model, and conducted a rigorous postmortem analysis of each model's ability to predict stock returns, Medeiros et al. (2006, p. 69) conclude that the "NN modelling strategy [...] is not any better than a linear model with a constant composition of variables. A nonlinear model cannot therefore be expected to do better than a linear one." See also Racine (2001) for an alternative example.

Indeed, there is no guarantee that a parametric model that passes a test for model adequacy will perform better than a nonparametric model because it is known that overspecified parametric models suffer efficiency losses and may perform worse than alternative specifications. However, focusing instead on out-of-sample predictive ability may provide the applied researcher with a potential avenue for discriminating among such models. Though a literature that advocates in-sample predictive evaluation in time-series settings has recently emerged (see Inoue and Kilian, 2004), this swims against the tide of a large body of literature that convincingly argues for the use of sample-splitting mechanisms whereby one splits the full sample into two subsamples and then uses one subsample for estimation and the other to guide predictive evaluation (see Corradi and Swanson (2007) and the references therein).

Out-of-sample predictive performance appears to be the metric of choice for time series researchers (see Diebold and Mariano (1995), West (1996), West and McCracken (1998), and McCracken (2000), among others). However, to the best of our knowledge, the insights underlying this literature have as yet to permeate cross-section applications. Furthermore, there remains scope for improvement in the time-series setting, as will be demonstrated. In this chapter we show how, through judicious use of an appropriate resampling mechanism, the proposed approach provides an appealing alternative to popular time-series tests for predictive accuracy by overcoming what we regard as limitations associated with such tests, namely, the reliance on a single split of the data and constraints placed on the minimum size of the hold-out sample driven by power considerations. As well, the approach is equally at home in cross-sectional settings.

In this chapter we take the view that fitted statistical models are approximations,[3] a perspective that differs from that of consistent model selection which posits a finite-dimensional "true model." That is, in this chapter we are not interested in tests that hypothesize one model being the "true model." Rather, our goal is instead to test whether one approximate model's expected performance is better than another on data drawn from the same DGP according to a prespecified loss function such as square or absolute error loss. The loss function is provided by the user; hence the method suggested herein is quite general.[4]

Our approach is firmly embedded in the statistics literature dealing with apparent versus true error estimation; for a detailed overview of "apparent," "true," and "excess" error, we direct the reader to Efron (1982, Chapter 7). In effect, within-sample measures of fit gauge "apparent error," which will be more optimistic than "true error," sometimes strikingly so, since a model is selected to *fit* the data best. For a given loss function, $\ell(u)$, one might compute the expected loss, $n^{-1}\sum_{i=1}^{n}\ell(\hat{u}_i)$, which provides an estimate of the apparent error arising from the modeling process. But all such within-sample measures are fallible, which is why they cannot be recommended as guides for model selection; for example, R^2 does not take into account model complexity, and adjusted R^2 measures are not defined for many semi- and nonparametric methods, whereas information-based measures such as AIC can be biased if the sequence of competing (parametric) models is non-nested (see Ye (1998) and Shen and Ye (2002)).

The approach we advocate involves constructing the distribution function of a model's true error and testing whether the expected true error is statistically smaller for one model than another. This will be accomplished by leveraging repeated splits of the data rather than just one as is commonly done and by computing the estimated loss for the hold-out data for each split. At the end of the day we will conclude that one model has statistically smaller estimated expected true error than another and therefore is expected to be closer to the true DGP and hence is preferred, though both models are, at best, approximations.

The basic idea is, of course, not new and involves splitting the data into two independent samples of size n_1 and n_2, fitting models on the first n_1 observations, then evaluating the models on the remaining $n_2 = n - n_1$ observations using, by way of example, average square prediction error (ASPE) (we know the outcomes for the evaluation data, hence this delivers an estimate of true error).[5] However, one might mistakenly favor one model when the estimate of true error is lower, but this in fact simply reflects a particular division of the data into two independent subsets that may not be representative of the DGP, that is, this can be overly influenced by which data points end up in each of the two samples. To overcome this limitation, one might consider repeating this process a large number of times, say $S = 10{,}000$ times, each time refitting the models on the "training" data (the n_1 observations) and evaluating the independent "evaluation" data (the $n_2 = n - n_1$ hold-out observations). This repeated sample-splitting experiment will thereby produce two vectors of length S which represent draws from the distribution of actual ASPEs for each model.[6] These two vectors of draws can then be used to discriminate between the two models.[7] For what follows we consider a simple (paired) test of differences in means for the two distributions, but also consider simple graphical tools that will help reveal stochastic dominance relationships, if present. Given that the test is a test of whether the data at hand reveal that the predictive performance of one econometric model is statistically different from that of another, we dub the test the "RP" test to denote "revealed performance."

A natural question to raise at this point is whether there are gains to be had by using sample splits with $n_2 > 1$ versus simply using $n_2 = 1$. Consider the cross-sectional setting, thus when $n_2 = 1$ there exist n unique splits of the data, and here our approach could boil down to computing the delete-one cross-validation (DOCV) function and using this to compare models. The problem that could arise here is that DOCV is a common method for nonparametric and semiparametric model selection (i.e., bandwidth selection), and it turns out that models which are fit using DOCV cannot be subsequently evaluated using the same criterion. This is similar in spirit to the recent work of Leeb and Pötscher (2006), who show that one cannot use the same data set for both model selection and post-model selection estimation and inference. Simulations not reported here for brevity show that indeed one cannot use the same criterion to both fit a model and then conduct inference across models. Given the popularity of DOCV for model fitting, it is only natural to think of using $n_2 > 1$ when sample splitting for the purpose of model evaluation. This also delivers a simple framework for inference as noted above. Finally, DOCV would clearly be inappropriate in time-series settings.

The statistics literature on cross-validated estimation of excess error is a well-studied field ("expected excess error" is the expected amount by which the true error exceeds the apparent error). However, this literature deals with model specification within a class of models (i.e., which predictor variables should be used, whether or not to conduct logarithmic transformations on the dependent variable, and so forth) and proceeds by minimizing excess error. Our purpose here is substantively different and is perhaps most closely related to the literature on non-nested model testing (see Davidson and MacKinnon 2002). Unlike this literature, however, we are asking an inherently different question that is not the subject of interest in the non-nested literature, namely, whether the expected true error associated with one model differs *significantly* from that for another model, whether nested or not.

Our test is quite flexible with regard to the types of models that can be compared. The flexibility of the test stems from the fact that it does not require both models to be of the same type (e.g., both parametric). In fact, while our focus here is on regression models, the insight here can be extended to predictions from count data or limited dependent variable models, probability models, quantile frameworks, and so forth—that is, any model for which we have a response and set of explanatory variables.[8] Moreover, our method overcomes two of the drawbacks associated with dominant time series model comparison approaches, namely, their reliance on a single split of the data and the need to have a sufficiently large hold-out sample in order for the test to have adequate power.

The rest of this chapter proceeds as follows. Section 10.2. outlines the basic approach and framework for our proposed test. Section 10.3. conducts several simulation exercises to assess the finite-sample performance of the proposed approach when the DGP is known. Section 10.4. presents several empirical examples, while Section 10.5. presents some concluding remarks.

10.2. Methodology

The method we describe here is closest in spirit to the original application of cross-validation in which the data set is randomly divided into two halves, the first of which is used for model fitting and the second for cross-validation where the regression model fitted to the first half of the data is used to predict the second half. The more common modern variant in which one leaves out one data point at a time, fits the model to the remaining points, and then takes the average of the prediction errors (each point being left out once), yielding a cross-validated measure of true error, has been widely studied, and we direct the interested reader to Stone (1974), Geisser (1975), and Wahba and Wold (1975) for detailed descriptions of this method. It is noteworthy that White (2000, p. 1108) argues that cross-validation represents a "more sophisticated use of "hold-out" data" and indicates that this "is a fruitful area for further research." Our approach indeed supports this claim as we demonstrate that the use of cross-validation to asses a model's expected true error can lead to substantial power improvements over existing, single-split techniques commonly used in the applied times-series literature.

Though we shall begin with cross-sectional i.i.d. data and pure sample splitting (i.e., resampling *without* replacement), we shall see how the same intuition carries over to a variety of dependent data structures as well. For (strictly) stationary dependent processes, we adopt resampling methods rather than pure sample splitting, and it will be seen that, with some care, each resample can respect dependence in the original series and can itself be split (e.g., Politis and Romano, 1992) thereby allowing us to apply our method in time-series settings.

In our regression problem the data consist of pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where X_i is a $1 \times p$ vector of predictor variables and Y_i is a real-valued response variable. We presume that $Z_i = (X_i, Y_i)$ represent random draws from a (strictly) stationary ergodic process with unknown distribution function F defined on $\mathcal{H} = \mathbb{R}^{p+1}$,

$$Z_1, Z_2, \ldots, Z_{n_1} \sim F. \tag{10.1}$$

We observe $Z_1 = z_1, Z_2 = z_2, \ldots, Z_{n_1} = z_{n_1}$, and for what follows we let $Z^{n_1} = (Z_1, Z_2, \ldots, Z_{n_1})$ and $z^{n_1} = (z_1, z_2, \ldots, z_{n_1})$. Having observed $Z^{n_1} = z^{n_1}$, we fit a regression model that will be used to predict some "future" values of the response variable, which we denote

$$\hat{g}_{z^{n_1}}(x^{n_2}), \tag{10.2}$$

where the superscript n_2 indicates a new set of observations, $z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \ldots, z_n)$, which are distinct from $z^{n_1} = (z_1, z_2, \ldots, z_{n_1})$ where $n_2 = n - n_1$. By way of example, simple linear regression would provide $\hat{g}_{z^{n_1}}(x^{n_2}) = x^{n_2}\hat{\beta}_{n_1}$ where $\hat{\beta}_{n_1} = (x^{n_1\,T} x^{n_1})^{-1} x^{n_1\,T} y^{n_1}$, T denotes transpose, and $y^{n_1} = (y_1, y_2, \ldots, y_{n_1})$.

We are interested in estimating a quantity known as "expected true error" (Efron 1982, p. 51).[9] Following Efron (1982), we first define the "true error" to be

$$E_{n_2,F}\big[\ell\big(Y^{n_2} - \hat{g}_{Z^{n_1}}(X^{n_2})\big)\big], \tag{10.3}$$

where $\ell(\,\cdot\,)$ denotes a loss function specified by the researcher satisfying regularity conditions given in Assumption 10.2 below. The notation $E_{n_2,F}$ indicates expectation over the new point(s)

$$Z_{n_1+1}, Z_{n_1+2}, \ldots, Z_n \sim F, \tag{10.4}$$

independent of $Z_1, Z_2, \ldots, Z_{n_1}$, the variables which determine $\hat{g}_{Z^{n_1}}(\,\cdot\,)$ in (10.3) (we refer to Z^{n_1} as the "training set," terminology borrowed from the literature on statistical discriminant analysis). Next, we define "expected true error,"

$$E\big(E_{n_2,F}[\ell(\,\cdot\,)]\big), \tag{10.5}$$

the expectation over all potential regression surfaces $\hat{g}_{Z^{n_1}}(\,\cdot\,)$, for the selected loss function $\ell(\,\cdot\,)$. When comparing two approximate models, the model possessing lower "expected true error" will be expected to lie closest to the true DGP given the loss function $\ell(\,\cdot\,)$ and would therefore be preferred in applied settings.[10]

A realization of "true error" (10.3) based upon the observed $z^{n_2} = (z_{n_1+1}, z_{n_1+2}, \dots, z_n)$, is given by

$$\frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \ell\big(y_i - \hat{g}_{z^{n_1}}(x_i)\big), \tag{10.6}$$

an average prediction error which, for square error loss, we denote ASPE ("average square prediction error").

Were we given S such splits of the data, we could then construct the empirical distribution function (EDF) of (10.6). Given two competing models and each model's respective EDF of realized true error, we can then use the respective EDFs to determine whether one model has statistically significantly lower expected true error than another. Note that here we have transformed the problem into a (paired) two-sample problem where we wish to test for equivalence of expected true error defined in (10.5) based upon two vectors of realizations of true error defined in (10.6).[11] Thus, the procedure we consider is strictly data-driven and nonparametric in nature.

10.2.1. The Empirical Distribution of True Error

Suppose we arbitrarily denote one approximate model "Model A" and the other "Model B." For the sake of concreteness, let us presume that one was interested in comparing, say, a nonparametric kernel regression model ("Model A") to a parametric regression model ("Model B"). In a time-series context, it might appear that there is only one possible split of the data, $\{z_i\}_{i=1}^{t}$ and $\{z_i\}_{i=t+1}^{n}$, and this one split underlies many tests for predictive accuracy (or forecast equality) such as Diebold and Mariano's (1995) test. But, there is nothing to preclude conducting repeated resampling with time-series data; we just need to use an appropriate resampling methodology.

We first consider the case where the data represent independent draws from the underlying DGP. When the data represent independent draws, we could proceed as follows:

(i) Resample without replacement pairwise from $z = \{x_i, y_i\}_{i=1}^{n}$ and call these resamples $z_* = \{x_i^*, y_i^*\}_{i=1}^{n}$.

(ii) Let the first n_1 of the resampled observations form a training sample, $z_*^{n_1} = \{x_i^*, y_i^*\}_{i=1}^{n_1}$, and let the remaining $n_2 = n - n_1$ observations form an evaluation sample, $z_*^{n_2} = \{x_i^*, y_i^*\}_{i=n_1+1}^{n}$.

(iii) Holding the degree of smoothing[12] (i.e., the bandwidth vector scaling factors) of Model A and the functional form of Model B fixed (i.e., at that for the full sample), fit each model on the training observations ($z_*^{n_1}$) and then obtain predicted values from the evaluation observations ($z_*^{n_2}$) that were not used to fit the model.

(iv) Compute the ASPE of each model which we denote $\text{ASPE}^A = n_2^{-1} \sum_{i=n_1+1}^{n} (y_i^* - \hat{g}_{z^{n_1}}^A(x_i^*))^2$ and $\text{ASPE}^B = n_2^{-1} \sum_{i=n_1+1}^{n} (y_i^* - \hat{g}_{z^{n_1}}^B(x_i^*))^2$.

(v) Repeat this a large number of times, say $S = 10{,}000$, yielding S draws, $\{\text{ASPE}_s^A, \text{ASPE}_s^B\}_{s=1}^S$. We refer to the respective EDFs as $\hat{F}_S^A$ and $\hat{F}_S^B$ where each places mass $1/S$ at ASPE_s^A and ASPE_s^B.

Step (i), which involves resampling without replacement from $z = \{x_i, y_i\}_{i=1}^n$, is valid for heteroskedastic processes; however, it does presume independence among the draws. That is, by resampling (x_i, y_i) pairs we avoid resorting to, for instance, the "wild bootstrap," which is a residual-based bootstrap that admits heteroskedastic errors. However, in a time-series context, independent pairwise resampling is clearly inappropriate. For one thing, univariate time-series models are quite popular but require a different treatment because we need to respect dependence in the series itself.

In the context of time-series prediction ("forecasting"), resampling methods are widely used. For instance, Corradi and Swanson (2002) propose a consistent test for nonlinear predictive accuracy for nested models where interest lies in testing whether the null model can outperform the nesting alternative model based upon "real-time forecasts" (i.e., one-step recursive forecasts for period $t+1$, $t+2$, and so on) and one split of the data. Corradi and Swanson (2004) examine finite-sample properties of their 2002 test where critical values are based on application of the block bootstrap. Corradi and Swanson (2004, Tables 1 and 2) employ manually set fixed block lengths and they note that the value of the test statistic(s) under consideration and the resulting power properties vary dramatically as the block length is changed. There is no reason to require the user to set block lengths manually, however, just as there is no reason to require users to manually specify bandwidths for kernel estimation; automatic methods possessing requisite statistical properties exist and are available for use.

In what follows, we shall exploit recent advances in time-series resampling methodology, and use geometric ("stationary") block bootstrapping to generate a bootstrap replication of the series of size n which then can itself be split into two samples of size n_1 and n_2, thereby preserving the dependence structure present in the full sample. That is, in a (univariate) time-series setting, we deploy a time-series bootstrap based on automatic block length selection where we resample from, say, $z = \{y_i\}_{i=1}^n$. By way of illustration, we elect to use the method of Politis and Romano (1992) for which Politis and White (2004) recently proposed a fully automatic method for choosing the block length that has excellent finite-sample properties.[13] This bootstrap preserves the underlying dependence structure by resampling the data in blocks of random length, where the lengths are derived from a geometric distribution, hence the name. See both Davison and Hinkley (1997, pp. 401–408) and Lahiri (2003, Sections 2.7.2 and 3.3) for more on the theoretical underpinnings underlying the geometric bootstrap.[14] In a time-series setting where the data represent draws from a (strictly) stationary ergodic process, we proceed as follows:

(i) Apply the stationary bootstrap to resample from $z = \{y_i\}_{i=1}^n$ and call these $z_* = \{y_i^*\}_{i=1}^n$.

(ii) Let the first n_1 of the resampled observations form a training sample, $z_*^{n_1} = \{y_i^*\}_{i=1}^{n_1}$; and let the remaining $n_2 = n - n_1$ observations form an evaluation sample, $z_*^{n_2} = \{y_i^*\}_{i=n_1+1}^{n}$.

(iii) Holding the degree of smoothing of the nonparametric Model A and the functional form of the parametric Model B fixed (i.e., at that for the full sample), fit each model on the training observations ($z_*^{n_1}$) and then generate predictions for the n_2 evaluation observations.

(iv) Compute the ASPE of each model, which we denote $\text{ASPE}^A = n_2^{-1} \sum_{i=n_1+1}^{n} (y_i^* - \hat{g}_{z^{n_1}}^A(y_{i-1}^*, \dots))^2$ and $\text{ASPE}^B = n_2^{-1} \sum_{i=n_1+1}^{n} (y_i^* - \hat{g}_{z^{n_1}}^B(y_{i-1}^*, \dots))^2$.

(v) Repeat this a large number of times, say $S = 10{,}000$, yielding S draws, $\{\text{ASPE}_s^A, \text{ASPE}_s^B\}_{s=1}^{S}$. We refer to the respective EDFs as $\hat{F}_S^A$ and $\hat{F}_S^B$, where each places mass $1/S$ at ASPE_s^A and ASPE_s^B.

We can now proceed to use $\hat{F}_S^A$ and $\hat{F}_S^B$ to discriminate between models. At this stage we point out that the choice $S = 1$ is typically used to discriminate among time-series models; that is, one split only of the data is the norm. By way of example, the popular time-series test for predictive accuracy of Diebold and Mariano (1995) is based on only one split, hence attention has shifted to determining how large n_2 need be (e.g., see Ashley (2003)). One might, however, be worried about basing inference on only one split, mistakenly favoring one model over another when this simply reflects a particular division of the data into two independent subsets that may not be representative of the underlying DGP; that is, this can be overly influenced by which data points end up in each of the two samples.

However, by instead basing inference on $S \gg 1$ (i.e., averaging over a large number of such splits), we can control for mistakes arising from divisions of the data that are not representative of the DGP. In fact, it will be seen that the power of our test increases with S, which is obvious in hindsight. Furthermore, by averaging over a large number of splits, we can base inference on much smaller evaluation sample sizes (i.e., n_2) thereby taking maximal advantage of the estimation data which would be particularly advantageous in time-series settings. Ashley (2003) clearly illustrates this dilemma in the $S = 1$ time-series context by highlighting that one may need n_2 to be quite large in order for such tests to have power; the dilemma is that for a time-series of fixed length n, increasing $n_2 = n - n_1$ means that the models are less efficiently estimated since they are based on fewer observations. Our approach will be seen to effectively overcome this limitation.

10.2.2. Validity of the Bootstrap

We now consider conditions that justify our use of the bootstrap for obtaining valid approximations to the unknown loss distributions for two competing approximate models, which we denote F^A and F^B, respectively. For what follows we leverage Lemma A.3 and Theorem 2.3 in White (2000) to establish consistency of our bootstrap approach. The conditions required (for competing parametric models) involve assumptions on the data, parameter estimates, behavior of the bootstrap, properties of the loss function, and some additional regularity conditions. Before proceeding we note that the proof we provide is for the time-series method we describe above. However, for i.i.d. data the automatic block length selection mechanism and geometric bootstrap that we use for our time-series approach (Patton et al., 2009) will in fact deliver an appropriate bootstrap for independent data since it will select a block length of one in probability in these settings and will cover the mechanism we consider for independent data. That is, the proof will cover both cases considered above as will the implementation. For concreteness, we focus our theoretical arguments on the case where the competing models are both of parametric form (but potentially nonlinear). Extensions to semiparametric and nonparametric estimators are easily handled with (minor) modifications to the requisite assumptions listed below. Becuase the conditions we impose involve theoretical arguments described in three separate papers, we shall outline each set of assumptions in turn and cite sources accordingly.

We begin with an assumption given in Politis and Romano (1994) that is required to demonstrate consistency of the stationary bootstrap under a range of settings. For what follows we have $f_s = f(u_s)$, where the index of s follows from the context. β^* represents an unknown parameter vector.

Assumption 10.1.

(i) Let q^* denote the probability of the geometric distribution used for the stationary bootstrap (q^* is equivalent to one over the block length). Assume that $q^* \to 0$ and that $nq^* \to \infty$ as $n \to \infty$.

(ii) Let $Z_1, Z_2, \ldots,$ be a strictly stationary process with $E|Z_1|^{6+\eta} < \infty$ for some $\eta > 0$.

(iii) Let $\{Z_n\}$ be α-mixing with $\alpha_Z(k) = O(k^{-r})$ for some $r > 3(6+\eta)/\eta$.

Assumption 10.1(i) establishes the rate at which the block length in the stationary bootstrap can grow. Assumptions 10.1(ii) and 10.1(iii) are required to ensure that the data behave in a manner consistent with the theoretical arguments of both Politis and Romano (1994) and White (2000). Of course, in cross-section settings these conditions are automatically satisfied. Note that Assumption 10.1 is the same as that used by Politis and Romano (1994) for much of their theoretical work in this area (see Theorems 2–4, Politis and Romano 1994).

Additionally, we require assumptions 1–4 in West (1996). We restate these and label them jointly as Assumption 10.2.

Assumption 10.2.

(i) Let the loss function be second-order continuously differentiable at $\beta^* \equiv \text{plim}\,\hat{\beta}$, where $\hat{\beta}$ is defined below. Additionally, the matrix of second-order derivatives is dominated by m_n, where $E[m_n] < D$ for $D < \infty$.

(ii) *Let the parameter estimates be linear combinations of orthogonality conditions used to identify the response. More formally we have that (for the parametric regression model* $y_j = X_j'\beta + \varepsilon_j$, $j = 1, \ldots, n$) $\hat{\beta} - \beta^* = B(n)H(n)$, *where* $B(n)$ *is* $(k \times q)$ *and* $H(n)$ *is* $(q \times 1)$ *with (a)* $B(n) \xrightarrow{a.s.} B$, *a matrix with rank k (in our regression example* $B = (EX'X)^{-1}$), *(b)* $H(n) = n^{-1}\sum_{j=1}^{n} h_j(\beta^*)$ *for a* $(q \times 1)$ *orthogonality condition* $h_j(\beta^*)$ ($H(n) = n^{-1}\sum_{j=1}^{n} X_j\varepsilon_j$ *in the regression context) and (c)* $E[h_j(\beta^*)] = 0$.

(iii) *Let*

$$f_j = f(,\beta^*), \qquad f_{j,\beta} = \frac{\partial f_j}{\partial \beta}(,\beta^*), \quad F = E[f_{j,\beta}].$$

(a) For some $d > 1$, $\sup_j E||[vec(f_{j,\beta})', f_j', h_j']'||^{4d} < \infty$, *where* $||\cdot||$ *signifies the Euclidean norm. (b)* $[vec(f_{j,\beta} - F)', (f_j - E[f_j])', h_j']'$ *is strong mixing with mixing coefficient of size* $-3d/(d-1)$. *(c)* $[vec(f_{j,\beta})', f_j', h_j']'$ *is covariance stationary. (d)* S_{ff} *is positive definite, where* $S_{ff} = \sum_{k=-\infty}^{\infty} \Gamma_{ff}(k)$ *and* $\Gamma_{ff}(k) = E[(f_j - E[f_j])(f_{j-k} - E[f_j])']$.

(iv) *Let* $n_1, n_2 \to \infty$ *as* $n \to \infty$ *and let* $\lim_{n\to\infty}(n_2/n_1) = c$, *for* $0 \le c \le \infty$.

Assumption 10.2(i) ensures that the loss function is well-behaved in a neighborhood of a specified parameter value. Essentially, the loss function evaluated at the prediction errors needs to be bounded and satisfy certain moment conditions in order to use White's (2000) bootstrap theory. As noted by West (1996), Assumption 10.2(ii) does not assume that either ε or $X\varepsilon$ is serially uncorrelated. Assumption 10.2(iii) is used to pin down the behavior of the mean of the losses for a particular model by suitable application of a law of large numbers applicable to mixing processes (see Section 3.4, White 2001). Assumption 10.2(iv) is needed to invoke asymptotic arguments related to either the estimation sample size (n_1) or the evaluation sample size (n_2).

In order to invoke either Lemma A.3 or Theorem 2.3 of White (2000), we need two additional conditions. In White (2000) these are Assumption A.5 and Assumption C. We state them here for convenience.

Assumption 10.3.

(i) *Let the spectral density of* $[(f_i - Ef_i)', h_i'B']'$, *where* $f_i = f(y_i - \hat{y}_i)$, *at frequency zero, multiplied by a scale factor, be positive definite.*

(ii) *Let the parameter estimates* ($\hat{\beta}$) *obey a law of iterated logarithm.*

Assumption 10.3(ii) is required to bound a pseudo-studentized term involving $\hat{\beta}$ in White's (2000) Theorem 2.3.

These conditions are sufficient to establish that the bootstrap distribution of any (parametric) candidate model's evaluation sample loss is consistent for the distribution of expected true error, which we now state formally.

Theorem 10.1. *Under Assumptions 10.1, 10.2 and 10.3, the stationary bootstrap estimates of the distributional laws F^A and F^B, denoted $\hat{F}_S^A$ and $\hat{F}_S^B$, converge in probability to F^A and F^B.*

Proof. Given that Assumptions 10.1, 10.2, and 10.3 are identical to those in White (2000), we can invoke his Theorem 2.3 (which follows immediately from Lemma A.3) directly to achieve the result. We mention that his Theorem 2.3 follows under the condition that the objective function used for estimation and loss function are equivalent. This is not a deterrent, because Corradi and Swanson (2007, Proposition 1, p. 77) generalize the results of White (2000) for the case where the loss function differs from the objective function used to obtain the parameter estimates. In our work, and certainly for most applications, they are identical. ∎

Theorem 10.1 allows us to therefore implement a variety of two-sample tests to assess revealed performance (pairwise) across a set of candidate models. Of course, we need to address the possibility that the realizations defined in (10.6) are correlated for Model A and Model B (i.e., that there may exist pairwise correlation of the realizations underlying $\hat{F}_S^A$ and $\hat{F}_S^B$); thus one's testing procedure must accommodate potential pairwise correlation. But such tests are widely available to practitioners, two popular examples being the paired t-test and the paired Mann–Whitney–Wilcoxon tests; see also Mehta and Gurland (1969) for an alternative to the paired t-test.

When considering two-sample tests it would be convenient to be able to guide the users' choice of n_2 and their choice of tests. It is known that a sufficient sample size for the sampling distribution of the paired Mann–Whitney–Wilcoxon test to be approximated by the normal curve is $n_2 \geq 10$ regardless of the distributions F^A and F^B, while the matched-pairs t-test is strictly speaking valid only when F^A and F^B are normally distributed, though it is known that the t-test is quite robust to departures from normality. For the simulations we consider there is no qualitative difference between rejection frequencies based upon the paired Mann–Whitney–Wilcoxon and the t-test, so for what follows we consider the popular paired t-test for equality in means to assess whether one distribution dominates the other (i.e., test equality (less than or equal) of means against the alternative hypothesis that the true difference in means is greater than zero). Full results for both tests beyond those reported here are available upon request.

Formally, we state the null and alternative as

$$H_0 : E\big(E_{n_2,F^A}[\ell(\,\cdot\,)]\big) - E\big(E_{n_2,F^B}[\ell(\,\cdot\,)]\big) \leq 0$$

and

$$H_A : E\big(E_{n_2,F^A}[\ell(\,\cdot\,)]\big) - E\big(E_{n_2,F^B}[\ell(\,\cdot\,)]\big) > 0,$$

which arises directly from our notation in (10.5).

This is, of course, not the only test available to practitioners. One might prefer, say, the Mann–Whitney–Wilcoxon test (i.e., test equality (less than or equal) of locations against the alternative hypothesis that the true location shift is greater than zero) (see

Bauer (1972)). Or perhaps one might undertake a more sophisticated test for, say, first-order stochastic dominance (e.g., Davidson and Duclos (2000)). We argue that this is not needed in the present context, and a simple test for equality of locations in conjunction with summary plots of the vectors of ASPEs is more than sufficient for our purposes. Indeed, one of the appealing aspects of the proposed approach lies in its simplicity, though nothing would preclude the practitioner from considering additional tests in this setting because they will all be based on $\hat{F}_S^A$ and $\hat{F}_S^B$ which have been pre-computed and are consistent given Theorem 10.1.

We now proceed to some Monte Carlo simulations designed to assess the finite-sample performance of the proposed method.

10.3. Monte Carlo Simulations

10.3.1. Finite-Sample Performance: Cross-Sectional Data

We begin with a series of simulations that assess the finite-sample performance of the proposed data in the presence of cross-sectional data, and we consider a DGP of the form

$$y_i = 1 + x_{i1} + x_{i2} + \delta\left(x_{i1}^2 + x_{i2}^2\right) + \varepsilon_i, \qquad (10.7)$$

where $X \sim U[-2,2]$ and $\varepsilon \sim N(0,1)$. By setting $\delta = 0$ we simulate data from a linear model, and by setting $\delta \neq 0$ we simulate data from a quadratic model with varying strength of the quadratic component.

For what follows, we estimate a range of parametric models starting with one that is linear in X_1 and X_2 and then ones that include higher-order polynomials in X_1 and X_2 along with local constant (LC) and local linear (LL) nonparametric models. We consider testing whether the LL nonparametric specification is preferred to an LC nonparametric specification and each of the parametric specifications that are linear in X_1 and X_2 ($P = 1$), linear and quadratic in X_1 and X_2 ($P = 2$), and so forth, through models that have quintic specifications. Clearly our test is designed to compare two models only, hence we intend this exercise to be illustrative in nature. Models with $P > 2$ are therefore *overspecified* parametric models for this DGP. The null is that the LL model has true error (as measured by ASPE) that is lower than or equal to a particular model, with the alternative that it has ASPE that exceeds that for the particular model. The nonparametric models use DOCV bandwidth selection while the parametric models are fit by the method of least squares. Before proceeding we point out that, when the underlying DGP is in fact linear, cross-validation will choose smoothing parameters that tend to infinity with probability one asymptotically hence the LL model will collapse to a globally linear one (i.e., linear least squares) (see Li and Racine (2004) for further discussion). We mention this point since some readers may naturally expect that if a DGP is linear and one estimates a linear parametric specification, then it must always dominate a nonparametric specification. However, given that a nonparametric LL specification can collapse to the linear parametric specification when the true DGP is in fact linear, this need not be the case as the following simulations reveal.

For what follows we set $n = 200$, set $S = 1000$ or 10,000 (to investigate the impact of increasing the number of sample splits), consider a range of values for n_2, and report empirical rejection frequencies at the 5% level in Table 10.1. Large rejection frequencies indicate that the model in the respective column heading has improved predictive accuracy over the LL model.

Table 10.1 reveals a number of interesting features. For example, overspecified parametric models (that would be expected to pass tests for correct parametric specification) can indeed be dominated by the nonparametric LL specification for the reasons outlined above (e.g., $P = 4$ and $P = 5$, $\delta = 0.0$), which may be surprising to some. Furthermore, the power of the proposed approach to discriminate against incorrectly underspecified parametric models approaches one as δ increases (the column with heading $P = 1$), suggesting that the test can correctly reveal that a nonparametric model is preferred to an incorrectly underspecified parametric model. Also, the results of the test appear to stabilize after $n_2 = 25$, indicating that the size of the hold-out sample is not a crucial parameter to be set by the user; see Ashley (2003) for more on the appropriate size of the hold-out sample for forecasting in time-series domains. It is easy for the user to investigate the stability of their results with respect to choice of n_2, and we encourage such sensitivity checks in applied settings.

Comparing the corresponding entries in Table 10.1 ($S = 1000$) to Table 10.2 ($S = 10{,}000$), we observe that power increases with S as expected. This suggests that when one fails to reject the null it may be advisable to increase S to confirm that this is not simply a consequence of too few splits of the data being considered. Our experience with this approach is that $S = 10{,}000$ is sufficient to overcome such concerns.

10.3.2. Finite-Sample Performance: Time-Series Data

The time-series literature dealing with predictive accuracy and forecasting is quite vast, and we make no claims at surveying this literature here.[15] Early work on forecast model comparison by Ashley, Granger, and Schmalensee (1980) and Granger and Newbold (1986) generated broad interest in this topic. However, only recently have formal tests that directly relate to forecast accuracy and predictive ability surfaced. Most notably the available tests include Diebold and Mariano (1995) (the "DM" test) and the size-corrected counterpart of Harvey, Leybourne, and Newbold (1997) (the "MDM" test) along with those proposed by Swanson and White (1997), Ashley (1998), Harvey, Leybourne, and Newbold (1998), West and McCracken (1998), Harvey and Newbold (2000), Corradi and Swanson (2002), van Dijk and Franses (2003), Hyndman and Koehler (2006), and Clark and West (2007), among others. Given the popularity of Diebold and Mariano's (1995) test, we perform a simple Monte Carlo simulation

Table 10.1 Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the LL model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy (large rejection frequencies indicate that the model in the respective column heading has improved predictive accuracy over the LL model).

δ	LC	$P=1$	$P=2$	$P=3$	$P=4$	$P=5$
$n=200$, $n_2=5$, $S=1000$						
0.0	0.005	0.187	0.019	0.008	0.012	0.006
0.2	0.012	0.001	0.723	0.497	0.270	0.131
0.4	0.026	0.000	0.890	0.837	0.750	0.615
$n=200$, $n_2=10$, $S=1000$						
0.0	0.004	0.225	0.017	0.011	0.007	0.006
0.2	0.018	0.000	0.716	0.524	0.308	0.157
0.4	0.046	0.000	0.914	0.869	0.813	0.657
$n=200$, $n_2=25$, $S=1000$						
0.0	0.005	0.298	0.034	0.013	0.007	0.007
0.2	0.016	0.001	0.787	0.614	0.377	0.155
0.4	0.042	0.000	0.949	0.915	0.855	0.701
$n=200$, $n_2=50$, $S=1000$						
0.0	0.005	0.399	0.046	0.014	0.004	0.007
0.2	0.019	0.001	0.850	0.633	0.337	0.157
0.4	0.066	0.000	0.976	0.953	0.883	0.763
$n=200$, $n_2=100$, $S=1000$						
0.0	0.012	0.537	0.064	0.039	0.022	0.011
0.2	0.025	0.033	0.909	0.619	0.267	0.086
0.4	0.066	0.002	1.000	0.988	0.936	0.715
$n=200$, $n_2=150$, $S=1000$						
0.0	0.036	0.675	0.135	0.071	0.032	0.017
0.2	0.069	0.243	0.873	0.374	0.115	0.046
0.4	0.149	0.045	0.999	0.996	0.744	0.188

similar to that presented in Section 10.3.1. but with stationary time-series models as opposed to cross-section ones.

We generate data from an AR(2) model given by

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t, \tag{10.8}$$

Table 10.2 Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the LL model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy (large rejection frequencies indicate that the model in the respective column heading has improved predictive accuracy over the LL model).

δ	LC	$P = 1$	$P = 2$	$P = 3$	$P = 4$	$P = 5$
$n = 200$, $n_2 = 5$, $S = 10{,}000$						
0.0	0.013	0.296	0.049	0.021	0.008	0.010
0.2	0.031	0.000	0.763	0.593	0.354	0.224
0.4	0.051	0.000	0.936	0.903	0.826	0.718
$n = 200$, $n_2 = 10$, $S = 10{,}000$						
0.0	0.005	0.328	0.039	0.029	0.022	0.012
0.2	0.025	0.000	0.764	0.614	0.381	0.214
0.4	0.046	0.000	0.936	0.902	0.845	0.765
$n = 200$, $n_2 = 25$, $S = 10{,}000$						
0.0	0.003	0.356	0.026	0.023	0.005	0.000
0.2	0.016	0.000	0.836	0.640	0.399	0.198
0.4	0.060	0.000	0.959	0.930	0.889	0.769
$n = 200$, $n_2 = 50$, $S = 10{,}000$						
0.0	0.023	0.424	0.083	0.025	0.010	0.013
0.2	0.017	0.005	0.880	0.694	0.376	0.187
0.4	0.066	0.000	0.990	0.975	0.900	0.748
$n = 200$, $n_2 = 100$, $S = 10{,}000$						
0.0	0.017	0.575	0.103	0.045	0.028	0.015
0.2	0.050	0.045	0.919	0.642	0.311	0.113
0.4	0.094	0.002	1.000	0.993	0.960	0.720
$n = 200$, $n_2 = 150$, $S = 10{,}000$						
0.0	0.022	0.721	0.105	0.044	0.024	0.011
0.2	0.082	0.266	0.881	0.419	0.122	0.027
0.4	0.198	0.062	1.000	1.000	0.775	0.225

where $\rho_1 = 0.9$ for all simulations, and ρ_2 varies from 0 (an AR(1) model) to -0.8 in increments of 0.4 and ε_t is $N(0, 1)$. For all simulations we conduct $M = 1000$ Monte Carlo replications using $S = 10{,}000$ sample splits for our revealed performance approach. We use sample sizes of $n = 200$ and $n = 400$ holding out the last $n_2 = 5$, 10, 25, or 50 observations of each resample for generating forecasts and restrict

Table 10.3 Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the AR(1) model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy (large rejection frequencies indicate that the model in the respective column heading has improved predictive accuracy over the AR(1) model).

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
$n = 200$, $n_2 = 5$, $S = 10{,}000$									
0.0	0.130	0.015	0.038	0.041	0.005	0.006	0.262	0.000	0.000
−0.4	0.290	0.083	0.295	0.094	0.022	0.102	0.981	0.647	0.983
−0.8	0.606	0.663	0.582	0.363	0.321	0.243	1.000	1.000	1.000
$n = 200$, $n_2 = 10$, $S = 10{,}000$									
0.0	0.080	0.009	0.014	0.042	0.006	0.008	0.149	0.000	0.000
−0.4	0.220	0.096	0.237	0.147	0.046	0.163	0.959	0.603	0.956
−0.8	0.633	0.707	0.620	0.558	0.586	0.531	0.993	0.993	0.993
$n = 200$, $n_2 = 25$, $S = 10{,}000$									
0.0	0.055	0.001	0.006	0.050	0.001	0.004	0.089	0.000	0.000
−0.4	0.306	0.078	0.299	0.276	0.062	0.263	0.944	0.555	0.937
−0.8	0.839	0.813	0.761	0.825	0.792	0.745	0.995	0.995	0.995
$n = 200$, $n_2 = 50$, $S = 10{,}000$									
0.0	0.042	0.000	0.001	0.038	0.000	0.000	0.051	0.000	0.000
−0.4	0.464	0.071	0.441	0.441	0.056	0.418	0.944	0.491	0.930
−0.8	0.979	0.915	0.913	0.979	0.909	0.909	0.996	0.996	0.996

attention to 1-step ahead forecasts.[16] When $\rho_2 = 0$ we can determine the extent to which our test predicts worse than or equivalent accuracy of the forecasts, while when $\rho_2 \neq 0$ we can assess how often our method determines that an AR(2) model predicts better than an AR(1) when indeed it should. We also compare the AR(1) to MA(1) and MA(2) specifications by way of comparison. We compare our results with the DM and MDM test, noting that the DM test has a tendency to over-reject when using k-step ahead forecasting, hence our inclusion of the size-corrected MDM results. We report empirical rejection frequencies at the 5% level in Tables 10.3 and 10.4.

We direct the interested reader to Harvey et al. (1998),[17] who report on the formal size and power properties of the DM and MDM tests for a variety of scenarios with a range of k-step ahead forecasts. All three approaches increase in power as n_2 increases, as expected; however, the RP test rejection frequencies approach one more quickly as $|\rho_2|$ increases, suggesting that smaller hold-out samples are required in order to

Table 10.4 Each entry represents rejection frequencies for a one-sided test at the 5% level of the hypothesis that the AR(1) model has predictive accuracy better than or equal to that for each model in the respective column heading, rejecting when the model in the respective column heading has improved predictive accuracy (large rejection frequencies indicate that the model in the respective column heading has improved predictive accuracy over the AR(1) model).

	DM Test			MDM Test			RP Test		
ρ_2	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)	AR(2)	MA(1)	MA(2)
				$n = 400$, $n_2 = 5$, $S = 10{,}000$					
0.0	0.108	0.018	0.028	0.018	0.003	0.008	0.269	0.000	0.000
−0.4	0.269	0.096	0.273	0.120	0.017	0.100	0.979	0.572	0.979
−0.8	0.570	0.545	0.545	0.335	0.250	0.190	1.000	1.000	1.000
				$n = 400$, $n_2 = 10$, $S = 10{,}000$					
0.0	0.051	0.015	0.019	0.034	0.012	0.010	0.167	0.000	0.000
−0.4	0.201	0.076	0.231	0.140	0.048	0.159	0.974	0.571	0.974
−0.8	0.619	0.675	0.582	0.552	0.557	0.469	0.995	0.995	0.995
				$n = 400$, $n_2 = 25$, $S = 10{,}000$					
0.0	0.054	0.000	0.002	0.045	0.000	0.002	0.130	0.000	0.000
−0.4	0.283	0.066	0.303	0.250	0.055	0.268	0.960	0.519	0.958
−0.8	0.820	0.791	0.762	0.791	0.767	0.743	0.990	0.990	0.990
				$n = 400$, $n_2 = 50$, $S = 10{,}000$					
0.0	0.062	0.000	0.002	0.055	0.000	0.002	0.092	0.000	0.000
−0.4	0.444	0.067	0.450	0.423	0.059	0.428	0.957	0.556	0.948
−0.8	0.956	0.912	0.912	0.952	0.912	0.903	0.996	0.996	0.996

discriminate between models in terms of their predictive accuracy. This feature of our approach overcomes one known drawback of the MDM and related tests, namely, the need to have a sufficiently large hold-out sample in order for the test to have power. Lastly, comparing the RP and MDM test, one will see immediately that as $|\rho_2|$ increases, our rejection frequencies approach one at a faster rate for the RP than for the MDM test, indicating that this approach is successful at detecting gains in predictive accuracy outside of an i.i.d. setting.

We note that the ability to choose one's loss function when using our approach may be appealing to practitioners. For instance, if interest lies in penalizing more heavily over or underprediction, the use of asymmetric loss functions may be of interest (LINEX, for example; Chang and Hung 2007). See Efron (1983, 1986) for more on issues related to prediction rules and apparent error in relation to cross-validation and bootstrapping.

10.4. Empirical Illustrations

10.4.1. Application to Wooldridge's Wage1 Data

For what follows, we consider an application that involves multiple regression analysis with both qualitative and real-valued regressors. This example is taken from Wooldridge (2003, p. 226).

We consider modeling an hourly wage equation for which the dependent variable is log(wage) (lwage) while the explanatory variables include three continuous variables, namely educ (years of education), exper (the number of years of potential experience), and tenure (the number of years with their current employer) along with two qualitative variables, female ("Female"/"Male") and married ("Married"/"Notmarried"). For this example there are $n = 526$ observations. We use Hurvich, Simonoff, and Tsai's (1998) AIC$_c$ approach for bandwidth selection.

We first test a parametric model that is linear in all variables for correct parametric specification using Ramsey's (1969) RESET test for functional form via the reset function in the R package lmtest (Zeileis and Hothorn (2002)).[18] We obtain a P-value of 0.0005104 and reject the null of correct specification. We then estimate a nonparametric LL model and test the null that the nonparametric model and a parametric model that is linear in all variables have equal ASPE versus the alternative that the parametric model has greater ASPE. This yields a P-value of $6.969649e;50$ hence we reject the null and conclude that the nonparametric model has statistically significantly improved performance on independent data and therefore represents a statistical improvement over the rejected parametric specification. If we consider a model that is popular in the literature (quadratic in experience), we again reject the null that the model is correctly specified based on the RESET test (P-value of 0.0009729) and again reject the null that the parametric and nonparametric specifications are equivalent in terms of their ASPE and conclude that the nonparametric specification is preferred (the P-value is $9.416807e - 05$). These results indicate that the proposed methodology can be successfully used in applied settings.

Figures 10.1 and 10.2 present box plots and EDFs of ASPEs for each model along with median and mean values for each.[19] It can be seen from Figure 10.2 that a stochastic dominance relationship exists, again indicating that the nonparametric model is to be preferred on the basis of its performance on independent data.

10.4.1.1. Implications of Nonoptimal Smoothing

It is surprisingly common to encounter practitioners applying kernel methods using nonoptimal rules of thumb for bandwidth choice, and in this section we briefly examine the issue of nonoptimal smoothing for the method proposed herein. We consider the two parametric models considered above, namely, one that is linear in all variables and one reflecting the popular specification that is quadratic in experience.

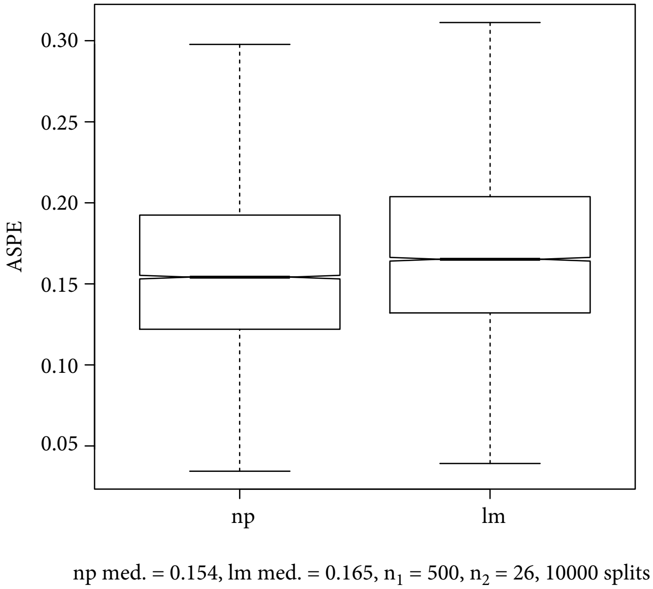


FIGURE 10.1 Box plots of the ASPE for the $S = 10{,}000$ splits of the data for the wage1 data set. Median values for each model appear below the figure.

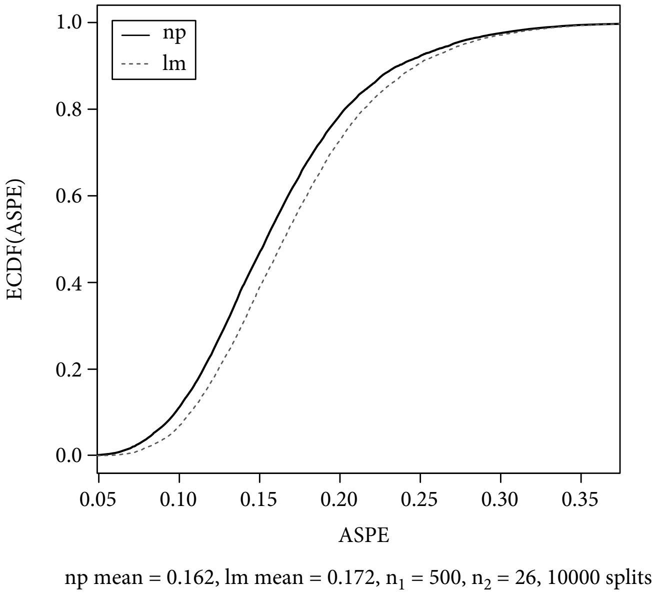


FIGURE 10.2 EDFs of ASPE for the $S = 10{,}000$ splits of the data for the wage1 data set. Mean values for each model appear below the figure.

Table 10.5 Estimated Apparent and True Errors for the Nonparametric and Parametric Models for Wooldridge's Wage1 Data.

	Apparent Error	True Error
	Nonparametric Model	
Smoothing		
Undersmoothed	0.1193826	0.1685334
AIC$_c$	0.1371530	0.1605222
Oversmoothed	0.1928813	0.1679134
	Parametric Model	
Experience		
Linear	0.1681791	0.1723598
Quadratic	0.1590070	0.1634519

We report three nonparametric models, one that is optimally smoothed according to Hurvich et al.'s (1998) AIC$_c$ criterion, one that is undersmoothed (25% of the bandwidth values given by the AIC$_c$ criterion), and one that is oversmoothed using the maximum possible bandwidths for all variables (0.5 for the discrete variables and ∞ for the continuous ones). We report the in-sample apparent error given by $\text{ASE} = n^{-1}\sum_{i=1}^{n}(y_i - \hat{g}_{z^n}(x_i))^2$, and the mean estimated true error taken over all sample splits, $S^{-1}\sum_{j=1}^{S}$ APSE$_j$ where $\text{ASPE}_j = n_2^{-1}\sum_{i=n_1+1}^{n}(y_i^* - \hat{g}_{z^{n_1}}(x_i^*))^2, j = 1, 2, \ldots, S$. Results are reported in Table 10.5.

It can be seen from Table 10.5 that the undersmoothed and optimally smoothed apparent errors are indeed overly optimistic as are those for the linear and quadratic parametric models, as expected. Interestingly, the oversmoothed nonparametric model is overly pessimistic. The tests provided in Section 10.4.1. above are tests that the value in column 3 of Table 10.5 for the AIC$_c$ model (0.1605222) is statistically significantly lower than that for the values in column 3 for both the linear (0.1723598) and quadratic (0.1634519) models. Thus, the nonparametric model is 7.4% more efficient than the linear model and 1.8% more efficient that the quadratic model as measured in terms of performance on independent data while the quadratic model is 5.5% more efficient than the linear model.

10.4.2. Application to CPS Data

We consider a classic data set taken from Pagan and Ullah (1999, p. 155), who consider Canadian cross-section wage data consisting of a random sample obtained from the 1971 Canadian Census Public Use (CPS) Tapes for male individuals having

Table 10.6 Ramsey's (1969) RESET Test for Correct Specification of the Parametric Models for the Canadian CPS Data.

P	RESET	df1	df2	P-Value
1	26.2554	2	201	7.406e-11
2	13.1217	2	200	4.42e-06
3	11.34	2	199	2.168e-05
4	2.1999	2	198	0.1135
5	0.8488	2	197	0.4295
6	1.0656	2	196	0.3465
7	1.4937	2	195	0.2271

common education (Grade 13). There are $n = 205$ observations in total, along with two variables; the logarithm of the individuals wage (logwage) and their age (age). The traditional wage equation is typically modeled as a quadratic in age.

For what follows we consider parametric models of the form

$$\log(\text{wage})_i = \beta_0 + \sum_{j=1}^{P} \beta_j \text{age}_i^j + \varepsilon_i.$$

When $P = 1$ we have a simple linear model, $P = 2$ quadratic and so forth. These types of models are ubiquitous in applied data analysis.

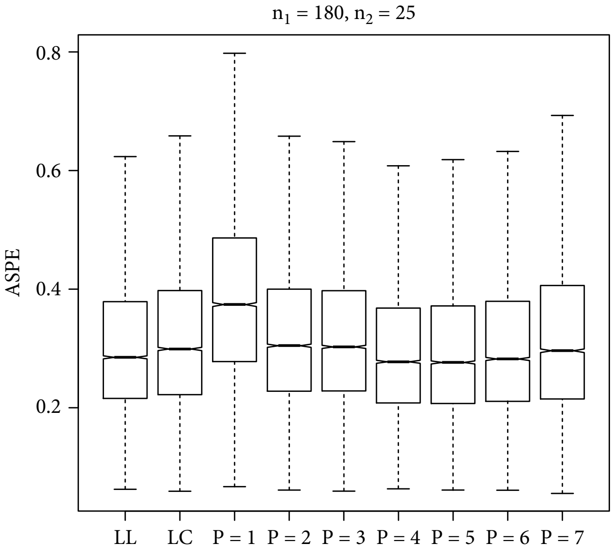
For each model we apply the RESET test. Table 10.6 summarizes the model specification tests for $P = 1$ through $P = 7$.

Models with $P > 3$ pass this specification test. However, this does not imply that the model will outperform other models on independent data drawn from this DGP. The model may be overspecified, and test results could potentially reflect lack of power.

We now consider applying the proposed method to this data set, considering parametric models of order $P = 1$ through 7 along with the LC and LL nonparametric specifications. We present results in the form of box plots and EDFs in Figures 10.3 and 10.4. The box plots and EDFs for $P = 4, 5$ or 6 reveal that these models exhibit visual stochastic dominance relationships with the parametric models for $P = 1, 2, 3$, and 7. This is suggestive that the models $P = 1, 2, 3$ may be underspecified while the model $P = 7$ is perhaps overspecified.

The (paired) t-statistics and P-values for the test that the mean ASPE is equal for each model versus the LL model are given in Table 10.7.

Table 10.7 reveals that the LL specification is preferred to the LC specification on true error grounds. Furthermore, the popular linear and quadratic specifications are dominated by the LL specification as is the less common cubic specification. The quartic and quintic parametric specifications dominate the LL specification as would be expected given the findings of Murphy and Welch (1990). Interestingly, the LL specification



LL: 0.304, LC: 0.318, P = 1: 0.391, P = 2: 0.323, P = 3: 0.321,
P = 4: 0.296, P = 5: 0.298, P = 6: 0.304, P = 7: 0.322

FIGURE 10.3 Box plots of the ASPE for the $S = 10{,}000$ splits of the Canadian CPS data. Median values for each model appear below the figure.

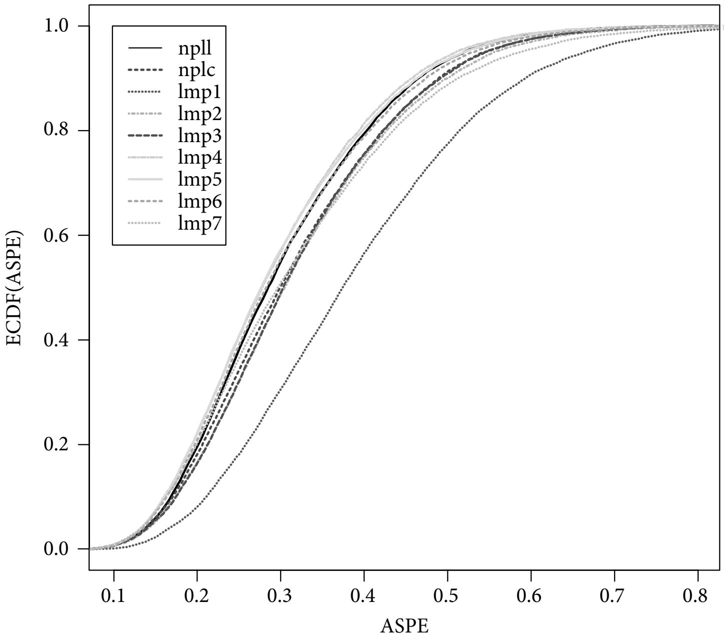


FIGURE 10.4 EDFs of ASPE for the $S = 10{,}000$ splits of the Canadian CPS data.

Table 10.7 RP Test Results for the Canadian CPS Data[a]

Model	t	P-Value
LC	7.847834	2.222817e-15
$P=1$	45.70491	0
$P=2$	10.85293	1.152248e-27
$P=3$	9.682618	1.999341e-22
$P=4$	−4.796251	0.9999992
$P=5$	−3.810738	0.9999305
$P=6$	−0.202236	0.580133
$P=7$	9.840635	4.257431e-23

[a] Small P-values indicate that the nonparametric LL model performs better than the model listed in column 1 according to the true error criterion.

dominates the overspecified parametric model ($P=7$), again underscoring the utility of the proposed approach.

10.4.3. Application to Housing Data

Hedonic analysis of housing data was studied in Anglin and Gençay (1996).[20] They argued that standard parametric models, which passed the RESET test, were outperformed based on overall fit against a partially linear model; two different tests of linearity versus a partially linear model rejected the null hypothesis of correct linear specification. Moreover, to further emphasize the superior performance of the partially linear model, they conducted two separate sampling exercises. First, they looked at price predictions for a "reference" house and plotted the change in price of the home as the number of bedrooms changed. Their results suggested that the price predictions from the semiparametric model were statistically different at the 99% level from the parametric predictions and the parametric model had wider confidence bounds than the partially linear model. Second, Anglin and Gençay (1996) performed a similar hold-out sample exercise as discussed here, however, they do not repeat this exercise a large number of times. From their paper it appears that they did this for *one* sampling of the data using first 10 hold-out homes and then using 20 hold-out homes where the holdout homes were randomly selected.

Recently, Parmeter, Henderson, and Kumbhakar (2007) challenged the partially linear specification of Anglin and Gençay (1996) and advocated for a fully nonparametric approach. Their findings suggested that the partially linear model fails to pass a test of correct specification against a nonparametric alternative, that Anglin and Gençay's (1996) measure of within-sample fit of the partially linear model was overstated, and the inclusion of categorical variables as continuous variables into the unknown function may produce a loss of efficiency; see also Gau, Liu, and Racine (forthcoming). This collection of results provides a useful conduit for examining the revealed performance of the parametric specification of Anglin and Gençay (1996, Table III), the partially linear specification of Anglin and Gençay (1996),[21] and the fully nonparametric specification of Parmeter et al. (2007).[22]

Formally, we model a hedonic price equation where our dependent variable is the logarithm of the sale price of the house while the explanatory variables include six nominal categorical variables, namely if the house is located in a preferential area in Windsor, if the house has air conditioning, if the house has gas heated water, if the house has a fully finished basement, if the house has a recreational room, and if the house has a driveway; four ordinal categorical variables, namely the number of garage places, the number of bedrooms, the number of full bathrooms, and the number of stories of the house; and a single continuous variable, namely the logarithm of the lot size of the house ($\ln(lot)$). There are a total of $n = 546$ observations for this data. All bandwidths are selected using the AIC_c criterion.[23]

Our three models are:

$$\ln(sell) = \gamma_{cat} z_{cat} + \gamma_{ord} z_{ord} + \beta \ \ln(lot) + \varepsilon_1, \tag{10.9}$$

$$\ln(sell) = \gamma_{cat} z_{cat} + g_{AG}(z_{ord}, \ \ln(lot)) + \varepsilon_2, \tag{10.10}$$

$$\ln(sell) = g_{PHK}(z_{cat}, z_{ord}, \ \ln(lot)) + \varepsilon_3, \tag{10.11}$$

where z_{cat} and z_{ord} are the vectors of categorical and ordered variables, respectively, described above.[24] We denote the unknown functions in Eqs. (10.10) and (10.11) by *AG* and *PHK* to refer to the models in Anglin and Gençay (1996) and Parmeter et al. (2007). As noted by Anglin and Gençay (1996), the parametric model is not rejected by a RESET test, suggesting correct specification.[25]

Our test of revealed performance begins with the estimation of all three models and then tests three distinct null hypotheses. First, we test if the nonparametric and linear models have equal ASPE, second we test if the nonparametric and partially linear models have equal ASPE and thirdly, we test if the linear and partially linear models have equal ASPE. For all three tests our alternative hypothesis is that the less general model has a greater ASPE. These tests yield P-values of 1, $2.2e - 16$ and 1, suggesting that the linear model has superior predictive performance over both the appropriately estimated semiparametric model of Anglin and Gençay (1996) and the fully nonparametric model of Parmeter et al. (2007), while the fully nonparametric model has performance that is at least as good as the semiparametric model. This is in direct contrast to Anglin and Gençay's (1996) finding that the semiparametric model provides lower MPSEs for their hold-out samples and is likely a consequence of the fact that they did not repeat their sampling process a large number of times. Additionally, Gençay and Yang (1996) and Bin (2004), also in a hedonic setting, compare semiparametric out-of-sample fits against parametric counterparts using only *one* sample.

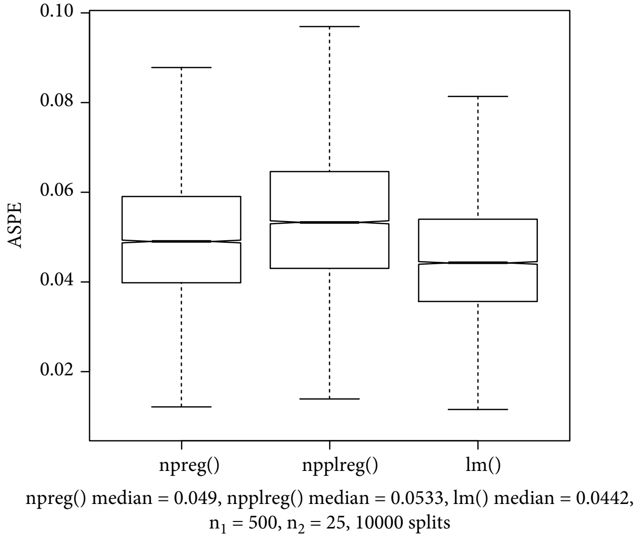


FIGURE 10.5 Box plots of the ASPE for the $S = 10{,}000$ splits of the housing data. Median values for each model appear below the figure.

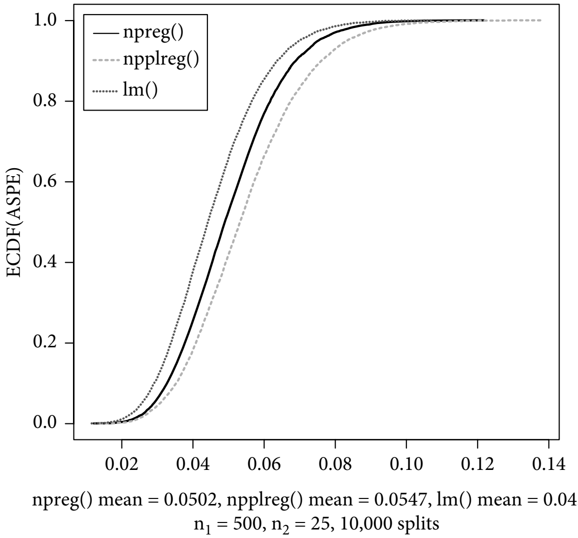


FIGURE 10.6 EDFs of ASPE for the $S = 10{,}000$ splits of the housing data. Mean values for each model appear below the figure.

These setups are entirely incorrect for assessing if one model produces substantially better out-of-sample predictions than another.

Figures 10.5 and 10.6 present box plots and EDFs of ASPEs for each of the three models along with median and mean values for each. It can be seen from Figure 10.6

that a stochastic dominance relationship exists between the linear model (lm()) and both the nonparametric and partially linear models (npreg() and npplreg()), again indicating that the linear model is to be preferred on the basis of its performance on independent data. Figure 10.5 is not suggestive of a stochastic dominance relationship between the linear model and the nonparametric model, whereas the plots of the EDFs in Figure 10.6 readily reveal that the parametric model dominates both the nonparametric and partly linear model, suggesting the use of both plots when assessing the performance of two competing models.

10.4.4. Application to Economic Growth Data

Recent studies by Maasoumi, Racine, and Stengos (2007) and Henderson, Papageorgiou, and Parmeter (2012) have focused on fully nonparametric estimation of "Barro regressions" (see Durlauf, Johnson, and Temple 2005) and argue that standard linear models of economic growth cannot adequately capture the nonlinearities that are most likely present in the underlying growth process. While both papers have soundly rejected basic linear specifications as well as several sophisticated parametric models, it is not evident that the nonparametric model explains the growth data any better than a parametric model.

For this example we use the data set "oecdpanel" available in the np package (Hayfield and Racine 2008) in R (R Core Team, 2012). This panel covers seven 5-year intervals beginning in 1960 for 86 countries.[26] Our dependent variable is the growth rate of real GDP per capita over each of the 5-year intervals, while our explanatory variables include an indicator if the country belongs to the OECD, an ordered variable indicating the year, and the traditional, continuous "Solow" variables: the initial real per capita GDP, the average annual population growth over the 5-year interval, the average investment-GDP ratio over the 5-year period, and the average secondary school enrolment rate for the 5-year period. This is the same data used in Liu and Stengos (1999) and Maasoumi et al. (2007).

We compare the baseline, linear specification, and a model that includes higher-order terms in initial GDP and human capital to a LL nonparametric model with bandwidths selected via AIC$_c$. The baseline linear model is rejected for correct specification using a RESET test (P-value = 0.03983), but the higher-order model cannot be rejected using a RESET test (P-value = 0.1551). However, this test result could be due to power problems related to overspecification with the inclusion of the additional quadratic, cubic, and quartic terms. Using the consistent model specification test of Hsiao, Li, and Racine (2007), the higher-order parametric model is rejected for correct specification with a P-value of 4.07087e-06. The question remains, however, Does the nonparametric model predict growth any better than this higher-order model?

Our test results, box plots, and EDFs all suggest that the nonparametric model significantly outperforms both the baseline "Barro" regression model and the

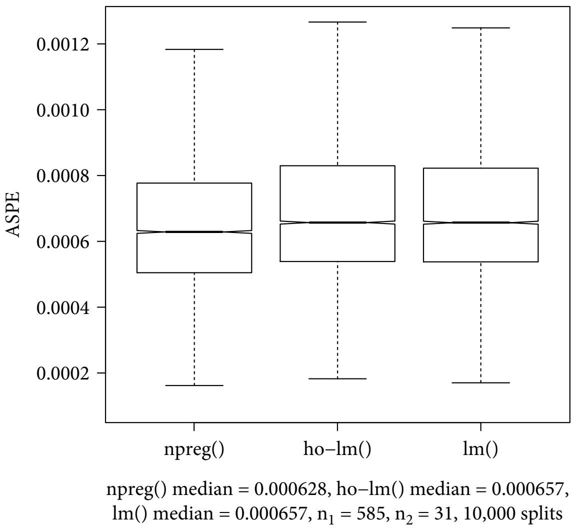


FIGURE 10.7 Box plots of the ASPE for the $S = 10{,}000$ splits of the oecd panel data. Median values for each model appear below the figure.

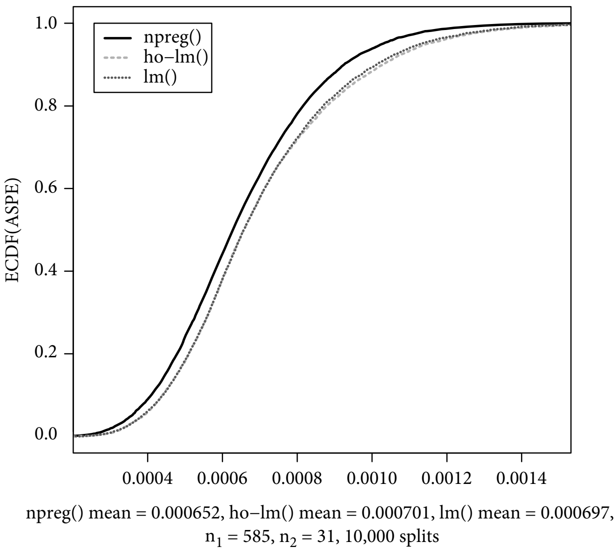


FIGURE 10.8 EDFs of ASPE for the $S = 10{,}000$ splits of the oecd panel data. Mean values for each model appear below the figure.

higher-order parametric model presented in Maasoumi et al. (2007). Our P-values for tests of equality between either the baseline linear model or the higher-order linear model and the LL nonparametric model are $3.475388e - 06$ and $1.542881e - 07$, respectively. This is suggestive that neither parametric model is revealing superior performance to the nonparametric model, which corroborates the findings of Maasoumi et al. (2007) and Henderson et al. (2012).

Figures 10.7 and 10.8 present box plots and EDFs of ASPEs for each of the three models along with median and mean values for each. It can be seen from Figure 10.8 that a stochastic dominance relationship exists between the nonparametric model (npreg()) and both of the linear models (lm() and ho-lm()), again indicating that the nonparametric model is to be preferred on the basis of its performance on independent draws from the data. What is interesting is that in terms of revealed performance given through the EDFs, the higher-order linear model *does not* exhibit any stochastic dominance over the standard "Barro" regression, suggesting that the hypothesized nonlinearities present are more complex than simple power terms of the individual covariates. Interestingly, Henderson et al. (2012) have uncovered marginal effects of the covariates consistent more so with interactions between covariates than with higher-order terms of individual covariates.

10.4.5. Application to the Federal Funds Interest Rate

In this section we consider modeling the monthly U.S. federal funds interest rate. The data come from the Board of Governors of the Federal Reserve. These data are a time series of monthly observations on the interest rate from July 1954 to September 2012 (a total of $n = 699$ observations). Figure 10.9 displays the raw interest rate series as well as autocovariance and partial autocorrelation functions for 20 lags. Figure 10.10 presents the time series of the first differenced series as well as autocovariance and partial autocorrelation plots. A test for stationarity reveals that the series is indeed nonstationary. First differencing produces a stationary series.

We note that the large spike of our differenced data for the first autocorrelation (lower right panel) suggests that an MA(1) process may be present. However, the presence of positive and significant autocorrelations past lag 1 and the regular pattern in the distant autocorrelations suggests that a more complex DGP may be present. Also, the partial autocorrelations (lower left panel), which have a positive and significant spike for lag 1 and a negative and significant spike for lag 2, would rule out the use of an AR(1) model but could be consistent with an AR(2) model.[27] Fitting the best ARIMA process to the first differenced data suggests that an MA(1) process with a seasonal component is appropriate.[28] For this exercise we use a hold-out sample of 24 observations, which corresponds to two years. The automatic block length selection of Politis and White (2004) suggests we use a random block length of 44 when resampling.

Our box plot and EDF comparison plots appear in Figures 10.11 and 10.12. Note that even though the MA(1) process was found to best fit the data within the ARIMA family, both the AR(1) and the ARMA(1,1) models appear to deliver superior predictions. This is clear from both the box plots and the EDFs of ASPE. Both our RP test and the DM test suggest that there is no difference in forecast prediction errors across the AR(1)

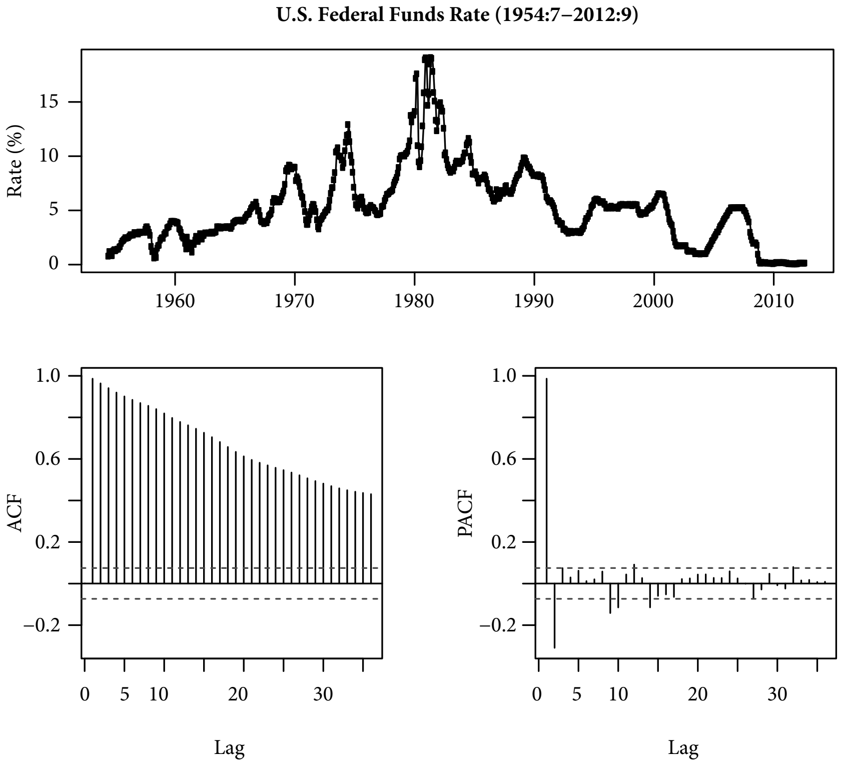


FIGURE 10.9 Time plot, autocovariance plot, and partial autocorrelation plot for the federal funds interest rate.

and the ARMA(1,1) specifications, while both dominate the MA(1) model. Thus, even though a close look at the partial autocorrelation plot reveals that an AR(1) model may be inappropriate, if the focus is on out-of-sample forecasting, then the AR(1) model does a better job than the MA(1) and one not significantly different from the ARMA(1,1).

10.5. Conclusion

In this chapter we propose a general methodology for assessing the predictive performance of competing approximate models based on resampling techniques. The approach involves taking repeated hold-out samples (appropriately constructed) from the data at hand to create an estimate of the expected true error of a model and then using this as the basis for a test. A model possessing lower expected true error than another is closer to the underlying DGP according to the specified loss function and is therefore to be preferred. Our approach allows practitioners to compare a broad range

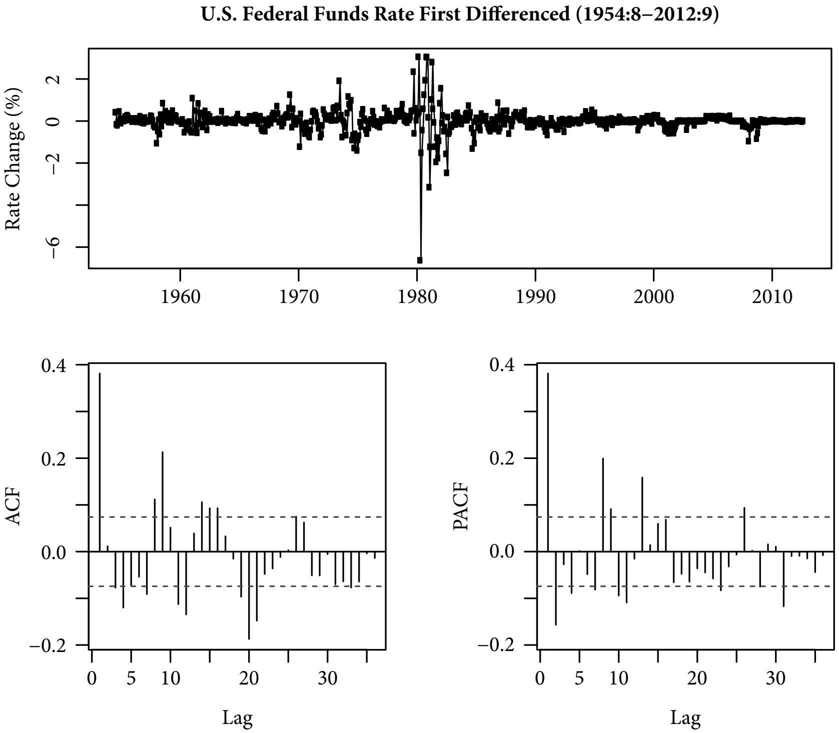


FIGURE 10.10 Time plot, autocovariance plot, and partial autocorrelation plot for the first differenced federal funds interest rate time series.

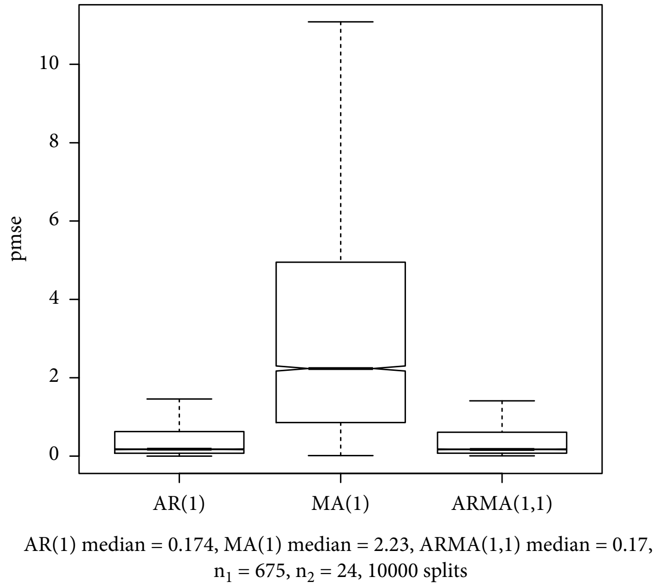


FIGURE 10.11 Box plots of the ASPE for the $S = 1000$ splits of the first differenced federal funds rate time-series data. Median values for each model appear below the figure.

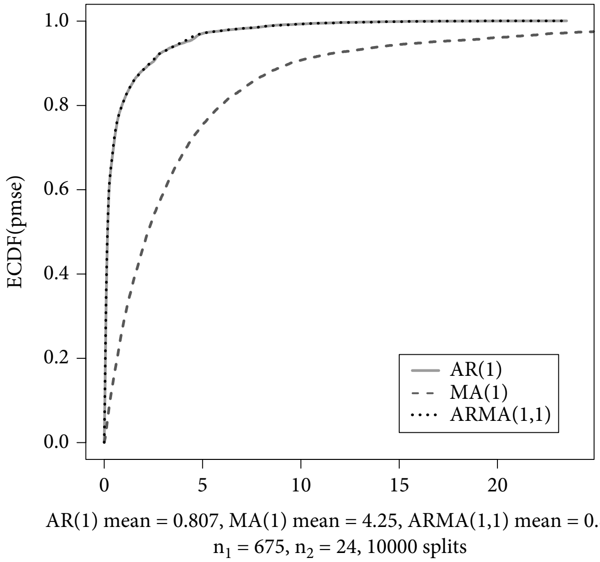


FIGURE 10.12 EDFs of ASPE for the $S = 1000$ splits of the first differenced federal funds rate time-series data. Mean values for each model appear below the figure.

of modeling alternatives and is not limited to the regression-based examples provided herein. The approach can be used to determine whether or not a more flexible model offers any gains in terms of expected performance than a less complex model and provides an alternative avenue for direct comparison of parametric and nonparametric regression surfaces (e.g, Härdle and Marron (1990), Härdle and Mammen (1993)).

We present both simulated and empirical evidence underscoring the utility of the proposed method in dependent data settings. Our simulation results indicate that, relative to popular time-series tests, our RP test is capable of delivering substantial gains when assessing predictive accuracy. The empirical examples highlight the ease with which the method can be deployed across a range of application domains (cross-section, panel, and time series). We also present telling empirical evidence as to how overspecified parametric and nonparametric models may not always provide the most accurate approximations to the underlying DGP. Thus, our method can be used as an auxiliary tool for assessing the accuracy of a selected model, thereby enhancing any insights one might otherwise glean from empirical exercises.

Fruitful extensions of this approach could include its use in nonregression settings such as the modeling of counts, survival times, or even probabilities. We leave rigorous analysis on optimal selection of the hold-out sample size and its impact on the resulting test statistic for future research. One could also trivially extend our testing idea to include formal tests of stochastic dominance as opposed to the visual arguments advocated in the chapter, though we leave this an an exercise for the interested reader.

Notes

[†] We would like to thank (but not implicate) Shahram Amini, Richard Ashley, Chris Bennett, Robert Hyndman, Nicolai Kuminoff, Esfandiar Maasoumi, Andrew Patton, Dimitris Politis, Silvia Gonçalves, and Zhiyuan Zheng for their thoughtful comments and suggestions, and we also acknowledge helpful feedback received at CESG 2012. Jeffrey S. Racine would like to gratefully acknowledge support from Natural Sciences and Engineering Research Council of Canada (NSERC.www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca). We would like to dedicate this chapter in memorian to Halbert L. White, Jr., whose insight insight and feedback had a profound effect on our understanding of the problem.

[1] Alternatively, as White (2000, p. 1097) discusses, resorting to extensive specification searches runs the risk that the observed good performance of a model is not the result of superior fit but rather luck, and he labels such practices "data snooping."

[2] Corradi and Swanson (2002, p. 356) underscore the importance of this issue when they discuss "...whether simple linear models (e.g., ARIMA models) provide forecasts which are (at least) as accurate as more sophisticated nonlinear models. If this were shown to be the case, there would be no point in using nonlinear models for out-of-sample prediction, even if the linear models could be shown to be incorrectly specified, say based on the application of *in-sample* nonlinearity tests..." (our italics).

[3] See Hansen (2005, pp. 62–63) for an eloquent discussion of this issue.

[4] This allows us to address how much more accurate one method is compared to another *on average* with respect to the chosen loss function. Indeed, this is in direct agreement with Goodwin (2007, p. 334): "[...] when comparing the accuracy of forecasting methods [...] The interesting questions are, how much more accurate is method A than method B, and is this difference of practical significance?" Our approach will allow a researcher to tackle both of these questions in a simple and easily implementable framework, though we take a broader view by considering "out-of-sample prediction" in cross-section settings and "forecasting" in time series ones.

[5] Readers familiar with Diebold and Mariano's (1995) test for predictive accuracy will immediately recognize this strategy.

[6] Clearly, for (strictly) stationary dependent processes, we cannot use sample splitting directly, however, we can use resampling methods that are appropriate for such processes. When each resample is the outcome of an appropriate bootstrap methodology, it will mimic dependence present in the original series and can itself be split; see Politis and Romano (1992) by way of example.

[7] For instance, we might perform a test of the hypothesis that the mean ASPE ("expected true error") for the $S = 10{,}000$ splits is equal (less than or equal to) for two models against a one-sided alternative (greater than) in order to maximize power. Or, one could test for stochastic dominance of one distribution over the other.

[8] Additionally, while we do not discuss it further, our test is not restricted to the case where the dependent variable is identical across models. One could use the insight of Wooldridge (1994) to transform the predictions from one model to that of another where the dependent variable was transformed (monotonically).

[9] Efron (1982, p. 51) considers estimation of "expected excess error," while we instead consider estimation of "expected true error."

10. Of course, we recognize that a model based on the true DGP may not deliver the best out-of-sample prediction due to parameter estimation error, so we highlight the fact that this is in the repeated sampling sense, hence the wording "expected to lie closest."
11. The pairing will be seen to arise from potential correlation between model predictions among competing models.
12. A "scaling factor" refers to the unknown constant c in the formula for the optimal bandwidth—for example, $h_{opt} = cn^{-1/(4+p)}$, where p is the number of continuous predictors for a kernel regression model using a second-order kernel function. Cross-validation can be thought of as a method for estimating the unknown constant c, where c is independent of the sample size n. This constant can then be rescaled for samples of differing size drawn from the same DGP, thereby ensuring that the same degree of smoothing is applied to the full sample and the subsample (see Racine (1993)). The rationale for so doing is as follows. Think of estimating a univariate density function where the data represent independent draws from the $N(0,1)$ distribution. The optimal bandwidth in this case is known to be $h_{opt} = 1.059n^{-1/5}$. If $n = 200$, then $h_{opt} = 0.3670$; while if $n = 100$, then $h_{opt} = 0.4215$. Cross-validation delivers an estimate of h_{opt} for a sample of size n (i.e., $\hat{h} = \hat{c}n^{-1/5}$), while it can be shown that $(\hat{h} - h_{opt})/h_{opt} \to 1$ asymptotically (see Stone (1984)). If you don't rescale the cross-validated bandwidth for subsamples of size $n_1 < n$ (i.e., adjust $\hat{h}$ when n falls to n_1), then you are in fact doing a different amount of smoothing on subsamples of size $n_1 < n$ (i.e., $h = 0.3670$ will undersmooth when $n_1 < 200$, so the estimate based on $h = 0.3670$ and $n_1 < 200$ will be overly variable). But, by using $\hat{c}$ corresponding to the cross-validated bandwidth for the full sample, $\hat{h}$, we can ensure that the same degree of smoothing is applied to the subsamples of size $n_1 < n$ and the full sample of size n. This keeps the baseline nonparametric model fixed at that for the full sample, in the same way that we hold the functional form of the parametric model fixed at that for the full sample.
13. See Patton, Politis, and White (2009) for a correction to several of the results in Politis and White (2004).
14. Our choice of the stationary block bootstrap is for expositional purposes. In practice we recommend that the user employ a bootstrap appropriate for the type of dependence apparent in the data. For example, additional types of bootstraps are the Markov conditional bootstrap (Horowitz, 2004), the circular bootstrap (Politis and Romano, 1992), and the sieve bootstrap (Bühlmann, 1997); see Lahiri (2003) for an up-to-date and detailed coverage of available block resampling schemes. One can easily implement a variety of block bootstrap procedures by using the tsboot() command available in the boot package (Canty and Ripley, 2012) in R (R Core Team, 2012).
15. See the review by De Gooijer and Hyndman (2006) for a thorough, up-to-date survey and bibliography on the subject.
16. For each Monte Carlo simulation, the initial data generated are passed through the automatic block length selection mechanism of Politis and White (2004) to determine the optimal block length. This block length is then used for generating each of the S artificial series in order to generate each of the S splits of the data. We do not investigate the behavior of our test with regard to alternate block length selection schemes, and we leave this for future investigation.
17. See also Harvey and Newbold (2000), Meade (2002), and van Dijk and Franses (2003).
18. We use default settings hence powers of 2 and 3 of the fitted model are employed.

19. A box-and-whisker plot (sometimes called simply a "box plot") is a histogram-like method of displaying data, invented by J. Tukey. To create a box-and-whisker plot, draw a box with ends at the quartiles Q_1 and Q_3. Draw the statistical median M as a horizontal line in the box. Now extend the "whiskers" to the farthest points that are not outliers (i.e., that are within 3/2 times the interquartile range of Q_1 and Q_3). Then, for every point more than 3/2 times the interquartile range from the end of a box, draw a dot.
20. The data from their paper is available on the JAE data archives webpage or can be found in the Ecdat package (Croissant 2011) in R (R Core Team 2012) under the name "housing."
21. We do not estimate the partially linear model as it appears in Anglin and Gençay (1996) since Parmeter et al. (2007) were unable to exactly replicate their results and Anglin and Gençay's (1996) handling of ordered discrete variables as continuous is erroneous given the current practice of using generalized kernel estimation.
22. See Haupt, Schnurbus, and Tschernig (2010) for further discussion of this setting.
23. See Gau et al. (forthcoming) for more on bandwidth selection in partially linear models.
24. We note that although the number of garage places is an ordered variable, Anglin and Gençay (1996) did not include it in the unknown function in their partially linear setup. To be consistent with their modeling approach, we follow suit and have the number of garage places enter in the linear portion of (10.10).
25. Anglin and Gençay (1996, p. 638) do note, however, that their benchmark model is rejected using the specification test of Wooldridge (1992). Also, we use the model estimated in Table III of Anglin and Gençay (1996) since this model has a higher $\bar{R}^2$; and as they note (Anglin and Gençay 1996, p. 638) the performance of this model is not substantially different from their benchmark model.
26. See Liu and Stengos (1999, Table 1) for a list of countries in the data set.
27. The positive and significant partial autocorrelations at lag 8, 13, and 26 are difficult to interpret.
28. This was done using the entire data set with the auto.arima() function in the forecast package (Hyndman, Razbash, and Schmidt, 2012) in R (R Core Team, 2012).

References

Anglin, P. M., and R. Gençay. 1996. "Semiparametric Estimation of a Hedonic Price Function." *Journal of Applied Econometrics*, **11**, pp. 633–648.

Ashley, R. 2003. "Statistically Significant Forecasting Improvements: How Much Out-of-Sample Data Is Likely Necessary?" *International Journal of Forecasting*, **19**, pp. 229–239.

Ashley, R. A. 1998. "A New Technique for Postsample Model Selection and Validation." *Journal of Economic Dynamics and Control*, **22**, pp. 647–665.

Ashley, R. A., C. W. J. Granger, and R. Schmalensee. 1980. "Advertising and Aggregate Consumption: An Analysis of Causality." *Econometrica*, **48**, pp. 1149–1167.

Bauer, D. F. 1972. "Constructing Confidence Sets Using Rank Statistics." *Journal of the American Statistical Association*, **67**, pp. 687–690.

Bin, O. 2004. "A Prediction Comparison of Housing Sales Prices by Parametric Versus Semi-parametric Regressions." *Journal of Housing Economics*, **13**, pp. 68–84.

Bühlmann, P. 1997. "Sieve Bootstrap for Time Series." *Bernoulli*, **3**, pp. 123–148.

Canty, A., and B. Ripley. 2012. *boot: Bootstrap R (S-Plus) Functions.* R package version 1.3-7. URL: http://www.r-project.org

Chang, Y.-C., and W.-L. Hung. 2007. "LINEX Loss Functions with Applications to Determining the Optimum Process Parameters." *Quality and Quantity*, **41**(2), pp. 291–301.

Clark, T. E., and K. D. West. 2007. "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models." *Journal of Econometrics*, **138**, pp. 391–311.

Corradi, V., and N. R. Swanson. 2002. "A Consistent Test for Nonlinear Out of Sample Predictive Accuracy." *Journal of Econometrics*, **110**, pp. 353–381.

Corradi, V., and N. R. Swanson. 2004. "Some Recent Developments in Predictive Accuracy Testing with Nested Models and (Generic) Nonlinear Alternatives." *International Journal of Forecasting*, **20**, pp. 185–199.

Corradi, V., and N. R. Swanson. 2007. "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes." *International Economic Review*, **48**(1), pp. 67–109.

Croissant, Y. 2011. *Ecdat: Data sets for econometrics*. R package version 0.1–6.1. URL: http://CRAN.R-project.org/package=Ecdat

Davidson, R., and J.-Y. Duclos. 2000. "Statistical inference for Stochastic Dominance and for the Measurement of Poverty and Inequality." *Econometrica*, **68**, pp. 1435–1464.

Davidson, R., and J. G. MacKinnon. 2002. "Bootstrap J Tests of Nonnested Linear Regression Models."*Journal of Econometrics*, **109**, pp. 167–193.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.

De Gooijer, J. G., and R. J. Hyndman. 2006. "25 years of Times Series Forecasting." *International Journal of Forecasting*, **22**, pp. 443–473.

Diebold, F. X., and R. S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business & Economic Statistics*, **13**(3), pp. 253–265.

Durlauf, S. N., P. Johnson, and J. Temple. 2005. "Growth Econometrics." In *Handbook of Economic Growth*, Vol. 1A, P. Aghion and S. N. Durlauf. Amsterdam: North-Holland.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. 1983. "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." *Journal of the American Statistical Association*, **78**(382), pp. 316–331.

Efron, B. 1986. "How Biased Is the Apparent Error Rate of a Prediction Rule." *Journal of the American Statistical Association*, **81**(394), pp. 461–470.

Gau, Q., L. Liu, and J. Racine. 2013. "A Partially Linear Kernel Estimator for Categorical Data." *Econometric Reviews*, forthcoming.

Geisser, S. 1975. "A Predictive Sample Reuse Method with Application." *Journal of the American Statistical Association*, **70**, pp. 320–328.

Gençay, R., and X. Yang. 1996. "A Prediction Comparison of Residential Housing Prices by Parametric Versus Semiparametric Conditional Mean Estimators." *Economics Letters*, **52**, pp. 129–135.

Goodwin, P. 2007. "Should We Be Using Significance Test in Forecasting Research?" *International Journal of Forecasting*, **23**, pp. 333–334.

Granger, C. W. J., and P. Newbold. 1986. *Forecasting Economic Time Series*. San Diego, CA: Academic Press.

Hansen, B. E. 2005. "Challenges for Econometric Model Selection." *Econometric Theory*, **21**, pp. 60–68.

Härdle, W., and E. Mammen. 1993. "Comparing nonparametric versus parametric regression fits." *Annals of Statistics*, **21**(4), pp. 1926–1947.

Härdle, W., and J. S. Marron. 1990. "Semiparametric Comparison of Regression Curves." *Annals of Statistics*, **18**(1), pp. 63–89.

Harvey, D. I., S. J. Leybourne, and P. Newbold. 1997. "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting*, **13**, pp. 281–291.

Harvey, D. I., S. J. Leybourne, and P. Newbold. 1998. "Tests of Forecast Encompassing." *Journal of Business & Economics Statistics*, **16**(2), pp. 254–259.

Harvey, D. I., and P. Newbold. 2000. "Tests for Multiple Forecast Encompassing." *Journal of Applied Econometrics*, **15**, pp. 471–482.

Haupt, H., J. Schnurbus, and R. Tschernig. 2010. "On Nonparametric Estimation of a Hedonic Price Function." *Journal of Applied Econometrics*, **25**, pp. 894–901.

Hayfield, T. and J. S. Racine. 2008. "Nonparametric Econometrics: The np Package." *Journal of Statistical Software*, **27**(5). URL: http://www.jstatsoft.org/v27/i05/

Henderson, D. J., C. Papageorgiou, and C. F. Parmeter. 2012. "Growth Empirics without Parameters." *The Economic Journal*, **122**, pp. 125–154.

Horowitz, J. L. 2004. "Bootstrap Methods for Markov Processes." *Econometrica*, **71**(4), pp. 1049–1082.

Hsiao, C., Q. Li, and J. S. Racine. 2007. "A Consistent Model Specification Test with Mixed Discrete and Continuous Data." *Journal of Econometrics*, **140**, pp. 802–826.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society Series B*, **60**, pp. 271–293.

Hyndman, R. J. and A. B. Koehler, 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting*, **22**, pp. 679–688.

Hyndman, R. J., S. Razbash, and D. Schmidt. 2012. *Forecast: Forecasting Functions for Time Series and Linear Models.* R package version 3.25. URL: http://CRAN.R-project.org/package=forecast

Inoue, A., and L. Kilian. 2004. "In-Sample and Out-of-Sample Tests of Predictability: Which One Should We Use?" *Econometric Reviews*, **23**, pp. 371–402.

Lahiri, S. N. 2003. *Resampling Methods for Dependent Data.* New York: Springer-Verlag.

Leeb, H., and B. M. Pötscher. 2006. "Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators?" *Annals of Statistics*, **34**, pp. 2554–2591.

Li, Q., and J. Racine. 2004. "Cross-Validated Local Linear Nonparametric Regression." *Statistica Sinica*, **14**(2), pp. 485–512.

Liu, Z., and Stengos, T. 1999. "Non-Linearities in Cross Country Growth Regressions: A Semiparametric Approach." *Journal of Applied Econometrics*, **14**, pp. 527–538.

Maasoumi, E., J. S. Racine, and T. Stengos. 2007. "Growth and Convergence: A Profile of Distribution Dynamics and Mobility." *Journal of Econometrics*, **136**, pp. 483–508.

McCracken, M. W. 2000. "Robust Out-of-Sample Prediction." *Journal of Econometrics*, **99**(2), pp. 195–223.

Meade, N. 2002. "A Comparison of the Accuracy of Short Term Foreign Exchange Forecasting Methods." *International Journal of Forecasting*, **18**, pp. 67–83.

Medeiros, M. C., T. Teräsvirta, and G. Rech. 2006. "Building Neural Network Models for Time Series: A Statistical Approach." *Journal of Forecasting*, **25**, pp. 49–75.

Mehta, J. S., and J. Gurland. 1969. "Testing Equality of Means in the Presence of Correlation." *Biometrika*, **56**, pp. 119–126.

Murphy, K. M., and F. Welch. 1990. "Empirical Age-Earnings Profiles." *Journal of Labor Economics*, **8**(2), pp. 202–229.

Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Parmeter, C. F., D. J. Henderson, and S. C. Kumbhakar. 2007. "Nonparametric Estimation of a Hedonic Price Function." *Journal of Applied Econometrics*, **22**, pp. 695–699.

Patton, A., D. N. Politis, and H. White. 2009. "Correction to 'Automatic Block-Length Selection for the Dependent Bootstrap' by D. Politis and H. White." *Econometric Reviews*, **28**(4), pp. 372–375.

Politis, D. N., and J. P. Romano. 1992. A Circular Block-Resampling Procedure for Stationary Data." In *Exploring the Limits of Bootstrap*, eds. R. LePage and R. Billard. New York: John Wiley & Sons, pp. 263–270.

Politis, D. N., and J. P. Romano. 1994. "The Stationary Bootstrap." *Journal of the American Statistical Association*, **89**(428), pp. 1303–1313.

Politis, D. N., and H. White. 2004. "Automatic Block-Length Selection for the Dependent Bootstrap." *Econometric Reviews*, **23**(1), pp. 53–70.

R Core Team. 2012. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org/

Racine, J. S. 1993. "An Efficient Cross–Validation Algorithm for Window Width Selection for Nonparametric Kernel Regression." *Communications in Statistics*, **22**(4), pp. 1107–1114.

Racine, J. S. 2001. "On the Nonlinear Predictability of Stock Returns Using Financial and Economic Variables." *Journal of Business and Economic Statistics*, **19**(3), pp. 380–382.

Ramsey, J. B. 1969. "Tests for Specification Error in Classical Linear Least Squares Regression Analysis." *Journal of the Royal Statistical Society, Series B*, **31**, pp. 350–371.

Shen, X., and Ye, J. 2002. "Model Selection." *Journal of the American Statistical Association*, **97**(457), pp. 210–221.

Stone, C. J. 1974. "Cross-Validatory Choice and Assessment of Statistical Predictions (with discussion)." *Journal of the Royal Statistical Society*, **36**, pp. 111–147.

Stone, C. J. 1984. "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates." *Annals of Statistics*, **12**, pp. 1285–1297.

Swanson, N. R., and H. White. 1997. "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks." *The Review of Economics and Statistics*, **79**, pp. 540–550.

van Dijk, D., and P. H. Franses. 2003. "Selecting a Nonlinear Time Series Model Using Weighted Tests of Equal Forecast Accuracy." *Oxford Bulletin of Economics and Statistics*, **65**, pp. 727–744.

Wahba, G., and S. Wold. 1975. "A Completely Automatic French Curve: Fitting Spline Functions by Cross-Validation." *Communications in Statistics*, **4**, pp. 1–17.

West, K. 1996. "Asymptotic Inference about Predictive Ability." *Econometrica*, **64**, pp. 1067–1084.

West, K. D., and M. W. McCracken. 1998. "Regression-Based Tests of Predictive Ability." *International Economic Review*, **39**(3), pp. 817–840.

White, H. 2000. "A Reality Check for Data Snooping." *Econometrica*, **68**(5), pp. 1097–1126.

White, H. 2001. *Asymptotic Theory for Econometricians*, San Diego, CA: Academic Press.

Wooldridge, J. M. 1992. "A Test of Functional Form Against Nonparametric Alternatives." *Econometric Theory*, **8**, pp. 452–475.

Wooldridge, J. M. 1994. "A Simple Specification Test for the Predictive Ability of Transformation Models." *The Review of Economics and Statistics*, **76**(1), pp. 59–65.

Wooldridge, J. M. 2003. *Introductory Econometrics*, third edition, Mason, OH: Thompson South-Western.

Ye, J. 1998. "On Measuring and Correcting the Effects of Data Mining and Data Selection." *Journal of the American Statistical Association*, **93**(441), pp. 120–131.

Zeileis, A., and Hothorn, T. 2002. "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), pp. 7–10. URL: http://CRAN.R-project.org/doc/Rnews/.

CHAPTER 11

SUPPORT VECTOR MACHINES WITH EVOLUTIONARY MODEL SELECTION FOR DEFAULT PREDICTION

WOLFGANG KARL HÄRDLE, DEDY DWI PRASTYO,
AND CHRISTIAN M. HAFNER

11.1. Default Prediction Methods

Default probability is defined as the probability that a borrower will fail to serve its obligation. Bonds and other tradable debt instruments are the main source of default for most individual and institutional investors. In contrast, loans are the largest and most obvious source of default for banks (Sobehart and Stein 2000).

Default prediction is becoming more and more important for banks, especially in risk management, in order to measure their client's degree of risk. The Basel Committee on Banking Supervision established the borrower's rating as a crucial criterion for minimum capital requirements of banks to minimize their cost of capital and mitigate their own bankruptcy risk (Härdle et al., 2009). Various methods to generate rating have been developed over the last 15 years (Krahnen and Weber, 2001).

There are basically two approaches dealing with default risk analysis: statistical model and market-based model. The statistical model was determined through an empirical analysis of historical data—for example, accounting data. The market-based model, also known as a structural model, uses time series of the company data to predict the probability of default. One of the common market-based approach is derived from an adapted Black–Scholes model (Black and Scholes (1973) and Vassalou and Xing (2004)). However, the most challenging requirement in market-based approach is the knowledge of market values of debt and equity. This precondition is a severe obstacle to using the Merton model (Merton 1974) adequately because it is only satisfied in a minority of cases (Härdle et al., 2009). The idea of Merton's model is that equity and debt could be considered as options on the value of the firm's assets. Unfortunately, long time series of market prices are not available for most companies. Moreover, for companies that are not listed, their market price is unknown. In that case, it is necessary to choose a model that relies on cross-sectional data, financial statements, or accounting data. Sobehart and Stein (2000) developed a hybrid model where the output is based on the relationship between default and financial statement information, market information, ratings (when they exist), and a variant of Merton's contingent claims model expressed as distance to default.

The early studies about default prediction identified the difference between financial ratios of default (insolvent) and nondefault (solvent) firms (Merwin (1942)). Then, discriminant analysis (DA), also known as Z-score, was introduced in default prediction; see Beaver (1966) and Altman (1968) for the univariate and multivariate case, respectively. The model separates defaulting from nondefaulting firms based on the discriminatory power of linear combinations of financial ratios. Afterward, the logit and probit approach replaced the usage of DA during the 1980s (see Martin (1977), Ohlson (1980), Lo (1986) and Platt et al. (1994)). The assumption in DA and logit (or probit) models often fails to meet the reality of observed data. In order to incorporate the conventional linear model and a nonparametric approach, Hwang et al. (2007) developed semiparametric logit model.

If there is evidence for the nonlinear separation mechanism, then the linear separating hyperplane approach is not appropriate. In that case, the Artificial Neural Network (ANN) is a nonparametric nonlinear classification approach that is appropriate to solve the problem. ANN was introduced to predict default in the 1990s (see Tam and Kiang (1992), Wilson and Sharda (1994), and Altman et al. (1994) for details). However, ANN is often criticized to be vulnerable to the multiple minima problem. ANN uses the principle of minimizing empirical risk, the same as in the Ordinary Least Square (OLS) and Maximum Likelihood Estimation (MLE) for linear models, which usually leads to poor classification performance for out-of-sample data (Haykin (1999), Gunn (1998), and Burges (1998)).

In contrast to the case of neural networks, where many local minima usually exist, Support Vector Machines (SVM) training always finds a global solution (Burges 1998). SVM is one of the most promising among nonlinear statistical techniques developed recently and is a state-of-the-art classification method. The idea of SVM was started in the late 1970s by Vapnik (1979), but it was receiving increasing attention after the work in statistical learning theory (Boser et al. (1992), Vapnik (1995) and Vapnik (1998)). The SVM formulation embodies the Structural Risk Minimization (SRM) principle (Shawe-Taylor et al., 1996). At the first stages, SVM has been successfully applied to classify (multivariate) observation (see Blanz et al. (1996), Cortes and Vapnik (1995), Schölkopf et al. (1995, 1996), Burges and Schölkopf (1997), and Osuna et al. (1997a)). Later, SVM has been used in regression prediction and time series forecasting (Müller et al., 1997).

The SVM has been applied to default prediction and typically outperformed the competing models (Härdle and Simar (2012), Härdle et al. (2009, 2011), Zhang and Härdle (2010), and Chen et al. (2011)). One of the important issues in SVM is the parameter optimization, which is also known as model selection. This chapter emphasizes the model selection of SVM for default prediction applied to a CreditReform database. The SVM parameters are optimized by using an evolutionary algorithm, the so-called Genetic Algorithm (GA), introduced by Holland (1975). Some recent papers that deal with GA are Michalewicz (1996), Gen and Cheng (2000), Mitchell (1999), Haupt and Haupt (2004), Sivanandam and Deepa (2008), and Baragona et al. (2011).

In the case of a small percentage of samples belonging to a certain class (label) compared to the other classes, the classification method may tend to classify every sample belong to the majority. This is the case in default and nondefault data sets, therefore such models would be useless in practice. He and Garcia (2009) provide a comprehensive and critical review of the development research in learning from imbalanced data.

Two of the methods to overcome the imbalanced problem are the *down-sampling* and *oversampling* strategies (Härdle et al., 2009). Down-sampling works with bootstrap to select a set of majority class examples such that both the majority and minority classes are balanced. Due to the random sampling of bootstrap, the majority sample might cause the model to have the highest variance. An oversampling scheme could be applied to avoid this unstable model building (Maalouf and Trafalis, 2011). The oversampling method selects a set of samples from the minority class and replicates the procedure such that both majority and minority classes are balanced.

At first glance, the down-sampling and oversampling appear to be functionally equivalent because they both alter the size of the original data set and can actually yield balanced classes. In the case of down-sampling, removing examples from the majority class may cause the classifier to miss important concepts pertaining to the majority class. With regard to oversampling, multiple instances of certain examples become "tied," which leads to overfitting (He and Garcia 2009). Although sampling methods and cost-sensitive learning methods dominate the current research in imbalanced learning, kernel-based learning, (e.g., SVM) have also been pursued. The representative SVMs can provide relatively robust classification results when applied to an imbalanced data set (Japkowicz and Stephen, 2002).

11.2. Quality of Default Prediction

One of the most important issues in classification is the discriminative power of classification methods. In credit scoring, the classification methods are used for evaluating the credit worthiness of a client. Assessing the discriminative power of rating systems is a very important topic for a bank because any misclassification can create damages to its resources.

Table 11.1 Contingency Table for Performance Evaluation of Two-Class Classification

		Sample (Y)	
		Default (1)	NonDefault (−1)
Predicted ($\widehat{Y}$)	(1)	True positive (TP)	False positive (FP)
	(−1)	False negative (FN)	True negative (TN)
Total		P	N

A representation of two-class classification performances can be formulated by a contingency table (confusion matrix) as illustrated in Table 11.1. The most frequent assessment metrics are accuracy (*Acc*) and misclassification rate (*MR*), defined as follow

$$Acc = P(\widehat{Y} = Y) = \frac{TP + TN}{P + N}. \tag{11.1}$$

$$MR = P(\widehat{Y} \neq Y) = 1 - Acc. \tag{11.2}$$

Acc and *MR* can be deceiving in certain situations and are highly sensitive to changes in data—for example, unbalanced two-class sample problems. *Acc* uses both columns of information in Table 11.1. Therefore, as class performance varies, measures of the performance will change even though the underlying fundamental performance of the classifier does not. In the presence of unbalanced data, it becomes difficult to do a relative analysis when the *Acc* measure is sensitive to the data distribution (He and Garcia, 2009).

Other evaluation metrics are frequently used to provide comprehensive assessments, especially for unbalanced data, namely, *specificity*, *sensitivity*, and *precision*, which are defined as

$$Spec = P(\widehat{Y} = -1 | Y = -1) = \frac{TN}{N}, \tag{11.3}$$

$$Sens = P(\widehat{Y} = 1 | Y = 1) = \frac{TP}{P}, \tag{11.4}$$

$$Prec = \frac{P(\widehat{Y} = 1 | Y = 1)}{P(\widehat{Y} = 1 | Y = 1) + P(\widehat{Y} = 1 | Y = -1)} = \frac{TP}{TP + FP}. \tag{11.5}$$

Precision measures an exactness, but it can not assert how many default samples are predicted incorrectly.

11.2.1. AR and ROC

Many rating methodologies and credit risk modeling approaches have been developped. The most popular validation techniques currently used in practice are Cumulative accuracy profile (CAP) and receiver operating characteristic (ROC) curve. Accuracy ratio (AR) is the summary statistic of the CAP curve (Sobehart et al., 2001). ROC has similar concept to CAP and has summary statistics, the area below the ROC curve (called AUC) (Sobehart and Keenan, 2001). Engelmann et al., (2003) analyzes the CAP and ROC from a statistical point of view.

Consider a method assigning to each observed unit a score S as a function of the explanatory variables. Scores from total samples, S, have cdf F and pdf f; scores from default samples, $S|Y = 1$, have cdf F_1; and scores from nondefault samples, $S|Y = -1$, have cdf F_{-1}.

The CAP curve is particularly useful because it simulataneously measures Type I and Type II errors. In statistical terms, the CAP curve represents the cumulative probability of default events for different percentiles of the risk score scale. The actual CAP curve is basically defined as the graph of all points $\{F, F_1\}$ where the points are connected by linear interpolation. A perfect CAP curve would assign the lowest scores to the defaulters, increase linearly, and then stay at one. For a random CAP curve without any discriminative power, the fraction x of all events with the lowest rating scores will contain x% of all defaulters, $F_i = F_{1,i}$.

Therefore, AR is defined as the ratio of the area between actual and random CAP curves to the area between the perfect and random CAP curves (Figure 11.1). The classification method is the better the higher is AR, or the closer it is to one. Formally,

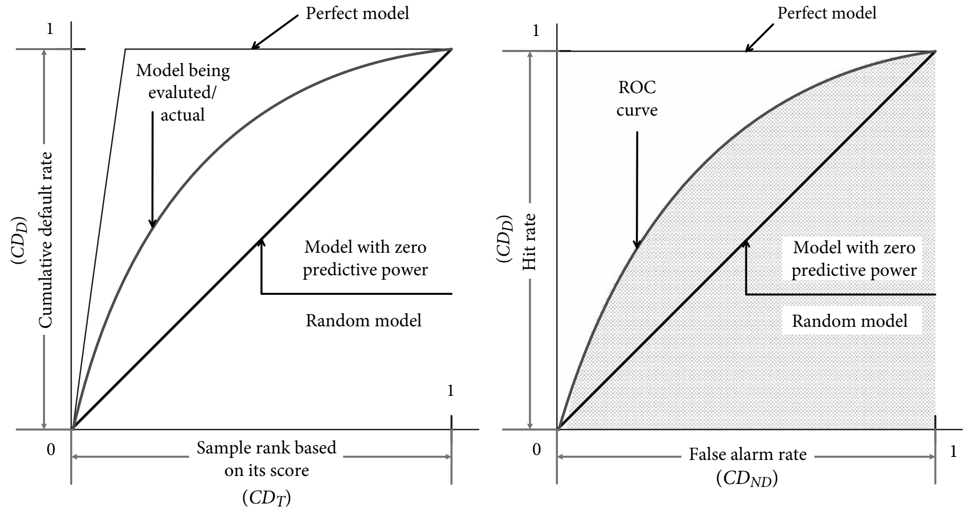


FIGURE 11.1 CAP curve (*left*) and ROC curve (*right*).

Table 11.2 Classification Decision Given Cutoff Value τ

		Sample (Y)	
		Default (1)	No default (-1)
Predicted rating score	$\leq \tau$ (default)	Correct prediction (hit)	Wrong prediction (false alarm)
	$> \tau$ (no default)	wrong prediction (mass)	correct prediction (correct rejection)

if $y = \{0, 1\}$, the AR value is defined as

$$AR = \frac{\int_0^1 y_{actual}\, F\, dF - \frac{1}{2}}{\int_0^1 y_{perfect}\, F\, dF - \frac{1}{2}}. \tag{11.6}$$

If the number of defaulters and nondefaulters is equal, the AR becomes

$$AR = 4 \int_0^1 y_{actual}\, F\, dF - 2. \tag{11.7}$$

In classification—for example, credit rating—assume that future defaulters and nondefaulters will be predicted by using rating scores. A decision maker would like to introduce a cut-off value τ, and an observed unit with rating score less than τ will be classified into potential defaulters. A classified nondefaulter in an observed unit would have rating score greater than τ. Table 11.2 summarizes the possible decisions.

If the rating score is less than the cutoff τ conditionally on a future default, the decision was correct and it is called a *hit.* Otherwise, the decision wrongly classified nondefaulters as defaulters (Type I error), called *false alarm.* The hit rate, $HR(\tau)$, and false alarm rate, $FAR(\tau)$, are defined as ((Engelmann et al., 2003) and (Sobehart and Keenan, 2001))

$$HR(\tau) = P(S|Y = 1 \leq \tau), \tag{11.8}$$

$$FAR(\tau) = P(S|Y = -1 \leq \tau). \tag{11.9}$$

Given a nondefaulter that has rating score greater than τ, the cassification is correct. Otherwise, a defaulter is wrongly classified as a nondefaulter (Type II error).

The ROC curve is constructed by plotting $FAR(\tau)$ versus $HR(\tau)$ for all given values τ. In other words, the ROC curve consists of all points $\{F_{-1}, F_1\}$ connected by linear interpolation (Figure 11.1). The area under the ROC curve (AUC) can be interpreted as the average power of the test on default or nondefault corresponding to all possible cutoff values τ. A larger AUC characterized a better classification result. A perfect model has an AUC value of 1, and a random model without discriminative power has an AUC value of 0.5. The AUC is between 0.5 and 1.0 for any reasonable rating model in practice. The ralationship between *AUC* and *AR* is defined as (Engelmann et al., (2003)

$$AR = 2AUC - 1. \tag{11.10}$$

Sing et al., (2005) developed package ROCR in R to calculate performance measures under the ROC curve for classification analysis.

Similarly, the ROC curve is formed by plotting FP_{rate} over TP_{rate}, where

$$FP_{rate} = \frac{FP}{N}, \qquad TP_{rate} = \frac{TP}{P}$$

and any point in the ROC curve corresponds to the performance of a single classifier on a given distribution. The ROC curve is useful because it provides a visual representation of the relative tradeoffs between the benefits (reflected by *TP*) and cost (reflected by *FP*) of classification (He and Garcia, 2009).

11.3. SVM Formulation

This section reviews the support vector machine (SVM) methodology in classification. We first discuss classical linear classification, both for linearly separable and nonseparable scenarios, and then focus on nonlinear classification (see Figure 11.2).

11.3.1. SVM in the Linearly Separable Case

Each observation consists of a pair of p predictors $x_i^\top = (x_i^1, \ldots, x_i^p) \in \mathbb{R}^p$, $i = 1, \ldots, n$, and the associated $y_i \in \mathcal{Y} = \{-1, 1\}$. We have a sequence

$$\mathcal{D}_n = \Big\{(x_1, y_1), \ldots, (x_n, y_n)\Big\} \in \mathcal{X} \times \{-1, 1\}, \tag{11.11}$$

of i.i.d. pairs drawn from a probability distribution $F(x, y)$ over $X \times Y$. The domain $\mathcal{X}$ is some non-empty set from which x_i are drawn, and y_i are *targets* or *labels*.

We have a machine learning, a classifier, whose task is to learn the information in a *training set*, $\mathcal{D}_n$, to predict y for any new observation. The label y_i from training set is then called a *trainer* or *supervisor*. A nonlinear classifier function f may be described by a function class $\mathcal{F}$ that is fixed *a priori*; for example, it can be the class of linear classifiers (hyperplanes).

First we will describe the SVM in the linearly separable case. A key concept to define a linear classifier is the dot product. The family $\mathcal{F}$ of classification functions, represented

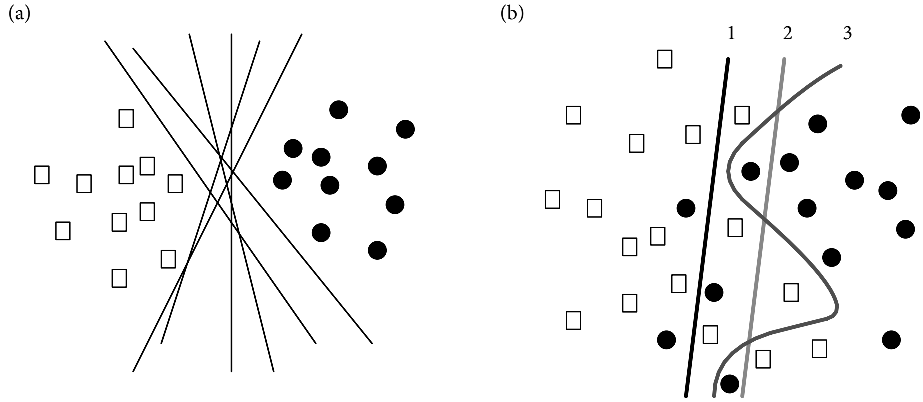


FIGURE 11.2 A set of classification function in the case of linearly separable data (*left*) and linearly nonseparable case (*right*).

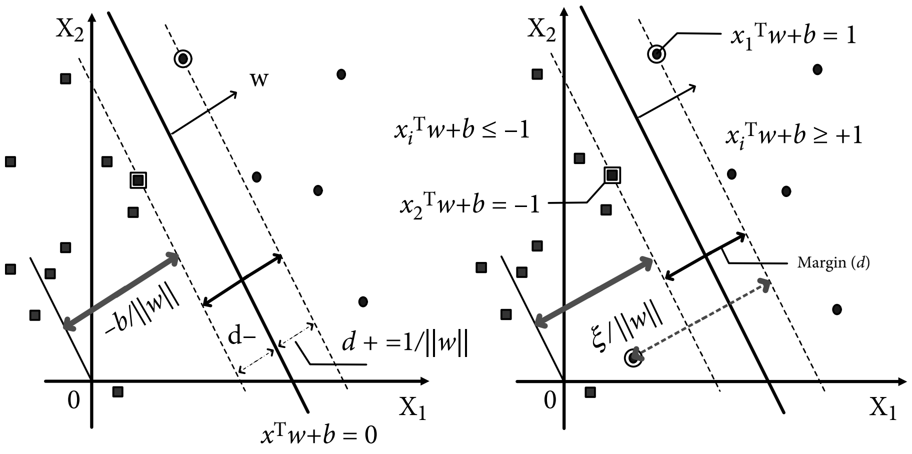

X2
w
$-b/||w||$
d–
$d+ = 1/||w||$
0
X1
$x^{\mathrm{T}}w+b = 0$
X2
$x_1{}^{\mathrm{T}}w+b = 1$
$x_i{}^{\mathrm{T}}w+b \leq -1$
$x_i{}^{\mathrm{T}}w+b \geq +1$
$x_2{}^{\mathrm{T}}w+b = -1$
Margin (*d*)
$\xi/||w||$
0
X1

FIGURE 11.3 The separating hyperplane $x^\top w + b = 0$ and the margin in the linearly separable case (*left*) and in the linearly nonseparable case (*right*).

in Figure 11.2, in the data space is given by

$$\mathcal{F} = \left\{ x^\top w + b, w \in \mathbb{R}^p, b \in \mathbb{R} \right\}, \tag{11.12}$$

where w is known as the *weight vector* and b is a deviation from the origin. .

The following decision boundary (separating hyperplane),

$$f(x) = x^\top w + b = 0, \tag{11.13}$$

divides the space into two regions as in Figure 11.3. The set of points x such that $f(x) = x^\top w = 0$ are all points that are perpendicular to w and go through the origin. The constant b translates the hyperplane away from the origin. The form of $f(x)$ is a line in two dimensions, a plane in three dimensions, and, more generally, a hyperplane in the higher dimension.

The sign of $f(x)$ determines in which regions the points lie. The decision boundary defined by a hyperplane is said to be linear because it is linear in the inputs x_i. A so-called *linear classifier* is a classifier with a linear decision boundary. Furthermore, a classifier is said to be a *nonlinear classifier* when the decision boundary depends on the data in a nonlinear way.

In order to determine the support vectors, we choose $f \in \mathcal{F}$ (or equivalently (w, b)) such that the so-called *margin*, the corridor between the separating hyperplanes, is maximal. The signs $(-)$ and $(+)$ in the margin, $d = d_- + d_+$, denote the two regions.

The classifier is a hyperplane plus the margin zone, where, in the separable case, no observations can lie. It separates the points from both classes with the highest "safest" distance (margin) between them. Margin maximization corresponds to the reduction of complexity as given by the Vapnik–Chervonenkis (VC) dimension (Vapnik 1998) of the SVM classifier.

The length of vector w is denoted by norm $\|w\| = \sqrt{w^\top w}$. A unit vector w, where $\|w\| = 1$, in the direction of w is given by $\frac{w}{\|w\|}$. Furthermore, the margin of a hyperplane $f(x)$ with respect to a data set $\mathcal{D}_n$ is as follows:

$$d_{\mathcal{D}}(f) = \frac{1}{2} w^\top (x_+ - x_-), \tag{11.14}$$

where the unit vector w is in the direction of w. It is assumed that x_+ and x_- are equidistant from the following separating hyperplane:

$$\begin{aligned} f(x_+) &= w^\top x_+ + b = a, \\ f(x_-) &= w^\top x_- + b = -a, \end{aligned} \tag{11.15}$$

with constant $a > 0$. Suppose to fix $a = 1$, hence $d_{\mathcal{D}}(f) = 1$. In order to make the geometric margin meaningful, divide $d_{\mathcal{D}}(f)\|w\|$ by norm of vector w to obtain

$$\frac{d_{\mathcal{D}}(f)}{\|w\|} = \frac{1}{\|w\|}. \tag{11.16}$$

Let $x^\top w + b = 0$ be a separating hyperplane and let $y_i \in \{-1, +1\}$ code a binary response for the ith observation. Then (d_+) and (d_-) will be the shortest distance between the separating hyperplane and the closest objects from the classes $(+1)$ and (-1). Since the separation can be done without errors, all observations $i = 1, 2, \ldots, n$ must satisfy

$$\begin{aligned} x_i^\top w + b &\geq +1 \qquad \text{for} \quad y_i = +1, \\ x_i^\top w + b &\leq -1 \qquad \text{for} \quad y_i = -1. \end{aligned}$$

We can combine both constraints into one as follows:

$$y_i(x_i^\top w + b) - 1 \geq 0, \qquad i = 1, \ldots, n. \tag{11.17}$$

Therefore the objective function of the linearly separable case would be a maximizing margin in (11.16) or, equivalently,

$$\min_w \frac{1}{2}\|w\|^2, \tag{11.18}$$

under the constraint (11.17). The Lagrangian for the primal problem in this case is

$$\min_{w,b} L_P(w, b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^n \alpha_i\{y_i(x_i^\top w + b) - 1\}. \tag{11.19}$$

The Karush–Kuhn–Tucker (KKT) (Gale et al., 1951) first-order optimality conditions are

$$\begin{aligned} \tfrac{\partial L_P}{\partial w_k} &= 0 : w_k - \textstyle\sum_{i=1}^n \alpha_i y_i x_{ik} = 0, \qquad k = 1, \ldots, d, \\ \tfrac{\partial L_P}{\partial b} &= 0 : \textstyle\sum_{i=1}^n \alpha_i y_i = 0, \\ y_i(x_i^\top w + b) - 1 &\geq 0, \quad i = 1, \ldots, n, \\ \alpha_i &\geq 0, \\ \alpha_i\{y_i(x_i^\top w + b) - 1\} &= 0. \end{aligned}$$

From these first-order conditions, we can derive $w = \sum_{i=1}^n \alpha_i y_i x_i$ and therefore the summands in (11.19) would be

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^\top x_j,$$

$$\begin{aligned} \sum_{i=1}^n \alpha_i\{y_i(x_i^\top w + b) - 1\} &= \sum_{i=1}^n \alpha_i y_i x_i^\top \sum_{j=1}^n \alpha_j y_j x_j - \sum_{i=1}^n \alpha_i, \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^\top x_j - \sum_{i=1}^n \alpha_i. \end{aligned}$$

By substituting the results into (11.19), we obtain the Lagrangian for the dual problem as follows:

$$\max_\alpha L_D(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^\top x_j. \tag{11.20}$$

Solving the primal and dual problems

$$\min_{w,b} L_P(w,b),$$

$$\max_{\alpha} L_D(\alpha)$$

gives the same solution since the optimization problem is convex.

Those points i for which the equation $y_i(x_i^\top w + b) = 1$ holds are called *support vectors.* In Figure 11.3 there are two support vectors that are marked in bold: one solid rectangle and one solid circle. Apparently, the separating hyperplane is defined only by the support vectors that hold the hyperplanes parallel to the separating one.

After solving the dual problem, one can classify an object by using the following classification rule:

$$\widehat{g}(x) = \operatorname{sign}\left(x^\top \widehat{w} + \widehat{b}\right), \tag{11.21}$$

where $\widehat{w} = \sum_{i=1}^{n} \widehat{\alpha}_i y_i x_i$.

11.3.2. SVM in the Linearly Nonseparable Case

In the linearly nonseparable case the situation is illustrated in Figure 11.3; the slack variables ξ_i represent the violation of strict separation that allow a point to be in the margin error, $0 \le \xi_i \le 1$, or to be misclassified, $\xi > 1$. In this case the following inequalities can be induced (see Figure 11.3):

$$\begin{aligned} w + b &\ge 1 - \xi_i && \text{for} \quad y_i = 1, \\ w + b &\le -(1 - \xi_i) && \text{for} \quad y_i = -1, \\ \xi_i &\ge 0, \end{aligned}$$

which could be combined into the two following constraints:

$$y_i(x_i^\top w + b) \ge 1 - \xi_i, \tag{11.22a}$$

$$\xi_i \ge 0. \tag{11.22b}$$

The penalty for misclassification is related to the distance of a misclassified point x_i from the hyperplane bounding its class. If $\xi_i > 0$, then an error in separating the two sets occurs. The objective function corresponding to penalized margin maximization is then formulated as

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i, \tag{11.23}$$

with constraints as in equation (11.22). This formulation, a convex optimization problem, is called *softmargin* ($x_i^\top w + b = \pm 1$), introduced by Cortes and Vapnik (1995). The parameter C characterizes the weight given to the misclassification. The minimization of the objective function with constraints (11.22a) and (11.22b) provides the highest possible margin in the case when misclassification are inevitable due to the linearity of the separating hyperplane.

Non-negative slack variables ξ_i allow points to be on the wrong side of their soft margin as well as on the separating hyperplane. Parameter C is a cost parameter that controls the amount of overlap. If the data are linearly separable, then for sufficiently large C the solution (11.18) and (11.23) coincide. If the data are linearly nonseparable as C increases the solution approaches the minimum overlap solution with largest margin, which is attained for some finite value of C (Hastie et al., 2004).

The Lagrange function for the primal problem is

$$\begin{aligned} L_P(w,b,\xi) = {} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ & - \sum_{i=1}^{n}\alpha_i\{y_i\left(x_i^\top w + b\right) - 1 + \xi_i\} - \sum_{i=1}^{n}\mu_i\xi_i, \end{aligned} \tag{11.24}$$

where $\alpha_i \geq 0$ and $\mu_i \geq 0$ are Lagrange multipliers. The primal problem is formulated by minimizing the Lagrange function as follows:

$$\min_{w,b,\xi} L_P(w,b,\xi). \tag{11.25}$$

The first-order conditions are given by

$$\frac{\partial L_P}{\partial w_k} = 0: \quad w_k - \sum_{i=1}^{n}\alpha_i y_i x_{ik} = 0, \tag{11.26a}$$

$$\frac{\partial L_P}{\partial b} = 0: \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \tag{11.26b}$$

$$\frac{\partial L_P}{\partial \xi_i} = 0: \quad C - \alpha_i - \mu_i = 0. \tag{11.26c}$$

with the following constraints:

$$\alpha_i \geq 0, \tag{11.27a}$$

$$\mu_i \geq 0, \tag{11.27b}$$

$$\alpha_i\{y_i(x_i^\top w + b) - 1 + \xi_i\} = 0, \tag{11.27c}$$

$$\mu_i\xi_i = 0. \tag{11.27d}$$

Note that $\sum_{i=1}^{n} \alpha_i y_i b = 0$, similar to the linearly separable case, therefore the primal problem translates into the following dual problem:

$$\begin{aligned} L_D(\alpha) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j - \sum_{i=1}^{n}\alpha_i y_i x_i^\top \sum_{j=1}^{n}\alpha_j y_j x_j \\ &\quad + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\mu_i\xi_i \\ &= \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j + \sum_{i=1}^{n}\xi_i(C - \alpha_i - \mu_i). \end{aligned}$$

The last term is equal to zero. The dual problem is formulated as follows:

$$\max_{\alpha} L_D(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^\top x_j, \tag{11.28}$$

subject to

$$0 \le \alpha_i \le C, \qquad \sum_{i=1}^{n}\alpha_i y_i = 0. \tag{11.29}$$

The sample x_i for which $\alpha > 0$ (support vectors) are those points that are on the margin, or within the margin when a soft margin is used. The support vector is often sparse and the level of sparsity (fraction of data serving as support vector) is an upper bound for the misclassification rate (Schölkopf and Smola 2002).

11.3.3. SVM in Nonlinear Classification

We have not made any assumptions on the domain $\mathcal{X}$ other than being a set. We need additional structure in order to study machine learning to be able to generalize to unobserved data points. Given some new point $x \in \mathcal{X}$, we want to predict the corresponding $y \in \mathcal{Y} = \{-1, 1\}$. By this we mean that we choose y such that (x, y) is in some sense similar to the training examples. To this end, we need similarity measures in $\mathcal{X}$ and in $\{-1, 1\}$ (see Chen et al. (2005)).

In order to be able to use a dot product as a similarity measure, we need to transform them into some dot product space, so-called *feature space* $\mathcal{H} \in \mathbb{H}$, which need not be identical to $\mathbb{R}^n$,

$$\psi : \mathcal{X} \to \mathcal{H}. \tag{11.30}$$

The nonlinear classifiers, as in Figure 11.4, maps the data with a nonlinear structure via a function $\psi : \mathbb{R}^p \mapsto \mathbb{H}$ into a high-dimensional space $\mathbb{H}$ where the classification

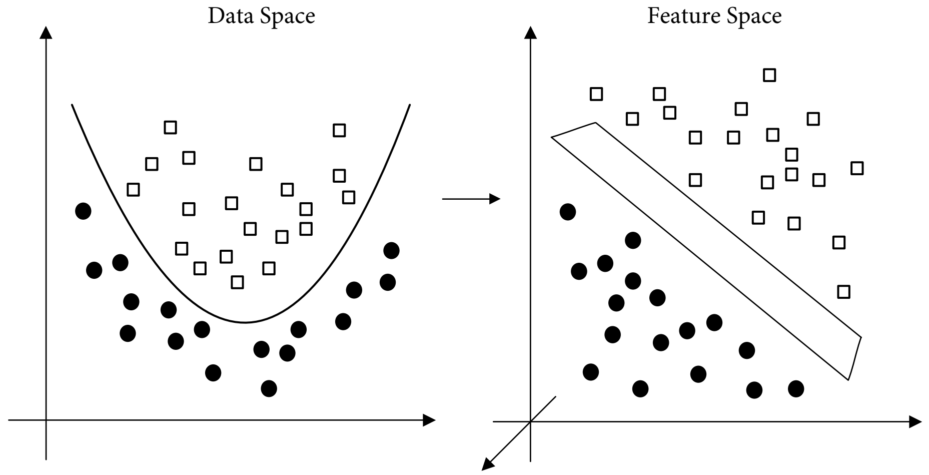


FIGURE 11.4 Mapping into a three-dimensional feature space from a two-dimensional data space $\mathbb{R}^2 \mapsto \mathbb{R}^3$. The transformation $\psi(x_1,x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top$ corresponds to the kernel function $K(x_i,x_j) = (x_i^\top x_j)^2$.

rule is (almost) linear. Note that all the training vectors x_i appear in L_D (eq. 11.28) only as dot products of the form $x_i^\top x_j$. In the nonlinear SVM, the dot product transforms to $\psi(x_i)^\top \psi\big(x_j\big)$.

The learning then takes place in the feature space, provided that the learning algorithm can be expressed so that the data points only appear inside dot products with other points. This is often referred to as the *kernel trick* (Schölkopf and Smola 2002). The *kernel trick* is to compute this scalar product via a kernel function. More precisely, the projection $\psi : \mathbb{R}^p \mapsto \mathbb{H}$ ensures that the inner product $\psi(x_i)^\top \psi\big(x_j\big)$ can be represented by kernel function

$$k(x_i,x_j) = \psi(x_i)^\top \psi(x_j). \tag{11.31}$$

If a kernel function k exists such that (11.31) holds, then it can be used without knowing the transformation ψ explicitly.

Given a kernel k and any data set $x_1,\ldots,x_n \in \mathcal{X}$, the $n \times n$ matrix

$$K = (k(x_i,x_j))_{ij} \tag{11.32}$$

is called the kernel or *Gram* matrix of k with respect to $x_1,\ldots,x_n$. A necessary and sufficient condition for a symmetric matrix K, with $K_{ij} = K(x_i,x_j) = K(x_j,x_i) = K_{ji}$, to be a kernel is, by Mercer's theorem (Mercer 1909), that K is positive definite:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j K(x_i,x_j) \geq 0. \tag{11.33}$$

The following is a simple example of a kernel trick which shows that the kernel can be computed without computing explicitly the mapping function ψ. To obtain the classifier $f(x) = w^\top \psi(x) + b$, consider the case of a two-dimensional input space with the following mapping function,

$$\psi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top,$$

such that

$$w^\top \psi(x) = w_1x_1^2 + \sqrt{2}w_2x_1x_2 + w_3x_2^2.$$

The dimensionality of the feature space $\mathcal{F}$ is of quadratic order of the dimensionality of the original space. Kernel methods avoid the step of explicitly mapping the data into a high-dimensional feature space by the following steps:

$$\begin{aligned} f(x) &= w^\top x + b \\ &= \sum_{i=1}^{n} \alpha_i x_i^\top x + b \\ &= \sum_{i=1}^{n} \alpha_i \psi(x_i)^\top \psi(x) + b \qquad \text{in feature space } \mathcal{F} \\ &= \sum_{i=1}^{n} \alpha_i k(x_i, x) + b, \end{aligned}$$

where the kernel is associated with the following mapping:

$$\begin{aligned} \psi(x_i)^\top \psi(x) &= (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2)(x_1^2, \sqrt{2}x_1x_2, x_2^2)^\top \\ &= x_{i1}^2x_1^2 + 2x_{i1}x_{i2}x_1x_2 + x_{i2}^2x_2^2 \\ &= (x_i^\top x)^2 \\ &= k(x_i, x). \end{aligned}$$

Furthermore, to obtain nonlinear classifying functions in the data space, a more general form is obtained by applying the kernel trick to (11.28) as follows:

$$\max_\alpha L_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j K(x_i, x_j), \tag{11.34}$$

subject to

$$0 \le \alpha_i \le C, \qquad i = 1, \ldots, n, \tag{11.35a}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{11.35b}$$

One of the most popular kernels used in SVM is the radial basis function (RBF) kernel given by

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right). \tag{11.36}$$

Furthermore, Chen et al. (2005) summarized the benefits of transforming the data into the feature space $\mathcal{H}$.

The resulting optimization problems (11.34), which is a typical quadratic problem (QP), are dependent upon the number of training examples. The problem can easily be solved in a standard QP solver, that is, package `quadprog` in `R` (quadprog 2004) or an optimizer of the interior point family (Vanderbei 1999; Schölkopf and Smola 2002) implemented to `ipop` in package `kernlab` in `R` (Karatzoglou et al., 2004).

Osuna et al., (1997b) proposed exact methods by presenting a decomposition algorithm that is guaranteed to solve QP problem and that does not make assumptions on the expected number of support vectors. Platt (1998) proposed a new algorithm called Sequential Minimal Optimization (SMO), which decomposes the QP in SVM without using any numerical QP optimization steps. Some work on decomposition methods for QP in SVM was done by, for example, Joachims (1998), Keerthi et al., (2001), and Hsu and Lin (2002). Subsequent developments were achieved by Fan et al., (2005) as well as by Glasmachers and Igel (2006).

Due to the fast development and wide applicability, the existence of many SVM software routines is not surprising. The SVM software, which is written in `C` or `C++`, includes `SVMTorch` (Collobert et al., 2002), `SVMlight` (Joachims 1998), `Royal Holloway Support Vector Machines` (Gammerman et al., 2001), and `libsvm` (Chang and Lin 2001), which provides interfaces to `MATLAB`, `mySVM` (Rüping 2004) and `M-SVM` (Guermeur 2004). The SVM is also available in `MATLAB` (Gunn (1998) and Canu et al. (2005)). Several packages in `R` dealing with SVM are `e1071` (Dimitriadou et al., 1995), `kernlab` (Karatzoglou et al., 2004), `svmpath` (Hastie et al., 2004) and `klaR` (Roever et al., 2005).

SVM has recently been developed by many researchers in various fields of application, that is, Least Squares SVM (Suykens and Vandewalle 1999), Smooth SVM or SSVM (Lee and Mangasarian 2001), 1-norm SVM (Zhu et al., 2004), Reduced SVM (Lee and Huang 2007), ν-SVM (Schölkopf et al., 2000; and Chen et al., 2005). Hastie et al., (2004) viewed SVM as a regularized optimisation problem.

11.4. Evolutionary Model Selection

During the learning process (training), an SVM finds the large margin hyperplane by estimating sets of parameters α_i and b. The SVM performance is also determined by another set of paramaters, the so-called *hypermarameters*. These are the soft margin constant C and the parameters of the kernel, σ, as in (11.36). The value of C determines the size of the margin errors. The kernel parameters control the flexibility of the classifier. If this parameter is too large, then overfitting will occur.

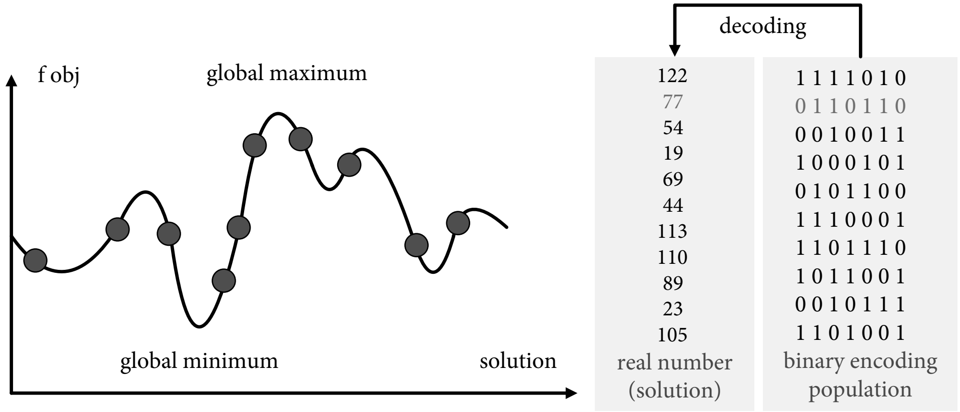
Hastie et al., (2004) argue that the choice of the cost parameter (C) can be critical. They derive an algorithm, so-called `SvmPath`, that can fit the entire path of SVM solutions for every value of the cost parameter, with essentially the same computational cost as fitting one SVM model. The `SvmPath` has been implemented in the `R` computing environment via the library `svmpath`. Chen et al., (2011) use grid search methods to optimize SVM hyperparamaters to obtain the optimal classifier for a credit data set. This chapter employs a Genetic Algorithm (GA) as an evolutionary algorithm to optimize the SVM parameters.

Lessmann et al., (2006) used GA for model selection applied on four well-known benchmark data sets from Statlog project (Australian credit and German credit data set) and UCI machine learning library (heart disease and Wisconsin breast cancer data sets). The SVM model selection used grid search and GA methods that were applied to two different fitness criteria: (i) cross-validation (CV) balanced classification accuracy (BCA) and (ii) CV BCA with simple bound for leave-one-out error. In general, GA gave better performance to guide the SVM model selection. Another paper discussing SVM model selection based on GA is Zhou and Xu (2009).

The idea of GA is based on the principle of *survival of the fittest*, which follows the evolution of a population of individuals through successive generations. Living beings are constructed by cells that carry the genetic information. Each cell contains a fixed number of chromosomes composed by several genes (information). A gene is conceptualized as a binary code. All information carried by genes of all chromosomes (so-called genotype) determines all characteristics of an individual (so-called phenotype). Each individual is evaluated to give measures of its fitness by means of genetic operation to form a new individual. There are two types of genetic operation: *mutation* and *crossover* (also known as *recombination*). Mutation creates a new individual by making changes in a single chromosome. Crossover creates new individuals by combining parts of chromosomes from two individuals. When sexual reproduction takes place, children (new chromosome) or offspring receive, for each pair, one chromosome from each of their parents (old chromosomes). The children are then evaluated. A new population is formed by selecting fitter individuals from the parent population and the children population. After several generations (iteration), the algorithm converges to the best individual, which hopefully represents a (global) optimal solution (Baragona et al., 2011; Gen and Cheng, 2000). See Figure 11.5 and 11.6 for illustration.

A binary string chromosome is composed of several genes. Each gene has a binary value (*allele*) and its position (*locus*) in a chromosome as shown in Figure 11.7. The binary string is decoded to the real number in a certain interval by the following equation:

$$\theta = \theta_{lower} + (\theta_{upper} - \theta_{lower}) \frac{\sum_{i=0}^{l-1} a_i 2^i}{2^l}, \tag{11.37}$$



real number (solution)	binary encoding population
122	1 1 1 1 0 1 0
77	0 1 1 0 1 1 0
54	0 0 1 0 0 1 1
19	1 0 0 0 1 0 1
69	0 1 0 1 1 0 0
44	1 1 1 0 0 0 1
113	1 1 0 1 1 1 0
110	1 0 1 1 0 0 1
89	0 0 1 0 1 1 1
23	1 1 0 1 0 0 1
105	

FIGURE 11.5 Generating binary encoding chromosomes to obtain the global optimum solution through GA.

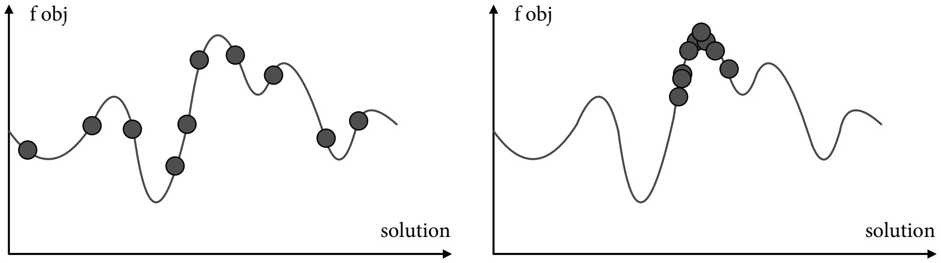


FIGURE 11.6 GA convergency: solutions at first generation (*left*) and *g*th generation (*right*).

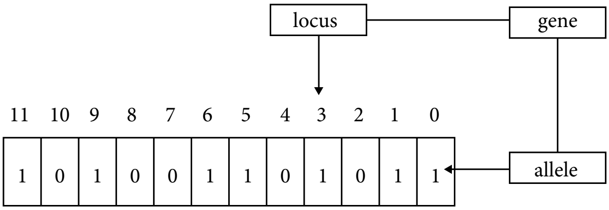


FIGURE 11.7 Chromosome.

where θ is the solution (i.e., parameter C or σ), a_i is binary value (*allele*), and l is the chromosome length. In the encoding issue, according to what kind of symbol is used as the alleles of a gene, the encoding methods can be classified as follows: *binary* encoding, *real-number* encoding, *integer* or *literal permutation* encoding, and *general data structure* encoding.

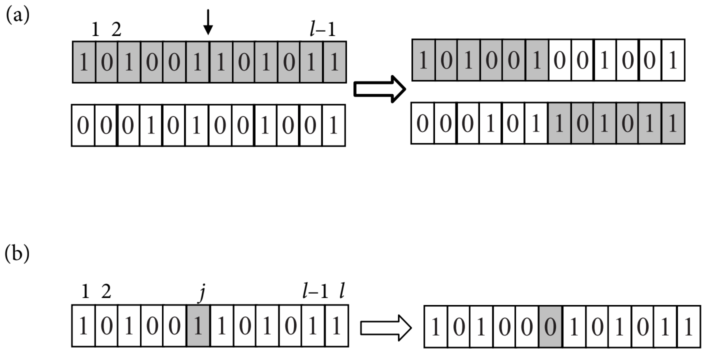


FIGURE 11.8 One-point crossover (*top*) and bit-flip mutation (*bottom*).

The current solution is evaluated to measure the fitness performance based on discriminatory power (AR or AUC), $f^*(C,\sigma)$. The next generation results from the reproduction process articulated in three stages: selection, crossover, and mutation (Figure 11.8). The selection step is choosing which chromosomes of the current population are going to reproduce. The most fitted chromosome should reproduce more frequently than the less fitted one.

If f_i^* is the fitness of ith chromosome, then its probability of being selected (relative fitness) is

$$p_i = \frac{f_i^*}{\sum_{i=1}^{popsize} f_i^*}, \tag{11.38}$$

where *popsize* is the number of chromosomes in the population or population size. The *roulette wheel* method selects a chromosome with probability proportional to its fitness (see Figure 11.9). To select the new chromosome, generate a random number $u \sim \mathrm{U}(0,1)$, then select ith chromosome if $\sum_{i=1}^{t} p_i < u < \sum_{i=1}^{t+1} p_i$, where $t = 1,\ldots,(popsize - 1)$. Repeat *popsize* times to get new population. The other popular selection operators are *stochastic universal sampling*, *tournament selection*, steady-state reproduction, sharing, ranking, and scaling.

The selection stage produces candidates for reproduction (iteration). Ordered pairs of chromosomes mate and produce a pair of offspring that may share genes of both parents. This process is called crossover (with fixed probability). One-point crossover can be extended to two-point or more crossover. Afterwards, the offspring is subject to the mutation operator (with small probability). Mutation introduces innovations into the population that cause the trapped local solutions to move out. The relationship of GA with evolution in nature is given in Table 11.3. The application of GA in SVM for model selection is represented by Figure 11.10.

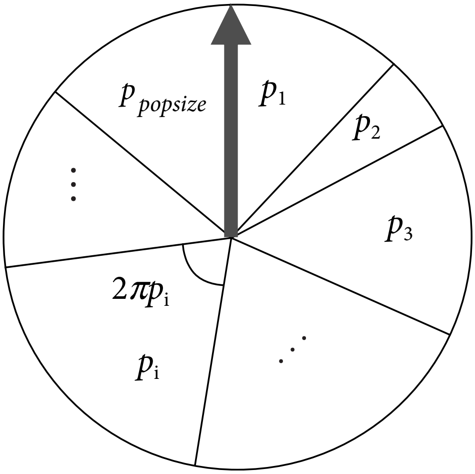


FIGURE 11.9 Probability of *i*th chromosome to be selected in the next iteration (generation).

Table 11.3 Nature to GA-SVM Mapping

Nature	GA-SVM
Population	Set of parameters
Individual (phenotype)	Parameters
Fitness	Discriminatory power
Chromosome (genotype)	Encoding of parameter
Gene	Binary encoding
Reproduction	Crossover
Generation	Iteration

A too-high crossover rate may lead to premature convergence of the GA as well as a too-high mutation rate may lead to the loss of good solutions unless there is elitist selection. In elitism, the best solution in each iteration is maintained in another memory. When the new population will replace the old one, check whether best solution exists in the new population. If not, replace any chromosomes in the new population with the best solution we saved in another memory.

It is natural to expect that the adaptation of GA is not only for finding solutions, but also for tuning GA to the particular problem. The adaptation of GA is to obtain an effective implemetation of GA to real-world problems. In general, there are two types of adaptations: adaptation to problems and adaptation to evolutionary processes (see Gen and Cheng (2000) for details).

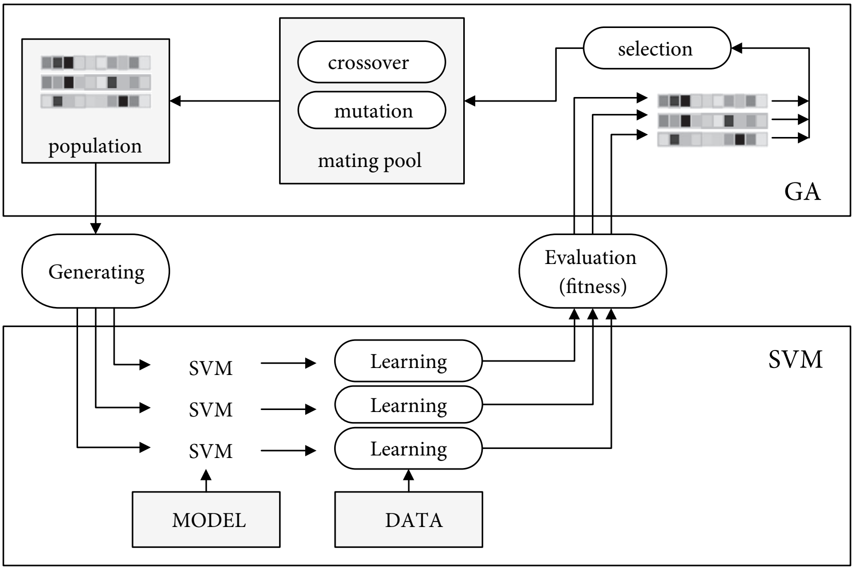


FIGURE 11.10 Iteration (generation) procedure in GA-SVM.

Table 11.4 Credit Reform Data Based on Industry Sector

Type	Solvent (%)	Insolvent (%)	Total (%)
Manufacturing	26.06	1.22	27.29
Construction	13.22	1.89	15.11
Wholesale and retail	23.60	0.96	24.56
Real estate	16.46	0.45	16.90
Total	79.34	4.52	83.86
Others	15.90	0.24	16.14

11.5. Application

The SVM with evolutionary model selection is applied to the CreditReform database consisting of 20,000 solvent and 1000 insolvent German companies in the period from 1996 to 2002. Approximately 50% of the data are from the years 2001 and 2002. Table 11.4 describes the composition of the CreditReform database in terms of industry sectors. In our study, we only used the observations from the following industry sectors: manufacturing, wholesale and retail, construction, and real estate.

Table 11.5 Filtered Credit Reform Data

Year	Solvent (Number (%))	Insolvent (Number (%))	Total
1997	872 (9.08)	86 (0.90)	958 (9.98)
1998	928 (9.66)	92 (0.96)	1020 (10.62)
1999	1005 (10.47)	112 (1.17)	1117 (11.63)
2000	1379 (14.36)	102 (1.06)	1481 (15.42)
2001	1989 (20.71)	111 (1.16)	2100 (21.87)
2002	2791 (29.07)	135 (1.41)	2926 (30.47)
Total	8964 (93.36)	638 (6.64)	9602 (100)

We excluded the observations of solvent companies in 1996 because of missing insolvencies in this year. The observations with zero values in those variables that were used as denominator to compute the financial ratios were also deleted. We also excluded the companies whose total assets were not in the range EUR 10^5-10^7. We replace the extreme financial ratio values by the following rule: If $x_{ij} > q_{0.95}(x_j)$, then $x_{ij} = q_{0.95}(x_j)$; and if $x_{ij} < q_{0.05}(x_j)$, then $x_{ij} = q_{0.05}(x_j)$, where q is quantile. Table 11.5 describes the filtered data used in this study.

Our data set is the same as used in Chen et al., (2011) and Härdle et al., (2009), who used grid search in model selection. A little difference in our filtered data set happened after the preprocessing step. We predict the default based on 28 financial ratio variables as predictors used in Chen et al., (2011). Härdle et al., (2009) used only 25 financial ratio variables as predictors. Grid search needs a large memory, in case of SVM model selection, to find the optimal solution in very large interval of parameters. Moreover, if open source software such as R is used free, memory may be limited. In order to overcome the problem, the grid search method can be applied in sequential interval of parameters. In this way, GA is a good solution to decide the initial interval of parameter.

In our work, the GA was employed as an evolutionary model selection of SVM. The population size is 20 chromosomes. We used a fixed number of iterations (generations) as a termination criterion. The number of generations is fixed at 100 with crossover rate 0.5, mutation rate 0.1, and elitism rate 0.2 of the population size. The obtained optimal parameters of GA-SVM are given by $\sigma = 1/178.75$ and $C = 63.44$.

We use classical methods such as discriminant analysis (DA), logit and probit models as benchmark (Table 11.6). Discriminant analysis shows a poor performance in both training and testing data set. The financial ratios variables are collinear such that the assumptions in DA are violated. Logit and probit models show that a perfect classification in training data set with several variables are not significant. The best models of logit and probit, by excluding the nonsignificant variables, still show not significant different from what would occur if we use the whole variables.

Table 11.6 Training Error and Testing Error (%) from Discriminant Analysis, Logit, and Probit

	Training Error (%)				Testing Error (%)		
Training	DA	Logit	Probit	Testing	DA	Logit	Probit
1997	10.01	0	0	1998	9.13	9.00	8.88
1998	9.25	0	0	1999	11.08	10.82	10.82
1999	10.43	0	0	2000	9.20	9.31	9.31
2000	8.62	0	0	2001	6.86	7.78	7.78
2001	6.64	0	0	2002	7.95	7.16	7.16

Table 11.7 Training Error (%), Discriminatory Power, Cross Validation (Fivefold) and Testing Error

Training	Training Error (%)	*Acc*, *Spec*, *Sens*, *Prec*, *AR*, *AUC*	Cross Validation	Testing	Testing Error (%)
1997	0	1	9.29	1998	9.02
1998	0	1	9.22	1999	10.38
1999	0	1	10.03	2000	6.89
2000	0	1	8.57	2001	5.29
2001	0	1	4.55	2002	4.75

The GA-SVM yields also a perfect classification in the training dataset as in Table 11.7 which shows an overfitting. Overfitting means that the classification boundary is too curved, therefore has less ability to classify the unobserved data (i.e. testing data) correctly. The misclassification is zero for all training data such that the other discriminatory power measures, *Acc*, *Spec*, *Sens*, *Prec*, *AR* and *AUC*, attain one. A fivefold cross-validation was used to measure the performance of GA-SVM in default prediction by omitting the overfitting effect. Overall, GA-SVM outperforms the benchmark models in both training and testing data sets.

11.6. Acknowledgments

This research is supported by the Deutsche Forschungsgemeinschaft through the SFB 649 "Economic Risk", Humboldt-Universität zu Berlin. Dedy Dwi Prastyo is also supported by Directorate General for Higher Education, Indonesian Ministry of Education and Culture through Department of Statistics, Institut Teknologi Sepuluh Nopember

(ITS), Indonesia. Christian Hafner acknowledges financial support of the Deutsche Akademische Auslandsdienst.

References

Altman, E. 1968. "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance*, **23**(4), pp. 589–609.

Altman, E., G. Marco, and F. Varetto. 1994. "Corporate Distress Diagnosis: Comparison Using Linear Discriminant Analysis and Neural Network (the Italian Experience)." *Journal of Banking and Finance*, **18**, pp. 505–529.

Baragona, R., F. Battaglia, and I. Poli. 2011. *Evolutionary Statistical Procedures*. Heidelberg: Springer.

Beaver, W. 1966. "Financial Ratios as Predictors of Failures." *Journal of Accounting Research. Empirical Research in Accounting: Selected Studies*, Supplement to Vol. 4, pp. 71–111.

Black, F. and M. Scholes. 1973. "The Pricing of Option and Corporate Liabilities." *The Journal of Political Economy* **81**(3), pp. 637–654.

Blanz, V., B. Schölkopf, H. Bülthoff, C. Burges, V. Vapnik, and T. Vetter. 1996. "Comparison of View-Based Object Recognition Algorithms Using Realistic 3d Models." *Proceedings of International Conference on Artificial Neural Networks—ICANN 96*.

Boser, B. E., I. M. Guyon, and V. Vapnik. 1992. A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, COLT '92*, ed. D. Haussler. Pittsburgh: ACM Press, pp. 144–152.

Burges, C. J. C. 1998. "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery*, **2**, pp. 121–167.

Burges, C. and B. Schölkopf. 1996. "Improving the accuracy and speed of support vector learning machines." In *Advances in Neural Information Processing System 9*, eds. M. Mozer, M. Jordan, and T. Petsche. Cambridge, MA: MIT Press, pp. 375–381.

Canu, S., Y. Grandvalet, and A. Rakotomamonjy. 2005. "SVM and Kernel Methods MATLAB Toolbox." Perception Systemes et Information. INSA de Rouen, Rouen, France. URL http://asi.insa-rouen.fr/enseignants/ arakoto/toolbox/index.html.

Chang, C. C. and C. J. Lin. 2001. "LIBSVM—A Library for Support Vector Machines." URL http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.

Chen, P.-H., C.-J. Lin, and B. Schölkopf. 2005. "A Tutorial on ν-Support Vector Machines." *Applied Stochastic Models in Business and Industry*, **21**, pp. 111–136.

Chen, S., W. Härdle, and R. Moro. 2011. "Modeling Default Risk with Support Vector Machines." *Quantitative Finance*, **11**, pp. 135–154.

Collobert, R., S. Bengio, and J. Mariethoz. 2002. "TORCH: A Modular Machine Learning Software Library." URL http://www.torch.ch/ and http://publications.idiap.ch/downloads/reports/2002/rr02-46.pdf

Cortes, C. and V. Vapnik. 1995. "Support Vector Networks." *Machine Learning*, **20**, pp. 273–297.

Dimitriadou, E., K. Hornik, F. Leisch, D. Meyer, and A. Weingessel. 1995. "e1071: misc Functions of the Department of Statistics (e1071), TU Wien." Version 1.5-11., URL http://CRAN.R-project.org/.

Engelmann, B., E. Hayden, and D. Tasche. 2003. "Measuring the Discriminative Power of Rating System." *Banking and Financial Supervision. Discussion Paper*, **2**(1), pp. 1–23.

Fan, D.R.E., P.-H. Chen, and C.-J. Lin. 2005. "Working Set Selection Using Second Order Information for Training SVM." *Journal of Machine Learning Research*, **6**, pp. 1889–1918.

Gale, D., H. W. Kuhn, and A. W. Tucker. 1951. "Linear Programming and the Theory of Games." *Proceedings: Activity Analysis of Production and Allocation*, ed. T. C. Koopmans. New York: John Wiley & Sons, pp. 317–329.

Gammerman, A., N. Bozanic, B. Schölkopf, V. Vovk, V. Vapnik, L. Bottou, A. Smola, C. Watkins, Y. LeCun, C. Saunders, M. Stitson, and J. Weston. 2001. "Royal Holloway Support Vector Machines." URL http://svm.dcs.rhbnc.ac.uk/dist/index.shtml.

Gen, M. and R. Cheng. 2000. *Genetic Algorithms and Engineering Design.* New York: John Willey & Sons.

Glasmachers, T., and C. Igel. 2006. "Maximum-Gain Working Set Selection for Support Vector Machines." *Journal of Machine Learning Research*, **7**, pp. 1437–1466.

Guermeur, Y. 2004. "M-SVM." Lorraine Laboratory of IT Research and Its Applications. URL http://www.loria.fr/la-recherche-en/equipes/abc-en.

Gunn, S. R., 1998. "Support Vector Machines for Classification and Regression." *Technical Report.* Department of Electronics and Computer Science, University of Southampton.

Härdle, W., L. Hoffmann, and R. Moro. 2011. *"Learning Machines Supporting Bankruptcy Prediction."* In eds. P. Cizek, W. Härdle, R. Weron. *Statistical Tools for Finance and Insurance, second edition*, Heidelberg: Springer Verlag, pp. 225–250.

Härdle, W., Y.-J. Lee, D. Schäfer, and Y.-R. Yeh. 2009. "Variable Selection and Oversampling in the Use of Smooth Support Vector Machine for Predicting the Default Risk of Companies." *Journal of Forecasting*, **28**, pp. 512–534.

Härdle, W. and L. Simar. 2012. *Applied Multivariate Statistical Analysis, third edition.* Heidelberg: Springer Verlag.

Haupt, R. L. and S. E. Haupt. 2004. *Practical Genetic Algorithms, second edition.* Hoboken, NJ: John Wiley & Sons.

Haykin, S. 1999. *Neural Network: A Comprehensive Foundation.* Engelwood Cliffs, NJ: Prentice-Hall.

Hastie, T., S. Rosset, R. Tibshirani, and J. Zhu. 2004. "The Entire Regularization Path for the Support Vector Machine." *Journal of Machine Learning Research*, **5**, pp. 1391–1415.

He, H. and E. A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering*, **21**(9), pp. 1263–1284.

Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press.

Hsu, C.-W. and C.-J. Lin. 2002. "A Simple Decomposition Method for Support Vector Machines." *Machine Learning*, **46**, pp. 291–314.

Hwang, R. C., K. F. Cheng, and J. C. Jee. 2007. "A Semiparametric Method for Predicting Bankruptcy." *Journal of Forecasting*, **26**, pp. 317–342.

Japkowicz, N., and S. Stephen. 2002. "The Class Imbalanced Problem: A systematic Study." *Intelligent Data Analysis*, **6**(5), pp. 429–449.

Joachims, T. 1998. "Making Large-Scale SVM Learning Practical." In *Advances in Kernel Methods—Support Vector Learning*, eds. B. Schölkopf, J.C. Burges, and A.J. Smola. Cambridge: MIT Press, pp. 169–184.

Karatzoglou, A., A. Smola, K. Hornik, and A. Zeileis. 2004. "Kernlab—an S4 Package for Kernel Methods in R." *Journal of Statistical Software*, **11**(**9**), pp. 1–20.

Keerthi, S. S., S. K. Shevade, C. Bhattacharya, and K. R. K. Murthy. 2000. "Improvements to Platt's SMO Algorithm for SVM Classifier Design." *Neural Computation*, **13**, pp. 637–649.

Krahnen, J. P. and M. Weber. 2001. "Generally Accepted Rating Principles: A Primer." *Journal of Banking and Finance*, **25**, pp. 3–23.

Lee, Y.-J. and S.-Y. Huang. 2007. "Reduced Support Vector Machines: A Statistical Theory." *IEEE Transactions on Neural Networks*, **18**(1), pp. 1–13.

Lee, Y.-J. and O. L. Mangasarian. 2001. "SSVM: A Smooth Support Vector Machine for Classification." *Computational Optimization and Application*, **20**(1), pp. 5–22.

Lessmann, S., R. Stahlbock, and S. F. Crone. 2006. "Genetic Algorithms for Support Vector Machine Model Selection." In *Proceedings of International Conference on Neural Networks*. Canada, Vancouver: IEEE, pp. 3063–3069.

Lo, A. W. 1986. "Logit Versus Discriminant Analysis: A Specification Test and Application to Corporate Bankruptcies." *Journal Econometrics*, **31**(2), pp. 151–178.

Maalouf, M., and T. B. Trafalis. 2011. "Robust Weighted Kernel Logistic Regression in Imbalanced and Rare Events Data." *Computational Statistics and Data Analysis*, **55**, pp. 168–183.

Martin, D. 1977. "Early Warning of Bank Failure: A Logit Regression Approach." *Journal of Banking and Finance*, **1**, pp. 249–276.

Mercer, J. 1909. Functions of Positive and Negative Type, and Their Connection with the Theory of Integral Equations." *Philosophical Transactions of the Royal Society of London*, **25**, pp. 3–23.

Merton, R. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *The Journal of Finance*, **29**, pp. 449–470.

Merwin, C. 1942. "Financing Small Corporations in Five Manufacturing Industries." Cambridge, MA: National Bureau of Economic Research, pp. 1926–36.

Michalewicz, Z. 1996. *Genetics Algorithm + Data Structures = Evolution Programs, third edition*. New York: Springer.

Mitchell, M. 1999. *An Introduction to Genetic Algorithms*. Cambridge, MA: MIT Press.

Müller, K.-R., A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik. 1997. "Predicting Time Series with Support Vector Machines." *Proceedings International Conference on Artificial Neural Networks* ICANN'97. Springer Lecture Notes in Computer Science, Berlin: Springer, pp. 999–1004.

Ohlson, J. 1980. "Financial Ratios and the Probabilistic Prediction of Bankruptcy." *Journal of Accounting Research*, **18**(1), pp. 109–131.

Osuna, E., R. Freund, and F. Girosi. 1997a. "Training Support Vector Machines: An Application to Face Detection." *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 130–136.

Osuna, E., R. Freund, and F. Girosi. 1997b. "An Improved Training Algorithm for Support Vector Machines." In *Proceedings of the 1997 IEEE Workshop*, eds. J. Principe, L. Gile, N. Morgan, and E. Wilson, *Neural Networks for Signal Processing VII*, New York, pp. 276–285.

Platt, H., M. Platt, and J. Pedersen. 1994. "Bankruptcy Discrimination with Real Variables." *Journal of Business Finance and Accounting*, **21**(4), pp. 491–510.

Platt, J. C. 1998. "Fast Training of Support Vector Machines Using Sequential Minimal Optimization." In eds. B. Schölkopf, J. C. Burges, and A. J. Smola. *Advances in Kernel Methods—Support Vector Learning*, Cambridge, Ma: MIT Press.

Roever, C., N. Raabe, K. Luebke, U. Ligges, G. Szepannek, and M. Zentgraf. 2005. "klaR—Classification and Visualization." R package, Version 0.4-1. URL http://CRAN.R-project.org/.

Rüping, S. 2004. "mySVM—A Support Vector Machine." University of Dortmund, Computer Science. URL http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/index.html.

Schölkopf, B., C. Burges, and V. Vapnik. 1995. "Extracting Support Data for a Given Task." In *Proceedings, First International Conference on Konwledge Discovery and Data Mining*, eds. U. M. Fayyad and R. Uthurusamy, Menlo Park, CA: AAAI Press.

Schölkopf, B., C. Burges, and V. Vapnik. 1996. "Incorporating Invariances in Support Vector Learning Machines." In Proceedings International Conference on Artificial Neural Networks, ICANN'96. Springer Lecture Note in Computer Science, Vol. 1112, Berlin: Springer, pp. 47–52.

Schölkopf, B., A. J. Smola, R. C. Williamson, and P. L. Bartlett. 2000. "New Support Vector Algorithm." *Neural Computation*, **12**, pp. 1207–1245.

Schölkopf, B. and A. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.

Shawe-Taylor, J., P. L. Bartlett, R. C. Williamson, and M. Anthony. 1996. "A Framework for Structural Risk Minimization." In *Proceedings 9th Annual Conference on Computational Learning Theory* pp. 68–76.

Sing, T., O. Sander, N. Beerenwinkel, and T. Lengauer. 2005. "ROCR: Visualizing Classifier Performance in R. *Bioinformatics*, **21**(20), pp. 3940–3941.

Sivanandam, S. N. and S. N. Deepa. 2008. *Introduction to Genetic Algorithms*. Heidelberg: Springer-Verlag.

Sobehart, J., S. Keenan, and R. Stein. 2001. *Benchmarking Quantitative Default Risk Models: A Validation Methodology*. Moody Investors Service.

Sobehart, J., and S. Keenan. 2001. "Measuring Default Accurately." *Risk*, **14**, pp. 31–33.

Sobehart, J., and R. Stein. 2000. *Moody's Public Firm Risk Model: A Hybrid Approach to Modeling Short Term Default Risk*. Moody Investors Service, Rating Methodology.

Suykens, J. A. K. and J. Vandewalle. 1999. "Least Squares Support Vector Machine Classifiers." *Neural Processing Letters*, **9**, pp. 293–300.

Tam, K., and M. Kiang. 1992. "Managerial Applications of Neural Networks: The Case of Bank Failure Predictions." *Management Science*, **38**, pp. 926–947.

Turlach, B. A., and A. Weingessel. 2004. "quadprog: Functions to Solve Quadratic Programming Problems." http://CRAN.R-project.org/package=quadprog

Vanderbei, R. 1999. "LOQO: An Interior Point Code for Quadratic Programming." *Optimization Methods and Software*, **11**(1–4), pp. 451–484.

Vapnik, V. 1979. *Estimation of Dependencies Based on Empirical Data. Russian Version.* Moscow: Nauka.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer Verlag.

Vapnik, V. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons.

Vassalou, M. and Y. Xing. 2004. "Default Risk in Equity Returns." *The Journal of Finance* **19**(2), pp. 831–868.

Wilson, R. L. and R. Sharda. 1994. "Bankruptcy Prediction Using Neural Network." *Decision Support System*, **11**, pp. 545–557.

Zhang, J. L. and W. Härdle. 2010. "The Bayesian Additive Classification Tree Applied to Credit Risk Modelling." *Computational Statistics and Data Analysis*, **54**, pp. 1197–1205.

Zhou, X., and J. Xu. 2009 . "A SVM Model Selection Method Based on Hybrid Genetic Algorithm and Emprirical Error Minimization Criterion." In *Advances in Intelligent and Soft Computing. The Sixth International Symposium on Neural Networks.* Berlin: Springer, pp. 245–253.

Zhu, J., S. Rosset, T. Hastie, and R. Tibshirani. 2004. "1-Norm Support Vector Machines. In eds. S. Thrun, L. K. Saul, and B. Schölkopf. *Advances in Neural Information Processing System 16.* Cambridge, MA: MIT Press, pp. 49–56.

PART V

TIME SERIES

CHAPTER 12

SERIES ESTIMATION OF STOCHASTIC PROCESSES: RECENT DEVELOPMENTS AND ECONOMETRIC APPLICATIONS[†]

PETER C. B. PHILLIPS AND ZHIPENG LIAO

12.1. Introduction

The explicit representation of stochastic processes has a long history in the probability literature with many applications in asymptotic statistics. For example, in early work Kac and Siegert (1947) showed that a Gaussian process can be decomposed as an infinite linear combination of deterministic functions. In fact, a much more powerful representation theory holds for any stochastic process that is continuous in quadratic mean, a result that was separately established in Karhunen (1946) and Loève (1955). In the modern literature, the explicit decomposition of a stochastic process in this way is known as the Karhunen–Loève (KL) representation or transformation. The deterministic functions used in this KL representation are orthonormal basis functions in a Hilbert space constructed on the same interval for which the stochastic process is defined.

The KL transformation was originally proposed to assist in determining the exact forms of certain asymptotic distributions associated with Cramér–von Mises-type statistics. These asymptotic distributions typically take the form of a quadratic functional of a Brownian motion (BM) or Brownian Bridge process, such as the integral over some interval of the square of the process. For example, the KL transformation reveals that the integral of the square of a Gaussian process is distributed as a weighted infinite sum of independent chi-square variates with one degree of freedom. Other examples are given in the work of Anderson and Darling (1952), Watson (1962), and

Stephens (1976); and Shorack and Wellner (1988) provide an overview of results of this kind.

The theory underlying the KL representation relies on Mercer's theorem, which represents the covariance function of any quadratic mean continuous stochastic process $\{X_t\}_{t\in\mathcal{T}}$ in terms of basis functions in a Hilbert space $L^2(\mathcal{T})$ defined under some measure on $\mathcal{T}$. The covariance function can be viewed as an inner product of the Hilbert space $L^2(X)$ generated by the stochastic process.[1] On the other hand, by Mercer's theorem, the covariance function has a representation that defines an inner product with respect to another Hilbert space $L^2_R(\mathcal{T})$. This new Hilbert space $L^2_R(\mathcal{T})$ has the attractive feature that any function in the space can be reproduced by its inner product with the covariance function. As a result, $L^2_R(\mathcal{T})$ is often called a reproducing kernel Hilbert space (RKHS) with the covariance function being the reproducing kernel. It was noted in Parzen (1959) that the two Hilbert spaces $L^2(X)$ and $L^2_R(\mathcal{T})$ are isometrically isomorphic, which implies that analysis of the stochastic process $\{X_t\}_{t\in\mathcal{T}}$ in $L^2(X)$ can be equivalently executed in $L^2_R(\mathcal{T})$. Sometimes a complicated problem in $L^2(X)$ space can be treated more easily in the RKHS space $L^2_R(\mathcal{T})$. More details about the analysis of time series in RKHS space can be found in Parzen (1959, 1961a, 1961b, 1963). Berlinet and Thomas-Agnan (2003) provide a modern introduction to RKHS techniques and their applications in statistics and probability.

While statisticians and probabilists have focused on the roles of the KL representation in determining asymptotic distributions of functionals of stochastic processes or rephrasing time series analysis issues equivalently in different spaces, econometric research has taken these representations in a new direction. In particular, econometricians have discovered that empirical versions of the KL representation are a powerful tool for estimation and inference in many econometric models. This chapter reviews some of these recent developments of the KL representation theory and its empirical application in econometrics.

First, the KL representation provides a bridging mechanism that links underlying stochastic trends with various empirical representations in terms of deterministic trend functions. This mechanism reveals the channel by which the presence of deterministic trends in a regression can affect tests involving stochastic trends, such as unit root and cointegration tests. For example, Phillips (2001) showed how the asymptotic distributions of coefficient-based unit root test statistics are changed in a material way as deterministic function regressors continue to be added to the empirical regression model. This work used KL theory to show that as the number of deterministic functions tends to infinity, the coefficient-based unit root tests have asymptotic normal distributions after appropriate centering and scaling rather than conventional unit root distributions. These new asymptotics are useful in revising traditional unit root limit theory and ensuring that tests have size that is robust to the inclusion of many deterministic trend functions or trajectory fitting by deterministic trends or trend breaks.

Secondly, the KL theory not only directly represents stochastic trends in terms of deterministic trends, it also provides a basis for linking independent stochastic trends.

This extension of the theory was studied in Phillips (1998), where it was established that a continuous deterministic function can be approximated using linear combinations of independent BMs with a corresponding result for the approximation of a continuous stochastic process. This latter result is particularly useful in analyzing and interpreting so-called spurious regressions involving the regression of an integrated process on other (possibly independent) integrated processes.

The KL theory and its empirical extensions in Phillips (1998) explain how regression of an integrated process on a set of basis functions can successfully reproduce the whole process when the number of basis functions expands to infinity with the sample size. An empirically important implication of this result that is explored in Phillips (2013) is that trend basis functions can themselves serve as instrumental variables because they satisfy both orthogonality and relevance conditions in nonstationary regression. For instance, in a cointegrated system, this type of trend IV estimator of the cointegrating matrix does not suffer from high-order bias problems because the basis functions are independent of the errors in the cointegrated system by virtue of their construction, thereby delivering natural orthogonality. Moreover, the IV estimator is asymptotically efficient because when the number of basis functions diverges to infinity, the integrated regressors in the cointegrating system are reproduced by the basis functions, thereby assuring complete relevance in the limit. In short, the long-run behavior of the endogenous variables in a cointegrated system is fully captured through a linear projection on basis functions in the limit while maintaining orthogonality of the instruments.

As the above discussion outlines, KL theory helps to answer questions about the asymptotic behavior of linear projections of integrated processes on deterministic bases. A related question relates to the properties of similar projections of the trajectory of a stationary process on deterministic bases. In exploring this question, Phillips (2005b) proposed a new estimator of the long-run variance (LRV) of a stationary time series. This type of estimator is by nature a series estimate of the LRV and has since been extensively studied in Chen, Liao, and Sun (2012), Chen, Hahn, and Liao (2012), Sun (2011, 2013), and Sun and Kim (2012, 2013).

The remainder of this chapter is organized as follows. Section 12.2 presents the KL representation theory for continuous stochastic processes together with some recent developments of this theory. Section 12.3 explores the implications of the KL theory for empirical practice, focusing on understanding and interpreting spurious regressions in econometrics. Section 12.4 investigates the implication of these representations for unit root tests when there are deterministic trends in the model. Section 12.5 considers the optimal estimation of cointegrated systems using basis functions as instruments. The optimal estimation method discussed in Section 12.5 assumes that the cointegration space of the cointegration system is known from the beginning. In Section 12.6, we present a new method that optimally estimates the cointegration system without even knowing the cointegration rank. Series estimation of LRVs and some of the recent applications of this theory are discussed in Section 12.7. Section 12.8 concludes and briefly describes some ongoing and future research in the field. Technical derivations are included in the Appendix.

12.2. Orthogonal Representation of Stochastic Processes

We start with a motivating discussion in Euclidean space concerned with the orthonormal representation of finite-dimensional random vectors. Such representations provide useful intuition concerning the infinite-dimensional case and are indicative of the construction of orthonormal representations of stochastic processes in Hilbert space.

Suppose X is a T-dimensional random vector with mean zero and positive definite covariance matrix Σ. Let $\{(\lambda_k, \varphi_k)\}_{k=1}^T$ be the pairs of eigenvalues and orthonormalized right eigenvectors of Σ. Define

$$Z_T' = X'\Phi_T = [z_1, \ldots, z_T],$$

where $\Phi_T = [\varphi_1, \ldots, \varphi_T]$, then Z_T is a T-dimensional random vector with mean zero and covariance matrix $\Lambda_T = \text{diag}(\lambda_1, \ldots, \lambda_T)$. We have the representation

$$X = \Phi_T Z_T = \sum_{k=1}^{T} z_k \varphi_k = \sum_{k=1}^{T} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k, \tag{12.1}$$

where the $\xi_k = \lambda_k^{-\frac{1}{2}} z_k$ have zero mean and covariances $E[\xi_k \xi_{k'}] = \delta_{kk'}$ where $\delta_{kk'}$ is the Kronecker delta. When X is a zero mean Gaussian random vector, $[\xi_1, \ldots, \xi_T]'$ is simply a T-dimensional standard Gaussian random vector. Expression (12.1) indicates that any T-dimensional ($T \in \mathbb{Z}_+ \equiv \{1, 2, \ldots,\}$) random vector can be represented by a weighted linear combination of T orthonormal real vectors, where the weights are random and uncorrelated across different vectors. Moreover, (12.1) shows that the spectrum of the covariance matrix of the random vector X plays a key role in the decomposition of X into a linear combination of deterministic functions with random coefficients.

The orthonormal representation of a random vector given in (12.1) can be generalized to a stochastic process $X(t)$ with $t \in [a, b]$ for $\infty < a < b < \infty$, and in this form it is known as the Kac–Siegert decomposition or KL representation. We can use heuristics based on those used to derive (12.1) to develop the corresponding KL representation of a general stochastic process. Without loss of generality, we assume that the random variables $\{X(t) : t \in [a, b]\}$ live on the same probability space $(\Omega, \mathcal{G}, P)$. The first and second moments of $X(t)$ for any $t \in [a, b]$ are given by

$$E[X(t)] = \int_\Omega X(t) dP \quad \text{and} \quad E\big[X^2(t)\big] = \int_\Omega X^2(t) dP.$$

The following assumption is used to derive the KL representation of $X(t)$.

Assumption 12.1. *The stochastic process $X(t)$ satisfies $E[X(t)] = 0$ and $E\big[X^2(t)\big] < \infty$ for all $t \in [a, b]$.*

The zero mean assumption is innocuous because the process $X(t)$ can always be recentered about its mean. The second moment assumption is important because it allows us to embed $X(t)$ in a Hilbert space and use the Hilbert space setting to establish the representation. Accordingly, let $L^2(X)$ denote the Hilbert space naturally generated by $X(t)$ so that it is equipped with the following inner product and semi-norm:

$$\langle X_1, X_2 \rangle \equiv \int_\Omega X_1 X_2 \, dP \quad \text{and} \quad \|X_1\|^2 = \int_\Omega X_1^2 \, dP,$$

for any $X_1, X_2 \in L^2(X)$. Let $L^2[a,b]$ be the Hilbert space of square integrable functions on $[a,b]$ with the following inner product and semi-norm

$$\langle g_1, g_2 \rangle_e \equiv \int_a^b g_1(s) g_2(s) \, ds \text{ and } \|g_1\|_e^2 = \int_a^b g_1^2(s) \, ds, \tag{12.2}$$

for any $g_1, g_2 \in L^2[a,b]$.

Under Assumption 12.1, the covariance/kernel function $\gamma(\cdot,\cdot)$ of the stochastic process $X(t)$ can be defined as

$$\gamma(s,t) \equiv E[X(s)X(t)] \tag{12.3}$$

for any $s, t \in [a,b]$. Let $\{(\lambda_k, \varphi_k)\}_{k \in K}$ be the collection of all different pairs (λ, φ) which satisfy the following integral equation:

$$\lambda \varphi(t) = \int_a^b \gamma(s,t) \varphi(s) \, ds \qquad \text{with } \|\varphi\|_e = 1, \tag{12.4}$$

where λ and φ are called the eigenvalue and normalized eigenfunction of the kernel $\gamma(\cdot,\cdot)$, respectively.

Using heuristics based on the procedure involved in deriving (12.1), one might expect to use the eigenvalues and eigenfunctions of the kernel function $\gamma(\cdot,\cdot)$ to represent the stochastic process $X(t)$ as a sum of the form

$$X(t) \overset{?}{=} \sum_{k=1}^{\bar{K}} \left[\int_a^b X(t) \varphi_k(t) \, dt \right] \varphi_k(t) = \sum_{k=1}^{\bar{K}} z_k \varphi_k(t) = \sum_{k=1}^{\bar{K}} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k(t), \tag{12.5}$$

where $z_k \equiv \int_a^b X(t) \varphi_k(t) \, dt$ and $\xi_k \equiv \lambda_k^{-\frac{1}{2}} z_k$ for $k = 1, \ldots, \bar{K}$ and some (possibly infinite) $\bar{K}$. To ensure that the expression in (12.5) is indeed an orthonormal representation of $X(t)$, we first confirm that the components ξ_k satisfy

$$E[\xi_k] = 0 \quad \text{and} \quad E[\xi_k \xi_{k'}] = \delta_{kk'} \qquad \text{for any } k, k' = 1, \ldots, \bar{K}, \tag{12.6}$$

where $\delta_{kk'}$ is Kronecker's delta, and that the process $X(t)$ can be written as

$$X(t) = \sum_{k=1}^{\bar{K}} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k(t) \qquad \text{a.s. } t \in [a,b] \text{ in quadratic mean.} \tag{12.7}$$

The following condition is sufficient to show (12.6) and (12.7).

Assumption 12.2. *The stochastic process $X(t)$ is continuous in quadratic mean (q.m.) on $[a,b]$; that is, for any $t_o \in [a,b]$ we have*

$$\|X(t) - X(t_o)\|^2 = E\big\{[X(t) - X(t_o)]^2\big\} \to 0 \tag{12.8}$$

as $|t - t_o| \to 0$, where we require $t \in [a,b]$ such that $X(t)$ is well defined in (12.8).

In this assumption, continuity in q.m. is well-defined at the boundary points a and b because we only need to consider the limits from the right to a and limits from the left to b. The following lemma is useful in deriving the KL representation of $X(t)$.

Lemma 12.1. *Suppose that Assumptions 12.1 and 12.2 are satisfied. Then the kernel function $\gamma(\cdot,\cdot)$ of the stochastic process $X(t)$ is symmetric, continuous, and bounded and it satisfies*

$$\int_a^b \int_a^b g(t)\gamma(t,s)g(s)\, dsdt \geq 0$$

for any $g \in L^2[a,b]$.

Under Assumptions 12.1 and 12.2, Lemma 12.1 implies that sufficient conditions for Mercer's theorem hold (see, e.g., Shorack and Wellner (1986, p. 208)). Thus, we can invoke Mercer's theorem to deduce that the normalized eigenfunctions of the kernel function $\gamma(\cdot,\cdot)$ are continuous on $[a,b]$ and form an orthonormal basis for the Hilbert space $L^2[a,b]$. Mercer's theorem ensures that the kernel function $\gamma(\cdot,\cdot)$ has the following series representation in terms of this orthonormal basis

$$\gamma(s,t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s)\varphi_k(t) \tag{12.9}$$

uniformly in s and t. The following theorem justifies the orthonormal representation of $X(t)$ in (12.5) with $\bar{K} = \infty$ and (12.6) and (12.7) both holding.

Theorem 12.1. *Suppose the stochastic process $X(t)$ satisfies Assumptions 12.1 and 12.2. Then $X(t)$ has the following orthogonal expansion*

$$X(t) = \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k(t) \qquad \text{with } \xi_k = \lambda_k^{-\frac{1}{2}} \int_a^b X(t)\varphi_k(t)dt, \tag{12.10}$$

where $E[\xi_k \xi_{k'}] = \int_a^b \varphi_k(t)\varphi_{k'}(t)dt = \delta_{kk'}$ and $\delta_{kk'}$ denotes the Kronecker delta, if and only if λ_k and φ_k ($k \in \mathbb{Z}_+$) are the eigenvalues and normalized eigenfunctions of $\gamma(\cdot,\cdot)$. The series in (12.10) converges in q.m. uniformly on $[a,b]$.

Just as a continuous function in $L^2[a,b]$ can be represented by series involving Fourier basis functions, Theorem 12.1 indicates that a continuous (in q.m.) stochastic process can also be represented by orthonormal basis functions that lie in $L^2[a,b]$. However, unlike the series representation of a continuous function, the coefficients of the basis functions in the KL representation are random variables and uncorrelated with each other. The representation of $X(t)$ in (12.10) converges in q.m. but may not necessarily converge pointwise.[2] For this reason, the equivalence in (12.10) is sometimes represented by the symbol "$\sim$" or "$\overset{d}{=}$", signifying that the series is convergent in the L^2 sense and that distributional equivalence applies. Importantly, the series (12.10) involves two sets of orthonormal components: the orthogonal random sequence $\{\xi_k\}$ and the orthogonal basis functions $\{\varphi_k\}$.

When the continuous time stochastic process $X(t)$ is covariance stationary, it is well known that $X(t)$ has the following spectral (SP) representation:

$$X(t) = \int_{-\infty}^{+\infty} \exp(i\lambda t)\, dZ(\lambda), \tag{12.11}$$

where i is the imaginary unit and $Z(\lambda)$ denotes the related complex spectral process that has orthogonal increments whose variances involve the corresponding increments in the spectral distribution function. In expression (12.11), $X(t)$ is represented as an uncountably infinite sum of the products of deterministic functions $\exp(i\lambda t)$ and random coefficients $dZ(\lambda)$ at different frequencies, which differs from the KL expression (12.10) in several ways. Most importantly, (12.10) represents in quadratic mean the trajectory of the process over a fixed interval $[a,b]$, whereas (12.11) is a representation of the entire stochastic process $X(t)$ in terms of the mean square limit of approximating Riemann Stieltjes sums (e.g., Hannan (1970, p. 41)).

When the stochastic process $X(t)$ is a BM, its KL representation has more structure. For example, the representation in (12.10) holds almost surely and uniformly in $[0,1]$ and the random coefficients $\{\xi_k\}$ are *i.i.d.* normal. These special structures are summarized in the following corollary.

Corollary 12.1. *Let $B_\sigma(t)$ be a BM with variance σ^2, then (i) $B_\sigma(t)$ has the following orthogonal expansion*

$$B_\sigma(t) = \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k(t), \tag{12.12}$$

where

$$\xi_k = \lambda_k^{-\frac{1}{2}} \int_a^b B_\sigma(t)\varphi_k(t)\, dt \tag{12.13}$$

and the above representation converges almost surely uniformly on $[a,b]$; (ii) the random sequence $\{\xi_k\}_k$ is i.i.d. $N(0,\sigma^2)$; *(iii) the random sequence $\{\eta_k\}_k$ defined by*

$$\eta_k = \int_a^b \varphi_k(t)\, dB_\sigma(t) \tag{12.14}$$

is also i.i.d. $N(0,\sigma^2)$.

It is easy to verify that $B_\sigma(t)$ satisfies Assumptions 12.1 and 12.2. Thus by Theorem 12.1, $B_\sigma(t)$ has a KL representation which converges in q.m. uniformly on $[a,b]$. The q.m. convergence of the series in (12.9) is strengthened to almost sure convergence in (12.12) by applying the martingale convergence theorem to the martingale formed by finite sums of (12.12). The normality of ξ_k or η_k ($k \in \mathbb{Z}_+$) holds directly in view of the representations (12.13) and (12.14) (the normal stability theorem, Loève, 1977) and the independence of the sequence $\{\xi_k\}$ or $\{\eta_k\}$ follows by their orthogonality. It is clear that the expression in (12.10) links the stochastic trend $X(t)$ with a set of deterministic functions $\{\varphi_k(\cdot)\}_{k=1}^{\infty}$ which might be regarded as trend functions on the interval $[a,b]$. Since the random wandering behavior of the stochastic trend $X(t)$ over $[a,b]$ is fully captured by the deterministic functions in its KL representation, throughout this chapter we shall call $\{\varphi_k(\cdot) : k \in \mathbb{Z}_+\}$ the trend basis functions.

Example 12.1. *Let $B(\cdot)$ be a standard BM on $[0,1]$. Then Corollary 12.1 ensures that $B(\cdot)$ has a KL representation. By definition, the kernel function of $B(\cdot)$ is $\gamma(s,t) = \min(s,t)$ and its eigenvalues and normalized eigenfunctions are characterized by the following integral equation*

$$\lambda\varphi(t) = \int_0^t s\varphi(s)\,ds + t\int_t^1 \varphi(s)\,ds \qquad \text{with} \quad \int_0^1 \varphi^2(s)\,ds = 1.$$

Direct calculation reveals that the eigenvalues and normalized eigenfunctions of $\gamma(\cdot,\cdot)$ are

$$\lambda_k = \frac{1}{(k-1/2)^2\pi^2} \quad \text{and} \quad \varphi_k(t) = \sqrt{2}\sin[(k-1/2)\pi t] \tag{12.15}$$

respectively for $k \in \mathbb{Z}_+$. Applying Corollary 12.1, we have the following explicit orthonormal representation:

$$B(t) = \sqrt{2}\sum_{k=1}^{\infty} \frac{\sin[(k-1/2)\pi t]}{(k-1/2)\pi}\xi_k, \tag{12.16}$$

which holds almost surely and uniformly in $t \in [0,1]$, where

$$\xi_k = \sqrt{2}(k-1/2)\pi \int_0^1 B(t)\sin[(k-1/2)\pi t]dt \qquad \text{for } k \in \mathbb{Z}_+. \tag{12.17}$$

Invoking Corollary 12.1, we know that $\{\xi_k\}_{k=1}^{\infty}$ are i.i.d. standard normal random variables.

Example 12.2. *Let $W(\cdot)$ be a Brownian bridge process corresponding to the standard BM $B(\cdot)$ on $[0,1]$, that is, $W(t) = B(t) - tB(1)$ for any $t \in [0,1]$. It is easy to show that $W(\cdot)$ is continuous in q.m. on $[0,1]$. Moreover, $W(\cdot)$ has kernel function $\gamma(s,t) = \min(s,t) - st$, which is continuous on $[0,1]$. The eigenvalues and normalized eigenfunctions are characterized by the following integral equation*

$$\lambda\varphi(t) = \int_0^t s\varphi(s)\,ds + t\int_t^1 \varphi(s)\,ds - \frac{t}{2} \qquad \text{with} \quad \int_0^1 \varphi^2(s)\,ds = 1.$$

Direct calculation shows that the eigenvalues and normalized eigenfunctions of $\gamma(\cdot,\cdot)$ *are*

$$\lambda_k = \frac{1}{k^2\pi^2} \quad \textit{and} \quad \varphi_k(t) = \sqrt{2}\sin(k\pi t),$$

respectively, for $k \in \mathbb{Z}_+$*. Applying Theorem 12.1, we have the following orthonormal representation*

$$W(t) = \sqrt{2}\sum_{k=1}^{\infty} \frac{\sin(k\pi t)}{k\pi}\xi_k \tag{12.18}$$

where

$$\xi_k = \sqrt{2}k\pi \int_0^1 B(t)\sin(k\pi t)\, dt \qquad \textit{for } k \in \mathbb{Z}_+. \tag{12.19}$$

Using similar arguments as those in Corollary 12.1, the representation in (12.18) is convergent almost surely and uniformly in $t \in [0,1]$*. Moreover,* $\{\xi_k\}_{k=1}^{\infty}$ *are i.i.d. standard normal random variables.*

The KL representation of a BM can be used to decompose other stochastic processes that are functionals of BMs. The simplest example is the Brownian bridge process studied in the above example. From the representation in (12.16),

$$W(t) = B(t) - tB(1) = \sqrt{2}\sum_{k=1}^{\infty} \frac{\sin[(k-1/2)\pi t] + (-1)^k t}{(k-1/2)\pi}\xi_{1,k}$$

where $\xi_{1,k}$ ($k \in \mathbb{Z}_+$) is defined in (12.17). Of course, one can also use the KL representation of the Brownian bridge process to decompose the process $B(t)$ into a series form, namely,

$$B(t) = tB(1) + W(t) = t\xi_{2,0} + \sqrt{2}\sum_{k=1}^{\infty} \frac{\sin(k\pi t)}{k\pi}\xi_{2,k} \tag{12.20}$$

where $\xi_{2,0} = B(1)$ and the $\xi_{2,k}$ ($k \in \mathbb{Z}_+$) are defined in (12.19).

The second example is the quadratic functional of a BM given by the integral $[B]_1 = \int_0^1 B^2(t)\, dt$. Using the KL representation (12.16), the following series expression for the functional is readily obtained:

$$[B]_1 = \int_0^1 B^2(t)\, dt = \sum_{k=1}^{\infty} \frac{1}{(k-1/2)^2\pi^2}\xi_k^2,$$

which implies that the random variable $[B]_1$ has a distribution equivalent to the weighted sum of independent chi-square random variables, each with unit degree of freedom.

The third example is the series representation of an Ornstein-Uhlenbeck (O-U) process. We provide two illustrations of how to construct such as series.

Example 12.3. *Let $J_c(t)$ be a stochastic process on $t \in [0,1]$ satisfying the following stochastic differential equation*

$$dJ_c(t) = cJ_c(t)\,dt + \sigma\,dB(t) \tag{12.21}$$

where c and $\sigma > 0$ are constants and $B(\cdot)$ denotes a standard BM. Set $\sigma = 1$ for convenience in what follows. It is clear that when $c = 0$, the process $J_c(t)$ reduces to standard BM $B(t)$. Under the initial condition $J_c(0) = B(0) = 0$, the above differential equation has the following solution

$$J_c(t) = B(t) + c\int_0^t \exp\left[(t-s)c\right]B(s)\,ds. \tag{12.22}$$

Using the series representation (12.20) and the solution (12.22), one obtains for $t \in [0,1]$

$$\begin{aligned} J_c(t) &= \frac{e^{ct}-1}{c}\xi_{2,0} + \sum_{k=1}^{\infty}\left[\sqrt{2}e^{ct}\int_0^t e^{-cs}\cos(k\pi s)ds\right]\xi_k \\ &= \frac{e^{ct}-1}{c}\xi_{2,0} + \sqrt{2}\sum_{k=1}^{\infty}\frac{ce^{ct} + k\pi\sin(k\pi t) - c\cos(k\pi t)}{c^2+k^2\pi^2}\xi_k, \end{aligned} \tag{12.23}$$

where ξ_k ($k \in \mathbb{Z}_+$) are i.i.d. standard normal random variables. The series representation (12.23) involves the orthogonal sequence $\{\xi_k\}$ associated with the Brownian bridge $W(t)$. An alternative representation that uses the series (12.16) is given in Phillips (1998) and in (12.72) below.

Example 12.4. *Suppose $X(t)$ is an O-U process with covariance kernel $\gamma(s,t) = e^{-|s-t|}$. In this case the process $X(t)$ satisfies the stochastic differential equation (12.21) with $c = -1$ and $\sigma = \sqrt{2}$. Then the KL representation of $X(t)$ over $t \in [0,1]$ is*

$$X(t) = \sqrt{2}\sum_{k=0}^{\infty}\frac{\sin\left\{\omega_k\left(t-\frac{1}{2}\right) + (k+1)\frac{\pi}{2}\right)\}}{(1+\lambda_k)^{1/2}}\xi_k, \tag{12.24}$$

where ξ_k ($k \in \mathbb{Z}_+$) are i.i.d. standard normal random variables, $\lambda_k = 2\left(1+\omega_k^2\right)^{-1}$, and $\omega_0, \omega_1, \ldots$ are the positive roots of the equation

$$\tan(\omega) = -2\frac{\omega}{1-\omega^2}$$

(Pugachev, 1959; see also Bosq, 2000, p. 27).

12.3. New Tools for Understanding Spurious Regression

Spurious regression refers to the phenomenon that arises when fitted least squares regression coefficients appear statistically significant even when there is no true relationship between the dependent variable and the regressors. In simulation studies, Granger and Newbold (1974) showed that the phenomenon occurs when independent random walks are regressed on one another. Similar phenomena occur in regressions of stochastic trends on deterministic polynomial regressors, as shown in Durlauf and Phillips (1988). Phenomena of this kind were originally investigated by Yule (1926) and the first analytic treatment and explanation was provided in Phillips (1986).

As seen in the previous section, the orthonormal representation (12.10) links the random function $X(\cdot)$ to deterministic basis functions $\varphi_j(\cdot)$ $(j \in \mathbb{Z}_+)$ on the Hilbert space $L^2[a,b]$. This linkage provides a powerful tool for studying relations between stochastic trends and deterministic trends, as demonstrated in Phillips (1998). The orthonormal representation (12.10) also provides useful insights in studying relations among stochastic trends.

Consider the normalized time series $B_n\left(\frac{t}{n}\right) = n^{-\frac{1}{2}} \sum_{s=1}^{t} u_s$, whose components u_t satisfy the following assumption.

Assumption 12.3. *For all* $t \geq 0$, u_t *has Wold representation*

$$u_t = C(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} j\left|c_j\right| < \infty \quad \textit{and} \quad C(1) \neq 0 \tag{12.25}$$

with $\varepsilon_t = \text{i.i.d. } (0, \sigma_\varepsilon^2)$ *with* $E\left(|\varepsilon_t|^p\right) < \infty$ *for some* $p > 2$.

Under the above assumption, one can invoke Lemma 3.1 of Phillips (2007), which shows that in a possibly expanded probability space we have the (in probability) approximation

$$\sup_{0 \leq t \leq n} \left| B_n\left(\frac{t}{n}\right) - B_{\sigma_u}\left(\frac{t}{n}\right) \right| = o_p(n^{-\frac{1}{2}+\frac{1}{p}}), \tag{12.26}$$

where $B_{\sigma_u}(\cdot)$ denotes a BM with variance $\sigma_u^2 = 2\pi f_u(0)$ and $f_u(\cdot)$ is the spectral density of u_t. Using the KL representation[3] in (12.12) and the uniform approximation in (12.26), we can deduce that

$$\sup_{0 \leq t \leq n} \left| B_n\left(\frac{t}{n}\right) - \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}} \varphi_k\left(\frac{t}{n}\right) \xi_k \right| = o_p(1), \tag{12.27}$$

where $\{(\lambda_k, \varphi_k(\cdot))\}_{k=1}^{\infty}$ is the set of all pairs of eigenvalues and orthonormalized eigenfunctions of the kernel function $\gamma(s,t) = \sigma_u^2 \min(s,t)$, and where ξ_k $(k \in \mathbb{Z}_+)$ are independent Gaussian random variables.

The result in (12.26) implies that the scaled partial sum $B_n\left(\frac{t}{n}\right) = n^{-\frac{1}{2}} \sum_{s=1}^{t} u_s$ can be uniformly represented in terms of the basis functions $\varphi_k(\cdot)$ ($k \in \mathbb{Z}_+$) in $L^2[a,b]$ for all $t \leq n$. Such a uniform approximation motivates us to study empirical LS regression estimation in which the scaled partial sum $B_n\left(\frac{t}{n}\right)$ is fitted using K orthonormal basis functions $\varphi_k(\cdot)$ ($k = 1,\ldots,K$), that is,

$$B_n\left(\frac{t}{n}\right) = \sum_{k=1}^{K} \widehat{a}_{k,n} \varphi_k\left(\frac{t}{n}\right) + \widehat{u}_{t,K}, \tag{12.28}$$

where

$$\widehat{A}_K = (\widehat{a}_{1,n},\ldots,\widehat{a}_{K,n})' = \left[\sum_{t=1}^{n} \Phi_K\left(\frac{t}{n}\right)\Phi_K'\left(\frac{t}{n}\right)\right]^{-1}\left[\sum_{t=1}^{n} \Phi_K\left(\frac{t}{n}\right) B_n\left(\frac{t}{n}\right)\right]$$

and $\Phi_K(\cdot) = [\varphi_1(\cdot),\ldots,\varphi_K(\cdot)]$. There are several interesting questions we would like to ask about the regression in (12.28). First, what are the asymptotic properties of the estimator $\widehat{A}_K$? More specifically, if we rewrite the uniform approximation (12.27) in the form

$$B_n\left(\frac{t}{n}\right) = \Phi_K\left(\frac{t}{n}\right)\Lambda_K \overline{\xi}_K + \sum_{k=K+1}^{\infty} \lambda_k^{\frac{1}{2}} \varphi_k\left(\frac{t}{n}\right)\xi_k,$$

where $\Lambda_K \equiv \operatorname{diag}(\lambda_1,\ldots,\lambda_K)$ and $\overline{\xi}_K = (\xi_1,\ldots,\xi_K)$, will the estimate $\widehat{A}_K$ replicate the random vector $\Lambda_K \overline{\xi}_K$ in the limit? In practical work an econometrician might specify a regression that represents an integrated time series such as $y_t = \sum_{s=1}^{t} u_s$ in terms of deterministic trends. Upon scaling, such a regression takes the form

$$B_n\left(\frac{t}{n}\right) = \Phi_K\left(\frac{t}{n}\right) A_{o,K} + v_{nk}, \tag{12.29}$$

which may be fitted by least squares to achieve trend elimination. To test the significance of the regressors $\Phi_K(\cdot)$ in such a trend regression, a natural approach would be to use a t-statistic for a linear combination of the coefficients $c_K' A_{o,K}$, such as

$$t_{c_K'\widehat{A}_K} = \frac{c_K'\widehat{A}_K}{\sqrt{\left(n^{-1}\sum_{i=1}^{n} \widehat{u}_{t,K}^2\right) c_K' \left[\sum_{t=1}^{n} \Phi_K\left(\frac{t}{n}\right)\Phi_K'\left(\frac{t}{n}\right)\right]^{-1} c_K}}$$

for any $c_K \in R^K$ with $c_K' c_K = 1$. Corresponding robust versions of $t_{c_K'\widehat{A}_K}$ using conventional HAC or HAR estimates of the variance of $c_K'\widehat{A}_K$ might also be used, options that we will discuss later. For now, what are the asymptotic properties of the statistic $t_{c_K'\widehat{A}_K}$ and how adequate is the test? Further, we might be interested in measuring goodness of fit using the estimated coefficient of determination

$$\widehat{R}_K^2 = \frac{\widehat{A}_K'\left[\sum\limits_{t=1}^{n} \Phi_K\left(\frac{t}{n}\right)\Phi_K'\left(\frac{t}{n}\right)\right]\widehat{A}_K}{n^{-1}\sum_{t=1}^{n} B_n^2\left(\frac{t}{n}\right)}.$$

What are the asymptotic properties of $\widehat{R}_K^2$ and how useful is this statistic as a measure of goodness of fit in the regression? The following theorem from Phillips (1998) answers these questions.

Theorem 12.2. *As $n \to \infty$, we have*

(a) $c_K'\widehat{A}_K \to_d c_K' \int_0^1 \Phi_K(r)B(r)\,dr \overset{d}{=} N\big(0, c_K'\Lambda_K c_K\big)$,

(b) $n^{-\frac{1}{2}} t_{c_K'\widehat{A}_K} \to_d c_K'\left[\int_0^1 \Phi_K(r)B(r)\,dr\right]\left[\int_0^1 B_{\varphi_K}^2(r)\,dr\right]^{-\frac{1}{2}}$,

(c) $\widehat{R}_K^2 \to_d 1 - \left[\int_0^1 B_{\varphi_K}^2(r)\,dr\right]\left[\int_0^1 B^2(r)\,dr\right]^{-1}$,

where $B_{\varphi_K}(\cdot) = B(\cdot) - \left[\int_0^1 B(r)\Phi_K(r)\,dr\right]\Phi_K'(\cdot)$ *is the projection residual of* $B(\cdot)$ *on* $\Phi_K(\cdot)$.

Theorem 12.2 explains the spurious regression phenomenon that arises when an integrated process is regressed on a set of trend basis functions. Part (a) implies that the OLS estimate $\widehat{a}_{k,n}$ has a limit that is equivalent to $\lambda_k^{\frac{1}{2}}\xi_k$ for $k = 1, \ldots, K$. Note that the weak convergence in part (a) leads to pointwise functional limits. In particular, it leads directly to the following pointwise functional convergence:

$$\Phi_K(t)\widehat{A}_K \to_d \sum_{k=1}^{K} \lambda_k^{\frac{1}{2}}\varphi_k(t)\xi_k, \qquad \text{for any } t \in [0,1]. \tag{12.30}$$

A corresponding uniform weak approximation, that is,

$$\sup_{t\in[0,1]}\left|\Phi_K(t)\widehat{A}_K - \sum_{k=1}^{K}\lambda_k^{\frac{1}{2}}\varphi_k(t)\xi_k\right| = o_p(1), \tag{12.31}$$

can be proved using bracketing entropy arguments and the rate of pointwise convergence in (12.30). We leave the theoretical justification of such a uniform approximation to future research. Part (b) confirms that trend basis functions are always significant when used in regressions to explain an integrated process because the related t-statistics always diverge as the sample size $n \to \infty$.[4] From the KL representation (12.10), we observe that for large K the Hilbert space projection residual $B_{\varphi_K}(\cdot)$ is close to zero with high probability. From part (c), we see that in such a case, $\widehat{R}_K^2$ is also close to 1 with large probability.

The results in Theorem 12.2 are derived under the assumption that the number of trend basis functions is fixed. A natural question to ask is, What are the asymptotic properties of $c_K'\widehat{A}_K$, $t_{c_K'\widehat{A}_K}$, and $\widehat{R}_K^2$ if the number of the trend basis functions K diverges to infinity with the sample size n. Note that if $K \to \infty$, then

$$\left[\int_0^1 B(r)\Phi_K(r)\,dr\right]\Phi_K'(t) = \sum_{k=1}^{K} \lambda_k^{\frac{1}{2}}\xi_k\varphi_k(t) \to_{a.s.} \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}}\xi_k\varphi_k(t) = B(t), \tag{12.32}$$

where the almost sure convergence follows by the martingale convergence theorem. The convergence in (12.32) immediately implies

$$B_{\varphi_K}(t) = B(t) - \left[\int_0^1 B(r)\Phi_K(r)\,dr\right]\Phi_K'(t) \to_{a.s.} 0 \tag{12.33}$$

as $K \to \infty$. Now, using (12.33) and sequential asymptotic arguments, we deduce that

$$\Phi_K(t)\widehat{A}_K \to_d \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}}\xi_k\varphi_k(t) = B(t), \tag{12.34}$$

$$\left|n^{-\frac{1}{2}} t_{c_K'\widehat{A}_K}\right| \to_p \infty \quad \text{and} \quad \widehat{R}_K^2 \to_p 1, \tag{12.35}$$

as $n \to \infty$ followed by $K \to \infty$. The result (12.34) indicates that the fitted value $\Phi_K(\cdot)\widehat{A}_K$ based on the OLS estimate $\widehat{A}_K$ fully replicates the BM $B(\cdot)$ as K goes to infinity. Moreover, (12.34) implies that all fitted coefficients are significant even when infinitely many trend basis functions are used in (12.27). Note that when more trend basis functions are added to the regression, the fitted coefficients become more significant, instead of being less significant, because the residual variance in the regression (12.28) converges to zero in probability when both K and n diverge to infinity. The second result in (12.35) implies that the model is perfectly fitted when $K \to \infty$, which is anticipated in view of (12.34).

The following theorem is due to Phillips (1998) and presents asymptotic properties of $c_K'\widehat{A}_K$, $t_{c_K'\widehat{A}_K}$ and $\widehat{R}_K^2$ under joint asymptotics when n and K pass to infinity jointly.

Theorem 12.3. *Suppose that $K \to \infty$, then $c_K'\Lambda_K c_K$ converges to a positive constant $\sigma_c^2 = c'\Lambda c$, where $c = (c_1, c_2, \ldots)$, $\Lambda \equiv diag(\lambda_1, \lambda_2, \ldots)$, and $c'c = 1$. Moreover, if $K \to \infty$ and $K/n \to 0$ as $n \to \infty$, then we have (a) $c_K'\widehat{A}_K \to_d N(0, \sigma_c^2)$, (b) $n^{-\frac{1}{2}} t_{c_K'\widehat{a}_K}$ diverges, and (c) $\widehat{R}_K^2 \to_p 1$.*

From Theorem 12.3 it follows that the asymptotic properties of $c_K'\widehat{A}_K$, $t_{c_K'\widehat{A}_K}$, and $\widehat{R}_K^2$ under joint limits are very similar to their sequential asymptotic properties. Thus, the above discussion about the results in (12.34) and (12.35) also applies to Theorem 12.3.

As this analysis shows, the KL representation is a powerful tool in interpreting regressions of stochastic trends on deterministic trends. The KL representation can also link different BMs, because different BMs can themselves each be represented in terms of the same set of orthonormal basis functions. This intuition explains spurious regressions that arise when an integrated process is regressed on other (possibly independent) integrated processes. The following theorem, again from Phillips (1998), indicates that any BM can be represented in terms of infinitely many independent standard BMs. This theory assists our understanding of empirical regressions among integrated processes that may be of full rank (or non-cointegrating). Such regressions are considered prototypical spurious regressions following the simulation study of Granger and Newbold (1974).

Theorem 12.4. *Let $B_\sigma(\cdot)$ be a BM on $[0,1]$ with variance σ^2 and let $\varepsilon > 0$ be arbitrarily small. Then we can find a sequence of independent BMs $\{B_i^*(\cdot)\}_{i=1}^N$ that are independent of $B_\sigma(\cdot)$ and a sequence of random variables $\{d_i\}_{i=1}^N$ defined on an augmented probability space $(\Omega, \mathcal{F}, P)$, such that as $N \to \infty$,*

(a) $\sup_{t\in[0,1]}\left|B_\sigma(t) - \sum_{i=1}^N d_i B_i^*(t)\right| < \varepsilon$ *a.s. P;*

(b) $\int_0^1 \left[B_\sigma(t) - \sum_{i=1}^N d_i B_i^*(t)\right]^2 dt < \varepsilon$ *a.s. P;*

(c) $B_\sigma(t) \overset{d}{=} \sum_{i=1}^\infty d_i B_i^*(t)$ *in* $L^2[a,b]$ *a.s. P.*

Part (c) of Theorem 12.4 shows that an arbitrary BM $B_\sigma(\cdot)$ has an L_2 representation in terms of independent standard BMs with random coefficients. It also gives us a model for the classic spurious regression of independent random walks. In this model, the role of the regressors and the coefficients becomes reversed. The coefficients d_i are random and they are co-dependent with the dependent variable $B_\sigma(t)$. The variables $B_i^*(t)$ are functions that take the form of BM sample paths, and these paths are independent of the dependent variable, just like the fixed coefficients in a conventional linear regression model. Thus, instead of a spurious relationship, we have a model that serves as a representation of one BM in terms of a collection of other BMs. The coefficients in this model provide the connective tissue that relates these random functions.

12.4. New Unit Root Asymptotics with Deterministic Trends

Since the mid-1980s it has been well understood that the presence of deterministic functions in a regression affects tests involving stochastic trends even asymptotically. This dependence has an important bearing on the practical implementation of unit root and cointegration tests. For example, the following model involves both an autoregressive component and some auxiliary regressors that include a trend component

$$Y_t = \rho_o Y_{t-1} + b_o' X_t + u_t. \qquad (12.36)$$

Here Y_t and u_t are scalars and X_t is a p-vector of deterministic trends. Suppose that u_t is i.i.d. $(0,\sigma^2)$ and X_t, Y_t satisfy

$$D_n \sum_{s=1}^{\lfloor nt \rfloor} X_s \to_d X(t) \quad \text{and} \quad n^{-\frac{1}{2}} Y_{\lfloor nt \rfloor} \to_d B_\sigma(t) \tag{12.37}$$

for any $t \in [0,1]$ as $n \to \infty$, where D_n is a suitable $p \times p$ diagonal scaling matrix, $X(\cdot)$ is a p-dimensional vector of piecewise continuous functions, and $B_\sigma(\cdot)$ is a BM with variance σ^2. By standard methods the OLS estimate $\widehat{\rho}_n$ of ρ_o in (12.36) has the following limiting distribution:

$$n(\widehat{\rho}_n - \rho_o) \to_d \left[\int_0^1 B_X(t)\, dB_\sigma(t)\right]\left[\int_0^1 B_X^2(t)\, dt\right]^{-1},$$

where

$$B_X(\cdot) \equiv B_\sigma(\cdot) - X'(\cdot)\left[\int_0^1 X(t)X'(t)\, dt\right]^{-1}\left[\int_0^1 X(t)B_\sigma(t)\, dt\right]$$

is the Hilbert space projection residual of $B_\sigma(\cdot)$ on $X(\cdot)$.

Figure 12.1 (from Phillips (2001)) depicts the asymptotic density of $n(\widehat{\rho}_n - \rho_o)$ with different numbers of deterministic (polynomial) trend functions. It is clear that the shape and location of the asymptotic density of $n(\widehat{\rho}_n - \rho_o)$ are both highly sensitive to the trend degree p. This sensitivity implies that critical values of the tests change

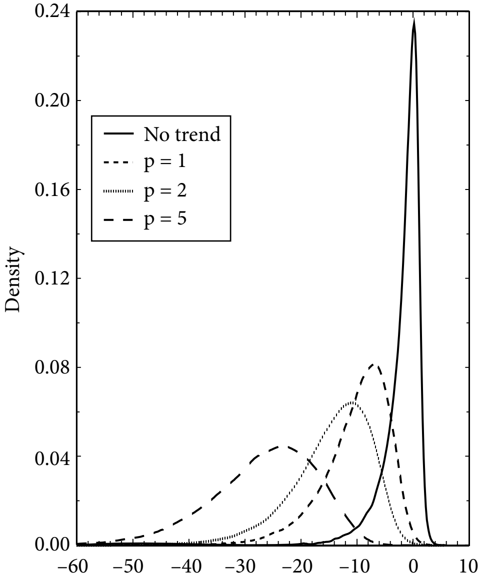


FIGURE 12.1 Densities of $\int_0^1 B_X(t)\, dB_\sigma(t) \Big/ \int_0^1 B_X^2(t)\, dt$ for $X = (1, t, \ldots, t^p)$.

substantially with the specification of the deterministic trend functions, necessitating the use of different statistical tables according to the precise specification of the fitted model. As a result, if the approach to modeling the time series were such that one contemplated increasing p as the sample size n increased, and to continue to do so as n goes to infinity, then a limit theory in which $p \to \infty$ as $n \to \infty$ may be more appropriate. In fact, even the moderate degree $p \sim 5$ produces very different results from $p = 0, 1$, and the large p asymptotic theory in this case produces a better approximation to the finite sample distribution. Entirely similar considerations apply when the regressor X_t includes trend breaks.

As we have seen in the previous section, the KL representation (12.10) of a stochastic process links the random function $B_\sigma(t)$ $(t \in [a, b])$ with the trend basis functions $\varphi_k(t)$ $(k \in \mathbb{Z}_+)$ of the Hilbert space $L^2[a, b]$, thereby enabling us to study the effects of deterministic functions on tests involving the stochastic trends. The present section reviews some of the findings in Phillips (2001), which shows how the asymptotic theory of estimation in unit root models changes when deterministic trends coexist with the stochastic trend.

Specifically, consider the following typical autoregression with a trend component:

$$\frac{1}{\sqrt{n}} Y_t = \frac{\widehat{\rho}_n}{\sqrt{n}} Y_{t-1} + \sum_{k=1}^{K} \widehat{a}_{k,n} \varphi_k\left(\frac{t}{n}\right) + \widehat{u}_{t,K}, \tag{12.38}$$

where $\varphi_k(\cdot)$ $(k \in \mathbb{Z}_+)$ are trend basis functions, and $\widehat{\rho}_n$ and $\widehat{a}_{k,n}$ are the OLS estimates by regressing $n^{-\frac{1}{2}} Y_t$ on the lagged variables $n^{-\frac{1}{2}} Y_{t-1}$ and $\varphi_k\left(\frac{t}{n}\right)$ $(k = 1, \ldots, K)$. The scaling in (12.38) is entirely innocuous and used only to assist in the asymptotics. As is apparent from regression (12.28) and Theorem 12.2, when there is no lagged dependent variable $n^{-\frac{1}{2}} Y_{t-1}$ in (12.38), the fitted value from the trend basis $\sum_{k=1}^{K} \widehat{a}_{k,n} \varphi_k(t)$ reproduces the KL component $\sum_{k=1}^{K} \lambda_k^{\frac{1}{2}} \xi_k \varphi_k(t)$ of the BM limit process of $n^{-\frac{1}{2}} Y_t$ as the sample size $n \to \infty$.

In particular, as the scaled partial sum $n^{-\frac{1}{2}} Y_t$ satisfies the functional central limit theorem (FCLT) in (12.37), we can invoke (12.26) to deduce that

$$\sup_{0 \le t \le n} \left| \frac{1}{\sqrt{n}} Y_t - \sum_{k=1}^{\infty} \lambda_k^{\frac{1}{2}} \varphi_k\left(\frac{t}{n}\right) \xi_k \right| = o_p(1). \tag{12.39}$$

From the partitioned regression in (12.38) and the series representation in (12.39), we see that $\widehat{\rho}_n$ is the fitted coefficient in the regression of $n^{-\frac{1}{2}} Y_t$ on the projection residual of $n^{-\frac{1}{2}} Y_{t-1}$ on the trend basis functions $\varphi_k(\cdot)$ $(k = 1, \ldots, K)$. The stochastic trend variable Y_{t-1} and the trend basis functions are highly correlated with large K, and there is a collinearity problem in the regression (12.38) as $K \to \infty$ because the lagged regressor is perfectly fitted by the trend basis. The asymptotic properties of $\widehat{\rho}_n$ are correspondingly affected by the presence of the deterministic trends and their influence is severe when $K \to \infty$. As a result, unit root tests and limit theory based on $\widehat{\rho}_n$ are affected by the presence of deterministic trends, the effects being sufficiently important as to alter the convergence rate. This point is confirmed in the next theorem. First, we have the following lemma (Phillips, 2001) which shows the effect of a finite number K of deterministic trends on the limit theory of semiparametric Z tests (Phillips, 1987; Phillips and Perron, 1988; and Ouliaris, Park and Phillips, 1988). These tests are either coefficient based (denoted here by $Z_{\rho,n}$) or t-ratio tests (denoted by $Z_{t,n}$). Readers may refer to the above references for their construction.

Lemma 12.2. *Suppose that u_t satisfies Assumption 12.3 and $Y_t = \sum_{s=1}^{t} u_s$. Then the unit root test statistic $Z_{\rho,n}$ and the t-ratio test statistic $Z_{t,n}$ satisfy*

$$Z_{\rho,n} \to_d \frac{\int_0^1 B_{\varphi_K}(r)\,dB_\sigma(r)}{\int_0^1 B_{\varphi_K}^2(r)dr} \quad \textit{and} \quad Z_{t,n} \to_d \frac{\int_0^1 B_{\varphi_K}(r)\,dB_\sigma(r)}{\left[\int_0^1 B_{\varphi_K}^2(r)dr\right]^{\frac{1}{2}}},$$

where $B_{\varphi_K}(\cdot) = B_\sigma(\cdot) - \left[\int_0^1 B_\sigma(r)\Phi_K(r)dr\right]\Phi_K'(\cdot)$.

From the KL representation, we see that

$$\begin{aligned}\int_0^1 B_{\varphi_K}^2(r)dr &= \int_0^1 \left[\sum\nolimits_{k=K+1}^{\infty} \lambda_k^{\frac{1}{2}} \varphi_k(r)\xi_k\right]^2 dr \\ &= \sum\nolimits_{k=K+1}^{\infty} \lambda_k \xi_k^2 \to_{a.s.} 0 \qquad \text{as } K \to \infty,\end{aligned}$$

which implies that when K is large, the asymptotic distributions of $Z_{\rho,n}$ and $Z_{t,n}$ are materially affected by a denominator that tends to zero and integrand in the numerator that tends to zero. This structure explains why the asymptotic distributions of $Z_{\rho,n}$ and $Z_{t,n}$ are drawn toward minus infinity with larger K. One may conjecture that when $K \to \infty$, $Z_{\rho,n}$ and $Z_{t,n}$ will diverge to infinity as $\int_0^1 B_{\varphi_K}^2(r)dr \to_p 0$ as $K \to \infty$. This conjecture is confirmed in the following theorem from Phillips (2001).

Theorem 12.5. *Suppose that u_t satisfies Assumption 12.3. If $K \to \infty$ and $K^4/n \to 0$ as $n \to \infty$, then*

$$K^{-\frac{1}{2}}\left(Z_{\rho,n} + \frac{\pi^2 K}{2}\right) \to_d N\left(0, \pi^4/6\right)$$

and

$$Z_{t,n} + \frac{\pi\sqrt{K}}{2} \to_d N\left(0, \pi^2/24\right).$$

When the lagged dependent variable and deterministic trend functions are included in the LS regression to model a stochastic trend, they are seen to jointly compete for the explanation of the stochastic trend in a time series. In such a competition, Theorem 12.5 implies that the deterministic functions will be successful in modeling the trend even in the presence of an autoregressive component. The net effect of including K deterministic functions in the regression is that the rate of convergence to unity of the autoregressive coefficient $\widehat{\rho}_n$ is slowed down. In particular, the theorem implies that

$\widehat{\rho}_n = 1 - \frac{\pi^2}{2}\frac{K}{n} + o_p\left(\frac{K}{n}\right) \to_p 1$ as $(n, K \to \infty)$. Thus, $\widehat{\rho}_n$ is still consistent for $\rho = 1$, but has a slower rate of approach to unity than when K is fixed. The explanation for the nonstationarity in the data is then shared between the deterministic trend regressors and the lagged dependent variable.

12.5. Efficient Estimation of Cointegrated Systems

The trend basis functions in the KL representation (12.10) are deterministic and accordingly independent of any random variables. Moreover, as shown in Theorem 12.3, a stochastic trend can be fully reproduced by its projection on the trend basis functions. These two properties indicate that trend basis functions provide a natural set of valid instrumental variables (IVs) to model stochastic processes that appear as endogenous regressors. This feature of the KL basis functions was pointed out in Phillips (2013), who proposed using trend basis functions as IVs to efficiently estimate cointegrated systems. We outline the essential features of this work in what follows.

Consider the cointegrated system

$$Y_t = A_o X_t + u_{y,t}, \tag{12.40}$$

$$\Delta X_t = u_{x,t}, \tag{12.41}$$

where the time series Y_t is $m_y \times 1$ and X_t is $m_x \times 1$ with initial conditions $X_0 = O_p(1)$ at $t = 0$. The composite error $u_t = (u'_{y,t}, u'_{x,t})'$ is a weakly dependent time series generated as a linear process

$$u_t = C(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \quad \sum_{j=0}^{\infty} j^a \left\| c_j \right\| < \infty, \qquad a > 3, \tag{12.42}$$

where $\varepsilon_t = \text{i.i.d.}(0, \Sigma)$ with $\Sigma > 0$ and $E\left[||\varepsilon_t||^p\right] < \infty$ for some $p > 2$ and matrix norm $\|\cdot\|$. The long-run moving average coefficient matrix $C(1)$ is assumed to be nonsingular, so that X_t is a full-rank integrated process. Under (12.42), the scaled partial sum $\frac{1}{\sqrt{n}} \sum_{s=0}^{t} u_t$ satisfies the following FCLT:

$$\frac{1}{\sqrt{n}} \sum_{s=0}^{\lfloor nt \rfloor} u_t \to_d B_u(t) \equiv \begin{pmatrix} B_y(t) \\ B_x(t) \end{pmatrix}, \tag{12.43}$$

for any $t \in [0, 1]$. The long-run variance matrix $\Omega = C(1)\Sigma C'(1)$ is partitioned conformably with u_t as

$$\Omega = \begin{bmatrix} \Omega_{yy} & \Omega_{yx} \\ \Omega_{xy} & \Omega_{xx} \end{bmatrix}.$$

The conditional long-run covariance matrix of u_y on u_x is $\Omega_{yy\cdot x} = \Omega_{yy} - \Omega_{yx}\Omega_{xx}^{-1}\Omega_{xy}$. In a similar way we define the one-sided long-run covariance matrix

$$\Delta = \sum_{j=0}^{\infty} E\left(u_0 u_{-j}'\right) = \begin{bmatrix} \Delta_{yy} & \Delta_{yx} \\ \Delta_{xy} & \Delta_{xx} \end{bmatrix}.$$

The rest of this section discusses and compares several different estimates of A_o. The comparison of different estimates helps in understanding the role that trend basis functions play in efficient estimation. For ease of notation and without loss of generality we henceforth assume that X_t and Y_t are scalar random variables. We first consider the OLS estimate of A_o, which is defined as $\widehat{A}_n = \left(\sum_{t=1}^n Y_t X_t'\right)\left(\sum_{t=1}^n X_t X_t'\right)^{-1}$. Under (12.42) it is easily seen that

$$n(\widehat{A}_n - A_o) = \frac{n^{-1}\sum_{t=1}^n u_{y,t}X_t}{n^{-2}\sum_{t=1}^n X_t^2} \to_d \frac{\int_0^1 B_x(t)\,dB_y(t) + \Delta_{yx}}{\int_0^1 B_x^2(t)\,dt}$$

where B_x and B_y are defined in (12.43). In view of the contemporaneous and serial correlation between $u_{x,t}$ and $u_{y,t}$, it is well known that OLS estimation suffers from two sources of high-order bias: endogeneity bias from the corresponding correlation of B_x and B_y and serial correlation bias that manifests in the one-sided long-run covariance Δ_{yx}.

We next consider the IV estimation of the augmented regression equation with K trend IVs (basis functions) $\varphi_k(\cdot)$ $(k = 1, \ldots, K)$:

$$Y_t = A_o X_t + B_o \Delta X_t + u_{y\cdot x,t}, \tag{12.44}$$

where $B_o = \Omega_{yx}\Omega_{xx}^{-1}$ and $u_{y\cdot x,t} = u_{y,t} - B_o u_{x,t}$. For this model, it is easy to show that the LS estimate of A_o continues to suffer from second-order bias effects and the LS estimate of B_o is not generally consistent. On the other hand, the IV estimate of A_o in the augmented equation has optimal properties. It can be written in projection form as

$$\widehat{A}_{IV} = \left(Y' R_{\Delta X,K} X\right)\left(X' R_{\Delta X,K} X\right)^{-1},$$

where $Y' = [Y_1, \ldots, Y_n]$ with similar definitions for the observation matrices X' and ΔX, the projector $P_K = \Phi_K\left(\Phi_K'\Phi_K\right)^{-1}\Phi_K'$, $\Phi_K = [\Phi_K'(\frac{1}{n}), \ldots, \Phi_K'(1)]'$, $\Phi_K(\cdot) = [\varphi_1(\cdot), \ldots, \varphi_K(\cdot)]$ and the composite projector $R_{\Delta X,K} = P_K - P_K \Delta X\left(\Delta X' P_K \Delta X\right)^{-1}$ $\Delta X' P_K$. Similarly, the IV estimate of B_o can be written as

$$\widehat{B}_{IV} = \left(Y' R_{X,K} \Delta X\right)\left(\Delta X' R_{X,K} \Delta X\right)^{-1},$$

where $R_{X,K} = P_K - P_K X\left(X' P_K X\right)^{-1} X' P_K$.[5]

The following lemma gives the asymptotic distributions of the IV estimates $\widehat{A}_{IV_K,n}$ and $\widehat{B}_{IV_K,n}$ when the number of the trend basis functions K is fixed.

Lemma 12.3. *Under the assumption (12.42), we have*

$$n(\widehat{A}_{IV} - A_o) \to_d \frac{\sum_{k=1}^K \eta_{x,k}^2 \sum_{k=1}^K \xi_{x,k}\eta_{y\cdot x,k} - \sum_{k=1}^K \eta_{x,k}\eta_{y\cdot x,k} \sum_{k=1}^K \xi_{x,k}\eta_{x,k}}{\sum_{k=1}^K \eta_{x,k}^2 \sum_{k=1}^K \xi_{x,k}^2 - \left[\sum_{k=1}^K \xi_{x,k}\eta_{x,k}\right]^2} \tag{12.45}$$

and

$$\widehat{B}_{IV} \to_d B_o + \frac{\sum_{k=1}^K \xi_{x,k}^2 \sum_{k=1}^K \eta_{x,k}\eta_{y\cdot x,k} - \sum_{k=1}^K \xi_{x,k}\eta_{y\cdot x,k} \sum_{k=1}^K \xi_{x,k}\eta_{x,k}}{\sum_{k=1}^K \xi_{x,k}^2 \sum_{k=1}^K \eta_{x,k}^2 - \left[\sum_{k=1}^K \xi_{x,k}\eta_{x,k}\right]^2}, \tag{12.46}$$

where $\eta_{y\cdot x,k} = \int_0^1 \varphi_k(r)dB_{y\cdot x}(r)$, *and* $\xi_{x,k}$, $\eta_{x,k}$, $\eta_{y,k}$ *are defined by*

$$\xi_{x,k} = \int_0^1 \varphi_k(t)B_x(t)\,dt, \quad \eta_{x,k} = \int_0^1 \varphi_k(t)dB_x(t), \quad \textit{and} \quad \eta_{y,k} = \int_0^1 \varphi_k(t)dB_y(t), \tag{12.47}$$

for all k.

From Lemma 12.3, we see that the IV estimate $\widehat{A}_{IV}$ of A_o in the augmented equation (12.40) is consistent, but it suffers second-order bias when the number of the trend basis functions K is fixed. Moreover, the IV estimate $\widehat{B}_{IV}$ of B_o, is not consistent when K is fixed. By Corollary 12.1, we get

$$\xi_{x,k}^2 = \left[\int_0^1 \varphi_k(r)\,dB_x(r)\right]^2 \overset{d}{=} \Omega_{xx}\chi_k^2(1) \qquad \text{for all } k \in \mathbb{Z}_+,$$

where Ω_{xx} is the long-run variance of $u_{x,t}$ and $\chi_k^2(1)$ denotes a chi-square random variable with degree of freedom 1. Moreover, $\chi_k^2(1)$ is independent of $\chi_{k'}^2(1)$ for any $k \neq k'$ and $k, k' \in \mathbb{Z}_+$. Using the law of large numbers, we have

$$\frac{1}{K}\sum_{k=1}^K \left[\int_0^1 \varphi_k(r)\,dB_x(r)\right]^2 \to_{a.s.} \Omega_{xx}. \tag{12.48}$$

Under sequential asymptotics, we see that

$$n(\widehat{A}_{IV} - A_o) = \frac{\sum_{k=1}^K \xi_{x,k}\eta_{y\cdot x,k} + O_p(K^{-1})}{\sum_{k=1}^K \xi_{x,k}^2 + O_p(K^{-1})} \tag{12.49}$$

and

$$\widehat{B}_{IV} = B_o + O_p(K^{-1}). \tag{12.50}$$

Results in (12.49) and (12.50) indicate that when the number of trend IVs diverges to infinity, the IV estimate $\widehat{A}_{IV}$ of A_o may be as efficient as the maximum likelihood (ML) estimate under Gaussianity (Phillips (1991a)) and the IV estimate $\widehat{B}_{IV}$ of B_o may be consistent. These conjectures are justified in Phillips (2012) and shown to hold under joint asymptotics.

Let $\widehat{\Omega}_{K,n} = K^{-1}\big(Y' - \widehat{A}_{IV}X' - \widehat{B}_{IV}\Delta X'\big)P_K\big(Y' - \widehat{A}_{IV}X' - \widehat{B}_{IV}\Delta X'\big)'$ and define $B_{y\cdot x}(t) = B_y(t) - B_o B_x(t)$. The following theorem is from Phillips (2013).

Theorem 12.6. *Under the assumption (12.42) and the rate condition*

$$\frac{1}{K} + \frac{K}{n^{(1-2/p)\wedge(5/6-1/3p)}} + \frac{K^5}{n^4} \to 0 \tag{12.51}$$

as $n \to \infty$, we have

(a) $n(\widehat{A}_{IV} - A_o) \to_d \left[\int_0^1 B_x(t)\, dB'_{y\cdot x}(t)\right]'\left[\int_0^1 B_x(t)B'_x(t)\, dr\right]^{-1}$,
(b) $\widehat{B}_{IV} \to_p B_o$
(c) $\widehat{\Omega}_{K,n} \to_p \Omega_{yy} - \Omega_{yx}\Omega_{xx}^{-1}\Omega_{xy}$.

Theorem 12.6 implies that the IV estimate $\widehat{A}_{IV}$ is consistent and as efficient as the ML estimate under Gaussian errors (see Phillips, 1991a, for the latter). Moreover, the IV estimates of the long-run coefficients are also consistent. It is easy to see that

$$E[\varphi_k(t)X_t] = \varphi_k(t)E[X_t] = 0$$

for any $k \in \mathbb{Z}_+$, which implies that trend IVs do not satisfy the relevance condition in the IV estimation literature. As a result, the fact that efficient estimation using trend IVs is possible may appear somewhat magical, especially in view of existing results on IV estimation in stationary systems where relevance of the instruments is critical to asymptotic efficiency and can even jeopardize consistency when the instruments are weak (Phillips, 1989; Staiger and Stock, 1997). Furthermore, the results in Theorem 12.5 make it clear that what is often regarded as potentially dangerous spurious correlation among trending variables can itself be used in a systematic way to produce rather startling positive results.

12.6. Automated Efficient Estimation of Cointegrated Systems

As illustrated in the previous section, the trend IV approach is very effective in efficient estimation of the cointegration systems. In reality, when the cointegration systems have the triangle representation (12.40) and (12.41), this method is very straightforward and easy to be implemented. However, when the cointegration rank of the cointegrated system is unknown, it is not clear how the trend IV approach can be applied to achieve optimal estimation. Determination of the cointegration rank is important for estimation and inference of cointegrated systems, because underselected cointegration rank produces inconsistent estimation, while overselected cointegration rank leads to second order bias and inefficient estimation (cf., Liao and Phillips (2010)). More recently, Liao and Phillips (2012) proposed an automated efficient estimation method for the cointegrated systems. The new method not only consistently selects the cointegration rank and the lagged differences in general vector error correction models (VECMs) in one step, but also performs efficient estimation of the cointegration matrix and nonzero transient dynamics simultaneously.

Liao and Phillips (2012) first studied the following simple VECM system:

$$\Delta Y_t = \Pi_o Y_{t-1} + u_t = \alpha_o \beta_o' Y_{t-1} + u_t, \tag{12.52}$$

where $\Pi_o = \alpha_o \beta_o'$ has rank, $0 \leq r_o \leq m$, α_o and β_o are $m \times r_o$ matrices with full rank, and $\{u_t\}$ is an m-dimensional i.i.d. process with zero mean and nonsingular covariance matrix Ω_u. The following assumption is imposed on Π_o.

Assumption 12.4 (RR). *(i) The determinantal equation $|I - (I + \Pi_o)\lambda| = 0$ has roots on or outside the unit circle; (ii) the matrix Π_o has rank r_o, with $0 \leq r_o \leq m$; (iii) if $r_o > 0$, then the matrix $R = I_{r_o} + \beta_o' \alpha_o$ has eigenvalues within the unit circle.*

The unknown parameter matrix Π_o is estimated in the following penalized GLS estimation

$$\widehat{\Pi}_{g,n} = \underset{\Pi \in R^{m \times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \| \Delta Y_t - \Pi Y_{t-1} \|^2_{\widehat{\Omega}^{-1}_{u,n}} + \sum_{k=1}^{m} \frac{n \lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \left\| \Phi_{n,k}(\Pi) \right\| \right\}, \tag{12.53}$$

where $\|A\|^2_B = A'BA$ for any $m \times 1$ vector A and $m \times m$ matrix B, $\widehat{\Omega}_{u,n}$ is some first-step consistent estimator of Ω_u, $\omega > 0$ is some constant, $\lambda_{r,k,n}$ $(k = 1, \ldots, m)$ are tuning parameters that directly control the penalization, $||\phi_k(\Pi)||$ denotes the kth largest modulus of the eigenvalues $\{\phi_k(\Pi)\}_{k=1}^m$ of the matrix Π,[6] $\Phi_{n,k}(\Pi)$ is the kth row vector of $Q_n \Pi$, and Q_n denotes the normalized left eigenvector matrix of $\widehat{\Pi}_{1st}$. The matrix $\widehat{\Pi}_{1st}$ is a first-step (OLS) estimate of Π_o. The penalty functions in (12.53) are constructed based on the so-called adaptive Lasso penalty (Zou, 2006) and they play the role of selecting the cointegrating rank in the penalized estimation. More importantly, if the cointegration rank is simultaneously determined in the estimation of Π_o, the selected rank structure will be automatically imposed on the penalized GLS estimate $\widehat{\Pi}_{g,n}$. As a result, $\widehat{\Pi}_{g,n}$ would be automatically efficient if the true cointegration rank could be consistently selected in the penalized GLS estimation (12.53).

The asymptotic properties of the penalized GLS estimate are given in the following theorem from Liao and Phillips (2012).

Theorem 12.7 Oracle Properties. *Suppose Assumption 12.4 holds. If $\widehat{\Omega}_{u,n} \to_p \Omega_u$ and the tuning parameter satisfies $n^{\frac{1}{2}} \lambda_{r,k,n} = o(1)$ and $n^{\omega} \lambda_{r,k,n} \to \infty$ for $k = 1, \ldots, m$, then as $n \to \infty$, we have*

$$\Pr\big(rank(\widehat{\Pi}_{g,n}) = r_o\big) \to 1, \tag{12.54}$$

where rank($\widehat{\Pi}_{g,n}$) denotes the rank of $\widehat{\Pi}_{g,n}$. Moreover $\widehat{\Pi}_{g,n}$ has the same limit distribution as the reduced rank regression (RRR) estimator, which assumes that the true rank r_o is known.

Theorem 12.6 shows that if the tuning parameters $\lambda_{r,k,n}$ $(k = 1,\ldots,m)$ converge to zero at certain rate, then the consistent cointegration selection and the efficient estimation can be simultaneously achieved in the penalized GLS estimation (12.53). Specifically, the tuning parameter $\lambda_{r,k,n}$ $(k = 1,\ldots,m)$ should converge to zero faster than $\sqrt{n}$ so that when $\Pi_o \neq 0$, the convergence rate of $\widehat{\Pi}_{g,n}$ is not slower than root-n. On the other hand, $\lambda_{r,k,n}$ should converge to zero slower than $n^{-\omega}$ so that the cointegration rank r_o is selected with probability approaching one.

The i.i.d. assumption on u_t ensures that Π_o is consistently estimated, which is usually required for consistent model selection in the Lasso model selection literature. But Cheng and Phillips (2009, 2012) showed that the cointegration rank r_o can be consistently selected by information criteria even when u_t is weakly dependent, in particular when u_t satisfies conditions such as LP below. We therefore anticipate that similar properties hold for Lasso estimation.

Assumption 12.5 (LP). *Let $D(L) = \sum_{j=0}^{\infty} D_j L^j$, where $D_0 = I_m$ and $D(1)$ has full rank. Let u_t have the Wold representation*

$$u_t = D(L)\varepsilon_t = \sum_{j=0}^{\infty} D_j \varepsilon_{t-j}, \qquad \textit{with} \sum_{j=0}^{\infty} j^{\frac{1}{2}} ||D_j|| < \infty, \tag{12.55}$$

where ε_t is i.i.d. $(0, \Sigma_{\varepsilon\varepsilon})$ *with $\Sigma_{\varepsilon\varepsilon}$ positive definite and finite fourth moments.*

It is clear that under Assumption 12.5, Π_o cannot be consistently estimated in general. As a result, the probability limit of the GLS estimate of Π_o may have rank smaller or larger than r_o. However, Liao and Phillips (2012) show that the cointegration rank r_o can be consistently selected by penalized estimation as in (12.53) even when u_t is weakly dependent and Π_o is not consistently estimated, thereby extending the consistent rank selection result of Cheng and Phillips (2009) to Lasso estimation.

Theorem 12.8. *Under Assumption LP, if $n^{\frac{1+\omega}{2}} \lambda_{r,k,n} = o(1)$ and $n^{\frac{1}{2}} \lambda_{r,k,n} = o(1)$ for $k = 1,\ldots,m$, then we have*

$$\Pr\big(rank(\widehat{\Pi}_{g,n}) = r_o\big) \to 1 \qquad \textit{as } n \to \infty. \tag{12.56}$$

Theorem 12.8 states that the true cointegration rank r_o can be consistently selected, even though the matrix Π_o is not consistently estimated. Moreover, even when the probability limit Π_1 of the penalized GLS estimator has rank less than r_o, Theorem 12.8 ensures that the correct rank r_o is selected in the penalized estimation. This result is new in the Lasso model selection literature as Lasso techniques are usually advocated because of their ability to shrink small estimates (in magnitude) to zero in penalized estimation. However, Theorem 12.8 shows that penalized estimation here does not shrink the estimates of the extra $r_o - r_1$ zero eigenvalues of Π_1 to zero.

Liao and Phillips (2012) also study the general VECM model

$$\Delta Y_t = \Pi_o Y_{t-1} + \sum_{j=1}^{p} B_{o,j} \Delta Y_{t-j} + u_t \tag{12.57}$$

with simultaneous cointegration rank selection and lag-order selection. To achieve consistent lag-order selection, the model in (12.57) has to be consistently estimable. Thus, we assume that given p in (12.57), the error term u_t is an m-dimensional i.i.d. process with zero mean and nonsingular covariance matrix Ω_u. Define

$$C(\phi) = \Pi_o + \sum_{j=0}^{p} B_{o,j}(1-\phi)\phi^j, \qquad \text{where } B_{o,0} = -I_m.$$

The following assumption extends Assumption 12.4 to accommodate the general structure in (12.57).

Assumption 12.6 (RR). *(i) The determinantal equation $|C(\phi)| = 0$ has roots on or outside the unit circle; (ii) the matrix Π_o has rank r_o, with $0 \leq r_o \leq m$; (iii) the $(m - r_o) \times (m - r_o)$ matrix*

$$\alpha'_{o,\perp}\left(I_m - \sum_{j=1}^{p} B_{o,j}\right)\beta_{o,\perp} \tag{12.58}$$

is nonsingular, where $\alpha_{o,\perp}$ and $\beta_{o,\perp}$ are the orthonormal complements of α_o and β_o respectively.

Let $B_o = [B_{o,1}, \ldots, B_{o,p}]$. The unknown parameters (Π_o, B_o) are estimated by penalized GLS estimation

$$\begin{aligned}(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n}) = \underset{\Pi, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} & \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2_{\widehat{\Omega}^{-1}_{u,n}} \right. \\ & \left. + \sum_{j=1}^{p} \frac{n\lambda_{b,j,n}}{\|\widehat{B}_{j,1st}\|^{\omega}} \left\| B_j \right\| + \sum_{k=1}^{m} \frac{n\lambda_{r,k,n}}{\|\phi_k(\widehat{\Pi}_{1st})\|^{\omega}} \left\| \Phi_{n,k}(\Pi) \right\| \right\},\end{aligned} \tag{12.59}$$

where $\lambda_{b,j,n}$ and $\lambda_{r,k,n}$ ($j = 1, \ldots, p$ and $k = 1, \ldots, m$) are tuning parameters, and $\widehat{B}_{j,1st}$ and $\widehat{\Pi}_{1st}$ are some first-step (OLS) estimates of $B_{o,j}$ and Π_o ($j = 1, \ldots, p$) respectively. Denote the index set of the zero components in B_o as $\mathcal{S}^c_B$ such that $\left\| B_{o,j} \right\| = 0$ for all $j \in \mathcal{S}^c_B$ and $\left\| B_{o,j} \right\| \neq 0$ otherwise. The asymptotic properties of the penalized GLS estimates $(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n})$ are presented in the following theorem from Liao and Phillips (2012).

Theorem 12.9. *Suppose that Assumption 12.6 is satisfied and $\widehat{\Omega}_{u,n} \to_p \Omega_u$. If $n^{\frac{1}{2}}(\lambda_{r,k,n} + \lambda_{b,j,n}) = O(1)$, $n^{\omega}\lambda_{r,k,n} \to \infty$ and $n^{\frac{1+\omega}{2}}\lambda_{b,j,n} \to \infty$ $(k = 1, \ldots, m$ and $j = 1, \ldots, p)$, then*

$$\Pr\left(r(\widehat{\Pi}_{g,n}) = r_o\right) \to 1 \quad \text{and} \quad \Pr\left(\widehat{B}_{g,j,n} = 0\right) \to 1 \tag{12.60}$$

for $j \in \mathcal{S}_B^c$ as $n \to \infty$; moreover, $\widehat{\Pi}_{g,n}$ and the penalized GLS estimate of the nonzero components in B_o have the same joint limiting distribution as that of the general RRR estimate, which assumes that the true rank r_o and true zero components in B_o are known.

From Theorem 12.7 and Theorem 12.9, we see that the tuning parameter plays an important role in ensuring that the penalized estimate is efficient and the true model is consistently selected in penalized GLS estimation. In empirical applications, the conditions stated in these two theorems do not provide a clear suggestion of how to select the tuning parameters. In the Lasso literature the tuning parameters are usually selected by cross-validation or information criteria methods. However, such methods of selecting the tuning parameter are computationally intensive and they do not take the finite sample properties of the penalized estimates into account. Liao and Phillips (2012) provide a simple data-driven tuning parameter selection procedure based on balancing first-order conditions that takes both model selection and finite sample properties of the penalized estimates into account. The new method is applied to model GNP, consumption and investment using U.S. data, where there is obvious co-movement in the series. The results reveal the effect of this co-movement through the presence of two cointegrating vectors, whereas traditional information criteria fail to find co-movement and set the cointegrating rank to zero for these data.

12.7. Series Estimation of the Long-Run Variance

Previous sections have shown how the long-run behavior of integrated processes can be fully reproduced in the limit by simple linear projections on trend basis functions. Motivated by this result, we are concerned to ask the following questions. First, let $\{u_t\}$ be a stationary process and let $\{\varphi_k(\cdot)\}_k$ be a set of trend basis functions. What are the asymptotic properties of the projection of $\{u_t\}_{t=1}^n$ on $\varphi_k(\cdot)$ with a fixed number K of basis functions? Further, what are the asymptotic properties of this projection when the number of basis functions goes to infinity?

As first observed in Phillips (2005b), such projections produce consistent estimates of the long-run variance (LRV) of the process $\{u_t\}$, when K goes to infinity with the sample size. This large K asymptotic theory justifies the Gaussian approximation of t-ratio statistics and chi-square approximations of Wald statistics in finite samples. More recently, Sun (2011, 2013) showed that when K is fixed, t-ratio statistics have an asymptotic Student-t distribution and Wald statistics have asymptotic F distributions.

The fixed-K asymptotic theory is argued in Sun (2013) to provide more accurate size properties for both t-ratio and Wald statistics in finite samples.

Formally, suppose that the process $\{u_t\}$ satisfies the following assumption.

Assumption 12.7. *For all $t \geq 0$, u_t has Wold representation*

$$u_t = C(L)\varepsilon_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j}, \qquad \sum_{j=0}^{\infty} j^a \left| c_j \right| < \infty, \qquad C(1) \neq 0 \quad \text{and} \quad a > 3 \tag{12.61}$$

with $\varepsilon_t =$ i.i.d. $(0,\sigma_\varepsilon^2)$ with $E\big(|\varepsilon_t|^p\big) < \infty$ for some $p > 2$.

Under Assumption 12.7, the scaled partial sum $n^{-\frac{1}{2}} \sum_{i=1}^{t} u_i$ satisfies the following FCLT

$$B_n(\cdot) \equiv \frac{\sum_{i=1}^{[n\cdot]} u_i}{\sqrt{n}} \to_d B_\omega(\cdot) \text{ as } n \to \infty, \tag{12.62}$$

where $B_\omega(\cdot)$ is a BM with variance $\omega^2 = \sigma_\varepsilon^2 C^2(1)$. Note that ω^2 is the LRV of the process $\{u_t\}$.

The projection of $\{u_t\}_{t=1}^{n}$ on $\varphi_k(\frac{t}{n})$ for some $k \in \mathbb{Z}_+$ can be written as

$$\left[\sum_{t=1}^{n} \varphi_k^2(\frac{t}{n}) \right]^{-1} \sum_{t=1}^{n} \varphi_k(\frac{t}{n}) u_t,$$

where

$$\sum_{t=1}^{n} \varphi_k(\frac{t}{n}) u_t \to_d \int_0^1 \varphi_k(r)\, dB_\omega(r) \qquad \text{as } n \to \infty \tag{12.63}$$

by standard functional limit arguments and Wiener integration, and

$$\frac{1}{n} \sum_{t=1}^{n} \varphi_k^2(\frac{t}{n}) \to \int_0^1 \varphi_k^2(r)\, dr = 1 \qquad \text{as } n \to \infty \tag{12.64}$$

by the integrability and normalization of $\varphi_k(\cdot)$. From the results in (12.63) and (12.64), we deduce that

$$\sqrt{n} \frac{\sum_{t=1}^{n} \varphi_k(\frac{t}{n}) u_t}{\sum_{t=1}^{n} \varphi_k^2(\frac{t}{n})} \to_d \int_0^1 \varphi_k(r)\, dB_\omega(r) \qquad \text{as } n \to \infty.$$

By Corollary 12.1, $\int_0^1 \varphi_k(r)\, dB_\omega(r) \overset{d}{=} N(0,\omega^2)$ and for any $k \neq k'$, the two random variables $\int_0^1 \varphi_k(r)\, dB_\omega(r)$ and $\int_0^1 \varphi_{k'}(r)\, dB_\omega(r)$ are independent with each other. These results motivate us to define the following orthonormal series estimate of the LRV:

$$\omega_{K,n}^2 = \frac{1}{K} \sum_{k=1}^{K} \left[n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_k(\frac{t}{n}) u_t \right]^2, \tag{12.65}$$

which leads to the following t-ratio test statistic

$$t_{K,n} = B_n(1)/\sqrt{\omega^2_{K,n}}. \tag{12.66}$$

Lemma 12.4. *Suppose that Assumption 12.7 is satisfied and the number K of trend basis functions are fixed. Then the series LRV estimate defined in (12.65) satisfies*

$$\omega^2_{K,n} \to_d \frac{\omega^2}{K}\chi^2(K), \tag{12.67}$$

where $\chi^2(K)$ is a chi-square random variable with degrees of freedom K. Moreover, the t-ratio test statistic defined in (12.66) satisfies

$$t_{K,n} \to_d t_K, \tag{12.68}$$

where t_K is a Student-t random variable with degree of freedom K.

While Lemma 12.4 applies to univariate processes, it is readily extended to the case where $\{u_t\}$ is a multiple time series. In that case, the series LRV estimate is defined as

$$\omega^2_{K,n} = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{t=1}^{n}\varphi_k(\frac{t}{n})u_t \sum_{t=1}^{n}\varphi_k(\frac{t}{n})u_t'}{n}$$

and the Wald-type test is defined as

$$W_{K,n} = B_n(1)'\left(\omega^2_{K,n}\right)^{-1}B_n(1).$$

Then using similar arguments to those in the proof of Lemma 12.4, we obtain

$$\frac{K - d_u + 1}{Kd_u}W_{K,n} \to_d \digamma_{d_u,K-d_u+1},$$

where $\digamma_{d_u,K-d_u+1}$ is a F random variable with degrees of freedom $(d_u, K - d_u + 1)$ and d_u denotes the dimensionality of the vector u_t.

The weak convergence in (12.67) implies that when the number of the trend basis functions is fixed, the series LRV estimate $\omega^2_{K,n}$ is not a consistent estimate of ω^2. However, the weak convergence in (12.68) indicates that the t-ratio test statistic is asymptotically pivotal. Using sequential asymptotic arguments, we see from (12.67) that when K goes to infinity, $\chi^2(K)/K$ converges to 1 in probability, which implies that $\omega^2_{K,n}$ may be a consistent estimate of ω^2 with large K. Similarly, from (12.67), we see that $t_{K,n}$ has an asymptotic Gaussian distribution under sequential asymptotics. These sequential asymptotic results provide intuition about the consistency of $\omega^2_{K,n}$ when K goes to infinity, as well as intuition concerning the improved size properties of the fixed K asymptotics in finite samples.

The following theorem from Phillips (2005b), which was proved using trend basis functions of the form (12.15) but which holds more generally, shows that $\omega^2_{K,n}$ is indeed a consistent estimate of ω^2 under the joint asymptotics framework.

Theorem 12.10. *Let $\gamma_u(\cdot)$ denote the autocovariance function of the process $\{u_t\}$. Suppose that Assumption 12.7 holds and the number of trend basis functions K satisfies*

$$\frac{n}{K^2} + \frac{K}{n} \to 0. \tag{12.69}$$

Then

(a) $\lim_{n\to\infty} \frac{n^2}{K^2} E\Big(\omega^2_{K,n} - \omega^2\Big) = -\frac{\pi^2}{6}\sum_{h=-\infty}^{\infty} h^2\gamma_u(h)$;

(b) if $K = o(n^{4/5})$, *then* $\sqrt{K}\Big(\omega^2_{K,n} - \omega^2\Big) \to_d N(0, 2\omega^4)$;

(c) if $K^5/n^4 \to 1$, *then* $\frac{n^4}{K^4} E\Big(\omega^2_{K,n} - \omega^2\Big)^2 = \frac{\pi^4}{36}\big[\sum_{h=-\infty}^{\infty} h^2\gamma_u(h)\big]^2 + 2\omega^4$.

Theorem 12.10.(a) implies that $\omega^2_{K,n}$ has bias of order K^2/n^2 as shown in

$$E\big[\omega^2_{K,n}\big] = \omega^2 - \frac{K^2}{n^2}\left[\frac{\pi^2}{6}\sum_{h=-\infty}^{\infty} h^2\gamma_u(h) + o(1)\right].$$

From (b), the variance of $\omega^2_{K,n}$ is of $O(K^{-1})$. Thus, given the sample size n, increases in the number of the trend basis functions K increases bias and reduces variance. The situation is analogous to bandwidth choice in kernel estimation.

The process $\{u_t\}$ studied above is assumed to be known. For example, u_t could be a function of data Z_t and some known parameter θ_o, i.e. $u_t = f(Z_t, \theta_o)$. However, in applications, usually we have to estimate the LRV of the process $\{f(Z_t, \theta_o)\}_t$, where θ_o is unknown but for which a consistent estimate $\widehat{\theta}_n$ may be available. As an illustration, in the rest of this section we use Z-estimation with weakly dependent data to show how the series LRV estimate can be used to conduct autocorrelation robust inference.

The Z-estimate $\widehat{\theta}_n$ can be defined as

$$n^{-\frac{1}{2}}\sum_{t=1}^{n} m(Z_t, \widehat{\theta}_n) = o_p(\varepsilon_n),$$

where $m(\cdot,\cdot) : R^{d_z} \times R^{d_\theta} \to R^{d_\theta}$ is a measurable function and ε_n is a $o(1)$ sequence. Let $M(\theta) = E[m(Z, \theta)]$. The following assumptions are convenient for the following development and exposition.

Assumption 12.8. *(i) $M(\theta)$ is continuous differentiable in the local neighborhood of θ_o and $\frac{\partial M(\theta_o)}{\partial \theta'}$ has full rank; (ii) the Z-estimate $\widehat{\theta}_n$ is root-n normal, that is,*

$$\sqrt{n}(\widehat{\theta}_n - \theta_o) \to_d N\big(0, M_-(\theta_o)V(\theta_o)M'_-(\theta_o)\big),$$

where $M_-(\theta_o) = \left[\frac{\partial M(\theta_o)}{\partial \theta'}\right]^{-1}$ *and* $V(\theta_o) = \lim_{n\to\infty} Var\left[n^{-\frac{1}{2}} \sum_{t=1}^{n} m(Z_t,\theta_o)\right]$*; (iii) let* N_n *denote some shrinking neighborhood of* θ_o*, then*

$$\sup_{\theta\in\mathcal{N}_n} n^{-\frac{1}{2}} \sum_{t=1}^{n} \phi_k(\frac{t}{n})\{m(Z_t,\theta) - m(Z_t,\theta_0) - E[m(Z_t,\theta) - m(Z_t,\theta_0)]\} = o_p(1);$$

(iv) the following FCLT holds:

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} \phi_k(\frac{t}{n}) m(Z_t,\theta_0) \to_d \int_0^1 \phi_k(r) dB_m(r) \qquad \textit{for } k = 1,\ldots,K,$$

where $B_m(\cdot)$ *denotes a vector BM with variance–covariance matrix* $V(\theta_o)$*; (v) we have*

$$M_{+,n}(\widehat{\theta}_n) \equiv n^{-1} \sum_{t=1}^{n} \frac{\partial m(Z_t,\widehat{\theta}_n)}{\partial \theta'} \to_p M_-^{-1}(\theta_o).$$

The conditions in Assumption 12.8 are mild and easy to verify. The series LRV estimate is defined as

$$V_{K,n}(\widehat{\theta}_n) = \frac{1}{K} \sum_{k=1}^{K} \Lambda_{k,n} \Lambda'_{k,n}, \tag{12.70}$$

where $\Lambda_{k,n} \equiv \sum_{t=1}^{n} \phi_k(\frac{t}{n}) m(Z_t,\widehat{\theta}_n)$ $(k = 1,\ldots,K)$. Under Assumption 12.8, we have the following lemma, which generalizes Lemma 12.4 to vector stochastic processes with unknown parameters.

Lemma 12.5. *Suppose that the number of the trend basis functions K is fixed and the basis functions satisfy* $\int_0^1 \phi_k(r)\,dr = 0$ $(k = 1,\ldots,K)$*. Then under Assumptions 12.7 and 12.8, we have*

$$\begin{aligned} \mathcal{F}_n &\equiv (\widehat{\theta}_n - \theta_o)' M_{+,n}(\widehat{\theta}_n) V_{K,n}^{-1}(\widehat{\theta}_n) M_{+,n}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_o)/d_\theta \\ &\to_d \frac{K}{K - d_\theta + 1} \mathcal{F}_{d_\theta, K-d_\theta+1}, \end{aligned}$$

where $\mathcal{F}_{d_\theta,K-d_\theta+1}$ *is a F random variable with degree of freedom* $(d_\theta, K - d_\theta + 1)$ *and* d_θ *denotes the dimensionality of* θ_o*.*

Lemma 12.5 shows that when the number of the trend basis functions K is fixed, the series LRV estimate $V_{K,n}(\widehat{\theta}_n)$ is inconsistent, but the Wald-type test statistic $\mathcal{F}_n$ is asymptotically pivotal. Autocorrelation robust inference about θ_o can be conducted using the statistic $\mathcal{F}_n^* \equiv (K - d_\theta + 1)\mathcal{F}_n/K$ and the asymptotic $\mathcal{F}_{d_\theta,K-d_\theta+1}$ distribution. As noted in Sun (2013), the restriction $\int_0^1 \phi_k(r)\ dr = 0$ $(k = 1,\ldots,K)$ helps to remove the estimation effect in $\widehat{\theta}_n$ from the asymptotic distribution of $V_{K,n}(\widehat{\theta}_n)$. As a result, the statistic $\mathcal{F}_n^*$ enjoys an exact asymptotic F-distribution. Using arguments similar to those in Phillips (2005b), it can be shown that under some suitable rate condition on K the series LRV estimate $V_{K,n}(\widehat{\theta}_n)$ is consistent, that is,

$$V_{K,n}(\widehat{\theta}_n) = \frac{1}{K}\sum_{k=1}^{K}\Lambda_{k,n}\Lambda'_{k,n} \rightarrow_p V(\theta_o),$$

as $n, K \to \infty$ jointly. In that case, the test statistic F_n has an asymptotic chi-square distribution with d_θ degrees of freedom.

Orthonormal series LRV estimates are becoming increasingly popular for autocorrelation robust inference in econometric models. Sun (2011) proposed a new testing procedure for hypotheses on deterministic trends in a multivariate trend stationary model, where the LRV is estimated by the series method. For empirical applications, the paper provides an optimal procedure for selecting K in the sense that the type II error is minimized while controlling for the type I error. Sun (2013) uses a series LRV estimate for autocorrelation robust inference in parametric M-estimation. This paper also shows that critical values from the fixed-K limit distribution of the Wald-type test statistic are second-order correct under conventional increased-smoothing asymptotics. Sun and Kim (2012, 2013) use the series LRV estimate for inference and specification testing in a generalized method of moments (GMM) setting. The series LRV estimate has also been used in inference for semi/nonparametric econometric models with dependent data. In particular, recent work of Chen, Hahn, and Liao (2012) uses the series method to estimate the LRV of a two-step GMM estimate when there are some infinite-dimensional parameters estimated by first-step sieve M-estimation. In related work, Chen, Liao, and Sun (2012) use series methods to estimate the LRVs of sieve estimates of finite-dimensional and infinite-dimensional parameters in semi-/nonparametric models with weakly dependent data.

12.8. Concluding Remarks

As explained in previous sections, the KL representation of stochastic processes can be very useful in modeling, estimation, and inference in econometrics. This chapter has outlined the theory behind the KL representation and some of its properties. The applications of the KL representation that we have reviewed belong to three categories:

(i) The link between stochastic trends and their deterministic trend representations. This link is a powerful tool for understanding the relationships between the two forms of trend and the implications of these relationships for practical work. As we have discussed, the KL representation provides new insights that help explain spurious regressions as a natural phenomena when an integrated or near-integrated process is regressed on a set of deterministic trend variables. And the representation helps to demonstrate the effect of adding deterministic trends or trend breaks to regressions in which unit root tests are conducted;

(ii) The KL representation reveals that traditional warnings of spurious regressions as uniformly harmful is unjustified. For example, as recovered in its KL representation, an integrated process can be perfectly modelled by trend basis functions. This relation, which in traditional theory is viewed as a spurious regression, turns out to be extremely useful in the efficient estimation of the cointegrated systems as discussed in Section 12.5.

(iii) Trend basis functions may be used to fit stationary processes, leading to a novel LRV estimation method that is simple and effective because of the natural focus on long-run behavior in the trend basis. The resulting series LRV estimate is automatically positive definite and is extremely easy to compute. Moreover, t-ratio and Wald-type test statistics constructed using the series LRV estimate are found to have standard limit distributions under both fixed-K and large-K asymptotics. These features make the use of series LRV estimation attractive for practical work in econometrics, as discussed in Section 12.7.

There are many potential research directions that seem worthy of future research. We mention some of these possibilities in what follows.

First, KL representations of nondegenerate or full-rank stochastic processes[7]are discussed in this chapter. It would be interesting to study KL forms of vector processes that are of deficient rank, such as multiple time series that are cointegrated. Phillips (2005a) gives some discussion of this idea and introduces the concept of coordinate cointegration in this context, which subsumes the usual cointegration concept. In this context, trend basis functions may be useful in testing for co-movement and efficient estimation of co-moving systems when system rank is unknown.

Second, trend basis representations of different stochastic processes differ. Such differences may be used to test if observed data are compatible with a certain class of stochastic processes. For example, one may be interested in testing a BM null against an O-U process alternative. From Section 12.2, we know that BM has the following KL representation:

$$B(t) = \sqrt{2}\sum_{k=1}^{\infty}\frac{\sin[(k-1/2)\pi t]}{(k-1/2)\pi}\xi_{\omega,k}, \tag{12.71}$$

where $\xi_{\omega,j}$ are i.i.d. $N(0,\omega^2)$ and ω^2 is the variance of $B(\cdot)$. Using the above representation and the expression in (12.22), we obtain the following alternate representation of an O-U process (cf. Phillips (1998)):

$$J_c(t) = \sqrt{2}\sum_{k=1}^{\infty}\frac{\xi_{\omega,k}}{(k-1/2)\pi}\left\{\sin[(k-1/2)\pi t] + c\int_0^t e^{(t-s)c}\sin[(k-1/2)\pi s]ds\right\}$$

$$
\begin{aligned}
&= \sqrt{2}\sum_{k=1}^{\infty}\frac{\xi_{\omega,k}}{(k-1/2)^2\pi^2+c^2}\{ce^{ct}-c\cos[(k-1/2)\pi t]\\
&\quad +(k-1/2)\pi\sin[(k-1/2)\pi t]\}. \qquad (12.72)
\end{aligned}
$$

If the data $\{X_t\}$ are projected on the trend IVs $\{\sin[(k-\frac{1}{2})\frac{\pi t}{n}],\ \cos[(k-\frac{1}{2})\frac{\pi t}{n}]:$ $k \le K\}$, then under the null, the projection will reproduce the representation in (12.71) when $K \to \infty$. However, under the alternative, as is apparent from (12.72), the projection has an asymptotic form that is very different from (12.71) and includes the cosine and exponential functions. It is of interest to see if significance tests on the coefficients in this regression can usefully discriminate integrated and locally integrated processes that have BM and O-U process limits after standardization.

Third, although trend basis functions are effective in modeling integrated processes and can be used to efficiently estimate cointegration systems, in finite samples it is not clear how many trend basis functions should be used. From the KL representation of BM in (12.71), it is apparent that the trend IVs $\{\sqrt{2}\sin[(k-1/2)\pi t]\}_k$ have a natural ordering according to the variances of their random coefficients $\{\frac{\xi_{\omega,k}}{(k-1/2)\pi}\}_{k=1}^{\infty}$. This ordering is useful in itself for selecting trend IVs, but it would also be useful to calculate the asymptotic mean square error (AMSE) of the trend IV estimate. Then an optimal IV selection criterion could be based on minimizing the empirical AMSE. However, calculation of the AMSE is complicated by the mixed normal limit theory of trend IV estimates and the presence of functional limits in the first-order asymptotics, so explicit formulae are not presently available.

In other recent work, Liao and Phillips (2011) propose to select trend IVs using Lasso penalized estimation. In particular, in the notation of Section 12.6 of the present chapter, trend IVs can be selected by means of the following penalized LS regression:

$$
\min_{\Pi\in R^{K\times 2m_x}} \|Z_n-\Phi_K\Pi\|^2+n\lambda_n\sum_{k=1}^{K}\|\Pi_k\|, \qquad (12.73)
$$

where $Z_n' = [n^{-\frac{1}{2}}X_1,\ldots,n^{-\frac{1}{2}}X_n]$, Π_k denotes the kth row ($k=1,\ldots,K$) of the $K \times m_x$ coefficient matrix Π and λ_n is a tuning parameter. The coefficient vector Π_k is related to the kth trend IV $\varphi_k(\cdot)$; and if Π_k is estimated as zero, then the kth trend IV $\varphi_k(\cdot)$ would not be used as an instrument for the "endogenous" variable Z. The tuning parameter λ_n determines the magnitude of the shrinkage effect on the estimator of Π_k. The larger the tuning parameter λ_n, the larger the shrinkage effect will be, leading to more zero coefficient estimates in Π_k. In consequence, the problem of trend IV selection becomes a problem of selecting the tuning parameter λ_n. Liao and Phillips (2011) provide data-driven tuning parameters in the penalty function, making Lasso IV selection fully adaptive for empirical implementation.

Fourth, as noted in Phillips (2005a), the KL representation, when restricted to a subinterval of $[0,1]$ such as $[0,r]$ $(r \in (0,1))$, is useful in studying the evolution of a trend process over time. For example, the KL representation of BM on $[0,r]$ has the following form

$$B(s) = \sum_{k=1}^{\infty} \varphi_k\Big(\frac{s}{r}\Big)\eta_k(r) \qquad \text{for any } s \in [0,r], \tag{12.74}$$

where $\eta_k(r) = r^{-1}\int_0^r B(s)\varphi_k\big(\frac{s}{r}\big)ds$. It follows that $B(r) = \sum_{k=1}^{\infty}\varphi_k(1)\eta_k(r)$, where $B(r)$ and $\eta_k(r)$ are both measurable with respect to the natural filtration $\mathcal{F}_r$ of the BM $B(\cdot)$. The process $\eta_k(r)$ describes the evolution over time of the coefficient of the coordinate basis $\varphi_k(\cdot)$. The evolution of these trend coordinates can be estimated by recursively regressing the sample data on the functions $\varphi_k(\cdot)$, and the resulting estimates deliver direct information on how individual trend coordinates have evolved over time.

The restricted KL representation in (12.74) may also be used for forecasting. In particular, setting $s = r$ in (12.74), the optimal predictor of $B(r)$ given $\mathcal{F}_p$ and coordinates up to K is

$$E\big[\,B(r)|\mathcal{F}_p, K\big] = \sum_{k=1}^{K}\varphi_k(1)E\big[\,\eta_k(r)|\mathcal{F}_p\big]. \tag{12.75}$$

By the definition of $\eta_k(\cdot)$ and using explicit formulae for φ_k, the conditional expectation in (12.75) can be written as

$$E\big[\,\eta_k(r)|\mathcal{F}_p\big] = \frac{1}{r}\int_0^p B(s)\varphi_k\Big(\frac{s}{r}\Big)ds + B(p)\frac{\sqrt{2}\cos\big[(k-1/2)\frac{\pi p}{r}\big]}{(k-1/2)\pi}. \tag{12.76}$$

Summing over $k = 1,\ldots,K$, we get

$$E\big[\,B(r)|\mathcal{F}_p, K\big] = \sum_{k=1}^{K}\varphi_k(1)\left[\frac{1}{r}\int_0^p B(s)\varphi_k\Big(\frac{s}{r}\Big)ds + \frac{\sqrt{2}\cos\big[\big(k-\frac{1}{2}\big)\frac{\pi p}{r}\big]B(p)}{\big(k-\frac{1}{2}\big)\pi}\right]. \tag{12.77}$$

Let $N = [np]$ and $N + h = [nr]$ so that (12.76) and (12.77) effectively provide h-step ahead optimal predictors of these components. $E\big[\,\eta_k(r)|\mathcal{F}_p\big]$ may be estimated from sample data by

$$\widehat{\eta}_k(r|p) = \sum_{t=1}^{N}\frac{X_t/\sqrt{n}}{N+h}\varphi_k\Big(\frac{t}{N+h}\Big) + \frac{\sqrt{2}\cos\Big[(k-1/2)\frac{\pi N}{N+h}\Big]X_N/\sqrt{n}}{(k-1/2)\pi},$$

which leads to the following h-step ahead predictor of the trend in the data:

$$\widehat{X}_{N+h,N} = \sum_{k=1}^{K}\varphi_k(1)\left[\sum_{t=1}^{N}\frac{X_t}{N+h}\varphi_k\Big(\frac{t}{N+h}\Big) + \frac{\sqrt{2}\cos\Big[(k-1/2)\frac{\pi N}{N+h}\Big]X_N}{(k-1/2)\pi}\right].$$

As pointed out in Phillips (2005a), this forecasting approach can be pursued further to construct formulae for trend components and trend predictors corresponding to a variety of long-run models for the data. Such formulae enable trend analysis and prediction in a way that captures the main features of the trend for K small and which can be related back to specific long-term predictive models for large K. The approach therefore helps to provide a foundation for studying trends in a general way, covering most of the trend models that are presently used for economic data.

Finally, in general semiparametric and nonparametric models, the series-based LRV estimation method described earlier also requires a selection procedure to determine the number of the trend basis functions. The test-optimal procedures proposed in Sun (2011, 2013) may be generalized to semiparametric and nonparametric models. Moreover, current applications of series LRV estimation methods involve semiparametric or nonparametric models of stationary data. It is of interest to extend this work on series LRV estimation and associated inference procedures to econometric models with nonstationary data.

12.9. Appendix

Proof of Lemma 12.1. The proof of this lemma is included for completeness. The symmetry of $\gamma(\cdot,\cdot)$ follows by its definition. To show continuity, note that for any $t_o, s_o, t_1, s_1 \in [a,b]$, by the triangle and Hölder inequalities we have

$$\begin{aligned}|\gamma(t_1,s_1)-\gamma(t_o,s_o)| &= |E[X(s_1)X(t_1)]-E[X(s_o)X(t_o)]| \\ &\leq \|X(t_1)\|\|X(s_1)-X(s_o)\| + \|X(s_o)\|\|X(t_1)-X(t_o)\|,\end{aligned}$$

which together with the q.m. continuity of $X(\cdot)$ implies that

$$|\gamma(t_1,s_1)-\gamma(t_o,s_o)| \to 0 \tag{12.78}$$

for any $t_o, s_o, t_1, s_1 \in [a,b]$ such that $t_1 \to t_o$ and $s_1 \to s_o$. The convergence in (12.78) implies that $\gamma(\cdot,\cdot)$ is a continuous function on $[a,b]\times[a,b]$ with $|\gamma(a,a)| < \infty$ and $|\gamma(b,b)| < \infty$. As a result, we get the following condition:

$$\max_{t\in[a,b]} |\gamma(t,t)| < \infty. \tag{12.79}$$

Furthermore, we see that for any $g \in L^2[a,b]$ we obtain

$$\begin{aligned}\int_a^b \int_a^b g(t)\gamma(t,s)g(s)\, dsdt &= \int_a^b \int_a^b E\big[g(t)X(t)g(s)X(s)\big]\, dsdt \\ &= E\left[\int_a^b g(t)X(t)\int_a^b g(s)X(s)dsdt\right] \\ &= E\left[\left(\int_a^b g(t)X(t)dt\right)^2\right] \geq 0, \end{aligned} \tag{12.80}$$

where the second equality is by (12.79) and Fubini's Theorem. ■

Proof of Theorem 12.1. The proof of this theorem is included for completeness. Let $Z_k \equiv \int_a^b X(t)\varphi_k(t)\,dt$. Then it is clear that

$$E[Z_k] = E\left[\int_a^b X(t)\varphi_k(t)\,dt\right] = \int_a^b E[X(t)]\varphi_k(t)\,dt = 0 \qquad (12.81)$$

and

$$\begin{aligned} E[Z_k Z_{k'}] &= E\left[\int_a^b \int_a^b X(s)X(t)\varphi_k(t)\varphi_{k'}(s)\,dsdt\right] \\ &= \int_a^b \int_a^b \gamma(s,t)\varphi_k(t)\varphi_{k'}(s)\,dsdt \\ &= \lambda_k \int_a^b \varphi_k(t)\varphi_{k'}(t)\,dt = \lambda_k\delta_{kk'}, \end{aligned} \qquad (12.82)$$

and moreover

$$E[Z_k X(t)] = E\left[X(t)\int_a^b X(t)\varphi_k(t)dt\right] = \int_a^b \gamma(t,s)\varphi_k(s)dt = \lambda_k\varphi_k(t), \qquad (12.83)$$

for any $k, k' \in \mathbb{Z}_+$. Note that the uniform bound of $\gamma(\cdot,\cdot)$ and Fubini's theorem ensure that we can exchange the integration and expectation in (12.81)–(12.83). Let M be some positive integer, then by definition, (12.83), and uniform convergence in (12.9), we deduce that

$$\begin{aligned} \left\| X(t) - \sum\nolimits_{k=1}^M Z_k\varphi_k(t) \right\|^2 &= \gamma(t,t) - 2\sum\nolimits_{k=1}^M \varphi_k(t)E[Z_kX(t)] + \sum\nolimits_{k=1}^M \lambda_k\varphi_k^2(t) \\ &= \gamma(t,t) - \sum\nolimits_{k=1}^M \lambda_k\varphi_k^2(t) \\ &= \sum\nolimits_{k=M+1}^{\theta} \lambda_k\varphi_k^2(t) \to 0, \qquad \text{as } M \to \infty \end{aligned} \qquad (12.84)$$

uniformly over $t \in [a,b]$, which proves sufficiency. Next suppose that $X(t)$ has the following representation

$$X(t) = \sum_{k=1}^{\infty} \alpha_k^{\frac{1}{2}}\xi_k^* g_k(t) \qquad \text{with } E\left[\xi_k^*\xi_{k'}^*\right] = \int_a^b g_k(t)g_{k'}(t)\,dt = \delta_{kk'}.$$

Then by definition we have

$$\gamma(s,t) = E\left[\sum_{j=1}^{\infty} \alpha_k^{\frac{1}{2}}\xi_k^* g_k(s) \sum_{j=1}^{\infty} \alpha_k^{\frac{1}{2}}\xi_k^* g_k(t)\right]$$

$$
\begin{aligned}
&= \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \alpha_j^{\frac{1}{2}} \alpha_k^{\frac{1}{2}} g_j(s) g_k(t) \delta_{jk} \\
&= \sum_{k=1}^{\infty} \alpha_k g_k(s) g_k(t).
\end{aligned}
$$

Hence for any $k \in \mathbb{Z}_+$ we obtain

$$
\int_a^b \gamma(t,s) g_k(s)\, dt = \int_a^b \left[\sum\nolimits_{j=1}^{\infty} \alpha_j g_j(t) g_j(s) g_k(s) \right] ds = \sum_{j=1}^{\infty} \alpha_j g_j(t) \delta_{jk} = \alpha_k g_k(t),
$$

which implies that $\{(\alpha_k, g_k)\}_{k=1}^{\infty}$ are the eigenvalues and orthonormal eigenfunctions of the kernel function $\gamma(\cdot,\cdot)$. This proves necessity. ■

Proof of Lemma 12.3. First, note that

$$
n(\widehat{A}_{K,n} - A_o) = \frac{\frac{1}{n} U'_{y\cdot x} R_{\Delta X,K} X}{\frac{1}{n^2} X' R_{\Delta X,K} X}.
$$

We next establish the asymptotic distributions of related quantities in the above expression.

$$
\begin{aligned}
\frac{X' P_K X}{n^2} &= \frac{X' \Phi_K \left(\Phi'_K \Phi_K\right)^{-1} \Phi'_K X}{n^2} \\
&= \frac{\sum_{t=1}^n X_t \Phi_K(\frac{t}{n})}{n^{\frac{3}{2}}} \left(\frac{\sum_{t=1}^n \Phi'_K(\frac{t}{n}) \Phi_K(\frac{t}{n})}{n} \right)^{-1} \frac{\sum_{t=1}^n X_t \Phi'_K(\frac{t}{n})}{n^{\frac{3}{2}}} \\
&\to_d \left[\int_0^1 B_x(r) \Phi_K(r)\, dr \right] \left[\int_0^1 B_x(r) \Phi'_K(r)\, dr \right] \\
&\overset{d}{=} \sum_{k=1}^K \left[\int_0^1 B_x(r) \varphi_k(r)\, dr \right]^2 \overset{d}{=} \sum_{k=1}^K \xi_{x,k}^2.
\end{aligned} \tag{12.85}
$$

$$
\begin{aligned}
\Delta X' P_K \Delta X &= \Delta X' \Phi_K \left(\Phi'_K \Phi_K\right)^{-1} \Phi'_K \Delta X \\
&= \frac{\sum_{t=1}^n \Delta X_t \Phi_K(\frac{t}{n})}{n^{\frac{1}{2}}} \left(\frac{\sum_{t=1}^n \Phi'_K(\frac{t}{n}) \Phi_K(\frac{t}{n})}{n} \right)^{-1} \frac{\sum_{t=1}^n \Delta X_t \Phi'_K(\frac{t}{n})}{n^{\frac{1}{2}}} \\
&\to_d \left[\int_0^1 \Phi_K(r)\, dB_x(r) \right] \left[\int_0^1 \Phi'_K(r)\, dB_x(r) \right] \\
&\overset{d}{=} \sum_{k=1}^K \left[\int_0^1 \varphi_k(r)\, dB_x(r) \right]^2 \overset{d}{=} \sum_{k=1}^K \eta_{x,k}^2.
\end{aligned} \tag{12.86}
$$

$$\begin{aligned}
\frac{X'P_K\Delta X}{n} &= \frac{X'\Phi_K\left(\Phi_K'\Phi_K\right)^{-1}\Phi_K'\Delta X}{n} \\
&= \frac{\sum_{t=1}^n X_t\Phi_K(\frac{t}{n})}{n^{\frac{3}{2}}}\left(\frac{\sum_{t=1}^n \Phi_K'(\frac{t}{n})\Phi_K(\frac{t}{n})}{n}\right)^{-1}\frac{\sum_{t=1}^n \Delta X_t\Phi_K'(\frac{t}{n})}{n^{\frac{1}{2}}} \\
&\to_d \left[\int_0^1 B_x(r)\Phi_K(r)\,dr\right]\left[\int_0^1 \Phi_K'(r)\,dB_x(r)\right] \\
&\overset{d}{=} \sum_{k=1}^K \int_0^1 B_x(r)\varphi_k(r)\,dr \int_0^1 \varphi_k(r)\,dB_x(r) \overset{d}{=} \sum_{k=1}^K \xi_{x,k}\eta_{x,k}. \qquad (12.87)
\end{aligned}$$

The results in (12.85), (12.86) and (12.87) imply that

$$\begin{aligned}
\frac{X'R_{\Delta X,K}X}{n^2} &= \frac{X'P_KX}{n^2} - \frac{X'P_K\Delta X}{n}\left(\Delta X'P_K\Delta X\right)^{-1}\frac{\Delta X'P_KX}{n} \\
&\to_d \sum_{k=1}^K \lambda_k\xi_{x,k}^2 - \frac{\left[\sum_{k=1}^K \xi_{x,k}\eta_{x,k}\right]^2}{\sum_{k=1}^K \eta_{x,k}^2}. \qquad (12.88)
\end{aligned}$$

Next, note that

$$\begin{aligned}
\frac{U_{y\cdot x}'P_KX}{n} &= \frac{U_{y\cdot x}'\Phi_K\left(\Phi_K'\Phi_K\right)^{-1}\Phi_K'X}{n} \\
&= \frac{\sum_{t=1}^n u_{y\cdot x,t}\Phi_K(\frac{t}{n})}{n^{\frac{1}{2}}}\left(\frac{\sum_{t=1}^n \Phi_K'(\frac{t}{n})\Phi_K(\frac{t}{n})}{n}\right)^{-1}\frac{\sum_{t=1}^n X_t\Phi_K'(\frac{t}{n})}{n^{\frac{3}{2}}} \\
&\to_d \left[\int_0^1 \Phi_K(r)\,dB_{y\cdot x}(r)\right]'\left[\int_0^1 \Phi_K'(r)B_x(r)\,dr\right] \\
&\overset{d}{=} \sum_{k=1}^K \xi_{x,k}\eta_{y\cdot x,k} \qquad (12.89)
\end{aligned}$$

and

$$\begin{aligned}
U_{y\cdot x}'P_K\Delta X &= U_{y\cdot x}'\Phi_K\left(\Phi_K'\Phi_K\right)^{-1}\Phi_K'\Delta X \\
&= \frac{\sum_{t=1}^n u_{y\cdot x,t}\Phi_K(\frac{t}{n})}{n^{\frac{1}{2}}}\left(\frac{\sum_{t=1}^n \Phi_K'(\frac{t}{n})\Phi_K(\frac{t}{n})}{n}\right)^{-1}\frac{\sum_{t=1}^n \Delta X_t\Phi_K'(\frac{t}{n})}{n^{\frac{1}{2}}} \\
&\to_d \left[\int_0^1 \Phi_K(r)dB_{y\cdot x}(r)\right]\left[\int_0^1 \Phi_K'(r)dB_x(r)\right] \\
&\overset{d}{=} \sum_{k=1}^K \eta_{x,k}\eta_{y\cdot x,k}. \qquad (12.90)
\end{aligned}$$

The results in (12.86), (12.87), (12.88), (12.89), and (12.90) imply that

$$\frac{U'_{y\cdot x}R_{\Delta X,K}X}{n} = \frac{U'_{y\cdot x}P_K X}{n} - U'_{y\cdot x}P_K\Delta X\big(\Delta X' P_K \Delta X\big)^{-1}\frac{\Delta X' P_K X}{n}$$
$$\to_d \sum_{k=1}^{K}\xi_{x,k}\eta_{y\cdot x,k} - \frac{\sum_{k=1}^{K}\eta_{x,k}\eta_{y\cdot x,k}\sum_{k=1}^{K}\xi_{x,k}\eta_{x,k}}{\sum_{k=1}^{K}\eta_{x,k}^2}. \qquad (12.91)$$

The result in (12.45) follows directly by (12.88) and (12.91).

For the second result, note that

$$\widehat{B}_{K,n} = B_o + \frac{U'_{y\cdot x}R_{X,K}\Delta X}{\Delta X' R_{X,K}\Delta X}.$$

The asymptotic distributions of the quantities in the above expression are obtained as follows. Under (12.85), (12.86) and (12.87), we have

$$\Delta X' R_{X,K}\Delta X = \Delta X' P_K \Delta X - \frac{\Delta X' P_K X}{n}\left(\frac{X' P_K X}{n^2}\right)^{-1}\frac{X' P_K \Delta X}{n}$$
$$\to_d \sum_{k=1}^{K}\eta_{x,k}^2 - \frac{\left[\sum_{k=1}^{K}\xi_{x,k}\eta_{x,k}\right]^2}{\sum_{k=1}^{K}\xi_{x,k}^2}. \qquad (12.92)$$

Similarly, under (12.85), (12.89) and (12.90), we have

$$U'_{y\cdot x}R_{X,K}\Delta X = U'_{y\cdot x}P_K\Delta X - \frac{U'_{y\cdot x}P_K X}{n}\left(\frac{X' P_K X}{n^2}\right)^{-1}\frac{X' P_K \Delta X}{n}$$
$$\to_d \sum_{k=1}^{K}\eta_{x,k}\eta_{y\cdot x,k} - \frac{\sum_{k=1}^{K}\xi_{x,k}\eta_{y\cdot x,k}\sum_{k=1}^{K}\xi_{x,k}\eta_{x,k}}{\sum_{k=1}^{K}\xi_{x,k}^2}. \qquad (12.93)$$

The result in (12.45) follows directly by (12.92) and (12.93). ■

Proof of Lemma 12.4. By (12.63) and the continuous mapping theorem (CMT), we obtain

$$\omega_{K,n}^2 \to_d \frac{\omega^2\sum_{k=1}^{K}\left[\frac{1}{\omega}\int_0^1 \phi_k(r)\,dB_\omega(r)\right]^2}{K} \overset{d}{=} \frac{\omega^2}{K}\chi^2(K), \qquad (12.94)$$

where the equivalence in distribution follows from the fact that $\frac{1}{\omega}\int_0^1 \phi_k(r)\,dB_\omega(r)$ is a standard normal random variable for any k and is independent of $\frac{1}{\omega}\int_0^1 \phi_{k'}(r)\,dB_\omega(r)$ for any $k \neq k'$. From (12.62), (12.94), and the CMT, we deduce that

$$t_{K,n} = \frac{B_n(1)}{\sqrt{\omega_{K,n}^2}} \to_d \frac{B_\omega(1)/\omega}{\sqrt{\chi^2(K)/K}} \overset{d}{=} t_K, \qquad (12.95)$$

where the equivalence in distribution follows by definition of the Student t and the fact that $B_{\omega}(1)$ is independent of $\int_0^1 \phi_{k'}(r)dB_{\omega}(r)$ for any k. ■

Proof of Lemma 12.5. First note that we can rewrite

$$
\begin{aligned}
& n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})m(Z_t,\widehat{\theta}_n) \\
&= n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})m(Z_t,\theta_0)+n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})E\big[m(Z_t,\widehat{\theta}_n)-m(Z_t,\theta_0)\big] \\
&\quad +n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})\big\{m(Z_t,\widehat{\theta}_n)-m(Z_t,\theta_0)-E\big[m(Z_t,\widehat{\theta}_n)-m(Z_t,\theta_0)\big]\big\}.
\end{aligned}
\tag{12.96}
$$

By Assumption 12.8.(i), (ii) and $\int_0^1 \phi_k(r)dr=0$, we have

$$
n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})E\big[m(Z_t,\widehat{\theta}_n)-m(Z_t,\theta_0)\big]=\frac{1}{n}\sum_{t=1}^{n}\phi_k(\frac{t}{n})O_p(1)=o_p(1). \tag{12.97}
$$

Hence, using the results in (12.96), (12.97) and Assumption 12.8.(iii)–(iv), we deduce that

$$
\begin{aligned}
n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})m(Z_t,\widehat{\theta}_n) &= n^{-\frac{1}{2}}\sum_{t=1}^{n}\phi_k(\frac{t}{n})m(Z_t,\theta_0)+o_p(1) \\
&\to_d \int \phi_k(r)dB_m(r)\equiv\xi_k.
\end{aligned}
\tag{12.98}
$$

Under Assumption 12.8.(i), (ii), and (v), we get

$$
\sqrt{n}V^{-\frac{1}{2}}(\theta_o)M_{+,n}(\widehat{\theta}_n)(\widehat{\theta}_n-\theta_o)\to_d N(0,I_{d_\theta})\overset{d}{=}\xi_0. \tag{12.99}
$$

Using the results in (12.98), (12.99) and the CMT, we deduce that

$$
\begin{aligned}
d_{\theta}F_n &= \Big[V^{-\frac{1}{2}}(\theta_o)M_{+,n}(\widehat{\theta}_n)\sqrt{n}(\widehat{\theta}_n-\theta_o)\Big]' \\
&\quad\times\left\{\frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{n}V^{-\frac{1}{2}}(\theta_o)\Lambda_{k,n}\Lambda_{k,n}'V^{-\frac{1}{2}}(\theta_o)\right]\right\}^{-1} \\
&\quad\times\Big[V^{-\frac{1}{2}}(\theta_o)M_{+,n}(\widehat{\theta}_n)\sqrt{n}(\widehat{\theta}_n-\theta_o)\Big] \\
&\to_d \xi_0'\left(\frac{1}{K}\sum_{k=1}^{K}\xi_k\xi_k'\right)^{-1}\xi_0,
\end{aligned}
$$

which has Hotelling's T^2-distribution. Using the relation between the T^2-distribution and $\mathcal{F}$-distribution, we get

$$\frac{K-d_\theta+1}{K}\mathcal{F}_n \to_d \mathcal{F}_{d_\theta, K-d_\theta+1},$$

which finishes the argument. ■

Notes

† Our thanks to the referee and editors for helpful comments on the original version of this paper.

1. The Hilbert space generated by the stochastic process $\{X_t\}_{t\in\mathcal{T}}$ is the completion of the space defined as the linear span of any finite elements $X_{t_1},\ldots,X_{t_n}$, where $t_k \in \mathcal{T}$, $k = 1,\ldots,n$ and $n = 1,2,\ldots$.
2. Similarly, the series representation of a continuous function may not converge pointwise unless the function has right and left derivatives at that point.
3. The specific orthonormal representation of BM given in (12.16) can of course be used here. But we use the representation in (12.12) to make the results of this section applicable to general basis functions.
4. The divergent behavior of the t-statistics might be thought to be a consequence of the use of OLS standard errors based on $n^{-1}\sum_{i=1}^{n}\widehat{u}_{t,K}^2$ which do not take account of serial dependence in the residuals. However, Phillips (1998) confirmed that divergence at a reduced rate continues to apply when HAC standard errors are used (employing an estimate of the long-run variance (LRV)). On the other hand, if HAR estimates rather than HAC estimates are used (for example, a series LRV estimate with fixed number of basis functions, see Section 12.7 for details), the t-statistics no longer diverge in general. Theorem 12.2 simply illustrates the spurious regression phenomenon when standard testing procedures based on OLS are employed.
5. The trend IV estimate is related to the spectral regression estimates proposed in Phillips (1991b), although those estimates are formulated in the frequency domain. Spectral regression first transfers the cointegration system (12.40) and (12.41) to frequency domain ordinates and then estimates A_o by GLS regression. The spectral transformation projects the whole model on the deterministic function $\exp(i\lambda t)$ at different frequencies $\lambda \in R$, which helps to orthogonalize the projections at different frequencies. However, optimal weights constructed using the empirical spectral density are used in this procedure. Phillips (1991b) also gives a narrow band spectral estimation procedure that uses frequency ordinates in the neighborhood of the origin. Trend IV estimation only projects the (endogenous) regressors on the deterministic functions (trend IVs) and does not need optimal weighting to achieve efficiency. It is more closely related to the narrow band procedure but does not involve frequency domain techniques.
6. For any $m \times m$ matrix Π, we order the eigenvalues of Π in decreasing order by their moduli, that is, $|\phi_1(\Pi)| \geq |\phi_2(\Pi)| \geq \cdots \geq |\phi_m(\Pi)|$. For complex conjugate eigenvalues, we order the eigenvalue a positive imaginary part before the other.

7. A full-rank or nondegenerate process refers to a random sequence that upon scaling satisfies a functional law with a nondegenerate limit process, such as a Brownian motion with positive definite variance matrix.

References

Anderson, T. W., and D. A. Darling. 1952. "Asymptotic Theory of Certain 'Goodness-of-fit' Criteria Based on Stochastic Processes." *Annals of Mathematical Statistics*, **23**, pp. 193–212.

Berlinet, A., and C. Thomas-Agnan. 2003. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Boston: Kluwer Academic Publishers.

Bosq, D. 2000. *Linear Processes in Function Spaces*, Lecture Notes in Statistics. Berlin: Springer.

Chen X., Z. Liao, and Y. Sun. 2012. "Sieve Inference on Semi-nonparametric Time Series Models." *Cowles Foundation Discussion Paper*, No 1849.

Chen X., J. Hahn, and Z. Liao. 2012. "Simple Estimation of Asymptotic Variance for Semiparametric Two-step Estimators with Weakly Dependent Data." Working paper, Yale University and UCLA.

Cheng X., and P. C. B. Phillips. 2009. "Semiparametric Cointegrating Rank Selection." *Econometrics Journal*, **12**, pp. 83–104,

Cheng X., and P. C. B. Phillips 2012. "Cointegrating Rank Selection in Models with Time-Varying Variance." *Journal of Econometrics*, **169**(2), pp. 155–165.

Durlauf, S. N., and P. C. B. Phillips. 1988. "Trends Versus Random Walks in Time Series Analysis." *Econometrica*, **56**, pp. 1333–1354.

Granger, C. W. J. and P. Newbold. 1974. "Spurious Regression in Econometrics." *Journal of Econometrics*, **2**, pp. 111–120.

Hannan, E. J. 1970. *Multiple Time Series*, New York: John Wiley & Sons.

Johansen S. 1998. "Statistical Analysis of Cointegration Vectors." *Journal of Economic Dynamics and Control*, **12**(2–3), pp. 231–254.

Johansen S. 1995. *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models.* New York: Oxford University Press.

Kac, M., and A. J. F. Siegert. 1947. "An Explicit Representation of a Stationary Gaussian Process." *Annals of Mathematical Statistics,* **18**, pp. 438–442.

Karhunen, K. 1946. "Zur Spektraltheorie Stochastischer Prozesse." *Annales Academiae Scientiarum Fennicae*, **37**.

Liao, Z., and P. C. B. Phillips. 2012. "Automated Estimation of Vector Error Correction Models." Working paper, Yale University and UCLA.

Liao Z., and P. C. B. Phillips. 2010. "Reduced Rank Regression of Partially Non-stationary Vector Autoregressive Processes under Misspecification." Working paper, Yale University and UCLA.

Liao Z. and P. C. B. Phillips. 2011. "A Lasso-type of IV Selection Approach in the Efficient Estimation of Cointegrated Systems." Working paper, Yale University and UCLA.

Loéve, M. M. 1955. *Probability Theory.* Princeton, NJ: Van Nostrand.

Loève, M. 1977. *Probability Theory*, New York: Springer.

Ouliaris, S., J. Y. Park, and P. C. B. Phillips. 1989. "Testing for a Unit Root in the Presence of a Maintained Trend." In *Advances in Econometrics and Modelling*, ed. B. Raj. Norwell, MA: Kluwer Academic Publishers, Chapter 1.

Ploberger, W., and P. C. B. Phillips. 2003. "Empirical Limits for Time Series Econometric Models." *Econometrica*, **71**(2), pp. 627–673.

Parzen, E. 1959. "Statistical Inference on Time Series by Hilbert Space Methods." *Technical Report*, 23, Statistics Department, Stanford University.

Parzen, E. 1961a. "An Approach to Time Series Analysis." *Annals of Mathematical Statistics*, **32**(4), pp. 951–989.

Parzen, E. 1961b. "Regression Analysis of Continuous Parameter Time Series." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press.

Parzen, E. 1963. "Probability Density Functionals and Reproducing Kernel Hilbert Spaces." In *Proceedings Symposium on Time Series Analysis*, ed. M. Rosenblatt. Wiley, New York: John Wiley & Sons.

Philips, P. C. B. 1986. "Understanding Spurious Regressions in Economics." *Journal of Econometrics*, **33**(3), 311–340.

Phillips, P. C. B. 1987. "Time Series Regression with a Unit Root." *Econometrica,* **55**, pp. 277–302.

Phillips, P. C. B. 1989. "Partially Identified Econometric Models." *Econometric Theory*, **5**, pp. 181–240.

Phillips, P. C. B. 1991a. "Optimal Inference in Cointegrated Systems." *Econometrica*, **59**(2), pp. 283–306.

Phillips, P. C. B. 1991b. "Spectral Regression for Cointegrated Time Series." In *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics*, eds. W. Barnett, J. Powell, and G. Tauchen. Cambridge University Press.

Phillips, P. C. B. 1995. "Fully Modified Least Squares and Vector Autoregression." *Econometrica*, **63**(5), pp. 1023–1078.

Phillips, P. C. B. 1998. "New Tools for Understanding Spurious Regressions." *Econometrica*, **66**(6), pp. 1299–1326.

Phillips, P. C. B. 2001. "New Unit Root Asymptotics in the Presence of Deterministic Trends." *Journal of Econometrics*, **11**, pp. 323–353.

Phillips, P. C. B. 2005a. "Challenges of Trending Time Series Econometrics." *Mathematics and Computers in Simulation*, **68**(5–6), pp. 401–416.

Phillips, P. C. B. 2005b. "HAC Estimation By Automated Regression." *Econometric Theory*, **21**(1), pp. 116–142.

Phillips, P. C. B. 2007. "Unit Root Log Periodogram Regression." *Journal of Econometrics*, **138**, pp. 104–124.

Phillips, P. C. B. 2013. "Optimal Estimation of Cointegrated Systems with Irrelevant Instruments." *Journal of Econometrics*, forthcoming.

Phillips, P. C. B., and B. E. Hansen. 1990. "Statistical Inference in Instrumental Variables Regression with I(1) Processes." *Review of Economics Studies*, **57**, pp. 99–125.

Phillips, P. C. B., and P. Perron. 1988. "Testing for a unit root in time series regression." *Biometrika*, **75**, pp. 335–346.

Phillips, P. C. B., and V. Solo. 1992. "Asymptotics for Linear Processes," *Annals of Statistics*, **20**(2), pp. 971–1001.

Pugachev V. S. 1959. "A Method for Solving the Basic Integral Equation of Statistical Theory of Optimum Systems in Finite Form." *Prikl. Math. Mekh.*, **23**(3–14), English translation, pp. 1–16.

Staiger, D., and J. H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, **65**, pp. 557–586.

Shorack, G. R., and J. A. Wellner. 1986. *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons.

Stephens, M. A. 1976. "Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters." *Annals of Statistics*, **4**, pp. 357–369.

Sun, Y. and Kim, M. S. (2012), "Simple and Powerful GMM Over-identification Tests," *Journal of Econometrics*, 166(2), pp. 267–281.

Sun, Y., and M. S. Kim. 2013. "Asymptotic F Test in a GMM Framework with Cross Sectional Dependence." Working paper, Department of Economics, UC San Diego.

Sun, Y. 2011. "Robust Trend Inference with Series Variance Estimator and Testing-optimal Smoothing Parameter." *Journal of Econometrics*, **164**(2), pp. 345–366.

Sun, Y. (2013), "Heteroscedasticity and Auto-correlation Robust F Test Using Orthonormal Series Variance Estimator." *Econometrics Journal*, 16, pp. 1–26.

Watson, G. S. 1962. "Goodness-of-fit Tests on a circle II." *Biometrika*, **49**, pp. 57–63

Yule, U. 1926. "Why Do We Sometimes Get Nonsense-Correlations Between Time Series? A Study in Sampling and the Nature of Time Series." *Journal of the Royal Statistical Society*, **89**, pp. 1–69.

Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association*, **101**(476), pp. 1418–1429.

CHAPTER 13

IDENTIFICATION, ESTIMATION, AND SPECIFICATION IN A CLASS OF SEMILINEAR TIME SERIES MODELS

JITI GAO

13.1. Introduction

Consider a class of semilinear (semiparametric) time series models of the form

$$y_t = x_t^\tau \beta + g(x_t) + e_t, \qquad t = 1, 2, \ldots, n, \tag{13.1}$$

where $\{x_t\}$ is a vector of time series regressors, β is a vector of unknown parameters, $g(\cdot)$ is an unknown function defined on R^d, $\{e_t\}$ is a sequence of martingale differences, and n is the number of observations. This chapter mainly focuses on the case of $1 \leq d \leq 2$. As discussed in Section 13.2.2 below, for the case of $d \geq 3$, one may replace $g(x_t)$ by a semiparametric single-index form $g(x_t^\tau \beta)$.

Various semiparametric regression models have been proposed and discussed extensively in recent years. Primary interest focuses on general nonparametric and semiparametric time series models under stationarity assumption. Recent studies include Tong (1990), Fan and Gijbels (1996), Härdle, Liang, and Gao (2000), Fan and Yao (2003), Gao (2007), Li and Racine (2007), and Teräsvirta, Tjøstheim, and Granger (2010), as well as the references therein. Meanwhile, model estimation and selection as well as model specification problems have been discussed for one specific class of semiparametric regression models of the form

$$y_t = x_t^\tau \beta + \psi(v_t) + e_t, \tag{13.2}$$

where $\psi(\cdot)$ is an unknown function and $\{v_t\}$ is a vector of time series regressors such that $\Sigma = E\big[(x_t - E[x_t|v_t])(x_t - E[x_t|v_t])^\tau\big]$ is positive definite. As discussed in the literature (see, for example, Robinson (1988), Chapter 6 of Härdle, Liang, and Gao (2000), Gao (2007), and Li and Racine (2007)), a number of estimation and specification problems have already been studied for the case where both x_t and v_t are stationary and the covariance matrix Σ is positive definite. In recent years, attempts have also been made to address some estimation and specification testing problems for model (13.2) for the case where x_t and v_t may be stochastically nonstationary (see, for example, Juhl and Xiao (2005), Chen, Gao, and Li (2012), Gao and Phillips (2011)).

The focus of our discussion in this chapter is on model (13.1). Model (13.1) has different types of motivations and applications from the conventional semiparametric time series model of the form (13.2). In model (13.1), the linear component in many cases plays the leading role while the nonparametric component behaves like a type of unknown departure from the classic linear model. Since such departure is usually unknown, it is not unreasonable to treat $g(\cdot)$ as a nonparametrically unknown function. In recent literature, Glad (1998), Martins–Filho, Mishra and Ullah (2008), Fan, Wu, and Feng (2009), Mishra, Su, and Ullah (2010), Long, Su, and Ullah (2011), and others have discussed the issue of reducing estimation biases through using a potentially misspecified parametric form in the first step rather than simply nonparametrically estimating the conditional mean function $m(x) = E[y_t|x_t = x]$. By comparison, we are interested in such cases where the conditional mean function $m(x)$ may be approximated by a parametric function of the form $f(x, \beta)$. In this case, the remaining nonparametric component $g(x) = m(x) - f(x, \beta)$ may be treated as an unknown departure function in our discussion for both estimation and specification testing. In the case of model specification testing, we treat model (13.1) as an alternative when there is not enough evidence to suggest accepting a parametric true model of the form $y_t = x_t^\tau \beta + e_t$. In addition, model (13.1) will also be motivated as a model to address some endogenous problems involved in a class of linear models of the form $y_t = x_t^\tau \beta + \varepsilon_t$, where $\{\varepsilon_t\}$ is a sequence of errors with $E[\varepsilon_t] = 0$ but $E[\varepsilon_t|x_t] \neq 0$. In the process of estimating both β and $g(\cdot)$ consistently, existing methods, as discussed in the literature by Robinson (1988), Härdle, Liang, and Gao (2000), Gao (2007), and Li and Racine (2007) for example, are not valid and directly applicable because $\Sigma = E\big[(x_t - E[x_t|x_t])(x_t - E[x_t|x_t])^\tau\big] = 0$. The main contribution of this chapter is summarized as follows. We discuss some recent developments for the stationary time series case of model (13.1) in Section 13.2 below. Sections 13.3 and 13.4 establish some new theory for model (13.1) for the integrated time series case and a nonstationary autoregressive time series case, respectively. Section 13.5 discusses the general case where $y_t = f(x_t, \beta) + g(x_t) + e_t$.

The organization of this chapter is summarized as follows. Section 13.2 discusses model (13.1) for the case where $\{x_t\}$ is a vector of stationary time series regressors. Section 13.2 also proposes an alternative model to model (13.1) for the case where $d \geq 3$. The case where $\{x_t\}$ is a vector of nonstationary time series regressors is discussed in Section 13.3. Section 13.4 considers an autoregressive case of $d = 1$ and $x_t = y_{t-1}$

and then establishes some new theory. Section 13.5 discusses some extensions and then gives some examples to show why the proposed models are relevant and how to implement the proposed theory and estimation method in practice. This chapter concludes with some remarks in Section 13.6.

13.2. Stationary Models

Note that the symbol "$\Longrightarrow_{\mathcal{D}}$" denotes weak convergence, the symbol "$\to_D$" denotes convergence in distribution, and "$\to_P$" denotes convergence in probability.

In this section, we give some review about the development of model (13.1) for the case where $\{x_t\}$ is a vector of stationary time series regressors. Some identification and estimation issues are then reviewed and discussed. Section 13.2.1 discusses the case of $1 \le d \le 2$, while Section 13.2.2 suggests using both additive and single–index models to deal with the case of $d \ge 3$.

13.2.1. Case of $1 \le d \le 2$

While the literature may mainly focus on model (13.2), model (13.1) itself has its own motivations and applications. As a matter of fact, there is also a long history about the study of model (13.1). Owen (1991) considers model (13.1) for the case where $\{x_t\}$ is a vector of independent regressors and then treats $g(\cdot)$ as a misspecification error before an empirical likelihood estimation method is proposed. Gao (1992) systematically discusses model (13.1) for the case where $\{x_t\}$ is a vector of independent regressors and then considers both model estimation and specification issues. Before we start our discussion, we introduce an identifiability condition of the form in Assumption 13.1.

Assumption 13.1.

(i) Let $g(\cdot)$ be an integrable function such that $\int ||x||^i \big|g(x)\big|^i dF(x) < \infty$ for $i = 1, 2$ and $\int x g(x)\, dF(x) = 0$, where $F(x)$ is the cumulative distribution function of $\{x_t\}$ and $|| \cdot ||$ denotes the conventional Euclidean norm.

(ii) For any vector γ, $\min_\gamma E\big[g(x_1) - x_1^\tau \gamma\big]^2 > 0$.

Note that Assumption 13.1 implies the identifiability conditions. In addition, Assumption 13.1(ii) is imposed to exclude any cases where $g(x)$ is a linear function of x. Under Assumption 13.1, parameter β is identifiable and chosen such that

$$E\big[y_t - x_t^\tau \beta\big]^2 \text{ is minimized over } \beta, \tag{13.3}$$

which implies $\beta = \big(E\big[x_1 x_1^\tau\big]\big)^{-1} E\big[x_1 y_1\big]$, provided that the inverse matrix does exist. Note that the definition of $\beta = \big(E\big[x_1 x_1^\tau\big]\big)^{-1} E\big[x_1 y_1\big]$ implies $\int x g(x)\, dF(x) = 0$, and vice versa. As a consequence, β may be estimated by the ordinary least squares estimator of the form

$$\widehat{\beta} = \left(\sum_{t=1}^{n} x_t x_t^{\tau} \right)^{-1} \left(\sum_{t=1}^{n} x_t y_t \right). \tag{13.4}$$

Gao (1992) then establishes an asymptotic theory for $\widehat{\beta}$ and a nonparametric estimator of $g(\cdot)$ of the form

$$\widehat{g}(x) = \sum_{t=1}^{n} w_{nt}(x) \big(y_t - x_t^{\tau} \widehat{\beta} \big), \tag{13.5}$$

where $w_{nt}(x)$ is a probability weight function and is commonly chosen as $w_{nt}(x) = \frac{K\left(\frac{x_t - x}{h}\right)}{\sum_{s=1}^{n} K\left(\frac{x_s - x}{h}\right)}$, in which $K(\cdot)$ and h are the probability kernel function and the bandwidth parameter, respectively.

As a result of such an estimation procedure, one may be able to determine whether $g(\cdot)$ is small enough to be negligible. A further testing procedure may be used to test whether the null hypothesis $H_0: \ g(\cdot) = 0$ may not be rejected. Gao (1995) proposes a simple test and then shows that under H_0,

$$\widehat{L}_{1n} = \frac{\sqrt{n}}{\widehat{\sigma}_1} \left(\frac{1}{n} \sum_{t=1}^{n} \big(y_t - x_t^{\tau} \widehat{\beta} \big)^2 - \widehat{\sigma}_0^2 \right) \to_D N(0,1), \tag{13.6}$$

where $\widehat{\sigma}_1^2 = \frac{1}{n} \sum_{t=1}^{n} \big(y_t - x_t^{\tau} \widehat{\beta} \big)^4 - \widehat{\sigma}_0^4$ and $\widehat{\sigma}_0^2 = \frac{1}{n} \sum_{t=1}^{n} \big(y_t - x_t^{\tau} \widehat{\beta} \big)^2$ are consistent estimators of $\sigma_1^2 = E[e_1^4] - \sigma_0^4$ and $\sigma_0^2 = E[e_1^2]$, respectively.

In recent years, model (13.1) has been commonly used as a semiparametric alternative to a simple parametric linear model when there is not enough evidence to suggest accepting the simple linear model. In such cases, interest is mainly on establishing an asymptotic distribution of the test statistic under the null hypothesis. Alternative models are mainly used in small sample simulation studies when evaluating the power performance of the proposed test. There are some exceptions that further interest is in estimating the $g(\cdot)$ function involved before establishing a closed–form expression of the power function and then studying its large-sample and small-sample properties (see, for example, Gao (2007) and Gao and Gijbels (2008)). Even in such cases, estimation of $g(\cdot)$ becomes a secondary issue. Therefore, there has been no primary need to rigorously deal with such an estimation issue under suitable identifiability conditions similar to Assumption 13.1.

To state some general results for $\widehat{\beta}$ and $\widehat{g}(\cdot)$, we introduce the following conditions.

Assumption 13.2.

(i) Let (x_t, e_t) be a vector of stationary and α-mixing time series with mixing coefficient $\alpha(k)$ satisfying $\sum_{k=1}^{\infty} \alpha^{\frac{\delta}{2+\delta}}(k) < \infty$ for some $\delta > 0$, where $\delta > 0$ is chosen such that $E\big[|x_1 \varepsilon_1|^{2+\delta}\big] < \infty$, in which $\varepsilon_t = e_t + g(x_t)$.

(ii) Let $E[e_1|x_1] = 0$ and $E[e_1^2|x_1] = \sigma_e^2 < \infty$ almost surely. Let also $\Sigma_{11} = E\big[x_1 x_1^\tau\big]$ be a positive definite matrix.

(iii) Let $p(x)$ be the marginal density of x_1. The first derivative of $p(x)$ is continuous in x.

(iv) The probability kernel function $K(\cdot)$ is a continuous and symmetric function with compact support.

(v) The bandwidth h satisfies $\lim_{n\to\infty} h = 0$, $\lim_{n\to\infty} nh^d = \infty$, and $\limsup_{n\to\infty} nh^{d+4} < \infty$.

Assumption 13.2 is a set of conditions similar to what has been used in the literature (such as Gao (2007), Li and Racine (2007), and Gao and Gijbels (2008)). As a consequence, its suitability may be verified similarly.

We now state the following proposition.

Proposition 13.1.

(i) Let Assumptions 13.1 and 13.2 hold. Then as $n \to \infty$ we obtain

$$\sqrt{n}\big(\widehat{\beta} - \beta\big) \to_D N\Big(0, \Sigma_{1\varepsilon}\Sigma_{11}^{-2}\Big), \tag{13.7}$$

where $\Sigma_{1\varepsilon} = E\big[x_1 x_1^\tau \varepsilon_1^2\big] + 2\sum_{t=2}^{\infty} E\big[\varepsilon_1 \varepsilon_t x_1 x_t^\tau\big]$.

(ii) If, in addition, the first two derivatives of $g(x)$ are continuous, then we have as $n \to \infty$

$$\sqrt{nh^d}\big(\widehat{g}(x) - g(x) - c_n\big) \to_D N\Big(0, \sigma_g^2(x)\Big) \tag{13.8}$$

at such x that $p(x) > 0$, where $c_n = \frac{h^2(1+o(1))}{2}\left(g''(x) + \frac{2g'(x)p'(x)}{p(x)}\right)\int u^\tau uK(u)\, du$ and $\sigma_g^2(x) = \frac{\int K^2(u)\, du}{p(x)}$, in which $p(x)$ is the marginal density of x_1.

The proof of Proposition 13.1 is relatively straightforward using existing results for central limit theorems for partial sums of stationary and α-mixing time series (see, for example, Fan and Yao (2003)). Obviously, one may use a local-linear kernel weight function to replace $w_{nt}(x)$ in order to correct the bias term involved in c_n. Since such details are not essential to the primary interest of the discussion of this kind of problem, we omit such details here.

Furthermore, in a recent paper by Chen, Gao, and Li (2011), the authors consider an extended case of model (13.3) of the form

$$y_t = f(x_t^\tau \beta) + g(x_t) + e_t \qquad \text{with } x_t = \lambda_t + u_t, \tag{13.9}$$

where $f(\cdot)$ is parametrically known, $\{\lambda_t\}$ is an unknown deterministic function of t, and $\{u_t\}$ is a sequence of independent errors. In addition, $g(\cdot)$ is allowed to be a sequence of functions of the form $g_n(\cdot)$ in order to directly link model (13.9) with a sequence of local alternative functions under an alternative hypothesis as has been widely discussed in the literature (see, for example, Gao (2007) and Gao, and Gijbels

(2008)). By the way, the finite-sample results presented in Chen, Gao, and Li (2011) further confirm that the pair $(\widehat{\beta},\widehat{g}(\cdot))$ has better performance than a semiparametric weighted least squares (SWLS) estimation method proposed for model (13.2), since the so-called "SWLS" estimation method, as pointed out before, is not theoretically sound for model (13.1). Obviously, there are certain limitations with the paper by Chen, Gao, and Li (2011), and further discussion may be needed to fully take issues related to endogeneity and stationarity into account.

As also briefly mentioned in the introduction, model (13.1) may be motivated as a model to address a kind of "weak" endogenous problem. Consider a simple linear model of the form

$$y_t = x_t^\tau \beta + \varepsilon_t \qquad \text{with} \quad E[\varepsilon_t | x_t] \neq 0, \tag{13.10}$$

where $\{\varepsilon_t\}$ is a sequence of stationary errors.

Let $g(x) = E[\varepsilon_t | x_t = x]$. Since $\{\varepsilon_t\}$ is unobservable, it may not be unreasonable to assume that the functional form of $g(\cdot)$ is unknown. Meanwhile, empirical evidence broadly supports either full linearity or semilinearity. It is therefore that one may assume that $g(\cdot)$ satisfies Assumption 13.1. Let $e_t = \varepsilon_t - E[\varepsilon_t | x_t]$. In this case, model (13.10) can be rewritten as model (13.1) with $E[e_t | x_t] = 0$. In this case, $g(x_t)$ may be used as an 'instrumental variable' to address a 'weak' endogeneity problem involved in model (13.10). As a consequence, β can be consistently estimated by $\widehat{\beta}$ under Assumption 13.1 and the so–called "instrumental variable" $g(x_t)$ may be asymptotically 'found' by $\widehat{g}(x_t)$.

13.2.2. Case of $d \geq 3$

As discussed in the literature (see, for example, Chapter 7 of Fan and Gijbels (1996) and Chapter 2 of Gao (2007)), one may need to encounter the so–called "the curse of dimensionality" when estimating high-dimensional (with the dimensionality $d \geq 3$) functions. We therefore propose using a semiparametric single-index model of the form

$$y_t = x_t^\tau \beta + g\big(x_t^\tau \beta\big) + e_t \tag{13.11}$$

as an alternative to model (13.1). To be able to identify and estimate model (13.11), Assumption 13.1 will need to be modified as follows.

Assumption 13.3.

(i) Let $g(\cdot)$ be an integrable function such that $\int ||x||^i \big|g(x^\tau \beta_0)\big|^i dF(x) < \infty$ for $i = 1, 2$ and $\int x g(x^\tau \beta_0) dF(x) = 0$, where β_0 is the true value of β and $F(x)$ is the cumulative distribution function of $\{x_t\}$.

(ii) For any vector γ, $\min_\gamma E\big[g\big(x_1^\tau \beta_0\big) - x_1^\tau \gamma\big]^2 > 0$.

Under Assumption 13.2, β is identifiable and estimable by $\widehat{\beta}$. The conclusions of Proposition 13.1 still remain valid except the fact that $\widehat{g}(\cdot)$ is now modified as

$$\widehat{g}(u) = \frac{\sum_{t=1}^{n} K\left(\frac{x_t^\tau \widehat{\beta} - u}{h}\right) y_t}{\sum_{s=1}^{n} K\left(\frac{x_s^\tau \widehat{\beta} - u}{h}\right)}. \tag{13.12}$$

We think that model (13.11) is a feasible alternative to model (13.1), although there are some other alternatives. One of them is a semiparametric single-index model of the form

$$y_t = x_t^\tau \beta + g\left(x_t^\tau \gamma\right) + e_t, \tag{13.13}$$

where γ is another vector of unknown parameters. As discussed in Xia, Tong, and Li (1999), model (13.13) is a better alternative to model (13.2) than to model (13.1). Another of them is a semiparametric additive model of the form

$$y_t = x_t^\tau \beta + \sum_{j=1}^{d} g_j\left(x_{tj}\right) + e_t, \tag{13.14}$$

where each $g_j(\cdot)$ is an unknown and univariate function. In this case, Assumption 13.1 may be replaced by Assumption 13.4.

Assumption 13.4.

(i) Let each $g_j(\cdot)$ satisfy $\max_{1\le j\le d} \int ||x||^i \left|g_j\left(x_j\right)\right|^i dF(x) < \infty$ for $i = 1,2$ and $\sum_{j=1}^{d} \int x g_j\left(x_j\right) dF(x) = 0$, where each x_j is the jth component of $x = (x_1, \ldots, x_j, \ldots, x_d)^\tau$ and $F(x)$ is the cumulative distribution function of $\{x_t\}$.

(ii) For any vector γ, $\min_\gamma E\left[\sum_{j=1}^{d} g_j\left(x_{tj}\right) - x_t^\tau \gamma\right]^2 > 0$, where each x_{tj} is the jth component of $x_t = \left(x_{t1}, \ldots, x_{tj}, \ldots, x_{td}\right)^\tau$.

Under Assumption 13.4, β is still identifiable and estimable by $\widehat{\beta}$. The estimation of $\{g_j(\cdot)\}$, however, involves an additive estimation method, such as the marginal integration method discussed in Chapter 2 of Gao (2007). Under Assumptions 13.2 and 13.4 as well as some additional conditions, asymptotic properties may be established for the resulting estimators of $g_j(\cdot)$ in a way similar to Section 2.3 of Gao (2007).

We have so far discussed some issues for the case where $\{x_t\}$ is stationary. In order to establish an asymptotic theory in each individual case, various conditions may be imposed on the probabilistic structure $\{e_t\}$. Both our own experience and the literature show that it is relatively straightforward to establish an asymptotic theory for $\widehat{\beta}$ and $\widehat{g}(\cdot)$ under either the case where $\{e_t\}$ satisfies some martingale assumptions or the case where $\{e_t\}$ is a linear process. In Section 13.3 below, we provide some necessary conditions before we establish a new asymptotic theory for the case where $\{x_t\}$ is a sequence of nonstationary regressors.

13.3. Nonstationary Models

This section focuses on the case where $\{x_t\}$ is stochastically nonstationary. Since the paper by Chen, Gao, and Li (2011) already discusses the case where nonstationarity is driven by a deterministic trending component, this section focuses on the case where the nonstationarity of $\{x_t\}$ is driven by a stochastic trending component. Due to the limitation of existing theory, we only discuss the case of $d = 1$ in the nonstationary case.

Before our discussion, we introduce some necessary conditions.

Assumption 13.5.

(i) Let $g(\cdot)$ be a real function on $R^1 = (-\infty, \infty)$ such that $\int |x|^i \big|g(x)\big|^i dx < \infty$ for $i = 1, 2$ and $\int xg(x)\, dx \neq 0$.

(ii) In addition, let $g(\cdot)$ satisfy $\int \big|\int e^{ixy} yg(y)\, dy\big| dx < \infty$ when $\int xg(x)\, dx = 0$.

Note in the rest of this chapter that we refer to $g(\cdot)$ as a 'small' function if $g(\cdot)$ satisfies either Assumption 13.5(i), or, Assumption 13.4(ii), or Assumption 4.2(ii) below. In comparison with Assumption 13.1, there is no need to impose a condition similar to Assumption 13.1(ii), since Assumption 13.5 itself already excludes the case where $g(x)$ is a simple linear function of x.

In addition to Assumption 13.5, we will need the following conditions.

Assumption 13.6.

(i) Let $x_t = x_{t-1} + u_t$ with $x_0 = 0$ and $u_t = \sum_{i=0}^{\infty} \psi_i \eta_{t-i}$, where $\{\eta_t\}$ is a sequence of independent and identically distributed random errors with $E[\eta_1] = 0$, $0 < E[\eta_1^2] = \sigma_\eta^2 < \infty$ and $E\big[|\eta_1|^{4+\delta}\big] < \infty$ for some $\delta > 0$, in which $\{\psi_i : i \geq 0\}$ is a sequence of real numbers such that $\sum_{i=0}^{\infty} i^2 |\psi_i| < \infty$ and $\sum_{i=0}^{\infty} \psi_i \neq 0$. Let $\varphi(\cdot)$ be the characteristic function of η_1 satisfying $|r|\varphi(r) \to 0$ as $r \to \infty$.

(ii) Suppose that $\{(e_t, \mathcal{F}_t) : t \geq 1\}$ is a sequence of martingale differences satisfying $E[e_t^2 | \mathcal{F}_{t-1}] = \sigma_e^2 > 0$, a.s., and $E\big[e_t^4 | \mathcal{F}_{t-1}\big] < \infty$ a.s. for all $t \geq 1$. Let $\{x_t\}$ be adapted to $\mathcal{F}_{t-1}$ for $t = 1, 2, \ldots, n$.

(iii) Let $E_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} e_t$ and $U_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$. There is a vector Brownian motion (E, U) such that $(E_n(r), U_n(r)) \Longrightarrow_{\mathcal{D}} (E(r), U(r))$ on $D[0,1]^2$ as $n \to \infty$, where $\Longrightarrow_{\mathcal{D}}$ stands for the weak convergence.

(iv) The probability kernel function $K(\cdot)$ is a bounded and symmetric function. In addition, there is a real function $\Delta(x, y)$ such that, when h is small enough, $\big|g(x + hy) - g(x)\big| \leq h\ \Delta(x, y)$ for all y and $\int K(y)\Delta(x, y)\ dy < \infty$ for each given x.

(v) The bandwidth h satisfies $h \to 0$, $nh^2 \to \infty$ and $nh^6 \to 0$ as $n \to \infty$.

Similar sets of conditions have been used in Gao and Phillips (2011), Li et al. (2011), and Chen, Gao, and Li (2012). The verification and suitability of Assumption 13.6 may be given in a similar way to Remark A.1 of Appendix A of Li et al. (2011).

Since $\{x_t\}$ is nonstationary, we replace Eq. (13.3) by a sample version of the form

$$\frac{1}{n}\sum_{t=1}^{n}\left[y_t - x_t\beta\right]^2 \text{ is minimized over } \beta, \tag{13.15}$$

which implies $\widehat{\beta} = \left(\sum_{t=1}^{n} x_t^2\right)^{-1}\left(\sum_{t=1}^{n} x_t y_t\right)$ as has been given in Eq. (13.4). A simple expression implies

$$n\left(\widehat{\beta} - \beta\right) = \left(\frac{1}{n^2}\sum_{t=1}^{n} x_t^2\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} x_t e_t\right) + \left(\frac{1}{n^2}\sum_{t=1}^{n} x_t^2\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} x_t g(x_t)\right). \tag{13.16}$$

Straightforward derivations imply as $n \to \infty$

$$\frac{1}{n^2}\sum_{t=1}^{n} x_t^2 = \frac{1}{n}\sum_{t=1}^{n} x_{tn}^2 \Longrightarrow_{\mathcal{D}} \int_0^1 U^2(r)\,dr, \tag{13.17}$$

$$\frac{1}{n}\sum_{t=1}^{n} x_t e_t = \frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_{tn} e_t \Longrightarrow_{\mathcal{D}} \int_0^1 U(r)\,dE(r), \tag{13.18}$$

where $x_{tn} = \frac{x_t}{\sqrt{n}}$.

In view of Eq. (13.16)–(13.18), in order to establish an asymptotic distribution for $\widehat{\beta}$, it is expected to show that as $n \to \infty$ we have

$$\frac{1}{n}\sum_{t=1}^{n} x_t g(x_t) \to_P 0. \tag{13.19}$$

To be able to show (13.19), we need to consider the case of $\int xg(x)\,dx = 0$ and the case of $\int xg(x)\,dx \neq 0$ separately. In the case of $\int xg(x)\,dx \neq 0$, existing results (such as Theorem 2.1 of Wang and Phillips (2009)) imply as $n \to \infty$

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t g(x_t) = \frac{d_n}{n}\sum_{t=1}^{n}(d_n x_{tn})g(d_n\, x_{tn}) \to_D L_U(1,0)\cdot\int_{-\infty}^{\infty} zg(z)dz, \tag{13.20}$$

where $d_n = \sqrt{n}$ and $L_U(1,0)$ is the local-time process associated with $U(r)$. This then implies as $n \to \infty$

$$\frac{1}{n}\sum_{t=1}^{n} x_t g(x_t) = \frac{1}{\sqrt{n}}\cdot\frac{1}{\sqrt{n}}\sum_{t=1}^{n} x_t g(x_t) \to_P 0. \tag{13.21}$$

In the case of $\int xg(x)\,dx = 0$, existing results (such as Theorem 2.1 of Wang and Phillips (2011)) also imply as $n \to \infty$

$$\sqrt{\frac{1}{\sqrt{n}}}\sum_{t=1}^{n} x_t g(x_t) = \sqrt{\frac{d_n}{n}}\sum_{t=1}^{n}(d_n x_{tn})g(d_n\, x_{tn}) \to_D \sqrt{L_U(1,0)}\cdot N(0,1) \cdot\sqrt{\int_{-\infty}^{\infty} z^2 g^2(z)dz}, \tag{13.22}$$

where $N(0,1)$ is a standard normal random variable independent of $L_U(1,0)$. This shows that Eq. (13.19) is also valid for the case of $\int xg(x)dx = 0$.

We therefore summarize the above discussion into the following proposition.

Proposition 13.2.

(i) Let Assumptions 13.5 and 13.6(i)–(iii) hold. Then as $n \to \infty$ we have

$$n\left(\widehat{\beta} - \beta\right) \to_D \left(\int_0^1 U^2(r)\,dr\right)^{-1}\int_0^1 U(r)\,dE(r). \tag{13.23}$$

(ii) If, in addition, Assumption 13.6(iv),(v) holds, then as $n \to \infty$

$$\sqrt{\sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)}\left(\widehat{g}(x) - g(x)\right) \to_D N\left(0, \sigma_g^2\right), \tag{13.24}$$

where $\sigma_g^2 = \sigma_e^2 \int K^2(u)\,du$.

The proof of (13.23) follows from equations (13.16)–(13.22). To show (13.24), one may be seen that

$$\widehat{g}(x) - g(x) = \sum_{t=1}^{n} w_{nt}(x)e_t + \sum_{t=1}^{n} w_{nt}(x)\left(g(x_t) - g(x)\right) + \sum_{t=1}^{n} w_{nt}(x)x_t\left(\beta - \widehat{\beta}\right). \tag{13.25}$$

The first two terms may be dealt with in the same way as in existing studies (such as the proof of Theorem 3.1 of Wang and Phillips (2009)). To deal with the third term, one may have the following derivations:

$$\sum_{t=1}^{n} w_{nt}(x)x_t = h\cdot\frac{\sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)\left(\frac{x_t - x}{h}\right)}{\sum_{t=1}^{n} K\left(\frac{x_t - x}{h}\right)} + x = O_P(1) \tag{13.26}$$

by the fact that $\int uK(u)\,du = 0$ and an application of Theorem 2.1 of Wang and Phillips (2011). Equations (13.25) and (13.26), along with (13.23), complete the proof of (13.24).

Meanwhile, as in the stationary case, model (13.1) can also be considered as an alternative model to a simple linear model of the form $y_t = x_t^\tau \beta + e_t$ in the nonstationary case. A nonparametric test of the form

$$\widehat{L}_{2n} = \frac{\sum_{t=1}^{n} \sum_{s=1,\neq t}^{n} \widehat{e}_s \, K\left(\frac{x_t - x_s}{h}\right) \widehat{e}_t}{\sqrt{2 \sum_{t=1}^{n} \sum_{s=1,\neq t}^{n} \widehat{e}_s^2 \, K^2\left(\frac{x_t - x_s}{h}\right) \widehat{e}_t^2}} \tag{13.27}$$

has been proposed to test $H_0 : \ P\big(g(x_t) = 0\big) = 1$ and studied in recent years (see, for example, Gao et al. (2009a); Li et al. (2011), and Wang and Phillips (2012)), where $\widehat{e}_t = y_t - x_t^\tau \widehat{\beta}$, in which $\widehat{\beta}$ is the ordinary least squares estimator based on model (13.1) under H_0. Obviously, Assumption 13.1 is no longer needed for this kind of testing problem.

This section has so far considered the case where $\{x_t\}$ is an integrated time series. In Section 13.4, we consider an autoregressive version of model (13.1) and then discuss stationary and nonstationary cases separately.

13.4. Nonlinear Autoregressive Models

Consider an autoregressive version of model (13.1) of the form

$$y_t = x_t^\tau \beta + g(x_t) + e_t, \tag{13.28}$$

where $x_t = \big(y_{t-1}, \ldots, y_{t-d}\big)^\tau$, and the others are the same as before.

As has been discussed in the literature (see, for example, Tong 1990; Masry and Tjøstheim 1995; Chapter 6 of Härdle, Liang and Gao 2000), $\{y_t\}$ can be stochastically stationary and α-mixing when β satisfies Assumption 13.7(i) and $g(\cdot)$ satisfies Assumption 13.7(ii).

Assumption 13.7.

(i) Let $\beta = (\beta_1, \cdots, \beta_d)^\tau$ satisfy $y^d - \beta_1 y^{d-1} - \cdots - \beta_{d-1} y - \beta_d \neq 0$ for any $|y| \geq 1$.

(ii) Let $g(x)$ be bounded on any bounded Borel measurable set and satisfy $g(x) = o(||x||)$ as $||x|| \to \infty$.

(iii) Let $\{e_t\}$ be a sequence of independent and identically distributed continuous random errors with $E[e_1] = 0$ and $0 < E[e_1^2] = \sigma_e^2 < \infty$. Let $\{e_t\}$ and $\{y_s\}$ be independent for all $s < t$. In addition, the probability density, $p(x)$, of e_1 satisfies $\inf_{x \in C_p} p(x) > 0$ for all compact sets C_p.

Under Assumption 13.7, $\{y_t\}$ is stationary and α-mixing. Assumption 13.7(iii) is needed, since $\{y_t\}$ can still be null recurrent when $E[\log(1 + |e_t|)] = \infty$ (see, for example, Zeevi and Glynn (2004)). This, along with Assumption 13.1, implies that the estimation of β and $g(\cdot)$ may be done in the same way as in Section 13.2.1 for the case of $1 \leq d \leq 2$ and in Section 2.2 for the case of $d \geq 3$. Therefore, discussion of model (13.28) is relatively straightforward.

In the rest of this section, we then focus on the case where $\{y_t\}$ is nonstationary and discuss about how to estimate β and $g(\cdot)$ consistently. To present the main ideas of our discussion, we focus on the case of $d = 1$ to imply a semiparametric autoregressive model of the form

$$y_t = \beta y_{t-1} + g(y_{t-1}) + e_t. \tag{13.29}$$

While model (13.29) might look too simple, as discussed below, the study of the nonstationarity of $\{y_t\}$ may not be so easy at all. This is mainly because the nonstationarity may be driven by either the case of $\beta = 1$ or the case where the functional form of $g(\cdot)$ may be too 'explosive,' or a mixture of both. Our interest of this section is to focus on the case where $g(\cdot)$ is a 'small' departure function and the true value of β is $\beta = 1$. In a recent paper by Gao, Tjøstheim, and Yin (2012), the authors discuss a threshold version of model (13.29), in which $g(\cdot)$ is being treated as a conventional unknown function (not necessarily a "small" function) defined on a compact subset.

In a fashion similar to (13.15), we estimate β by minimizing

$$\frac{1}{n}\sum_{t=1}^{n}\left[y_t - y_{t-1}\beta\right]^2 \text{ over } \beta, \tag{13.30}$$

which implies $\widehat{\beta} = \left(\sum_{t=1}^{n} y_{t-1}^2\right)^{-1}\left(\sum_{t=1}^{n} y_{t-1}y_t\right)$. The unknown departure function $g(\cdot)$ can then be estimated by

$$\widehat{g}(y) = \sum_{t=1}^{n} w_{nt}(y)\left(y_t - \widehat{\beta}y_{t-1}\right) \qquad \text{with} \qquad w_{nt}(y) = \frac{K\left(\frac{y_{t-1}-y}{h}\right)}{\sum_{s=1}^{n} K\left(\frac{y_{s-1}-y}{h}\right)}. \tag{13.31}$$

When $\beta = 1$, we have

$$\widehat{\beta} - 1 = \left(\sum_{t=1}^{n} y_{t-1}^2\right)^{-1}\left(\sum_{t=1}^{n} y_{t-1}e_t\right) + \left(\sum_{t=1}^{n} y_{t-1}^2\right)^{-1}\left(\sum_{t=1}^{n} y_{t-1}g(y_{t-1})\right). \tag{13.32}$$

To establish an asymptotic distribution for $\widehat{\beta}$, we will need to understand the probabilistic structure of $\{y_t\}$. Obviously, $\{y_t\}$ is not integrated unless $g(\cdot) \equiv 0$. Thus, existing theory for the integrated time series case is not applicable here. We therefore impose some specific conditions on $g(\cdot)$ and $\{e_t\}$ to ensure that certain probabilistic structure can be deduced for $\{y_t\}$.

Assumption 13.8.

(i) Let Assumption 13.7(iii) hold.

(ii) Let $g(y)$ be twice differentiable and let the second derivative of $g(y)$ be continuous in $y \in R^1 = (-\infty, \infty)$. In addition, $\int \left|g(y)\right|^i \pi_s(dy) < \infty$ for $i = 1, 2$, where $\pi_s(\cdot)$ is the invariant measure of $\{y_t\}$.

(iii) Furthermore, $\int \left|yg(y)\right|^i \pi_s(dy) < \infty$ for $i = 1, 2$.

Assumption 13.8(i) is needed to show that $\{y_t\}$ can be a λ–null recurrent Markov chain with $\lambda = \frac{1}{2}$. Assumption 13.8(ii) is required to ensure that the functional form of $g(\cdot)$ is not too "explosive" in a fashion similar to Assumption 13.8(ii). If the functional form of $g(\cdot)$ is too "explosive" in this case, the nonstationarity of $\{y_t\}$ may be too strong to be controllable. Assumption 13.8(iii) imposes additional integrability conditions on $yg(y)$ in a way similar to Assumption 13.5(i) for the integrated case. Note that we need not require $\int yg(y)\pi_s(dy) = 0$ and then discuss this case specifically as in Section 13.3.

In order to establish an asymptotic theory for $(\widehat{\beta},\widehat{g}(\cdot))$, we need to introduce the following proposition.

Proposition 13.3. *Let Assumption 13.8(i), (ii) hold. Then $\{y_t\}$ is a λ–null recurrent Markov chain with $\lambda = \frac{1}{2}$.*

The proof of Proposition 13.3 follows similarly from that of Lemma 3.1 of Gao, Tjøstheim, and Yin (2013). More details about null recurrent Markov chains are available in Karlsen and Tjøstheim (2001) and Appendix A of Gao, Tjøstheim, and Yin (2013). Proposition 13.3 shows that $\{y_t\}$ is a nonstationary Markov chain, although it cannot be an integrated time series when $g(\cdot) \neq 0$. As a consequence, one may establish the following asymptotic theory in Proposition 13.4.

Proposition 13.4.

(i) *Let Assumption 13.8 hold. Then as $n \to \infty$ we obtain*

$$n\left(\widehat{\beta} - 1\right) \to_D \frac{\left(Q^2(1) - \sigma_e^2\right)}{2\int_0^1 Q^2(r)dr}, \tag{13.33}$$

where $Q(r) = \sigma_e B(r) + M_{\frac{1}{2}}(r)\mu_g$, in which $B(r)$ is the conventional Brownian motion, $M_{\frac{1}{2}}(t)$ is the Mittag–Leffler process as defined in Karlsen and Tjøstheim (2001, p. 388), and $\mu_g = \int g(y)\pi_s(dy)$.

(ii) *If, in addition, Assumption 13.6(iv),(v) holds, then as $n \to \infty$ we have*

$$\sqrt{\sum_{t=1}^{n} K\left(\frac{y_{t-1} - y}{h}\right)}\left(\widehat{g}(y) - g(y)\right) \to_D N(0,\sigma_g^2), \tag{13.34}$$

where $\sigma_g^2 = \sigma_e^2 \int K^2(u)du$.

The proof of Proposition 13.4 is given in Appendix A below. Note that Proposition 13.4 shows that the super rate-n of convergence is still achievable for $\widehat{\beta}$ even when $\{y_t\}$ is not an integrated time series. In addition, $Q(r) = \sigma_e B(r)$ when $\mu_g = 0$. In other words, $\widehat{\beta}$ retains the same asymptotic behavior as if $\{y_t\}$ were integrated when the 'small' departure function $g(\cdot)$ satisfies $\int g(y)\pi_s(dy) = 0$. Meanwhile, the asymptotic theory of $\widehat{g}(\cdot)$ remains the same as in the integrated case (see, for example, Proposition 13.2(ii)).

Remark 13.1.

(i) While Assumptions 13.7 and 13.8 are assumed respectively for the stationary and nonstationary cases, there are some common features in both assumptions. To present the main ideas in this discussion, we focus on the case of $d = 1$ in Assumption 13.7(i). When $|\beta| < 1$, Assumption 13.7(ii) basically requires that the rate of $g(y)$ decaying to infinity is slower than that of $|y| \to \infty$ in order to ensure that $\{y_t\}$ is stochastically stationary. In the case of $\beta = 1$, in addition to the 'smallness' condition in Assumption 13.8(iii), Assumption 13.8(ii) also imposes certain conditions on the rate of divergence of $g(\cdot)$ to deduce that $\{y_t\}$ is a nonstationary Markov chain, although, in the case of $g(\cdot) \neq 0$, $\{y_t\}$ is not an integrated time series. This is mainly because it may be impossible to study such nonlinear autoregressive models when $g(\cdot)$ behaves too "explosive."

(ii) $\{y_t\}$ could be generated recursively by a nonlinear autoregressive time series of the form $y_t = y_{t-1} + g(y_{t-1}) + e_t$ if $\beta = 1$ and $g(\cdot)$ were known. In the paper by Granger, Inoue, and Morin (1997), the authors propose some parametric specifications for $g(\cdot)$ and treat $g(\cdot)$ as a stochastic trending component. The authors then suggest estimating $g(\cdot)$ nonparametrically before checking whether $g(\cdot)$ is negligible. Gao et al. (2009b) further consider this model and propose a nonparametric unit–root test for testing $H_0: \ g(\cdot) = 0$. As pointed out above, what we have been concerned about in this section is to deal with the case where $g(\cdot)$ is not negligible, but is a "small" departure function satisfying Assumption 13.8(ii),(iii). Proposition 13.4 implies that model (13.29) may generate a class of null-recurrent time series models when $\beta = 1$ and $\int g(y)\pi_s(dy) = 0$. This may motivate us to further develop some general theory for such a class of time series models.

13.5. Extensions and Examples of Implementation

Since many practical problems (see, for example, Examples 13.1 and 13.2 below) may require the inclusion of a general polynomial function as the main mean function of y_t, model (13.1) may need to be extended to accommodate a general class of parametric functions. In this case, model (13.1) can be written as

$$y_t = f(x_t, \beta) + g(x_t) + e_t, \tag{13.35}$$

where $f(x, \beta)$ is a parametrically known function indexed by a vector of unknown parameters β. In the stationary case, Eq. (13.3) now becomes

$$E\big[y_t - f(x_t, \beta)\big]^2 \qquad \text{is minimized over } \beta. \tag{13.36}$$

In the integrated time series case, Eq. (13.2) can be replaced by minimising

$$\frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(x_t,\beta)\right]^2 \qquad \text{over } \beta, \tag{13.37}$$

which is similar to the discussion used in Park and Phillips (2001). Obviously, various other identifiability conditions imposed in Sections 13.2 and 13.3 can be modified straightforwardly. Thus, further discussion is omitted here.

In the autoregressive time series case, model (13.35) becomes

$$y_t = f(y_{t-1},\beta) + g(y_{t-1}) + e_t, \tag{13.38}$$

and Eq. (13.30) is now replaced by minimizing

$$\frac{1}{n}\sum_{t=1}^{n}\left[y_t - f(y_{t-1},\beta)\right]^2 \qquad \text{over } \beta. \tag{13.39}$$

In the threshold case where $g(y) = \psi(y)I\left[y \in C_\tau\right]$ and $f(y,\beta) = \beta y I\left[y \in D_\tau\right]$, in which $\psi(\cdot)$ is an unknown function, C_τ is a compact set indexed by parameter τ, and D_τ is the complement of C_τ, Gao, Tjøstheim, and Yin (2013) show that $\{y_t\}$ is a sequence of $\frac{1}{2}$-null recurrent Markov chains under Assumption 13.26(i),(ii). In general, further discussion about model (13.38) is needed and therefore left for future research.

Examples 13.1–13.3 show why the proposed models and estimation methods are relevant and how the proposed estimation methods may be implemented in practice.

Example 13.1. This data set consists of quarterly consumer price index (CPI) numbers of 11 classes of commodities for 8 Australian capital cities spanning from 1994 to 2008 (available from the Australian Bureau of Statistics at www.abs.gov.au). Figure 13.1 gives the scatter plots of the log food CPI and the log all–group CPI.

Figure 13.1 shows that either a simple linear trending function or a second–order polynomial form may be sufficient to capture the main trending behavior for each of the CPI data sets. Similarly, many other data sets available in climatology, economics, and finance also show that linearity remains the leading component of the trending component of the data under study. Figure 13.2 clearly shows that it is not unreasonable to assume a simple linear trending function for a disposable income data set (a quarter data set from the first quarter of 1960 to the last quarter of 2009 available from the Bureau of Economic Analysis at http://www.bea.gov).

The following example is the same as Example 5.2 of Li et al. (2011). We use this example to show that in some empirical models, a second–order polynomial model is more accurate than a simple linear model.

Example 13.2. In this example, we consider the 2–year (x_{1t}) and 30–year (x_{2t}) Australian government bonds, which represent short-term and long-term series in the

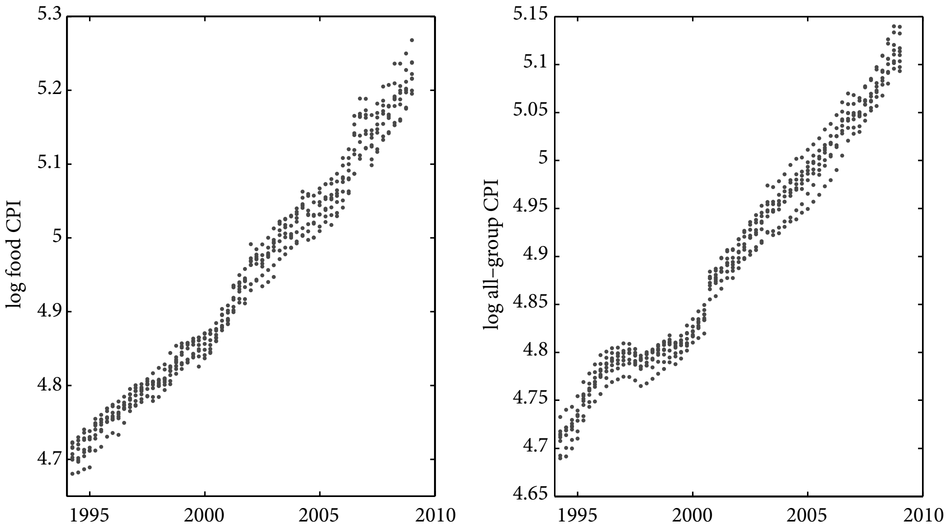


FIGURE 13.1 Scatter plots of the log food CPI and the log all–group CPI.

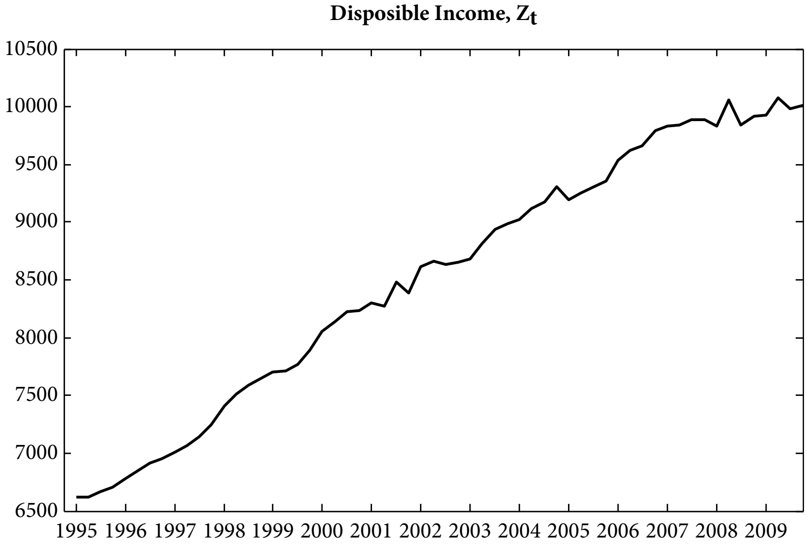


FIGURE 13.2 Plot of the disposable income data.

term structure of interest rates. Our aim is to analyze the relationship between the long-term data $\{x_{2t}\}$ and short-term data $\{x_{1t}\}$. We first apply the transformed versions defined by $y_t = \log(x_{2t})$ and $x_t = \log(x_{1t})$. The time frame of the study is during January 1971 to December 2000, with 360 observations for each of $\{y_t\}$ and $\{x_t\}$.

Consider the null hypothesis defined by

$$H_0:\ y_t = \alpha_0 + \beta_0 x_t + \gamma_0 x_t^2 + e_t, \tag{13.40}$$

where $\{e_t\}$ is an unobserved error process.

In case there is any departure from the second-order polynomial model, we propose using a nonparametric kernel estimate of the form

$$\widehat{g}(x) = \sum_{t=1}^{n} w_{nt}(x)\big(y_t - \widehat{\alpha}_0 - \widehat{\beta}_0 x_t - \widehat{\gamma}_0 x_t^2\big), \tag{13.41}$$

where $\widehat{\alpha}_0 = -0.2338$, $\widehat{\beta}_0 = 1.4446$, and $\widehat{\gamma}_0 = -0.1374$, and $\{w_{nt}(x)\}$ is as defined in (13.13), in which $K(x) = \frac{3}{4}(1 - x^2)I\{|x| \le 1\}$ and an optimal bandwidth $\widehat{h}_{\text{optimal}}$ is chosen by a cross-validation method.

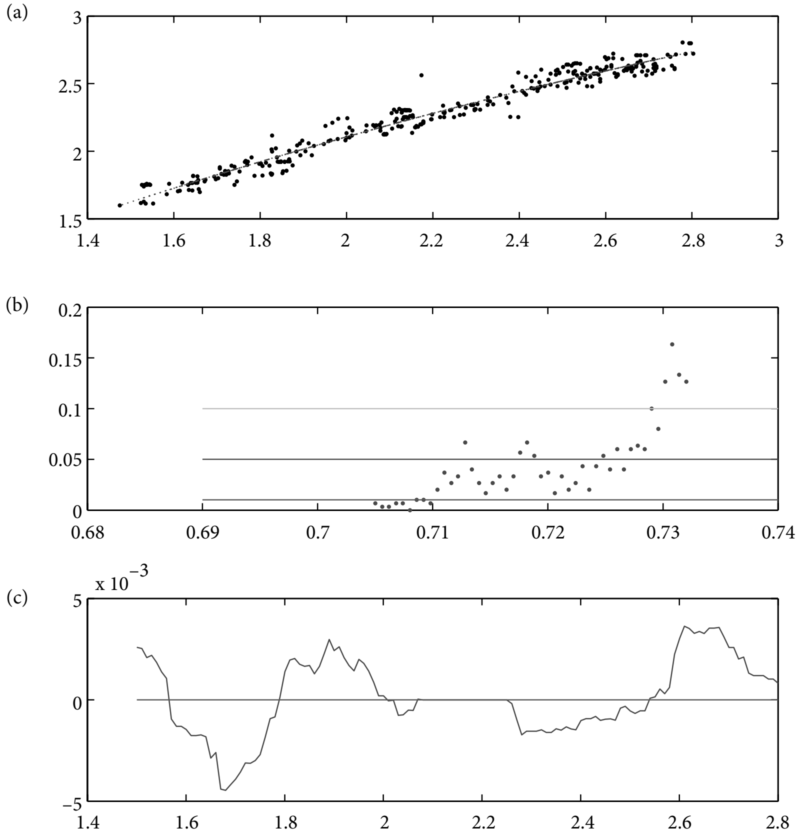


FIGURE 13.3 (a) Scatter chart of (y_t, x_t) and a nonparametric kernel regression plot $\widehat{y} = \widehat{m}(x)$; (b) p-values of the test for different bandwidths; and (c) plot of $\widehat{g}(x)$, whose values are between -5×10^{-3} and 5×10^{-3}.

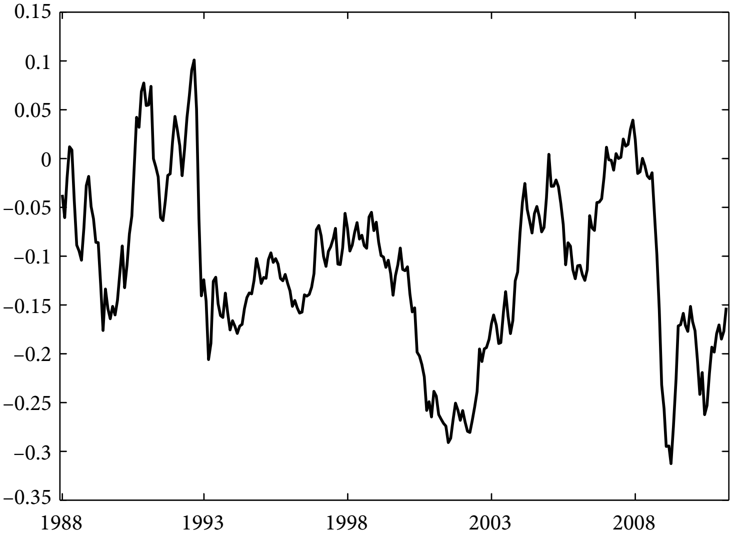


FIGURE 13.4 $y_t = \log(e_t) + \log(p_t^{UK}) - \log(p_t^{USA})$.

Figure 13.3 shows that the relationship between y_t and x_t may be approximately modeled by a second-order polynomial function of the form $y = -0.2338 + 1.4446\,x - 0.1374\,x^2$.

The following example is the same as Example 4.5 of Gao, Tjøstheim, and Yin (2013). We use it here to show that a parametric version of model (13.29) is a valid alternative to a conventional integrated time series model in this case.

Example 13.3. We look at the logarithm of the British pound/American dollar real exchange rate, y_t, defined as $\log(e_t) + \log(p_t^{UK}) - \log(p_t^{USA})$, where $\{e_t\}$ is the monthly average of the nominal exchange rate, and $\{p_t^i\}$ denotes the consumption price index of country i. These CPI data come from website: http://www.rateinflation.com/ and the exchange rate data are available at http://www.federalreserve.gov/, spanning from January 1988 to February 2011, with sample size $n = 278$.

Our estimation method suggests that $\{y_t\}$ approximately follows a threshold model of the form

$$y_t = y_{t-1} - 1.1249\,y_{t-1} I[|y_{t-1}| \leq 0.0134] + e_t. \tag{13.42}$$

Note that model (13.41) indicates that while $\{y_t\}$ does not necessarily follow an integrated time series model of the form $y_t = y_{t-1} + e_t$, $\{y_t\}$ behaves like a "nearly integrated" time series, because the nonlinear component $g(y) = -1.1249\ y\ I[|y| \leq 0.0134]$ is a 'small' departure function with an upper bound of 0.0150.

13.6. Conclusions and Discussion

This chapter has discussed a class of "nearly linear" models in Sections 13.1–13.4. Section 13.2 has summarized the history of model (13.1) and then explained why model (13.1) is important and has theory different from what has been commonly studied for model (13.2). Sections 13.3 and 13.4 have further explored such models to the nonstationary cases with the cointegrating case being discussed in Section 13.3 and the autoregressive case being discussed in Section 13.4. As shown in Sections 13.3 and 13.4, respectively, while the conventional "local-time" approach is applicable to establish the asymptotic theory in Proposition 13.2, one may need to develop the so-called "Markov chain" approach for the establishment of the asymptotic theory in Proposition 13.4.

As discussed in Remark 13.1, model (13.29) introduces a class of null-recurrent autoregressive time series models. Such a class of nonstationary models, along with a class of nonstationary threshold models proposed in Gao, Tjøstheim, and Yin (2013), may provide existing literature with two new classes of nonlinear nonstationary models as alternatives to the class of integrated time series models already commonly and popularly studied in the literature. It is hoped that such models proposed in (13.29) and Gao, Tjøstheim, and Yin (2013) along with the technologies developed could motivate us to develop some general classes of nonlinear and nonstationary autoregressive time series models.

Acknowledgments

The author would like to thank the Co–Editors, Liangjun Su in particular, and a referee for their constructive comments. The author also acknowledges benefits from collaborations and discussions with his co–authors in the last few years, particularly with Jia Chen, Maxwell King, Degui Li, Peter Phillips, Dag Tjøstheim and Jiying Yin. Thanks go to Biqing Cai and Hua Liang for some useful discussions. Thanks also go to the Australian Research Council for its financial support under the Discovery Grants Scheme under grant number DP1096374.

Appendix

In order to help the reader of this chapter, we introduce some necessary notation and useful lemmas for the proof of Proposition 13.4.

Let $\{y_t\}$ be a null-recurrent Markov chain. It is well known that for a Markov chain on a countable state space that has a point of recurrence, a sequence split by the regeneration times becomes independent and identically distributed (i.i.d.) by the Markov property (see, for example, Chung (1967)). For a general Markov process that does not have an obvious point of recurrence, as in Nummelin (1984), the Harris recurrence allows one to construct a split chain that decomposes the partial sum of the Markov process $\{y_t\}$ into blocks of i.i.d. parts and the negligible remaining parts.

Let z_t take only the values 0 and 1, and let $\{(y_t,\ z_t),\ t \geq 0\}$ be the split chain. Define

$$\tau_k = \begin{cases} \inf\{t \geq 0 :\ z_t = 1\}, & k = 0, \\ \inf\{t > \tau_{k-1} :\ z_t = 1\}, & k \geq 1, \end{cases} \tag{13.A1}$$

and denote the total number of regenerations in the time interval $[0,\ n]$ by $T(n)$, that is,

$$T(n) = \begin{cases} \max\{k :\ \tau_k \leq n\}, & \text{if } \tau_0 \leq n, \\ 0, & \text{otherwise.} \end{cases} \tag{13.A2}$$

Note that $T(n)$ plays a central role in the proof of Proposition 13.4 below. While $T(n)$ is not observable, it may be replaced by $\frac{T_C(n)}{\pi_s(\mathrm{I}_C)}$ (see, for example, Lemma 3.6 of Karlsen and Tjøstheim 2001), where $T_C(n) = \sum_{t=1}^{n} I\big[y_t \in C\big]$, C is a compact set, and $\mathbf{I}_C$ is the conventional indicator function. In addition, Lemma 3.2 of Karlsen and Tjøstheim (2001) and Theorem 2.1 of Wang and Phillips (2009) imply that $T(n)$ is asymptotically equivalent to $\sqrt{n}L_B(1,0)$, where $L_B(1,0) = \lim_{\delta\to 0} \frac{1}{2\delta} \int_0^1 I[|B(s)| < \delta]\ ds$ is the local-time process of the Brownian motion $B(r)$.

We are now ready to establish some useful lemmas before the proof of Proposition 13.4. The proofs of Lemmas 13.A.1 and 13.A.2 below follow similarly from those of Lemmas 2.2 and 2.3 of Gao, Tjøstheim, and Yin (2012), respectively.

Lemma 13.1. *Let Assumption 13.2(i),(ii) hold. Then as $n \to \infty$ we obtain*

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} e_t + \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} g(y_{t-1}) \Longrightarrow_D \sigma_e\ B(r) + M_{\frac{1}{2}}(r)\mu_g \equiv Q(r), \tag{13.A3}$$

where $M_{\frac{1}{2}}(r)$, μ_g, and $Q(r)$ are the same as defined in Proposition 13.4.

Lemma 13.2. *Let Assumption 13.8 hold. Then as $n \to \infty$ we obtain*

$$\frac{1}{T(n)} \sum_{t=1}^{n} y_{t-1} g(y_{t-1}) \to_P \int_{-\infty}^{\infty} y g(y) \pi_s(dy), \tag{13.A4}$$

$$\frac{1}{n^2} \sum_{t=1}^{n} y_{t-1}^2 \to_D \int_0^1 Q^2(r)\ dr, \tag{13.A5}$$

$$\frac{1}{n} \sum_{t=1}^{n} y_{t-1} e_t \to_D \frac{1}{2}\big(Q^2(1) - \sigma_e^2\big). \tag{13.A6}$$

Proof of Proposition 13.4. The proof of the first part of Proposition 13.4 follows from Lemma A.2 and

$$n(\widehat{\beta}-1) = \left(\frac{1}{n^2}\sum_{t=1}^{n} y_{t-1}^2\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} y_{t-1}e_t\right) + \left(\frac{1}{n^2}\sum_{t=1}^{n} y_{t-1}^2\right)^{-1}\left(\frac{T(n)}{n}\,\frac{1}{T(n)}\sum_{t=1}^{n} y_{t-1}g(y_{t-1})\right). \qquad (13.A7)$$

The proof of the second part of Proposition 13.4 follows similarly from that of Proposition 13.2(ii).

References

Chen, Jia, Jiti Gao, and Degui Li. 2011. "Estimation in Semiparametric Time Series Models (invited paper)." *Statistics and Its Interface*, **4**(4), pp. 243–252.

Chen, Jia, Jiti Gao, and Degui Li. 2012. "Estimation in Semiparametric Regression with Nonstationary Regressors." *Bernoulli*, **18**(2), pp. 678–702.

Chung, Kailai. 1967. *Markov Chains with Stationary Transition Probabilities*, second edition. New York: Springer-Verlag.

Fan, Jianqing, and Iréne Gijbels. 1996. *Local Polynomial Modeling and Its Applications*. London: Chapman & Hall.

Fan, Jianqing, Yichao Wu, and Yang Feng. 2009. "Local Quasi–likelihood with a Parametric Guide." *Annals of Statistics*, **37**(6), pp. 4153–4183.

Fan, Jianqing, and Qiwei Yao. 2003. *Nonlinear Time Series: Parametric and Nonparametric Methods*. New York: Springer.

Gao, Jiti. 1992. *Large Sample Theory in Semiparametric Regression*. Doctoral thesis at the Graduate School of the University of Science and Technology of China, Hefei, China.

Gao, Jiti. 1995. "Parametric Test in a Partially Linear Model." *Acta Mathematica Scientia* (English Edition), **15**(1), pp. 1–10.

Gao, Jiti. 2007. *Nonlinear Time Series: Semi- and Non-Parametric Methods*. London: Chapman & Hall/CRC.

Gao, Jiti, and Iréne Gijbels. 2008. "Bandwidth Selection in Nonparametric Kernel Testing." *Journal of the American Statistical Association*, **103**(484), pp. 1584–1594.

Gao, Jiti, Maxwell King, Zudi Lu, and Dag Tjøstheim. 2009a. "Nonparametric Specification Testing for Nonlinear Time Series with Nonstationarity." *Econometric Theory*, **25**(4), pp. 1869–1892.

Gao, Jiti, Maxwell King, Zudi Lu, and Dag Tjøstheim. 2009b. "Specification Testing in Nonlinear and Nonstationary Time Series Autoregression." *Annals of Statistics*, **37**(6), pp. 3893–3928.

Gao, Jiti, and Peter C. B. Phillips. 2011. "Semiparametric Estimation in Multivariate Nonstationary Time Series Models." Working paper at http://www.buseco.monash.edu.au/ebs/pubs/wpapers/2011/wp17-11.pdf.

Gao, Jiti, Dag Tjøstheim, and Jiying Yin. 2013. "Estimation in Threshold Autoregressive Models with a Stationary and a Unit Root Regime." *Journal of Econometrics* **172**(1), pp. 1–13.

Glad, Ingrid. 1998. "Parametrically Guided Nonparametric Regression." *Scandinavian Journal of Statistics*, **25**(4), pp. 649–668.

Granger, Clive, Tomoo Inoue, and Norman Morin. 1997. "Nonlinear Stochastic Trends." *Journal of Econometrics*, **81**(1), pp. 65–92.

Härdle, Wolfang, Hua Liang, and Jiti Gao. 2000. *Partially Linear Models.* Springer Series in Economics and Statistics. New York: Physica-Verlag.

Juhl, Ted, and Zhijie Xiao. 2005. "Partially Linear Models with Unit Roots." *Econometric Theory*, **21**(3), pp. 877–906.

Karlsen, Hans, and Dag Tjøstheim. 2001. "Nonparametric Estimation in Null Recurrent Time Series." *Annals of Statistics*, **29**(2), pp. 372–416.

Li, Degui, Jiti Gao, Jia Chen, and Zhengyan Lin. 2011. "Nonparametric Estimation and Specification Testing in Nonlinear and Nonstationary Time Series Models." Working paper available at http://www.jitigao.com/page1006.aspx.

Li, Qi, and Jeffrey Racine. 2007. *Nonparametric Econometrics: Theory and Practice.* Princeton, NJ: Princeton University Press.

Long, Xiangdong, Liangjun Su, and Aman Ullah. 2011. "Estimation and Forecasting of Dynamic Conditional Covariance: A Semiparametric Multivariate Model." *Journal of Business & Economic Statistics*, **29**(1), pp. 109–125.

Martins–Filho, Carlos, Santosh Mishra, and Aman Ullah. 2008. "A Class of Improved Parametrically Guided Nonparametric Regression Estimators." *Econometric Reviews*, **27**(4), pp. 542–573.

Masry, Elias, and Dag Tjøstheim. 1995. "Nonparametric Estimation and Identification of Nonlinear ARCH Time Series." *Econometric Theory* **11**(2), pp. 258–289.

Mishra, Santosh, Liangjun Su, and Aman Ullah. 2010. "Semiparametric Estimator of Time Series Conditional Variance." *Journal of Business & Economic Statistics*, **28**(2), pp. 256–274.

Nummelin, Esa. 1984. *General Irreducible Markov Chains and Non-negative Operators.* Cambridge: Cambridge University Press.

Owen, Art. 1991. "Empirical Likelihood for Linear Models." *Annals of Statistics*, **19**(6), pp. 1725–1747.

Park, Joon, and Peter C. B. Phillips. 2001. "Nonlinear Regressions with Integrated Time Series." *Econometrica*, **69**(1), pp. 117–162.

Robinson, Peter M. 1988. "Root–N–Consistent Semiparametric Regression." *Econometrica*, **56**(4), pp. 931–964.

Teräsvirta, Timo, Dag Tjøstheim, and Clive Granger. 2010. *Modelling Nonlinear Economic Time Series.* Oxford: Oxford University Press.

Tong, Howell. 1990. *Nonlinear Time Series: A Dynamical System Approach.* Oxford: Oxford University Press.

Wang, Qiying, and Peter C. B. Phillips. 2009. "Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression." *Econometric Theory*, **25**(3), pp. 710–738.

Wang, Qiying, and Peter C. B. Phillips. 2011. "Asymptotic Theory for Zero Energy Functionals with Nonparametric Regression Application." *Econometric Theory*, **27**(2), pp. 235–259.

Wang, Qiying, and Peter C. B. Phillips. 2012. "Specification Testing for Nonlinear Cointegrating Regression." *Annals of Statistics*, **40**(2), pp. 727–758.

Xia, Yingcun, Howell Tong, and Waikeung Li. 1999. "On Extended Partially Linear Single-Index Models. *Biometrika*, **86**(4), pp. 831–842.

Zeevi, Assf, and Peter Glynn 2004. "Recurrence Properties of Autoregressive Processes with Super-Heavy-Tailed Innovations." *Journal of Applied Probability*, **41**(3), pp. 639–653.

CHAPTER 14

NONPARAMETRIC AND SEMIPARAMETRIC ESTIMATION AND HYPOTHESIS TESTING WITH NONSTATIONARY TIME SERIES

YIGUO SUN AND QI LI

14.1. Introduction

Nonparametric and semiparametric estimation and hypothesis testing methods have been intensively studied for cross-sectional independent data and weakly dependent time series data (see, for example, Pagan and Ullah (1999), Gao (2007), and Li and Racine (2007)). However, many important macroeconomics and financial data are found to exhibit a stochastic trend and/or a deterministic trend, and the trend can be nonlinear in nature. For example, a univariate nonlinear growth model studied by Granger, Inoue, and Morin (1997) can be extended to a non-/semiparametric autoregressive model with nonstationary time series.

This chapter focuses on econometric modeling and conceptual issues only for nonstationary time series with stochastic trend. While a linear model may provide a decent approximation to a nonlinear model for weakly dependent data, the linearization can result in severely biased approximation to a nonlinear relation for the nonstationary data (e.g., Marmer (2008)). Park and Phillips (1999) derived limit results for nonlinearly transformed integrated time series whose sample average converges at different rates depending on the forms of nonlinear transformation functions. Therefore, it is utterly important to understand how to properly explore potential nonlinear relation when integrated time series are to be analyzed.

The rest of this chapter is organized as follows. Because I(0) and I(1) are concepts defined in the linear model framework, a nonlinear transformation of an I(1) variable could lose its I(1) meaning even though the transformed data continues to exhibit strong persistency. Therefore, Section 14.2 discusses extensions of the traditional textbook definition of I(0) process to general nonlinear time series data. Section 14.3 deals with parametric nonlinear models with nonstationary data. Section 14.4 covers nonparametric estimation and test results for models considered in Section 14.3 and Phillips' (2009) kernel analysis of a spurious nonparametric regression of nonstationary time series. Semiparametric extensions of cointegrating models of nonstationary time series are discussed in Section 14.5. Section 14.5 also includes the bandwidth selection via cross-validatory method and consistent estimation for semiparametric varying coefficient models with correlated but not cointegrated data. Section 14.6 presents some newly developed model specification tests for parametric functional forms with nonstationary data. In Section 14.7, we discuss the co-summability concept proposed by Berenguer-Rico and Gonzalo (2013), which is useful in explaining nonlinear co-movement of non-stationary time series data. We apply the co-summability concept to analyze the co-movement and co-summability of semiparametric functional coefficient models discussed in Section 14.5. We conclude the chapter in Section 14.8. For ease of reference, an Appendix section includes some useful technical tools developed for nonstationary time series analysis.

Throughout this chapter, we use $[a]$ to denote the integer part of $a > 0$, $W(\cdot)$ to denote a standard Brownian motion, $B(\cdot)$ to denote a general Brownian motion, "$\Rightarrow$" to denote the weak convergence on the Skorohod space $\mathcal{D}[0,1]^m$ for some integer $m \geq 1$, and "$\xrightarrow{d}$" and "$\xrightarrow{p}$" to denote the convergence in distribution and in probability, respectively. In addition, "Δ" is the lag operator such that $\Delta X_t = X_t - X_{t-1}$ for any t, and $I(A)$ is an indicator function taking a value of one if event A holds and zero otherwise. $M > 0$ is a generic constant that can take different values at different places. Finally, $X_n = O_e(a_n)$ denotes an exact probability order of a_n. It means that $X_n = O_p(a_n)$, but $X_n \neq o_p(a_n)$. In addition, a superscript T in A^T denotes the transpose of A.

14.2. Nonlinear Nonstationary Data

Many macroeconomic and finance variables exhibit some kind of growth trend—for example, CPI, real GDP, money supply, oil prices, stock price indexes, and so on. In the linear model framework, if level data show strong persistency and their first difference becomes an I(0) process, the level data series is called an I(1) process. ARIMA models are developed for univariate time series analysis, and linear cointegrating model is developed to explore stable relations among integrated time series; (see Brockwell and Davis (1991) and Hamilton (1994) for details). However, macroeconomic and finance theories usually suggest nonlinear relationships among aggregate level data (e.g., Chapter 2 of Teräsvirta et al. (2010)). For example, the weak form stock market efficiency hypothesis states that stock prices are nonpredictable given existing publicly available information. Statistically, the weak-form market efficiency hypothesis means that $E\big[g(P_t)|\mathcal{F}_{t-1}\big] = g(P_{t-1})$ holds true, where P_t is the stock price at time t and $\mathcal{F}_{t-1}$ contains all the publicly available information at the end of time $t-1$. In practice, both $g(x) = x$ and $g(x) = \ln x$ are popularly used. Can both $\ln(P_t)$ and P_t be an I(1) process defined in the linear framework?

The answer to the above question is as follows: A nonlinear transformation of an I(1) process in general will not be an I(1) process any more. In some cases, its order of integration may not be well-defined, and such an example will be provided below. In other cases, the process may become an I(0) process; see Nicolau (2002) for a bounded random walk process becoming a stationary process under certain conditions. In addition, even though a nonlinear transformed integrated time series keeps the persistency feature embedded in the original data, the first difference of the nonlinear transformed data may exhibit shorter memory than the level data but may not be an I(0) process as defined in the linear framework; an example is given in the next paragraph. This makes the traditional I(1) concept improper in labeling nonlinearly transformed integrated time series data, and Granger and Hallman (1991) and Granger (1995) are among early ones to address this issue. We will discuss the concept of co-summability (Berengner-Rico and Gonzalo, 2012) in Section 14.7, which can be used to describe co-movement of nonlinear nonstationary time series data.

Below, we first use a simple example to address why new concepts are needed for nonlinearly transformed integrated time series data. Let us consider a pure random walk process without drift. Specifically, we assume $X_t = X_{t-1} + u_t$ with $X_0 \equiv 0$ and $u_t \sim$ i.i.d. $\big(0,\sigma_u^2\big)$ and u_t is independent of all the past X_s' $(s < t)$. Define $Y_t = X_t^2$, then $\Delta Y_t = Y_t - Y_{t-1} = 2X_{t-1}u_t + u_t^2$. As $E(Y_t) = \sigma_u^2 t$, $\{Y_t\}$ is not a covariance stationary process. For any $t > s \geq 0$, simple calculations give $\mathrm{Corr}(Y_t, Y_s) = \sqrt{\big(s\kappa + 2s^2\sigma_u^4\big)/\big(t\kappa + 2t^2\sigma_u^4\big)}$ and $\mathrm{Var}(Y_t) = 2t^2\sigma_u^4 + t\kappa$, where $\kappa = E\big(u_t^4\big) - 3\sigma_u^4$. It implies that $\mathrm{Corr}\big(Y_t, Y_{t-h}\big) \approx (t-h)/t \to 1$ for a given finite $h > 0$ and a sufficiently large t. In addition, $\{X_t\}$, as an I(1) process, has $\mathrm{Corr}\big(X_t, X_{t-h}\big) \approx \sqrt{(t-h)/t} \to 1$ for a finite $h > 0$ and a sufficiently large t. Therefore, the autocorrelation functions are close to one for both $\{X_t\}$ and $\{Y_t\}$, although the correlation coefficients converge to one at faster speed for $\{Y_t\}$ than $\{X_t\}$. In other words, $\{Y_t\}$ exhibits stronger persistency than $\{X_t\}$. Now, if we take a first difference of $\{Y_t\}$, simple calculations give $E(\Delta Y_t) = \sigma_u^2$, $\mathrm{Cov}(\Delta Y_t, \Delta Y_s) = 0$ for any $t \neq s$, and $\mathrm{Var}(\Delta Y_t) = 2\sigma_u^4(2t-1) + \kappa$. Therefore, first differencing does completely remove the serial correlations in $\{\Delta Y_t\}$. Unlike $\{\Delta X_t\}$, $\{\Delta Y_t\}$ is not a covariance stationary process because its variance is explosive as time increases. In fact, it can be shown that the variance of $\Delta^d Y_t$ is of order t for a finite integer $d \geq 1$. Hence, $\{Y_t\}$ is not an $I(d)$ process for any finite value of d. This example shows that a nonlinear transformation of an I(1) variable may have different degree of persistency than the original I(1) series, and some new concepts are needed to extend the definition of $I(1)$ (or $I(d)$) processes in describing the co-movement of nonlinearly transformed nonstationary data.

The time series data analysis distinguishes itself from cross-sectional data analysis due to its temporal relation, which leads to extensive modeling to capture the temporal impact of a shock occurring to a variable today on future values of the variable. Roughly speaking, a shock occurring to an I(0) variable will gradually render its impact, while a shock occurring to an I(1) variable will have forever impact on the variable. Davidson (2009) listed five widely used I(0) concepts, which are defined from population point of view and emphasize on covariance stationarity, short memory, and finite variance features of an I(0) process. Among the five concepts, the most popularly used is Engle and Granger's (1987, p. 252) concept of an I(0) process, which should have a stationary, invertible ARMA representation. However, the ARMA model is a linear regression model, which is too restrictive to apply to nonlinear time series analysis. In practice, a pragmatic way of defining an I(0) process is based on the asymptotic behavior of its sample average (e.g., Stock (1994), Davidson (2009), and Müller (2008)), which is stated as follows.

Definition. A time series $\{X_t\}_{t=1}^{\infty}$ is I(0) if its partial sum process obeys the functional central limit theorem (or FCLT); that is,

$$S_T(r) = \omega_T^{-1} \sum_{t=1}^{[Tr]} (X_t - EX_t) \Rightarrow W(r) \qquad \text{for all } r \in [0,1], \tag{14.1}$$

where $W(\cdot)$ is a standard Brownian motion and $\omega_T^2 = \mathrm{Var}\left(\sum_{t=1}^{T} X_t\right)$.

It is easy to see that a trend stationary process satisfies (14.1) as well as a stationary ARMA process. It is well known that, without recognizing trending features, one may end up with a spurious regression when exploring even linear relations among trend stationary time series. Therefore, the $I(0)$ concept given above is too broad to differentiate a trend stationary process from a stationary ARMA process. Adding a requirement on finite $E(X_t)$ and finite $Var(X_t)$ to (14.1) may provide a more appropriate definition of an I(0) process, which excludes processes with unbounded means/variances.

To understand when $\{X_t\}_{t=1}^{\infty}$ is an I(0) process based on the above definition, one needs to know what (sufficient) conditions delivers the FCLT. Among plenty works of FCLTs under variant conditions, we just mention a few here. For example, the FCLT is derived by Hall (1979) for martingale difference processes, by Herndorf (1984) and Kuelbs and Philipp (1980) for strong mixing processes, by de Jong and Davidson (2000), McLeish (1975), and Wooldridge and White (1988) for near-epoch dependent (or NED) functions of mixing processes (the NED concept was first introduced to econometricians by Gallant and White (1988)), and by Phillips and Solo (1992) for linear processes with i.i.d. innovations. Because a stationary ARMA(p,q) process and a stationary linear process are a stationary α-mixing process under some conditions (e.g., Athreya and Pantula (1986) and Withers (1981)) and are a NED on an α-mixing process (Davidson, 2002), and an α-mixing (or a strong mixing) process is of course a near-epoch dependence function of an α-mixing process, one can enlarge the class of

I(0) processes to the class of near-epoch dependent functions of an α-mixing process (see Escribano and Mira (2002)).

Now, let us go back to the example given above, where Y_t is the square of a random walk process. We can see that even under the enlarged I(0) class, a $\{\Delta Y_t\}$ sequence is not an I(0) process, as $\omega_T^{-1}\sum_{t=1}^{[Tr]}\Delta y_t \Rightarrow \sqrt{2}\big(\int_0^r W(r)dW(r)+r/2\big)$, which is not a Brownian motion process. If $\{\Delta Y_t\}$ is not an I(0) process, neither is $\{Y_t\}$ an I(1) process. Consequently, enlarging the I(0) class will not solve the problem that $\{Y_t\}$ is not an I(1) process. Through this example, we observe that the nonlinear transformation of an I(1) process creates an urgent need to classify time series data properties.

As explained above, although the I(0) concept building upon NED on an underlying strong mixing process is invariant to nonlinear transformation according to White and Domowitz (1984, Lemma 2.1), this nonlinear invariance property is not passed on to the concept of I(1), because first differencing is a linear operator. The example given above shows that $\left\{Y_t : Y_t = X_t^2\right\}$ reserves the converging-to-one autocorrelations as the random walk process, $\{X_t\}$, does. So, the first thought is to use the length of memory to classify data. However, correlation coefficient is not an adequate concept even for stationary series, because nonlinear transformation of an I(0) variable could generate completely different autocorrelation functions than those from the original data. Granger and Hallman (1991) and Granger (1995) provided early efforts in extending the linear concepts of I(1) and I(0) to a general nonlinear framework and introduced the concepts of short memory in distribution (SMD) versus extended memory in distribution (EMD) from forecasting point of view, where an I(0) process is SMD, but a SMD process may not be I(0); an I(1) process is EMD, but an EMD process may not be I(1). Granger and Hallman (1991) and Granger (1995) also provided the concepts of short memory in mean (SMM) versus extended memory in mean (EMM).

An alternative concept of nonlinear nonstationarity (or nonlinear persistency) has been introduced by Karlsen, and Tjøstheim (2001) via the class of null recurrent Markov chains. A null recurrent Markov chains is nonstationary, and Myklebust et al. (2011, Theorem 2, p. 10) showed that the irreducible (β-) null recurrence is invariance with respect to measurable one-to-one functional transformations, where the parameter β can be interpreted as the expected number of times that a Markov process $\{X_t\}$ visits a small set, say C. Or mathematically, $E_\lambda\left[\sum_{t=1}^T I(X_t \in C)\right] = T^\beta L(T)(1+o(1))$ for $\beta \in (0,1)$, where $L(x)$ is a positive, slowing varying function defined over $[a,\infty)$ for $a>0$ such that

$$\lim_{x\to\infty} L(kx)/L(x) = 1 \qquad (14.2)$$

for all $k>0$. Detailed concept of the β-null recurrence can be found in Karlsen and Tjøstheim (2001, p. 380). Kallianpur and Robbins (1954) showed that a random walk process is β-null recurrent with $\beta = 1/2$, while Myklebust et al. (2011) extended this result to more general $ARIMA(p,1,0)$ processes (with a finite p). Therefore, the concept of the β-null recurrence consolidates linear I(1) and nonlinear nonstationarity within the Markov chain framework, but it fails to consolidate nonlinearity with I(2) and general cointegrating relations as Myklebust et al. (2011) showed that linear I(2) processes and some linear cointegrating models are not null recurrent.

Instead of conceptually extending the I(1) concept into nonlinear time series models, Park and Phillips (1999), on the other hand, derived asymptotic sampling results for sample averages of nonlinear transformed I(1) time series with zero mean via the concept of local time of a Brownian motion process. By Corollary 1.9 in Revuz and Yor (2005, p. 227), a Brownian motion process, B, as a continuous martingale process, has a well-defined local time process

$$L_B(t,s) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t I\{|B(r) - s| < \epsilon\} dr \tag{14.3}$$

for every $s \in R$ and $t \in [0,1]$. Roughly speaking, $L_B(t,s)$ measures the time that the Brownian process B spends in the vicinity of s over the interval $[0,t]$. Under some conditions, Park and Phillips (1999) showed that both the convergence rate and the limiting distribution of $\sum_{t=1}^{[Tr]} g(X_t)$ depend on the functional form of g. Consider that $X_{[Tr]} \Rightarrow B(r)$, a Brownian motion, by some FCLT. Then, $T^{-3/2} \sum_{t=1}^{T} X_t \xrightarrow{d} \int_0^1 B(s)ds$, $T^{-2} \sum_{t=1}^{T} X_t^2 \xrightarrow{d} \int_0^1 B^2(s)ds = \int_{-\infty}^{\infty} s^2 L_B(1,s)ds$ by the occupation times formula, and $T^{-1/2} \sum_{t=1}^{T} g(X_t) \xrightarrow{d} \big(\int_{-\infty}^{\infty} g(s)ds\big) L_B(1,0)$ if g is integrable and satisfies some Lipschitz condition. Hence, a nonlinear transformation of an I(1) variable does not share the same sampling properties as the original I(1) variable.

Finally, Zeevi and Glynn (2004) showed that an AR(1) process generated from $X_t = \rho X_{t-1} + u_t$ $(t = 1, 2, \ldots, T)$ can be null recurrent or explosive if $\{u_t\}$ is a zero-mean i.i.d. innovation sequence with $E\big[\log(1 + |u_t|)\big] = \infty$ even if $|\rho| < 1$. It follows logically that a stable ARMA(p,q) structural with infinite variance arising from an ARCH-type error term may actually be nonstationary. However, we will not go further to include the literature on (G)ARCH models with infinite variance because the current literature focuses mainly on the extension of nonlinear modeling to conditional mean regression models of nonstationarity time series, which are nonlinear transformation of I(1) variables with finite increment variance. We close this section by emphasizing that there does not exist *satisfactory* extension of I(d) ($d = 0, 1, \ldots$) processes to nonlinear time series data. The concept of co-summability proposed by Berengner-Rico and Gonzalo (2012) (see Section 14.7) provides a useful step toward generalizing I(d) concepts to describe nonlinear time series data.

14.3. Nonlinear Econometrics Models with Nonstationary Data

In this section we discuss cointegration models with nonlinear nonstationary data. By nonlinear cointegrating models, we mean that error terms exhibit less persistency than dependent variable or the sample averages of the first few moments of the error terms are dominated by those of the (nonlinear) regressors. In a general sense, a nonlinear cointegrating model has to be balanced such that its dependent variable in the left-hand side of the model has the same persistency or the degree of nonstationarity as the (nonlinearly) transformed regressors at the right-hand side of the model, and the error term is I(0) using the general definition given in Section 14.2.

14.3.1. Univariate Nonlinear Modeling

We start with nonlinear modeling of univariate time series. For stationary time series, nonlinearity has been successfully built into linear stationary ARMA models, including self-exciting threshold autoregressive (SETAR) models (e.g., Tong and Lim (1980), Tong (1990), and Chan (1993)), smooth transition autoregressive (STAR) models (van Dijk, Teräsvirta, and Franses (2002) and references therein), bilinear models (e.g., Granger and Anderson, (1978) and Subba Rao (1981)), and functional coefficient autoregressive (FAR) models (Nicholls and Quinn, 1982; Chen and Tsay, 1993). Further reference can also be found in Fan and Yao (2003) and Teräsvirta, Tjøstheim, and Granger (2010). This section contains several popularly used nonlinear dynamic models of nonstationary data including I(1) and nonlinear transformation of an I(1) process. We recognize that the literature has been growing across time, partially due to space restriction, the current chapter is limited in its selection of existing models into this review. We also include some results on whether time series generated from some nonlinear autoregressive models can be characterized as I(0) processes or not.

We first consider a nonlinear dynamic model given by

$$Y_t = g(Y_{t-1}) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.4}$$

where $\{u_t\}$ is a stationary process with a zero mean and at least a finite second moment. If there exists a non-negative test function $V(\cdot)$ satisfying a Foster–Lyapunov drift criterion,

$$E\big[V(Y_t)|Y_{t-1} = y\big] = E\big[V\big(g\big(y,\theta\big) + u_t\big)\big] < V\big(y\big) \tag{14.5}$$

for large values of $\big|y\big|$, an initial distribution of Y_0 can be found such that $\{Y_t\}$ is (strictly) stationary; for details, see Meyn and Tweedie (1993). Tjøstheim (1990, Theorem 4.1) showed that $\{Y_t\}$ is null recurrent under the conditions that g is bounded over compact sets and $\big|g(x)\big| - |x| \to 0$ as $|x| \to \infty$, which implies an I(1) feature for extremely large realization of Y's. It follows that

$$Y_t = m(Y_{t-1},\theta)I(|Y_{t-1}| < c) + \alpha Y_{t-1}I(|Y_{t-1}| \geq c) + u_t, \qquad \text{with } \alpha = 1, \tag{14.6}$$

generates a null recurrent process for a constant $c > 0$ if $\sup_{|y|\leq c}\big|m\big(y,\theta\big)\big| \leq M < \infty$, where Y_t enters into a regime of I(1) process if $|Y_{t-1}| \geq c$ and switches to a bounded process when $|Y_{t-1}| < c$. If we replace Y_{t-1} by Y_{t-m} for some finite $m \geq 1$, model

(14.4) nests some SETAR and STAR models with delay as special cases. Cline and Pu (1999) studied stability and instability of nonlinear AR(1) time series with delay, which includes both SETAR(1) and FAR(1) processes as special cases.

If $\{Y_t\}$ is generated from a SETAR(1) model, then we have

$$Y_t = (\gamma + \alpha_1 Y_{t-1}) I(Y_{t-1} > \lambda) + (\delta + \alpha_2 Y_{t-1}) I(Y_{t-1} \leq \lambda) + u_t, \qquad t = 1,2,\ldots,T, \tag{14.7}$$

where $u_t \sim$ i.i.d. $\big(0,\sigma^2\big)$, and $\gamma,\alpha_1,\delta,\alpha_2$, and λ are unknown parameters to be estimated. The above model is stationary and geometrically ergodic if and only if $\alpha_1 < 1$, $\alpha_2 < 1$, and $\alpha_1\alpha_2 < 1$, is transient if (α_1,α_2) lies outside this region, and is an I(1) process with break in intercept term if $\alpha_1 = \alpha_2 = 1$. For stationary $\{Y_t\}$, the strong consistency result of the least squares estimator (LSE) of model (14.7) has been shown by Chan (1993) and Chan and Tsay (1998) for the continuous case (i.e., $\gamma + \alpha_1\lambda = \delta + \alpha_2\lambda$). When $\gamma = \delta = \lambda = 0$ is imposed, Pham, Chan, and Tong (1991) derived the strong consistency result of the LSE for more general cases where $\{Y_t\}$ can be stationary or nonstationary. Liu, Ling, and Shao (2011) showed that the LSE of α_1 is T-consistent if $\alpha_1 = 1, \alpha_2 < 1$, and $\lambda \leq 0$ and that of α_2 is T-consistent if $\alpha_1 < 1, \alpha_2 = 1$, and $\lambda \geq 0$, and they also derived the limiting distribution results. Setting $\gamma = \delta = 0$ and assuming the true parameter value of α_2 equal to 1, Gao, Tjøstheim, and Yin (2011) showed that $\{Y_t\}$ is a β-null recurrent Markov chains with $\beta = 1/2$ and derived the limiting distribution of the LSE estimator of (α_1,α_2) when λ is known, where the LSE of α_2 is T-consistent but that of α_1 is $T^{1/4}$-consistent. Moreover, under certain conditions, Gao, Tjøstheim, and Yin (2011) showed that model (14.6) with $m(Y_{t-1},\theta)$ replaced by an unknown smooth function $m(Y_{t-1})$ is also a β-null recurrent Markov chains with $\beta = 1/2$ and derived the limiting results of the kernel estimator of $m\big(y\big)$ and of the LSE estimator of α.

Caner and Hansen (2001) considered a two-regime TAR(p) model with an autoregressive unit root

$$\Delta Y_t = \theta_1^T X_{t-1} I(Z_{t-1} < \lambda) + \theta_2^T X_{t-1} I(Z_{t-1} \geq \lambda) + u_t, \qquad t = 1,2,\ldots,T,$$

where $X_{t-1} = (Y_{t-1} \quad 1 \quad \Delta Y_{t-1} \quad \ldots \quad \Delta Y_{t-p})^T$, $Z_t = Y_t - Y_{t-m}$ for some finite $m \geq 1$, and $u_t \sim$ i.i.d.$\big(0,\sigma^2\big)$. Caner and Hansen (2001) derived the limiting distribution of (a) a sup-Wald test statistic to test for the existence of a threshold when $\{Y_t\}$ is an I(1) process and (b) a unit root test to test for a unit root null aganist a stationary alternative and against a one-regime unit root alternative when there is no threshold and when there is a threshold.

The growth models considered by Granger et al. (1997) have $g(Y_{t-1}) = Y_{t-1} + m(Y_{t-1})$, which gives

$$Y_t = Y_{t-1} + m(Y_{t-1}) + u_t, \qquad t = 1,2,\ldots,T. \tag{14.8}$$

In particular, Granger et al. (1997) considered a model of $Y_t = Y_{t-1} + \alpha Y_{t-1}^{\gamma} + u_t\sqrt{\delta Y_{t-1}^{\beta}}$ with $u_t \sim$ i.i.d. $N(0,1)$ and $\gamma < 1$, where $\{Y_t\}$ exhibits positive growth by Theorem 2 in Granger et al. (1997) if $\beta = 1+\gamma$ and $\delta < 2\alpha$ or if $\beta < 1+\gamma$.

Granger and Swanson (1997) studied the stability and instability conditions of a stochastic unit-root process generated by $Y_t = a_t Y_{t-1} + u_t$, t = 1,2,...,T, where u_t is I(0) and $a_t = \exp(\alpha_t)$ is a random function of an exogenous stationary AR(1) process, α_t.

14.3.2. Nonlinear Error Correction Models and Nonlinear Cointegrating Models

In this subsection, we review some nonlinear error correction (NEC) models and nonlinear cointegrating models studied in the nonlinear modeling literature. Teräsvirta, Tjøstheim, and Granger (2010, Chapter 11) and Dufrénot and Mignon (2002) are good references to start with.

NEC models introduce nonlinearity into a linear cointegrating model by allowing the speed of adjustment parameters to be a function of a stationary linear combination of the I(1) variables appearing in the model. Specifically, Escribano and Mira (2002) defined a NEC model by

$$\Delta Y_t = \Psi_1 \Delta Y_{t-1} + g(Y_{t-1}) + u_t, \qquad t = 1,2,\ldots,T, \tag{14.9}$$

where Y_t and u_t are both $d\times 1$ vectors, and $\{u_t\}$ is a stationary strong mixing sequence with a zero mean and finite variance. Under some conditions including $g(Y_{t-1}) \equiv m\big(\beta^T Y_{t-1}\big)$ for some $\beta \in \Theta \subset R^d$, Escribano and Mira (2002) showed that $\{\Delta Y_t\}$ and $\{u_t\}$ are simultaneously near-epoch dependence of the strong mixing sequences $\big\{\big(u_t^T, \beta^T u_t\big)\big\}$. Therefore, $\{Y_t\}$ is an I(1) process if we extend the linear I(0) concept to a NED on strong mixing sequences as discussed in Section 14.2.

Saikkonen (2005) derived the stability condition of $\{\Delta Y_t\}$ $(d\times 1)$ for a more general NEC model,

$$\begin{aligned}\Delta Y_t = \alpha\beta^T Y_{t-1} + \sum_{j=1}^{p}\Gamma_j \Delta Y_{t-j} + \sum_{i=1}^{k} I(\phi(Z_{t-1},\eta_t) \in R_s)\\ \times \big[g_s(Z_{t-1}) + H_s(Z_{t-1})^{1/2}u_t\big], \qquad t = 1,2,\ldots,T,\end{aligned} \tag{14.10}$$

where $\alpha(d\times r)$, $\beta(d\times r)$, and $\Gamma_j(d\times d)$ are parameter matrices with α and β having full column rank $r \le d-1$, $Z_t = \begin{bmatrix} \big(\beta^T Y_t\big)^T & \Delta Y_t^T & \cdots & \Delta Y_{t-p+2}^T \end{bmatrix}^T$, $g_s : R^{d(p-1)+r} \to R^d$ and H_s: $R^{d(p-1)+r} \to R^{d\times d}$ are nonlinear functions, and $\phi : R^{d(p-1)+r+l} \to R^d$ is a (Borel) measurable function with the sets $R_s \subseteq R^d$ from a partition of R^d. Also, Y_{t-j} is independent of the innovations (η_t, u_t). Evidently, model (14.10) nests Escribano and Mira's (2002) NEC model, Balke and Fomby's (1997) three-regime threshold cointegrating models, and a nonlinear smooth NEC model of Kapetanios, Shin, and Snell (2006). Connecting to Tjøstheim (1990, Theorem 4.1) mentioned in Section 14.1, one necessary condition required for $\{Y_t\}$ to be an I(1) process is that $\left\|g_s(x)\right\| = O(\|x\|)$ for large $\|x\|$, where $\|\cdot\|$ is the Euclidean norm. See Saikkonen (2005) for details. Kapetanios, Shin, and Snell (2006) constructed an F-type test statistic to test a spurious null against a globally stationary smooth transition autoregressive cointegrating alternative model, where their test is based on a special case of model (14.10) with the last term of model (14.10) replaced by $g\left(\beta^T Y_{t-1}\right) + u_t$.

By imposing proper conditions on the functional form of g, the NEC models cited above introduce nonlinearity to linear cointegrating models by allowing short-term nonlinear dynamics of the speed of adjustment parameters without changing the I(1) properties of the nonstationary processes involved. Consequently, these models do not impose a challenge to the concept of linear cointegrating relations. The real challenge comes from specifying and building suitable nonlinear models to describe nonlinear long-run stable relationships among nonstationary time series—for example, nonlinear cost and production functions in macroeconomics. In the past two decades, research interests in this area have been gradually building up; see, for example, Chang, Park, and Phillips (2001), Granger and Hallman (1991), Granger (1995), Park and Phillips (1999, 2001), and Saikkonen and Choi (2004), among many others. Essentially, if both $\{Y_t\}$ and $\{X_t\}$ are nonstationary time series exhibiting strong persistence, and there exists a nonlinear transformation function $g : R \times R^d \to R$ such that $u_t = g(Y_t, X_t)$ is stationary or captures dominated or subdued stochastic component of Y_t and X_t, for easy reference, we call such models *nonlinear cointegrating models*.

Because no consensus has been reached in extending the linear I(0) and I(1) and cointegrating concepts to the nonlinear framework, the specification, estimation, and hypothesis testing of nonlinear cointegrating models developed in the literature has shown variations. In this chapter, we focus on three approaches corresponding to our explanation given in Section 14.2. The first approach follows from Granger and Hallman's (1991) and Granger's (1995) classification of short and long (or extended) memory in mean or distribution, which is originated from forecasting point of view. Specifically, a time series with a short or long (or extended) memory in mean or distribution is classified by whether remote shocks have a persistent influence on the level forecasts of the time series. The second approach is based on the null recurrence of Markov chains studied by Karlsen and Tjøstheim (2001). Based on the traditional definitions of I(0) and I(1), the third approach is carried on via sampling theories developed for nonlinear transformations of the traditional I(1) series as popularized by Park and Phillips (1999, 2001). No matter which approach is applied, they all share the same essence in building a nonlinear cointegrating model; that is, more than one variable share one or more common *stochastic trends*; and in the long run, these variables reach jointly to one or more nonlinear equilibrium relationships. For example, both $\{X_t\}$ and $\{Y_t\}$ exhibit persistent stochastic trend, and a nonlinear equilibrium relationship could be the one that $m(X_t, Y_t) = 0$ holds for some nonlinear function m, while the short-term dynamics could be described by $u_t = m(X_t, Y_t)$ for all t, a zero-mean mean reverting process with a finite variance and short memory, for example.

Granger and Hallman (1991) proposed to generate a sequence of extended memory in mean (EMM) time series as a monotonic nondecreasing function of the traditionally defined I(1) process plus a zero mean short memory in mean (SMM) process, and they considered the following nonlinear cointegrating model:

$$h(Y_t) = g(X_t) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.11}$$

where u_t is a zero mean SMM with a finite variance. Because an I(1) process is EMM, and an I(0) process is SMM, Breitung's (2001) rank-based cointegrating test works for a special case of model (14.11) when $h(Y_t)$ and $g(X_t)$ both are I(1) processes, and $h : R \to R$ and $g : R \to R$ are both monotonically nondecreasing functions, where Breitung (2001) used the I(0) concept defined by (14.1) given in Section 14.2. Although exploratorily attractive, Granger and Hallman's (1991) and Granger's (1995) definitions of short and extended memory processes are not accompanied with LLN, CLT, or FCLT results that can be directly used for estimation and hypothesis test purpose. Aparicio and Escribano (1998) attempted to quantify the concepts of SMM and LMM (or EMM) for bivariate cases via the information-theoretic approach; however, such an idea is still in the trial stage and needs more elaboration. Therefore, the rest of this chapter focuses on the other two approaches, which have experienced promising estimation and hypothesis test applications in practice.

Park and Phillips (1999) considered a regression model with nonlinear transformed I(1) covariate,

$$Y_t = \theta g(X_t) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.12}$$

where $\{u_t\}$ is a stationary martingale difference sequence and is independent of the I(1) process $\{X_t\}$, $g : R \to R$ is a *known* nonlinear function such as $g(x) = \ln(x)$, and θ is the parameter to be estimated. Here, $X_t = X_{t-1} + w_t$, where w_t is a stationary linear process with a zero mean and finite pth ($p > 2$) moment and its partial sum processes obey the FCLT derived in Phillips and Solo (1992). The stochastic property of Y_t should balance that of $g(X_t)$, which depends on the functional form of g. Park and Phillips (1999) considered three functional classes. Because these three classes are to be cited repeatedly in the following sections, we state their definitions below for easy reference.

(a) A function g is in Class $\mathcal{T}(I)$ if and only if it is integrable.

(b) A function g is in Class $\mathcal{T}(H)$ if and only if $g(\lambda x) = \nu(\lambda)H(x) + R(x, \lambda)$, where H is locally integrable and $R(x, \lambda)$ is asymptotically dominated by $\nu(\lambda)\ H(x)$ when $\lambda \to \infty$ and/or $|x| \to \infty$.

(c) A function g is in Class $\mathcal{T}(E)$ if and only if $g(x) = E(x) + R(x)$, where, roughly speaking, R is dominated by the monotonic function E for large $|x|$. If E is increasing (decreasing), then it is positive and differentiable on R^+ (R^-). Let $E(x) = \exp(G(x))$ on R^+ (R^-), then as $\lambda \to \infty$, $G'(\lambda x) = \nu(\lambda)D(x) + o(\nu(\lambda))$

holds uniformly in a neighborhood of x, where D is a positive (negative) and continuous and $\lambda v(\lambda) \to \infty$.

For example, $g(x) = \text{sgn}(x)$ and $g(x) = x^k$ both belong to Class $\mathcal{T}(H)$. For $g \in \mathcal{T}(E)$, E denotes the asymptotically dominating exponential component of g. For example, $g(x) = \exp\left(x^k\right)$ for $k > 0$. Because the stochastic order of $\sum_{t=1}^{T} g(X_t)$ depends on to which class g belongs, the choice of g has to at least match the convergence order of $\sum_{t=1}^{T} g(X_t)$ with that of $\sum_{t=1}^{T} Y_t$. Under certain conditions, Park and Phillips (1999) showed that

$$T^{-1/2}\sum_{t=1}^{T} g(X_t) \xrightarrow{d} \left(\int_{-\infty}^{\infty} g(x)dx\right) L_B(1,0) \qquad \text{for } g \in \mathcal{T}(I) \text{ and } p > 4 \tag{14.13}$$

$$\frac{1}{Tv(T)}\sum_{t=1}^{T} g(X_t) \xrightarrow{d} \int_{-\infty}^{\infty} H(s)L_B(1,s)ds \qquad \text{for } g \in \mathcal{T}(H) \tag{14.14}$$

$$\frac{v\left(\sqrt{T}\right)}{\sqrt{T}g\left(\max_{1\le t\le T} X_t\right)}\sum_{t=1}^{T} g(X_t) \xrightarrow{d} L_B(1,s_{\max})/D(s_{\max}) \qquad \text{for } g \in \mathcal{T}(E) \text{ and } v(\lambda) = \lambda^m \tag{14.15}$$

$$\frac{v\left(\sqrt{T}\right)}{\sqrt{T}g\left(\min_{1\le t\le T} X_t\right)}\sum_{t=1}^{T} g(X_t) \xrightarrow{d} L_B(1,s_{\max})/D(s_{\max}) \qquad \text{for } g \in \mathcal{T}(E) \text{ and } v(\lambda) = \lambda^m \tag{14.16}$$

where $T^{-1/2}\sum_{t=1}^{[Tr]} X_t \Rightarrow B(r)$ in Skorokhod space $D[0,1]$, $s_{\max} = \sup_{0\le r\le 1} B(r)$ and $s_{\min} = \inf_{0\le r\le 1} B(r)$, and $m < \left(p-8\right)/6p$ and $E|\Delta X_t|^p < M < \infty$. Therefore, for $g \in \mathcal{T}(I)$, neither $\left\{g(X_t)\right\}$ nor $\{Y_t\}$ is I(1); for $g \in \mathcal{T}(H)$, $\{Y_t\}$ will contain stochastic trend. When $g \in \mathcal{T}(E)$, the convergence results are path-dependent. Christopeit (2009), de Jong and Wang (2005), and Berkes and Horváth (2006) are some other relevant references.

Assuming that $\{u_t\}$ in model (14.12) is a martingale difference sequence and is independent of $\{X_t\}$, Park and Phillips (1999) showed that the OLS estimator of model (14.12), $\widehat{\theta}$, is a consistent estimator of θ for the three classes of functionals, and its asymptotic distribution in general is not normal distribution any more. This can be seen from the OLS formula,

$$\widehat{\theta} = \left[\sum_{t=1}^{T} g^2(X_t)\right]^{-1}\sum_{t=1}^{T} Y_t g(X_t) = \theta + \left[\sum_{t=1}^{T} g^2(X_t)\right]^{-1}\sum_{t=1}^{T} g(X_t)u_t. \tag{14.17}$$

For example, if both g and g^2 are in class $\mathcal{T}(I)$, then $T^{-1/2}\sum_{t=1}^{T} g^2(X_t) \xrightarrow{d} \left(\int_{-\infty}^{\infty} g^2(x)dx\right)L_B(1,0) \equiv \omega_1$, and

$$T^{-\frac{1}{4}}\sum_{t=1}^{T} g(X_t)u_t \xrightarrow{d} W\left(\left(\int_{-\infty}^{\infty} g^2(x)dx\right)L_B(1,0)\right) \equiv W(\omega_2), \tag{14.18}$$

a standard Brownian motion with random time. A joint convergence in the distribution step is needed to show ω_1 and ω_2 are the same so that $T^{\frac{1}{4}}\left(\widehat{\theta}-\theta\right) \overset{d}{\to} W(1)\left(\left(\int_{-\infty}^{\infty} g^2(x)dx\right)L_B(1,0)\right)^{-1/2}$, where $W(1)$ is a standard normal random variable independent of the Brownian process B. For detailed proofs, see Park and Phillips (1999, pp. 296–297). The usual asymptotic normal distribution of the OLS estimator does not hold because the denominator does not converge in probability to a constant, and such a mixed normal asymptotic distribution result highlights the limiting estimation result when a model contains nonlinearly transformed integrated covariates.

For other two cases, Park and Phillips (1999) showed that $\widehat{\theta}=\theta+O_p\left([T\nu(T)]^{-1}\right)$ if $g\in\mathcal{T}(H)$ and that $\widehat{\theta}=\theta+O_p\left(\left[\frac{\nu\left(\sqrt{T}\right)}{\sqrt{T}g^2(\max_{1\le t\le T}X_t)}\right]^{-1}\right)$ or $\widehat{\theta}=\theta+O_p\left(\left[\frac{\nu\left(\sqrt{T}\right)}{\sqrt{T}g^2(\min_{1\le t\le T}X_t)}\right]^{-1}\right)$ if $g\in\mathcal{T}(E)$. Because the choice of the functional form g is crucial in the sample theories quoted above, it will be interesting to see results on how the OLS estimator performs when model (14.12) is misspecified in the functional form of g. Kasparis' (2011) study on functional form misspecification with respect to model (14.20) is also applicable to model (14.12), although his study assumes that one correctly specifies, out of the three classes, $\mathcal{T}(I)$, $\mathcal{T}(H)$, and $\mathcal{T}(E)$, to which class g belongs.

However, in practice, it is possible that researchers may assume an incorrect functional form for the g function. In the literature of environmental Kuznet curve (EKC) study, one links pollution with economic growth. The theoretical hypothesis is that environment deteriorates at fast speed in the early stage of economic industrialization, but the deteriorating process will be reverted as an economy grows. This inverted-U shaped relationship is termed as the environmental Kuznet curve. In empirical studies, CO_2 and SO_2 emissions are usually chosen as the dependent variable, and real DGP and real GDP squared, real GDP cubed, and other variables are chosen as the regressors, where CO_2 emission and real GDP are in general believed to be I(1) variables (see, e.g., Narayan (2010) and references therein). Our discussion in Section 14.2 showed that real DGP squared and real GDP cubed series are not I(1) if real GDP is an I(1) process. Applying the partitioned OLS to a simple linear regression model such as $\mathrm{CO}_{2,t}=\alpha_0+\alpha_1\mathrm{GDP}_t+\alpha_2\mathrm{GDP}_t^2+\alpha_3\mathrm{GDP}_t^3+u_t$, one runs an unbalanced model as $g(x)=x^2$ and $g(x)=x^3$ belong to class $\mathcal{T}(H)$ with $\nu(T)=T$ and $T^{3/2}$, respectively. If CO_2 emission is I(1), one naturally expects $\alpha_2=\alpha_3=0$, which theoretically makes the inverted-U EKC impossible. One may observe the EKC in reality, but the ongoing econometric models used to capture this phenomenal are not properly designed from theoretical econometrics point of view.

Because many empirical studies are policy-oriented, forecasts based on a misspecified model can be harmful in guiding policymaking. In such circumstances, a nonparametric estimation technique designed to recover the unknown functional form from the data becomes more valuable in avoiding imposing the wrong functional form in the framework of nonlinear modeling of nonstationary time series than of stationary time series. We will discuss the nonparametric estimation technique in Section 14.4.

Sticking to the parametric setup, Kasparis (2008) attempted to test functional form misspecification. Assuming that the true data-generating mechanism for $\{Y_t\}$ is given by

$$Y_t = \sum_{j=1}^{p} \theta_j m_j\big(X_{jt}\big) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.19}$$

where $m_j(\cdot)$ has a known functional form $(j = 1, \ldots, p)$, Kasparis (2008, Lemma 1) gave the limit result of the fully modified OLS estimator of θ's derived by de Jong (2002). Kasparis (2008) derived two consistent model misspecification tests when both the true functional forms m_j's and users fully specified functional forms g_j's all belong to class $\mathcal{T}(H)$, where $\{X_t\}$ is a vector of $p(\geq 1)$-dimensional I(1) processes whose increments are a stationary linear process, and $\{u_t\}$ is an I(0) process. Model (14.19) extends model (14.12) in two aspects: It allows contemporaneous endogeneity in X's and is a *parametric* additive model of more than one covariate. With given m_j's, model (14.19) is an apparent linear regression model, but we call it a parametric "additive" model, aiming to emphasize that each I(1) covariate is nonlinearly transformed separately. For the proposed test statistics, Monte Carlo simulation results indicate that strong size distortion occurs if the serial correlation in error terms is too strong. Also, when the null hypothesis is rejected, it does not reveal the source of the rejection—no relation at all or wrong functional form specification.

Choi and Saikkonen (2004) tested for a linear cointegrating relation against a STR cointegrating model. The alternative model is given by $Y_t = \mu + \nu g(\gamma(X_{st} - c)) + \alpha^T X_t + \beta^T X_t g(\gamma(X_{st} - c)) + u_t$, $t = 1, 2, \ldots, T$, for some $s \in \{1, \ldots, d\}$, where $g(0) = 0$, $\{X_t : X_t = [X_{1t}, \ldots, X_{dt}]^T\}$ is a d-dimensional I(1) process, $\{u_t\}$ is an I(0) process with a zero mean and finite variance, $\mu, \nu, \gamma \neq 0, c, \alpha$, and β are the parameters to be estimated. The test relies on one important assumption imposed on the three-time differentiable smooth function g; that is, $g(\gamma(X_{st} - c)) \approx b\gamma(X_{st} - c)$ when X_{st} takes value close to c. Under H_0, $\nu = 0$ and $\beta = 0$, which gives a linear cointegration model, and the alternative hypothesis assumes a smooth transition cointegrating model.

Park and Phillips (2001) considered a nonlinear cointegrating model given by

$$Y_t = g(X_t, \theta) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.20}$$

where $g : R \times R^d \to R$ is a known function of an exogenous I(1) series, $\{X_t\}$, $\{u_t\}$ is a stationary martingale difference process, and θ is a d-dimensional parameter vector to be estimated. Again, the nonlinear least squared estimator $\widehat{\theta}$ converges to the true parameter vector, θ, at different rates, depending on the functional form of g. For integrable and asymptotically homogeneous function $g(\cdot, \theta)$ over all $\theta \in \Theta \subset R^d$, where

Θ is a compact subset of R^d, the NLS estimator has the same convergence rate as the OLS estimator for model (14.12). Under the same conditions, Kasparis (2011) found that the NLS estimator of model (14.20) can converge to a random variable θ^* even when $g(\cdot,\theta) \neq m(\cdot)$ for any $\theta \in \Theta$ on a set of positive Lebesgue measure, when the true data generating mechanism is given by $Y_t = m(X_t) + u_t$ for all t, and $m(\cdot)$ and $g(\cdot,\theta)$ are both class $\mathcal{T}(I)$ (or $\mathcal{T}(H)$) regular functions as defined in Park and Phillips (2001).

Choi and Saikkonen (2010) developed a KPSS type test statistic to test for a nonlinear cointegrating null model (14.20) aganist a spurious alternative hypothesis via subsamples of NLS residuals. Kasparis (2010), on the other hand, studied Bierens type of test statistic to test the validity of a *parametric* additive nonlinear regression model, $Y_t = c + \sum_{i=1}^{d} g_i(X_{it},\theta_i) + u_t$, against a general nonlinear model $Y_t = c + m(X_t) + u_t$, where $m(x) = \sum_{i=1}^{d} m_i(x_i)$ with $m_i(\cdot)$ all being class $T(I)$ functions and $\{X_t\}$ is a vector of $d(\geq 1)$-dimensional I(1) processes. Again, we call the nonlinear regression model a parametric additive nonlinear model to emphasize that each I(1) covariate has its own nonlinear parametric transformation function as in model (14.19).

Extending Granger and Swanson's (1997) stochastic unit root models to cointegrating models, Park and Hahn (1999) considered a time-varying cointegrating model

$$Y_t = \alpha_t^T X_t + u_t, \qquad t = 1,2,\ldots,T, \tag{14.21}$$

where $\alpha_t = \alpha(t/n)$ with $\alpha(\cdot)$ being an unknown smooth function defined on $[0,1]$, $\{X_t\}$ is a vector of d-dimensional I(1) processes, and its increments and $\{u_t\}$ are both stationary linear processes. Approximating α_t by linear combinations of the Fourier flexible form (or FFF) functions and then applying the least squares estimation method, Park and Hahn (1999) showed that α_t can be estimated consistently. Two tests are considered. The first is to test constant coefficients against time-varying coefficients in the presence of cointegrating or stable long-run relationship, and the second is to test a time-varying cointegrating model (14.21) against a spurious regression model.

14.4. Nonparametric Econometric Models with Nonstationary Data

In Section 14.3 we discussed some popular nonlinear parametric models, where we see that the choice of nonlinear transformation functional form is crucial in balancing the left-hand-side variable with the right-hand-side covariates. In practice, one may not know the true nonlinear functional form. Any misspecification of the unknown functional form $g(\cdot)$ may lead to spurious regression. It would be attractive to let data speak out about the form of g. Nonparametric techniques are designed to let the data reveal the underlying structure. Therefore, this section is devoted to studying the consistency of nonparametric estimators without imposing explicitly the nonlinear functional transformation form. Parallel to the models discussed in Section 14.3, this section first considers nonparametric autoregressive models of nonstationary data in Section 14.1, followed by the estimation of nonparametric cointegrating models in Section 14.2.

14.4.1. Nonparametric Autoregressive Models

Firstly, we consider model (14.4), when the functional form g is unknown. The local constant kernel estimator of $g(\cdot)$ at an interior point $y \in R$ is defined by

$$\widehat{g}(y) = \frac{\sum_{t=1}^{T} Y_t K\left(\frac{Y_t - y}{h}\right)}{\sum_{t=1}^{T} K\left(\frac{Y_t - y}{h}\right)}, \tag{14.22}$$

where $K(\cdot)$ is a second-order kernel function, and h is the smoothing parameter. It is well known that $\widehat{g}(y) - g(y) = O_p(h^2) + O_p\left((nh)^{-1/2}\right)$ for a twice continuously differentiable function g when $\{Y_t\}$ is weakly stationary and u_t is I(0) and uncorrelated with Y_{t-1}. When the true model is $Y_t = Y_{t-1} + u_t$ with $u_t \sim$ i.i.d. $\left(0, \sigma^2\right)$, Wang and Phillips (2009b) showed that, under certain conditions,

$$\sqrt{\sum_{t=1}^{T} K\left(\frac{Y_t - y}{h}\right)} [\widehat{g}(y) - g(y)] \xrightarrow{d} N\left(0, \sigma^2 \int_{-\infty}^{\infty} K^2(v) dv\right), \tag{14.23}$$

where $g(y) \equiv y$ for this case. As $\sum_{t=1}^{T} K\left(\frac{Y_t - y}{h}\right) = O_p\left(\sqrt{T} h\right)$ for integrated time series, $\widehat{g}(y) - g(y) = O_p\left(T^{1/4}(Th)^{-1/2}\right)$. Comparing this with the kernel estimator of $g(y)$ when Y_t is I(0), we notice three different features: (i) The "asymptotic variance" is deflated by a factor of $\sqrt{T}$, when g is an unknown function of an integrated time series, from the usual asymptotic variance of order $O_p\left((Th)^{-1}\right)$, when g is an unknown function of weakly dependent time series; (ii) $\left(\sqrt{T} h\right)^{-1} \sum_{t=1}^{T} K\left(\frac{Y_t - y}{h}\right) \xrightarrow{d} L_B(1, 0)$, a random variable, for an integrated Y_t series rather than $(Th)^{-1} \sum_{t=1}^{T} K\left(\frac{Y_t - y}{h}\right) \xrightarrow{d} f(y)$, a constant, the marginal density of Y_t at y for a stationary Y_t series; (iii) the difference between the kernel estimator and the true value follows a mixed normal distribution for integrated series instead of a normal distribution for weakly dependent data. Although the convergence rates are or stationary and integrated time series, combining nonparametric theory for stationary time series and (14.23) indicates that the kernel estimator is a consistent estimator of the unknown function g whether Y_t is I(0) or I(1).

In the framework of recurrent Markov chains, Karlsen and Tjøstheim (2001) derived the pointwise strong consistency of the kernel estimator of $E\left[g(Y_t) | Y_{t-1} = y\right]$ and $\mathrm{Var}\left[g(Y_t) | Y_{t-1} = y\right]$ and the limiting distribution result of the kernel estimator of $E\left[g(Y_t) | Y_{t-1} = y\right]$, where $\{Y_t\}$ can be either a β-null recurrent or a positive recurrent Markov chain (or a stationary process), although the convergence rate is slower in the null recurrent case than in the stationary case. Karlsen and Tjøstheim (2001) did not pay particular attention to the kernel estimation of the conditional variance curve in their paper other than simply stating that they use the formula $\mathrm{Var}\left[g(Y_t)|Y_{t-1}=y\right] = E\left[g^2(Y_t)|Y_{t-1}=y\right] - \left\{E\left[g(Y_t)|Y_{t-1}=y\right]\right\}^2$ to estimate the conditional variance. However, as pointed out in the literature that the kernel estimator based on this variance formula can produce negative values, one may find it attractive to apply Fan and Yao's (1998) efficient kernel estimator of conditional variance functions via a formula of $\mathrm{Var}\left[g(Y_t)|Y_{t-1}=y\right] = E\left(u_t^2|Y_{t-1}=y\right)$, where u_t is replaced by the nonparametric residuals.

14.4.2. Nonparametric Cointegrating Models

Estimating models (14.12) and (14.20) introduced in Section 14.3.2 can suffer serious functional form misspecification problem. This section discusses how to estimate the relation between nonstationary time series $\{(X_t, Y_t)\}$ via kernel-based nonparametric method.

Wang and Phillips (2009a,b) and Karlsen, Myklebust, and Tjøstheim (2007, 2010) considered the following nonparametric cointegrating model:

$$Y_t = g(X_t) + u_t, \qquad t = 1,2,\ldots,T, \tag{14.24}$$

where $g(\cdot)$ is an unknown function to be estimated. The kernel estimator of $g(x)$ is given by

$$\widehat{g}(x) = \left[\sum_{t=1}^{T} Y_t K_h(X_t - x)\right] \Bigg/ \left[\sum_{t=1}^{T} K_h(X_t - x)\right]. \tag{14.25}$$

The consistency of the kernel estimator depends on whether there are sufficient observations falling into a small neighborhood of each (interior) point x. Wang and Phillips (2009a,b) and Karlsen, Myklebust, and Tjøstheim (2007, 2010) apply different methods in dealing with localness. Specifically, Karlsen and Tjostheim (2001) and Karlsen et al. (2007, 2010), in the framework of recurrent Markov chains, studied the localness feature of a class of nonstationary time series called β-null recurrent Markov chains according to its average number of visits to a small neighborhood of each point, while Wang and Phillips (2009a,b) relied on a local time density (see Section 14.2).

For easy reference, we briefly summarize the key assumptions imposed on $\{(X_t, u_t)\}$ in these papers. In Wang and Phillips (2009a,b), $\{X_t\}$ is an I(1) or a near I(1) process and $\{u_t\}$ is an I(0) process, where Wang and Phillips (2009a) required $\{u_t\}$ to be a martingale difference sequence, while Wang and Phillips (2009b) allowed serially correlated $\{u_t\}$ and some correlation between X_t and u_s for $|t-s| \le m$ for some finite $m > 0$. In Karlsen, Myklebust, and Tjøstheim (2007, 2010), $\{X_t\}$ is a β-null recurrent time series and $\{u_t\}$ is a stationary Markov chain satisfying some mixing conditions in their 2007 paper and a strictly stationary linear process in their 2010 paper, where $\{X_t\}$ can be a random walk, unit root process, and other nonlinear processes according to Myklebust, Karlsen, and Tjøstheim (2011), and $\{u_t\}$ and $\{X_t\}$ are not required to be independent in their 2007 paper.

Wang and Phillips (2009a,b) and Karlsen, Myklebust, and Tjøstheim (2007, 2010) studied the limiting result of kernel estimator (14.25) when model (14.24) represents a meaningful long-run relation between nonstationary times series $\{X_t\}$ and $\{Y_t\}$. However, as usual, the model can be misspecified. For example, the model can be spurious if there is no true relation between the nonstationary time $\{X_t\}$ and $\{Y_t\}$. Kaparis and Phillips (2012), on the other hand, considered another interesting misspecification—temporal or dynamic misspecification. The following two paragraphs summarize the findings of Phillips (2009) and Kasparis and Phillips (2012) for the performance of the kernel estimator in the presence of spurious regression and dynamic misspecification, respectively.

Phillips (2009) investigated the asymptotic performance of the kernel estimator in a spurious nonparametric regression model of nonstationary time series. Specifically, he studied model (14.24) when it is spurious, where $\{(X_t, Y_t)\}$ obeys a FCLT result. If $\{X_t\}$ and $\{Y_t\}$ are independent I(1) processes, Phillips (2009, Theorem 2) implies that the kernel estimator $\widehat{g}(x) = O_p\left(\sqrt{T}\right)$ for x satisfying $\lim_{n\to\infty} x/\sqrt{T} = c$ ($c = 0$ if x is a fixed finite constant). Phillips (1986) derived asymptotic results for the t-statistic for significance of X_t in linear spurious regression of $Y_t = \alpha + \beta X_t + u_t$, and he found that the t-statistic is explosive, R^2 has a nondegenerate limiting distribution as $n \to \infty$, and DW converges to zero, which provided a theoretical foundation for Granger and Newbold's (1974) numerical observations on under-rejection of the significance test based on the t-statistic for linear spurious regression models of two independent integrated time series. Parallel to the linear spurious regression case, for a given c, Phillips (2009) constructed a local version of the t-statistic, R^2, and the DW statistic via the kernel method, and also he found that the local t-statistic is explosive, the local R^2 is nondegenerately distributed, and the local DW statistic converges to zero in large samples for the nonparametric spurious regression case. Phillips' (2009) study therefore brings the need for a "cointegrating" test to our current research agenda for econometrics modeling of nonlinearly transformed integrated time series.

In terms of dynamic misspecification, it highlights additional difference between linear cointegrating modeling aganist nonlinear cointegrating modeling. For the linear cointegrating case, if Y_t and X_t are linearly cointegrated, so are Y_t and X_{t-m} for a finite $m > 0$. Kasparis and Phillips (2012) described this property as the *invariance of time translation* of a linear cointegrating relation. However, such an invariance property may not hold for a nonlinear cointegrating relation. For example, if a true model is given by

$$Y_t = \theta X_t^2 + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.26}$$

where $X_t = X_{t-1} + v_t$, it follows that

$$Y_t = \theta E\big(v_t^2\big) + \theta X_{t-1}^2 + \xi_t, \tag{14.27}$$

where $\xi_t = 2\theta X_{t-1} v_t + \theta\big(v_t^2 - Ev_t^2\big) + u_t$. To simplify our discussion, we assume $X_0 \equiv 0$, $v_t = \Delta X_t \sim$ i.i.d. $\big(0, \sigma_v^2\big)$, $u_t \sim$ i.i.d. $\big(0, \sigma_u^2\big)$, and $\{u_t\}$ and $\{v_t\}$ are independent of each other. Then, it is easy to calculate $E(\xi_t) = 0$, $\text{Var}(\xi_t) = \sigma_u^2 + \theta^2 \text{Var}\big(v_t^2\big) + 4(t-1)\theta^2\sigma_v^4$, and $E(\xi_t \xi_s) = 0$ for $t \neq s$. One can show that the OLS estimator of θ for the temporally or dynamically misspecified model still converges to the true value θ at the rate of T^{-1}; however, the temporal misspecification in the covariate X does cause the compounded error ξ_t to be nonstationary, so that model (14.27) does not have an I(0) error any more.

Kasparis and Phillips (2012) considered the following two models

$$Y_t = g(X_{t-r}) + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.28}$$

$$Y_t = g(X_{t-s}) + \xi_t, \tag{14.29}$$

where both r and s are some positive integers but $r \neq s$, $g : R \to R$ is a locally integrable function, $\{u_t\}$ is a stationary martingale difference sequence, and $X_t = \sum_{i=1}^{t} v_i$ is I(1), although it can be relaxed to the data such that a properly scaled X_t obeys a FCLT result. Most importantly, model (14.28) is the true model, while model (14.29) is not the true model but wrongly estimated by a researcher. The misspecification problem is said to be mild if $|r - s|$ is finite and server if $|r - s| \to \infty$ as $T \to \infty$.

Closely following the assumptions imposed in Wang and Phillips (2009a) and for the case that $|r - s|$ is finite, Kasparis and Phillips (2012) showed that $\widehat{g}(x)$ defined by (14.25) converges in probability to $E\big[g\big(x + \sum_{rs} v_i\big)\big]$ if $\sqrt{T}h \to \infty$ and $h \to 0$ as $T \to \infty$, where $v_t = X_t - X_{t-1}$ is I(0) and $\sum_{rs} v_i = X_r - X_s$ if $r > s$ and $X_s - X_r$ if $r < s$. This result is the same as that derived for stationary time series. Under certain conditions, they further showed that

$$\sqrt{\sum_{t=s+1}^{T} K\left(\frac{X_{t-s} - x}{h}\right)}\left[\widehat{g}(x) - g\left(x + \sum_{rs} v_i\right)\right] \xrightarrow{d} N\left(0, \sigma^2(x)\int_{-\infty}^{\infty} K^2(v)\,dv\right), \tag{14.30}$$

where $\sigma^2(x) = \sigma_u^2 + \text{Var}\big[g\big(x + \sum_{rs} v_i\big)\big]$.

Consider two examples here. First, if $g(x) = x^2$, we have $E\big[g\big(x + \sum_{rs} v_i\big)\big] = g(x) + \text{var}\big(\sum_{rs} v_i\big) \neq g(x)$, where the asymptotic bias term of the kernel estimator depends on the accumulated variance of $\{X_t\}$ between time r and s. Second, if $g(x) = \theta x$, the kernel estimator is consistent but its asymptotic variance will be larger for $r \neq s$ than for $r = s$ as $\sigma^2(x) = \sigma_u^2 + \text{Var}\big(\sum_{rs} v_i\big)$.

When $|r - s| \to \infty$ as $T \to \infty$, Kasparis and Phillips (2012) obtained the convergence result setting $s = s_T = [cT]$ for some $c \in (0,1)$ in model (14.29). Interested readers are directed to their paper, but in general sense, $\widehat{g}(x)$ is inconsistent and converges to 0 if g is integrable and explodes if g is an unbounded local integrable class $\mathcal{T}(H)$ function. However, $\widehat{g}(x) \xrightarrow{p} E\big[g(X_t)\big]$ holds true for weakly dependent data.

When model (14.12) suffers dynamic misspecification in its covariate, the OLS estimator of the misspecified model can be consistent for some functional classes; see model (14.26) and linear case, for example. Based on this observation, Kasparis and Phillips (2012) considered a t-statistic based on the kernel estimation of model (14.29) to test a linear null of $g(x) = \theta_0 + \theta_1 x$ aganist a nonlinear alternative. For a given point x, the nonparametric t-statistic $\widehat{t}(x,\widehat{\theta}) \overset{d}{\to} N(0,1)$ under H_0 whether there is a temporal or dynamic misspecification or not. They then constructed two test statistics $\widehat{F}_{sum} = \sum_{x\in\mathcal{X}_k}\left[\widehat{t}(x,\widehat{\theta})\right]^2$ and $\widehat{F}_{\max} = \max_{x\in\mathcal{X}_k}\left[\widehat{t}(x,\widehat{\theta})\right]^2$, where $\mathcal{X}_k = \{\bar{x}_1,\ldots,\bar{x}_k\} \subset R$ for some $k \in N$ is a set of preselected and isolated points. Under H_0, $\widehat{F}_{sum} \overset{d}{\to} \chi^2(k)$ and $\widehat{F}_{\max} \overset{d}{\to} Z$, where the cumulative distribution function of the random variable Z is $F_Z(z) = \left[\Pr(\chi^2(1) \leq z)\right]^k$. Under H_1, both statistics are of stochastic order $O_p\left(\sqrt{T}h\right)$ whether g belongs to class $\mathcal{T}(I)$ or $\mathcal{T}(H)$. Therefore, both test statistics are consistent. The choice of $\mathcal{X}_k$ and k evidently will affect the size and power of both tests, but it is not discussed in the paper. Under H_0, the Monte Carlo simulations (Table 2 in their paper) suggest to use smaller bandwidth for larger $|r-s|$, and the rejection rates actually increase as sample size increases given α and for large $|r-s|$, where $h = T^{-\alpha}$. More careful study is needed to understand how to select the bandwidth given an unknown $|r-s|$, and an alternative bootstrap method may be worth considering.

14.5. Semiparametric Models with Nonstationary Data

Parallel to the models discussed in Sections 14.3 and 14.4, this section first considers semiparametric autoregressive models of nonstationary data in Section 14.5.1, followed by the estimation of semiparametric varying coefficient cointegrating models in Section 14.5.2. Section 14.5.3 presents an estimation method for semiparametric varying coefficient models with correlated but not cointegrated data. Section 14.4 contains some recent developments in parametric and semiparametric discrete choice models. Section 14.5.4 discusses a semiparametric time trend model. Due to space limitation, the current chapter does not include recent works on partially linear regression models and we refer interested readers to Chen, Gao, and Li (2012), Gao and Phillips (2011), and Juhl and Xiao (2005).

14.5.1. Semiparametric Autoregressive Models

Consider a functional-coefficient conditional mean regression model of the following form:

$$E\left(X_t|X_{t-1},\ldots,X_{t-p},Z_t\right) = a_1(Z_t)X_{t-1} + \cdots + a_p(Z_t)X_{t-p}, \qquad (14.31)$$

where $a_j(Z_t)$ can be nonlinear functions of Z_t. This model nests Chen and Tsay's (1993) functional coefficient autoregression (FAR) models given by

$$X_t = a_1\big(X_{t-1}^*\big)X_{t-1} + \cdots + a_p\big(X_{t-1}^*\big)X_{t-p} + u_t, \qquad t = p+1, \ldots, T, \tag{14.32}$$

where $X_{t-1}^* = \big(X_{t-i_1}, \ldots, X_{t-i_k}\big)^T$ and $\{u_t\}$ is a sequence of i.i.d. random variables independent of $\{X_s : s < t\}$. The FAR models include self-exciting threshold autoregressive (SETAR) models and smooth transition AR (STAR) models as special cases. Cai, Fan, and Yao (2000) developed asymptotic results for local linear estimator of model (14.31) for stationary time series, where the nonparametric estimator of the unknown functional coefficient curves is asymptotically normally distributed with the usual nonparametric convergence rate of $O_p\big(h^2 + (Th)^{-1/2}\big)$.

Let $\mathcal{F}_t$ be the smallest sigma field that containing all the past history of the data up to time t. If Z_t is $\mathcal{F}_{t-1}$ measurable, stationary, and geometrically absolutely regular, and X_t is generated from a linear $ARIMA(p,1,0)$ model, Juhl (2005) showed that the local linear estimator of $a_1(z)$ converges to the true value of one at a speed faster than the stationary case by an order of $O_p\big(T^{-1/2}\big)$ such that the "asymptotic variance" of the local linear estimator of $a_1(z)$ is of order $O_p\big(T^{-1}(Th)^{-1}\big)$. He also obtained a mixed normal limit distribution of the estimator under the assumption that the model error terms are assumed to be an i.i.d. sequence and independent of past X's. Combining the results of Cai et al. (2000) and Juhl (2005), we see the local linear estimator of the unknown functional coefficient curves consistent whether $\{X_t\}$ is an I(0) or an I(1) process.

14.5.2. Semiparametric Varying Coefficient Cointegrating Models

Cai, Li, and Park (2009) and Xiao (2009b) studied a semiparametric functional coefficient models for nonstationary time series data,

$$Y_t = X_t^T g(Z_t) + u_t, \qquad 1 \le t \le n, \tag{14.33}$$

where Y_t, Z_t, and u_t are scalar, $X_t = (X_{t1}, \ldots, X_{td})^T$ is a $d \times 1$ vector, $\beta(\cdot)$ is a $d \times 1$ column vector of unknown functions. When all the variables are stationary or independent, this model has been studied by Cai et al. (2000) and Li et al. (2002), among others. For stationary u_t and $Z_t = t/n$, Robinson (1989), Cai (2007), and Chen and Hong (2012) derived estimation theory (for stationary X_t) of local constant and local linear estimators, respectively; Park and Hahn (1999) approximated the unknown coefficient curves by trigonometric series for integrated X_t.

Assuming stationary u_t and weakly exogenous covariates X_t and Z_t, Cai, Li, and Park (2009) derived limit results for local linear estimator of $g(\cdot)$ for two cases: (i) $\{(u_t, Z_t, X_{1t})\}$ are stationary α-mixing processes, where $X_t = \big(X_{1t}^T, X_{2t}^T\big)^T$, $\{X_{1t}\}$ is I(0), and $\{X_{2t}\}$ is I(1); (ii) $\{Z_t\}$ is an I(1) process, $\{(u_t, X_t, \Delta Z_t)\}$ are all stationary α-mixing processes, and $\{u_t\}$ is a martingale difference process with respect to the smallest σ-field of $\{X_t, Z_t, X_{t-1}, Z_{t-1}, \ldots\}$. Case (i) is also studied independently by Xiao (2009b) via local polynomial regression approach. Complementing to the existing literature, Sun, Cai, and Li (2013) considered the case that both $\{X_t\}$ and $\{Z_t\}$ are I(1) processes.

In all cases mentioned above, the local linear estimator consistently estimates $g(\cdot)$, although the convergence rate varies with respect to the stochastic properties of X_t and Z_t. It is wellknown for stationary and independent data cases that the kernel and local linear estimator are expressed as $\widehat{g}(z) = g(z) + O_p\big(h^2 + (Th)^{-1/2}\big)$ for any interior point z, where h is the bandwidth parameter and T is the sample size. The "asymptotic variance" of $\widehat{g}(z)$ is of order $O_e\big(T^{1/2}(Th)^{-1}\big)$ when $\{X_t\}$ is I(0) and $\{Z_t\}$ is I(1), of order $O_e\big(T^{-1}(Th)^{-1}\big)$ when $\{X_t\}$ is I(1) and $\{Z_t\}$ is I(0) by Cai et al. (2009), and of order $O_e\big(T^{-1/2}(Th)^{-1}\big)$ when both $\{X_t\}$ and $\{Z_t\}$ are I(1) by Sun, Cai, and Li (2013). Therefore, the unknown function with an integrated argument, Z_t, inflates the "asymptotic variance" by an order of $T^{1/2}$ (for a given stochastic process X_t), while an integrated X_t deflates the asymptotic variance by an order of T^{-1} (for a given stochastic process Z_t).

Sun and Li (2011) derived limit result for data-driven bandwidth selected via cross-validation method for model (14.33). When X_t contains both I(0) and I(1) variables and Z_t is I(0), Sun and Li (2011) found that the CV-selected bandwidth converges in distribution to a random variable at different rates for the local constant kernel estimator and the local linear kernel estimator. Specifically, for the local constant estimator, the CV-selected bandwidth is $\widehat{h}_{lc} = O_e\big(T^{-1/2}\big)$, and for the local linear estimator, the CV-selected bandwidth is $\widehat{h}_{ll} = O_e\big(T^{-2/5}\big)$, where $O_e(1)$ means an exact order probability of $O_p(1)$ and it is not $o_p(1)$. The local constant estimator gives a larger average squared error than the local linear estimator. This result implies a sharp contrast to the existing results derived for stationary and independent data cases. The reason behind the different performance of the local constant and linear estimators lies in the leading squared bias term of the local constant estimator, which is $O_e(h/T)$ and is larger than the squared bias term of order $O_e\big(h^4\big)$ for the local linear estimator. Their results favor the usage of the local linear estimator over the local constant estimator if one is interested in obtaining better estimation result. When X_t contains both I(0) and I(1) components, the cross-validation selected bandwidth is of order $O_e\big(T^{-2/5}\big)$, not $O_e\big(T^{-1/5}\big)$, as the I(1) components of X_t dominate the I(0) components in asymptotic performance. Hence, it is not surprising to see that, in their Table 4, the CV-selected bandwidth produces smaller mean squared errors for the estimator of the coefficient curve in front of the I(1) component of X_t than that of the I(0) component of X_t. The estimation accuracy of the coefficient in front of the I(0) component of X_t can be improved further, though. Let $Y_t = X_{1t}^T g_1(Z_t) + X_{2t}^T g_2(Z_t) + u_t$, where X_{1t} is I(0) and X_{2t} is I(1). One first estimates the model by the local linear estimation approach, using the CV-selected bandwidth. Name the estimator $\widehat{g}_1(\cdot)$ and $\widehat{g}_2(\cdot)$. Then, one estimates $\tilde{Y}_t = X_{1t}^T g_1(Z_t) + v_t$, where $\tilde{Y}_t = Y_t - X_{2t}^T \widehat{g}_2(Z_t)$, by the local linear regression approach with a new CV-selected bandwidth from this model and name this estimator $\tilde{g}_1(\cdot)$. Sun and Li (2011) showed that $\tilde{g}_1(\cdot)$ performs better than $\widehat{g}_1(\cdot)$ in their Monte Carlo designs when both X_{1t} and Z_t are strictly exogenous.

14.5.3. Semiparametric Varying Coefficient Models with Correlated but not Cointegrated Data

Sun, Hsiao, and Li (2011, SHL hereafter) constructed a consistent estimator of $g(\cdot)$ when Z_t is I(0), but both $\{X_t\}$ and $\{u_t\}$ are I(1) processes in model (14.33). That is, SHL considered a case that Y_t and X_t are not cointegrated in model (14.33) even with a varying coefficient function $g(Z_t)$, where the I(1) error term u_t may be due to omitted variables and/or measurement errors. For example, the true model is a partially linear cointegrating model,

$$Y_t = X_t^T g(Z_t) + W_t\delta + \epsilon_t = X_t^T g(Z_t) + u_t, \tag{14.34}$$

where both $\{X_t\}$ and $\{W_t\}$ are I(1), $\{Z_t\}$ and $\{\epsilon_t\}$ are I(0), and $\left[1, -\theta(Z_t)^T, \delta\right]^T$ is the varying cointegrating vector. However, if W_t is not observable, then the composite error in model (14.34), $u_t = W_t\delta + \epsilon_t$, is an I(1) process if $\delta \neq 0$, and Y_t and X_t do not form a stable relation with W_t missing from the model. Under some conditions, it is easy to show that $\widehat{\theta}_0 - E\left[g(Z_t)\right] \xrightarrow{d} \bar{\theta}_1$ and $\breve{g}(z) - g(z) \xrightarrow{d} \bar{\theta}_2$, where $\widehat{\theta}_0$ is the OLS estimator of the linear model $Y_t = X_t^T\theta_0 + error_t$, $\breve{g}(z)$ is the kernel estimator of model $Y_t = X_t^T g(Z_t) + error_t$, and $\bar{\theta}_1$ and $\bar{\theta}_2$ have the same nondegenerate distribution. Motivated by these results, they constructed two estimators for $\alpha(z) = g(z) - c_0$, where $c_0 = E\left[g(Z_t)\right]$; that is, $\tilde{\alpha}(z) = \breve{g}(z) - \widehat{\theta}_0$ and $\widehat{\alpha}(z) = \breve{g}(z) - n^{-1}\sum_{t=1}^{T}\breve{g}(Z_t)M_{nt}$, where $M_{nt} = M_n(Z_t)$ is the trimming function that trims away observations near the boundary of the support of Z_t. SHL showed the consistency of the two proposed estimators and derived the limiting results.

To obtain an estimator for c_0, one needs to rewrite model (14.33) as

$$Y_t = X_t^T c_0 + X_t^T\alpha(Z_t) + u_t, \tag{14.35}$$

and subtracting $X_t^T\tilde{\alpha}(Z_t)$ in (14.35) gives

$$\tilde{Y}_t \overset{def}{=} Y_t - X_t^T\tilde{\alpha}(Z_t) = X_t^T c_0 + v_t, \tag{14.36}$$

where $v_t = X_t^T[\alpha(Z_t) - \tilde{\alpha}(Z_t)] + u_t$. Because $\tilde{\alpha}(Z_t)$ is a consistent estimator of $\alpha(Z_t)$, $\tilde{Y}_t$ mimics the stochastic properties of $Y_t - X_t^T\alpha(Z_t)$, and the stochastic property of v_t is dominated by that of u_t. Taking a first difference of (14.36) gives $\Delta\tilde{Y}_t = \zeta_t^T c_0 + \Delta v_t$, from which one can calculate the OLS estimator $\tilde{c}_0$. It gives $\tilde{g}(z) = \tilde{\alpha}(z) + \tilde{c}_0$ to estimate $g(z)$. Similarly, replacing $\tilde{\alpha}(\cdot)$ in (14.36) by $\widehat{\alpha}(\cdot)$ and following the same procedure, one obtains the OLS estimator $\widehat{c}_0$ and another estimator for $g(z)$ by $\widehat{g}(z) = \widehat{\alpha}(z) + \widehat{c}_0$. Under some mixing conditions of $\{(\Delta X_t, \Delta u_t, Z_t)\}$ and $h \to 0$, $nh \to \infty$ and $nh^5 = O(1)$ as

$n \to \infty$ and other regularity conditions, Sun, Hsiao, and Li (2011) derived the limiting result for $\tilde{\alpha}(z)$ in their Theorem 3.1, but deriving the limiting result for $\widehat{\alpha}(z)$ requires the independence between $\{(\Delta X_t, \Delta u_t)\}$ and $\{Z_t\}$ as indicated in their Theorem 3.2. Overall, SHL showed that both $\tilde{\alpha}(z)$ and $\widehat{\alpha}(z)$ converge to $\alpha(z)$ with a convergence rate of $O_p\big(h^2\big) + O_p\big((Th)^{-1/2}\big)$, and the same convergence rates hold for $\tilde{c}_0$, $\widehat{c}_0$, $\tilde{g}(z)$, and $\widehat{g}(z)$ as shown in their Theorem 3.3. Monte Carlo simulations show that $\widehat{g}(z)$ performs better than the simple kernel estimator (14.25) when both $\{X_t\}$ and $\{u_t\}$ are near I(1) process in model (14.33), where both estimators have the same convergence rate. An empirical study of national market spillover effects across US/UK and US/Canadian markets can be found in Sun, Hsiao, and Li (2012).

14.5.4. Semiparametric Binary Choice Models

We start with a standard binary choice model, where the observed data, $\{(X_t, Y_t)\}_{t=1}^T$, is generated by

$$Y_t = I\big\{Y_t^* > 0\big\}, \tag{14.37}$$

$$Y_t^* = X_t^T \beta_0 + u_t, \qquad t = 1, 2, \ldots, T, \tag{14.38}$$

with Y_t^* being a latent variable and $\{u_t\}$ being an i.i.d. sequence with a zero mean and unit variance. When all the data are weakly dependent or independent, it is well known that the maximum likelihood (ML) estimator of β_0 is $\sqrt{T}$-consistent and has an asymptotic normality distribution (see, e.g., Dhrymes (1986) and Wooldridge (1994)). However, when $\{X_t\}$ is an I(1) process and

$$E(Y_t|\mathcal{F}_{t-1}) = \Pr(Y_t = 1|\mathcal{F}_{t-1}) = F\Big(X_t^T \beta_0\Big), \tag{14.39}$$

where $\mathcal{F}_{t-1}$ is some natural filtration with respect to which u_t is measurable and $F(\cdot)$ is the known marginal cumulative distribution of $\{u_t\}$, Park and Phillips (2000) showed that the MLE of β_0 (with $\beta_0 \neq 0$) is $T^{1/4}$-consistent and has a mixed normal distribution, while Guerre and Moon (2002) showed that the MLE of β_0, when $\beta_0 = 0$, is T-consistent. Note that different components of $\{X_t\}$ are assumed not to be cointegrated among themselves.

Park and Phillips (2000) and Guerre and Moon (2002) both assume that the distribution of $\{u_t\}$ is known and correctly specified, which may not hold true in practice. For independent data case, nonparametric methods have been introduced to estimate β_0 with unknown error distributions, including the local maximum likelihood estimator of Fan, Farmen, and Gijbels (1998), maximum score estimator of Manski (1985, 1988) and Horowitz (1992), the semiparametric least-squares (SLS) estimator of Ichimura (1993) and Härdle, Hall, and Ichimura (1993), and the semiparametric efficient estimator of Klein and Spady (1993), where Klein and Spady (1993) considered a case that $\Pr(Y_t = 1|\mathcal{F}_{t-1}) = F(v(X_t, \beta_0))$ has an unknown F but known $v(\cdot,\cdot)$ up to an unknown β_0. The connection between the binary choice model and single-index model is well known, and one can construct a semiparametric single-index model from (14.39),

$$Y_t = F\Big(X_t^T\beta_0\Big) + \varepsilon_t, \qquad t = 1,2,\ldots,T, \tag{14.40}$$

where $\{(\varepsilon_t, \mathcal{F}_t)\}$ is a martingale difference sequence with conditional variance $E\big(\varepsilon_t^2|\mathcal{F}_{t-1}\big) = \sigma^2\big(X_t^T\beta_0\big)$ and $\sigma^2(z) = F(z)[1-F(z)]$. The SLS estimator of Ichimura (1993) and Härdle et al. (1993) is proposed to estimate the unknown parameter β_0 by the nonparametric least squares method from model (14.40) for independent data. Out of different interest, Chang and Park (2003) derived the consistency and limit results of the nonlinear least squares (NLS) estimator of a simple neutral network model with integrated time series, which is defined as $Y_t = \mu + \alpha G\big(X_t^T\beta_0\big) + \varepsilon_t$, $t = 1,2,\ldots,T$, with a known, bounded, three-time differentiable function $G: R \to R$ satisfying $\lim_{x\to-\infty} G(x) = 0$ and $\lim_{x\to\infty} G(x) = 1$; this model is a parametric single-index model more general than model (14.40).

When $\{X_t\}$ is an I(1) process and $E(\Delta X_t) = 0$ in (14.38), Moon (2004) and Guerre and Moon (2006) studied Manski's maximum score estimator of $\beta_0 \neq 0$ when the error distribution, F, is unknown. Moon (2004) obtained the identification condition and consistency result of the estimator, and Guerre and Moon (2006) showed that both Manski's maximum score estimator and Horowitz's (1992) smoothed maximum score estimator are $\sqrt{T}$-consistent under some conditions. Parallel to the literature on stationary and independent data, it will be interesting to see how the local MLE of Fan et al. (1998) and the SLS estimator of Ichimura (1993) and Härdle et al. (1993) perform when the covariates are I(1).

Further, Hu and Phillips (2004) and Phillips, Jin, and Hu (2007, PJH hereafter) extended the study of nonstationary binary choice models to nonstationary discrete choice models with each threshold value equal to $\sqrt{T}$ times a constant, while assuming a correctly specified error distribution. In the presence of non-zero threshold values, PJH showed that both the threshold parameters and β_0 are $T^{3/4}$-consistent. As the error distribution is generally unknown in practice, more research work is needed to study the performance of quasi-maximum likelihood estimation with misspecified error distribution function and semiparametric discrete choice models with unknown error distribution function.

14.5.5. A Time Trend Variable Coefficient Model

Liang and Li (2012) considered the estimation of the following semiparametric time trend varying coefficient model

$$Y_t = X_t^T\beta_1(Z_t) + t\beta_2(Z_t) + u_t,$$

where X_t and $\beta(Z_t)$ are $d \times 1$ vectors, Y_t, Z_t, u_t are scalars, and t is the time trend variable. X_t, Z_t, and u_t are all stationary I(0) variables. The nonstationarity comes from the time trend variable. The functional forms of $\beta_1(\cdot)$ and $\beta_2(\cdot)$ are not specified. Liang and Li (2012) showed a surprising result that the local constant estimation method leads to inconsistent estimation result for $\beta_2(z)$. They then suggested to use a local polynomial method to approximate the $\beta_2(\cdot)$ function, while keeping the local constant approximation to the $\beta_1(\cdot)$ function. They derived the rate of convergence and asymptotic normal distribution result for their proposed estimators of $\beta(z)$, and they proposed some test statistics for testing whether $\beta(z)$ is a constant vector or has some parametric functional forms.

14.6. Model Specification Tests with Nonstationary Data

With nonstationary time series, functional form misspecification is tested under two general setups. The first one conducts tests when a model represents a meaningful relation under both null and alternative hypotheses. The second one conducts tests when the model is meaningful under one hypothesis and is spurious under another hypothesis. In many cases, one finds that a wrongly specified functional form can lead to a spurious regression. However, the rejection of a cointegrating null hypothesis can be resulted from functional misspecification or spurious regression if no prior knowledge is known. In this sense, testing for spurious regression or functional form misspecification is nonseparable in many occasions. Some consolidation method can be useful (e.g., Harvey, Leybourne, and Xiao (2008)).

The RESET test of Ramsey (1969), White's information criterion test, Wald and LM tests are the most popularly used classical model misspecification tests; see Godfrey (1988) for an excellent survey. However, Lee et al. (2005 and references therein) showed that the traditional RESET test, White's information criterion test and other tests for nonlinearity derived for weakly dependent and independent data are not proper tests for I(1) data, as the limit results are completely different. Consequently, Hong and Phillips (2010) constructed a modified RESET test to correct for finite sample bias arising from endogeneity of integrated covariates and showed that the modified RESET test has a standard normal limiting distribution under the null hypothesis of a linear cointegrating model, but the test statistic is of order $O_p(T/M)$ under the alternative linear spurious regression or nonlinear cointegrating regression model (14.24) when g belongs to the functional class $\mathcal{T}(H)$, where M is the bandwidth used to calculate the kernel estimation of the long-run (co)variance(s). However, when g belongs to $\mathcal{T}(I)$, the RESET test fails to detect a linear cointegrating model from a nonlinear cointegrating model.

Gao, King, Lu, and Tjøstheim (2009a) constructed a unit root test via nonparametric method from model (14.4); that is, they considered a class of nonlinear autoregressive models, $Y_t = g(Y_{t-1}) + u_t$, $t = 1,2,\ldots,T$, where $\{u_t\}$ is an i.i.d. sequence with a zero mean and a finite fourth moment. The null and alternative hypotheses are given by

$$H_0 : \Pr\big\{g(Y_{t-1}) = Y_{t-1}\big\} = 1, \tag{14.41}$$

$$H_1 : \Pr\big\{g(Y_{t-1}) = Y_{t-1} + \Delta_T(Y_{t-1})\big\} = 1, \tag{14.42}$$

where $\big\{\Delta_T\big(y\big)\big\}$ is a sequence of unknown functions. Under H_0, $\{Y_t\}$ is a random walk process; under H_1, $\{Y_t\}$ can be a nonlinear stationary process if $\big|g\big(y\big)\big| \le c\big|y\big|$ for extremely large $\big|y\big|$ and $c < 1$. Estimating the unknown curve $g\big(y\big) = E\big(Y_t|Y_{t-1} = y\big)$ by the kernel estimator $\widehat{g}\big(y\big)$ given by (14.22) and calculating the nonparametric estimated residuals, $\widehat{u}_t = Y_t - \widehat{g}(Y_{t-1})$, Gao et al. (2009a) constructed a standardized residual-based test statistic

$$\widehat{L}_T(h) = M_T(h)/\sqrt{\widehat{\sigma}_T^2(h)}, \tag{14.43}$$

where $M_T(h) = \sum_{t=1}^T \sum_{s=1,s\neq t}^T \widehat{u}_t\widehat{u}_s K_{h,s-1,t-1}$, $\widehat{\sigma}_T^2(h) = 2\sum_{t=1}^T \sum_{s=1,s\neq t}^T \widehat{u}_t^2\widehat{u}_s^2 K_{h,s-1,t-1}^2$, and $K_{h,s-1,t-1} = h^{-1}K\Big(\frac{Y_{s-1}-Y_{t-1}}{h}\Big)$. They showed that $\widehat{L}_T(h) \overset{d}{\to} N(0,1)$ under H_0 under some conditions including that u_t has a symmetric probability density function, where we have the bandwidth $h \to 0$, $Th \to \infty$, $Th^4 \to 0$ as $T \to \infty$. Such standardized residual-based tests have been widely used in testing model specification for weakly dependent and independent data (see, e.g., Zheng (1996), Li and Wang (1998), Fan and Li (1999)). Gao et al.'s (2009a) test is a test alternative to Dickey and Fuller's (1979) unit root test as both test a unit root null against a stationary alternative, but is confined to a random walk null instead of a general unit root process. In addition, we are not sure whether the residual-based test statistic can be extended to test for null recurrent time series against positive recurrent time series in the framework of Markov chains, and more research needs to be done in the future.

Applying the standardized residual-based test technique, Gao, King, Lu, Tjøstheim (2009b) tested a parametric null aganist a nonparametric alternative hypothesis; that is,

$$H_0 : \Pr\big\{g(X_t) = g(X_t,\theta_0)\big\} = 1 \qquad \text{for some } \theta_0 \in \Theta \subset R^d,$$

$$H_0 : \Pr\big\{g(X_t) \neq g(X_t,\theta)\big\} > 0 \qquad \text{for all } \theta \in \Theta \subset R^d$$

when model (14.20) and model (14.24) both represent meaningful relation between $\{X_t\}$ and $\{Y_t\}$, where $\Delta X_t \sim$ i.i.d. $\big(0,\sigma_u^2\big)$ with a symmetric marginal density function is independent of the martingale difference errors in the model. The test statistic converges to a standard normal distribution under H_0. As explained in Park and Phillips (2001), the convergence rate of the NLS estimator under H_0 depends on the functional form, Gao et al. (2009b, Assumption 2.2 (ii)) imposes a restriction on the convergence rate of the NLS estimator and the functional form g. Wang and Phillips (2012) considered the same test statistic but for the case that X_t is a predetermined variable relative to Y_t.

The current research results seem to suggest that the standardized residual-based test for functional misspecification is consistent as long as the model under the null is nested within the model under alternative and the alternative model is not spurious no matter the data are I(0) or I(1). There is no result pushed to null recurrent time series. In addition, as the estimation results are still limited comparing with what the literature has developed for weakly and independent data, many commonly used statistics for various hypotheses testing problems have unknown asymptotic properties when applied with nonstationary data—for example, variable selection with non-nested hypotheses, linear against partially linear regression, and so on.

In the semiparametric varying coefficient framework, Xiao (2009b) provided a maximum chi-squared test to test a linear cointegrating model against a functional-coefficient cointegrating model. Sun, Cai, and Li (2012) considered the problem of testing $g(z) = g_0$ for all $z \in R$, where g_0 is a constant vector of parameters in a semiparametric varying coefficient model $Y_t = X_t^T g(Z_t) + u_t$, where X_t can contain both stationary I(0) and nonstationary I(1) components, and Z_t is a stationary variable.

14.7. Co-summability: Cointegration of Nonlinear Processes

After reviewing several non-/semiparametric regression models of nonstationary and persistent time series data, we now discuss an extension of cointegrating relation from linear to nonlinear setup. In this section we discuss the co-summability concept introduced by Berenguer-Rico and Gonzalo (BG) (2013), who attempted to consolidate the linear and nonlinear cointegrating concepts via the sampling approach. We re-state BG's (2013) definitions 2 and 4 below for readers' convenience.

Definition. A stochastic process $\{Y_t\}$ with positive variance is said to be summable of order δ, represented as $S(\delta)$, if

$$S_T = T^{-\left(\frac{1}{2}+\delta\right)} L(T) \sum_{t=1}^{T} (Y_t - m_t) = O_p(1) \text{ as } n \to \infty, \tag{14.44}$$

where δ is the minimum real number that makes S_T bounded in probability, m_t is a deterministic sequence, and $L(T)$ is a slowly varying function as defined by Eq. (14.2) in Section 2.

The above definition is defined according to the limit result of the sample average of $(Y_t - m_t)$, where m_t is not explicitly defined in BG (2013). One reasonable possibility is to select m_t that corresponds to a smallest possible δ. We provide some examples to illustrate that the concept needs to be further elaborated via population properties of $\{Y_t\}$, or population moment conditions. Apparently, if $\{Y_t\}$ is an I(1) process with a zero mean and finite Var(ΔY_t), Y_t is $S(1)$ with $m_t = 0$. However, suppose that $\{Y_t\}$ is a random walk process with a nonzero drift generated from a model, $Y_t = \mu + Y_{t-1} + u_t$ with $Y_0 \equiv 0$, $\mu \neq 0$, and $u_t \sim$ i.i.d. $\left(0,\sigma_u^2\right)$ independent of all the past X_s' $(s < t)$. For this process, taking $m_t \equiv 0$ yields $Y_t \sim S(1.5)$, while taking $m_t \equiv \mu t$ gives $Y_t \sim S(1)$. Without explaining what m_t is, especially in a nonlinear setup, the uniqueness of δ is of question. Let $Z_t = Y_t^2$ and $X_t = \sum_{i=1}^{t} u_i$. Then, $Z_t = (\mu t + X_t)^2 = \mu^2 t^2 + 2\mu t X_t + X_t^2$. With $m_t = 0$, Z_t is $S(2.5)$; with $m_t = \mu^2 t^2$, Z_t is $S(2)$; with $m_t = E(Z_t) = \mu^2 t^2 + \sigma_u^2 t$, Z_t is $S(2)$. Again, the choice of the deterministic term, m_t, affects the value of δ. For general nonlinear transformation of an I(1) process, the discussion given in Granger, Inoue, and Morin (1997) can be useful in determining the dominant deterministic trend, m_t, before determining the value of δ. To highlight the potential usage of BG's (2013) definition, from this point on, we assume that all the I(1) variables to be discussed have a zero mean and finite increment variance.

In a linear setup, Engle and Granger (1987) defined a general cointegration relation as follows. Let Y_t be a $k \times 1$ $(k \geq 2)$ vector of $I(d)$ processes. The components of Y_t are cointegrated of order (d, b) with $d \geq b$ or $C(d, b)$, if there exists at least one $k \times 1$ nonzero constant vector, α, such that $\alpha^T Y_t$ is $I(d-b)$. Parallel to Engle and Granger's cointegrating concept, BG (2013) extends the cointegrating relation to the nonlinear setup based on definition (14.44):

Definition. Two summable stochastic processes, $Y_t \sim S\left(\delta_y\right)$ and $X_t \sim S(\delta_x)$, will be said to be *co-summable* if there exists $g\left(X_t,\theta_g\right) \sim S\left(\delta_g\right)$ such that $u_t = Y_t - g\left(x,\theta_g\right)$ is $S(\delta_u)$, with $\delta_u = \delta_y - \delta$ and $\delta > 0$. In short, $\left(Y_t, g\left(x,\theta_g\right)\right) \sim CS\left(\delta_y,\delta\right)$.

Of course, one can replace the parametric function $g\left(\cdot,\theta_g\right)$ with a general nonlinear function. Because in practice, δ_y, δ_x, and δ are unknown, BG (2013) proposed a consistent estimator with a slow convergence rate of $\left[\log(T)\right]^{-1}$. Now, we will apply BG's co-summability concepts to model (14.33) in Section 14.5.2 and model (14.34) in Section 14.5.3. For easy reference, we use δ_g to refer to the limit order of $T^{-\left(\frac{1}{2}+\delta_g\right)} L(T) \sum_{t=1}^{T} X_t^T g(Z_t)$.

Consider model (14.33) first. As the error term is I(0), we have $\delta_u = 0$. Cai et al. (2009) considered two cases: (i) $\{Z_t\}$ is I(0) and $X_t = \left(X_{1t}^T, X_{2t}^T\right)^T$ with $\{X_{1t}\}$ being I(0) and $\{X_{2t}\}$ being I(1); (ii) $\{Z_t\}$ is I(1) and $\{X_t\}$ is I(0). Partition $g(Z_t) = \left[g_1^T(Z_t), g_2^T(Z_t)\right]^T$ conformably with respect to $X_t = \left(X_{1t}^T, X_{2t}^T\right)^T$. For case (i), $\sum_{t=1}^{T} X_t^T g(Z_t)$ is dominated by $\sum_{t=1}^{T} X_{2t}^T g_2(Z_t)$, and we obtain $\delta_g = 0.5$ if $E\left[g_2(Z_t)\right] = 0$ and $\delta_g = 1$ if $E\left[g_2(Z_t)\right] \neq 0$, applying Hansen (1992, Theorems 3.1 and 3.3). Therefore, Y_t is $S(0.5)$ or $S(1)$, depending on whether $E\left[g_2(Z_t)\right] = 0$ or not. Therefore, model (14.33) in case (i) defines a co-summable relation $\left(Y_t, X_t^T g(Z_t)\right) \sim CS(0.5, 0.5)$ if $E\left[g_2(Z_t)\right] = 0$ or $\left(Y_t, X_t^T g(Z_t)\right) \sim CS(1,1)$ if $E\left[g_2(Z_t)\right] \neq 0$.

For case (ii), the limit result of $\sum_{t=1}^{T} X_t^T g(Z_t)$ is determined by the functional form $g(\cdot)$ and whether or not $E(X_t) = 0$. Equations (14.13)–(14.16) given in Section 14.3.2 give the limit results of $\sum_{t=1}^{T} g(Z_t)$ when $g(\cdot)$ belongs to class $\mathcal{T}(I)$, $\mathcal{T}(H)$, and $\mathcal{T}(E)$, respectively. Because $\sum_{t=1}^{T} X_t^T g(Z_t) = \sum_{t=1}^{T} E(X_t)^T g(Z_t) + \sum_{t=1}^{T} [X_t - E(X_t)]^T g(Z_t)$, under some regularity conditions, we can show that the second term is dominated by the first term if $g(\cdot)$ is in class $\mathcal{T}(H)$ and $\mathcal{T}(E)$ and $E(X_t) \neq 0$. Also, by Eq. (14.13), when $g(\cdot)$ is in class $\mathcal{T}(I)$ (say, $g(z) = \sin(z)$) and $\{X_t\}$ has a zero mean and independent of $\{Z_t\}$, $\delta_g = \delta_y = 0$. So, one will not discuss co-summability at all.

Sun et al. (2013) considered model (14.33) when both $\{X_t\}$ and $\{Z_t\}$ are I(1) processes with a zero mean, but $\{u_t\}$ is an I(0) process. Again, we have $\delta_u = 0$. We discuss the case that $\int g(x)dx \neq 0$ below. To determine $\delta_y = \delta_g$, we will apply Phillips' (2009) limit result (Eq. (14.A.6) in the Appendix) to $\sum_{t=1}^{T} X_t^T g(Z_t)$ and obtain

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{X_t^T}{\sqrt{T}} g(Z_t) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \frac{X_t^T}{\sqrt{T}} g\left(\sqrt{T} \frac{Z_t}{\sqrt{T}}\right) \xrightarrow{d} \left(\int g(x)dx\right) \int_0^1 B_x(p)\, dL_{B_z}(p,0) \tag{14.45}$$

where $\left(T^{-1/2} X_{[Tr]}, T^{-1/2} Z_{[Tr]}\right) \Longrightarrow (B_x(r), B_z(r))$, a bivariate Brownian process with a zero mean and finite positive definite variance–covariance matrix, for all $r \in [0,1]$, $L_{B_z}(t,s)$ is the local time process of B_z, and $g(x)$ and $g^2(x)$ are Lebesgue integrable functions on R with $\int g(x)dx \neq 0$. Hence, with $g(\cdot)$ belonging to class $\mathcal{T}(I)$, then $\sum_{t=1}^{T} X_t^T g(Z_t) = O_e(T)$ and $\delta_y = \delta_g = 0.5$. If $g(\cdot)$ belongs to class $\mathcal{T}(H)$ with $g(\lambda z) = v(\lambda)H(z) + R(z,\lambda)$, where H is locally integrable and $R(z,\lambda)$ is asymptotically dominated by $v(\lambda)H(z)$ when $\lambda \to \infty$ and/or $|z| \to \infty$, the dominant term of $T^{-1} \sum_{t=1}^{T} \left(T^{-1/2} X_t\right)^T g(Z_t)$ will be $v\left(\sqrt{T}\right) T^{-1} \sum_{t=1}^{T} \left(T^{-1/2} X_t\right)^T H\left(T^{-1/2} Z_t\right) = O_e\left(v\left(\sqrt{T}\right)\right)$, so $\delta_y = \delta_g > 0.5$ is determined by $v\left(\sqrt{T}\right)$. If $\beta(z) = z^2$, then $v\left(\sqrt{T}\right) = T$ and $\delta_y = \delta_g = 2$, and $\{Y_t\}$ in this case has the same order as a squared random walk process (without drift) that we analyzed in Section 14.2. Model (14.34) is studied by Sun, Hsiao, and Li (2011), where $\delta_u = 1$ as $\{u_t\}$ is an I(1) process with a zero mean. In this model, $\{X_t\}$ is an I(1) process, but $\{Z_t\}$ is an I(0) process. As discussed for model (14.33) under case (i), we have $\delta_g = 0.5$ if $E\left[g(Z_t)\right] = 0$ and $\delta_g = 1$ if $E\left[g(Z_t)\right] \neq 0$. Therefore, if $E\left[g(Z_t)\right] \neq 0$, we have $\delta_y = \delta_g = \delta_u = 1$, and model (14.34) represents no co-summability. However, when $E\left[g(Z_t)\right] = 0$, we have $\delta_g = 0.5$, so that $\left(Y_t, X_t^T g(Z_t)\right)$ is not co-summable, either.

14.8. Conclusion

When a model contains a nonlinear transformation of an integrated time series, the kernel-based nonparametric estimator is still consistent as long as the model is not spurious, but the convergence rate of the "asymptotic variance" of the kernel estimator is reduced by a factor of $\sqrt{T}$ when the unknown curve is a function of an integrated series, compared to the case of weakly dependent and independent data. In addition, applying the local linear regression approach can be more beneficial than applying the local constant (or kernel) estimator. Moreover, the optimal bandwidth chosen via the cross-validatory method is random even asymptotically. This is in sharp contrast to the weakly dependent and independent data case where it is known that the CV-selected optimal smoothing parameters, after multiplying by some normalization constant, converge to nonstochastic constants.

Finally, unless the error terms in models (14.24) and (14.32), for example, are homescedastic, the conditional mean regression model does not provide the full picture of how the dependent variable responds to the changes of the regressors. Also, in the presence of outliers or fat-tailed data distribution, estimating the conditional mean regression model can result small sample biases. In such circumstances, conditional quantile regression models are found to be nice alternatives. Since Koenker and Bassett (1978, 1982) derived estimation and hypothesis tests for linear quantile regression models, quantile regression models have become widely used in various disciplines. Koenker and Xiao (2004) extended quantile estimation techniques to unit root quantile autoregressive (QAR) models and provided unit root test statistics and Xiao (2009a) estimated linear quantile cointegrating model. Cai (2002) provided limit result for kernel estimation of nonparametric quantile regression models of stationary time series, and Cai and Xu (2009) considered nonparametric quantile estimations of dynamic functional-coefficient models for stationary time series. So far, for null recurrent time series, Lin, Li, and Chen's (2009) local linear M-estimator of nonparametric cointegrating model (14.24) can be applied to quantile regression, but we are not aware any literature on estimating non-/semiparametric quantile cointegrating models with nonlinearly transformed integrated covariates.

Acknowledgments

We would like to thank the three editors, Jeff Racine, Liangjun Su, and Aman Ullah, and the anonymous referee for their constructive comments and suggestions. Yiguo Sun acknowledges the financial support from the Social Science and Humanities Research Council of Canada (grant 410-2009-0109). Qi Li thanks National Natural Science Foundation of China (project # 71133001 and project # 71131008) for partially financial supports.

Appendix: Some Results on Kernel Estimators with I(1) Data

When integrated time series data are considered, the development of both consistent estimators and test statistics, in general, heavily relies on the development of proper functional central limit theorems (FCLTs), limiting results of sample averages and covariances of nonlinearly transformed (scaled) integrated time series. This section summarizes some of the recently developed technical results for nonlinearly transformed integrated time series. Billingsley (1999), Davidson (1994), Jacod and Shiryaev (2003), and Revuz and Yor (2005) are among excellent references in explaining continuous martingales, Brownian motions, functional central limit theorems, convergence of stochastic integrals, and so on.

Suppose that one observes nonstationary time series $\{(Y_t, X_t)\}_{t=1}^T$. Assume that there exist continuous stochastic processes $\big(G_y(r), G_x(r)\big)$ for which the weak convergence result

$$\left(d_T^{-1} Y_{[Tr]}, d_T^{-1} X_{[Tr]}\right) \Longrightarrow \big(G_y(r), G_x(r)\big) \tag{14.A.1}$$

holds with respect to the Skorohod topology on $D[0,1]^2$ for all $r \in [0,1]$, where d_T is a certain sequence of positive numbers with $d_T \to \infty$ as $T \to \infty$. For an I(1) process with a zero mean, $d_T = \sqrt{T}$. In addition, the process $G_x(\cdot)$ has a continuous local time process (see Revuz and Yor (2005, Chapter 7) for details)

$$L_G(t,s) = \lim_{\epsilon \to 0} \frac{1}{2\epsilon} \int_0^t I\{|G_x(r) - s| < \epsilon\} dr \tag{14.A.2}$$

for $s \in R$ and $t \in [0,1]$, where $L_G(t,s)$ is a local time that the process $G_x(r)$ stays around a point s over the time interval $[0,t]$.

Under different regularity conditions, Jeganathan (2004) and Wang and Phillips (2009a) have shown that, for any $c_T \to \infty$, $c_T/T \to 0$, as $T \to \infty$, and $r \in [0,1]$, we obtain

$$\frac{c_T}{T} \sum_{t=1}^{[Tr]} g\left(c_T\left(\frac{X_t}{d_T} + x\right)\right) \xrightarrow{d} \left(\int_{-\infty}^{\infty} g(x)dx\right) L_{G_x}(t, -x). \tag{14.A.3}$$

Jeganathan (2004) allowed $\{X_t\}$ to be a fractional ARIMA process with possibly heavy-tailed innovations, and Wang and Phillips (2009a) assumed $\{X_t\}$ to be an I(1) or a near I(1) process. Park and Phillips (1999), Pötscher (2004), and de Jong (2004) studied the case with $d_T = \sqrt{T}$, $c_T = 1$, and $x = 0$. For integrated time series $\{X_t\}$, Kasparis and Phillips (2012, Proposition A, p. 19) showed that under certain conditions for $r \neq s$ with finite $|r - s|$ we have

$$\frac{c_T}{T} \sum_{k=1}^{[Tr]} g(X_{t-r}) K\left(c_T\left(\frac{X_{t-s} - x}{\sqrt{T}}\right)\right) \xrightarrow{d} E\left[g\left(x + \sum_{rs} v_i\right)\right] \int_{-\infty}^{\infty} K(s) ds L_{G_x}(r, 0). \tag{14.A.4}$$

Moreover, complementing to (14.A.3), Wang and Phillips (2011, Theorem 2.1) derived the limit result when $\int_{-\infty}^{\infty} g(x)dx = 0$; and under certain conditions, they showed that

$$\sqrt{\frac{c_T}{T}}\sum_{t=1}^{[Tr]} g\left(c_T\frac{X_t}{d_T}\right) \Longrightarrow \left(\int g^2(x)\,dx\right)^2 N\sqrt{L_{G_x}(r,0)}, \tag{14.A.5}$$

where N is a standard normal random variate independent of the local time process $L_{G_x}(r,0)$ for $r \in [0,1]$.

Phillips (2009, Theorem 1) derived a limiting result for the sample covariance, for a positive sequence $c_T \to \infty$, $c_T/T \to 0$, and $r \in [0,1]$:

$$\frac{c_T}{T}\sum_{t=1}^{[Tr]} \frac{Y_t}{d_T} g\left(c_T\frac{X_t}{d_T}\right) \Longrightarrow \left(\int g(x)\,dx\right)\int_0^r G_y(p)\,dL_{G_x}(p,0), \tag{14.A.6}$$

where c_T is a certain sequence of positive numbers, and $g(x)$ and $g^2(x)$ are Lebesgue integrable functions on R with $\int g(x)\,dx \neq 0$. And, if Eq. (14.A.1) can be strengthened to strong convergence result, we have

$$\sup_{r\in[0,1]}\left|\frac{c_T}{T}\sum_{t=1}^{[Tr]} \frac{Y_t}{d_T} g\left(c_T\frac{X_t}{d_T}\right) - \int g(x)\,dx\int_0^r G_y(p)\,dL_{G_x}(p,0)\right| \xrightarrow{p} 0. \tag{14.A.7}$$

References

Aparicio, F. M., and A. Escribano. 1998. "Information-Theoretic Analysis of Serial Dependence and Cointegration." *Studies in Nonlinear Dynamics & Econometrics*, **3**, pp. 119–140.

Athreya, K. B., and S. G. Pantula. 1986. "A Note on Strong Mixing of ARMA Processes." *Statistics & Probability Letters*, **4**, pp. 187–190.

Balke, N. S., and T. B. Fomby. 1997. "Threshold cointegration." *International Economic Review*, **38**, pp. 627–645.

Berenguer-Rico, V., and J. Gonzalo. 2013. "Summability of stochastic processes: A generalization of integration for non-linear processes" Forthcoming to Journal of Econometrics.

Berkes, I., and L. Horváth. 2006. "Convergence of integral Functionals of Stochastic Processes." *Econometric Theory*, **22**, pp. 304–322.

Billingsley, P. 1999. *Convergence of Probability Measures*, second edition. Toronto: John Wiley & Sons.

Breitung, J. 2001. "Rank Tests for Nonlinear Cointegration." *Journal of Business and Economic Statistics*, **19**, pp. 331–340.

Brockwell, P. J., and R. A. Davis. 1991. *Time Series: Theory and Methods*, second edition. Springer Series in Statistics. New York: Springer.

Cai, Z. 2002. "Regression Quantiles for Time Series." *Econometric Theory*, **18**, pp. 169–192.

Cai, Z. 2007. "Trending Time Varying Coefficient Time Series Models with Serially Correlated Errors." *Journal of Econometrics*, **137**, pp. 163–188.

Cai, Z., J. Fan, and Q. Yao. 2000. "Functional-Coefficient Regression Models for Nonlinear Time Series." *Journal of the American Statistical Association*, **95**, pp. 941–956.

Cai, Z., Q. Li, and J. Y. Park. 2009. “Functional-Coefficient Models for Nonstationary Time Series Data.” *Journal of Econometrics*, **148**, pp. 101–113.

Cai, Z., and X. Xu. 2009. “Nonparametric Quantile Estimations for Dynamic Smooth Coefficient Models.” *Journal of the American Statistical Association*, **104**, pp. 371–383.

Caner, M., and B. E. Hansen. 2001. “Threshold Autoregression with a Unit Root.” *Econometrica*, **69**, pp. 1555–1596.

Chan, K. S. 1993. “Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model.” *The Annals of Statistics*, **21**, pp. 521–533.

Chan, K. S., and R. S. Tsay. 1998. “Limiting Properties of the Least Squares Estimator of a Continuous Threshold Autoregressive Model.” *Biometrika*, **85**, pp. 413–426.

Chang, Y., and J. Y. Park. 2003. “Index Models with Integrated Time Series.” *Journal of Econometrics*, **114**, pp. 73–106.

Chang, Y., J. Y. Park, and P. C. B. Phillips. 2001. “Nonlinear Econometrics Models with Cointegrated and Deterministically Trending Regressors.” *Econometrics Journal*, **4**, pp. 1–36.

Chen, J., J. Gao, and D. Li. 2012 . “Estimation in Semiparametric Null-Recurrent Time Series.” *Bernoulli*, **18**, pp. 678–702.

Chen, B., and Y. Hong. 2012. “Testing for Smooth Structural Changes in Time Series Models via Nonparametric Regression. ” *Econometrica*, **80**, pp. 1157–1183.

Chen, R., and R. S. Tsay. 1993. “Functional-Coefficient Autoregressive Models.” *Journal of the American Statistical Association*, **88**, pp. 298–308.

Choi, I., and P. Saikkonen. 2004. “Testing Linearity in Cointegrating Smooth Transition Regressions.” *Econometrics Journal*, **7**, pp. 341–365.

Choi, I., and P. Saikkonen. 2010. “Testing for Nonlinear Cointegration.” *Econometrics Theory*, **26**, pp. 682–709.

Christopeit, N. 2009. “Weak Convergence of Nonlinear Transformations of Integrated Processes: The Multivariate Case.” *Econometric Theory*, **25**, pp. 1180–1207.

Cline, D. B. H., and H. H. Pu. 1999. “Stability of Nonlinear AR(1) Time Series with Delay.” *Stochastic Processes and Their Applications*, **82**, pp. 307–333.

Davidson, J. 1994. *Stochastic Limit Theory*. New York: Oxford University Press.

Davidson, J. 2002. “Establishing Conditions for the Functional Central Limit Theorem in Nonlinear and Semiparametric Time Series Processes.” *Journal of Econometrics*, **106**, pp. 243–269.

Davidson, J. 2009. “When Is a Time Series I(0)?” Chapter 13 In *The Methodology and Practice of Econometrics*, eds. Jennifer Castle and Neil Shephard. New York: Oxford University Press, pp. 322–342.

de Jong, R. M. 2002. “Nonlinear Estimators with Integrated Regressors but Without Exogeneity.” Mimeo, Ohio State University.

de Jong, R. M. 2004. “Addendum to ‘Asymptotics for Nonlinear Transformation of Integrated Time Series.’ ” *Econometric Theory*, **21**, pp. 623–635.

de Jong, R. M., and J. Davidson. 2000. “The Functional Central Limit Theorem and Weak Convergence to Stochastic Integrals I: Weakly Dependent Processes.” *Econometric Theory*, **16**, pp. 621–642.

de Jong, R., and C. Wang. 2005. “Further Results on the Asymptotics for Nonlinear Transformations of Integrated Time Series.” *Econometric Theory*, **21**, pp. 413–430.

Dhrymes, P. 1986. “Limited Dependent Variables.” Chapter 27 In *Handbook of Econometrics* 3, eds. Z. Griliches and M. D. Intriligator. Amsterdam: North Holland.

Dickey, D. A., and W. A. Fuller. 1979. "Distribution of Estimators for Autoregressive Time Series with a Unit Root." *Journal of American Statistics Association*, **74**, pp. 427–431.

Dufrénot, G., and V. Mignon. 2002. *Recent Developments in Nonlinear Cointegration with Applications to Macroeconomics and Finance*. Dordrecht: Kluwer Academic Publishers.

Engle, R. F., and C. W. J. Granger. 1987. *Long-Run Economic Relationships: Readings in Co-integration*. New York: Oxford University Press.

Escribano, A., and S. Mira. 2002. "Nonlinear Error Correction Models." *Journal of Time Series Analysis*, **23**, pp. 509–522.

Fan, J., M. Farmen, and I. Gijbels. 1998. "Local Maximum Likelihood Estimation and Inference." *Journal of the Royal Statistical Society*, Series B, **60**, pp. 591–608.

Fan, Y., and Q. Li. 1999. "Central Limit Theorem for Degenerate U-Statistics of Absolutely Regular Processes with Applications to Model Specification Testing." *Journal of Nonparametric Statistics*, **10**, pp. 245–271.

Fan, J., and Q. Yao. 1998. "Efficient Estimation of Conditional Variance Functions in Stochastic Regression." *Biometrika*, **85**, pp. 645–660.

Fan, J., and Q. Yao. 2003. *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Gallant, A. R., and H. White. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Basil Blackwell.

Gao, J. 2007. *Nonlinear Time Series: Semiparametric and Nonparametric Methods*. London: Chapman & Hall.

Gao, J., M. King, Z. Lu, and G. Tjøstheim. 2009a. "Specification Testing in Nonlinear and Nonstationary Time Series Autoregression." *The Annals of Statistics*, **37**, pp. 3893–3928.

Gao, J., M. King, Z. Lu, and G. Tjøstheim. 2009b. "Nonparametric Specification Testing for Nonlinear Time Series with Nonstationarity." *Econometrics Theory*, **25**, pp. 1869–1892.

Gao, J., and P. C. B. Phillips. 2011. "Semiparametric Estimation in Multivariate Nonstationary Time Series Models." Working paper available at http://ideas.repec.org/p/msh/ebswps/2011-17.html.

Gao, J., D. Tjøstheim, and J. Yin. 2011. "Estimation in Threshold Autoregression Models with a Stationary and a Unit Root Regime." *Journal of Econometrics*, **172**, pp. 1–13.

Godfrey, L. G. 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Econometric Society Monographs 16. New York: Cambridge University Press.

Granger, C. W. J. 1995. "Modelling Nonlinear Relationships Between Extended-Memory Variables." *Econometrica*, **63**, pp. 265–279.

Granger, C. W. J., and A. P. Anderson. 1978. *An Introduction to Bilinear Time Series*. Göttingen: Vandenhoeck and Ruprecht.

Granger, C. W. J., and J. Hallman. 1991. "Long memory series with Attractors." *Oxford Bulletin of Economics and Statistics*, **53**, pp. 11–26.

Granger, C. W. J., T. Inoue, and N. Morin. 1997. "Nonlinear Stochastic Trends." *Journal of Econometrics*, **81**, pp. 65–92.

Granger, C. W. J., and P. Newbold. 1974. "Spurious Regressions in Econometrics." *Journal of Econometrics*, **2**, pp. 111–120.

Granger, C. W. J., and N. R. Swanson. 1997. "An Introduction to Stochastic Unit-Root Processes." *Journal of Econometrics*, **80**, pp. 35–62.

Guerre, E., and H. R. Moon. 2002. "A Note on the Nonstationary Binary Choice Logit Model." *Economics Letters*, **76**, pp. 267–271.

Guerre, E., and H. R. Moon. 2006. "A Study of a Semiparametric Binary Choice Model with Integrated Covariates." *Econometric Theory*, **22**, pp. 721–742.

Hall, P. 1979. "On the Skorokhod Representation Approach to Martingale Invariance Principles." *The Annals of Probability*, **7**, pp. 371–376.

Hamilton, J. D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Hansen, B. E. 1992. "Convergence to Stochastic Integrals for Dependent Heterogeneous Processes." *Econometric Theory*, **8**, pp. 489–500.

Härdle, W., P. Hall, and H. Ichimura. 1993. "Optimal Smoothing in Single-Index Models." *The Annals of Statistics*, **21**, pp. 157–178.

Harvey, D. I., S. J. Leybourne, and B. Xiao. 2008. "A Powerful Test for Linearity When the Order of Integration Is Unknown." *Studies in Nonlinear Dynamics & Econometrics*, **12**, pp. 1–22.

Herndorf, N. 1984. "A Functional Central Limit Theorem for weakly Dependent Sequences of Random Variables." *The Annals of Probability*, **12**, pp. 141–153.

Hong, S. H., and P. C. B. Phillips. 2010. "Testing Linearity in Cointegrating Relations with an Application to Purchasing Power Parity." *Journal of Business & Economic Statistics*, **28**, pp. 96–114.

Horowitz, J. 1992. "A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica*, **60**, pp. 505–531.

Hu, L., and P. C. B. Phillips. 2004. "Nonstationary Discrete Choice." *Journal of Econometrics*, **120**, pp. 103–138.

Ichimura, H. 1993. "Semiparametric Least-Squares (SLS) and Weighted SLS Estimation of Single-Index Models." *Journal of Econometrics*, **58**, pp. 71–120.

Jacod, J., and A. N. Shiryaev. 2003. *Limit Theorems for Stochastic Processes*, second edition. New York: Springer-Verlag.

Jeganathan, P. 2004. "Convergence of Functionals of Sums of Random Variables to Local Times of Fractional Stable Motions." *The Annals of Probability*, **32**, pp. 1771–1795.

Juhl, T. 2005. "Functional-Coefficient Models Under Unit Root Behavior." *Econometrics Journal*, **8**, pp. 197–213.

Juhl, T., and Z. Xiao. 2005. "Testing for Cointegration Using Partially Linear Models." *Journal of Econometrics*, **124**, pp. 363–394.

Kallianpur, G., and H. Robbins. 1954. "The Sequence of Sums of Independent Random Variables." *Duke Mathematical Journal*, **21**, pp. 285–307.

Karlsen, H. A., T. Myklebust, and D. Tjøstheim. 2007. "Nonparametric Estimation in a Nonlinear Cointegration Type Model." *The Annals of Statistics*, **35**, pp. 252–299.

Karlsen, H., T. Myklebust, and D. Tjøstheim. 2010. "Nonparametric Regression Estimation in a Null Recurrent Time Series." *Journal of Statistical Planning and Inference*, **140**, pp. 3619–3626.

Karlsen, H., and D. Tjøstheim. 2001. "Nonparametric Estimation in Null Recurrent Time Series Models." *The Annals of Statistics*, **29**, pp. 372–416.

Kapetanios, G., Y. Shin, and A. Snell. 2006. "Testing for Cointegration in Nonlinear Smooth Transition Error Correction Models." *Econometric Theory*, **22**, pp. 279–303.

Kasparis, I. 2008. "Detection of Functional Form Misspecification in Cointegrating Relations." *Econometric Theory*, **24**, pp. 1373–1403.

Kasparis, I. 2010. "The Bierens Test for Certain Nonstationary Models." *Journal of Econometrics*, **158**, pp. 221–230.

Kasparis, I. 2011. "Functional Form Misspecification in Regressions with a Unit Root." *Econometric Theory*, **27**, pp. 285–311.

Kasparis, I., and P. C. B. Phillips. 2012. "Dynamic Misspecification in Nonparametric Cointegration." *Journal of Econometrics*, **170**, pp. 15–31.

Klein, R. W., and R. H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica*, **61**, pp. 387–421.

Koenker, R., and G. W. Bassett. 1978. "Regression Quantiles." *Econometrica*, **46**, pp. 33–50.

Koenker, R., and G. W. Bassett. 1982. "Robust Tests for Heteroscedasticity Based on Regression Quantiles." *Econometrica*, **50**, pp. 43–61.

Koenker, R., and Z. Xiao. 2004. "Unit Root Quantile Autoregression Inference." *Journal of the American Statistical Association*, **99**, pp. 775–787.

Kuelbs, J., and W. Philipp. 1980. "Almost Sure Invariance Principles for Partial Sums of Mixing B-Valued Random Variables." *The Annals of Probability*,**8**, pp. 1003–1036.

Lee, Y., T. Kim, and P. Newbold. 2005. "Spurious Nonlinear Regression in Econometrics." *Economics Letters*, **87**, pp. 301–306.

Li, Q., C. Huang, D. Li, and T. Fu. 2002. "Semiparametric smooth Coefficient Models." *Journal of Business and Economic Statistics*, **20**, pp. 412–422.

Li, Q., and J. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton, NJ: Princeton University Press.

Li, Q., and S. Wang. 1998. "A Simple Consistent Bootstrap Test for a Parametric Regression Functional Form." *Journal of Econometrics*, **87**, pp. 145–165.

Liang, Z., and Q. Li. 2012. "Functional Coefficient Regression Model with Time Trend." *Journal of Econometrics*, forthcoming.

Lin, Z., D. Li, and J. Chen. 2009. "Local Linear M-Estimators in Null Recurrent Time Series." *Statistica Sinica*, **19**, pp. 1683–1703.

Liu, Q., S. Ling, and Q. Shao. 2011. "On Non-stationary Threshold Autoregressive Models." *Bernoulli*, **17**, pp. 969–986.

Manski, C. 1985. "Semiparametric Analysis of Discrete Response." *Journal of Econometrics*, **27**, pp. 313–333.

Manski, C. 1988. "Identification of Binary Response Models." *Journal of American Statistical Association*, **83**, pp. 729–738.

Marmer, V. 2008. "Nonlinearity, Nonstationarity, and Spurious Forecasts." *Journal of Econometrics*, **142**, pp. 1–27.

McLeish, D. L. 1975. "Invariance Principles for Dependent Variables." *Zeitschrift fur Wahrscheinlichskeitstheorie und Verwandete Gebiete*, **32**, pp. 165–178.

Meyn, S., and R. Tweedie. 1993. *Markov Chains and Stochastic Stability*. New York: Springer.

Moon, H. R. 2004. "Maximum Score Estimation of a Nonstationary Binary Choice Model." *Journal of Econometrics*, **122**, 385–403.

Müller, U. K. 2008. "The Impossibility of Consistent Discrimination Between I(0) and I(1) Processes." *Econometric Theory*, **24**, pp. 616–630.

Myklebust, T., H. Karlsen, and D. Tjøstheim. 2011. "Null Recurrent Unit Root Processes." *Econometric Theory*, **27**, pp. 1–41.

Narayan, P. 2010. "Carbon Dioxide Emissions and Economic Growth: Panel Data Evidence from Developing Countries." *Energy Policy*, **38**, pp. 661–666.

Nicolau, J. 2002. "Stationary Processes that Look Like Random Walks: The Bounded Random Walk Process in Discrete and Continuous Time." *Econometric Theory*, **18**, pp. 99–118.

Nicholls, D. F., and B. G. Quinn. 1982. *Random Coefficient Autoregression Models: An Introduction.* Lecture Notes in Statistics 11. New York: Springer.

Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics.* New York: Cambridge University Press.

Park, J. Y., and S. B. Hahn. 1999. "Cointegrating Regressions with Time Varying Coefficients." *Econometric Theory*, **15**, pp. 664–703.

Park, J. Y., and P. C. B. Phillips. 1999. "Asymptotics for Nonlinear Transformations of Integrated Time Series." *Econometric Theory*, **15**, pp. 269–298.

Park, J., Phillips, P. C. B. 2000. "Nonstationary Binary Choice." *Econometrica*, **68**, pp. 1249–1280.

Park, J. Y., and P. C. B. Phillips. 2001. "Nonlinear Regressions with Integrated Time Series." *Econometrica*, **69**, pp. 117–161.

Pham, D.T., K. S. Chan, and H. Tong. 1991. "Strong Consistency of the Least Squares Estimator for a Nonergodic Threshold Autoregressive Model." *Statistic Sinica*, **1**, pp. 361–369.

Phillips, P. C. B. 1986. "Understanding Spurious Regressions in Econometrics." *Journal of Econometrics*, **33**, pp. 311–340.

Phillips, P. C. B. 2009. "Local Limit Theory and Spurious Nonparametric Regression." *Econometric Theory*, **25**, pp. 1466–1497.

Phillips, P. C. B., S. Jin, and L. Hu. 2007. "Nonstationary Discrete Choice: A Corrigendum and Addendum." *Journal of Econometrics*, **141**, pp. 1115–1130.

Phillips, P. C. B., and V. Solo. 1992. "Asymptotics for Linear Processes." *The Annals of Statistics*, **20**, pp. 971–1001.

Pötscher, B. M. 2004. "Nonlinear Functions and Convergence to Brownian Motion: Beyond the Continuous Mapping Theorem." *Econometric Theory*, **20**, pp. 1–22.

Ramsey, J. B. 1969. "Tests for Specification Errors in classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society* (Series B) **31**, pp. 350–371.

Revuz, D., and M. Yor. 2005. *Continuous Martingales and Brownian Motion*, third edition. Fundamental Principles of Mathematical Sciences 293. New York: Springer-Verlag.

Robinson, P. M. 1989. "Nonparametric Estimation of Time-Varying Parameters." In *Statistical Analysis and Forecasting of Economic Structural Change*, ed. Hackl, P. Berlin: Springer, pp. 164–253.

Saikkonen, P. 2005. "Stability Results for Nonlinear Error Correction Models." *Journal of Econometrics*, **127**, pp. 69–81.

Saikkonen, P., and I. Choi. 2004. "Cointegrating Smooth Transition Regressions." *Econometric Theory*, **20**, pp. 301–340.

Stock, J. H. 1994. "Deciding Between I(1) and I(0)." *Journal of Econometrics*, **63**, pp. 105–131.

Subba Rao, T. 1981. "On the Theory of Bilinear Time Series Models." *Journal of the Royal Statistical Society*, Series B, **43**, pp. 244–255.

Sun, Y., Z. Cai, and Q. Li. 2012. "Consistent Nonparametric Test on Semiparametric Smooth Coefficient Models with Integrated Time Series." manuscript.

Sun, Y., Z. Cai, and Q. Li. 2013. "Functional Coefficient Models with Integrated Covariates." *Econometric Theory*, **29**, pp. 659–672.

Sun, Y., C. Hsiao, and Q. Li. 2011. "Measuring Correlations of Integrated but Not Cointegrated Variables: A Semiparametric Approach." *Journal of Econometrics*, **164**, pp. 252–267.

Sun, Y., C. Hsiao, and Q. Li. 2012. "Volatility Spillover Effect: A Semiparametric Analysis of Non-Cointegrated Process." *Econometrics Review*, forthcoming.

Sun, Y., and Q. Li. 2011. "Data-Driven Bandwidth Selection for Nonstationary Semiparametric Models." *Journal of Business and Economic Statistics*, **29**, pp. 541–551.

Teräsvirta, T., D. Tjøstheim, and C. W. J. Granger. 2010. *Modelling Nonlinear Economic Time Series*. Oxford, UK: Oxford University Press.

Tjøstheim, D. 1990. "Non-linear Time Series and Markov Chains." *Advances in Applied Probability*, **22**, pp. 587–611.

Tong, H., and K. S. Lim. 1980. "Threshold Autoregression, Limit Cycles and Cyclical Data (with Discussion)." *Journal of the Royal Statistical Society*, Series B, **42**, pp. 245–292.

Tong, H. 1990. *Nonlinear Time Series: A Dynamic Approach*. Oxford, UK: Oxford University Press.

van Dijk, D., T. Teräsvirta, and P. H. Franses. 2002. "Smooth Transition Autoregressive Models—A Survey of Recent Developments." *Econometric Reviews*, **21**, pp. 1–47.

Wang, Q., and P. C. B. Phillips. 2009a. "Asymptotic Theory for Local Time Density Estimation and Nonparametric Cointegrating Regression." *Econometric Theory*, **25**, pp. 710–738.

Wang, Q., and P. C. B. Phillips. 2009b. "Structural Nonparametric Cointegrating Regression." *Econometrica*, **77**, pp. 1901–1948.

Wang, Q., and P. C. B. Phillips. 2011. "Asymptotic Theory for Zero Energy Functionals with Nonparametric Regression Application." *Econometric Theory*, **27**, pp. 235–259.

Wang, Q., and P. C. B. Phillips. 2012. "Specification Testing for Nonlinear Cointegrating Regression." *The Annals of Statistics*, **40**, 727–758.

White, H., and I. Domowitz. 1984. "Nonlinear Regression Wit Dependent Observations." *Econometrica*, **52**, pp. 143–162.

Withers, C. S. 1981. "Conditions for Linear Processes to be Strong-Mixing." *Z. Wahrscheinlichkeitstheorie View, Gebiete*, **57**, pp. 477–480.

Wooldridge, M. 1994. "Estimation and Inference for Dependent Processes." Chapter 45 In *Handbook of Econometrics* 4, eds. R. F. Engle and D. L. McFadden. Amsterdam: North Holland.

Wooldridge, J. M., and H. White. 1988. "Some Invariance Principles and Central Limit Theorems for Dependent Heterogeneous Processes." *Econometric Theory*, **4**, pp. 210–230.

Xiao, Z. 2009a. "Quantile Cointegration Regression." *Journal of Econometrics*, **150**, pp. 248–260.

Xiao, Z. 2009b. "Functional Coefficient Co-integration Models." *Journal of Econometrics*, **152**, pp. 81–92.

Zeevi, A., and P. W. Glynn. 2004. "Recurrence Properties of Autoregressive Processes with Super-Heavy-Tailed Innovations." *Journal of Applied Probability*, **41**, pp. 639–653.

Zheng, J. X. 1996. "A Consistent Test of Functional Form Via Nonparametric Estimation Techniques." *Journal of Econometrics*, **75**, pp. 263–289.

PART VI

CROSS SECTION

CHAPTER 15

NONPARAMETRIC AND SEMIPARAMETRIC ESTIMATION OF A SET OF REGRESSION EQUATIONS

AMAN ULLAH AND YUN WANG

15.1. Introduction

It is well known that the weighted least squares (WLS), also known as the generalized least squares (GLS) estimator in a parametric regression model with a known nonscalar covariance matrix of errors, is the best linear unbiased estimator. This also holds asymptotically for an operational WLS estimator in which the nonscalar covariance matrix is replaced by a consistent estimator (see Greene (2007, p. 157) and Hayashi (2000, p. 138)). Further, in small samples it is known to be unbiased for the symmetric errors (see Kakwani (1967)), and its efficiency properties are analyzed in Taylor (1977). In the case of a single-equation nonparametric regression model with a nonscalar covariance, various local linear weighted least squares (LLWLS) estimators have been developed for the pointwise local linear regression and its derivative estimators (see Welsh and Yee (2006), Ullah and Roy (1998), Henderson and Ullah (2005, 2012), and Lin and Carroll (2000), among others). However, it has been shown in Henderson and Ullah (2012), Welsh and Yee (2006), and Lin and Carroll (2000), among others, that such LLWLS estimators may not be efficient even when the covariance matrix is known. In fact, often they are even beaten by the local linear least squares (LLLS) estimator ignoring the existence of a nonscalar covariance matrix. In view of this, Ruckstuhl, Welsh, and Carroll (2000) proposed a two-step estimator to a nonparametric model in which the dependent variable suitably filtered and the nonscalar covariance matrix is transformed to be a scalar covariance matrix (also see Su and Ullah (2007)). Martins-Filho and Yao (2009) proposed a two-step estimator with another filtered dependent variable but with a nonscalar covariance matrix consisting of heteroscedasticity (also see You, Xie, and Zhou (2007) for a similar estimator). Su, Ullah, and Wang (2013) then suggested a new two-step estimator in which the filtered dependent variable is different from that of Martins-Filho and Yao (2009), but with a scalar covariance matrix. They showed that their two-step estimator is asymptotically more efficient than both the LLLS and the two-step estimator proposed by Martins-Filho and Yao (2009). In a simulation study they also show that their two-step estimator is also more efficient in small samples compared to both the LLLS and the Martins-Filho and Yao's two-step estimator.

In this chapter we consider a set of regression equations (SRE) models. As we know, the SRE models have been extensively studied in parametric framework and widely used in empirical economic analysis, such as the wage determinations for different industries, a system of consumer demand equations, capital asset pricing models, and so on. However, it hasn't been well developed within the nonparametric estimation framework, although see, for example, Smith and Kohn (2000) and Koop, Poirier, and Tobias (2005), where nonparametric Bayesian methods are used to estimate multiple equations, Wang, Guo, and Brown (2000), where a penalized spline estimation method is considered, and Welsh and Yee (2006), where LLWLS estimators are used.

The objective of this chapter is to systematically develop a set of estimation results for SRE regression analysis within nonparametric and semiparametric framework. Specifically, we explore conventional LLLS and LLWLS in nonparametric SRE, and we develop efficient two-step estimation for various nonparametric and semiparametric SRE models following Su, Ullah, and Wang (2013) estimation results in the context of the single-equation model. The models considered include the partially linear semiparametric model, the additive nonparametric model, the varying coefficient model, and the model with endogeneity.

The structure of this chapter is as follows. In section 15.2, we introduce SRE nonparametric estimators including an LLLS estimator, a general two-step estimator, and various LLWLS estimators. In section 15.3 we propose the estimation procedures for a variety of nonparametric and semiparametric SRE models. Section 15.4 briefly discusses NP SRE models with conditional error covariance. Section 15.5 concludes.

15.2. Nonparametric Set of Regression Equations

We start with the following basic nonparametric set of regression equations (SRE)

$$y_{ij} = m_i(X_{ij}) + u_{ij}, \qquad i = 1,\ldots,M,\ j = 1,\ldots,N. \tag{15.1}$$

The economic variable y_{ij} is the jth observation on the ith cross-sectional unit, X_{ij} is the jth observation on the ith unit on a $q_i \times 1$ vector of exogenous regressors which may differ for different regression equations, $m_i(\cdot)$ is an unknown function form, which can differ across the cross-sectional units, and $E(u_{ij}|X_{ij})=0$. For simplicity, the equal number of observations N is assumed across M cross-section units.

The examples of such models are the cross-country economic growth model, the regional consumption model across clusters, the wage determination model for different industries, and a system of consumer demand equations, among others. In a special case, where $m_i(X_{ij})=X_{ij}\beta_i$, (15.1) is the standard Zellner's (1962) parametric seemingly unrelated regressions (SUR) system, in which $E(u_{ij}u_{i'j}|X_{ij},X_{i'j})=\sigma_{ii'}$ if $i\neq i'$ and it is σ_{ii} if $i=i'$. The system (15.1) is a VAR system in which X variables are lagged of y variables. When $m_i(X_{ij})=X_{ij}\beta_i$ with $q_i=q$ for all i and $\sigma_{ii'}=0$, we get the set of regression equations model (see Pesaran, Smith, and Im (1996)). Further, in the case where $m_i(X_{ij})=X_{ij}\beta$ with $q_i=q$, $\sigma_{ii}=\sigma^2$, and $\sigma_{ii'}=\rho\sigma^2$, we get a set of cluster regression equations. Also, if u_{ij} is treated as $\alpha_i+\epsilon_{ij}$ disturbances, we get the set of equations with error components.

15.2.1. Estimation with Unconditional Error Variance–Covariance Ω

15.2.1.1. Local Linear Least Squares (LLLS) Estimator

First, by first-order Taylor expansion, we obtain

$$\begin{aligned} y_{ij} &= m_i(X_{ij})+u_{ij} \\ &\simeq m_i(x_i)+(X_{ij}-x_i)m_i^{(1)}(x_i)+u_{ij} \\ &= X_{ij}(x_i)\delta_i(x_i)+u_{ij}, \end{aligned}$$

where $\delta_i(x_i)=\begin{pmatrix} m_i(x_i) & m_i^{(1)\prime}(x_i)\end{pmatrix}'$, which is a $(q_i+1)\times 1$ vector, and

$$X_{ij}(x_i)=\begin{pmatrix}1 & (X_{ij}-x_i)'\end{pmatrix}.$$

Let $y_i=(y_{i1},\ldots,y_{iN})'$, $X_i(x_i)=\begin{pmatrix}X_{i1}(x_i), & \ldots, & X_{iN}(x_i)\end{pmatrix}'$, which has a dimension of $N\times(q_i+1)$, and $u_i=\begin{pmatrix}u_{i1}, & \ldots, & u_{iN}\end{pmatrix}'$. In a vector representation, for each regression i, we can write

$$y_i\simeq X_i(x_i)\delta_i(x_i)+u_i,$$

which can be further written compactly as

$$\begin{aligned} \mathbf{y} &= \mathbf{m}(X)+\mathbf{u} \\ &\simeq X(x)\delta(x)+\mathbf{u}, \end{aligned} \tag{15.2}$$

where $\mathbf{y}=(y_1^{'},\ldots,y_M^{'})'$ is an $MN\times 1$ vector, $\mathbf{m}(X)=(\mathbf{m}_1(X_1),\ldots,\mathbf{m}_M(X_M))'$, $\mathbf{m}_i(X_i)=(m_i(X_{i1}),\ldots,m_i(X_{iN}))'$, $\mathbf{u}=(u_1^{'},\ldots,u_M^{'})'$, $X(x)=\text{diag}\begin{pmatrix}X_1(x_1), & \ldots, & X_M(x_M)\end{pmatrix}$, which has $MN\times(\Sigma_{i=1}^{M}q_i+M)$ dimension, and $\delta(x)=\begin{pmatrix}\delta_1(x_1), & \ldots, & \delta_M(x_M)\end{pmatrix}$, a

$(\sum_{i=1}^{M} q_i + M) \times 1$ vector. Further, we have $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_{MN\times 1}$ and $\Omega \equiv \mathrm{Var}(\mathbf{u}|\mathbf{X}) = \Omega(\theta)$ is an $MN \times MN$ unconditional covariance matrix, where θ is a vector of unknown parameters—for example, in the case of SUR $\Omega(\theta) = \Sigma \otimes I_N$, where Σ is an $M \times M$ matrix with typical diagonal element σ_{ii} and off-diagonal element $\sigma_{ii'}$ for i, $i' = 1, \ldots, M$.

Then the LLLS estimator of $\delta(x)$ is obtained by minimizing $\mathbf{u}'K(x)\mathbf{u}$,

$$\hat{\delta}(x) = (X'(x)K(x)X(x))^{-1}X'(x)K(x)\mathbf{y},$$

where $K(x) \equiv \mathrm{diag}\big(K_{h_1}(X_1 - x_1), \quad \ldots, \quad K_{h_M}(X_M - x_M)\big)$ is a $MN \times MN$ diagonal matrix, $K_{h_i}(X_i - x_i) \equiv \mathrm{diag}\big(K_{h_i}(X_{i1} - x_i), \quad \ldots, \quad K_{h_i}(X_{iN} - x_i)\big)$, and $K_{h_i}(X_{ij} - x_i) = \frac{1}{h_i}k(\frac{X_{ij}-x_i}{h_i})$. From the standard results on the asymptotic normality of the LLLS estimator (Li and Racine (2007)) in a single equation it is straightforward to show that $\hat{\delta}(x)$ is asymptotically normally distributed (Wang (2012) for the case of the SUR model).

15.2.1.2. Local Linear Weighted Least Squares (LLWLS) Estimator

Another class of local linear estimator in nonparametric literature is called an LLWLS estimator. By minimizing the following weighted sum of squared residuals

$$(\mathbf{y} - X(x)\delta(x))'\Phi_r(x)(\mathbf{y} - X(x)\delta(x)),$$

the LLWLS can be obtained as

$$\hat{\delta}_r(x) = (X'(x)\Phi_r(x)X(x))^{-1}X'(x)\Phi_r(x)\mathbf{y},$$

where $\Phi_r(x)$ is a weight matrix based on kernel smoothing function and covariance matrix of errors. For $r = 1,2,3,4$, $\Phi_1(x) = K^{1/2}(x)\Omega^{-1}K^{1/2}(x)$, $\Phi_2(x) = \Omega^{-1}K(x)$, $\Phi_3(x) = K(x)\Omega^{-1}$, and $\Phi_4(x) = \Omega^{-1/2}K(x)\Omega^{-1/2}$, respectively. $\Phi_1(x)$ and $\Phi_2(x)$ are given in Lin and Carroll (2000) for nonparametric panel data models with random effect, and $\Phi_4(x)$ is discussed in Ullah and Roy (1998) for random effect models.

Welsh and Yee (2006) give all these four types of LLWLS estimators, but only study the bias and variance of LLWLS estimator $\hat{\delta}_1(x)$ (not asymptotic distribution) with weight $\Phi_1(x)$ for a SUR with $M = 2$ for both unconditional and conditional variance–covariance (Section 15.4) of errors. For a single equation panel model, Henderson and Ullah (2012) compare the efficiency among LLWLS estimators and find that these LLWLS estimators do not always perform well, and sometimes they are even beaten by the LLLS estimator, which ignores weights. A possible reason for this is that LLWLS estimators described above are estimating regression and its derivative at a point $X_{ij} = x$ to the data which is already transformed by $K^{1/2}(x)\Omega^{-1/2}$ or $\Omega^{-1/2}K^{1/2}(x)$. Thus not too many local transformed data points are around x. In view of this, a two-step estimator, given in the next subsection, may be more appropriate.

However, we note that when Ω is a diagonal matrix, we have $W_1(x) = W_2(x) = W_3(x) = W_4(x)$. Thus, in this case the estimators $\hat{\delta}_r(x)$ for $r = 1, \ldots, 4$, are equivalent and they are equal to an LLLS estimator. In addition, if the equations have identical explanatory variables (i.e., $X_i = X_j$), then LLLS and LLWLS are identical. Both of these results correspond to the similar results found in the parametric SUR model.

15.2.1.3. Two-Step Estimator

In Section 15.2.1.1 we have observed that the LLWLS estimators do not tend to perform well always compared to LLLS. In view of this, Wang (2012) for a SUR case and Su, Ullah, and Wang (2013) and Martins-Filho and Yao (2009) for a single-equation case proposed the following two-step estimator to improve the estimation. The transformation required for the second step is made as follows:

$$\begin{aligned}
\mathbf{y} &= \mathbf{m}(X) + \mathbf{u}, \\
\Omega^{-1/2}\mathbf{y} + (\mathbf{H}^{-1} - \Omega^{-1/2})\mathbf{m}(X) &= \mathbf{H}^{-1}\mathbf{m}(X) + \Omega^{-1/2}\mathbf{u}, \\
\overrightarrow{\mathbf{y}} &= \mathbf{H}^{-1}\mathbf{m}(X) + \mathbf{v}, \\
&= \mathbf{H}^{-1}X(x)\delta(x) + \mathbf{v},
\end{aligned} \tag{15.3}$$

where $\overrightarrow{\mathbf{y}} \equiv \Omega^{-1/2}\mathbf{y} + (\mathbf{H}^{-1} - \Omega^{-1/2})\mathbf{m}(X)$, $\mathbf{v} \equiv \Omega^{-1/2}\mathbf{u}$, and we have used (15.2). It is obvious that the transformed errors are now independent and identically distributed.

Assume that $\Omega = PP'$ for some $MN \times MN$ matrix P. Let p_{ij} and v_{ij} denote the (i, j)th element of P and P^{-1}, respectively. Let $\mathbf{H} \equiv \text{diag}(v_{11}^{-1}, \ldots, v_{MNMN}^{-1})$ and $R^*(x) = \mathbf{H}^{-1}X(x)$, then by minimizing $\mathbf{v}'K(x)\mathbf{v}$ the two-step estimator would be

$$\hat{\delta}_{2\text{-}step}(x) = (R^{*\prime}(x)K(x)R^*(x))^{-1}R^{*\prime}(x)K(x)\overrightarrow{\mathbf{y}}. \tag{15.4}$$

The intuition behind this two-step estimator is that we are estimating, at a given x, $m_i(x_i)/v_{ii}$ instead of a combination of m functions. This may provide a better estimator of $m(x)$ from the data of x_{ij} close to x. Also, it is interesting to note that if the errors are uncorrelated across equations, and $K(x) \to K(0)$, the nonparametric two-step estimator $\hat{\delta}_{2\text{-}step}$ will become the parametric GLS estimator. Wang (2012) shows that $\hat{\delta}_{2\text{-}step}(x)$ is asymptotic normal.

Some of the special cases of the above two-step estimators are as follows. Martins-Filho and Yao (2009) considered the case where the two-step estimator is proposed based on premultiplying (15.3) on both sides by H. In this case the covariance matrix of Hv is H^2 and their estimator becomes inefficient compared to the above estimator (see Wang (2012)). Further, comparing the asymptotic covariance of $\hat{\delta}_{2\text{-}step}(x)$ with the one of $\widehat{\delta}(x)$, it is easy to see that $\hat{\delta}_{2\text{-}step}(x)$ is asymptotically more efficient than $\widehat{\delta}(x)$. Also, Ruckstuhl, Welsh, and Carroll (2000) and Su and Ullah (2007) considered a class of two-step estimator for nonparametric panel data models with random effects in which $\mathbf{H} = \tau I$ in (15.3) and (15.4). Note that $\mathbf{H} = \tau I$ implies that all the diagonal elements in $\Omega^{-1/2}$ contain identical information; that is, $v_{ii} = \tau^{-1}$ for $i = 1, \ldots, MN$. However, in (15.4), $\mathbf{H}$ can incorporate both heteroskedastic and correlation information in errors.

The two-step estimator described above is infeasible, since $\overrightarrow{\mathbf{y}}$ is unobservable, and Ω and $\mathbf{H}$ are unknown. In practice, $\Omega = \Omega(\theta)$ is replaced by its estimator $\hat{\Omega} = \Omega(\hat{\theta})$, where $\hat{\theta}$ is a consistent estimator of θ. For example, in the case of SUR equations, the operational two-step estimator can be written as follows. First, obtain a preliminary consistent estimator of m_i by first-order local polynomial smoothing y_{ij} on X_{ij} for each equation i. Denote $\hat{u}_{ij} = y_{ij} - \hat{m}_i(X_{ij})$. Second, we can obtain a consistent estimator of $\hat{\Omega}$, $\hat{H}$ by estimating

$$\hat{\sigma}_{ii'} = \frac{1}{N-1}\sum_{j=1}^{N}\left(\hat{u}_{ij} - \bar{\hat{u}}_{ij}\right)\left(\hat{u}_{i'j} - \bar{\hat{u}}_{i'j}\right),$$

$$\hat{\sigma}_{ii} = \frac{1}{N-1}\sum_{j=1}^{N}\left(\hat{u}_{ij} - \bar{\hat{u}}_{ij}\right)^2.$$

Further we can obtain the feasible $\overrightarrow{\mathbf{y}} = \hat{\Omega}^{-1/2}\mathbf{y} + (\hat{\mathbf{H}}^{-1} - \hat{\Omega}^{-1/2})\hat{\mathbf{m}}(X)$. Third, by first-order local polynomial smoothing feasible $\overrightarrow{\mathbf{y}}$ on X, obtain the two-step estimator $\hat{\delta}_{2\text{-}step}(x) = (R^{*\prime}(x)K(x)R^*(x))^{-1}R^{*\prime}(x)K(x)\overrightarrow{\mathbf{y}}$.

15.3. Alternative Specifications of NP/SP Set of Regressions

Up to now all estimators are discussed for the basic NP SRE models. In reality, we may have various specifications for the system—for example, a partially linear semiparametric model, a model with NP autocorrelated errors, an additive nonparametric model, a varying coefficient model, and a model with endogeneity. These models are well discussed in either cross-sectional or panel data framework. However, within the SRE system framework, they haven't been studied. So we discuss them below.

15.3.1. Partially Linear Semiparametric SRE Models

We consider the partially linear semiparametric set of regression equations

$$y_{ij} = m_i(X_{ij}) + Z_{ij}\gamma_i + u_{ij}, \qquad i = 1,\ldots,M,\ j = 1,\ldots,N, \tag{15.5}$$

where Z_{ij} is a vector of exogenous variables such that $E(u_{ij}|X_{ij}, Z_{ij}) = 0$, and the assumptions on errors remain the same as in (15.1). A method using profile least squares can be used to estimate such a model. For this we write

$$y^*_{ij} \equiv y_{ij} - Z_{ij}\gamma_i = m_i(X_{ij}) + u_{ij}, \tag{15.6}$$

or in a vector form

$$y^* = y - Z\gamma = m(X) + u, \tag{15.7}$$

where $Z = \text{diag}(Z_1, Z_2, \ldots, Z_M)$ and $\gamma = (\gamma_1', \ldots, \gamma_M')$.

For the two-step estimation this can be rewritten from (15.3) as

$$\overrightarrow{y}^* = H^{-1}X(x)\delta(x) + v, \tag{15.8}$$

where $\overrightarrow{y}^*$ is the same as $\overrightarrow{y}$ in (15.3) except that y is replaced by y^*. Then the two-step estimator of $\delta(x)$ is the same as the $\hat{\delta}_{2\text{-}step}(x)$ in (15.4) with $\overrightarrow{y}$ replaced by $\overrightarrow{y}^*$. However, it is not operational.

It follows that $\hat{m}_{2\text{-}step}(x) = [1\ 0]\hat{\delta}_{2\text{-}step}(x) = s(x)\overrightarrow{y}^*$ where

$$s(x) = [10](R^{*\prime}(x)K(x)R^*(x))^{-1}R^{*\prime}(x)K(x).$$

Thus $\hat{m}_{2\text{-}step}(x) = S\overrightarrow{y}^*$, where $S = (s(x_{11})', \ldots, s(x_{MN})')'$. Substituting this in (15.7), we can write

$$\begin{aligned} y &= \hat{m}_{2\text{-}step}(x) + Z\gamma + u \\ &= S[\Omega^{-1/2}(y - Z\gamma) + (H^{-1} - \Omega^{-1/2})m(x)] + Z\gamma + u \end{aligned}$$

or

$$\begin{aligned} [I - S\Omega^{-1/2}]y - S(H^{-1} - \Omega^{-1/2})m(x) &= [I - S\Omega^{-1/2}]Z\gamma + u \\ \bar{y} &= \bar{Z}\gamma + u. \end{aligned}$$

The GLS estimator of γ is

$$\hat{\gamma} = \left(\bar{Z}'\Omega^{-1}\bar{Z}\right)^{-1}\bar{Z}'\Omega^{-1}\bar{y}.$$

Then the operational estimator $\hat{\delta}_{2\text{-}step}(x)$ can be written by substituting γ by $\hat{\gamma}$. The asymptotic properties of both $\hat{\gamma}$ and $\hat{\delta}_{2\text{-}step}(x)$ can be developed by following Su and Ullah (2007).

Alternatively, we can estimate (15.5) using the idea of partial residual procedure by Clark(1977), Denby (1984), and Robinson (1988). For this we note from (15.7) that

$$E(y|x) = E(Z|x)\gamma + m(x)$$

and

$$y - E(y|x) = (Z - E(Z|x))\gamma + u.$$

Now, considering the local linear estimators of $E(y|x)$ and $E(Z|x)$, we can write

$$\begin{aligned} \tilde{y}_i &= y_i - \widehat{E(y_i|x_i)} = (I - S_i)y_i, \qquad i = 1, \ldots, M, \\ \tilde{Z}_i &= Z_i - \widehat{E(Z_i|x_i)} = (I - S_i)Z_i, \end{aligned}$$

where $S_i = [1\ 0]\big(Z'(x_i)\big)K(x_i)Z(x_i))^{-1}Z'(x_i)K(x_i)$ and $\tilde{y}_i$ and $\tilde{Z}_i$ are residuals in the ith local linear regressions of y_i on x_i and Z_i on x_i, respectively. So we get the set of linear regressions

$$\tilde{y}_i = \tilde{Z}_i\gamma_i + u_i.$$

Further, the GLS estimator of γ is

$$\hat{\gamma} = \left(\tilde{Z}'\Omega^{-1}\tilde{Z}\right)^{-1}\tilde{Z}'\Omega^{-1}\tilde{y}.$$

With this a two-step estimator of $m(x)$, given in (15.4), follows from the model

$$y - Z\hat{\gamma} = m(x) + u$$

by a nonparametric regression of $y - Z\hat{\gamma}$ on x.

15.3.2. NP Regressions with NP Autocorrelated Errors

Let us consider the NP system with nonparametric autocorrelated errors as follows:

$$y_{ij} = m_i(X_{ij}) + g_i(U_{i,j-1}, \ldots, U_{i,j-d}) + \varepsilon_{ij}, \qquad i = 1, \ldots, M,\ j = d+1, \ldots, N, \quad (15.9)$$

where d is assumed to be known, and $U_{i,j} = g_i(U_{i,j-1}, \ldots, U_{i,j-d}) + \varepsilon_{ij}$ in (15.9) is the AR(d) process of unknown form in the ith equation. To obtain efficient two-step estimation, we propose the following procedure for estimating the model (15.9).

In *the first step*, we follow Su and Ullah (2006) to estimate each regression i as below: First, a preliminary consistent estimator of m_i can be obtained, which gives $\hat{U}_{ij} = y_{ij} - \hat{m}_i(X_{ij})$. Second, we obtain a consistent estimator of g_i, $\hat{g}_i(\hat{U}_{i,j-1}, \ldots, \hat{U}_{i,j-d})$ via first-order local polynomial smoothing $\hat{U}_{ij}$ on $\underline{\hat{U}}_{i,j-1} \equiv (\hat{U}_{i,j-1}, \ldots, \hat{U}_{i,j-d})$. Third, replacing $g_i(U_{i,j-1}, \ldots, U_{i,j-d})$ by $\hat{g}_i(\hat{U}_{i,j-1}, \ldots, \hat{U}_{i,j-d})$, we obtain the objective function

$$\hat{Q}_i \equiv Nh_0^{-q_i}\sum_{j=1}^{N} K((x_i - X_{ij})/h_0) \times \hat{I}_{ij}.$$

$$\left[y_{ij} - \hat{g}_i(\hat{U}_{i,j-1}, \ldots, \hat{U}_{i,j-d}) - m_i(x_i) - (X_{ij} - x_i)m_i^{(1)}(x_i)\right]^2,$$

where $\hat{I}_{ij} = 1\left\{\hat{f}_{i,\underline{U}}(\underline{\hat{U}}_{i,j-1}) \geq b_i\right\}$ for some constant $b_i = b_i(N) > 0$ and $\hat{f}_{i,\underline{U}}$ is the nonparametric kernel estimator for the density $f_{i,\underline{U}}$ of $\underline{U}_{i,j-1}$. h_0 is the bandwidth based on $\{X_{ij}\}$ in the first-step estimation. Bandwidth h_U and kernel K_U based on the residual series $\left\{\hat{U}_{i,j}\right\}$ are used to estimate the density $f_{i,\underline{U}}$. By minimizing $\hat{Q}_i$, we obtain $\hat{m}_i(x_i)$ as the desirable estimator of $m_i(x_i)$ for the current step. As Su and Ullah (2006) indicated, $\hat{I}_{ij}$ is used to trim out small values of $\hat{f}_{i,\underline{U}}$ to obtain a desirable $\hat{m}_i(x_i)$, and we can set $b_i \propto (\ln N)^{-1/2}$.

In *the second step*, let $y_{ij}^* = y_{ij} - \hat{g}_i(\hat{U}_{i,j-1}, \ldots, \hat{U}_{i,j-d})$. Define $\mathbf{y}_E^{**} = \Omega^{-1/2}\mathbf{y}^* + (\mathbf{H}^{-1} - \Omega^{-1/2})\mathbf{m}(X)$. Then the efficient estimator $\hat{\delta}_E = \left(\hat{m}(x) \quad \hat{m}^{(1)}(x)'\right)'$ can be obtained by minimizing

$$\hat{Q}_{i,E} \equiv Nh_1^{-q_i} \sum_{j=1}^{N} K((x_i - X_{ij})/h_1)$$
$$\times \hat{I}_{ij} \Big[\mathbf{y}_{ij,E}^{**} - v_{(i-1)N+j} \Big(m_i(x_i) + (X_{ij} - x_i) m_i^{(1)}(x_i) \Big) \Big]^2,$$

where $v_{(i-1)N+j}$ is the $\big((i-1)N+j\big)$th element in P^{-1}, and h_1 is the second-step bandwidth.

After obtaining both $\hat{g}_i(\hat{U}_{i,j-1}, \cdots \hat{U}_{i,j-d})$ and $\hat{m}_i(x_i)$ in the first step, we can have the estimated residuals $\hat{\varepsilon}_{ij} = y_{ij} - \hat{g}_i(\hat{U}_{i,j-1}, \cdots \hat{U}_{i,j-d}) - \hat{m}_i(x_i)$. Further, we can estimate $\Omega^{-1/2}$ and $\mathbf{H}^{-1}$ by using $\hat{\varepsilon}_{ij}$ to obtain the feasible $\hat{\mathbf{y}}_E^{**} = \hat{\Omega}^{-1/2}\mathbf{y}^* + (\hat{\mathbf{H}}^{-1} - \hat{\Omega}^{-1/2})\hat{\mathbf{m}}(X)$. Finally, we obtain our two-step estimator $\hat{\delta}_E$ for the model (15.9). The asymptotic normality of this two-step estimator remains to be investigated in a future work.

15.3.3. Additive NP Models

The additive model is useful to conquer the notorious "curse of dimension" issue in nonparametric literature. In this section, we consider the following additive NP models:

$$y_{ij} = m_i(X_{ij,1}, \ldots, X_{ij,d}) + \varepsilon_{ij}$$
$$= c_i + \sum_{\alpha=1}^{d} m_{i,\alpha}(X_{ij,\alpha}) + \varepsilon_{ij}, \qquad i = 1, \ldots, M,\ j = 1, \ldots, N,$$

where $X_{ij,\alpha}$ is the αth regressor. To stack the regression models into one, we have

$$y = c + \sum_{\alpha=1}^{d} m_\alpha(X_\alpha) + \varepsilon, \tag{15.10}$$

where $y = \big(y_{11}, \quad \ldots, \quad y_{MN}\big)$, $m_\alpha(X_\alpha) = \big(m_{1,\alpha}(X_{1,\alpha}), \quad \ldots, \quad m_{M,\alpha}(X_{M,\alpha})\big)'$, and $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{MN})$. To estimate the above additive NP regression model, we propose the following procedure.

1. We use single-equation additive model estimator techniques (e.g., Yang, Härdle, and Nelson (1999), Linton and Härdle (1996)), to estimate $m_{i,\alpha}(X_{i,\alpha})$.
2. Obtain $\hat{m}_i(x_i) = \hat{c}_i + \sum_{\alpha=1}^{d} \hat{m}_{i,\alpha}(x_{i,\alpha})$, where $\hat{c}_i = \frac{1}{N}\sum_{j=1}^{N} y_{ij}$.

3. By applying the transformation proposed in two-step estimation, we can transfer (15.10) into

$$\begin{aligned}\Omega^{-1/2}\mathbf{y}+(\mathbf{H}^{-1}-\Omega^{-1/2})\left(c+\sum_{\alpha=1}^{d}m_{\alpha}(X_{\alpha})\right)&=\mathbf{H}^{-1}\left(c+\sum_{\alpha=1}^{d}m_{\alpha}(X_{\alpha})\right)+v\\ \overrightarrow{\mathbf{y}}&=\mathbf{H}^{-1}c+\mathbf{H}^{-1}\sum_{\alpha=1}^{d}m_{\alpha}(X_{\alpha})+v\\ &=c^{*}+\sum_{\alpha=1}^{d}m_{\alpha}^{*}(X_{\alpha})+v.\end{aligned}$$

Then employing the procedure proposed above, we can estimate the transformed model to obtain $\hat{m}_{\alpha,2\text{-}step}(X_{\alpha})$. Specifically, the feasible transformed response variable can be obtained from the estimated residuals $\hat{\varepsilon}_{ij}=y_{ij}-\hat{m}_i(x_i)$ to estimate Ω, H, and P. This gives

$$\overrightarrow{\mathbf{y}}=\hat{\Omega}^{-1/2}\mathbf{y}+(\hat{\mathbf{H}}^{-1}-\hat{\Omega}^{-1/2})\left(\hat{c}+\sum_{\alpha=1}^{d}\hat{m}_{\alpha}(X_{\alpha})\right).$$

The two-step estimator of $m_{i,\alpha}(x_{i,\alpha})$ can now be obtained as in step 1 above, considering the additive model in $\overrightarrow{\mathbf{y}}$ with respect to $x_{i,\alpha}$. The asymptotic properties remain to be developed.

15.3.4. Varying Coefficient NP Models

Varying coefficient NP models are practically useful in applied works (see Cai and Li (2008) and Su and Ullah (2011)). In this section, we consider the following varying coefficient NP model for the set of regressions,

$$y_{ij}=\beta_i\big(Z_{ij}\big)X_{ij}+\varepsilon_{ij},\qquad i=1,\ldots,M,\ j=1,\ldots,N. \tag{15.11}$$

When $E(\varepsilon_{ij}|x_{ij},z_{ij})=0$, by local linearizing the coefficient, we have

$$\begin{aligned}y_{ij}&=\left[\beta_i(z_i)+\big(Z_{ij}-z_i\big)\beta_i^{(1)}(z_i)\right]X_{ij}+u_{ij}\\&=Z_{ij}(z_i,X_{ij})\delta_i(z_i)+u_{ij},\end{aligned}$$

where $\beta_i^{(1)}(z_i)\equiv\partial\beta_i(z_i)/\partial z_i$, $Z_{ij}(z_i,X_{ij})\equiv\big(1\quad (Z_{ij}-z_i)\big)X_{ij}$,

$$Z_i(z_i,X_i)=\big(Z_{i1}(z_i,X_{i1}),\quad\ldots,\quad Z_{iN}(z_i,X_{iN})\big)',$$

which has dimension $N \times (q_i + 1)$. Stack the above models $j = 1, \ldots, M$ in a matrix form as

$$\begin{aligned} \mathbf{y} &= \beta(Z)X + \mathbf{u} \\ &= Z(z)\delta(z) + u, \end{aligned}$$

where $\mathbf{Z}(z) = \text{diag}\big(Z_1(z_1, X_1) \quad ,\ldots, \quad Z_M(z_M, X_M)\big), \delta(z) = \big(\delta_1(z_1) \quad ,\ldots, \quad \delta_M(z_M)\big)$.

The local linear least squares estimator for the varying coefficient NP models in (15.11) is

$$\hat{\delta}(z) = (\mathbf{Z}'(z)\mathbf{K}(z)\mathbf{Z}(z))^{-1}\mathbf{Z}^{'}(z)\mathbf{K}(z)\mathbf{y}.$$

Then we apply the two-step estimator as follows:

$$\begin{aligned} \Omega^{-1/2}\mathbf{y} + (\mathbf{H}^{-1} - \Omega^{-1/2})\beta(Z)X &= \mathbf{H}^{-1}\beta(Z)X + v \\ \overrightarrow{\mathbf{y}}_{VF} &= \mathbf{H}^{-1}\beta(Z)X + v. \end{aligned}$$

The corresponding two-step estimator can be written as

$$\hat{\boldsymbol{\delta}}_{2\text{-}step}(z) = (\mathbf{Z}^{*\prime}(z)\mathbf{K}(z)\mathbf{Z}^{*}(z))^{-1}\mathbf{Z}^{*\prime}(z)\mathbf{K}(z)\overrightarrow{\mathbf{y}}_{VF}, \tag{15.12}$$

where $\mathbf{Z}^{*}(\mathbf{z}) = \mathbf{H}^{-1}\mathbf{Z}(\mathbf{z})$. To obtain the operational estimator in the first step, we can estimate each equation by local linear least squares to get residuals. Then use the residuals to get a consistent estimator of covariance, further, obtain the feasible $\overrightarrow{\mathbf{y}}_{VF} = \hat{\Omega}^{-1/2}\mathbf{y} + (\hat{\mathbf{H}}^{-1} - \hat{\Omega}^{-1/2})\mathbf{Z}(z)\hat{\delta}(z)$. In the second step, we regress the feasible $\overrightarrow{\mathbf{y}}_{VF}$ on $\mathbf{H}^{-1}\beta(Z)X$ to get the two-step estimator.

15.3.5. Varying Coefficient IV Models

Let us consider the varying coefficient model with endogenous variables as

$$\begin{aligned} y_{ij} &= \beta_i\big(Z_{ij}\big)X_{ij} + \varepsilon_{ij}, \qquad i = 1, \ldots, M,\ j = 1, \ldots, N \\ E(\varepsilon_{ij}|W_{ij}, Z_{ij}) &= 0 \qquad \text{almost surely (a.s.),} \end{aligned}$$

where X_{ij} is an endogenous regressor, Z_{ij} denotes a $q_i \times 1$ vector of continuous exogenous regressors, and W_{ij} is a $p_i \times 1$ vector of instrument variables and the orthogonality condition $E(\varepsilon_{ij}|W_{ij}, Z_{ij}) = 0$ provides the intuition that the unknown functional coefficients can be estimated by nonparametric generalized method of moments (NPGMM). Let $V_{ij} = (W_{ij}^{\prime}, Z_{ij})^{\prime}$, we can write the orthogonality condition as

$$E[Q_{z_i}(V_{ij})\varepsilon_{ij}|V_{ij}] = E[Q_{z_i}(V_{ij})\{y_{ij} - Z_{ij}(z_i, X_{ij})\delta_i(z_i)\}|Z_{ij}] = 0,$$

where $Q_{z_i}(V_{ij})$ may also contain Ω^{-1} for improving efficiency.

Define

$$g(z) = \frac{1}{N} Q(z)' K_h(z)[y - Z(z)\delta(z)].$$

The dimension of $g(z)$ is $\sum_{i=1}^{M} k_i \times 1$, $Q(z) \equiv \operatorname{diag}\big(Q_{z_1}(V_1), \ldots, Q_{z_M}(V_M)\big)$, which has dimension $MN \times \left(\sum_{i=1}^{M} k_i\right)$. To obtain $\delta(z)$, we can minimize the following local linear GMM criterion function:

$$\big[Q(z)'K(z)\big(\mathrm{y}{-}Z(z)\delta(z)\big)\big]'\Psi(z)^{-1}\big[Q(z)'K(z)\big(\mathrm{y}{-}Z(z)\delta(z)\big)\big],$$

where

$$\Psi(z) = \frac{1}{N^2} Q(z)' K_h(z) \Omega K_h(z) Q(z),$$

which is a symmetric $\sum_{i=1}^{M} k_i \times \sum_{i=1}^{M} k_i$ weight matrix that is positive definite. Then the local linear GMM estimator of $\delta(z)$ is given by $\hat{\delta}_{GMM}(z)$ as

$$\hat{\delta}_{GMM}(z) = \Big\{Z(z)'K(z)Q(z)\big[Q(z)'K_h(z)\Omega K_h(z)Q(z)\big]^{-1}Q(z)'K(z)Z(z)\Big\}^{-1}$$
$$Z(z)'K(z)Q(z)\big[Q(z)'K_h(z)\Omega K_h(z)Q(z)\big]^{-1}Q(z)'K(z)\mathrm{y}. \qquad (15.13)$$

To obtain the optimal choice of weight matrix, we can first get the preliminary estimator $\tilde{\delta}_{GMM}(z)$ of $\delta_{GMM}(z)$ by setting $\Psi(z)$ as an identity matrix. Then we define the local residual $\tilde{\varepsilon}_{ij}(z_i) = y_{ij} - Z_{ij}(z_i, X_{ij})\tilde{\delta}_{GMM,i}(z_i)$. Using this, we can estimate $g(z)$ to obtain the optimal choice of weight matrix $\tilde{\Psi}(z) = \sum_{i=1}^{N} \hat{g}(z_i)\hat{g}(z_i)/N$. Alternatively, we can directly estimate the local variance–covariance matrix Ω by $\hat{\Omega} = \hat{\Sigma} \otimes I_N$. $\sigma_{ii'}$, the $\big(i, i'\big)$th element of Σ, can be estimated by

$$\hat{\sigma}_{ii'} = \sum_{j=1}^{N}\Big(\tilde{\varepsilon}_{ij}(z_i) - \bar{\tilde{\varepsilon}}_i(z_i)\Big)\Big(\tilde{\varepsilon}_{i'j}(z_{i'}) - \bar{\tilde{\varepsilon}}_{i'}(z_{i'})\Big)/(N-1),$$

where $\bar{\tilde{\varepsilon}}_i(z_i) = \frac{1}{N}\sum_{j=1}^{N} \tilde{\varepsilon}_{ij}(z_i)$, $i, i' = 1, \ldots, M$. Then the feasible local linear GMM estimator is obtained by substituting Ω with $\hat{\Omega}$ in (15.13).

The choice of instrument vector $Q(z)$, which implies choosing $Q_{z_i}(v_{ij}) = [Q_i(v_i)'(Q_i(v_i) \otimes (\bar{z}_i - z)\,/h_i)]$, is important in applications. For example, one can consider the union of w_i and z_i (say $Q_i(v_i) = v_i$) such that some identification is satisfied. In a cross-section model, Su, Murtazashrilli, and Ullah (2013) consider an optimal choice of $Q(v_i)$ by minimizing the asymptotic covariance matrix for the local linear GMM estimator. Another point to note is that the local constant GMM estimator, a special case of the local linear GMM, has been studied in Lewbel (2007), Tran and Tsionas (2009), and Cai and Li (2008) for the cross-section or panel model.

There are several special cases of the varying coefficient model considered above. For example, $\beta_i(z_i)x_i = \beta_{i1}(z_i)x_{i1} + \beta_{i2}(z_i)x_{i2}$, where x_{i1} and x_{i2} are subsets of x_i. Thus, we may test $\beta_{i1}(z_i) = \theta_1$ with respect to x_{i1} in each equation. This could be developed following Su, Murtazashvili, and Ullah (2013).

Now we consider an alternative estimation procedures that is free from the instruments w_{ij} for x_{ij}. This is based on an idea that $y_{ij} = \beta_i(z_{ij})x_{ij} + m_i(x_{ij}) + u_{ij}$, where we have written $\epsilon_{ij} = m_i(x_{ij}) + u_{ij}$ because $E(\epsilon_{ij}|x_{ij}) \neq 0$. Thus the parameter of interest $\beta_i(z_{ij})$ is not always identified since $m_i(x_{ij})$ could be a linear function of x_{ij}. The GMM estimation based on the instruments w_{ij} is one way to identify and estimate $\beta(z_{ij})$. Assuming $m_i(x_{ij}) \neq x_{ij}$ (linear), Gao (2012) provides a way to estimate both $\beta_i(z_{ij}) = \beta$ and $m(x_{ij})$ in a single equation model. However, it should be noted that an economic parameter of interest is derivative of y_{ij} with respect to x_{ij}, which is $\beta_i(z_{ij}) + \partial m_i(x_{ij})/\partial x_{ij}$ and not merely $\beta_i(z_{ij})$. This derivative is identifiable even if $\beta_i(z_{ij})$ is not identifiable. The estimation of $\beta_i(z_{ij}) + \partial m_i(x_{ij})/\partial x_{ij}$ can be obtained by a two-step procedure, which involves first estimating $y_{ij} = \beta_i(z_{ij})x_{ij} + u_{ij}$ by the estimator given in (15.12) and then doing a local linear NP regression of $\hat{\epsilon}_{ij}$ on x_{ij} to get an estimate of $\hat{m}_i(x_{ij})$. This is the Martins-Filho, Mishra, and Ullah (2008)-type estimator, and its properties need to be developed. The $\hat{m}_i(x_{ij})$ here works as an instrument. If the dimension of z_{ij} is large, then we can consider additive model or consider the model as $y_{ij} = \beta_i(z_{ij}\gamma_j)x_{ij} + m_i(x_{ij}) + \epsilon_{ij}$ (see Gao (2012)).

15.4. Estimation with Conditional Error Variance–Covariance $\mathbf{\Omega}(x)$

All the aforementioned estimations are based on the unconditional error variance covariance of errors. This section discusses the asymptotic properties for local linear least squares estimator and the two-step estimator for the NP set of regressions with conditional error variance–covariance of errors. Now we assume that $E(u_{ij}|X_{ij}) = 0$, and $\text{Var}(\varepsilon_{ij}|X_{ij}) = \sigma_{ii}\big(X_{ij}\big)$ for each equation. Also, we assume that the disturbances are uncorrelated across observations but correlated across equations; that is, $E(\varepsilon_{ij}\varepsilon_{i'j}|X_{ij}, X_{i'j}) = \sigma_{ii'}\big(X_{ij}, X_{i'j}\big)$ for $i, i' = 1, \ldots, M$ and $i \neq i'$, and $j = 1, \ldots, N$. In a matrix form, the conditional variance–covariance is $\Omega(x) \equiv \Sigma(x) \otimes I$ for a given evaluated point x. It is straightforward to show that both $\hat{\delta}(x)$ and $\hat{\delta}_{2\text{-}step}(x)$ are asymptotic normally distributed (for more details, see Wang (2012)).

To obtain feasible two-step estimation in this scenario, the estimated conditional variance–covariance is required. One can estimate the conditional covariance by a standard approach as follows:

$$\hat{\sigma}_{ii}(x) = \frac{\frac{1}{N}\sum\limits_{j=1}^{N} K_{\mathbf{h}}(x_i - X_{ij})\varepsilon_{ij}^2}{\frac{1}{N}\sum\limits_{j=1}^{N} K_{\mathbf{h}}(x_i - X_{ij})} \qquad \text{for } i = 1, \ldots, M,$$

$$\hat{\sigma}_{ii'}(x) = \widehat{\mathrm{Cov}}(\varepsilon_{ij}, \varepsilon_{i'j}) = \frac{\frac{1}{N}\sum\limits_{j=1}^{N} K_{\mathbf{h}}(x - X_j)\varepsilon_{ij}\varepsilon_{i'j}}{\frac{1}{N}\sum\limits_{j=1}^{N} K_{\mathbf{h}}(x - X_j)} \qquad \text{for } i, i' = 1, \ldots, M \text{ and } i \neq i',$$

where $X_j \in \mathbb{R}^d$ is a disjoint union of $\{X_{ij}\}$, $\mathbf{h} = \mathrm{diag}(h_{1,}, \ldots, h_q)$, $K_{\mathbf{h}}(x - X_j) = |\mathbf{h}|^{-1} \cdot K\big(\mathbf{h}^{-1}(x - X_j)\big)$, and $K_{\mathbf{h}}(x_i - X_{ij}) = |\mathbf{h}|^{-1} K\big(\mathbf{h}^{-1}(x_i - X_{ij})\big)$. For $i = i'$ we get $\hat{\sigma}^2_{ii}(x)$. Using this estimate of $\Omega(x)$, we can write the two-step estimators for all the models in the above subsections. The results here apply for the time-series data also, but the asymptotic theory can follow from the results in Long, Su, and Ullah (2011), where tests for the multivariate GARCH or univariate ARCH are also given (also see Mishra, Su, and Ullah (2010)).

15.5. Concluding Remarks

In this chapter, we survey some recent developments on NP and SP estimation for SRE models. The procedures of estimation for various nonparametric and semiparametric SRE models are proposed, including the partially linear semiparametric model, the model with nonparametric autocorrelated errors, the additive nonparametric model, the varying coefficient model, and the model with endogeneity. These results could also be extended for the NP SRE model with panel data; for example, see Wang (2012). The asymptotic properties of the estimators for many estimators in such models need to be developed in future studies. Also, the results on various testing problems—for example, testing for cross-equation correlations—need to be developed in future works.

References

Cai, Z., and Q. Li. 2008. "Nonparametric Estimation of Varying Coefficient Dynamic Panel Data Models." *Econometric Theory*, **24**, pp. 1321–1342.

Clark, R. M. 1977. "Non-parametric Estimation of a Smooth Regression Function." *Journal of the Royal Statistical Society*, Ser. B, **39**, pp. 107–113.

Denby, L. 1984. "Smooth Regression Functions." Unpublished Ph.D. dissertation, University of Michigan.

Gao, J. 2012. "Identification, Estimation and Specification in a Class of Semi-linear Time Series Models. Monash University, working paper.

Greene, W. H. 2007. *Econometric Analysis*, sixth edition. Upper Saddle River, NJ: Prentice Hall, 1216 pages.

Hayashi, F. 2000. *Econometrics*. Princeton, NJ: Princeton University Press, 690 pages.

Henderson, D., and A. Ullah. 2005. "A Nonparametric Random Effect Estimator." *Economics Letters*, **88**, pp. 403–407.

Henderson, D., and A. Ullah. 2012. “Nonparametric Estimation in a One-Way Error Component Model: A Monte Carlo Analysis.” In *ISI Statistical Jubilee, Statistical Paradigms: Recent Advances and Reconciliations*, forthcoming.

Kakwani, N. 1967. “The Unbiasedness of Zellner’s Seemingly Unrelated Regression Equations Estimators.” *Journal of the American Statistical Association*, **62**(317), pp. 141–142.

Koop, G., D. Poirier, and J. Tobias. 2005. “Semiparametric Bayesian Inference in Multiple Equation Models.” *Journal of Applied Econometrics*, **20**(6), pp. 723–747.

Lewbel, A. 2007. “A Local Generalized Method of Moments Estimator.” *Economics Letters*, **94**(1), pp. 124–128.

Li, Q., and J. S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice.* Princeton, NJ: Princeton University Press, 768 pages.

Lin, X., and R. J. Carroll. 2000. “Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured without/with Error.” *Journal of the American Statistical Association*, **95**, pp. 520–534.

Linton, O. B., and W. Härdle. 1996. “Estimation of Additive Regression Models with Known Links.” *Biometrika*, **83**(3), pp. 529–540.

Long, X., L. Su, and A. Ullah. 2011. Estimation and Forecasting of Dynamic Conditional Covariance: A Semiparametric Multivariate Model.” *Journal of Business Economics & Statistics*, **29**(1), pp. 109–125.

Martins-Filho, C., S. Mishra, and A. Ullah. 2008. “A Class of Improved Parametrically Guided Nonparametric Regression Estimators.” *Econometric Reviews*, **27**, pp. 542–573.

Martins-Filho, C., and F. Yao. 2009. “Nonparametric Regression Estimation with General Parametric Error Covariance.” *Journal of Multivariate Analysis*, **100**(3), pp. 309–333.

Mishra, S., L. Su, and A. Ullah. 2010. “Semiparametric Estimator of Time Series Conditional Variance.” *Journal of Business Economics & Statistics*, **28**(2), pp. 256–274.

Pesaran, M. H., R. P. Smith, and K. S. Im. 1996. “Dynamic Linear Models for Heterogeneous Panels.” Chapter 8 In *The Econometrics of Panel Data*, eds. L. Mátyás and P. Sevestre. Dordrecht: Kluwer Academic Publishers, pp. 145–195.

Robinson, P. M. 1988. “Root-N-Consistent Semiparametric Regression.” *Econometrica*, **56**(4), pp. 931–954.

Ruckstuhl, A. F., A. H. Welsh, and R. J. Carroll. 2000. “Nonparametric Function Estimation of the Relationship Between Two Repeatedly Measured Variables.” *Statistica Sinica*, **10**, pp. 51–71.

Smith, M. S., and R. J. Kohn. 2000. “Nonparametric Seemingly Unrelated Regression.” *Journal of Econometrics*, **98**(2), pp. 257–281.

Su, L., I. Murtazashvili, and A. Ullah. 2013. Local Linear GMM Estimation of Functional Coefficient IV Models with Application to the Estimation of Rate of Return to Schooling. *Journal of Business and Economic Statistics*, **31**(2), pp. 184–207.

Su, L., and A. Ullah. 2006. “More Efficient Estimation in Nonparametric Regression with Nonparametric Autocorrelated Errors.” *Econometric Theory*, **22**(1), pp. 98–126.

Su, L., and A. Ullah. 2007. “More Efficient Estimation of Nonparametric Panel Data Models with Random Effects.” *Economics Letters*, **96**(3), pp. 375–380.

Su, L., and A. Ullah. 2011. “Nonparametric and Semiparametric Panel Econometric Models: Estimation and Testing.” In *Handbook of Empirical Economics and Finance*, eds. A. Ullah and D. E. A. Giles. New York: Taylor & Francis, pp. 455–497.

Su, L., A. Ullah, and Y. Wang. 2013. "Nonparametric Regression Estimation with General Parametric Error Covariance: A More Efficient Two-Step Estimator." Forthcoming in *Empirical Economics*.

Taylor, W. 1977. "Small Sample Properties of a Class of Two Stage Aitken Estimators." *Econometrica*, **45**(2), pp. 497–508.

Tran, K. C., and E. G. Tsionas. 2009. "Local GMM Estimation of Semiparametric Panel Data with Smooth Coefficient Models." *Econometric Reviews* **29**(1), pp. 39–61.

Ullah, A., and N. Roy. 1998. "Nonparametric and Semiparametric Econometrics of Panel Data." In *Handbook of Applied Economics Statistics*, 1, eds. A. Ullah and D. E. A. Giles, New York: Marcel Dekker pp. 579–604.

Wang, Y. 2012. "Essays on Nonparametric and Semiparametric Models and Continuous Time Models." Ph.D. thesis (Chapter 3), University of California, Riverside.

Wang, Y., W. Guo, and B. Brown. 2000. "Spline Smoothing for Bivariate Data with Application Between Hormones." *Statistica Sinica*, **10**(2), pp. 377–397.

Welsh, A. H., and T. W. Yee. 2006. "Local Regression for Vector Responses." *Journal of Statistical Planning and Inference*, **136**(9), pp. 3007–3031.

Yang, L., W. Härdle, and J. Nielsen. 1999. "Nonparametric Autoregression with Multiplicative Volatility and Additive Mean." *Journal of Time Series Analysis*, **20**(5), pp. 579–604.

You, J., S. Xie, and Y. Zhou. 2007. "Two-Stage Estimation for Seemingly Unrelated Nonparametric Regression Models." *Journal of Systems Science and Complexity*, **20**(4), pp. 509–520.

Zellner, A. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association*, **57**(298), pp. 348–368.

CHAPTER 16

SEARCHING FOR REHABILITATION IN NONPARAMETRIC REGRESSION MODELS WITH EXOGENOUS TREATMENT ASSIGNMENT†

DANIEL J. HENDERSON AND ESFANDIAR MAASOUMI

16.1. Introduction

As Gary Becker (1981) has argued, “education” is not just an investment good, but also a consumption stream. An educated person will likely drive a larger stream of consumption from reading a book, or a page of NewsWeek, than an uninformed person. This greater benefit is likely related, nonlinearly, to many attributes of the individual and the characteristics of the “goods.”

In a parametric model for treatment effects, fixed coefficients for the treatment variable, as well as other variables and attributes, imposes severe, and possibly inadvertent, restrictions that may exclude the possibility of observing some aspects of the program or treatment. These may include different distributions of any treatment effect, different distributions of behavioral or response changes due to treatment, and others.

Consider the stylized linear in parameters model:

$$y = \beta X + Z'\gamma + u, \tag{16.1}$$

where X is an exogenous treatment and Z is a vector of other variables and attributes. A constant coefficient β has several implications, two of which are of immediate concern in this chapter. The first is that constant coefficients force a large degree of homogeneity on the individuals irrespective of treatment level (if X is not a binary variable). Everyone in each group (treated and nontreated) has the same response. The second is that changes in X (and/or Z) have no impact on “β” or “γ.” We call this a no rehabilitation assumption since it will not allow any behavior modification, either as a direct response to changes in X or as modifications in γ arising from treatment, or different amounts of treatment. To impose these restrictions, a priori, will exclude the possibility of learning from the observed data whether there are changes in both the outcome variable, y, and perhaps sustainable changes in behavior. In other words, a model of this kind is biased toward observing temporal/local responses, at most, to treatment programs and policies. There is no redemption!

Some aspects of these restrictions may be dealt with by considering variable parameter models when data allow it (e.g., panels are available). Indeed, parameters may be allowed to depend on the observed variables in a parametric way, effectively permitting a priori specific forms of nonlinearity in treatment responses, and changes in other coefficients. Alternatively, we could estimate nonparametric regressions, letting the data settle (a) the degree of nonlinearity (b) and the form of dependence of the responses as well as attribute effects (analogous forms of β and γ).

We attempt to address some of these problems in the present study in which we consider the impact of a well-known (exogenously assigned treatment) program, Greater Avenues for Independence (GAIN), on labor market outcomes. Our aim is to examine changes in the gradients which will vary continuously with the values of the variables in the model. The traditional approach in this literature focuses on the average treatment effect in the conditional distribution of y. More recent work removes this "veil of ignorance" by looking at the distribution of the treatment effects on various individuals/households, and so on.

The heterogeneous estimates allowed by the nonparametric approach pose new challenges. In effect, we now have a *distribution* of responses that need to be examined. We could of course report several aspects of this latter distribution, such as the mean value and quantiles of (the analogous version of β), say, for a range of observed variables and characteristics. Alternatively, dominance criteria may be used, as we intend in this work. Suppose that a job or drug treatment program is intended to be life-enhancing, so that one values higher responses ("β"). This will be compatible with the class of increasing utility/valuation functions, including "dollar valuations." This is all that is needed for first-order stochastic dominance (FSD) rankings. Failing to find it is as informative as when it is found to a statistical degree of confidence, using tests such as in Linton, Maasoumi, and Whang (2005). When FSD is found, we do not necessarily need a cardinal valuation function to inform the decision maker that the program is effective, or failed, whatever the criterion function. Only a decision needing to quantify the impact of the treatment by a scalar value will need to select a function for that purpose. Dollar values are useful and sensible, so long as we acknowledge that they reflect only one particular valuation function that gives one particular "complete" ranking. On the other hand, failing to find FSD makes it clear that decisions based on any cardinal valuation function, including dollars and averages, will inevitably assign different weights to different members of the population and are completely subjective. Different people will legitimately differ on both the usefulness of the treatment and the magnitude of the effects.

When FSD does not hold, it may be the case that second-order stochastic dominance (SSD) or higher orders hold to a statistical degree of confidence. Higher-order dominance rankings are interesting when *concave* valuation functions are justified, reflecting aversion to too much "dispersion," or "inequality" in treatment outcomes, or aversion to the risk of leaving behind some seriously at risk groups, or overly benefitting the less needy.

In our work, we examine such rankings for the distribution of the "responses" which are derivatives of the nonparametric version of model (16.1). Our approach does not compete with quantile techniques, as such. Indeed, SD rankings are equivalent to *joint* testing of ranking "all," or a desired subset of, quantiles. The difference is that a comparison based on individual quantiles may leave one in a quandary when the outcome direction is different for different quantiles. This will be equivalent to not finding FSD on the outcomes. But with SD rankings, comparing quantiles is not the end of the road, as it were. We can look for higher-order rankings with meaningful welfare theoretic interpretations that are essential to policy debate and decision making. While the treatment literature has begun to move beyond the "average treatment effect" in the distribution of the outcome variable, to our knowledge, our work is the first exploration of the distributed effects on responses.

In our empirical analysis, we find that future earnings are only (significantly) impacted by a handful of variables. Specifically, we find that enrollment in GAIN as well as higher test scores lead to higher earnings. More importantly, we find that enrollment in GAIN leads to heterogeneous impacts across the sample, with females having larger returns to GAIN than males, those whose primary language is English over those whose primary language is not English, older individuals over younger, those with no previous earnings over those with previous earnings, and those with higher test scores over those with lower test scores on average. However, even though we see higher returns at the quartiles, we find relatively few cases of stochastic dominance. In fact, we only find one case of FSD (English as the primary language versus English not being the primary language). However, we do find SSD for those age 21 and older over those under 21, those with children over those without children, and those with above median reading skills over those with below median reading skills. From a policy standpoint, this would suggest providing additional training in English reading skills, generally, and prior to enrollment in programs such as GAIN.

The remainder of the chapter proceeds as follows: Section 16.2 describes the stochastic dominance procedure. Section 16.3 briefly outlines the GAIN program, while Section 16.4 gives the empirical results of our study. Section 6.5 concludes the chapter.

16.2. Stochastic Dominance Procedure

In this section we outline our stochastic dominance procedure for gradient estimates. This methodology will also work for nonlinear parametric models, but we discuss a procedure for obtaining the gradient estimates nonparametrically. In our empirical application, we employ local-linear kernel regression for mixed data (Li and Racine, 2004; Racine and Li, 2004) using *AICc* selected bandwidth vectors (Hurvich, Simonoff, and Tsai, 1998), but other regression methods and bandwidth selectors are clearly feasible. We should note here that while we have a relatively large sample of data (6460 observations), we do have a large number of covariates (14) and hence we should keep the curse of dimensionality in mind.

Nonparametric estimation generates unique gradient estimates for each observation (individual) for each variable. This feature of nonparametric estimation enables us to compare (rank) several distributed effects of the exogenous treatment for subgroups and make inferences about who benefits most from the treatment. Here we propose using stochastic dominance tests for empirical examination of such comparisons.[1] The comparison of the effectiveness of a policy on different subpopulations based on a particular index (such as a conditional mean) is highly subjective; different indices may yield substantially different conclusions. Quantile regressions offer a limited solution that can be conclusive only when first-order dominance holds. In contrast, finding different orders of stochastic dominance provides uniform ranking regarding the impact of the policy among different groups and offers robust inferences. It is known to be simpler and more powerful than the corresponding tests of joint ranking of simple/marginal quantiles (see Maasoumi (2001)).

To proceed, consider a nonparametric version of the treatment regression

$$y = m(X, Z) + u,$$

where $m(\cdot)$ is an unknown smooth function of (the exogenous treatment) X and (covariates) Z. We are particularly interested in the change in the conditional expectation of y with respect to a change in the exogenous treatment variable X. We will denote this change as $\beta(X)$ $(= \nabla_X m(X, Z))$, but wish to emphasize that (as with all nonlinear regression functions with interactions) this gradient will likely depend on the values taken by the control variables Z. While it is possible to fix these control variables at their means (or other values), we prefer to allow them to remain at their individual observed values both because employing fixed values for Z would result in counterfactual estimates not representing any particular individual (see Henderson, Parmeter, and Kumbhakar (2012) for a discussion on the problems of such methods) and because in our case X is binary and thus fixing the Z would lead to scalar estimates and not allow for a distributional analysis.

If distinct and known groups are selected within the sample, we can examine the differences in returns between any two groups, say w and v. Here w and v might refer to males and females, respectively. Denote $\beta_w(X)$ as the effect of the treatment specific to an individual in group w. $\beta_v(X)$ is defined similarly. Again, note that the remaining covariates are not constrained to be equal across or within groups.

In practice, the actual treatment effect is unknown, but the nonparametric regression gives us an estimate of this effect. $\{\widehat{\beta}_{w,i}(X)\}_{i=1}^{N_w}$ is a vector of N_w estimates (one for each individual in group w) of $\beta_w(X)$ and $\{\widehat{\beta}_{v,i}(X)\}_{i=1}^{N_v}$ is an analogous vector of estimates of $\beta_v(X)$. $F[\beta_w(X)]$ and $G[\beta_v(X)]$ represent the cumulative distribution functions of $\beta_w(X)$ and $\beta_v(X)$, respectively.

Consider the null hypotheses of interest as follows.

Equality of distributions:

$$F[\beta(X)] = G[\beta(X)] \quad \forall \beta(X). \tag{16.2a}$$

First-order stochastic dominance: F dominates G if

$$F[\beta(X)] \le G[\beta(X)] \quad \forall \beta(X), \tag{16.2b}$$

Second-order stochastic dominance: F dominates G if

$$\int_{-\infty}^{\beta(X)} F(t)\,dt \le \int_{-\infty}^{\beta(X)} G(t)\,dt \qquad \forall \beta(X), \tag{16.2c}$$

Third-order stochastic dominance: F dominates G if

$$\int_{-\infty}^{\beta(X)} \int_{-\infty}^{s} F(t)\,dt\,ds \le \int_{-\infty}^{\beta(X)} \int_{-\infty}^{s} G(t)\,dt\;\;ds \quad \forall \beta(X), \tag{16.2d}$$

and so on. To test the null hypotheses, we define the empirical cumulative distribution function for $\beta_w(X)$ as

$$\widehat{F}[\beta_w(X)] = \frac{1}{N_w} \sum_{i=1}^{N_w} 1\big[\widehat{\beta}_{w,i}(X) \le \beta_w(X)\big], \tag{16.3}$$

where $1[\cdot]$ denotes the indicator function and $\widehat{G}[\beta_v(X)]$ is defined similarly. Next, we define the Kolmogorov–Smirnov statistics

$$T_{EQ} = \max\begin{pmatrix} \big\{\widehat{F}[\beta(X)] - \widehat{G}[\beta(X)]\big\}, \\ \big\{\widehat{G}[\beta(X)] - \widehat{F}[\beta(X)]\big\} \end{pmatrix}, \tag{16.4a}$$

$$T_{FSD} = \min\begin{pmatrix} \max\big\{\widehat{F}[\beta(X)] - \widehat{G}[\beta(X)]\big\}, \\ \max\big\{\widehat{G}[\beta(X)] - \widehat{F}[\beta(X)]\big\} \end{pmatrix}, \tag{16.4b}$$

$$T_{SSD} = \min\begin{pmatrix} \max\Big\{\int_{-\infty}^{\beta(X)} \big[\widehat{F}(t) - \widehat{G}(t)\big]dt\Big\}, \\ \max\Big\{\int_{-\infty}^{\beta(X)} \big[\widehat{G}(t) - \widehat{F}(t)\big]dt\Big\} \end{pmatrix}, \tag{16.4c}$$

$$T_{TSD} = \min\begin{pmatrix} \max\Big\{\int_{-\infty}^{\beta(X)} \int_{-\infty}^{s} \big[\widehat{F}(t) - \widehat{G}(t)\big]dt\;ds\Big\}, \\ \max\Big\{\int_{-\infty}^{\beta(X)} \int_{-\infty}^{s} \big[\widehat{G}(t) - \widehat{F}(t)\big]dt\;ds\Big\} \end{pmatrix}, \tag{16.4d}$$

for testing the equality, first-order stochastic dominance (FSD), second-order dominance (SSD), and third-order dominance (TSD), respectively.

Consistent estimation of $\beta(X)$ does not require us to split the sample for groups w and v, but our bootstrap procedure does. Specifically, we suggest to split the sample into two distinct groups and run separate nonparametric regressions on each (including estimating bandwidths for each group separately). These estimates of $\beta(X)$ will also be consistent (this is analogous to running separate regressions for a Chow test) and will allow us to compare the distributions of the two groups without the information from one affecting the other. In essence, this is equivalent to setting the bandwidth on the variable we are comparing (say gender) to zero (which will occur asymptotically, in any case).

Based on these estimates, we can construct our test statistics in (16.4a)–(16.4d). The asymptotic distributions of these nonparametric statistics are generally unknown because they depend on the underlying distributions of the data. We propose resampling approximations for the empirical distributions of these test statistics to overcome this problem. Our bootstrap strategy is as follows:

(i) Using nonparametric regression methods, obtain the estimates of $\beta(X)$ $(\widehat{\beta}(X) = \nabla_X \widehat{m}(X,Z))$ for each group.

(ii) Let T be a generic notation for T_{EQ}, T_{FSD}, T_{SSD}, and T_{TSD}. Compute the test statistics T from the original gradient estimates $\{\widehat{\beta}_{w,1}(X), \widehat{\beta}_{w,2}(X), \ldots, \widehat{\beta}_{w,N_w}(X)\}$ and $\{\widehat{\beta}_{v,1}(X), \widehat{\beta}_{v,2}(X), \ldots, \widehat{\beta}_{v,N_v}(X)\}$.

(iii) For each observation in group w, construct the centered bootstrapped residual u^*, where $u^* = \frac{1-\sqrt{5}}{2}\left(\widehat{u} - \overline{\widehat{u}}\right)$ with probability $\frac{1+\sqrt{5}}{2\sqrt{5}}$ and $u^* = \frac{1+\sqrt{5}}{2}\left(\widehat{u} - \overline{\widehat{u}}\right)$ with probability $1 - \frac{1+\sqrt{5}}{2\sqrt{5}}$. Then construct the bootstrapped left-hand-variable as $y^* = \widehat{m}(X,Z) + u^*$ for each observation in group w. Call $\left\{y_i^*, X_i, Z_i\right\}_{i=1}^{N_w}$ the bootstrap sample. Repeat this process for group v.

(iv) Re-estimate $\beta(X)$ for each group using the same nonparametric procedure and bandwidths in (i), but replace the data with the bootstrap data obtained in (iii). Call these estimates $\widehat{\beta}^*(X)$.

(v) Compute (centered^2) bootstrapped test statistics T_b from the bootstrapped estimates, where (for FSD, the others follow similarly)

$$T_b = \min \left[\begin{array}{c} \max \left(\begin{array}{c} \left\{\widehat{F}^*[\beta(X)] - \widehat{G}^*[\beta(X)]\right\} \\ -\left\{\widehat{F}[\beta(X)] - \widehat{G}[\beta(X)]\right\} \end{array} \right), \\ \max \left(\begin{array}{c} \left\{\widehat{G}^*[\beta(X)] - \widehat{F}^*[\beta(X)]\right\} \\ -\left\{\widehat{G}[\beta(X)] - \widehat{F}[\beta(X)]\right\} \end{array} \right) \end{array} \right],$$

where $\widehat{F}^*[\beta(X)]$ is the analogous estimate of (16.3) for the bootstrap estimates.

(vi) Repeat steps (iii)–(v) B times.

(vii) Calculate the "p-values" of the tests based on the percentage of times the centered bootstrapped test statistic is negative. Reject the null hypotheses if the p-value is smaller than some desired level α, where $\alpha \in (0, 1/2)$.

The careful reader will notice that the main departure from typical SD tests is that the data in question ($\beta(X)$) is unknown and thus must be estimated. Therefore, instead of bootstrapping from $\widehat{\beta}(X)$, it is important to bootstrap from the data and re-estimate $\beta(X)$ in each replication.[3] This allows us to to approximate the distribution of the derivatives. By resampling, we take into account the fact that we are dealing with the estimates of the gradients and not the actual gradients.

The most important steps above are the third through fifth. In (iii), we emphasize that we do not impose the least favorable case. Instead we separate the groups and resample from each separately. This can be achieved several ways (which we have done), but our preferred procedure is to use a wild bootstrap (to avoid issues with respect to potential heteroskedasticity). Then proceeding to step (iv), we re-estimate each model (using the same bandwidths as in step (i)). Note that we evaluate the bootstrapped gradient estimates at the original X and Z values. In the fifth step, we calculate the bootstrapped-based test statistic by evaluating over the same grid we did in step (ii).[4]

We wish to note here that in our empirical example, the gradient in question comes from a binary regressor. Hence, we only achieve a gradient estimate for those observations for which the dummy variable is equal to unity. Therefore, we construct our empirical CDF's with fewer observations than if we had a continuous regressor, but the basic methodology remains the same.

16.3. Greater Avenues for Independence

The Greater Avenues for Independence (GAIN) program was started in California in 1986 in order to help long-term welfare recipients "find employment, stay employed, and move on to higher-paying jobs, which will lead to self-sufficiency and independence." It is a mandatory (excluding female heads of households with children under age six) program for adults receiving Aid to Families with Dependent Children (AFDC).

The program initially administers screening tests to determine basic math and reading skills. Those deemed to be below a given level are targeted to receive basic education. Those above a given level are moved into either a job search assistance program or a vocational training program. This decision largely falls on the county with some counties preferring one over another.

Starting in 1988, a randomly assigned subset of GAIN registrants in six California counties (Alameda, Butte, Los Angeles, Riverside, San Diego, and Tulare) were assigned to a treatment group and the remaining were selected into a control group. Those in the treatment group were allowed to participate in the GAIN program and the remaining were not, but were still allowed to receive standard AFDC benefits. Those in the control group were allowed, but not required, after two years, to join the GAIN program.

Table 16.1 Descriptive Statistics

Variables By Type	All	Treatment Group	Control Group
Dependent Variable			
Earnings	10079.5147	10696.7075	7790.6829
Unordered Categorical Variables			
Experimental (GAINS)	0.7876	1.0000	0.0000
Female	0.6819	0.7040	0.5999
Employment or training (prior year)	0.2376	0.2390	0.2325
White	0.5500	0.5513	0.5452
Not White	0.4500	0.4487	0.4548
Hispanic	0.2464	0.2435	0.2573
Black	0.1551	0.1584	0.1429
Asian	0.0334	0.0311	0.0423
Primary language English	0.9610	0.9636	0.9512
Primary language Spanish	0.0197	0.0181	0.0255
Ordered Categorical Variables			
Age	32.2918	32.3143	32.2085
Highest school grade completed	11.1642	11.1733	11.1305
Number of children	2.0193	2.0161	2.0313
Continuous Variables			
Earnings previous 12 quarters	2335.7927	2293.0782	2494.1975
CASAS reading score	232.6416	232.6085	232.7646
CASAS math score	219.5249	219.5871	219.2945
Number of Observations	6460	5088	1372

Notes: Average values are listed. The first column of numbers is for the entire Riverside sample, the second is for the treatement group, and the final is for the control group.

From the econometrician's standpoint, this data set is ideal because the participants were randomly assigned to either the treatment or the control group. Table 16.1 shows that for Riverside County, nearly all means are the same between the two groups, perhaps with the exception of females in the control group. The results are similar for the other counties.

We choose Riverside County for several reasons. It has been highlighted by many as the best performing county. In fact, it has often been referred to as the "Riverside Miracle" (e.g., see Nelson (1997)). This result has led many to study this case (e.g., see Dehejia (2003)) and thus our findings can be compared to past studies. Finally, the sample is relatively large, and given the large number of covariates, our estimation procedure benefits greatly from the relatively large sample size.

Although these data have previously been studied using rigorous econometric techniques (e.g., Dehejia (2003); Hotz, Imbens, and Klerman (2006)), to our knowledge, no one has used nonparametric methods. The need for these methods with this particular data set has been hinted at before. Dehejia (2003, p. 9) mentions that "an estimator or a functional form that is more flexible in terms of pretreatment covariates should yield a more reliable prediction of the treatment impact."

In addition to having a more flexible approach, we are also able to get a treatment effect for each GAIN recipient in the program. This allows us to look at heterogeneity both across and within groups. Further, it allows us to use the stochastic dominance methods discussed earlier to look for relationships amongst the returns for prespecified groups in order to better inform policy decisions.

16.4. Empirical Results

We begin by looking at the cross-validated bandwidths from the regression of earnings on pretreatment attributes (Table 16.2). These bandwidths can lead to knowledge about whether or not variables are relevant and whether or not they enter the model linearly. We then turn our attention to the gradient estimates (Table 16.3). Although our primary concern is with respect to the GAIN participation variable, we will also

Table 16.2 Bandwidths

Variables By Type	Bandwidth	Upper Bound	Interpretation
Unordered Categorical Variables			
Experimental (GAINS)	0.2630	0.5000	Relevant
Sex	0.2382	0.5000	Relevant
Employment or training (prior year)	0.3451	0.5000	Relevant
Ethnic group	0.7721	0.8750	Relevant
Primary language English	0.4993	0.5000	Most likely irrelevant
Primary language Spanish	0.4993	0.5000	Most likely irrelevant
Family status	0.7590	0.8000	most likely irrelevant
Ordered Categorical Variables			
Age	0.9986	1.0000	Most likely irrelevant
Highest school grade completed	0.9986	1.0000	Most likely irrelevant
Number of children	0.9986	1.0000	Most likely irrelevant
Random assignment month	0.9986	1.0000	Most likely irrelevant
Continuous Variables			
Earnings previous 12 quarters	$1.35E-01$	∞	Nonlinear
CASAS reading score	$3.94E+06$	∞	Most likely linear
CASAS math score	$5.22E+07$	∞	Most likely linear

Notes: Bandwidths selected via AICc. Aitchison and Aitken (1976) kernel used for unordered data, Wang and van Ryzin (1981) kernel used for ordered data and second-order Gaussian kernel used for continuous data.

Table 16.3 Significant Nonparametric Gradient Estimates at the Quartiles

Variable	Q1	Q2	Q3
Unordered Categorical Variables			
Treatment (GAIN)		184.1634	644.8306
Continuous Variables			
CASAS reading score		89.6598	123.8145
CASAS math score		37.6296	60.5834

Notes: Significant gradient estimates for the first, second, and third quartiles are listed above (standard errors obtained via bootstrapping are available upon request). For those variables with no significant quartiles, the estimates are excluded. For discrete regressors, the lowest value taken by the gradient is exactly zero by definition.

analyze other gradients. We then turn our focus to our primary interest. We split the sample amongst the prespecified groups and look at their returns distributions to the GAIN program (Table 16.4). Finally, we perform stochastic dominance tests to determine whether or not we have first- or higher-order dominance relationships (Tables 16.5 and 16.6).

16.4.1. Bandwidth Estimates

Table 16.2 presents the bandwidths for the nonparametric model. The bandwidths reveal three salient points. First, the bandwidths on the CASAS reading and math score variables each exceed 3.94E+06. Since continuous regressors behave linearly as the bandwidths approach infinity, this suggests that a linear approximation for these two variables may be reasonable. The bandwidth on the "previous earnings" in the past 12 quarters is relatively small, indicating nonlinear effects. Employing a model that is linear in this variable would most likely lead to inconsistent estimates. Second, the bandwidths on the treatment, gender, prior employment or training, and ethnic group are much smaller than their respective upper bounds, implying that these variables are relevant in the model. Finally, the bandwidths on the primary language variables, as well as family status, age, highest school grade completed, number of children, and random assignment month are each close to their respective upper bounds; thus, these variables are (likely) statistically irrelevant in explaining treatment effect on earnings.

In sum, examination of the bandwidths suggest that some variables are relevant and some variables are irrelevant. Further, it suggests that some variables enter the model nonlinearly and some variables enter the model linearly. However, this does not mean we should automatically switch to a semiparametric estimation procedure. Linearity is not synonymous with homogeneous effects of the covariates. Consequently, while

Table 16.4 Significant returns to GAIN by group at the quartiles

Variable	*Q*1	*Q*2	*Q*3
Unordered Categorical Variables			
Gender			
Female	691.3410	1125.3420	1652.9889
Male		103.3956	331.2337
Previous training			
Employment or training (prior year)	338.7188	1122.1333	1974.1487
No Employment or training (prior year)	379.5400	897.4375	1457.1979
Ethnic group			
White	173.4314	971.4750	1754.1438
Not white	202.2289	897.1133	1368.9425
Hispanic	127.9419	362.3600	504.5695
Black	385.7882	996.9691	1412.9833
Asian			
Language			
Primary language English	554.1186	1035.6361	1513.4771
Primary language is not English			
Primary language Spanish			
Ordered Categorical Variables			
Age			
Under 21			386.0289
21 and over	545.5721	1090.4551	1696.0552
Highest school grade completed			
Less than high school		48.5136	80.2157
High school diploma and over		32.7699	46.8052
Number of children			
Zero			
One or more	344.6863	723.9139	1074.3786
Continuous Variables			
Previous earnings			
Positive earnings previous 12 quarters	281.0717	604.8498	790.5349
No earning in previous 12 quarters	716.2675	1133.7201	1580.6813
Test scores			
CASAS reading score above median	864.1227	1400.7653	1848.2009
CASAS reading score below median	225.2025	486.8775	719.7718
CASAS math score above median	481.5905	1066.8121	1674.2816
CASAS math score below median		247.4605	740.0365

Notes: Returns to GAIN for the first, second, and third quartiles for particular subgroups are listed above. Only those that are significant are listed (standard errors obtained via bootstrapping are available upon request). Each estimate is obtained by splitting the sample and running a separate regression (including cross-validation routine) on the prespecified group.

Table 16.5 Stochastic Dominance Test Statistics

Comparison	EQ	FSD	SSD	TSD
Unordered Categorical Variables				
Female vs. Male	0.6862	0.0336	3.3447	383.2913
Previous employment or training vs. no previous employment or training	0.1520	0.0423	11.0134	3381.1571
White vs. Not white	0.1397	0.0187	0.8180	24.1794
White vs. Black	0.1422	0.0728	3.9184	−0.0219
White vs. Hispanic	0.5074	0.0829	24.1267	7098.3774
White vs. Asian	0.6407	−0.0042	−0.0340	−0.0340
Black vs. Hispanic	0.5790	0.0824	31.0026	11533.7441
Black vs. Asian	0.6989	0.0091	−0.0120	−0.0120
Hispanic vs. Asian	0.6296	0.0275	−0.0435	−0.0435
Primary language English vs. Primary language not English	0.8821	−0.0102	−0.3854	−0.3854
Primary language Spanish vs. primary language not Spanish	0.7590	0.1264	1.4339	15.6832
Ordered Categorical Variables				
21 and Over vs. Under 21	0.5492	0.0029	−0.0230	−0.0230
High school diploma vs. No high school diploma	0.2761	0.0609	0.5873	6.0183
Children vs. No Children	0.7084	0.0091	−0.1010	−0.1010
Continuous Variables				
No earnings in previous 12 quarters vs. Earnings in previous 12 quarters	0.4951	0.0077	0.2199	14.0987
CHASS reading score above median vs. CHASS reading score below median	0.5799	−0.0014	−0.0041	−0.0041
CHASS math score above median vs. CHASS math score below median	0.4135	0.0017	0.0182	0.5662

Notes: The number in each cell is the test statistic for the comparison of the returns to enrollment in GAIN between two prespecified groups for a particular test. The first column is a test for equality. The second through fourth columns are tests for stochastic dominance (first, second, and third order, respectively). For the stochastic dominance tests, those test statistics which are negative are possible cases where dominance may exist. For the negative test statistics, the *p*-values in Table 16.6 will determine whether or not dominance exists.

Table 16.6 Stochastic Dominance Test *p*-Values

Comparison	EQ	FSD	SSD	TSD
Unordered Categorical Variables				
Female vs. Male	0.0000			
Previous employment or training vs. No previous employment or training	0.0000			
White vs. Not white	0.0000			
White vs. Black	0.0000			**0.4810**
White vs. Hispanic	0.0000			
White vs. Asian	0.0000	0.0253	0.0506	0.0506
Black vs. Hispanic	0.0000			
Black vs. Asian	0.0000		0.2025	0.2405
Hispanic vs. Asian	0.0000		0.0886	0.0886
Primary language English vs. Primary language not English	0.0000	**0.8734**	**0.8861**	**0.8861**
Primary language Spanish vs. Primary language not Spanish	0.0000			
Ordered Categorical Variables				
21 and Over vs. Under 21	0.0000		**0.8228**	**0.8608**
High school diploma vs. No high school diploma	0.0000			
Children vs. No children	0.0000		**0.7722**	**0.8101**
Continuous Variables				
No earnings in previous 12 Quarters vs. Earnings in previous 12 quarters	0.0000			
CHASS reading score above median vs. CHASS reading score below median	0.0000	0.1013	0.3924	**0.5443**
CHASS math Score above median vs. CHASS math score below median	0.0000			

Notes: The number in each cell is the *p*-value for the comparison of the returns to enrollment in GAIN between two prespecified groups for a particular test. For the stochastic dominance tests (columns 2–4), the *p*-value is included only if the corresponding test statistic in Table 16.5 is negative. Cases where we fail to reject the null of dominance are listed in bold. 399 bootstrap replications are performed for each SD test.

the assumption of linearity receives the most attention, heterogeneity may be just as problematic. We now turn to the actual results, as well as more formal statistical tests.

16.4.2. Parameter Estimates

16.4.2.1. All Covariates

Table 16.3 presents the results for the gradient estimates. We present the nonparametric estimates corresponding to the 25th, 50th, and 75th percentiles of the estimated gradient distributions (labeled $Q1$, $Q2$, and $Q3$). Estimates that are statistically significant at the 5% level are listed. To conserve space, we exclude any regressor for which each of the quartiles are insignificant. The full set of estimates with corresponding standard errors is available from the authors upon request.

In terms of the unordered categorical variables, several findings stand out. First, nonparametric estimates of the treatment (enrollment in GAIN) are positive and significant at the median and upper quartile. Perhaps more important for this study is that the third quartile is over three times the value of the second quartile. This shows prevalence of heterogeneity in the effect of the treatment across the sample. Finally, while some of the bandwidths suggest relevance, we did not find significance of any of the other unordered categorical regressors at the quartile values.

Likewise, for the ordered categorical variables, none of the quartile gradient estimates are significant. Again, these results are expected because, as was observed before, their bandwidths approached their upper bounds of unity. The implication is that they are not important in prediction of earnings. However, this does not mean that they do not play a role in terms of the impact of the treatment, as we will check later.

Finally, for the continuous variables, it is seen that CASAS reading and math scores have effects on earnings. The partial effect at the median for reading scores is 89.6598 (s.e. = 22.4854) and the partial effect at the median for the math score is 37.6296 (s.e. = 12.4125). This result suggests that improving basic reading and math scores would lead to higher earnings (with improvements in reading skills typically being more beneficial than mathematics). While the bandwidths suggest that each of these variables enter linearly, they do not shed light on possible heterogeneity. The results at the quartiles show heterogeneity in the partial effects and re-emphasize the importance of a nonlinear estimation procedure.

16.4.2.2. Treatment Variable

The results across different covariates are interesting, but a main purpose of this study and the GAIN experiment is to determine the effect of the treatment. In most studies, a single coefficient is obtained for the (average) treatment, and its magnitude determines whether or not the treatment was successful. Here we obtain a separate estimate for each person receiving the treatment. Thus, we can examine the effect of the treatment among prespecified groups.

Table 16.4 gives the nonparametric estimates corresponding to the 25^{th}, 50^{th}, and 75^{th} percentiles of the distribution for the treatment (GAIN) for specific subgroups. Specifically, we broke the sample across each prespecified group and ran separate nonparametric regressions on each subgroup (including calculating bandwidths for each). The quartile estimates for those in the GAIN program are given in the table.

The results for the groupings by unordered categorical variables are quite striking. First, the effect of the treatment on women is larger than the effect of the treatment on men at the median and at the first and the third quartiles (note that the first quartile estimate for men is negative and insignificant). Although the results at the quartiles are strong, we cannot determine whether or not the effect of the treatment for women dominates the effect of the treatment for men. We will examine this further in the next subsection. Second, there is some evidence that individuals who received the treatment and had previous training or work experience in the prior year experienced larger returns to GAIN than treated individuals who did not have employment or training in the prior year. However, these results only hold at the median and at the upper quartile. This would suggest that no dominance relation exists. Third, Asians are the only ethnic group who did not experience significant returns at these quartiles. Finally, although the bandwidth for English or Spanish as a first language was near its upper bound, treated individuals who spoke English as their native language had positive and significant treatment effects for enrollment in GAIN and those whose primary language was not English did not experience significant returns to GAIN at any of the quartiles. This result may suggest that immigrants are not benefitting from the program. This result is consistent with other inferences below related to spoken or written English.

For the ordered categorical variables, we see that treated individuals aged 21 and over had larger effects than did treated individuals under 21. The level of schooling seemed to make little difference on who benefitted the most from the program. Finally, treated individuals who have one or more children have larger treatment effects at each quartile than individuals who did not have any children (perhaps a sign of necessity). Again, these results at these selective quartiles are strong, but it is premature to conclude that any of these groups “dominate” one another in terms of the partial effect of the treatment variable.

Finally, for groupings corresponding to the continuous variables, treated individuals with no earnings in the previous 12 quarters had larger effects of the treatment than did treated individuals who had positive earnings in the previous 12 quarters at each quartile. The test scores results are as expected. Treated individuals obtaining scores above the median (either in math or reading) have larger treatment effects as compared to their counterparts who scored below the median. This shows that higher ability individuals are able to benefit more from the treatment. We return to the Gary Becker argument that we paraphrased in the Introduction: education is both an investment and a consumption good. The greater benefit of education is likely related, nonlinearly, to many attributes of the individual and the characteristics of the “goods.”

16.4.3. Dominance Tests

The results of the previous subsection showed that certain groups appeared to have higher returns from the treatment than did other groups, at certain quantiles. Here we use formal tests to compare the effect of the treatment between two prespecified groups across all quantiles. Tables 16.5 and 16.6 break down the results for tests of equality, first order, second order and third order dominance. Table 5 gives the test statistics. A negative sign of a test statistic is a sign of possibly significant dominance relation. The entries in Table 16.6 are the "*p*-values" for the corresponding tests.

16.4.3.1. Test Statistics

In Table 16.5, the entries are the sample value of the test statistics. The left-hand side of the table gives the prespecified groups being compared. In each case we are comparing the treated individuals in each group. The first column of numbers gives the test statistic for the equality of the distributions of the gradient of the conditional mean with respect to the treatment (GAIN). The second through fourth columns give the test statistic for first-, second-, and third-order dominance, respectively. In order for a dominance relation to exist, the test statistic must be negative. For example, for the first-order dominance case, if the test statistic is negative, then first-order dominance is observed. If the test statistic is positive, then there is no observed ranking in the first-order sense. Similar interpretations are given to higher-order dominance relations.

When examining the test statistics for first-order dominance, there is only the possibility of FSD for three of the 17 comparisons. The comparisons with negative FSD test statistics are: white versus Asian, primary language English versus primary language not being English, and CHASS reading score above the median versus score below the median. The lack of negative test statistics for the comparison between those with and without previous earnings may be surprising given the results at the quartiles, but these suggest crossing of the distributions closer to the tails.

As expected, more cases of second order dominance are observed. The third column of numbers in Table 16.5 gives the test statistics for the null of second-order dominance (noting that first-order dominance implies second-order dominance, and so on). Here we also find negative test statistics for each ethnic group versus Asians, those 21 and over versus those under 21, and those with children over those not having children. For third-order dominance, we also find a negative test statistic for white versus black. These higher-order dominance rankings imply that policy makers with an aversion, or increasing aversion to earnings poverty, would find the program to be beneficial, whatever cardinal weighting function/utility is adopted.

16.4.3.2. Probability Values

Each value in Table 16.6 is the *p*-value associated with a particular test. The first column rejects the null of equality of the distributions of the treatment effects when the

p-value is below α. In columns 2–4, the respective order of dominance is rejected (when its associated test statistic is negative) if the p-value is less than 0.400 (see Maasoumi, Millimet, and Sarkar (2009)). Substantial coverage probability for negative values of the statistic supports an inference of dominance to a degree of statistical confidence.

In Table 16.6, we reject each null that the pairs of treatment effect estimates are equal. These results are not surprising given what we have seen thus far. For the dominance tests (in Table 16.5) with negative sample statistics, there are cases where there is significant evidence of dominance. The strongest ranking is the finding of first-order dominance. We find that those whose primary language is English have uniformly higher returns to GAIN than those whose first language is not English. First-order dominance implies higher-order dominance, and we see that the p-values for second- and third-order dominance are larger in magnitude than that of the first order test. In two other cases where we found negative test statistics for first-order dominance (white versus Asian and above median reading score versus below median), both have p-values much less than 0.40.

We find three strong cases for second-order dominance. In addition to white versus Asian, we also see that those who received the treatment and were 21 years and older gained more than those under 21; similar results occurred for those with children versus those without children. It may be that older individuals and those who have dependents took better advantage of the program. Finally, we have one test statistic with a p-value near the border of 0.40. Reading score above the median versus reading score below the median (p-value = 0.3924) is likely related to the result of language ability. This, along with the previous results, suggest that the program may want to focus more on basic language and reading skills.

Finally, for third-order dominance, in addition to those listed above, we find a further ranking of white versus black treatment outcomes. Those with increasing aversion to inequality of earnings at the lower end of the earnings distribution would infer a greater benefit to whites versus blacks treated in GAIN.

16.5. Conclusions

In this chapter we outlined a method to compare gradient estimates from a nonparametric regression via stochastic dominance techniques. Our goal here was to look at the impact of an exogenous treatment across different prespecified groups.

To showcase the methodology, we applied our procedure to the California GAIN program. Here we found that relatively few inputs commonly used in determining labor outcomes are significant. Specifically, we only found significant quartile estimates for improving earnings for enrollment in GAIN and for test scores. Although many results were insignificant, we did find that certain groups had higher returns to GAIN. For example, we found that females, those whose primary language was English, those individuals over the age of 21, and those with higher test scores had higher returns to the treatment. However, we only found one case of first-order dominance: English as the primary language versus English not being the primary language. We also found some evidence of second- and higher-order dominance—for example, for above median versus below median reading scores. From a policy standpoint, this suggests that improving basic reading skills can increase the impact of GAIN.

An interesting extension to our work would be to calculate "collateral effects," which we define as changes to the gradients of the other regressors (Z) arising from the treatment, or different amounts of the treatment (if the treatment were continuous). These can be calculated as the cross-partial derivatives with respect to X and any element in Z. In other words, we would like to allow for the treatment to have effects on other attributes of the individual.

Notes

[†] The authors would like to thank an anonymous reviewer and Liangjun Su for excellent comments. Manpower Demonstration Research Corporation and its funders are not responsible for the use or interpretation of the data.

1. For an empirical application of stochastic dominance tests on estimated outcome values obtained via nonparametric regression, see Maasoumi, Racine, and Stengos (2007).
2. The centering of the bootstrap test statistic is performed by subtracting the initial sample estimates of the empirical CDF differences. We do not impose the null hypothesis (specifically, we do not impose the least-favorable case) in step (iii). In this way we obtain consistent estimates of the sampling distributions and coverage probabilities, with i.i.d samples. Standard results for centered bootstrap validity apply here. We have also conducted extensive experiments when the null of the least favorable case is imposed, in addition to centering on the initial test statistics themselves. Our empirical findings are generally the same.
3. Eren and Henderson (2008) and Henderson (2010) simply resample the gradient estimates. If the distribution functions are sufficiently well separated, this should lead to the same conclusions, but we recommend re-estimating the gradients in practice.
4. We performed simulations to determine the size and power of our bootstrap-based test and found that it did well in relatively small samples. These results are available from the author upon request.

References

Aitchison, John, and Colin G. G. Aitken. 1976. Multivariate binary discrimination by kernel method. *Biometrika*, **63**, pp. 413–420.

Becker, Gary S. 1981. *A Treatise on the Family* (enlarged edition). Cambridge: Harvard University Press.

Dehejia, Rajeev H. 2003. "Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data." *Journal of Business and Economic Statistics*, **21**, pp. 1–11.

Eren, Ozkan, and Daniel J. Henderson. 2008. "The Impact of Homework on Student Achievement." *Econometrics Journal*, **11**, pp. 326–348.

Henderson, Daniel J. 2010. "A Test for Multimodality of Regression Derivatives with Application to Nonparametric Growth Regressions." *Journal of Applied Econometrics*, **25**, pp. 458–480.

Henderson, Daniel J., Christopher F. Parmeter, and Subal C. Kumbhakar. 2012. "A Simple Method to Visualize Results in Nonlinear Regression Models." *Economics Letters*, **117**, pp. 578–581.

Hotz, V. Joseph, Guido W. Imbens, and Jacob A. Klerman. 2006. "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program." *Journal of Labor Economics*, **24**, pp. 521–566.

Hurvich, Clifford M., Jeffrey S. Simonoff, and Chih-Ling Tsai. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society, Series B*, **60**, pp. 271–293.

Li, Qi, and Jeffrey S. Racine. 2004. "Cross-Validated Local Linear Nonparametric Regression." *Statistica Sinica*, **14**, pp. 485–512.

Linton, Oliver, Esfandiar Maasoumi, and Yoon-Jae Whang. 2005. "Consistent Testing for Stochastic Dominance under General Sampling Schemes." *Review of Economic Studies*, **72**, pp. 735–765, Also the Corrigendum to the same, 2007.

Maasoumi, Esfandiar. 2001. "Parametric and Nonparametric Tests of Limited Domain and Ordered Hypotheses in Economics." Chapter 25 in *A Companion to Econometric Theory*, ed. Badi Baltagi. Malden: Basil Blackwell Publishers.

Maasoumi, Esfandiar, Jeffrey S. Racine, and Thanasis Stengos. 2007. "Growth and Convergence: A Profile of Distribution Dynamics and Mobility." *Journal of Econometrics*, **136**, pp. 483–508.

Maasoumi, Esfandiar, Daniel L. Millimet, and Dipa Sarkar. 2009. "A Distributional Analysis of Marriage Effects." *Oxford Bulletin of Economics and Statistics*, **71**, pp. 1–33.

Nelson, Doug. 1997. "Some 'Best Practices' and 'Most Promising Models' for Welfare Reform." memorandum, Baltimore: Annie E. Casey Foundation.

Racine, Jeffrey S., and Qi Li. 2004. "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data." *Journal of Econometrics*, **119**, pp. 99–130.

Wang, Min-Chiang, and John van Ryzin. 1981. A class of smooth estimators for discrete estimation. *Biometrika*, **68**, pp. 301–9.

Author Index

Subject Index