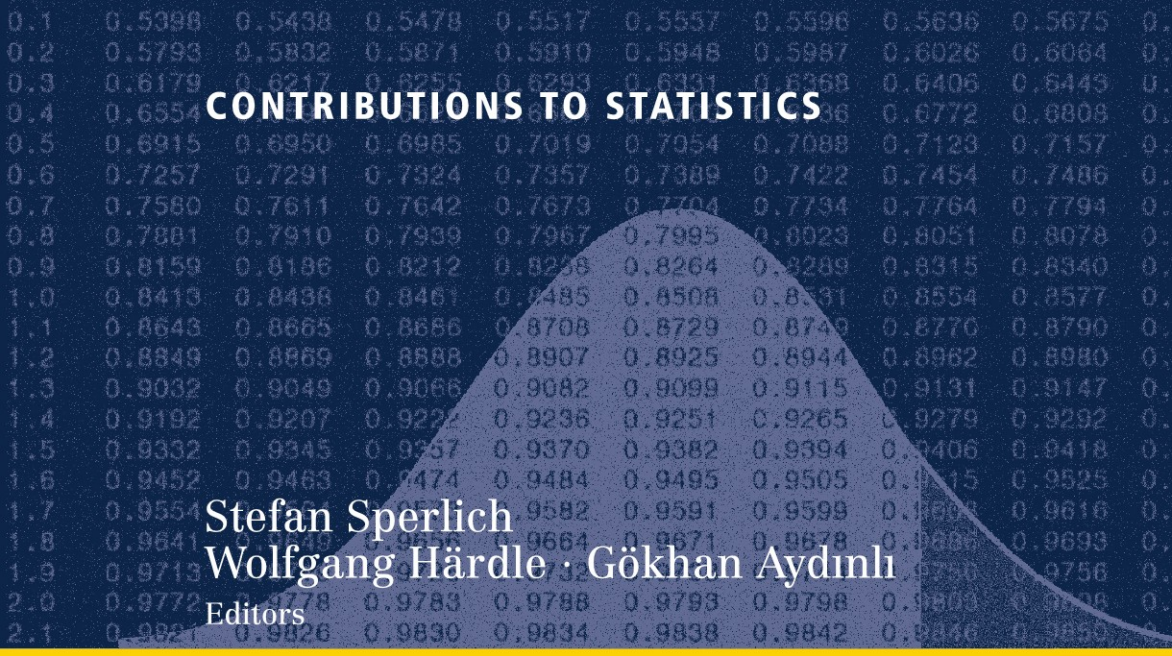
CONTRIBUTIONS TO STATISTICS



Stefan Sperlich
Wolfgang Härdle · Gökhan Aydınlı
Editors

The Art of Semiparametrics

Physica-Verlag
A Springer Company

The Art of Semiparametrics

Contributions to Statistics

V. Fedorov/W.G. Müller/I. N. Vuchkov (Eds.)
Model-Oriented Data Analysis,
XII/248 pages, 1992

J. Antoch (Ed.)
Computational Aspects of Model Choice,
VII/285 pages, 1993

W. G. Müller/H. P. Wynn/A. A. Zhigljavsky (Eds.)
Model-Oriented Data Analysis,
XIII/287 pages, 1993

P. Mandl/M. Hušková (Eds.)
Asymptotic Statistics,
X/474 pages, 1994

P. Dirschedl/R. Ostermann (Eds.)
Computational Statistics,
VII/553 pages, 1994

C. P. Kitsos/W. G. Müller (Eds.)
MODA 4 – Advances in Model-Oriented Data Analysis,
XIV/297 pages, 1995

H. Schmidli
Reduced Rank Regression,
X/179 pages, 1995

W. Härdle/M. G. Schimek (Eds.)
Statistical Theory and Computational Aspects of Smoothing,
VIII/265 pages, 1996

S. Klinke
Data Structures for Computational Statistics, VIII/284 pages, 1997

C. P. Kitsos/L. Edler (Eds.)
Industrial Statistics,
XVIII/302 pages, 1997

A. C. Atkinson/L. Pronzato/H. P. Wynn (Eds.)
MODA 5 – Advances in Model-Oriented Data Analysis and Experimental Design,
XIV/300 pages, 1998

M. Moryson
Testing for Random Walk Coefficients in Regression and State Space Models,
XV/317 pages, 1998

S. Biffignandi (Ed.)
Micro- and Macrodata of Firms,
XII/776 pages, 1999

W. Härdle/Hua Liang/J. Gao
Partially Linear Models,
X/203 pages, 2000

A. C. Atkinson/P. Hackl/W. Müller (Eds.)
MODA 6 – Advances in Model-Oriented Design and Analysis,
XVI/283 pages, 2001

W. G. Müller
Collecting Spatial Data, 2nd edition,
XII/195 pages, 2001

C. Lauro/J. Antoch/V. Esposito Vinzi/G. Saporta (Eds.)
Multivariate Total Quality Control,
XIII/236 pages, 2002

P.-A. Monney
A Mathematical Theory of Arguments for Statistical Evidence
XIII/154 pages, 2003

Y. Haitovsky/H. R. Lerche/Y. Ritov (Eds.)
Foundations of Statistical Inference,
XI/230 pages, 2003

A. Di Bucchiancio/H. Läufer/H. P. Wynn (Eds.)
MODA 7 – Advances in Model-Oriented Design and Analysis,
XIII/240 pages, 2004

Stefan Sperlich · Wolfgang Härdle
Gökhan Aydınlı (Editors)

The Art of Semiparametrics

With 33 Figures and 17 Tables

Physica-Verlag
A Springer Company

Editors
Professor Dr. Stefan Sperlich
Georg-August-Universität Göttingen
Institut für Statistik und Ökonometrie
Platz der Göttinger Sieben 5
37073 Göttingen
Germany
stefan@eco.uc3m.es

Professor Dr. Wolfgang Härdle
Humboldt-Universität zu Berlin
Institut für Statistik und Ökonometrie
CASE – Center for Applied Statistics and Economics
Spandauer Straße 1
10178 Berlin
Germany
stat@wiwi.hu-berlin.de

Gökhan Aydınlı
Hudson Advisors Germany GmbH
Hamburger Allee 14
60486 Frankfurt a. M.
Germany
gokhan@aydınlı.net

ISSN 1431-1968
ISBN-10 3-7908-1700-7 Physica-Verlag Heidelberg New York
ISBN-13 978-3-7908-1700-3 Physica-Verlag Heidelberg New York

Physica is part of Springer Science+Business Media GmbH

springer.com

Printed in Germany

Soft-Cover-Design: Erich Kirchner, Heidelberg

SPIN 11674382 154/3153-5 4 3 2 1 0 – Printed on acid-free and non-aging paper

Preface

This selection of articles has emerged from different works presented at the conference "THE ART OF SEMIPARAMETRICS" held in 2003 in Berlin. The idea was to bring together junior and senior researchers but also practitioners working on semiparametric statistics in rather different fields. The meeting succeeded in welcoming a group that presented a broad range of areas where research on, respectively with semiparametric methods is going on. It contains mathematical statistics, applied economics, finance, business statistics, etc. and thus combines theoretical contributions with applied statistics and finally empirical studies. Although each article represents an original contribution to its own field, they all are written in a self-contained way to be read also by non-experts of the particular topic. This volume therefore offers a collection of individual works that together show the actual large spectrum of semiparametric statistics. We hope very much you will enjoy reading this special collection of selected articles.

Madrid, February 2006

Stefan Sperlich

Wolfgang Härdle

Gökhan Aydınlı

Contents

1 Asymptotic Theory for M-Estimators of Boundaries

Keith Knight[1]

Department of Statistics, University of Toronto, ON, Canada

Summary

We consider some asymptotic distribution theory for M-estimators of the parameters of a linear model whose errors are non-negative; these estimators are the solutions of constrained optimization problems and their asymptotic theory is non-standard. Under weak conditions on the distribution of the errors and on the design, we show that a large class of estimators have the same asymptotic distributions in the case of i.i.d. errors; however, this invariance does not hold under non-i.i.d. errors.

Keywords: Constrained optimization, epi-convergence, linear programming estimator, M-estimator, point processes.

1.1 Introduction

Consider the linear regression model

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + W_i \quad (i = 1, \cdots, n) \tag{1.1.1}$$

where $\mathbf{x}_i$ is a vector of covariates (of length p) whose first component is always 1, $\boldsymbol{\beta}$ is a vector of parameters and $W_1, \cdots, W_n$ are independent, identically distributed (i.i.d.) non-negative random variables whose essential infimum is 0. Thus $\mathbf{x}^T\boldsymbol{\beta}$ can be interpreted as the conditional minimum of the response Y given covariate values $\mathbf{x}$. (The assumption that the model (1.1.1) has an intercept is not always necessary in the sequel but will be assumed throughout as its inclusion reflects common practice.)

Suppose that the W_i's have common density

$$f(w) = \exp[-\rho(w)] \quad \text{for } w > 0$$

[1]Research supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

where $\rho(w) \to +\infty$ as $w \to \infty$. If ρ is assumed known and is lower semicontinuous (note that a lower semicontinuous version of ρ typically exists) then the maximum likelihood estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}_n$, minimizes

$$\sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^T \boldsymbol{\varphi}) \quad \text{subject to } Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi} \text{ for } i = 1, \cdots, n. \tag{1.1.2}$$

This type of estimator seems to have first been considered by Aigner & Chu (1968) for estimating the so-called "efficient frontier"; they considered $\rho(w) = w^2$ and $\rho(w) = w$. In a recent paper, Florens & Simar (2002) comment on the lack of development of statistical properties of these estimators. An estimator minimizing (1.1.2) seems to be sensible estimator of $\boldsymbol{\beta}$ generally for non-negative W_i's.

In fact, the asymptotics of $\widehat{\boldsymbol{\beta}}_n$ appear to have only been considered in the case where $\rho(w) = w$; in this case, $\widehat{\boldsymbol{\beta}}_n$ minimizes

$$-\sum_{i=1}^{n} \mathbf{x}_i^T \boldsymbol{\varphi} \quad \text{subject to } Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi} \text{ for } i = 1, \cdots, n, \tag{1.1.3}$$

which is a linear program. This estimator can also be viewed as a minimum regression quantile estimator as defined by KB78 (1978). Limit theory for the estimator minimizing (1.1.3) can be derived under weak assumptions on the behaviour of the distribution of the W_i's near 0 and the behaviour of the empirical distribution of the $\mathbf{x}_i$'s; see, for example, Smith (1994), Portnoy & Jurečková (1999) and Knight (2001). A similar estimation method has been studied by (among others) Anděl (1989), An & Huang (1993) and Feigen & Resnick (1994) in the context of estimation in stationary autoregressive models with non-negative innovations; Feigen & Resnick (1994) derive limiting distributions of the estimators using an approach that relies heavily on point process arguments. For the general first order autoregressive model, Nielsen & Shephard (2003) derive the exact distribution of this estimator when the innovations have an exponential distribution.

In this paper, we will study the dependence of estimators minimizing (1.1.2) on the loss function ρ. Figure 1.1 shows the yearly best men's 1500m times from 1957 to 2002 with lower boundaries (which might be interpreted as the best possible time for a given year) estimated using $\rho(w) = w$ and $\rho(w) = w^2$; in both cases, we use a b-spline basis with four knots, which means that the parameter vector $\boldsymbol{\beta}$ has five elements (including an intercept). For these data, the two estimates are quite close although not identical; it is natural to ask whether this phenomenon occurs more generally. Note that the estimate for $\rho(w) = w$ is not strictly decreasing; depending on our interpretation of the lower boundary, it may be more natural to constrain the estimation so the estimate of the lower boundary is strictly decreasing.

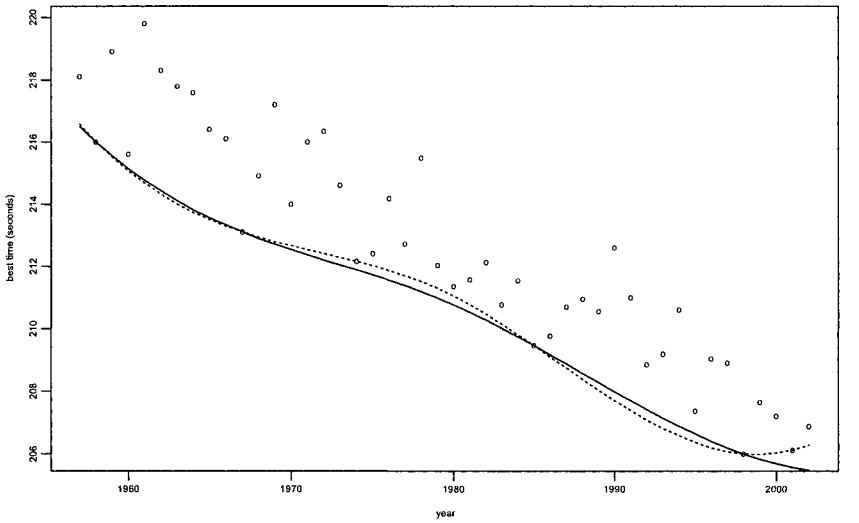


Figure 1.1: Yearly best men's outdoor 1500m times (in seconds) from 1957 to 2002 with estimated boundary lines using $\rho(w) = w$ (dotted) and $\rho(w) = w^2$ (solid).

In the i.i.d. setting (where $Y_i = \theta + W_i$), the analysis is straightforward to do. If $\rho(w)$ is increasing for $w \geq 0$ then the estimator is simply $\widehat{\theta}_n = \min(Y_1, \cdots, Y_n)$. More generally, suppose that $\rho(w)$ is convex and differentiable though not necessarily increasing for $w \geq 0$. Then the estimator is again $\min(Y_1, \cdots, Y_n)$ unless there exists $\widehat{\theta}_n < \min(Y_1, \cdots, Y_n)$ such that

$$\sum_{i=1}^{n} \rho'(Y_i - \widehat{\theta}_n) = 0.$$

Suppose that $\min(Y_1, \cdots, Y_n) \xrightarrow{p} \theta$ and set $W_i = Y_i - \theta$. If $E[\rho'(W_i)] > 0$ then by convexity of ρ, it follows that $E[\rho'(W_i + t)] > 0$ for $t \geq 0$ and so

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} [\rho(W_i + t) - \rho(W_i)] &= \int_0^t \frac{1}{n}\sum_{i=1}^{n} \rho'(W_i + s)\, ds \\ &\xrightarrow{a.s.} \int_0^t E[\rho'(W_i + s)]\, ds > 0. \end{aligned}$$

From this we can conclude that $\widehat{\theta}_n$ is eventually equal to $\min(Y_1, \cdots, Y_n)$ if $E[\rho'(W_i)] > 0$.

The purpose of this paper is to extend the equivalence in an asymptotic sense to the regression case under general conditions on the $\mathbf{x}_i$'s and the distribution of the W_i's; in particular, we will not assume any relationship between the density of the W_i's and the loss function ρ. We will also show that the asymptotic equivalence does not necessarily hold for non-i.i.d. errors.

1.2 Asymptotics

As in Knight (2001), the key tools used in deriving the limiting distribution of $\widehat{\boldsymbol{\beta}}_n$ minimizing (2) are epi-convergence in distribution (Pflug 1994, Pflug 1995, Geyer 1994, Geyer 1996, Knight 1999) and point process convergence for extreme values (Kallenberg 1983, Leadbetter *et al.* 1983). Point processes defined on an appropriate space can be characterized by random measures that count the (random) number of points lying in subsets of that space; point process convergence is characterized by the weak convergence of integrals of bounded continuous functions with compact support with respect to the random measures (Kallenberg 1983). Under appropriate regularity conditions (described below), the configuration of points $\{(\mathbf{x}_i, Y_i)\}$ generated from (1.1.1) lying in a neighbourhood of the plane $\mathbf{x}^T\boldsymbol{\beta}$ can be approximated (in a distributional sense) by a Poisson process when n is large and the asymptotic behaviour of $\widehat{\boldsymbol{\beta}}_n$ (perhaps not surprisingly) turns out to depend on this Poisson process.

Epi-convergence in distribution gives us an elegant way of proving convergence in distribution of "argmin" (and "argmax") estimators, and is particularly useful for constrained estimation procedures. A sequence of random lower semicontinuous functions $\{Z_n\}$ epi-converges in distribution to Z ($Z_n \xrightarrow{e-d} Z$) if for any closed rectangles $R_1, \cdots, R_k$ with open interiors $R_1^o, \cdots, R_k^o$ and any real numbers $a_1, \cdots, a_k$,

$$
\begin{aligned}
& P\left\{\inf_{\mathbf{u}\in R_1} Z(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z(\boldsymbol{u}) > a_k\right\} \\
& \leq \liminf_{n\to\infty} P\left\{\inf_{\mathbf{u}\in R_1} Z_n(\boldsymbol{u}) > a_1, \cdots, \inf_{\mathbf{u}\in R_k} Z_n(\boldsymbol{u}) > a_k\right\} \\
& \leq \limsup_{n\to\infty} P\left\{\inf_{\mathbf{u}\in R_1^o} Z_n(\boldsymbol{u}) \geq a_1, \cdots, \inf_{\mathbf{u}\in R_k^o} Z_n(\boldsymbol{u}) \geq a_k\right\} \\
& \leq P\left\{\inf_{\mathbf{u}\in R_1^o} Z(\boldsymbol{u}) \geq a_1, \cdots, \inf_{\mathbf{u}\in R_k^o} Z(\boldsymbol{u}) \geq a_k\right\}.
\end{aligned}
$$

For an extended real-valued lower-semicontinuous function g, define

$$\begin{aligned} \text{argmin}(g) &= \left\{ \boldsymbol{u}_0 : g(\boldsymbol{u}_0) = \inf_{\mathbf{u}} g(\boldsymbol{u}) \right\} \\ \epsilon - \text{argmin}(g) &= \left\{ \boldsymbol{u}_0 : g(\boldsymbol{u}_0) \leq \inf_{\mathbf{u}} g(\boldsymbol{u}) + \epsilon \right\}. \end{aligned}$$

Suppose that $\boldsymbol{U}_n \in \text{argmin}(Z_n)$ where $Z_n \xrightarrow{e-d} Z$ and $\boldsymbol{U}_n = O_p(1)$; then $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U} = \text{argmin}(Z)$ provided that $\text{argmin}(Z)$ is (with probability 1) a singleton. (The condition that $\boldsymbol{U}_n \in \text{argmin}(Z_n)$ can be weakened to $\boldsymbol{U}_n \in \epsilon_n - \text{argmin}(Z_n)$ where $\epsilon_n \xrightarrow{p} 0$.) If the Z_n's are convex (as will be the case here) then epi-convergence is quite simple to prove; finite dimensional convergence in distribution of Z_n to Z ($Z_n \xrightarrow{f-d} Z$) is sufficient for epi-convergence in distribution provided that Z is finite on an open set. (In fact, it is sufficient to prove this finite dimensional convergence on a countable dense subset.) Moreover in the case of convexity, if $\text{argmin}(Z)$ is a singleton then $\boldsymbol{U}_n = O_p(1)$ is implied by $Z_n \xrightarrow{e-d} Z$.

In order to consider the asymptotics of estimators minimizing (1.1.2), we need to make some mild assumptions. We will assume that ρ in (1.1.2) is a convex function with

$$\rho(w) = \int_0^w \psi(t)\, dt \tag{1.2.1}$$

for some non-decreasing function ψ satisfying

$$|\psi(w+t) - \psi(w)| \leq M(w)|t|^{\delta} \tag{1.2.2}$$

for $w > 0$ and $|t| \leq \epsilon$ where $\delta > 0$. In addition, we will make the following assumptions about the design and the distributions of the W_i's:

(A0) For ψ defined in (4), $E[\psi(W_i)] > 0$ and $E[\psi^2(W_i)] < \infty$.

(A1) For some sequence $a_n \to \infty$, we have

$$|nP\left(a_n W_i \leq t\right) - t^{\alpha}| \leq \tau_n t^{\alpha}$$

where $\alpha > 0$ and $\tau_n \to 0$.

(A2) There exists a sequence of matrices $\{C_n\}$ and a probability measure μ on R^p such that for each set B with $\mu(\partial B) = 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(C_n^{-1}\mathbf{x}_i \in B) = \mu(B).$$

(A3) $\int \|\mathbf{x}\| \, \mu(d\mathbf{x}) < \infty$ with

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} C_n^{-1}\mathbf{x}_i &= \int \mathbf{x} \, \mu(d\mathbf{x}) = \boldsymbol{\gamma}, \\ \lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^{n} \|C_n^{-1}\mathbf{x}_i\|^2 &= 0. \end{aligned}$$

(A4) $\mu(D_\gamma) = 0$ where

$$D_\gamma = \left\{ \mathbf{x} : \mathbf{x}^T \boldsymbol{c} = 0 \text{ for some } \boldsymbol{c} \neq \mathbf{0} \text{ with } \boldsymbol{\gamma}^T \boldsymbol{c} = 0 \right\}$$

where $\boldsymbol{\gamma}$ is defined in (A3).

(A5) The (closed) set

$$K = \left\{ \boldsymbol{u} : \int (\boldsymbol{u}^T \mathbf{x})_+^\alpha \, \mu(d\mathbf{x}) < \infty \right\}$$

has an open interior and for each $\boldsymbol{u} \in \text{int}(K)$,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{u}^T C_n^{-1}\mathbf{x}_i)_+^\alpha &= \int (\boldsymbol{u}^T \mathbf{x})_+^\alpha \, \mu(d\mathbf{x}) \\ \lim_{n\to\infty} \frac{1}{n} \max_{1\le i\le n} (\boldsymbol{u}^T C_n^{-1}\mathbf{x}_i)_+^\alpha &= 0 \end{aligned}$$

where $x_+ = \max(x, 0)$ denotes the positive part of x.

(A6) $E[M(W_i)] < \infty$ and

$$\lim_{n\to\infty} \frac{1}{a_n^\delta} \max_{1\le i\le n} \|C_n^{-1}\mathbf{x}_i\|^{\delta+1} = 0$$

where $M(\bullet)$ and δ are defined as in (1.2.2).

It is worth commenting at this point on the *raison d'être* of these conditions. The first part of condition (A0) is essentially necessary for consistency; if $E[\psi(W_i)] < 0$ then $\widehat{\boldsymbol{\beta}}_n$ will not converge to $\boldsymbol{\beta}$. Condition (A1) generalizes the condition on the density of the W_i's assumed in Smith (1994) and implies that the W_i's are in the domain of attraction of a type III extreme value distribution. Condition (A2) is effectively a weak convergence condition for the empirical distribution of the $\mathbf{x}_i$'s; if the $\mathbf{x}_i$'s are a random sample from some distribution then we would have $C_n = I$ and μ equal to the underlying probability measure of the $\mathbf{x}_i$'s. Even for fixed designs, (A2) is a

reasonable condition although C_n need not equal I (although it is typically a diagonal matrix). For example, if $\mathbf{x}_i = (1, i, i^2)^T$ for $i = 1, \cdots, n$ then the diagonal elements of C_n are $(1, n, n^2)$ and μ is the probability measure of the random vector $(1, U, U^2)$ where U is uniformly distributed on $[0, 1]$. More importantly, (A2) implies similar weak convergence results about the empirical distribution of $\boldsymbol{u}^T C_n^{-1} \mathbf{x}_i$ $(i = 1, \cdots, n)$ for a given $\boldsymbol{u}$ (or finite number of $\boldsymbol{u}$'s). Moreover, if $C_n^{-1} \mathbf{x}_i$ is bounded then condition (A1) can be replaced by

$$nP\left(a_n W_i \leq t\right) \rightarrow t^\alpha$$

for each $t > 0$. Conditions (A3)–(A5) are used to facilitate the proof of epi-convergence in distribution of an appropriate sequence of objective functions; for example, (A4) will imply that the limiting objective function has a unique minimizer (with probability 1) while (A5) will imply that the limiting objective function is finite on a open set and so finite dimensional weak convergence will imply epi-convergence in distribution. (In fact, condition (A5) is not necessary and is included only to simplify the proof.) Condition (A6) together with condition (A3) allows us to approximate the finite part of the objective function by a linear function.

Note that conditions (A3), (A5), and (A6) are essentially moment conditions on the $\mathbf{x}_i$'s (or, more precisely, on the $C_n^{-1} \mathbf{x}_i$'s); depending on the value of α, one of these conditions may imply all or part of the others. The conditions as stated are certainly far from minimal and can be weakened or modified.

THEOREM 1.1 *Assume the model (1.1.1) and suppose that $\widehat{\boldsymbol{\beta}}_n$ minimizes (1.1.2) where ρ is convex and satisfies (1.2.1) and (1.2.2). If conditions (A0)–(A6) hold then*

$$a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ is the solution of the linear programming problem:

$$\textit{maximize} \quad \boldsymbol{u}^T \boldsymbol{\gamma} \quad \textit{subject to } \Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i \textit{ for } i = 1, 2, 3, \cdots$$

where

(i) $\Gamma_i = (E_1 + \cdots + E_i)^{1/\alpha}$ *for unit mean i.i.d. exponential random variables* $E_1, E_2, \cdots$;

(ii) $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots$ *are i.i.d. with distribution* $P(\boldsymbol{X}_i \in A) = \mu(A)$;

(iii) the $\boldsymbol{X}_i$'s are independent of the E_i's (and hence of the Γ_i's).

Proof. The proof follows much along the lines of the proof of Theorem 1 of Knight (2001). First of all, note that $\boldsymbol{U}_n = a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ is the solution to the linear programming problem:

$$\begin{aligned} \text{minimize} \quad & \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n) - \rho(W_i)\right] \\ \text{subject to} \quad & a_n W_i \geq \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i \quad \text{for } i = 1, \cdots, n. \end{aligned}$$

Defining $\varphi_n(\boldsymbol{u})$ to be 0 when the constraints above are all satisfied and $+\infty$ otherwise, $\boldsymbol{U}_n$ minimizes

$$Z_n(\boldsymbol{u}) = \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n) - \rho(W_i)\right] + \varphi_n(\boldsymbol{u}). \tag{1.2.3}$$

Z_n is a convex function for each n and so to prove that $\boldsymbol{U}_n \xrightarrow{d} \boldsymbol{U}$, it suffices to show that $Z_n \xrightarrow{e-d}$ some Z where $\boldsymbol{U} = \text{argmin}(Z)$; we will show that

$$Z(\boldsymbol{u}) = -E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + \varphi(\boldsymbol{u}) \tag{1.2.4}$$

where $\varphi(\boldsymbol{u}) = 0$ if $\Gamma_i \geq \boldsymbol{u}^T\boldsymbol{X}_i$ for all i and $\varphi(\boldsymbol{u}) = +\infty$ otherwise.

Using the integral representation (1.2.1) for ρ and condition (1.2.2), we obtain

$$\begin{aligned} & \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n) - \rho(W_i)\right] \\ &= -\frac{1}{n}\sum_{i=1}^{n}\psi(W_i)\boldsymbol{u}^T C_n^{-1}\mathbf{x}_i + o_p(1) \\ &= -E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + o_p(1) \end{aligned}$$

using condition (A3) to establish the weak law of large numbers and condition (A6) to establish the asymptotic linearity. From the convexity of Z_n, $Z_n \xrightarrow{e-d} Z$ follows from $Z_n \xrightarrow{f-d} Z$ provided that Z is finite on an open set with probability 1; the latter follows since $\Gamma_i \sim i^{1/\alpha}$ (with probability 1) as $i \to \infty$ and so by the first Borel-Cantelli lemma $P(\boldsymbol{u}^T\boldsymbol{X}_i > \Gamma_i \text{ infinitely often}) = 0$ for any $\boldsymbol{u} \in K$ (since $E[(\boldsymbol{u}^T\boldsymbol{X}_i)_+^\alpha] < \infty$ on K); for $\boldsymbol{u} \notin K$, we also have $P(\boldsymbol{u}^T\boldsymbol{X}_i > \Gamma_i \text{ infinitely often}) = 1$ (since $E[(\boldsymbol{u}^T\boldsymbol{X}_i)_+^\alpha] = \infty$) by the second Borel-Cantelli lemma. Thus for a given $\boldsymbol{u} \in K$, at most a finite number of constraints are violated, the rest being trivially satisfied. Since $\boldsymbol{u} \in K$ implies that $t\boldsymbol{u} \in K$ for $t > 0$, taking t sufficiently small guarantees that all the constraints are satisfied. Since $\text{int}(K)$ is open (by condition (A5)), it is possible (with probability 1) to find a finite number of points in K such that all the constraints are satisfied and the convex hull of these points contains

an open set. Since Z is finite at these points, it is necessarily finite on the convex hull (since Z is convex).

To show the finite dimensional weak convergence of φ_n, we first define the following point process (random measure) on R^{p+1}:

$$N_n(A \times B) = \sum_{i=1}^{n} I(a_n W_i \in A, C_n^{-1}\mathbf{x}_i \in B).$$

It is easy to verify that N_n tends in distribution with respect to the vague topology (Kallenberg 1983) to a Poisson process (random measure) N_0 whose mean measure is

$$E[N_0(A \times B)] = \mu(B) \int_{A \cap (0,\infty)} \alpha x^{\alpha-1}\, dx.$$

We can represent the points of this Poisson process by $\{(\Gamma_i, \boldsymbol{X}_i) : i \geq i\}$ where the Γ_i's and $\boldsymbol{X}_i$'s are as defined above. Thus it suffices to show that

$$P\left[\varphi_n(\boldsymbol{u}_1) = 0, \cdots, \varphi_n(\boldsymbol{u}_k) = 0\right] \to P\left[\varphi(\boldsymbol{u}_1) = 0, \cdots, \varphi(\boldsymbol{u}_k) = 0\right]$$

where $\varphi(\boldsymbol{u}) = 0$ if $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i$ for all i and ∞ otherwise. Exploiting the convergence in distribution of N_n to the Poisson random measure N_0, we have

$$\begin{aligned}
&P\left[\varphi_n(\boldsymbol{u}_1) = 0, \cdots, \varphi_n(\boldsymbol{u}_k) = 0\right] \\
&= P\left\{\sum_{i=1}^{n} I\left[0 \leq a_n W_i < \max_{1\leq j\leq k} (\boldsymbol{u}_j^T C_n^{-1}\mathbf{x}_i)_+\right] = 0\right\} \\
&\to \exp\left[-\int \max_{1\leq j\leq k} (\boldsymbol{u}_j^T \mathbf{x})_+^{\alpha}\, \mu(d\mathbf{x})\right] \\
&= P\left[\varphi(\boldsymbol{u}_1) = 0, \cdots, \varphi(\boldsymbol{u}_k) = 0\right].
\end{aligned}$$

Hence for Z_n given in (1.2.3), we have $Z_n \xrightarrow{f-d} Z$ where Z is defined in (1.2.4). Finally, to show that Z has a unique minimizer (with probability 1), we note that if $\boldsymbol{U}$ minimizes Z then for some indices $i_1 < i_2 < \cdots < i_p$, we have $\boldsymbol{U}^T \boldsymbol{X}_{i_k} = \Gamma_{i_k}$ with $\Gamma_j > \boldsymbol{U}^T \boldsymbol{X}_j$ for $j \notin \{i_1, i_2, \cdots, i_p\}$. If $\boldsymbol{U}$ and $\boldsymbol{U}^*$ both minimize Z then $\boldsymbol{U}^* = \boldsymbol{U} + t\boldsymbol{c}$ for some vector $\boldsymbol{c}$ with $\boldsymbol{c}^T\boldsymbol{\gamma} = 0$ and so $t\boldsymbol{c}^T \boldsymbol{X}_{i_k} = 0$ for $k = 1, \cdots, p$. However, condition (A4) says that $P(\boldsymbol{c}^T \boldsymbol{X}_i = 0) = 0$ (when $\boldsymbol{c}^T\boldsymbol{\gamma} = 0$) and so Z is a unique minimizer (with probability 1). □

As mentioned above, the conclusion of Theorem 1.1 holds even if the set K defined in condition (A5) does not have an open interior. In this case,

the limiting objective function Z will not be finite on an open set and so $Z_n \xrightarrow{e-d} Z$ will not follow immediately from $Z_n \xrightarrow{f-d} Z$. However, the epi-convergence in distribution will still hold; we need to establish that the sequence of functions $\{\varphi_n\}$ (which describe the constraints) are stochastically equi-lower-semicontinuous (Knight 1999). For this, we need to show that for any bounded set B and $\delta > 0$, there exist points $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m$ in B and open neighbourhoods $V(\boldsymbol{u}_1), \cdots, V(\boldsymbol{u}_m)$ of these points such that

$$B \subset \bigcup_{i=1}^{m} V(\boldsymbol{u}_i)$$

(that is, B is covered by the neighbourhoods) and

$$\limsup_{n\to\infty} P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in V(\mathbf{u}_i)} \varphi_n(\boldsymbol{u}) = 0, \varphi_n(\boldsymbol{u}_i) = \infty\right]\right\} < \delta.$$

This turns out to be reasonably straightforward to show since

$$\begin{aligned} & P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in V(\mathbf{u}_i)} \varphi_n(\boldsymbol{u}) = 0, \varphi_n(\boldsymbol{u}_i) = \infty\right]\right\} \\ \leq\ & P\left\{\bigcup_{i=1}^{m}\left[\inf_{\mathbf{u}\in V(\mathbf{u}_i)} \varphi_n(\boldsymbol{u}) = 0\right]\right\} - P\left\{\bigcup_{i=1}^{m}\left[\varphi_n(\boldsymbol{u}_i) = 0\right]\right\}. \end{aligned}$$

The right hand side above can be made arbitrarily small by making the neighbourhoods uniformly small.

In the case where $\rho(w) = w$, Smith (1994) as well as Portnoy & Jurečková (1999) determine the limiting distribution by finding the limiting density of $\boldsymbol{U}_n = \text{argmin}(Z_n)$; however, they need to assume a specific form for the density of the W_i's, from which the density of $\boldsymbol{U}_n$ can be approximated. The conclusion of Theorem 1.1 holds under a weak assumption (condition (A1)) about the distribution of the W_i's, which in particular does not imply the existence of a density function. Chernozhukov (2000) also uses an epi-convergence approach to study the asymptotic behaviour of "near extreme" regression quantile estimators. Other estimation problems in which the limiting objective function is related to a Poisson process are considered by Pflug (1994).

In the case where the set K defined in (A5) satisfies $K = \text{cl}(\text{int}(K))$, we can determine the limiting joint density, that is, the density of $\boldsymbol{U} = \text{argmin}(Z)$. Using the Poisson process representation of Z, it follows that the density of $\boldsymbol{U}$ is

$$g(\boldsymbol{u}) = \kappa(\boldsymbol{u}; \alpha, p, \mu) \int \cdots \int |D(\mathbf{x}_1, \cdots, \mathbf{x}_p)| \prod_{i=1}^{p} \left\{(\boldsymbol{u}^T \mathbf{x}_i)_+^{\alpha-1} \mu(d\mathbf{x}_i)\right\} \quad (1.2.5)$$

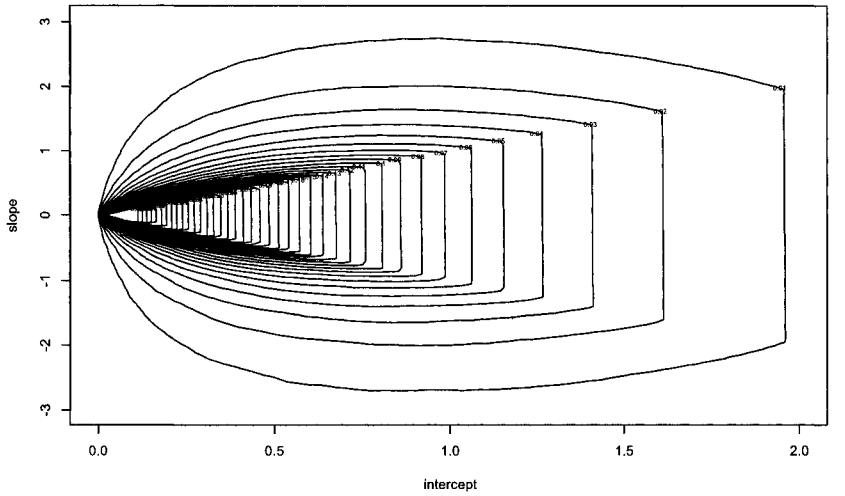


Figure 1.2: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0$; the interval between adjacent contours is 0.01.

where

$$\kappa(\boldsymbol{u}; \alpha, p, \mu) = \frac{\alpha^p}{p!} \exp\left[-\int (\boldsymbol{u}^T \mathbf{x})_+^\alpha \, \mu(d\mathbf{x})\right]$$

and $D(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is the determinant of the matrix with columns $\mathbf{x}_1, \cdots, \mathbf{x}_p$ if $\boldsymbol{\gamma}$ lies in the convex hull of $\mathbf{x}_1, \cdots, \mathbf{x}_p$ and $D(\mathbf{x}_1, \cdots, \mathbf{x}_p) = 0$ otherwise. (If there is no intercept in the model (1.1.1) then $D(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is the determinant if $\boldsymbol{\gamma} = \sum_{j=1}^p t_j \mathbf{x}_j$ for non-negative t_j's with $D(\mathbf{x}_1, \cdots, \mathbf{x}_p) = 0$ otherwise.) The density $g(\boldsymbol{u})$ is not easy to evaluate in closed-form (except in special cases) but can be approximated quite easily using Monte Carlo techniques to sample from the probability measure μ. However, it seems that this density does not provide as much insight into the limiting distribution as does the representation of $\boldsymbol{U}$ as the solution of a linear programming problem.

Theorem 1.1 implies that we obtain the same limiting distribution for any convex ρ satisfying some mild regularity conditions so that all such estimators differ by $o_p(a_n^{-1} C_n^{-1})$. However, an examination of the proof of Theorem 1.1 suggests that this asymptotic equivalence is a consequence of the i.i.d. assumption on the W_i's.

Suppose instead that we assume the W_i's in (1.1.1) are independent with the

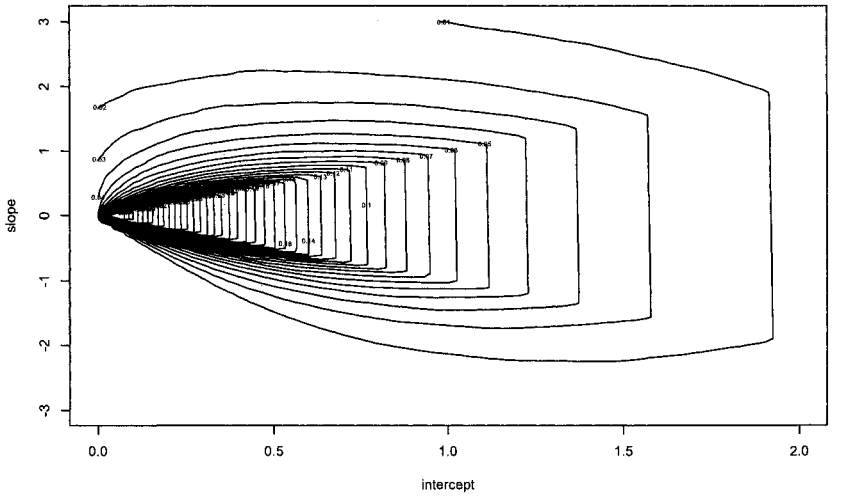


Figure 1.3: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.25$; the interval between adjacent contours is 0.01.

distribution of W_i depending on $\mathbf{x}_i$ such that

$$|nP(a_n W_i \le t|\mathbf{x}_i) - \lambda(\mathbf{x}_i)t^\alpha| \le \tau_n(\mathbf{x}_i)t^\alpha$$

where

$$\max_{1\le i\le n} |\tau_n(\mathbf{x}_i)| \to 0.$$

Under condition (A2) on the $\mathbf{x}_i$'s, it then follows that the point process

$$N_n(A\times B) = \sum_{i=1}^{n} I(a_n W_i \in A, C_n^{-1}\mathbf{x}_i \in B)$$

converges in distribution to a point process N_0 whose mean measure is given by

$$E[N_0(A\times B)] = \int_A \int_B \alpha\lambda(\mathbf{x})t^{\alpha-1}\,\mu(d\mathbf{x})\,dt.$$

The points of N_0 can be represented by $\{(\Gamma_i/\lambda(\boldsymbol{X}_i), \boldsymbol{X}_i) : i = 1, 2, \cdots\}$ where the Γ_i's and $\boldsymbol{X}_i$'s are defined as in Theorem 1.1. Assuming that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\psi(W_i)\boldsymbol{u}^T C_n^{-1}\mathbf{x}_i &\xrightarrow{p} \int E(\psi(W)|\mathbf{x})\boldsymbol{u}^T\mathbf{x}\,\mu(d\mathbf{x}) \\ &= \boldsymbol{u}^T\boldsymbol{\gamma}_\psi\end{aligned}$$

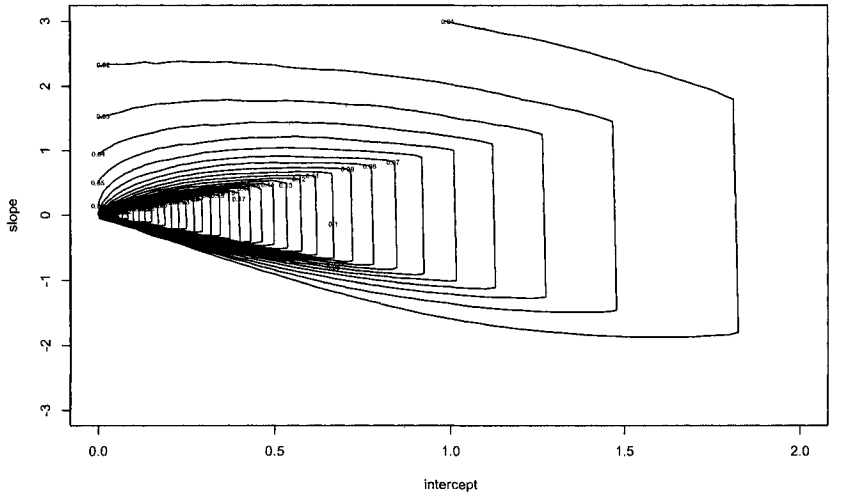


Figure 1.4: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.5$; the interval between adjacent contours is 0.01.

it will follow (under appropriate modifications of the regularity conditions) that

$$a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ maximizes $\boldsymbol{u}^T\boldsymbol{\gamma}_\psi$ subject to $\Gamma_i \geq \lambda(\boldsymbol{X}_i)\boldsymbol{u}^T\boldsymbol{X}_i$ for all i. Note that $\boldsymbol{U} = \boldsymbol{U}(\boldsymbol{\gamma}_\psi, N_0)$ where the point process N_0 does not depend on the loss function ρ (nor its "derivative" ψ).

We can extend (1.2.5) to obtain the density of $\boldsymbol{U}$ in this case:

$$g(\boldsymbol{u}) = \kappa_\lambda(\boldsymbol{u}; \alpha, p, \mu) \int\cdots\int |D_\lambda(\mathbf{x}_1, \cdots, \mathbf{x}_p)| \prod_{i=1}^{p} \left\{ [\lambda(\mathbf{x}_i)\boldsymbol{u}^T\mathbf{x}_i]_+^{\alpha-1} \mu(d\mathbf{x}_i) \right\} \tag{1.2.6}$$

where

$$\kappa_\lambda(\boldsymbol{u}; \alpha, p, \mu) = \frac{\alpha^p}{p!} \exp\left[-\int [\lambda(\mathbf{x})\boldsymbol{u}^T\mathbf{x}]_+^\alpha \mu(d\mathbf{x}) \right]$$

and $D_\lambda(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is the determinant of the matrix whose columns are $\lambda(\mathbf{x}_1)\mathbf{x}_1, \cdots, \lambda(\mathbf{x}_p)\mathbf{x}_p$ if

$$\boldsymbol{\gamma}_\psi = \sum_{j=1}^{p} t_j \lambda(\mathbf{x}_j)\mathbf{x}_j$$

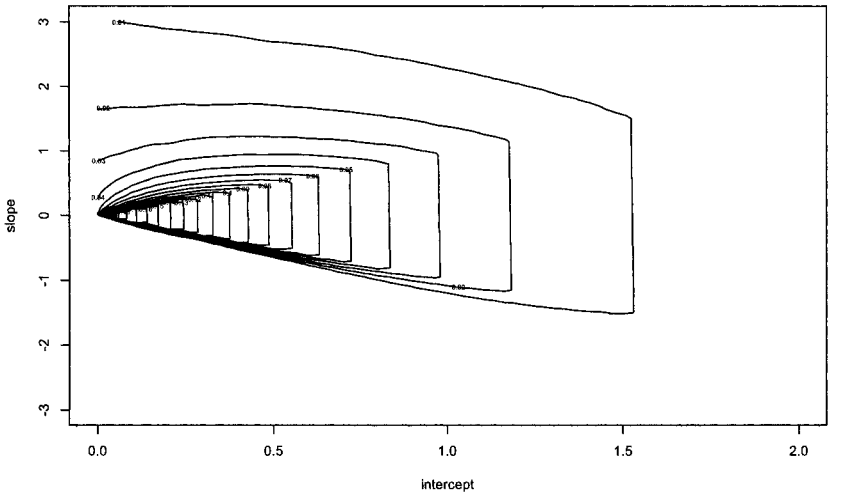


Figure 1.5: Contours of the joint density of $\boldsymbol{U}$ in Example 1 for $c_\psi = 0.75$; the interval between adjacent contours is 0.01.

for some non-negative $t_1, \cdots, t_p$ and $D_\lambda(\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is 0 otherwise.

EXAMPLE 1. Consider the simple regression model

$$Y_i = \beta_0 + \beta_1 x_i + W_i \quad (i = 1, \cdots, n)$$

where $W_1, \cdots, W_n$ are independent (but identically distributed) random variables with the distribution of W_i depending on x_i, and we will assume that the x_i's are uniformly distributed on the interval $[-1, 1]$, which implies that μ is a uniform distribution on $[-1, 1]$. For a given loss function ρ (with "derivative" ψ), the vector $\boldsymbol{\gamma}_\psi$ is simply

$$\boldsymbol{\gamma}_\psi = \frac{1}{2}\int_{-1}^{1} E[\psi(W)|x]\begin{pmatrix}1\\ x\end{pmatrix} dx = \left\{\int_{-1}^{1} E[\psi(W)|x]\, dx\right\}\begin{pmatrix}1\\ c_\psi\end{pmatrix}$$

where $-1 < c_\psi < 1$; note that for $\rho(x) = x$, $c_\psi = 0$. For simplicity, we will take $\alpha = 1$ and set $\lambda(x) = 1$ (which is possible even when the W_i's are not identically distributed). Thus for a given ρ (and corresponding ψ), we have

$$n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U} = \begin{pmatrix}U_0\\ U_1\end{pmatrix}$$

where $\boldsymbol{U}$ maximizes $u_0 + c_\psi u_1$ subject to $\Gamma_i \geq u_0 + u_1 X_i$ for $i \geq 1$ where the Γ_i's are partial sums of i.i.d. unit mean exponential random variables and the X_i's are i.i.d. uniform random variables on $[-1, 1]$. Thus the limiting distribution depends only on the constant c_ψ (which depends on ψ and the dependence between the W_i's and the x_i's). Figures 1.2 to 1.5 show contour plots of the joint density of $\boldsymbol{U}$ (using (1.2.6)) for $c_\psi = 0, 0.25, 0.5, 0.75$. In all cases, the distribution of U_0 (intercept) is concentrated on the positive part of the real line. As c_ψ increases, more probability mass is shifted to the positive part of the distribution of U_1, that is, the bias of the slope estimator becomes more positive as c_ψ increases; likewise, the bias becomes more negative as c_ψ decreases from 0 to -1.

1.3 Barrier Regularization

Estimators minimizing (1.1.2) are inherently biased upwards since necessarily we have

$$\sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n) < \sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^T \boldsymbol{\beta})$$

and so $\mathbf{x}^T \widehat{\boldsymbol{\beta}}_n$ tends to be systematically smaller than $\mathbf{x}^T \boldsymbol{\beta}$ (since ρ is "on average" increasing under condition (A0)). In general, reducing bias is a tricky proposition since such a reduction often leads to an increase in variance. In this problem, the bias typically manifests itself in the estimation of the intercept and so one might consider reducing bias simply by adjusting (downwards) the intercept estimator.

An alternative approach to reducing bias is to replace the constraints $Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi}$ $(i = 1, \cdots, n)$ in (1.1.2) by a "barrier" function that pushes the estimator away from the boundary of the constraint region. Specifically, we will define $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ to minimize

$$\sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^T \boldsymbol{\varphi}) + \epsilon \sum_{i=1}^{n} \tau(Y_i - \mathbf{x}_i^T \boldsymbol{\varphi}) \quad \text{subject to } Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi} \text{ for all } i \quad (1.3.1)$$

where ϵ is a positive constant and the barrier function $\tau(w)$ is a convex function on $(0, \infty)$ satisfying

$$\lim_{w \downarrow 0} \tau(w) = +\infty,$$

for example, $\tau(w) = w^{-r}$ for $r > 0$ or $\tau(w) = -\ln(w)$. It is easy to see that, for any $\epsilon > 0$, the minimizer of (1.3.1) will lie in the interior of the set $\{\boldsymbol{\varphi} : Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi} \text{ for } i = 1, \cdots, n\}$ and so if $\rho(w)$ and $\tau(w)$ are differentiable

for $w > 0$, it follows that $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ satisfies

$$\sum_{i=1}^{n} \left[\rho'(Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n(\epsilon)) + \epsilon\tau'(Y_i - \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_n(\epsilon))\right] \mathbf{x}_i = \mathbf{0}.$$

More importantly, by choosing $\epsilon = \epsilon_n$ appropriately, we may be able to reduce the bias of $\widehat{\boldsymbol{\beta}}_n(\epsilon_n)$ while retaining many of the otherwise attractive properties possessed by $\widehat{\boldsymbol{\beta}}_n$.

There is a connection between the estimators minimizing (1.1.2) and (1.3.1). If $\widehat{\boldsymbol{\beta}}_n(\epsilon)$ minimizes (1.3.1) and $\widehat{\boldsymbol{\beta}}_n$ minimizes (1.1.2) then

$$\lim_{\epsilon \downarrow 0} \widehat{\boldsymbol{\beta}}_n(\epsilon) = \widehat{\boldsymbol{\beta}}_n.$$

(This follows since the objective function implied by (1.3.1) epi-converges to the objective function implied by (1.1.2) as $\epsilon \downarrow 0$ for each fixed n.) This observation turns out to be useful in the computation of $\widehat{\boldsymbol{\beta}}_n$ minimizing (1.1.2). For each $\epsilon > 0$, (1.3.1) can be minimized using "standard" optimization techniques (for example, Newton and quasi-Newton methods) and so we can obtain an arbitrarily good approximation to $\widehat{\boldsymbol{\beta}}_n$ minimizing (1.1.2) by computing a sequence of minimizers of (1.3.1), $\widehat{\boldsymbol{\beta}}_n(\epsilon_k)$ with $\epsilon_k \downarrow 0$. Such numerical methods for solving constrained optimization problems are commonly referred to as barrier or interior point methods; some theory and discussion of these methods can be found in Fiacco & McCormick (1990).

By taking $\tau(w) = w^{-r}$ for r sufficiently large, we obtain the following analogue of Theorem 1.

THEOREM 1.2 *Assume the model (1.1.1) and suppose that $\widehat{\boldsymbol{\beta}}_n(\epsilon_n)$ minimizes (1.3.1) (with $\epsilon = \epsilon_n$) where ρ is convex and satisfies (1.2.1) and (1.2.2). If conditions (A0)-(A6) hold and $\tau(w) = w^{-r}$ where $r > \alpha$ and*

$$\lim_{n\to\infty} \frac{a_n^{r+1}}{n} \epsilon_n = \epsilon_0$$

then

$$a_n C_n(\widehat{\boldsymbol{\beta}}_n(\epsilon_n) - \boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{U}$$

where $\boldsymbol{U}$ minimizes

$$-E[\psi(W_1)]\boldsymbol{u}^T\boldsymbol{\gamma} + \epsilon_0 \sum_{i=1}^{\infty} (\Gamma_i - \boldsymbol{u}^T \boldsymbol{X}_i)^{-r}$$

subject to $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i$ for all i with $\{\Gamma_i\}$ and $\{\boldsymbol{X}_i\}$ defined as in Theorem 1.1.

Proof. The proof follows along the same lines as the proof of Theorem 1.1. We redefine Z_n in (1.2.3) by

$$\begin{aligned} Z_n(\boldsymbol{u}) &= \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n) - \rho(W_i)\right] \\ &\quad + \frac{a_n}{n}\epsilon_n \sum_{i=1}^{n}\left(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n\right)^{-r} \\ &= \frac{a_n}{n}\sum_{i=1}^{n}\left[\rho(W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i/a_n) - \rho(W_i)\right] \\ &\quad + \frac{a_n^{r+1}}{n}\epsilon_n \sum_{i=1}^{n}\left(a_n W_i - \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i\right)^{-r} \\ &= Z_n^{(1)}(\boldsymbol{u}) + Z_n^{(2)}(\boldsymbol{u}) \end{aligned}$$

provided that $a_n W_i \geq \boldsymbol{u}^T C_n^{-1}\mathbf{x}_i$ for all i with $Z_n(\boldsymbol{u}) = +\infty$ otherwise. The only technical complication lies in showing that $Z_n^{(2)} \xrightarrow{f-d} Z^{(2)}$ where

$$Z^{(2)}(\boldsymbol{u}) = \epsilon_0 \sum_{i=1}^{\infty}(\Gamma_i - \boldsymbol{u}^T \boldsymbol{X}_i)^{-r}$$

when $\Gamma_i \geq \boldsymbol{u}^T \boldsymbol{X}_i$ for all i with $Z^{(2)}(\boldsymbol{u}) = +\infty$ otherwise; this can be done by truncating the barrier function $\tau(w) = w^{-r}$ to make it bounded with compact support and then using Slutsky-type arguments to take care of the difference. □

The assumption that $r > \alpha$ is inconvenient but seems to be necessary in order to obtain non-degenerate asymptotic results, at least, with the "right" rate of convergence; if $\tau(w) \to \infty$ too slowly as $w \downarrow 0$ then typically we will obtain a slower convergence rate for the resulting estimators. In particular, it rules out the barrier function $\tau(w) = -\ln(w)$, which is quite useful for numerical computation.

Figure 1.6 shows the estimated boundaries for the 1500m data discussed in section 1 using $\rho(w) = w$ and $\tau(w) = w^{-2}$ in (1.3.1) with $\epsilon = 0.05$ and $\epsilon = 0.5$. The choice of ϵ for a given value of r is an open question; however, for these data, the estimates seem somewhat insensitive to the value of ϵ.

1.4 Final Comments

Models such as (1.1.1) fit into framework considered by Chernozhukov & Hong (2002), Donald & Paarsch (2002), and Hirano & Porter (2003), who

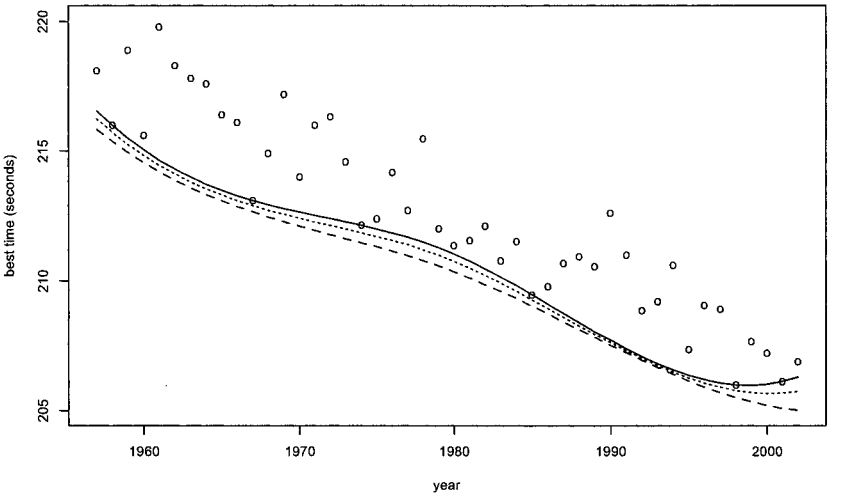


Figure 1.6: Estimated boundary lines for the 1500m data using $\rho(w) = w$ and $\tau(w) = w^{-2}$ for $\epsilon = 0.05$ (dotted) and $\epsilon = 0.5$ (dashed). The solid line is estimate given in Figure 1.1 and corresponds to the limit as $\epsilon \downarrow 0$.

consider asymptotic theory for estimation in models with parameter-dependent support. Unlike classical statistical models (where the support is independent of the parameters), maximum likelihood estimation does not have any particular asymptotic optimality. Both Chernozhukov & Hong (2002) and Hirano & Porter (2003) consider the asymptotics of Bayes estimators for a given loss function and prior distribution on the parameter space. Such estimators have the advantage of being admissible (with respect to loss function) and have asymptotic distributions that are independent of the prior distribution. Of course, these admissibility results are dependent on the model being correctly specified although one might expect Bayes estimators to be useful more generally.

It is also possible to extend the results to estimators $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}}_n)$ minimizing

$$\sum_{i=1}^{n} \rho(Y_i - \mathbf{x}_i^T \boldsymbol{\varphi}; \boldsymbol{\zeta}) \quad \text{subject to } Y_i \geq \mathbf{x}_i^T \boldsymbol{\varphi} \text{ for } i = 1, \cdots, n$$

where $\rho(w; \boldsymbol{\zeta})$ is a three times differentiable (or otherwise sufficiently smooth) function in $\boldsymbol{\zeta}$; the support of the response depends on $\boldsymbol{\beta}$ but not on the

"nuisance" parameter $\boldsymbol{\theta}$. We assume that for some matrices $A(\boldsymbol{\theta})$ and $B(\boldsymbol{\theta})$, we have

$$\begin{aligned} E\left[\nabla_{\boldsymbol{\zeta}}\rho(W_i;\boldsymbol{\theta})\right] &= 0, \\ E\left[\nabla_{\boldsymbol{\zeta}}\rho(W_i;\boldsymbol{\theta})\nabla_{\boldsymbol{\zeta}}^T\rho(W_i;\boldsymbol{\theta})\right] &= A(\boldsymbol{\theta}), \\ \text{and} \quad E\left[\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}}\rho(W_i;\boldsymbol{\theta})\right] &= B(\boldsymbol{\theta}) \end{aligned}$$

where $\nabla_{\boldsymbol{\zeta}}$ and $\nabla_{\boldsymbol{\zeta}\boldsymbol{\zeta}}$ are, respectively, the gradient and Hessian operators with respect to $\boldsymbol{\zeta}$. Then under additional regularity conditions (including, for example, appropriate modifications of (A0)–(A6)), we have the same limiting behaviour for $a_n C_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ as given in Theorem 1.1; moreover,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, B^{-1}(\boldsymbol{\theta})A(\boldsymbol{\theta})B^{-1}(\boldsymbol{\theta})\right)$$

with the two limiting distributions being independent.

We can also consider non-parametric estimation of boundaries by fitting parametric models (for example, polynomials) locally in the neighbourhood of a given point; the asymptotic behaviour of such non-parametric estimators can be determined using the theory discussed in Sections 1.2 and 1.3 with appropriate modifications. An alternative non-parametric approach to boundary estimation is given by Bouchard *et al.* (2003). This approach defines the boundary as a linear combination of kernel functions with non-negative weights estimated as the solution of a linear programming problem. In the context of production frontier estimation, a good survey of non-parametric estimation methods can be found in Florens & Simar (2002).

Bibliography

Aigner, D.J. & Chu, S.F. (1968) On estimating the industry production function. *American Economic Review* **58**, 826-839.

An, H.Z. & Huang, F.C. (1993) Estimation for regressive and autoregressive models with nonnegative residual errors. *Journal of Time Series Analysis* **14**, 179-191.

Anděl, J. (1989) Nonnegative autoregressive processes. *Journal of Time Series Analysis* **10**, 1-11.

Bouchard, G., Girard, S., Iouditski, A. & Nazin, A. (2003) Linear programming problems for frontier estimation. *Rapport de Recherche INRIA RR–4717.*

Chernozhukov, V. (2000) Conditional extremes and near extremes: estimation, inference and economic applications. Ph.D. thesis, Department of Economics, Stanford University.

Chernozhukov, V. & Hong, H. (2002) Likelihood inference in a class of nonregular econometric models. *MIT Department of Economics Working Paper 02-05.*

Donald, S.G. & Paarsch, H.J. (2002) Superconsistent estimation and inference in structural econometric models using extreme order statistics. *Journal of Econometrics* **109**, 305-340.

Feigen, P.D. & Resnick, S.I. (1994) Limit distributions for linear programming time series estimators. *Stochastic Processes and their Applications* **51**, 135-166.

Fiacco, A.V. & McCormick, G.P. (1990) *Nonlinear Programming: Sequential Unconstrained Minimization Techniques* SIAM, Philadelphia.

Florens, J.-P. & Simar, L. (2002) Parametric approximations of nonparametric frontiers. *Institut de Statistique, Université Catholique de Louvain, DP0222.*

Geyer, C.J. (1994) On the asymptotics of constrained M-estimation. *Annals of Statistics* **22**, 1993-2010.

Geyer, C.J. (1996) On the asymptotics of convex stochastic optimization. (unpublished manuscript)

Hirano, K. & Porter, J. (2003) Asymptotic efficiency in parametric structural models with parameter-dependent support. *Econometrica* **71**, forthcoming.

Kallenberg, O. (1983) *Random Measures.* (third edition) Akademie-Verlag, Berlin.

Knight, K. (1999) Epi-convergence in distribution and stochastic equi-semicontinuity. (unpublished manuscript)

Knight, K. (2001) Limiting distributions of linear programming estimators. *Extremes* **4**, 87-104.

Koenker, R. & Bassett, G. (1978) Regression quantiles. *Econometrica* **46**, 33-50.

Leadbetter, M.R., Lindgren, G. & Rootzén, H. (1983) *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

Nielsen, B. & Shephard, N. (2003) Likelihood analysis of a first order autoregressive model with exponential innovations. *Journal of Time Series Analysis* **24**, 337-344.

Pflug, G.Ch. (1994) On an argmax-distribution connected to the Poisson process, in *Asymptotic Statistics* (P. Mandl & M. Hušková, eds) 123-130, Physica-Verlag, Heidelberg.

Pflug, G. Ch. (1995) Asymptotic stochastic programs. *Mathematics of Operations Research* **20**, 769-789.

Portnoy, S. & Jurečková, J. (1999) On extreme regression quantiles. *Extremes* **2**, 227-243.

Smith, R.L. (1994) Nonregular regression. *Biometrika* **81**, 173-183.

2 A Simple Deconvolving Kernel Density Estimator when Noise Is Gaussian[1]

Isabel Proença[2]

Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa, Portugal

Summary

Deconvolving kernel estimators when noise is Gaussian entail heavy calculations. In order to obtain the density estimates numerical evaluation of a specific integral is needed. This work proposes an approximation to the deconvolving kernel which simplifies considerably calculations by avoiding the typical numerical integration. Simulations included indicate that the lost in performance relatively to the true deconvolving kernel, is almost negligible in finite samples.

Keywords: Deconvolution, density estimation, errors-in-variables, kernel, simulations.

2.1 Introduction

The estimation of a density by deconvolution consists in the estimation of the density of a random variable that is observed with an added unknown random noise. A typical example is the estimation of a density of a variable observed with measurement error. Another example is the estimation of the mixing distribution in a duration model. In what concerns applications, Fan and Truong (1993) introduce deconvolution techniques in the context of nonparametric regression with errors in variables. Calvet and Comon (2000) perform the deconvolution estimation of the joint density of spendig and

[1]Thanks are due to Hidehiko Ichimura and João Santos Silva for valuable comments. The usual disclaimer applies. This work was partially done while the author was visiting the Economics Department of University College, London, due to the support of *Centre for Microdata Methods and Practice (CEMMAP)* which is gratefully appreciated. Financial support from Fundação para a Ciência e Tecnologia/MCT under FCT/POCTI and BFAB-212/2000 is also acknowledge.

[2]Address for correspondence: R. do Quelhas 2, 1200-781 Lisboa, Portugal.E-mail: isabelp@iseg.utl.pt. Fax: 351 213922781.

tastes in presence of measurement error, and Horowitz and Markatou (1996) analyze earnings mobility using nonparametric deconvolution estimation of a density in the context of a random effects model for panel data.

To describe the problem, suppose a random variable Y such that $Y = X + U$, where X and U are independent random variables. Suppose more that Y is observable while X (the target) and U (the noise) are non-observable, and the aim here is to estimate the density of X, $f(x)$, also called the *target* density when $g(y)$, the density of Y is completely unknown. Usually the *noise* density, $q(u)$, is assumed to belong to a given family, most frequently the Normal, with zero mean. Observe that $g(y)$ is equal to the convolution of the densities of X and U, verifying,

$$g(y) = \int_{-\infty}^{\infty} f(y-u)q(u)du. \tag{2.1.1}$$

Then the density $f(x)$ may be obtained by deconvolution. The usual procedure is to obtain first the characteristic function of X, $\dot{f}$, using $\dot{g}$ and $\dot{q}$ (the characteristic functions of Y and U respectively) and then inverse Fourier transform leads to $f(x)$ according to

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \frac{\dot{g}(t)}{\dot{q}(t)} dt \tag{2.1.2}$$

An estimator of $f(x)$ can be obtained by substituting the unknown quantities in (2.1.2) by consistent estimators. However, in practice the corresponding calculations can entail problems leading to high fluctuations in the aimed estimate. To avoid this a suitable damping factor is incorporated in the corresponding integral leading to the following deconvolving kernel density estimator introduced by Stefanski and Carroll (1990),

$$\hat{f}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \dot{k}(th) \frac{\widehat{\dot{g}}(t)}{\dot{q}(t)} dt \tag{2.1.3}$$

with $\dot{k}(t)$ the Fourier transform of a kernel function $K(x)$ (such that $\dot{k}(0) = 1$), h the bandwidth which tends to 0 (so that the damping factor tends to 1), and $\widehat{\dot{g}}(t)$ the empirical characteristic function of Y. Fan (1991) obtains convergence rates of this estimator for several *noise* distributions.

Stefanski and Carrol (1990) show that in case the function $\dot{k}(t)/\dot{q}(t/h)$ is integrable then expression (2.1.3) can be rewritten as,

$$\hat{f}(x) = \frac{1}{nh} \sum_{1}^{n} K_h^* \left(\frac{x - Y_j}{h} \right) \tag{2.1.4}$$

where

$$K_h^*(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} \frac{\dot{k}(t)}{\dot{q}(t/h)} dt \tag{2.1.5}$$

Observe that by equation (2.1.4) the deconvolving kernel estimate is just an ordinary kernel estimate but with specific kernel function equal to (2.1.5) where the shape of this kernel function depends on the bandwidth. For certain types of distributions for the noise variable, (2.1.5) has a closed-form expression and calculations are as hard as the ordinary kernel estimation. Unfortunately this is not the case when the noise is normally distributed as it is often assumed namely in econometric applications. When $q(u)$ belongs to the normal family the integral (2.1.5) has to be evaluated and calculations are much harder. Moreover, the damping kernel K has to be carefully chosen, to guarantee that the integral exists. Consequently, deconvolving kernel density estimation for Gaussian noise suffers from the drawback of being subject to Monte Carlo error and computationally very burdensome.

In this paper, an approximation of (2.1.5) is proposed to estimate $f(x)$ by kernel deconvolution when noise is Gaussian. It avoids the typical numerical integration making calculations incredibly easier, being as difficult as an ordinary kernel density estimator. The next section introduces the simple deconvolving Kernel. Section 3 presents a simulation study that analyzes the performance of the new estimator compared to the exact one for several sample sizes and target distributions. Section 4 concludes.

2.2 The Simple Deconvolving Kernel Density Estimator

The main idea behind the simple deconvolving kernel estimator is to substitute in (2.1.5) the inverse of the true characteristic function of the normal density (which is an exponential function) by an approximation given by the first-order term of the respective Taylor series expansion around $\sigma_u^2 = 0$. Therefore, considering as damping kernel the standard normal the approximate deconvolving kernel is equal to,

$$K_h^{a*}(z) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itz} e^{-0.5t^2} \left(1 + \frac{t^2 \sigma_u^2}{2h^2}\right) dt \tag{2.2.1}$$

with σ_u^2 the variance of the noise variable.

Elementary calculations simplify (2.2.1) to the following expression,

$$K_h^{a*}(z) = \varphi(z) - \frac{\sigma_u^2}{2h^2} \varphi''(z) \tag{2.2.2}$$

where $\varphi(z)$ is the standard normal density function and $\varphi''(z)$ is its second

derivative. Finally, the density estimate is obtained with,

$$\hat{f}^a(x) = \frac{1}{nh} \sum_{1}^{n} K_h^{a*} \left(\frac{x - Y_j}{h} \right) \tag{2.2.3}$$

where $K_h^{a*}(\bullet)$ is given in (2.2.2).

Using the same arguments as Stefanski and Carroll (1990) is easy to show that $K_h^{a*}(\bullet)$ is symmetric and $\int \hat{f}^a(x)dx = 1$.

2.3 Simulation Results

In this section the performance of the approximated deconvolving kernel in finite samples is examined by a detailed simulation study. The main goal is to evaluate the deterioration in the accuracy of the deconvolving kernel density estimates due to the use of the much simpler approximated deconvolving kernel introduced in this paper instead of the exact one of Stefanski and Carroll (1990). With this aim the average integrated squared error (AISE) calculated for the optimal bandwidth (optimal in the sense that minimizes the AISE) is compared for the two procedures. The optimal bandwidth was found by grid search (with increment equal to 0.02). Depending on the particular design, the grid intervals were chosen wide enough in order to assure that they contain within the optimal bandwidth. It is also an aim to analyze whether the performance depends on the shape of the target density or on the importance of the disturbing noise measured by the reliability ratio equal to,

$$r = \frac{Var(X)}{Var(Y)} = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2}. \tag{2.3.1}$$

For the last issue, two situations were considered. One, with a relatively small noise with corresponding reliability ratio equal to 0.89 and another where r is 0.62 which expresses a severe perturbation of X (these particular quantities where chosen in order that σ_U^2 and σ_X^2 have suitable values). Observe that in this last situation the exact deconvolving kernel density estimate looses accuracy (in the sense that the corresponding AISE has tendency to be considerably bigger).

For $r = 0.89$ three different designs were chosen in order to illustrate three of the most important types of shapes of the densities of the target variable that are more frequently found in practice resulting respectively in a symmetric, a skewed, and a bimodal densities. The designs are,

DESIGN 1 - $X \sim N(0, 16)$, $U \sim N(0, 2)$ and $Y = X + U$.

DESIGN 2 - $X \sim \chi^2(8)$, $U \sim N(0,2)$ and $Y = X + U$.

DESIGN 3 - X is a mixture of a $N(\sqrt{44}, 20)$ with a $N(-\sqrt{44}, 20)$, being $U \sim N(0,8)$ and $Y = X + U$.

For each design 1000 replications were calculated for samples with size respectively of 100, 250 and 500 observations. The results can be seen in table 2.1. It is clear that the simpler approximated deconvolving kernel has a good performance given that the deterioration in the AISE is almost negligible even for small samples and for all designs tried. The worst case refers to an increase in the AISE of 7% for sample size of 100 obtained with Design 1. There are not remarkable differences among all the different density shapes tried. The fact that the rates are slightly less favorable for Design 1 may be due to the general better performance (in AISE) of the exact Kernel for this type of *target* densities (relatively to the skewed and bimodal) as is analyzed in Wand 1998.

Av. Best AISE $\times 10^6$, $r = 0.89$			
	Exact	Approx.	App/Exa
Normal density			
n = 100	68.25	73.07	1.0706
n = 250	39.64	42.19	1.0643
n = 500	26.71	28.14	1.0535
Chi-square density			
n = 100	99.18	104.44	1.0530
n = 250	62.01	64.99	1.0481
n = 500	44.40	45.89	1.0336
Bimodal density			
n = 100	32.63	33.92	1.0395
n = 250	20.86	21.93	1.0513
n = 500	15.88	16.65	1.0485

Table 2.1: Average Best AISE in 1000 simulated samples of 100, 250 and 500 observations.

The good performance of the approximated deconvolving kernel can be seen also in figures 2.1 to 2.3. These figures represent for each design the true density together with the exact and approximated deconvolving kernel density estimates (calculated each for the respective optimal bandwidth) for one sample randomly selected with 500 observations. Both estimates are remarkable close in all the graphics.

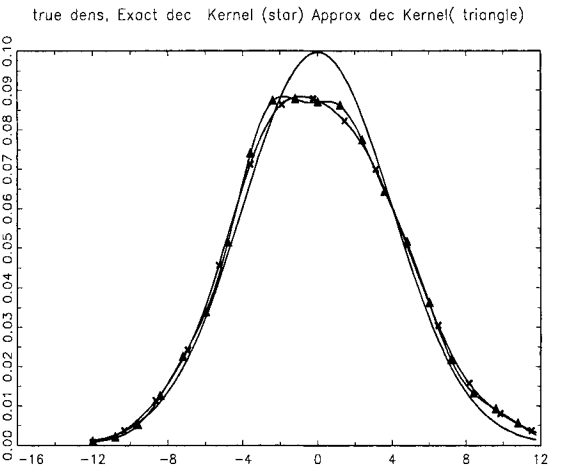


Figure 2.1: True density, exact deconvolving kernel (star) approximated deconvolving kernel (triangle). One random sample for design 1, n=500.

Given the heavy computations required by the exact density kernel estimator, and the fact that when $r = 0.89$ the performance of the approximated procedure for the different designs was very similar, only the symmetric shape for target variable was analyzed when $r = 0.62$, leading to

DESIGN 4 - $X \sim N(0, 16)$, $U \sim N(0, 10)$ and $Y = X + U$.

The results are included in table 2.2. The performance of the simple deconvolving kernel is better in approximating the exact deconvolving kernel density estimate when the variance of the error is bigger. So that, it seems that the deterioration in quality of the exact deconvolving density estimate caused by the increase of the importance of the disturbing noise is less significant in the simple approximated estimator.

Table 2.3 shows the ratio in the average computation time for each h between the exact deconvolving kernel density estimate and the simple approximated one. The gain in computation time of the simple deconvolving kernel over the exact can be impressively large. For instance, with data from Design 2 the exact calculations take more 296 times the time spent with the approximated,

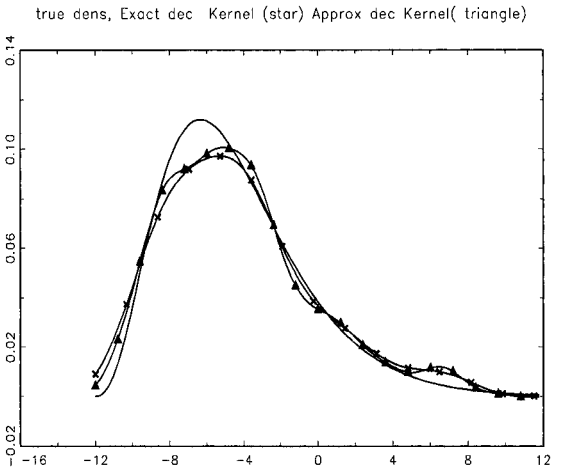


Figure 2.2: True density, exact deconvolving kernel (star) approximated deconvolving kernel (triangle). One random sample for design 2, n=500.

Av. Best AISE $\times 10^6$ for Normal density			
	Exact	Approx.	App/Exa
	$r = 0.62$		
n = 100	150.16	157.75	1.0505
n = 250	108.51	110.79	1.0210
n = 500	85.70	85.06	0.9925

Table 2.2: Average Best AISE for normal *target* density in 1000 simulated samples of 100, 250 and 500 observations.

while for Design 4 the ratio is 28 which is already considerable, specially if one needs to replicate calculations for several bandwidths.

2.4 Concluding Remarks

This work introduces a simple deconvolving kernel to estimate a density by deconvolution when the noise variable is Normally distributed. It avoids the

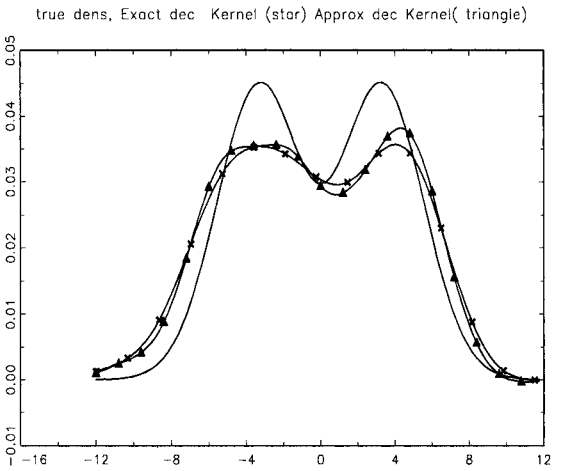


Figure 2.3: True density, exact deconvolving kernel (star) approximated deconvolving kernel (triangle). One random sample for design 3,n=500.

Ratio of calculation times Exact over Approximated	
Design 1	150.73
Design 2	295.68
Design 3	128.55
Design 4	27.78

Table 2.3: Ratio of the average calculation time for each h in one sample randomly selected.

typical numerical integration necessary to obtain the ordinary deconvolving kernel density estimate in this situation, making calculations noticeably faster and less subject to numerical error. It has a simple and direct application being as difficult as an ordinary kernel density estimator.

A simulation study shows that the lost in performance of this simpler estimator in finite samples is reasonable low. Moreover, in situations where the accuracy of the exact ordinary deconvolving kernel tends to deteriorate (because of a more complex shape of the true density or a low reliability ratio)

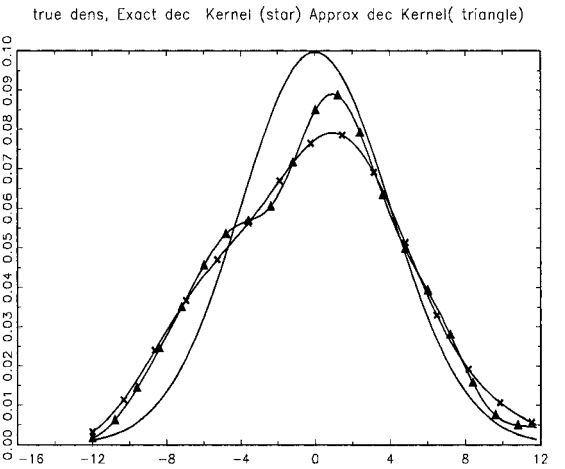


Figure 2.4: True density, Exact deconvolving kernel (plus) Approximated deconvolving kernel (triangle). One random sample for Design 4,n=500.

the simpler deconvolving kernel has a relatively better performance. Therefore, the use of this procedure seems to be beneficial when calculations of the deconvolving kernel density estimate have to be replicated several times, or numerical integration is a problem. On the other hand, it could be even relatively more beneficial in situations where the exact kernel is less accurate because of the shape of the density or an important noise with a low reliability ratio.

Bibliography

Calvet, L. E. & Comon, E. (2000) Behavioral Heterogeneity and the Income Effect, *Harvard Institute of Economic Research Working Papers* No 1892.

Diggle, P.J. & Hall, P. (1993) A Fourier Approach t Nonparametric Deconvolution of a Density Estimate, *J.R. Statist. Soc.* B **55**, 2, 523–531.

Fan, J. (1991) On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems, *The Annals of Statistics* **19**, 1257–1272.

Fan, J. & Truong, Y. K. (1993) Nonparametric Regression with Errors in Variables, *The Annals of Statistics* **21**, 1900–1925.

Horowitz, J.L. & Markatou, M. (1996) Semiparametric Estimation of Regression Models for Panel Data, *Review of Economic Studies* **63**, 145–168.

Stefanski, L. & Carrol, R.J. (1990) Deconvoluting Kernel Density Estimators, *Statistics* **21**, 169–184.

Wand, M.P. (1998) Finite Sample Performance of Deconvolving Density Estimators, *Statistics and Probability Letters* **37**, 131–139.

3 Nonparametric Volatility Estimation on the Real Line from Low Frequency Data

Markus Reiß

Institute of Mathematics, Humboldt University Berlin,
Unter den Linden 6, D-10099 Berlin, Germany

Summary

We estimate the volatility function of a diffusion process on the real line on the basis of low frequency observations. The estimator is based on spectral properties of the estimated Markov transition operator of the embedded Markov chain. Asymptotic risk estimates for a growing number of observations are provided without assuming the observation distance to become small.

Keywords: Scalar diffusion, discrete observations, warped wavelets, spectral approximation

3.1 Introduction

Diffusion processes are widely used in physical, chemical or economical applications to model random fluctuations of some quantity over time. Especially in mathematical finance it has become very popular to model asset prices by diffusion processes because this allows the use of strong tools from stochastic analysis for option pricing or risk analysis. Removing seasonal effects and long-term growth results in time-homogeneous diffusion processes. A typical time-homogeneous scalar diffusion $(X_t,\, t \geq 0)$ solves the Itô stochastic differential equation

$$dX_t = b(X_t)\, dt + \sigma(X_t)\, dW_t, \quad t \geq 0, \tag{3.1.1}$$

with drift coefficient $b(\bullet)$, volatility or diffusion coefficient $\sigma(\bullet)$ and with a one-dimensional Brownian motion $(W_t,\, t \geq 0)$.

Statistical inference for the volatility function has attracted a lot of interest recently, see the discussions in (Kleinow 2002) or (Gobet, Hoffmann,

& Reiß 2002) for an overview. Especially, in (Kleinow 2002) it is argued on the basis of empirical data that common parametric assumptions on the coefficients are highly misspecified in models for financial markets. Moreover, statistical methods developed for high frequency observations, that is small observation distances, have been typically applied to daily asset price data over periods of several years, which should be qualified rather as low-frequency observations. Therefore, the work (Kessler & Sørensen 1999) on low-frequency statistical methods became a popular alternative, but remains restricted to certain parametric models.

Here, we consider the case of nonparametric inference for the volatility function $\sigma(\bullet)$ in the case of an unknown drift function $b(\bullet)$ and equidistant observations $(X_{n\Delta})_{0\leq n\leq N}$ with some fixed $\Delta > 0$. In (Gobet, Hoffmann, & Reiß 2002) it was shown that for diffusions with reflections on a compact interval the nonparametric estimation problems can be solved using ideas in (Hansen, Scheinkman, & Touzi 1998), but it involves some ill-posedness such that the minimax rate of convergence is $N^{-s/(2s+3)}$ for $N \to \infty$ and regularity $s \geq 1$ of $\sigma(\bullet)$. Moreover, first numerical simulations in the reflected setting have shown that the spectral estimator outperforms the traditional quadratic variation estimator already for rather small observation distances Δ. We generalize this approach to cope also with diffusions on the entire real line.

The basic idea is that we can only draw inference on the law of the embedded Markov chain $(X_{n\Delta})_{n\geq 0}$, that by spectral calculus its transition operator determines the infinitesimal generator of the diffusion process and that this generator encodes rather explicitly the two coefficients $b(\bullet)$ and $\sigma(\bullet)$. More specifically, the spectral estimator we propose is based on estimates of the invariant density and of one eigenfunction and its eigenvalue of the transition operator of $(X_{n\Delta})_{n\geq 0}$, see formula (3.2.3) below. Leaving the case of a compact state space, we face several new problems compared with the situation treated in (Gobet, Hoffmann, & Reiß 2002): (1) the observation design is degenerate, (2) the invariant densities are not uniformly comparable and (3) the eigenfunctions are unbounded. Point (1) is overcome by using warped wavelet functions or equivalently a suitable state transformation. To avoid problem (2) we work on parameter-dependent function spaces and problem (3) is treated by smoothing differently at the boundaries. By this approach we obtain that our spectral estimator also attains the rate $N^{-s/(2s+3)}$ as in the simpler case of reflected diffusions, provided the coefficients guarantee that the process is well mixing and the first eigenfunction exists and does not grow too fast to infinity.

For the proof we assume the invariant law of the diffusion to be known. This is, of course, not realistic, but the estimation of the invariant density is standard and contributes less to the overall risk than the spectral estimations,

as can also be seen from the lower bound proof in (Gobet, Hoffmann, & Reiß 2002).

Section 3.2 introduces the diffusion model and recalls some theory for diffusions, Section 3.3 presents and discusses the estimator and Section 3.4 provides the mathematical results. We adopted (hopefully) standard notation. In particular, $C^r(\mathbb{R})$ denotes the space of r-times continuously differentiable functions and $C_b^r(\mathbb{R})$ its subspace such that all derivatives are uniformly bounded including the function itself. The relation $A \lesssim B$ means that A is bounded by a multiple of B, independent of the quantities appearing in the expression B. The relation $A \sim B$ stands for $A \lesssim B$ and $B \lesssim A$. A sequence of random variables that is bounded in probability will be abbreviated by $O_P(1)$. Vectors and matrices are usually set in bold fonts.

3.2 The Diffusion Model

In this section fundamental results for one-dimensional diffusions are recalled, for more details and proofs see e.g. (Karlin & Taylor 1981) or (Bass 1998). We consider diffusion processes $(X_t,\, t \geq 0)$ solving (3.1.1). The drift $b(\bullet)$ and diffusion coefficient or volatility $\sigma(\bullet)$ are assumed to be Lipschitz continuous functions such that a strong solution exists. We shall henceforth assume the uniform ellipticity condition

$$\exists\, \sigma_0,\, \sigma_1 > 0 :\ \sigma_0 \leq \sigma(x) \leq \sigma_1 \text{ for all } x \in \mathbb{R} \tag{3.2.1}$$

and the mixing condition

$$\lim_{x \to +\infty} b(x) = -\infty \text{ and } \lim_{x \to -\infty} b(x) = +\infty. \tag{3.2.2}$$

These conditions imply the existence of a stationary solution X with invariant marginal density

$$\mu(x) = \frac{2C}{\sigma^2(x)} \exp\Big(\int_0^x \frac{2b(x)}{\sigma^2(x)}\, dx \Big), \quad x \in \mathbb{R},$$

where $C > 0$ is a suitable norming constant. Moreover, the solution process is time-reversible and β-mixing with exponential speed such that for statistical purposes the hypothesis of stationary observations is reasonable and will be assumed henceforth.

Diffusions are efficiently described by their Markov transition operators $(P_t)_{t \geq 0}$ with

$$P_t f(x) = \mathbb{E}[f(X_t) \,|\, X_0 = x] = \int_{-\infty}^{\infty} f(\xi) p_t(x, \xi)\, d\xi, \quad x \in \mathbb{R},\ f \in C_b(\mathbb{R}),$$

where $p_t(x,\xi)$ denotes the transition probability density. The operators $(P_t)_{t\geq 0}$ can be extended to the Hilbert space

$$L^2(\mu) = \Big\{ f : \mathbb{R} \to \mathbb{R} \,|\, \int f^2(x)\mu(x)\,dx < \infty \Big\},$$

on which they form a strongly continuous, self-adjoint semigroup of contraction operators with infinitesimal generator

$$Lf(x) = \tfrac{1}{2}\sigma^2(x)f''(x) + b(x)f'(x), \quad x \in \mathbb{R},$$

for functions f in the domain (with natural boundary conditions)

$$\mathcal{D}(L) = \{ f \in L^2(\mu) \,|\, Lf \in L^2(\mu) \}.$$

L is a closed selfadjoint operator with spectrum on the negative real axis and the spectral mapping theorem asserts $P_t = \exp(tL)$. In particular, the eigenfunctions of P_t and L coincide and the eigenvalues are transformed like the operators. The Markov semigroup can be described equivalently by the invariant density $\mu(\bullet)$ and the inverse scale density $S^{-1}(\bullet)$ given by

$$S(x) = \tfrac{1}{2}\sigma^2(x)\mu(x), \quad x \in \mathbb{R}.$$

Then the infinitesimal generator can be written in divergence form

$$Lf(x) = \mu^{-1}(x)(Sf')'(x), \quad x \in \mathbb{R}.$$

Any eigenfunction $u \in L^2(\mu)$ of L with eigenvalue ν satisfies

$$S(x)u'(x) = \nu \int_{-\infty}^{x} u(\xi)\mu(\xi)\,d\xi, \quad x \in \mathbb{R},$$

which yields

$$\sigma^2(x) = \frac{2\nu \int_{-\infty}^{x} u(\xi)\mu(\xi)\,d\xi}{u'(x)\mu(x)}, \quad x \in \mathbb{R}. \tag{3.2.3}$$

This identity allows to determine the volatility $\sigma(\bullet)$ from quantities accessible from the embedded Markov chain $(X_{n\Delta})_{n\geq 0}$, namely from the invariant density and a spectral pair $(u, e^{\nu\Delta})$ of the transition operator. This approach was first proposed by (Hansen, Scheinkman, & Touzi 1998) and statistically analyzed in (Gobet, Hoffmann, & Reiß 2002).

For this method to work we have to ensure that at least parts of the spectrum are discrete, that is proper eigenvalues exist. In the sequel we shall only need that the largest nontrivial (i.e., nonzero) spectral value is discrete, but to avoid any technicalities we assume $\sigma \in C^1(\mathbb{R})$ and

$$\lim_{|x|\to\infty} \Big(\sigma'(x) - \frac{2b(x)}{\sigma(x)} \Big)^2 = \infty, \tag{3.2.4}$$

which by Section 4.2 in (Hansen, Scheinkman, & Touzi 1998) ensures that the entire spectrum of L is discrete. In view of our previous assumptions this is already satisfied if $\sigma'(\bullet)$ is uniformly bounded.

The mathematical analysis of our proposed estimators relies on some additional growth restrictions for the first nontrivial eigenfunction u_1 of L, namely

$$u_1 \in L^p(\mu) \text{ and } u_1' \in L^p(\mu) \tag{3.2.5}$$

for some arbitrary $p > 2$. For $p = 2$ this condition is always satisfied because u_1 is in the domain of L and thus also of $(-L)^{1/2}$:

$$\|(-L)^{1/2}u_1\|_\mu^2 = \langle(-L)u_1, u_1\rangle_\mu = \langle Su_1', u_1'\rangle \sim \langle u_1', u_1'\rangle_\mu.$$

Observe the different norms and scalar products employed, where the index μ always refers to $L^2(\mu)$ and no index to L^2 with respect to the Lebesgue measure. For the canonical example of a stationary Ornstein-Uhlenbeck process all eigenfunctions satisfy condition (3.2.5) even for exponential moments. It is plausible that this behaviour remains the same whenever the tails of the invariant densities are equally small which is to say that the negative drift $-b(\bullet)$ grows linearly. A formal mathematical result in this direction still lacks and we can merely provide an example of a sufficient result under nonasymptotic conditions on the coefficients.

PROPOSITION 3.1 *Condition* (3.2.5) *is satisfied for all* $p < \infty$ *if the coefficients* $\sigma^2 \in C^2(\mathbb{R})$, $b \in C^1(\mathbb{R})$ *of the diffusion satisfy*

$$\inf_{x\in\mathbb{R}}\Big(\frac{(\sigma^2(x))'}{\sigma^2(x)}\Big(b - \frac{(\sigma^2(x))'}{2\sigma^2(x)}\Big) + \frac{1}{2}(\sigma^2(x))'' - 2b'(x)\Big) > 0.$$

For constant volatility this reduces to $\sup_x b'(x) < 0$.

PROOF:
By definition 1.2 in (Ledoux 1998) the diffusion process X satisfies the condition $CD(R,\infty)$ for some $R > 0$ under our assumption, which implies that the Markov semigroup is hypercontractive. It is proved in (Bakry 1994) that any eigenfunction u of a hypercontractive semigroup operator has exponential moments, that is satisfies $\int \exp(cu^\alpha(x))\mu(x)\,dx < \infty$ for some $c, \alpha > 0$.

By Lemma 1.3 in (Ledoux 1998) the condition $CD(R,\infty)$ is equivalent to

$$|\sigma(x)(P_tf)'(x)| \le e^{-Rt}P_t(|\sigma(x)f'(x)|), \quad x \in \mathbb{R},$$

for all sufficiently smooth f. For any eigenfunction u of L with eigenvalue ν we thus obtain

$$|\sigma(x)u'(x)|e^{t\nu} \le e^{-Rt}P_t(|\sigma(x)u'(x)|), \quad x \in \mathbb{R}.$$

The hypercontractivity of (P_t) and $\sigma u' \in L^2(\mu)$ therefore imply $\sigma u' \in L^p(\mu)$ for all $p < \infty$ and by ellipticity also $u' \in L^p(\mu)$. $\square$

3.3 Construction of the Estimators

We describe the spectral estimation procedure using the projection method in detail. The use of projection methods has the advantage of approximating the abstract operators by finite-dimensional matrices, for which the spectrum is easy to calculate numerically. In addition, mathematical results for spectral approximation by kernel-smoothed operators seem to be difficult to obtain. A projection approach was already suggested by (Chen, Hansen, & Scheinkman 1997) and adopted by (Gobet, Hoffmann, & Reiß 2002). More specifically, we make use of compactly supported wavelets on the interval $[0,1]$. For the notion of wavelet bases on compact intervals and their properties we refer to (Cohen 2000).

DEFINITION 3.1 *Let (ψ_λ) with multi-indices $\lambda = (j,k)$ be a compactly supported orthonormal wavelet basis of $L^2(0,1)$ including the scaling function $\psi_{-1,0} = \mathbf{1}$. For $\lambda = (j,k)$ we set $|\lambda| := j$. The approximation spaces (V_Λ) are defined as the linear span of the wavelets indexed with Λ:*

$$V_\Lambda := span\{\psi_\lambda \,|\, \lambda \in \Lambda\}.$$

The L^2-orthogonal projection onto V_Λ will be called Π_Λ. For a function $M : \mathbb{R} \to [0,1]$ we introduce the warped wavelets $\psi_\lambda^M(x) := \psi_\lambda(M(x))$, $x \in \mathbb{R}$.

Note that (ψ_λ^M) constitutes an orthonormal basis of $L^2(\mu)$ with $\mu(x) = M'(x)$, if M is (weakly) differentiable.

The first main idea is to use wavelets warped by the empirical stationary distribution function of the diffusion process X in order to obtain a regular autoregressive design, see (Kerkyacharian & Picard 2003) for a similar approach in classical regression with random design, but note that our density does not define a Muckenhoupt weight. An equivalent viewpoint is that we consider the data

$$\hat{Y}_n := \hat{M}(X_{n\Delta}), \qquad \text{where } \hat{M}(x) := \frac{1}{N+1} \sum_{n=0}^{N} \mathbf{1}_{(-\infty, X_{n\Delta}]}(x)$$

is the empirical stationary distribution function. The transformed observations $(\hat{Y}_n)_{0 \le n \le N}$ form a permutation of the set $\{n/(N+1) \,|\, 1 \le n \le N+1\}$. For such equispaced data Mallat's pyramidal algorithm for computing wavelet

coefficients is very efficient and widely available. Since the marginal density does not determine the diffusion process if drift and volatility are both unknown, we use the dependency structure in the data $(\hat{Y}_n)$ in order to draw further inference. By the Markov property of diffusion processes, it suffices to consider the empirical distribution of the transitions $X_{(n-1)\Delta} \mapsto X_{n\Delta}$ or $\hat{Y}_{n-1} \mapsto \hat{Y}_n$, respectively. Furthermore, the time reversibility asserts that the laws of $(X_{(n-1)\Delta}, X_{n\Delta})$ and $(X_{n\Delta}, X_{(n-1)\Delta})$ coincide such that we may symmetrize our estimators.

Under the stationarity assumption, $\hat{M}$ converges for $N \to \infty$ uniformly to the true distribution function $M(x) := \int_{-\infty}^{x} \mu(\xi)\, d\xi$, $x \in \mathbb{R}$. We are thus naturally lead to consider the diffusion process $Y_t = M(X_t)$, $t \geq 0$, with values in the open unit interval $(0,1)$, which has natural boundaries and satisfies by Itô's formula

$$dY_t = \mu_M(Y_t)\Big(b_M(Y_t) + \frac{1}{2}(\mu_M)'(Y_t)\sigma_M^2(Y_t)\Big)\, dt + \mu_M(Y_t)\sigma(Y_t)\, dW(t).$$

The process Y is equivalently described by the following quantities, where we write $f_M(y) := f(M^{-1}(y))$ for any function $f : \mathbb{R} \to \mathbb{R}$:

invariant measure:	$\mu_Y(y) = \mathbf{1}_{(0,1)}(y)$, (uniform),
scale density:	$S_Y^{-1}(y) = 2\mu_M^{-2}(y)\sigma_M^{-2}(y)$,
transition density:	$p_{t,Y}(y,\eta) = p_t(M^{-1}(y), M^{-1}(\eta))\mu_M(\eta)^{-1}$,
inf. generator:	$L_Y f(y) = \big(\frac{1}{2}\sigma_M^2\mu_M^2 f'\big)'(y)$,
domain of L_Y:	$\mathcal{D}(L_Y) = \{f \in L^2(0,1) \mid Lf \in L^2(0,1)\}$.

Note that quantities without index usually refer to X, whereas those related to Y carry an index. From the formula for the transition operator $(P_{t,Y} f_M)(M(x)) = P_t f(x)$ it follows that any eigenvalue ν_Y of L_Y with eigenfunction u_Y is also an eigenvalue of L, but with the rescaled eigenfunction $u = u_Y \circ M$ and vice versa.

We have thus separated the estimation problem for the volatility function $\sigma(\bullet)$ of the original process X into the two subproblems of estimating the invariant density $\mu(\bullet)$ of X and of drawing inference on the Markov transitions of the transformed diffusion process Y. Of course, the latter is the much more demanding task, because the invariant density can be estimated classically under a suitable mixing hypothesis on X.

EXAMPLE 3.1 *The stationary Ornstein-Uhlenbeck process with parameters α, $\sigma > 0$ satisfies the stochastic differential equation*

$$dX_t = -\alpha X_t\, dt + \sigma\, dW_t.$$

It is a Gaussian process with normal stationary law $N(0, \frac{\sigma^2}{2\alpha})$. *Its generator* L *has discrete spectrum* $\Sigma(L) = \{-\alpha n \,|\, n \geq 0\}$ *and the eigenfunctions are given by Hermite-type polynomials.*

The transformed process Y *satisfies the stochastic differential equation*

$$dY_t = -2\alpha\mu_M(Y_t)M^{-1}(Y_t)\,dt + \sigma\mu_M(Y_t)\,dW_t,$$

where by normality $\mu_M(y)$ *is up to logarithmic terms of order* y *for* y *near zero and of order* $(1-y)$ *for* y *close to one. The eigenfunctions of* L_Y *are polynomials in* $M^{-1}(y)$ *such that they have logarithmic singularities and their derivatives of order* r *have polynomial singularities of order* r *at the boundary.*

Recall that by formula (3.2.3) we can estimate the volatility function $\sigma(\bullet)$ by a plug-in from estimates of the invariant density $\mu(\bullet)$ and the inverse scale density $S(\bullet)$ of the process X. Hence, we make use of the transformation of this formula

$$\sigma_M^2(y) = \frac{2\nu_1 \int_0^y u_{1,Y}(\eta)\,d\eta}{(u_{1,Y})'(y)\mu_M^2(y)}, \tag{3.3.1}$$

where $u_{1,Y}$ denotes the eigenfunction of L_Y corresponding to the largest non-trivial eigenvalue ν_1. By the spectral mapping theorem $(e^{\Delta\nu_1}, u_{1,Y})$ is the corresponding spectral pair of the transition operator $P_{\Delta,Y}$.

Consequently, we are interested in obtaining spectral information about the transition operator $P_{\Delta,Y}$ of Y. Its expansion in the wavelet basis (ψ_λ) of $L^2(0,1)$ can be estimated by the symmetrized empirical operator coefficients

$$(\hat{\mathbf{P}}_\Delta)_{\lambda,\lambda'} := \frac{1}{2N}\sum_{n=1}^{N}\Big(\psi_\lambda(\hat{Y}_{n-1})\psi_{\lambda'}(\hat{Y}_n) + \psi_\lambda(\hat{Y}_n)\psi_{\lambda'}(\hat{Y}_{n-1})\Big).$$

Note that this is equivalent to estimating the transition operator P_Δ of X in terms of the empirically warped wavelet basis $(\psi_\lambda^{\hat{M}})$:

$$(\hat{\mathbf{P}}_\Delta)_{\lambda,\lambda'} = \frac{1}{2N}\sum_{n=1}^{N}\Big(\psi_\lambda^{\hat{M}}(X_{(n-1)\Delta})\psi_{\lambda'}^{\hat{M}}(X_{n\Delta}) + \psi_\lambda^{\hat{M}}(X_{n\Delta})\psi_{\lambda'}^{\hat{M}}(X_{(n-1)\Delta})\Big).$$

If we had $\hat{M} = M$, this would give an unbiased estimate because of

$$\begin{aligned}\mathbb{E}[\psi_\lambda(M(X_{(n-1)\Delta}))\psi_{\lambda'}(M(X_{n\Delta}))] &= \int_0^1\int_0^1 \psi_\lambda(y)\psi_{\lambda'}(\eta)p_{\Delta,Y}(y,\eta)\,d\eta\,dy \\ &= \langle P_{\Delta,Y}\psi'_\lambda, \psi_\lambda\rangle.\end{aligned}$$

The eigenfunction $u_1 \in L^2(\mu)$ of P_Δ with eigenvalue κ_1 satisfies for any multi-index λ the coefficient equation

$$\sum_{\lambda'}\langle P_\Delta\psi_\lambda^M, \psi_{\lambda'}^M\rangle_\mu\langle u_1, \psi_{\lambda'}^M\rangle_\mu = \kappa_1\langle u_1, \psi_\lambda^M\rangle_\mu.$$

Furthermore, we have $u_{1,Y} = \sum_\lambda \langle u_1, \psi_\lambda^M \rangle_\mu \psi_\lambda$. We therefore calculate the largest nontrivial eigenvalue $\hat{\kappa}_1$ (i.e. $\hat{\kappa}_1 < 1$) with eigenvector $\hat{\mathbf{u}}_1$ of the symmetric $|\Lambda| \times |\Lambda|$-matrix $\hat{\mathbf{P}}_{\Delta,\Lambda} := (\hat{\mathbf{P}}_\Delta)_{\lambda,\lambda' \in \Lambda}$ and use the estimators

$$\hat{u}_{1,Y}(x) := \sum_{\lambda \in \Lambda} (\hat{\mathbf{u}}_1)_\lambda \psi_\lambda(x), \quad \hat{\nu}_1 := \Delta^{-1} \log(\hat{\kappa}_1).$$

Observe that by construction $\hat{\mathbf{P}}_{\Delta,\Lambda}$ always has eigenvector $\hat{\mathbf{u}}_0 = (1, 0, \ldots, 0)$ corresponding to the constant scaling function $\psi_{-1,0}$ with eigenvalue 1.

Even though formula (3.3.1) is valid for any nontrivial spectral pair of L_Y, we prefer taking the first nontrivial eigenfunction $u_{1,Y}$ for two reasons: first, all other eigenfunctions oscillate such that the denominator vanishes at some point and the estimate in its neighbourhood is worthless. Second, the spectral estimation quality depends very much on the separation of the eigenvalue from the remaining spectrum (cf. Proposition 3.5) and the spectrum $\Sigma(P_{\Delta,Y}) = \{e^{\Delta\nu} \,|\, \nu \in \Sigma(L_Y)\}$ is such that it becomes rapidly very dense for smaller eigenvalues. Nevertheless, it might be reasonable to use the information about the other spectral pairs, compare also the embeddability discussion in (Hansen, Scheinkman, & Touzi 1998).

The usage of warped basis functions simplifies the design and thus the analysis of the stochastic error term, but does not overcome the complex structure of the deterministic approximation error. As proved later, the eigenfunctions of L_Y have logarithmic singularities at the boundary of the unit interval and its derivatives have even polynomial-type singularities. This is why, theoretically and in practice, the finite index set Λ employed in the construction of $\hat{\mathbf{P}}_{\Delta,\Lambda}$ has to be chosen carefully. On the one hand, we have the usual bias-variance balance that lets us choose the highest resolution level J in accordance with the smoothness s of the eigenfunction and the number N of observations. On the other hand we have to take into account the singular behaviour such that we shall refine more in the neighbourhood of the boundary points. We roughly choose a maximal frequency level $J(y)$ for wavelets with support in the point $y \in (0,1)$ that satisfies $2^{J(y)} \sim 2^J \min(y, 1-y)^{-1+\varepsilon}$ with some small $\varepsilon > 0$, see Proposition 3.3 for details.

It remains to estimate μ_M, which we propose to do by the – up to transformation – classical projection estimate

$$\hat{\mu}_M(y) := \sum_{|\lambda| \le J} \hat{\mu}_\lambda \psi_\lambda, \quad \hat{\mu}_\lambda := \frac{1}{N+1} \sum_{n=0}^{N} \psi_\lambda(\hat{Y}_n).$$

Equipped with these estimates we use formula (3.3.1) in order to derive an estimate $\hat{\sigma}_M^2$ of σ_M^2 and use the estimated invariant law to transform it to an estimator $\hat{\sigma}^2$ of σ^2, which is our proposed spectral estimator.

Let us summarize our estimation procedure:

1. Form the empirical distribution function $\hat{M}$ and the transformed observations $\hat{Y}_n = \hat{M}(X_{n\Delta})$, $n = 0, 1, \ldots, N$.
2. Estimate the transition operator by the matrix $\hat{\mathbf{P}}_{\Delta,\Lambda}$ of empirical wavelet coefficients.
3. Calculate the first nontrivial spectral pair $(\kappa_1, \mathbf{u}_1)$ of $\hat{\mathbf{P}}_{\Delta,\Lambda}$ and build the estimate $\hat{u}_{1,M}$ of the eigenfunction.
4. Estimate the invariant density μ by some classical method.
5. Derive the estimator $\hat{\sigma}^2_M$ by inserting the preceding estimates in formula (3.3.1) and transform it back to the real line.

As already mentioned in the introduction, we provide a proof in the case that the invariant law of X is known, that is M and μ are available exactly. In this case our spectral estimator is given by

$$\hat{\sigma}^2(x) := \frac{2\hat{\nu}_1 \int_0^{M(x)} \hat{u}_{1,Y}(\eta)\, d\eta}{\hat{u}'_{1,Y}(M(x))\mu^2(x)}, \tag{3.3.2}$$

derived from formula (3.3.1) by plug-in and transformation. To avoid theoretical complications we must keep $\hat{\nu}_1$, $\|\hat{u}_{1,Y}\|_{L^2}$ and $\|\mu_M \hat{u}'_{1,Y}\|^2_{L^2}$ uniformly bounded, e.g. by applying a cut-off for unreasonably large values. Similarly, we guarantee that the a priori knowledge $\hat{\sigma}^2(x) \geq \sigma_0^2$ is fulfilled by changing the denominator if necessary. Then our main result is the following:

THEOREM 3.1 *Let us assume that the invariant distribution function M and its derivative μ are known. Then for $\sigma \in C^s_b(\mathbb{R})$ and $b \in C^{s-1}(\mathbb{R})$ the spectral volatility estimator $\hat{\sigma}^2$ from* (3.3.2) *satisfies for any $\delta > 0$*

$$(2^{-2Js} + N^{-1}2^{3J})^{-1} \int_{M^{-1}(\delta)}^{M^{-1}(1-\delta)} |\hat{\sigma}^2(x) - \sigma^2(x)|^2 \mu(x)\, dx = O_P(1).$$

In particular, we obtain with the asymptotically optimal choice $2^J \sim N^{1/(2s+3)}$ that $N^{s/(2s+3)}\|\hat{\sigma}^2 - \sigma^2\|_{L^2(K)}$ is bounded in probability for any compact set $K \subset \mathbb{R}$.

REMARK 3.1 *For true minimax results we should have a uniform constant for all parameters in some smoothness class. This might be possible, although very technical, but requires also uniform estimates on the separation of the spectrum which are usually difficult to obtain. It is not clear whether it is*

possible to get rid of the restriction to bounded intervals. In the case of reflected diffusion a lower bound proof shows that estimation at the boundary is definitely more difficult, but whether this holds also in our situation with a μ-weighted loss function is an open question. Following the approach in (Gobet, Hoffmann, & Reiß 2002) we can extend our procedure to estimate also the drift coefficient $b(\bullet)$.

3.4 Mathematical Results

LEMMA 3.1 *Suppose $\sigma \in C_b^s(\mathbb{R})$ and $b \in C^{s-1}(\mathbb{R})$ with $b' \in C_b^{s-2}(\mathbb{R})$ for some $s \geq 2$. Then the inverse of the scale density S_Y of Y is s-times differentiable and satisfies*

$$|S_Y^{(r)}(y)| \lesssim S_Y(y) \left| \frac{b_M(y)}{\mu_M(y)} \right|^r \quad 0 \leq r \leq s.$$

PROOF:
The derivatives $S^{(r)}(x)$ for $r \leq s$ are given by

$$S(x) = \tfrac{1}{2}\sigma^2(x)\mu(x) =: a(x)\mu(x), \qquad S^{(r)}(x) = \sum_{k=0}^{r} \binom{r}{k} a^{(k)}(x)\mu^{(r-k)}(x).$$

Applying iteratively the formula $\mu'(x) = 2(-\sigma'(x)+b(x))\mu(x)/\sigma(x)$ and using that $\sigma^{(r)}(\bullet)$, $0 \leq r \leq s$, and $b^{(r)}(\bullet)$, $1 \leq r \leq s$, are uniformly bounded, we obtain

$$|S^{(r)}(x)| \lesssim |b(x)|^r \mu(x).$$

If we now use $S_Y(y) = S_M(y)\mu_M(y)$ and thus $S_Y'(y) = (S')_M(y)$, we arrive at

$$|S_Y^{(r)}| \lesssim \left| (S^{(r)})_M(y)\mu_M^{-r+1} \right| + \left| (S')_M(y)(\mu^{(r-1)})_M \mu_M^{-r+1} \right| \lesssim b_M^r \mu_M^{-r+2}.$$

By the uniform ellipticity condition on $\sigma(\bullet)$ the assertion follows. □

PROPOSITION 3.2 *Suppose $\sigma \in C_b^s(\mathbb{R})$ and $b \in C^{s-1}(\mathbb{R})$ for some $s \geq 2$, and the eigenfunction u of L satisfies $u, u' \in L^p(\mu)$ for some $p \geq 2$. Then the derivatives of the corresponding eigenfunction u_Y of L_Y exist up to order $s+1$ and satisfy for any $1 \leq r \leq s+1$*

$$\mu_M w^{r-1} u_Y^{(r)} \in L^p(0,1) \qquad \textit{with the weight function } w(y) := \min(y, 1-y).$$

PROOF:
The $L^p(\mu)$-integrability of u and u' translates via $u_Y = u_M$ into u_Y, $u'_Y \mu_M \in L^p(0,1)$. We now apply the eigenfunction relation

$$S_Y u''_Y + S'_Y u'_Y = (S_Y u'_Y)' = \nu u_Y \in L^p(0,1).$$

From $S'_Y \mu_M^{-1} = (S'\mu^{-1})_M = b_M$ we conclude $\|(S'_Y)^{-1}\mu_M\|_\infty < \infty$ and

$$S_Y (S'_Y)^{-1} \mu_M u''_Y = \mu_M u'_Y - \nu (S'_Y)^{-1} \mu_M u_Y \in L^p(0,1).$$

Consequently the estimate $|S'_Y(y)| \lesssim |S_Y(y) b_M(y) \mu_M^{-1}(y)|$ from Lemma 3.1 shows that $\mu_M^2 b_M^{-1} u''_Y \in L^p(0,1)$ holds. More generally, we use $(S_Y u'_Y)^{(r)} = \nu u_Y^{(r-1)}$ and $\|(S_Y^{(r)})^{-1}\mu_M\|_\infty < \infty$ to obtain inductively over $0 \le r \le s$

$$S_Y (S_Y^{(r)})^{-1} \mu_M u_Y^{(r+1)} \in L^p(0,1).$$

Hence, Lemma 3.1 yields that $\mu_M^{r+1} b_M^{-r} u_Y^{(r+1)}$ lies in $L^p(0,1)$ for $0 \le r \le s$. While $b_M \mu_M^{-1}$ is obviously bounded on compact subintervals of $(0,1)$, we have by L'Hopital's rule

$$\lim_{y \to 0+} \frac{b_M(y) y}{\sigma_M^2(y) \mu_M(y)} = \lim_{x \to -\infty} \frac{b(x) M(x)}{\sigma^2(x) \mu(x)} = \lim_{x \to -\infty} \frac{b'(x) M(x) + b(x) \mu(x)}{2 b(x) \mu(x)}.$$

Due to $M(x)\mu^{-1}(x) \to 0$ and $b'(x)/b(x) \to 0$ (μ decays faster than exponentially because of $|b(x)| \to \infty$ and b' is bounded) we obtain

$$b_M(y) y \sim \sigma_M^2(y) \mu_M(y) \sim \mu_M(y) \text{ for } y \to 0. \tag{3.4.1}$$

Together with the symmetric argument for $y \to 1$ we obtain the assertion. □

PROPOSITION 3.3 *The projection $\Pi_\Lambda u_Y$ of the eigenfunction u_Y of L_Y with*

$$\Lambda := \Lambda(J,\varepsilon) := \{(j,k) \,|\, j \le J \text{ or } w(k2^{-j}) \in (2^{-J/\varepsilon}, 2^{(J-j)/(1-\varepsilon)})\}$$

satisfies $\|(\boldsymbol{I} - L_Y)^{1/2}(\boldsymbol{I} - \Pi_\Lambda) u_Y\|_{L^2(0,1)} \lesssim 2^{-Js}$ for any $J \in \mathbb{N}$, provided $\sigma \in C_b^s(\mathbb{R})$, $b' \in C_b^{s-2}(\mathbb{R})$ and $\varepsilon \in (0, (p-2)/2ps)$.

REMARK 3.2 *By construction of Λ, we only use wavelet coefficients in $\Pi_\Lambda u_Y$ up to the maximal frequency level J/ε.*

PROOF:
Due to $\|(\boldsymbol{I} - L)^{1/2} f\|_{L^2}^2 = \|f\|_{L^2}^2 + \|S_Y^{1/2} f'\|_{L^2}^2$ we can separately bound

the norms of $(\boldsymbol{I} - \Pi_\Lambda)u_Y$ and its derivative. Since the first norm bound is a much simpler version of the second, we only present the estimate for $\|S_Y^{1/2}((\boldsymbol{I} - \Pi_\Lambda)u_Y)'\|_{L^2}$. For this note that due to inequalities of the type

$$\|S_Y^{1/2}u_Y'\mathbf{1}_{[0,\delta]}\|_{L^2} \le \|S_Y^{1/2}u_Y'\|_{L^p}\delta^{(p-2)/2p}, \quad p > 2,$$

we only need to bound the $L^2(\delta, 1-\delta)$-norm with $\delta^{(p-2)/2p} \sim 2^{-Js}$, that is $\delta \sim 2^{-2Jsp/(p-2)}$ and thus $\delta/2^{-J/\varepsilon} \to \infty$.

We use the compact support and the vanishing moment property of the wavelet functions and its derivatives following the classical approximation estimates via Taylor expansion. Denoting the supporting interval of ψ_λ by $\mathfrak{S}_\lambda$, that is $\mathfrak{S}_{j,k} = [k2^{-j}, (s_0+k)2^{-j}]$, and its length by $|\mathfrak{S}_\lambda|$, we obtain

$$\begin{aligned}
\|S_Y^{1/2}((\boldsymbol{I} - \Pi_\Lambda)u_Y)'\|_{L^2}^2 &= \int_0^1 S_Y(y)\Big(\sum_{\lambda\notin\Lambda}\langle u_Y, \psi_\lambda\rangle\psi_\lambda'(y)\Big)^2 dy \\
&\lesssim \int_0^1 \Big(\sum_{\lambda\notin\Lambda}|\mathfrak{S}_\lambda|^s\Big|\int_{\mathfrak{S}_\lambda} u_Y^{(s+1)}\Big|\|\psi_\lambda\|_{L^1}S_Y^{1/2}(y)\psi_\lambda'(y)\Big)^2 dy \\
&\lesssim \Big\|\sum_{\lambda\notin\Lambda}2^{-(s+1)|\lambda|}\Big(\int_{\mathfrak{S}_\lambda}|u_Y^{(s+1)}|^2\Big)^{1/2}\big(\max_{y\in\mathfrak{S}_\lambda}S_Y^{1/2}(y)\big)\psi_\lambda\Big\|_{H^1}^2 \\
&\sim \sum_{\lambda\notin\Lambda}2^{-2s|\lambda|}\big(\max_{y\in\mathfrak{S}_\lambda}S_Y(y)\big)\int_{\mathfrak{S}_\lambda}|u_Y^{(s+1)}|^2.
\end{aligned}$$

Since we only need to consider wavelet coefficients $(j,k) \notin \Lambda$ satisfying additionally $w(k2^{-j}) \ge \delta$ with $\delta/2^{-J/\varepsilon} \to \infty$, the corresponding support intervals $\mathfrak{S}_{j,k}$ have a distance of at least $\max(\delta, 2^{(J-j)/(1-\varepsilon)}) - s_0 2^{-j} \gtrsim 2^{-j}$ from the boundary. The estimates $S_Y(y) \sim \mu_M^2(y)$ and $\mu_M(y) \gtrsim w(y)$ yield

$$\frac{\sup_{y\in S_{j,k}}S_Y(y)}{\inf_{y\in S_{j,k}}S_Y(y)} \lesssim \Big(1 + \frac{S2^{-j}}{2^{-j}}\Big)^2 \sim 1,$$

which gives the bound

$$\|S_Y^{1/2}((\boldsymbol{I} - \Pi_\Lambda)u_Y)'\|_{L^2(\delta,1-\delta)}^2 \lesssim \sum_{\lambda\notin\Lambda}2^{-2s|\lambda|}\int_{S_\lambda}S_Y(\zeta)|u_Y^{(s+1)}(\zeta)|^2\,d\zeta.$$

We apply the Hölder inequality with $\frac{p}{2}$ and $q = \frac{p}{p-2} > 1$ and obtain for $j \geq J$ by Proposition 3.2

$$\begin{aligned}
&\sum_{k:\,(j,k)\notin\Lambda} \int_{S_{j,k}} S_Y(\zeta)|u_Y^{(s+1)}(\zeta)|^2\,d\zeta \\
&\leq \Big(\sum_k \int_{S_{j,k}} S_Y(\zeta)^{p/2}|u_Y^{(s+1)}(\zeta)|^p w^{sp}(\zeta)\,d\zeta\Big)^{2/p}\Big(\sum_k \int_{S_{j,k}} w^{-2sq}(\zeta)\,d\zeta\Big)^{1/q} \\
&\lesssim \|S_Y^{1/2} w^s u_Y^{(s+1)}\|_{L^p}^2 \Big(\sum_k 2^{-j} w(k2^{-j})^{-2sq}\Big)^{1/q} \\
&\lesssim \big(2^{-j}2^{2jqs}\big(2^{(J-j)/(1-\varepsilon)}2^j\big)^{1-2qs}\big)^{1/q} \\
&= 2^{(J-j)(q^{-1}-2s)/(1-\varepsilon)}.
\end{aligned}$$

Consequently, $\|S_Y^{1/2}((\boldsymbol{I}-\Pi_\Lambda)u)'\|_{L^2}$ is of order 2^{-Js}, provided

$$\sum_{j\geq J} 2^{-2(j-J)s} 2^{(J-j)(q^{-1}-2s)/(1-\varepsilon)} = \sum_{j\geq 0} 2^{-j(2s+(q^{-1}-2s)/(1-\varepsilon))}$$

is finite, which is ensured for $\varepsilon < (p-2)/2ps$. □

PROPOSITION 3.4 *For any function $v \in V_\Lambda$, $\|v\|_{L^2} = 1$, we have*

$$\mathbb{E}[\|(\boldsymbol{I}-L_Y)^{1/2}(\hat{P}_{\Delta,\Lambda}-P_{\Delta,\Lambda})v\|_{L^2}^2] \lesssim N^{-1}2^{3J},$$

where we have introduced the operators $P_{\Delta,\Lambda} := \Pi_\Lambda P_{\Delta,Y}$ and

$$\hat{P}_{\Delta,\Lambda}v := \sum_{\lambda\in\Lambda} \big(\hat{\mathbf{P}}_{\Delta,\Lambda}(\langle v,\psi_{\lambda'}\rangle)_{\lambda'\in\Lambda}\big)_\lambda \psi_\lambda.$$

PROOF:
The bound on $\|(\hat{P}_{\Delta,\Lambda}-P_{\Delta,\Lambda})v\|_{L^2}$ is again easy and therefore omitted.

We obtain by the mixing properties of Y, cf. Lemma 5.2 in (Gobet, Hoffmann, & Reiß 2002):

$$\begin{aligned}
&\mathbb{E}\Big[\|S_Y^{1/2}((\hat{P}_{\Delta,\Lambda}-P_{\Delta,\Lambda})v)'\|_{L^2}^2\Big] \\
&= \int_0^1 S(y)\,\mathrm{Var}\Big[\sum_{\lambda\in\Lambda}\frac{1}{N}\sum_{n=1}^N \psi_\lambda(Y_{(n-1)\Delta})v(Y_{n\Delta})\psi_\lambda'(y)\Big]\,dy \\
&\lesssim N^{-1}\int_0^1 S_Y(y)\mathbb{E}\Big[\Big(\sum_{\lambda\in\Lambda}(\psi_\lambda(Y_0)v(Y_\Delta)\psi_\lambda'(y)\Big)^2\Big]\,dy \\
&= N^{-1}\sum_{\lambda,\lambda'\in\Lambda}\Big(\int_0^1 S_Y(y)\psi_\lambda'(y)\psi_{\lambda'}'(y)\,dy\Big)\mathbb{E}\Big[\psi_\lambda(Y_0)\psi_{\lambda'}(Y_0)v^2(Y_\Delta)\Big].
\end{aligned}$$

Because of (3.4.1) and the logarithmic growth bound on b_M we obtain for $\lambda, \lambda' \in \Lambda$ with $j' := |\lambda'| \geq |\lambda| =: j$ and $j' > J$

$$\left| \int_0^1 S_Y(y)\psi'_\lambda(y)\psi'_{\lambda'}(y)\,dy \right| \lesssim |S_{\lambda'}|\mu_M^2(2^{(J-j')/(1-\varepsilon)})2^{3j'/2}2^{3j/2}$$
$$\lesssim 2^{2J}(j'-J)^2 2^{(j'-J)(\frac{1}{2}-\frac{2}{1-\varepsilon})}2^{3(j-J)/2}.$$

For $j \leq j' \leq J$ the same term is evidently bounded by $2^{(3j+j')/2}$.

Inserting these estimates and then proceeding similarly for the expectation we obtain

$$\mathbb{E}\Big[\|S_Y^{1/2}((\hat{P}_{\Delta,\Lambda} - P_{\Delta,\Lambda})v)'\|_{L^2}^2\Big]$$
$$\lesssim N^{-1}\sum_{\substack{(j,k),(j',k')\in\Lambda \\ j'\geq\max(j,J+1)}} 2^{2J}(j'-J)^2 2^{(j'-J)(\frac{1}{2}-\frac{2}{1-\varepsilon})}2^{3(j-J)/2}\mathbb{E}\Big[\psi_{j'k'}(Y_0)$$
$$\psi_{jk}(Y_0)v^2(Y_\Delta)\Big] + N^{-1}\sum_{\substack{(j,k),(j',k')\in\Lambda \\ j\leq j'\leq J}} 2^{(3j+j')/2}\mathbb{E}\Big[\psi_{jk}(Y_0)\psi_{j'k'}(Y_0)v^2(Y_\Delta)\Big]$$
$$\lesssim N^{-1}2^{2J}\sum_{j'\geq\max(j,J+1)}(j'-J)^2 2^{(j'-J)(\frac{1}{2}-\frac{2}{1-\varepsilon})}2^{3(j-J)/2}\|v\|_{L^2}^2$$
$$2^{(J-j')/(1-\varepsilon)}2^{j/2}2^{j'/2} + N^{-1}\sum_{j\leq j'\leq J} 2^{(3j+j')/2}\|v\|_{L^2}^2 2^{j/2}2^{j'/2}$$
$$\lesssim N^{-1}2^{3J}\Big(\sum_{j\leq J,\, j'>0}(j')^2 2^{j'(1-\frac{3}{1-\varepsilon})}2^{2(j-J)} + \sum_{j>0,\, j'\geq j}(j')^2 2^{j'(1-\frac{3}{1-\varepsilon})}$$

$$\times 2^{2j} + \sum_{j\leq j'\leq J} 2^{2j+j'-3J}\Big) \lesssim N^{-1}2^{3J}. \quad \square$$

The next result is essential for the spectral approximation to work. It is stated as Proposition 2.9 and Corollary 2.13 in (Gobet, Hoffmann, & Reiß 2002).

PROPOSITION 3.5 *Suppose a selfadjoint bounded linear operator T on a Hilbert space has a simple eigenvalue κ such that κ has distance ρ from the remaining spectrum. Let T_ε be a second linear operator with $\|T_\varepsilon - T\| < \frac{1}{2}\rho^{-1}$. Then the operator T_ε has a simple eigenvalue κ_ε and there are normalized eigenvectors u and u_ε with $Tu = \kappa u$, $T_\varepsilon u_\varepsilon = \kappa_\varepsilon u_\varepsilon$ satisfying*

$$|\kappa_\varepsilon - \kappa| + \|u_\varepsilon - u\| \lesssim \rho\|(T_\varepsilon - T)u\|.$$

PROOF:
[Proof of Theorem 3.1.] We apply the preceding proposition to the Hilbert space $H = \mathcal{D}((\boldsymbol{I} - L_Y)^{1/2})$, the domain of the operator $(\boldsymbol{I} - L_Y)^{1/2}$ on $L^2(0,1)$, and with the operators $P_{\Delta,Y}$ and $\hat{P}_{\Delta,\Lambda}$. The functional calculus shows that L_Y and $P_{\Delta,Y}$ are selfadjoint on H. For any normalized eigenfunction u_Y of $P_{\Delta,Y}$ we obtain from Proposition 3.3 using $P_{\Delta,Y}u_Y = \kappa u_Y$ and from

Proposition 3.4 that

$$\mathbb{E}\big[\|(\boldsymbol{I}-L)^{1/2}(\hat{P}_{\Delta,\Lambda}-P_{\Delta,Y})u_Y\|_{L^2}^2\big] \lesssim 2^{-2Js}+N^{-1}2^{3J}.$$

The spectral approximation result in Proposition 3.5 thus gives

$$\mathbb{E}\big[\big(|\hat{\kappa}_1-\kappa_1|^2+\|(\boldsymbol{I}-L)^{1/2}(\hat{u}_{1,Y}-u_{1,Y})\|_{L^2}^2\big)\mathbf{1}_A\big] \lesssim 2^{-2Js}+N^{-1}2^{3J}$$

on the random set $A=\{\|\hat{P}_{\Delta,\Lambda}-P_{\Delta,Y}\|<\frac{1}{2}\rho^{-1}\}$. By the strong law of large numbers for mixing sequences and the smoothing property of $P_{\Delta,Y}$ it follows $\mathbb{P}(A)\to 1$ for $N,J\to\infty$. Keeping $|\hat{\kappa}_1|+\|(\boldsymbol{I}-L)^{1/2}\hat{u}_{1,Y}\|_{L^2}$ uniformly bounded, we obtain using $S_Y\sim\mu_M^2$

$$\mathbb{E}\big[|\hat{\nu}_1-\nu_1|^2\big]+\mathbb{E}\big[\|\hat{u}_{1,Y}-u_{1,Y}\|_{L^2}^2\big]+\mathbb{E}\big[\|\mu_M(\hat{u}'_{1,Y}-u'_{1,Y})\|_{L^2}^2\big] \lesssim 2^{-2Js}+N^{-1}2^{3J}\quad.$$

Note that we have bounded the estimation risk for ν_1 by that of κ_1 due to the continuity of the transformation involved. From

$$\mathbb{E}\Big[\Big\|\int_0^{\bullet}(\hat{u}_{1,Y}-u_{1,Y})\Big\|_{L^2}^2\Big] \lesssim 2^{-2Js}+N^{-1}2^{3J}$$
$$\mathbb{E}\big[\|(\hat{u}'_{1,Y}-u'_{1,Y})\mu_M^2\|_{L^2}^2\big] \lesssim 2^{-2Js}+N^{-1}2^{3J}$$

and the fact that $u'_{1,Y}$ does not vanish inside $(0,1)$ we infer for any fixed $R,\delta>0$ by the usual triangle inequality argument and the exclusion of explosions

$$(2^{-2Js}+N^{-1}2^{3J})^{-1}\mathbb{E}\Big[\int_\delta^{1-\delta}|\hat{\sigma}_M^2(y)-\sigma_M^2(y)|^2\,dy\wedge R\Big]\lesssim 1.$$

Hence, transforming back to the real line gives

$$(2^{-2Js}+N^{-1}2^{3J})^{-1}\mathbb{E}\Big[\int_{M^{-1}(\delta)}^{M^{-1}(1-\delta)}|\hat{\sigma}^2(x)-\sigma^2(x)|^2\mu(x)\,dx\wedge R\Big]\lesssim 1.$$

The fact that $\mathbb{E}[|X-Y|\wedge R]$ is a metric for convergence in probability then gives the result. □

Bibliography

Bakry, D. (1994) L'hypercontractivite et son utilisation en theorie des semigroupes. (Hypercontractivity and its usage in semigroup theory)., in *Bakry, Dominique (ed.) et al., Lectures on probability theory. Ecole d'Ete de Probabilites de Saint-Flour XXII-1992. Berlin: Springer. Lect. Notes Math. 1581, 1-114* .

Bass, R. F. (1998) *Diffusions and elliptic operators.* Probability and Its Applications. New York, NY: Springer.

Chen, X., L. P. Hansen, & J. A. Scheinkman (1997) Shape preserving estimation of diffusions, Manuscript.

Cohen, A. (2000) Wavelet methods in numerical analysis., in *Handbook of numerical analysis. Vol. 7*, ed. by P. G. Ciarlet. North-Holland/ Elsevier, Amsterdam.

Gobet, E., M. Hoffmann, & M. Reiß (2002) Nonparametric estimation of scalar diffusions based on low-frequency data is ill-posed, Discussion paper 57, Sonderforschungsbereich 373, Humboldt University Berlin (to appear in Annals of Statistics).

Hansen, L. P., J. A. Scheinkman, & N. Touzi (1998) Spectral methods for identifying scalar diffusions, *Journal of Econometrics*, 86, 1–32.

Karlin, S., & H. M. Taylor (1981) *A second course in stochastic processes.* New York etc.: Academic Press.

Kerkyacharian, G., & D. Picard (2003) Regression in random design and warped wavelets., Preprint, Laboratoire de probabilités et modèles aléatoires, Universités de Paris 6 et 7.

Kessler, M., & M. Sørensen (1999) Estimating equations based on eigenfunctions for a discretely observed diffusion process., *Bernoulli*, 5(2), 299–314.

Kleinow, T. (2002) Testing the diffusion coefficient, Discussion paper 38, Sonderforschungsbereich 373, Humboldt University Berlin.

Ledoux, M. (1998) The geometry of Markov diffusion generators, Lecture notes, ETH Zürich.

4 Linear Regression Models for Functional Data[1]

Hervé Cardot[2] and Pascal Sarda[3,4]

[2] Unité Biométrie et Intelligence Artificielle, INRA Toulouse, BP 27, F-31326 Castanet-Tolosan Cedex, France
[3] LSP, UMR C5583, Université Paul Sabatier, 118, route de Narbonne, F-31062 Toulouse Cedex, France
[4] GRIMM, EA2254, Université Toulouse-le-Mirail, 5 Allées Antonio-Machado, F-31058 Toulouse Cedex, France

Summary

This paper addresses a specific case of regression analysis: the predictor is a random curve and the response is a scalar. We consider three models: the functional linear model, the functional generalized linear model and functional linear regression on quantiles. Spline functions are used to build estimators which minimize a penalized criterion. The method is illustrated by means of real data examples. Then, we give asymptotics results for these estimators.

Keywords: Functional linear model, generalized functional linear model, conditional quantiles, regularization, roughness penalty, splines

4.1 Introduction

This paper is concerned with statistical modelization for Regression Analysis in the case of functional data: more precisely we consider the case where the predictor is a curve linked to a scalar response variable. This arises in the three situations analyzed below: the first one concerns wheat yield estimation, the second one deals with remote sensing whereas the third one is an application to ozone prediction. Other examples may be found in the literature: in chemometrics (Osborne *et al.*, 1984), climatology (Ramsay and Silverman, 1997) or linguistics (Marx and Eilers, 1996).

[1] We would like to thank all the members and participants of the working group on functional data STAPH from Toulouse for helpful discussions.

In this context we are faced with the problem of estimating the link between a real random response Y and a square integrable random function X defined on some compact set $\mathcal{C}$ of $\mathbb{R}$ through a sample (X_i, Y_i), $i = 1, \ldots, n$ drawn from (X, Y). Of course, in practice the curves X_i's are discretized (possibly not at the same points) and then statistical multivariate methods may be applied to these random vectors with special attention to the problems of high dimension and multicollinearity inherent to this kind of data. Frank and Friedman (1993) summarize the main (chemometrics) regression tools *i.e.* Partial Least Squares (PLS), Principal Components Regression (PCR) and Ridge Regression (RR).

These methods may perform well in several applications as it is for instance the case in chemometrics where the predictive curves are quite smooth (typically sinusoidal signal curves). However, the functional nature of the data is not taken into account by these methods and Marx and Eilers (1999) show that functional methods might work better especially when the curve predictors are not smooth. More generally, a part of literature has been recently concerned with functional data in a variety of statistical problems and authors aim at developing *ad hoc* procedures based on smoothing techniques. The monographs from Ramsay and Silverman (1997, 2002) give good insights into a variety of models dealing with data taken as curves.

After giving some notations and definitions in section 2, we present three regression models for the situation described above: the linear regression model is defined in section 3 and its extension to the generalized linear model in section 4; section 5 deals with the problem of estimating a conditional quantile. For each model we define an estimator based on a B-splines basis expansion of the functional coefficient to be estimated. Problems linked to the implementation of these procedures are discussed by means of real data examples. Then, in section 6 we give some L^2 convergence results for estimators and especially we derive an upper bound for the rate of convergence. It shows the importance of introducing a regularization (or penalty) in the criterion to be minimized.

4.2 Notations and Definitions for Functional Data

In the following we suppose that the random variable X takes values into some real separable Hilbert space H and we consider that this space is the space of square integrable functions defined on $[0, 1]$. Let $< \varphi, \psi >$ denote the usual inner product of functions φ and ψ in H and $\|\varphi\|$ the norm of φ.

If we suppose that the H-valued random variable X, assumed to be centered ($EX(t) = 0$, for t a.e.) has a finite second moment ($E(\|X\|^2) < \infty$), the covariance operator Γ is defined as

$$\Gamma x(t) = \int_0^1 E\left[X(t)X(s)\right] x(s)ds, \; x \in H, \; t \in [0,1].$$

Note that Γ is an integral operator whose kernel is the covariance function of X and it may be shown that the operator Γ is nuclear, self-adjoint and non-negative (Dauxois and Pousse, 1976, and Dauxois, Pousse and Romain, 1982).

In the same way, we define the cross covariance operator Δ of (X, Y). It is the linear functional defined as

$$\Delta x = \int_0^1 E\left[X(t)Y\right] x(t)dt, \; x \in H.$$

In the following, we denote by λ_j, $j = 1, 2, \ldots$ the eigenvalues of Γ and by v_j, $j = 1, 2, \ldots$ a complete orthonormal system of eignefunctions.

4.3 The Functional Regression Linear Model

4.3.1 Definition of the Model and Spline Estimators

The *functional regression linear model with scalar response* is defined as

$$Y = \int_0^1 \alpha(t)X(t)dt + \varepsilon, \tag{4.3.1}$$

where α is a square integrable function defined on $[0,1]$ and ε is a random variable such that $E\varepsilon = 0$ and $EX(t)\varepsilon = 0$, for t a.e. Model (4.3.1) may be written as

$$Y = \Psi(\{X(t), t \in [0,1]\}) + \varepsilon, \tag{4.3.2}$$

where Ψ is some continuous linear functional. This model traces back to Hastie and Mallows (1993) (see also Ramsay and Silverman, 1997). Cardot, Ferraty and Sarda (2003) shows that condition 1 below insures existence and unicity of α in model (4.3.1) and the unique solution α of the model satisfies

$$\alpha = \sum_{j=1}^{\infty} \frac{< E\left(XY\right), v_j >}{\lambda_j} v_j. \tag{4.3.3}$$

Condition 1 The random variables X and Y satisfy

$$\sum_{j=1}^{\infty} \frac{< E(XY), v_j >^2}{\lambda_j^2} < \infty.$$

Condition 1 is known as the Picard condition in the field of linear inverse problems (see e.g. Kress, 1989).

Several procedures have been proposed in the literature to estimate the functional coefficient α (sometimes called the *contrast template*) and/or the functional Ψ from a "functional point of view". Hastie and Mallows (1993) propose for α an estimator that minimizes a penalized least squares criterion, the solution being a cubic spline. This method is studied by Ramsay and Silverman (1997) which discuss various computational aspects. A second approach, proposed by Hastie and Mallows, is based on a smooth basis expansion of the function α. Marx and Eilers (1999) use a smooth B-spline expansion for α and introduce a difference penalty in a log-likelihood criterion in the context of smoothed generalized linear regression. Direct estimation of the functional Ψ has been achieved in Cardot *et al.* (1999) by means of a functional PCR, in the setting of a predictor valued in a general real separable Hilbert space. A smooth version of this Functional Principal Components Regression has been studied in Cardot, Ferraty and Sarda (2003).

We define below the spline estimators proposed by Cardot, Ferraty, and Sarda (2003) which combines ideas from Marx and Eilers (1996) and Hastie and Mallows (1993). Suppose that q and k are integers and let S_{qk} be the space of *splines* defined on $[0,1]$ with degree q and $k-1$ equispaced interior knots. The space S_{qk} has dimension $q+k$ and one can derive a basis by means of normalized B-splines $\{B_{k,j}, j = 1, \ldots, k+q\}$ (see de Boor (1978)). In the following we denote by $\mathbf{B}_k$ the vector of all the B-splines and by $\mathbf{B}_k^{(m)}$ the vector of derivatives of order m of all the B-splines for some integer m ($m < q$).

The penalized B-splines estimator of α is thus defined as

$$\widehat{\alpha}_{PS} = \sum_{j=1}^{q+k} \widehat{\theta}_j B_{k,j} = \mathbf{B}_k' \widehat{\theta}, \tag{4.3.4}$$

where $\widehat{\boldsymbol{\theta}}$ is a solution of the minimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{q+k}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q+k} < \theta_j B_{k,j}, X_i > \right)^2 + \rho \left\| \mathbf{B}_k^{(m)'} \boldsymbol{\theta} \right\|^2, \tag{4.3.5}$$

with smoothing parameter $\rho > 0$. The solution $\widehat{\boldsymbol{\theta}}$ of the minimization problem

(4.3.5) is given by

$$\widehat{\boldsymbol{\theta}} \;=\; \widehat{\mathbf{C}}_{\rho}^{-1}\widehat{\mathbf{b}} \;=\; \left(\widehat{\mathbf{C}} + \rho \mathbf{G}_k\right)^{-1} \widehat{\mathbf{b}}, \tag{4.3.6}$$

where $\widehat{\mathbf{C}}$ is the $(q+k)\times(q+k)$ matrix with elements $n^{-1}\sum_{i=1}^{n} < B_{k,j}, X_i > < B_{k,l}, X_i >$, $\widehat{\mathbf{b}}$ is the vector in $\mathbb{R}^{q+k}$ with elements $n^{-1}\sum_{i=1}^{n} < B_{k,j}, X_i > Y_i$, and where $\mathbf{G}_k$ is the $(q+k)\times(q+k)$ matrix with elements $< B_{k,j}^{(m)}, B_{k,l}^{(m)} >$. In the special case $m = 0$, the minimization criterion (4.3.5) becomes

$$\frac{1}{n}\sum_{i=1}^{n} \left(Y_i - < \mathbf{B}_k'\boldsymbol{\theta}, X_i >\right)^2 \;+\; \rho \left\|\mathbf{B}_k'\boldsymbol{\theta}\right\|^2,$$

which is a functional generalization of the ridge regression criterion.

4.3.2 An Application to Wheat Yield Estimation

During the last decades, crop simulations models became more and more accurate to predict development and growth of several species, such as corn or wheat. Crop models are now able to help farmers to reach a decision about irrigation, sowing time or nitrogen fertilization. These models are highly complex dynamic models generally involving more than fifty variables such as soil characteristics, climatic variations, ..., with daily time step, that simulate the behaviour of the soil-crop system within one year. The crop model considered here is Déciblé (Meynard, 1997) and deals with winter wheat crop management.

We have a sample of size $n = 198$ of yield measures resulting from Déciblé model. We aim at giving a simple statistical model which can summarize the most important effects of the climate, simulated by a generator, on the wheat yield. For that purpose we have, for each crop, the daily cumulative temperatures (denoted by T) and the daily cumulative precipitations (denoted by P) measured between the first day of September and the end of August of the following year. Let us notice that the sowing time is a random variable and the harvest depends on climatic variations and thus are different from one sample to one another. These variables are measured between the first day of October which is the beginning of the sowing phase and the first week of August which is the last date for the harvest. Thus, the considered period in our statistical model begins in October, ends in August and lasts 309 days. We have also defined the real variable L_i which is the duration of cultivation for sample i, that is to say the difference between harvest and sowing date.

The following model was considered in order to explain yield as a linear function of the duration of the crop and climatic variations :

$$Y_i = aL_i + \int \alpha_1(t)T_i(t)\,dt + \int \alpha_2(t)P_i(t)\,dt + \epsilon_i, \quad i = 1, \ldots, 198.$$

The parameters of this model are $a \in \mathbb{R}$ associated to the crop duration effect and functions α_1 and α_2, supposed to be twice continuously differentiables, for the climatic effects. It is a kind of hybrid functional linear model which incorporates parametric and nonparametric effects simultaneously. The parameters of this model are estimated by minimizing the following criterion

$$\min_{a,\alpha_1,\alpha_2} \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - aL_i - \int \alpha_1(t)T_i(t)\,dt - \int \alpha_2(t)P_i(t)\,dt\right)^2$$

$$+\rho_1\|\alpha_1^{(2)}\|^2 + \rho_2\|\alpha_2^{(2)}\|^2, \tag{4.3.7}$$

where $\alpha_1 \in \mathcal{S}_{qk}$ and $\alpha_2 \in \mathcal{S}_{qk}$ with $q = 4$ and $k = 25$.

Cross validation was used to choose these smoothing parameters and leaded us to select $\rho_1 = \rho_2 = 0.001$. The cross validation error is mapped in figure 4.1. With these smoothing parameters values, the explained variance of Y, that is to say the squared correlation coefficient, is 0.56. This result can be considered to be rather good taking into account the sample size $n = 198$, the smoothing parameters which act as regularization parameters and the small number of explanatory variables compared to the number of variables involved in the agronomical Déciblé model.

Estimated value for a is $\widehat{a} = 0.46$ with estimated standard deviation $\widehat{\sigma}_a = 0.08$. Thus, the longer the duration of the culture is, the higher the wheat yield. Estimated functions α_1 and α_2 are drawn in fig. 4.2. Let us note that period around the 210th day, that is to say around May, is important for the winter wheat yield. In fact, high temperatures and/or low precipitations during spring and the beginning of summer lead to high yields. On the other hand, a rainy winter implies higher yields whereas temperature effect seems to be negligible during this period. From an agronomical point of view, this can be interpreted as follows: if the winter is rainy, then soil has sufficient water resources to allow wheat to grow rapidly, on the other hand if spring and the beginning of summer are too rainy then wheat is likely to rot or to catch diseases. Furthermore, during this period wheat needs energy to grow and thus higher temperatures lead to higher yields. On the contrary, bad yields may be the consequence of drought if during July (around the 285th day) and the beginning of August temperatures are too high and precipitations are too low.

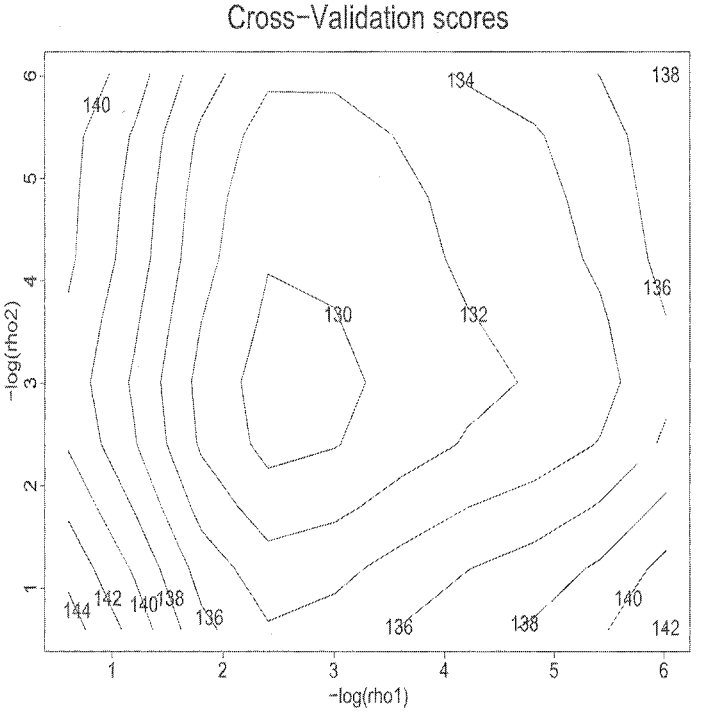


Figure 4.1: Map of cross-validation error for both smoothing parameters ρ_1 and ρ_2. The minimum is attained for $\rho_1 \approx 1e\text{-}03$ and $\rho_2 \approx 1e\text{-}03$.

4.4 The Functional Generalized Linear Model

4.4.1 Definition of the Model and Spline Estimators

In this section, the conditional distribution of Y given $X = x$ is supposed to belong to the exponential family of the form

$$\exp\{b_1(\eta)y + b_2(\eta)\}\,\nu(dy), \tag{4.4.1}$$

where ν is a nonzero measure on $\mathbb{R}$ which is not concentrated at a single point and where the function b_1 is twice continuously differentiable and b_1' is strictly positive on $\mathbb{R}$. Then, the function b_1 is strictly increasing and b_2 is twice continuously differentiable on $\mathbb{R}$. The mean μ of the distribution is

$$\mu \;=\; b_3(\eta) \;=\; -\frac{b_2'(\eta)}{b_1'(\eta)},$$

where b_3 is continuously differentiable and b_3' is strictly positive on $\mathbb{R}$. The function b_3^{-1} is called the *link function* and one has $\eta = b_3^{-1}(\mu)$.

The following *functional generalized linear model* is suppose to holds, that is we assume the existence of a function $\alpha \in H$ such that

$$E(Y|X = x) = b_3(< \alpha, x >), \qquad x \in H. \tag{4.4.2}$$

To insure identifiability of the parameter of the model, we assume that the following condition 2 holds (see Cardot and Sarda, 2003, for details).

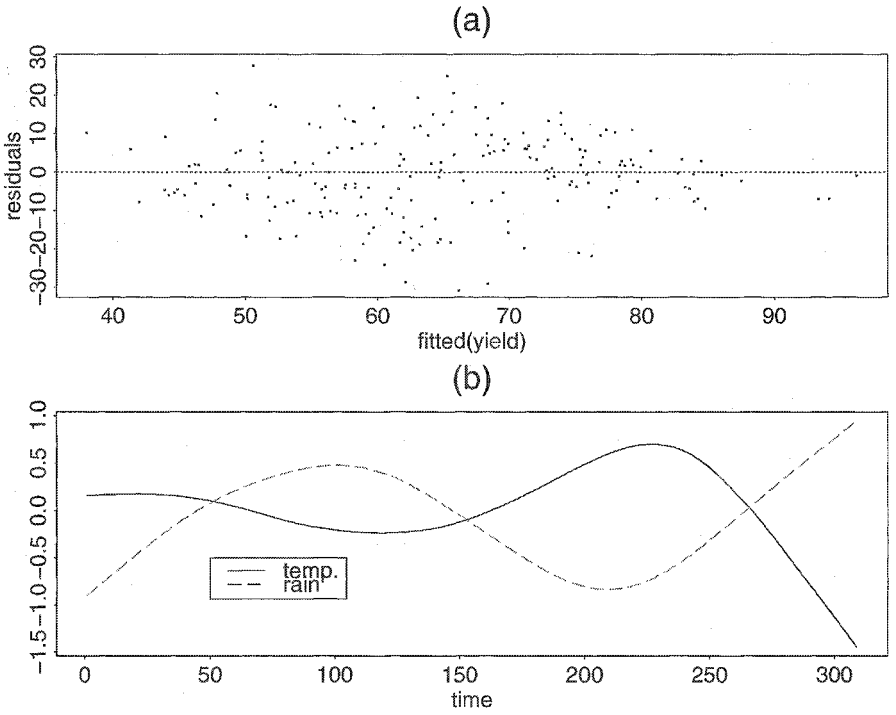


Figure 4.2: (a) Residuals versus fitted wheat yield (smoothing parameter values are those obtained by minimizing the cross-validation error). (b) Plots of estimated functions α_1 and α_2 ($t = 0$ is 1 October and $t = 309$ is the beginning of August of the following year).

Condition 2 There is an interval S in $\mathbb{R}$ such that ν is concentrated on S and

$$b_1''(\eta)y + b_2''(\eta) \ < \ 0, \forall \eta \in R, \forall y \in S.,$$

and

The eigenvalues of Γ are non zero.

We refer to Stone (1986) for examples of exponential families, such as Bernouilli or gamma distribution, satisfying the first part of condition 2.

The penalized B-splines estimator of α introduced in Cardot and Sarda (2003) is defined as

$$\widehat{\alpha}_{PS} \ = \ \sum_{j=1}^{q+k} \widehat{\theta}_j B_{k,j} \ = \ \mathbf{B}_k' \widehat{\boldsymbol{\theta}},$$

where $\widehat{\boldsymbol{\theta}}$ is a solution of the following maximization problem

$$\max_{\boldsymbol{\theta} \in R^{q+k}} \frac{1}{n} \sum_{i=1}^{n} (b_1(< \mathbf{B}_k' \boldsymbol{\theta}, X_i >) Y_i + b_2(< \mathbf{B}_k' \boldsymbol{\theta}, X_i >)) \ - \ \frac{1}{2} \rho \left\| \mathbf{B}_k^{(m)'} \boldsymbol{\theta} \right\|^2,$$

with smoothing parameter $\rho > 0$. The estimator $\widehat{\alpha}_{PS}$ is of the same type as the one introduced by Marx and Eilers (1999), with however a different roughness penalty.

4.4.2 A Remote Sensing Application

On board SPOT 4, a satellite launched in March 1998, the Végétation sensor gives, at a high temporal resolution, daily images of Europe at a coarse spatial resolution, each pixel corresponding to a ground area of 1 km^2. The information given by this sensor are the reflectances, *i.e* the proportion of reflected radiation, in the four spectral bands Blue (B), Red (R), Near Infra-Red (NIR) and Short Wave Infra-Red (SWIR). We also considered two vegetation indices, that are frequently used in bioclimatology and remote sensing (Tucker, 1979), the NDVI (Normalized Difference Vegetation Index), $NDVI = (NIR - R)/(NIR + R)$, and the PVI (Perpendicular Vegetation Index), $PVI = (NIR - 1.2R)/(\sqrt{1 + (1.2)^2})$, which are functions of the reflectances in the Red (R) and NIR channels. This information allows to characterize the developpement of vegetation and crops at the scale of a small country (Tucker, 1979). Because in Europe, and particularly in France, the size of plots is much less than 1 km^2, the observed reflectances are a mixture of different informations since they contain different agricultural plots (maize, wheat, forest, ...).

We aim at estimating the land use, *i.e.* the proportion of each types of culture or land cover inside each mixed pixel. A multilogit model with functional

covariates is proposed to achieve that (see Cardot, Faivre and Goulard 2003 for more details). This is the first step in predicting regional crop productions. Despite the medium spatial resolution, we take advantage of the high temporal resolution of such a sensor to derive estimations of the land use. The proportions are assumed to be drawn from a multinomial distribution whose parameters depend on the temporal evolution of the reflectance.

Let us denote by π_{ij}, $j = 1, \ldots, p$, the proportion of land use of crop j in pixel i of 1 km^2. In our application the observed area is about 40km $\times$40km so that $i = 1, \ldots, n = 1554$. Ten ($p = 10$) different classes of crops were present. The curves of reflectance for each pixel i, in each channel and index, are denoted by $\boldsymbol{X_i} = [X_i(t_1), \ldots, X_i(t_K)]^T$ where $t_1 < \ldots < t_k < \ldots < t_K$ are the instants of measure. The images in which the clouds were too important were removed to finally get $K = 39$ different images from March to August 1998. We assume that the land use is fixed during the observation period.

We suppose the proportions π_{ij} given the temporal evolution of the reflectance $\{X_i(t), t \in T\}$ can be modelled as resulting from a multinomial distribution whose parameters satisfy

$$\mathbb{E}(\pi_{ij}|X_i) = \frac{\exp\left(\delta_j + \int_T \alpha_j(t) X_i(t)\, dt\right)}{\sum_{\ell=1}^{p} \exp\left(\delta_\ell + \int_T \alpha_\ell(t) X_i(t)\, dt\right)}. \qquad (4.4.3)$$

For identifiability reasons we take $\alpha_p = 0$ and $\delta_p = 0$. Each functional coefficient α_j may have an interpretation by comparison to the reference function $\alpha_p = 0$. For instance, if α_j is a positive function, then the ratio of the proportion will be higher than the mean value and thus the class j will be more important in the pixel i, if the centered reflectance curve is positive.

We aim at estimating the vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{p-1})^T$ and the functional coefficients $\alpha_j(t)$, $j = 1, \ldots, p-1$. The estimations are obtained by means of the maximum likelihood criterion.

For computational purposes, we prefered the dimension reduction approach based on a functional principal components analysis. The number of covariates (the principal components) still may be large and we decided to select the most significative parameters by means of the likelihood ratio test with an ascendant procedure. More details may be found in Cardot, Faivre and Goulard (2003).

The initial sample was split into a learning sample composed of 1055 pixels and a test sample composed of 499 pixels.

The estimators for functions α_j in the multilogit model are shown in Fig. (4.3). The reference curve is taken for the theme "Urban" since we expect that it

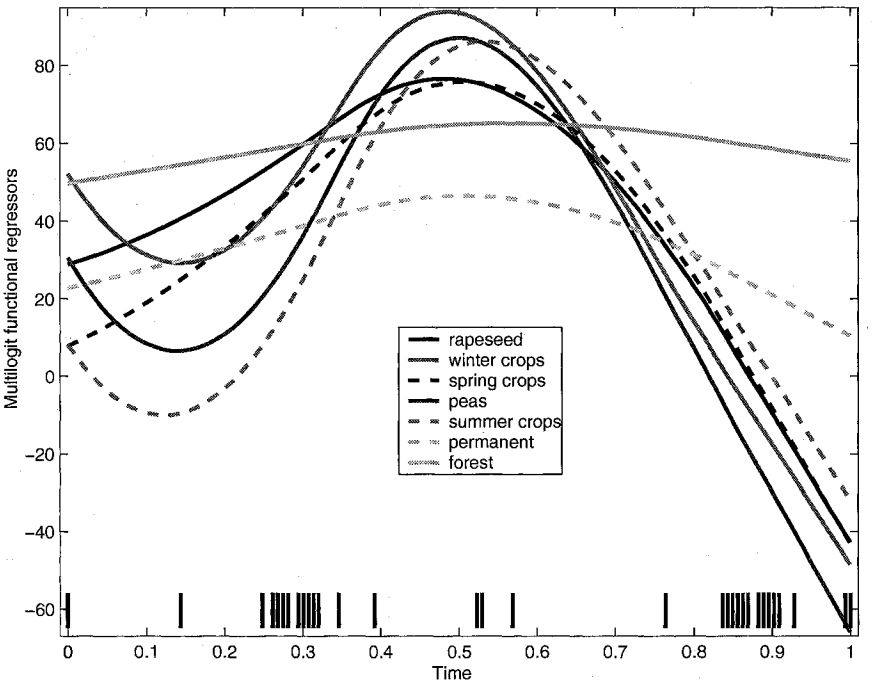


Figure 4.3: Estimated regression functions for the multilogit model with the temporal NDVI index. Theme "urban" taken as reference.

varies less along time. We recognize a biological cycle for the curves associated to crops such as "Winter crops" or "Peas" and a rather flat coefficient for themes such as "Forest" or "Permanent crops".

We defined the following criterion to evaluate, on the test sample, the skill of this approach:

$$R_{ij} = \frac{|\pi_{ij} - \widehat{\pi}_{ij}|}{\frac{1}{n}\sum_{i=1}^{n} \pi_{ij}},$$

where $\widehat{\pi}_{ij}$ is the predicted proportion of theme j in pixel i. We also considered the most simple model, named M_0, as a benchmark to indicate if it is worth building sophisticated statistical models. It consists in predicting the land use of one crop by its empirical mean in the learning sample. This is a particular case of the multilogit model with no covariates.

We noticed (see Table 4.1) that the functional multilogit model, even if it can appear to be less natural since it has no direct physical interpretation, gave generally better predictions than other more intuitive methods (see Cardot, Faivre and Goulard 2003).

Themes	NDVI	PVI	Blue	Red	NIR	SWIR	M_0
urban	0.49	**0.36**	0.47	0.54	0.41	0.51	0.86
water	0.43	**0.29**	0.78	0.62	0.61	0.31	1.30
rapeseed	0.48	0.46	**0.45**	0.50	0.47	0.47	0.59
winter crops	0.20	0.21	**0.19**	0.20	0.22	**0.19**	0.30
spring crops	0.58	**0.56**	0.60	0.61	0.65	0.61	0.69
peas	0.50	**0.43**	0.45	**0.43**	0.48	0.46	0.63
summer crops	0.61	0.68	0.61	0.60	0.76	**0.53**	0.88
permanent crops	0.47	**0.46**	0.52	0.49	**0.46**	0.50	0.61
forest	0.34	0.36	0.34	**0.31**	0.45	0.35	0.98
potatoes	0.90	0.93	0.94	0.90	1.06	**0.85**	1.31

Table 4.1: Median value of the criterion error when predicting land use in the test sample with the GLM approach. Bold face numbers correspond to the best predictions. Model M_0 is used as a benchmark.

The best predictions seem to be obtained when using the PVI index. For instance, the errors were reduced of about 60 % compared to the reference model M_0. Thus combinations of the original wavelengths may be more appropriate to predict the land use and our future work will deal with finding optimal combinations of the available canals.

4.5 Functional Linear Regression on Quantiles

4.5.1 Definition of the Model and Spline Estimators

Let $\beta \in]0,1[$ and $x \in H$; the conditional quantile $g_\beta(x)$ of Y given $X = x$ is defined as

$$P_x\left(Y \leq g_\beta(x)\right) \;=\; \beta, \tag{4.5.1}$$

where P_x is the conditional distribution of Y given $X = x$. Alternatively (see Koenker and Bassett, 1978), it is defined as solution of

$$g_\beta(x) \;=\; \underset{a \in \mathbb{R}}{argmin} \;\; E(l_\beta(Y_i - a)|X_i = x), \tag{4.5.2}$$

with

$$l_\beta(u) \;=\; |u| + (2\beta - 1)u.$$

The reader is referred to Poiraud-Casanova and Thomas-Agnan (1998) for a review of nonparametric estimators of conditional quantiles.

Assuming now that the functional g_β is linear and continuous, Cardot, Crambes and Sarda (2003) get the model

$$Y = < \alpha, X > + \epsilon, \tag{4.5.3}$$

where ϵ is a random variable such that $P(\epsilon \leq 0) = \beta$. This variable is also supposed to be independent with X. Identifiability of the model is a consequence of condition 3

Condition 3. The distribution of ϵ is supposed to have density f_ϵ satisfying

$$f_\epsilon(0) > 0.$$

Cardot, Crambes and Sarda (2003) propose to estimate α with the following spline estimator

$$\widehat{\alpha}_{PS} = \sum_{\ell=1}^{k+q} \widehat{\theta}_\ell B_\ell = {}^t\mathbf{B}_{k,q}\widehat{\boldsymbol{\theta}},$$

with $\widehat{\boldsymbol{\theta}}$ solution of the minimization problem

$$\min \left\{ \frac{1}{n} \sum_{i=1}^{n} l_\beta(Y_i - <{}^t \mathbf{B}_{k,q}\boldsymbol{\theta}, X_i>) + \rho \left\| ({}^t\mathbf{B}_{k,q}\boldsymbol{\theta})^{(m)} \right\|^2 \;\middle|\; \boldsymbol{\theta} \in \mathbb{R}^{k+q} \right\}$$

This is a L_1 type optimization problem and it is not possible to exhibit any explicit solution. A numerical approximation is needed to get solutions (see Lejeune and Sarda, 1988 and Ruppert and Carroll, 1988).

4.5.2 Application to Ozone Prediction

The data come from the ORAMIP, the air quality observatory for the Midi-Pyrénées area in south west of France. They are collected around the city of Toulouse and we have hourly measurements of ozone concentration (O_3), NO concentration (NO) and NO_2 concentration (NO_2), wind speed (WS), and wind direction (WD), during the summers (15th May - 15th September) of the years 1997-2000. We aim at forecasting daily maximum O_3 concentrations with the help of the functional covariates measured the day before until 5 pm.

Data consist in a sample of 474 days during the four summers 1997-2000, 22 days are missing days because of technical reasons. Unfortunately important variables such as the temperature or the nebulosity are not available yet. The discrete trajectories were approximated with the help of splines functions in order to get curves. Even if it was not the case in this study, this allows

to deal with time measurements that differ from one day to another. The response is defined by

$$Y_i = \max_{t\in[0,24]} O_{3,i}(t),$$

and it corresponds to the maximum of the B-spline approximation of O_3 concentration during the i^{th} day. There are 5 explanatory variables:

$$\begin{aligned} \mathbf{X}_i(t) &= (X_i^1(t),\ldots,X_i^5(t)) \\ &= (O_{3,i-1}(t), NO_{i-1}(t), NO_{2,i-1}(t), WS_{i-1}(t), WD_{i-1}(t)) \end{aligned} \tag{4.5.4}$$

The latter vector is precisely the 5-dimensional functional vector of the B-splines approximations of the covariates $O_{3,i-1}(t),\ldots,WD_{i-1}(t)$, where time $t \in D$ varies from $t = 6$ *p.m.* of day $i-2$ to $t = 5$ *p.m.* of day $i-1$.

We first consider the general following additive model:

$$P(Y_i \leq c + g_1(X_i^1) + \ldots + g_5(X_i^5)/X_i^1 = x_i^1,\ldots,X_i^5 = x_i^5) = \beta \tag{4.5.5}$$

where $g_1,\ldots,g_v$ are continuous linear operator mapping $L^2(D)$ to $\mathbb{R}$. Equivalently we have

$$\begin{cases} Y_i = c + \displaystyle\int_D \alpha_1(t)X_i^1(t)\,dt + \ldots + \int_D \alpha_5(t)X_i^5(t)\,dt + \epsilon_i \\ \\ P(\epsilon_i \leq 0/X_i^1 = x_i^1,\ldots,X_i^5 = x_i^5) = \alpha \end{cases}$$

where functions $\alpha_1,\ldots,\alpha_5$ are supposed to be twice continuously differentiable. We have expanded the estimators into a B-splines basis of order $q = 3$ with $k = 8$ knots and we considered a penalty proportional to the norm of the second derivative, $m = 2$. A backfitting algorithm is used to estimate iteratively each functional component of the model. For identifiability reasons we have centred the vectors $\mathbf{X}_i$. We noticed that the algorithm converges quite rapidly.

Because of the heterogeneity of the four summers under study we did not take, as it is usually done, the last year as a validation year. The data were split randomly into a test sample, say I_T, composed of $n_T = 142$ observations and a learning sample, say I_L, with $n_L = 332$ observations. The training sample is used to select and estimate the parameters of the models and the test sample is used to compare the predictions of the different models.

To evaluate the performance of a model/estimator we considered different criterions. Let us denote by $\widehat{Y}_i$ the quantile prediction for observation i. The first criterion is a $L1$ distance

$$C_1 = \frac{1}{n_T}\sum_{i=1}^{n_T} \mid Y_i - \widehat{Y}_i \mid \tag{4.5.6}$$

Covariates	C_1	C_2
NO	16,998	0,911
NO_2	16,800	0,900
O_3	**12,332**	**0,661**
WD	16,836	0,902
WS	18,222	0,976
O_3, NO	12,007	0,643
O_3, NO_2	11,936	0,640
O_3, WD	12,109	0,649
O_3, WS	11,823	0,633
O_3, NO, NO_2	11,935	0,639
O_3, NO, WD	12,024	0,644
O_3, NO, WS	11,832	0,634
O_3, WD, WS	11,976	0,642
O_3, NO, WD, WS	11,954	0,641
O_3, NO, NO_2, WD	11,921	0,639
O_3, NO, NO_2, WS	**11,712**	**0,628**
O_3, $N0_2$, WD, WS	11,952	0,640
O_3, NO, NO_2, WD, WS	11,978	0,642

Table 4.2: Values of the criterion C_1 and C_2 on the pollution data for different models.

and the second one is the ratio

$$C_2 = \frac{\frac{1}{n_T}\sum_{i=1}^{n_T} l_\beta(Y_i - \widehat{Y}_i)}{\frac{1}{n_T}\sum_{i=1}^{n_T} l_\beta(Y_i - q_\beta(Y_L))} \tag{4.5.7}$$

where $q_\beta(Y_L)$ is the empirical quantile of order β in the learning sample. These criterions take small values for "good" prediction whereas values larger than one indicate, for criterion C_2 that the use of a model deteriorates the prediction accuracy compared to $q_\beta(Y_L)$.

The results, for models with different numbers of covariates, are gathered in Table (4.2) for the prediction of the median ($\beta = 0.5$). When considering only one functional covariate, it is clear that the best prediction is obtained when using the O_3 variable of the previous day. The gain, in terms of C_2, is about 34 % compared to the empirical quantile. Furthermore, the use of the four covariates O_3, NO, NO_2 and WS allows to improve the predictions.

4.6 Asymptotic Results

Asymptotic results for the spline estimators in models (4.3.1), (4.4.2) and (4.5.3) have been studied in terms of the asymptotic behavior of the L^2 norm in H with respect to the distribution of X defined as

$$\|\varphi\|_2^2 = <\Gamma\varphi, \varphi>, \ \varphi \in H.$$

To get existence, unicity and L^2 convergence of $\widehat{\alpha}_{PS}$, one needs, for all previous models, the following assumptions

(H.1) $\|X\| \leq C_1 < \infty$, a.s.

(H.2) The eigenvalues of Γ are strictly positive.

The functional coefficient α is supposed to have p' derivatives for some integer p' with $\alpha^{(p')}$ satisfying

(H.3) $|\alpha^{(p')}(y_1) - \alpha^{(p')}(y_2)| \leq C_4|y_1 - y_2|^{\nu}$, $C_4 > 0$, $\nu \in [0, 1]$.

In the following, we note $p = p' + \nu$ and assume that the degree q of splines is such that $q \geq p$.

Under assumptions (H.1)-(H.3) and if $\rho \sim n^{-(1-\delta_0)/2}$ for some $0 < \delta_0 < 1$, and $\rho k^{2(m-p)} = o(1)$, one gets the following result

A unique solution $\widehat{\alpha}_{PS}$ exists except on an event whose probability goes to zero as $n \to \infty$.

For models (4.3.1) and (4.5.3) and under additional condition in model (4.3.1) that the conditional expectation and the conditional variance are bounded for all $x \in H$, for $k \sim n^{1/(4p+1)}$ and $\rho \sim n^{-2p/(4p+1)}$ one gets, for $m \leq p$,

$$E\left(\|\widehat{\alpha}_{PS} - \alpha\|_2^2 | X_1, \ldots, X_n\right) = O_P(n^{-2p/(4p+1)}). \tag{4.6.1}$$

We refer to Cardot, Ferraty and Sarda (2003) and to Cardot, Crambes and Sarda (2003) for the proofs and comments of this result for model (4.3.1) and (4.5.3) respectively.

For model (4.4.2), Cardot and Sarda (2003) get for $k \sim n^{1/(2p+1)}$, $\rho \sim n^{-(1-\delta)/2}$ and for $m \leq p$

$$E\left(\|\widehat{\alpha}_{PS} - \alpha\|_2^2 | X_1, \ldots, X_n\right) = O_P(n^{-2p/(2p+1)}) + O(\rho). \tag{4.6.2}$$

Bibliography

de Boor, C. (1978) *A practical guide to Spline.* Springer, New York.

Cardot, H., Crambes, C. and Sarda, P. (1999) Spline estimator of conditional quantiles for functional covariates (in French) *Preprint.*

Cardot, H., Faivre, R. and M. Goulard (2003) Functional Approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, **30**, 1185-1199.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional Linear Model. *Statist. & Prob. Letters*, **45**, 11-22.

Cardot, H., Ferraty, F. and P. Sarda (2003) Spline Estimators for the Functional Linear Model. *Statistica Sinica*, **13**, 571-591.

Cardot, H. and Sarda, P. (2003) Estimation in Generalized Linear Models for Functional Data via Penalized Likelihood. *Journal of Multivariate Analysis*, to appear.

Dauxois, J. and Pousse, A. (1976) Les analyses factorielles en calcul des probabilités et en statistique. Essai d'étude synthétique (in French) Thèse, Université Paul Sabatier, Toulouse, France.

Dauxois, J., Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a random vector function: some applications to statistical inference. *Journal of Mult. Analysis*, **12**, 136-154.

Frank, I.E. and Friedman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-148.

Hastie, T.J. and Mallows, C., (1993) A discussion of "A statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman. *Technometrics*, **35**, 140-143.

Koenker, R. and Bassett, G. (1978) Regression Quantiles. *Econometrica*, **46**, 33-50.

Kress, R. (1989) *Linear Integral Equations*, Springer Verlag, New York.

Lejeune, M. and Sarda, P. (1988) Quantile Regression: A Nonparametric Approch. *Computational Statistics and Data Analysis*, **6**, 229-239.

Marx, B.D. and Eilers P.H. (1996) Generalized Linear Regression on Sampled Signals with penalized likelihood. In: Forcina, A., Marchetti, G.M., Hatzinger, R., Galmacci, G. (Eds), Statistical Modelling, Proceedings of the Eleventh International Workshop on Statistical Modelling, Orvietto.

Marx, B.D. and Eilers P.H. (1999) Generalized Linear Regression on Sampled Signals and Curves: A *P*-Spline Approach. *Technometrics*, **41**, 1-13.

Meynard, J.M. (1997) Which crop models for decision support in crop management: example of the Déciblé system. *Quantitative Approaches in System Analysis*, **15**, 107-112.

Osborne, B. G., Fearn, T., Miller, A. R. and Douglas, S. (1984) Application of near infrared reflectance spectroscopy to the compositinla analysis of biscuits and biscuit dough. *J. Sci. Food Agriculture*, **35** 99-105.

Poiraud-Casanova, S. et Thomas-Agnan, C. (1998) Quantiles Conditionnels. *Journal de la Société Française de Statistique*, **139**, 31-44.

Ramsay, J.O. and Silverman, B.W. (1997) *Functional Data Analysis.* Springer-Verlag.

Ramsay, J.O. and Silverman, B.W. (2002) *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag.

Ruppert, D. and Caroll, J. (1988) *Transformation and Weighting in Regression.* Chapman and Hall.

Stone, C.J. (1986) The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14**, 590-606.

Tucker, C.J. (1979) Red and Photographic Infrared Linear Combinations for Monitoring Vegetation. *Remote Sensing of Environment*, **8**, 127-150.

5 Penalized Binary Regression as Statistical Learning Tool for Microarray Analysis[1]

Michael G. Schimek

Medical University of Graz, Institute for Medical Informatics, Statistics and Documentation, A-8036 Graz, Austria

Summary

Statistical learning is an essential strategy in the analysis of microarray data. A typical task we are discussing here is the classification of biological samples into two alternative categories. We prefer a procedure that, based on the expression levels measured, allows us to compute the probability that a new sample belongs to a certain class. This is in contrast to other learning approaches like support vector machines. An approach providing us with probability statements and not just a classification rule is binary regression. High-dimensionality and at the same time small sample sizes are the challenge. Standard logit or probit regression fails because of condition problems and poor predictive performance. Binary regression based on penalized log-likelihood is considered here instead. Further the role of cross-validation for regularization and feature selection is discussed. Finally we illustrate penalized logit regression in the context of statistical learning on a well-known gene expression data set.

Keywords: Classification, cross-validation, logit regression, microarray analysis, penalization, prediction, probit regression, statistical learning

5.1 Introduction

Typical for microarray gene expression analysis (for an introduction see e.g. (McLachlan, Do, & Ambroise 2004)) are the following characteristics: (i) The number of observations (samples) is smaller than the number of input variables (genes). Hence more parameters have to be estimated than esti-

[1]The author wishes to thank Dennis Kostka and Rainer Spang (Max-Planck-Institute for Molecular Genetics, Berlin) for valuable discussions about statistical learning.

mation equations are available, resulting in non-unique solutions. (ii) The bias-variance dilemma where we have to decide between a model of low complexity with high bias and low variance, and a model of high complexity with low bias and high variance. Statistical learning provides a framework for handling these problems (see e.g. (Hastie, Tibshirani, & Friedman 2001)). It is assumed that the input vector x and the outcome y are generated by sampling from an unknown underlying distribution P. The objective of learning is then to find that predictor function f which minimizes the expectation of the loss function $\mathcal{L}(y, f)$. Learning f is based on data and is in the simplest case a two-step procedure applying a training set and a validation set. Training data are used to infer a function f^* that is close to the target function f. By the approximation of a minimization problem, f^* is chosen as the minimizing function of an empirical risk criterion. Since a function f^* which fits well to the training data can still be distant from the target function f, restrictions are required for the set F^* of possible solutions f^*. Statistical learning enforces such restrictions on F^* in dependence of the complexity of F^*. Well-known statistical learning concepts are kernel-based techniques like support vector machines (for a comparison with penalized logistic regression see (Zhu & Hastie 2004)).

A typical statistical learning task is the classification of biological samples into two alternative categories. Based on registered expression levels the goal is to compute the probability that a new sample belongs to a certain class. This can be achieved by binary regression. Both logit and probit regression are candidates. As pointed out above there are far more parameters (say m) than samples (say n). Because of $m \gg n$ the necessary reduction in dimension can be achieved by penalization. To avoid overfitting and poor prediction we impose a penalty on large fluctuations of the estimated parameters. Quadratic regularization is know as ridge regression for continuous responses. It can also be applied to binary responses. The n samples are represented by points in an m-dimensional space. In practice, however, there are only n relevant points which all lie in a linear subspace of maximum dimension n. For the projection onto the subspace singular value decomposition has been proposed (e.g. (Eilers et al. 2001) for penalized logit regression). When penalizing, a ridge parameter needs to be chosen.

In this paper two relevant binary regression models, logit (logistic) and probit regression are described. Both can be characterized by their log-likelihood. For gene expression data direct log-likelihood estimation is inadequate. However we can show that penalization of the log-likelihood is possible in both instances. Given a ridge parameter representing an adequate bias-variance tradeoff, stable estimates can be obtained. Finally we illustrate our approach applying penalized logit regression to a well-known gene expression data set from (West et al. 2001).

5.2 Binary Regression

Binary regression is well established in the context of generalized linear models (GLM) as described in (McCullagh & Nelder 1989). In biostatistics it is a suitable tool for dose response modelling, going back as far as 1973 (Finney 1973). Nowadays it is increasingly becoming an important linear classification method (for an overview see (Hastie, Tibshirani, & Friedman 2001)).

In a GLM a binary response can be predicted by a so-called predictor function consisting of a transformed (depending on the nature of the response variable) linear combination of explanatory variables. As stated in the introduction, binary regression would be an appropriate tool to classify biological samples belonging to one of two alternative classes.

The typical situation is that a number of biological samples has been collected, preprocessed and hybridized to microarrays. It is assumed that n_1 microarrays belong to the one and n_2 microarrays belong to the other class (e.g. to patients having the disease versus not having the disease) and $n_1 + n_2 = n$. The statistical task is to compute the probability that a specific sample belongs to one of the two alternatives based on the expression levels recorded. The final goal is the classification of new microarray data.

Why is the application of GLMs not satisfying in this setting? A basic assumption, typical for standard regression and also classification techniques, is not fulfilled: $m \ll n$, where m denotes the predictors (i.e. genes in the above example). On the contrary, here we always have $m \gg n$. For estimation the consequence is that there are many more unknowns than equations and infinitely many solutions exist! Hence we end up with an ill-conditioned problem. A naive application of binary regression, whatever the model or the algorithm, would result in unstable, non-unique estimates.

Another problem in the context of classification is the low discrimination power due to the high variance of the data fit.

How can we cope with these detrimental features? Binary regression models can be characterized by their log-likelihood (see later). A likelihood can always be penalized to obey certain conditions. Let us start with the description of binary regression in the GLM context.

Suppose y_i, $i = 1, \ldots, n$, a binary response variable and x_{ij}, $j = 1, \ldots, m$, continuous predictor variables. To investigate the relationship between the predictor variables and the response probability p we assume a linear combination

$$\eta_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij},$$

where the β_j's are unknown coefficients, and β_0 is the offset. Let p_i be the probability of observing $y_i = 1$. The connection between η_i and p_i for the Binomial family of the GLMs is the canonical link which is non-linear. Unless restrictions are imposed on the β's we have $-\infty < \eta_i < \infty$. Hence a transformation, also called link function, $g(p)$ is introduced that maps the unit interval onto the whole real line, i.e. $g(p_i) = \eta_i$. Suitable transformations for the purpose of gene profiling are the logit (logistic) link function

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

and the probit (inverse Normal) link function

$$g(p_i) = \Phi^{-1}(p_i),$$

where $\Phi(\bullet)$ is the cumulative distribution function of the standard Normal distribution. The transformations ensure that the probabilities of the two classes (0/1 response y_i) sum to one and remain in $[0, 1]$. In logit regression we have

$$p(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)},$$

(logistic curve) and in probit regression $p(\eta) = \Phi(\eta)$. Assuming a Binomial error distribution in both instances the log-likelihood can be written

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^{n}\{y_i \log p(x_i;\beta) + (1-y_i)\log(1-p(x_i;\beta))\} \\ &= \sum_{i=1}^{n}\{y_i\beta^T x_i + \log(1+\exp(\beta^T x_i))\}. \qquad (5.2.1)\end{aligned}$$

Here we have assumed $\beta = \{\beta_0, \beta_j\}$ for $j = 1, \ldots, m$, and that the vector of predictor values x_i includes the constant term one to accommodate the offset.

To maximize the log-likelihood, we set its derivatives to zero. The score equations are

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i(y_i - p(x_i;\beta)) = 0.$$

Their number is $m+1$ and they are non-linear in β. Since the first element of x_i is one, the first score equation specifies that

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} p(x_i;\beta).$$

This is to say that the mean of p has to be equal to the fraction of ones in y.

For logit as well as probit regression the score equations can be solved via the Newton-Raphson algorithm (iteratively reweighted least squares in the GLM framework). Typically the algorithm converges since the log-likelihood is concave (for details see (Hastie, Tibshirani, & Friedman 2001), p.98f).

5.3 Penalized Binary Regression

As pointed out already, binary regression cannot be immediately applied in a situation of $m \gg n$. A remedy are the so-called shrinkage methods. The key idea is that overfitting is avoided by imposing a penalty on large fluctuations of the estimated parameters. In 1970 Hoerl and Kennard (Hoerl & Kennard 1970) introduced ridge regression for ill-conditioned ordinary regression problems. It shrinks the regression coefficients by imposing a penalty on their size. A complexity parameter λ ($\lambda \geq 0$) controls the amount of shrinkage. The regularization in (Hoerl & Kennard 1970) is quadratic, i.e. based on the sum of squares of the regression coefficients. This ridge penalty has later been applied to linear discriminant analysis (Friedman 1989), logistic regression (le Cessie & van Houwelingen 1992), and neural networks (Girosi, Jones, & Poggio 1995). Regularization is closely related to certain spline concepts in nonparametric regression (Schimek 1988). In smoothing splines we are talking about penalty functions controlled by a smoothing parameter similar to the above complexity parameter.

The general motivation behind penalizing the likelihood is the following: Avoid arbitrary coefficient estimates $\hat{\beta}$ and a classification that appears to be perfect in the training set but is poor in the validation set. Penalization aims at an improved predictive performance in a new data set by balancing the fit to the data and the stability of the estimates.

Suppose $\beta^* = \{\beta_j\}$ for $j = 1, \ldots, m$, the log-likelihood $\ell(\beta)$ can be penalized in the following way:

$$\ell^*(\beta) = \ell(\beta) - \frac{\lambda}{2} J(\beta^*) \tag{5.3.1}$$

where

$$J(\beta^*) = \|\beta^*\|^2 = \sum_{j=1}^{m} \beta_j^2$$

is a quadratic (ridge) penalty. As can be seen, only the regression coefficients β_j are subject to penalization, not the offset β_0. The complexity (ridge) parameter λ controls the size of the coefficients (increasing λ values decrease their size). There are also other penalties known in the literature (e.g. (Donoho & Johnstone 1994), (Tibshirani 1995)) we cannot discuss here.

To maximize the penalized log-likelihood, we set its derivatives to zero,

$$\frac{\partial \ell^*(\beta)}{\partial \beta_0} = 0,$$

$$\frac{\partial \ell^*(\beta)}{\partial \beta_j} = 0.$$

We obtain the penalized likelihood equations

$$u^T(y - p) = 0$$

and

$$X^T(y - p) = \lambda\beta.$$

The m-dimensional vector u consists of ones. As in classical binary regression the equations are non-linear. A first-order Taylor expansion gives (we abbreviate $p_i = p(x_i; \beta)$)

$$p_i = \tilde{p_i} + \frac{\partial p_i}{\partial \beta_0}(\beta_0 - \tilde{\beta_0}) + \sum_{j=1}^{m} \frac{\partial p_i}{\partial \beta_j}(\beta_j - \tilde{\beta_j}).$$

Tilde denotes the approximation (e.g. $\tilde{p_i}$ for p_i). The partial derivatives are

$$\frac{\partial p_i}{\partial \beta_0} = p_i(1 - p_i)$$

and

$$\frac{\partial p_i}{\partial \beta_j} = p_i(1 - p_i)x_{ij}.$$

In writing $p_i(1 - p_i) = w_i$ we finally have

$$u^T\tilde{W}u\beta_0 + u^T\tilde{W}X\beta = u^T(y - \tilde{p} - \tilde{W}\tilde{\eta}) \tag{5.3.2}$$

and

$$X^T\tilde{W}u\beta_0 + (X^T\tilde{W}X + \lambda I)\beta = X^T(y - \tilde{p} - \tilde{W}\tilde{\eta}), \tag{5.3.3}$$

where W is a diagonal matrix consisting of the elements w_i. As before, the parameter β_0 is determined by the fraction of ones in y.

The above linearized system can be solved by means of iterative techniques. Under the requirements of gene expression analysis the system of equations is huge. In the statistical environment R (Hornik et al. 2004) binary regression can be fitted by means of singular value decomposition, an expensive but reliable numerical approach. In the contributed R package `Design` there are functions for penalized logit regression.

5.4 Cross-Validation

In the previous sections we have seen how to cope with the severe problem of far more parameters than samples. Regression approaches are prone to overfitting (being to optimistic in the training set) which leads to poor predictions (in the validation set). Hence the control of the model complexity is crucial.

In binary regression models Cross-Validation (CV) is the standard technique both for the choice of the regularization (ridge) parameter λ and for feature selection aiming at a stable model fit.

The simplest type of CV is leave-one-out: An observed value is predicted from the remaining observations and that parameter is chosen which yields the best prediction. Because the computational demand is proportional to $m(m-1)$, it is quite high when applied to all m genes in expression profiling. The generalization of leave-one-out CV is leave-k-out CV (computationally less demanding).

Outside statistics (e.g. in chemometrics) often the concept of *k-fold* CV is used. It means that the training set is split into k parts of approximately equal size. In each run one of the k parts is left out and used as an independent validation set for optimizing the parameters.

As far as penalized regression is concerned smaller sets require stronger regularization. This can lead to sub-optimal model fits. Another problem is the large number of possible divisions of the training set into k groups each of size $g = m/k$ (i.e. $m!/k!(g!)^k$). Different partitions may yield different performance assessments (e.g. in feature selection). These problems have been studied in great detail in (Jonathan, Krzanowski, & McCarthy 2000). For a thorough discussion of instabilities in feature (model) selection see (Breiman 2003). In practice one tries to have balanced groups and decides for parameters which work best across the runs. Criteria for the performance of CV are either the misclassification rate or the strength of prediction.

5.5 An Example

The example is based on the Breast Cancer Data Set of (West et al. 2001). The data originate from a gene expression study with patients suffering from breast cancer. There are two different response variables, status of the estrogen receptor (ER) and lymph nodal status (LN). Here we study the full data set with respect to ER. It consists of $n = 49$ samples, $n_1 = 25$ belonging to the ER+ (patient samples 1–15, 28–30, 32, 34–35, 45–48) and $n_2 = 24$ ER- (patient samples 16–27, 31, 33, 36–44, 49). We are not splitting the

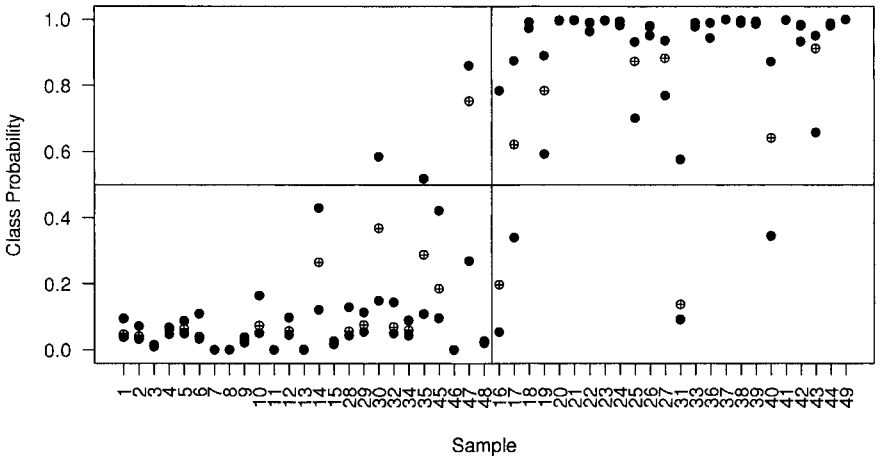


Figure 5.1: Penalized logistic regression regression results for two cross-validation schemes: predicted probability for each sample belonging to ER+ or ER-

data into a learning and a validation set. Instead we resort to pre-validation ideas (Tibshirani & Efron 2002) in connection with ten-fold CV applied to all available samples.

The data comprise the preprocessed expression of 7129 human genes (Affymetrix gene chip probes) and can be obtained from the Web location `http://compdiag.molgen.mpg.de/ngfn/data/2004/mar/`. Further we have a vector of tumor class labels with elements 0 for ER+ and 1 for ER-. In each evaluation step the 100 most informative genes (criterion t-statistic from Bioconductor's `multtest`; (Gentleman et al. 2004)) were chosen for model fitting. Obviously this set of genes can change from step to step.

The procedure for penalized logit regression was implemented in R, taking advantage of functions in the library `Design`. It is maximum likelihood-based.

The complexity (ridge) parameter $\hat{\lambda} = 5.29$ for the logit model was chosen by ten-fold CV (based on balanced groups) with respect to log-likelihood prediction. For the evaluation of the regression model outside loop feature selection was applied. In each step the most informative 100 hundred genes were used out of 7129.

In Fig.5.1 for each sample the logit class probability is plotted based (i) on simple ten-fold CV for the choice of the ridge parameter (crossed circles) and (ii) on a two-step (i.e. in two loops) ten-fold CV for additional feature selection (black circles) for improved prediction. The vertical line separates

the ER+ (left) from the ER- (right) samples according to external evidence. It can be clearly seen that the simple approach is overoptimistic. When feature selection is adopted we have missclassification for the samples 30, 35, 47, 16, 17, 31, and 40 and an overall missclassification rate of 14.29%. This is a typical value for penalized binary classification procedures in microarray data.

5.6 Conclusions

Penalized binary regression is a well-motivated statistical concept which has proven to be useful in classifying gene expression data. The approach described in this paper has, compared with kernel methods such as support vector machines from machine learning, the advantage of providing class probabilities, thus yielding a quantification of the specific contribution of each gene to prediction.

Bibliography

Breiman, L. (1996) Heuristics of instability and stabilization in model selection, *Annal. Statist.* **24**, 2350–83.

le Cessie, S. & van Houwelingen J.C. (1992) Ridge estimators in logistic regression, *Appl. Statist.* **41**, 191–201.

Donoho D, Johnstone I.(1994) Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**, 425–55.

Dudoit S, Fridlyand J, Speed TP.(2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *J Amer Statist Assoc* **97**, 77–87.

Eilers PHC et al.(1992) Classification of microarray data with penalized logistic regression, *Proceedings of SPIE* **4266**, 187–98.

Finney D.(1973) *Statistical Method in Biological Assay*, New York: Hafner, (2nd edition).

Friedman JH.(1989) Regularized discriminant analysis, *J Amer Statist Assoc* **84**, 165–75.

Gentleman R et al.(2004) The Bioconductor FAQ.
`http://www.bioconductor.org/`

Girosi F, Jones M, & Poggio T.(1992) Regularization theory and neural networks architecture, *Neural Computation* **7**, 219–69.

Hastie T, Tibshirani R, & Friedman J.(2001) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, New York: Springer-Verlag.

Hoerl AE, & Kennard RW.(1970) Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12**, 55–67.

Hornik K et al.(2004) The R FAQ. `http://www.r-project.org/`

van Houwelingen JC, & le Cessie S.(1990) Predictive value of statistical models, *Statistics in Medicine* **9**, 1303–25.

Jonathan P, Krzanowski WJ, & McCarthy WV.(2000) On the use of cross-validation to assess performance in multivariate prediction, *Statist and Comput* **10**, 209–29.

McCullagh P, & Nelder JA.(1989) *Generalized Linear Models*, London: Chapman & Hall, (2nd edition).

McLachlan, GJ, Do, K-A, & Ambroise, C.(2004), *Analyzing microarray gene expression data*, New York: Wiley.

Park P, Pagano M, & Bonetti M.A (2001) nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac Symp Biocomput* **6**, 52–63.

Schimek MG.(1992) A roughness penalty approach for statistical graphics. In Edwards D, Raun NE eds. *Proceedings in Computational Statistics*, Heidelberg: Physica-Verlag, 37–43.

Tibshirani R.(1995) Regression shrinkage and selection via the lasso, *J Royal Statist Soc B* **57**, 267–88.

Tibshirani R, & Efron B.(1992) Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology* **1**, article 1. `http://www.bepress.com/sagmb/vol1/iss1/art1`

West M et al.(2001) Predicting the clinical status of human breast cancer by using gene expression profiles, *Proc Nat Academy Scien* **98**, 11462–7.

Zhu, J, & Hastie, T.(2004) Classification of gene microarrays by penalized logistic regression, *Biostatistics* **5**, 427–443.

6 A Relaxed Iterative Projection Algorithm for Rank-Deficient Regression Problems[1]

Michael G. Schimek[2] and Haro Stettner[3]

[2] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, A-8036 Graz, Austria
[3] Department of Mathematics, Alpen-Adria-University Klagenfurt, A-9020 Klagenfurt, Austria

Summary

A relaxed iterative projection (abb. RIP) algorithm for arbitrary linear equation systems is described. It has favorable properties with respect to many statistical applications. A major advantage is that convergence can be established without restrictions on the system matrix. Hence certain characteristics of the system matrix such as diagonal dominance are not required. As a result RIP fitting can be applied where backfitting tends to fail, e.g. when regression predictors are substantially correlated (problem of multicollinearity respectively concurvity). Convergence under a suitable choice of the relaxation parameter is derived for general $n \times m$ system matrices. The RIP solution of typical equation systems is studied with respect to the correct (analytical) solution. Empirical findings for the practical selection of the relaxation parameter are reported.

Keywords: Backfitting, concurvity, linear equation systems, nonparametric regression, relaxed iterative projection fitting, relaxation parameter, semiparametric regression, successive over-relaxation, singularity

6.1 Introduction

For numerous statistical applications we have to solve large linear equation systems of the form $A\mathbf{x} = \mathbf{b}$ in $\mathbf{x}$, often arising in the context of parametric, nonparametric or semiparametric regression. In some instances we can

[1]The financial support of the FWF Austrian Science Fund in the research project "Konvergenz und Numerik des Backfitting-Algorithmus" is greatly acknowledged.

assume a square (i.e. $n \times n$) system matrix A and n-dimensional vectors $\mathbf{x}$ and $\mathbf{b}$. However there are many instances, where the matrix A is rectangular (i.e. $n \times m$), $m < n$, $\mathbf{x}$ is m- and $\mathbf{b}$ is n-dimensional. The number of equations is not necessarily equal to the number of regressors. $m < n$ and $rank([A \mid \mathbf{b}]) > rank(A)$ is typical for systems leading to normal equations in least squares problems. Such systems are said to be overdetermined. There is no $\mathbf{x}$ that satisfies such a system, but approximate solutions can be obtained and are useful. A system for which $rank([A \mid \mathbf{b}]) = rank(A)$ is said to be consistent and solutions exist. For an overview of numerical methods for least squares problems see Bjorck (1996).

In this paper we consider overdetermined consistent systems (underdetermined systems are not of interest here). Further we assume a general $n \times m$ equation system with $m \leq n$. The dimension of n is usually much higher than that of m (under the assumption that there are substantially more observations than predictor variables resulting in an overdetermined system). In general A is assumed to be a full matrix. In certain cases A may be sparse, e.g. in non- and semiparametric regression problems depending on the smoothing technique.

Outside the parametric world in which direct numerical techniques dominate (Choleski for square and QR factorization for rectangular systems, see Gentle, 1998, p.93ff) iterative backfitting is the dominating algorithm (see e.g. Hastie and Tishirani (1990), p.90f for generalized additive models, and Green and Silverman (1994), p.68 for partial spline models). In this paper we propose a new iterative algorithm with improved convergence characteristics compared to backfitting, not requiring specific assumptions about the system matrix A. It is called relaxed iterative projection (abb. RIP) fitting and can be seen in the context of successive over-relaxation (abb. SOR). RIP has a simple geometric interpretation, is easier to implement and computationally more efficient than SOR.

6.2 Backfitting

Backfitting is a basic iterative method of Gauss-Seidel type. Given starting values x_0, such methods generate a sequence of x_k converging to the solution $A^{-1}\mathbf{b}$ of

$$A\mathbf{x} = \mathbf{b}, \tag{6.2.1}$$

where x_{k+1} is cheap to compute from x_k. An alternative to Gauss-Seidel iterations are Jacobi iterations (see Schimek and Turlach, 2000, p. 288), though hardly used in statistical computing. Both methods are equation oriented and have been derived for solving square ($m = n$) non-singular linear systems. The main difference between the two iteration types is that

at the ith step of the loop improved values of the first $i-1$ components of the solution are used in Gauss-Seidel which means a faster information update.

Gauss-Seidel backfitting is defined in the $(k+1)$th step for $i = 1, 2, \ldots, n$ by

$$x_{k+1,i} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_{k+1,j} - \sum_{j=i+1}^{n} a_{ij} x_{k,j} \right]$$

with iterative solutions $x_{k,i}$ and starting values $x_{0,i} = 0$.

Statistically speaking the idea is to determine estimates for the regression covariates in a successive manner, taking advantage of specific features of A (a question of ordering of the grid points we loop through). This allows us to be most effective for the choice of certain smoothers (e.g. cubic smoothing splines or regression splines in non- and semiparametric problems), because we can avoid the explicit calculation of all smoother matrices which are associated with the covariates.

Convergence depends on the eigenvalues of the iteration matrix. If A is strictly row diagonally dominant, then backfitting (holds true for Gauss-Seidel and Jacobi iterations) always converges. Strictly row diagonal dominance means that each diagonal entry of A is larger than the sum of the magnitudes of the other entries in its row. This is usually not satisfied in regression problems. A weaker form of diagonal dominance is irreducibility which concerns the pattern of non-zero entries of A. For a proof of convergence under this condition see Demmel (1997, p.286f).

A different variant of backfitting is the so-called modified backfitting algorithm introduced by Buja, Hastie and Tibshirani (1989). The solution is obtained in a subspace of the vector space constituting A. As a result there is a gain in efficiency. Because of $m < n$ the Gauss-Seidel iterations are replaced by QR decomposition. Modified backfitting can be applied in (generalized) additive regression models for specific smoothers allowing for a decomposition into a projection part and a shrinking part. In S-Plus this is the default algorithm because non-decomposable smoothers are not recommended. Yet standard backfitting is most general as it can be adopted for any kind of linear smoother.

Our main criticism of backfitting for the solution of linear equations is the fact that certain characteristics of the system matrix are required but not always met. Features like diagonal dominance or regularity cannot be taken for granted in all statistical estimation problems of interest. The algorithm has been shown to converge only in special cases of additive regression models (Opsomer and Ruppert, 1997). When generalized additive models are evaluated by means of kernel smoothers or local polynomial smoothers, ill-posed normal equations cannot be ruled out because of the weighting scheme

imposed on the data. In addition, from our experience we can say that a correlation larger than |0.3| between two predictors is sufficient to hamper convergence of backfitting (for more information on rank deficiency see Wood, 2004). This problem is known as multicollinearity in parametric models and concurvity in nonparametric models (for a detailed discussion see Ramsay, Burnett and Krewski, 2003). Despite the fact that generalized additive models were not designed for time series data, they are often applied to them (e.g. Dominici et al., 2002). Apart from (auto)correlation certain features of the data such as sparseness can also cause poor convergence.

The above mentioned disadvantages can be overcome by a non-standard iterative method which is column-oriented instead of equation-oriented. Its convergence can be established independent of specific features of the system matrix. Hence we can deal with many applications in which backfitting shows slow or no convergence (breaks down).

6.3 Relaxed Iterative Projection Fitting

In contrast to standard iterative techniques for linear systems which are equation-oriented, the relaxed iterative projection method is column-oriented. This idea has been first brought up by de la Garza (1951). It is only recent that projection concepts have become numerically feasible due to modern computer power. Schimek (1996) has introduced iterative projection fitting for a square system matrix and a relaxation concept to improve the computational efficiency of the therein proposed algorithm.

Here we derive the most general case assuming an arbitrary rectangular ($m \leq n$) matrix A consisting of m nontrivial column vectors of dimension n. Let us denote by $col(A)$ the linear space generated by $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ and by $\mathbf{x}$ a solution vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$. We assume

$$\mathbf{b} \in col(A). \tag{6.3.1}$$

The assumption of nontriviality of the columns of A for practical purposes means no restriction. For a trivial column $\mathbf{a}_i = 0$ the corresponding component x_i of each solution vector can be chosen arbitrarily. Likewise we could add any number of zero columns to A without influencing the "essential part" of the solution vector, i.e. these components of $\mathbf{x}$ which correspond to nontrivial columns.

We define two real sequences, the one $\{\mu_j\}$ is

$$\left(\mathbf{b} - \sum_{i=1}^{j} \mu_i \mathbf{a}_i, \mathbf{a}_j\right) = 0, \qquad j = 1, 2 \ldots,$$

where $\mathbf{a}_i = \mathbf{a}_{i+lm}$ (resp. $\mathbf{a}_j = \mathbf{a}_{j+lm}$) with $l = 0, 1, \ldots$ are the column vectors of A as defined above. But we could also consider $\mathbf{a}_i := \mathbf{a}_p{}^*$ with $p = 1 + (i-1) \bmod m$ for $i > m$. The $\mathbf{a}_1{}^*, \mathbf{a}_2{}^*, \ldots, \mathbf{a}_m{}^*$ are a permutation of the $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$. According to Murty (1983, p.457) this can help to improve the speed of convergence.

The other sequence $\{s_{kl}\}$ is

$$s_{kl} = \sum_j \mu_j, \qquad j = k + lm, \qquad k = 1, 2, \ldots, m, \qquad l = 0, 1, \ldots .$$

In the terminology of Maess (1988, p.113ff) this numerical method produces an instationary iteration process which is geometrically motivated. In the jth iteration step μ_j is determined by the orthogonal projection of the previous residual component $\mathbf{u}_{j-1}$ onto the dimension $\mathbf{a}_j$ and the coefficients μ_j can be calculated by dot (inner) products. Hence

$$\mu_j = \frac{(\omega \mathbf{u}_{j-1}, \mathbf{a}_j)}{(\mathbf{a}_j, \mathbf{a}_j)},$$

where ω is a relaxation parameter with admissible values $0 < \omega < 2$ and

$$\mathbf{u}_{j-1} = \mathbf{b} - \sum_{i=1}^{j-1} \mu_i \mathbf{a}_i,$$

which makes the geometric interpretation clear. The parameter ω should influence the sequence of orthogonal projections (dot products) in a way that less iteration steps are necessary till convergence (Euclidean norm $\|\mathbf{u}_j\|$ less than some ϵ).

The norm (length) of μ is shrinking for admissible ω values (for the proof see later) and convergence can be expected. Because the above s_{kl} tend to the x_k for $l \to \infty$, each element x_k of the solution vector $\mathbf{x} = (x_1, x_2, \ldots, x_m)^T$ – assuming l sufficiently large – can be obtained by

$$x_k = \sum_j \mu_j, \qquad j = k + ml, \qquad k = 1, 2, \ldots, m, \qquad l = 0, 1, \ldots .$$

The necessary condition is that the residual components $\mathbf{u}_j$ tend to zero. To simplify notation (e.g. for $\mathbf{u}$ below) let us have *single prime* denoting the next step, and *double prime* denoting the step after the next. Admissible ω values can be derived from the expression

$$\|\mathbf{u}'\|^2 = \|\mathbf{u} - \frac{\omega(\mathbf{u}, \mathbf{a})}{\|\mathbf{a}\|^2}\mathbf{a}\|^2 = \|\mathbf{u}\|^2 - (2\omega - \omega^2)\frac{(\mathbf{u}, \mathbf{a})^2}{\|\mathbf{a}\|^2}, \tag{6.3.2}$$

requiring $0 < \omega < 2$ in the case of convergence.

It is possible to establish convergence for RIP fitting without restrictions on the system matrix A.

Theorem: If (2) and $0 < \omega < 2$ is fulfilled, then for j sufficiently large $\mathbf{u}_j \rightarrow 0$.

Proof: Let $M := max\|\mathbf{a}_i\|$. Let us now study the sequence of the residual components. Because of (3) the sequence $\{\|\mathbf{u}_j\|\}$ is monotone decreasing in weak sense. We have to show that this sequence tends towards zero to prove the theorem.

If $\{\|\mathbf{u}_j\|\} \rightarrow C \neq 0$, then we could choose a subsequence

$$\mathbf{u}_{n_k} \rightarrow \mathbf{u}_0$$

with $\|\mathbf{u}_0\| = C$ because of the subset $\{\mathbf{u} : \|\mathbf{u}\| \leq 2C\}$ being compact in R^n.

Consider the space $col(A)$. It not only includes $\mathbf{b}$ and all $\mathbf{a}_i$ but also all $\mathbf{u}_{n_k}$, thus $\mathbf{u}_0$ too. As a result the set of $\mathbf{a}_i$ for which $(\mathbf{u}_0, \mathbf{a}_i) \neq 0$ holds, cannot be empty. Let k be the number of $\mathbf{a}_i$ with $(\mathbf{u}_0, \mathbf{a}_i) = 0$ and

$$\epsilon := min\{\|(\mathbf{u}_0, \mathbf{a}_i)\| : (\mathbf{u}_0, \mathbf{a}_i) \neq 0\} > 0.$$

Let us suppose an arbitrary $\delta > 0$ and $\mathbf{u}$ a fixed residual in the subsequence $\{\mathbf{u}_{n_k}\}$ with $\|\mathbf{u} - \mathbf{u}_0\| < \delta$. Starting from $\mathbf{u}$ we continue the iterative projections. For those columns $\mathbf{a}$ for which $(\mathbf{u}_0, \mathbf{a}) = 0$ we have

$$\|\mathbf{u}' - \mathbf{u}_0\| = \|\mathbf{u} - \frac{\omega(\mathbf{u}, \mathbf{a}')}{\mathbf{a}'^2}\mathbf{a}' - \mathbf{u}_0\| \leq \|\mathbf{u} - \mathbf{u}_0\| + \omega\|\frac{(\mathbf{u} - \mathbf{u}_0, \mathbf{a}')}{\mathbf{a}'^2}\mathbf{a}'\| \leq \delta + 2\delta < 4\delta,$$

further

$$\|\mathbf{u}'' - \mathbf{u}_0\| \leq \|\mathbf{u}' - \mathbf{u}_0\| + \omega\|\frac{(\mathbf{u}' - \mathbf{u}_0, \mathbf{a}'')}{\mathbf{a}''^2}\mathbf{a}''\| \leq 4\delta + 2 \bullet 4\delta < 16\delta,$$

and finally

$$\|\mathbf{u}^{(k)} - \mathbf{u}_0\| < 2^{2k}\delta$$

unless the procedure has not yet encountered an $\mathbf{a}$ with $(\mathbf{a}, \mathbf{u}_0) \neq 0$. The next step in analogy to (3) is

$$\|\mathbf{u}^{(k+1)}\|^2 = \|\mathbf{u}^{(k)}\|^2 - (2\omega - \omega^2)\frac{(\mathbf{u}^{(k)}, \mathbf{a})^2}{\mathbf{a}^2}.$$

We can assess

$$\|\mathbf{u}^{(k)}\| \leq \|\mathbf{u}_0\| + 2^{2k}\delta = C + 2^{2k}\delta$$

and

$$\|\mathbf{u}^{(k)}, \mathbf{a}\| = \|(\mathbf{u}_0, \mathbf{a}) + (\mathbf{u}^{(k)} - \mathbf{u}_0, \mathbf{a})\| \geq \|(\mathbf{u}_0, \mathbf{a})\| - \|(\mathbf{u}^{(k)} - \mathbf{u}_0, \mathbf{a})\| \geq \epsilon - 2^{2k}\delta M,$$

together yielding

$$\|\mathbf{u}^{(k+1)}\|^2 \leq (C + 2^{2k}\delta)^2 - (2\omega - \omega^2)\frac{(\epsilon - 2^{2k}\delta M)^2}{M^2}. \qquad (6.3.3)$$

Here the values for ω, k, ϵ, C and M are fixed. If δ is chosen sufficiently small, the righthand side of equation (4) is always less than C^2. This contradiction proves the theorem.

Extending the above considerations one can show that during a full cycle of projections $\|\mathbf{u}\|$ is reduced at least by a factor $f < 1$ only depending on A.

We have seen that for $0 < \omega < 2$ the residual components $\mathbf{u}_j$ tend to zero without specific assumptions on A. This is the main advantage over backfitting as described earlier.

The relaxation parameter ω as defined above operates on the solution in the following way: $\omega = 1$ characterizes the unrelaxed solution; $1 < \omega < 2$ is the interval of values of practical relevance improving the fit and at the same time reducing the required number of iterations; and finally $0 < \omega < 1$ denoting values under which convergence is still obtained yet requiring an increased number of iterations compared to $\omega = 1$. Increasing degeneracy needs larger ω values. There is no mathematical argument how to select an optimal value, however empirical evidence gives some guidance (see later).

Further we would like to state the following lemma.

Lemma 1: If $rank([A \mid \mathbf{b}]) = rank(A)$ (consistency) and $\mathbf{b} \in col(A)$ it follows from the theorem that

$$s_{k,l} \rightarrow x_k$$

for $k = 1, 2, \ldots, m$ and $l \rightarrow \infty$, thus $x = (x_1, x_2, \ldots, x_m)^T$ solves equation (1) (compare with Schimek, 1996).

For $\mathbf{b} \in col(A)$ a unique solution is obtained. In the case of $rank([A \mid \mathbf{b}]) > rank(A)$ (overdetermination) we do not obtain a unique minimum norm solution. However this does not imply any restriction because here we are also interested in the solution of overdetermined statistical problems of (penalized) least squares type such as (nonparametric smoothing spline) regression.

Lemma 2: If $\mathbf{b} \notin col(A)$, i.e. $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2, \mathbf{b}_1 \in col(A), \mathbf{b}_2 \perp col(A)$ with $\mathbf{b}_2 \neq 0$, the result of the iteration process tends to a solution of $\|A\mathbf{x} - \mathbf{b}\| \rightarrow min$, i.e. $A\mathbf{x} = \mathbf{b}_1$.

This can be seen by applying the above theorem to $A\mathbf{x} = \mathbf{b}_1$. We notice that the coefficients μ_j are not influenced by $\mathbf{b}_2$, since they are linear functions of $(\mathbf{b}, \mathbf{a}_j) = (\mathbf{b}_1, \mathbf{a}_j)$.

In summary, the proposed iterative projection method can be characterized as follows: Firstly, it works for square as well as rectangular system matrices. This means that it can cover all the situations where backfitting is currently applied. Secondly, it always converges because convergence does not depend on special features of the system matrix, such as positive definiteness or diagonal dominance. There are numerous statistical computations for which such assumptions do not hold neither can be checked during execution (e.g. in data mining tasks). Thirdly, even for the special case of a singular or near singular system a solution can be obtained. This is also an important aspect because the condition of the matrix A can deteriorate during execution when backfitting or RIP fitting forms the core of a more complicated algorithm (e.g. in generalized additive models where local scoring takes place at the same time, see Schimek and Turlach, 2000). As already pointed out, there are various statistical problems which tend to produce rank-deficient systems.

Finally it should be noted, that the RIP procedure discussed above is indeed a modification of Gauss-Seidel iterations, but is using a different approach in the derivation of the equations to solve. So although there are numerical equivalencies, the projection-oriented view allows extended insights.

6.4 RIP Fitting and Successive Overrelaxation

As already mentioned there is a connection between RIP fitting and successive overrelaxation (SOR), although evaluated via different algorithms. Its superior numerical features (for algorithmic details see next section) are the advantage of the RIP procedure.

Assuming A has full column rank, let us decompose $A^TA = L + D + L^T$, where D is the diagonal matrix of A^TA and L is the lower triangular part of A^TA. Let $x_{l+1} = [\mu_{lm+1}, \mu_{lm+2}, \ldots, \mu_{(l+1)m}]^T$ for $l = 0, 1, 2, \ldots$ and let $x_0 = 0$. Then one can show that

$$(D + \omega L)x_{l+1} = \omega A^T(\mathbf{b} - Ax_0 - Ax_1 - \ldots - Ax_l) \tag{6.4.1}$$

for $l = 0, 1, 2, \ldots$. From equation (5) and $A^TA = L + D + L^T$ it follows that

$$x_1 + x_2 + \ldots + x_{l+1} =$$

$$(D + \omega L)^{-1}((1 - \omega)D + \omega L^T)(x_1 + x_2 + \ldots + x_l) + (D + \omega L)^{-1}(\omega A^T\mathbf{b})$$

for $l = 0, 1, 2, \ldots$. Now let $x^{(l)} \equiv x_1 + x_2 + \ldots + x_l$, where in fact $x^{(l)} = [s_{1l}, s_{2l}, \ldots, s_{ml}]^T$. What we obtain is equivalent to the SOR iteration for solving the normal equation $A^TAx = A^T\mathbf{b}$, i.e.

$$x^{(l+1)} = (D + \omega L)^{-1}((1 - \omega)D + \omega L^T)x^{(l)} + (D + \omega L)^{-1}(\omega A^T\mathbf{b})$$

for $l = 0, 1, 2, \ldots$ and $0 < \omega < 2$ the SOR relaxation parameter (compare with Gentle, 1998, p.103f). SOR in combination with conjugate gradient methods is very useful for solving large linear sparse systems.

6.5 An Algorithm for the RIP Procedure

Let us have the following starting values: $u = b, x = 0$, where u and x are vectors, $mu = 0$, $k = 0$, where mu and k are scalars. A stopping rule for the iteration process is required. The iterations are limited by $MaxIter$. For the RIP procedure without vector permutation the algorithm can be characterized as follows:

```
while not break
  uTemp = u
  for i = 1 to m
     mu(i) = InnerProd(omega * uTemp, a(i))/InnerProd(a(i), a(i));
     uTemp = uTemp - mu(i) * a(i);
  for j = 1 to m
     x(j) = x(j) + mu(j);
  u = uTemp;
  term = EuclidNorm(u);
  if term fulfills stopping rule
     break;
  k = k + 1
  if k > MaxIter
     break;
```

The algorithm has a simple recursive structure and is built on inner (dot) products. There are highly reliable ways to calculate inner products. The operations are carried out in double precision and the dot product itself has a very good relative numerical error (see Golub and van Loan 1989, p.65 for details). Apart from that the RIP procedure has the valuable feature of self-correction. This means that numerical errors are automatically taken care of in the next iteration cycle, a positive side effect of the projection concept. Last but not least the memory requirements are small compared to standard procedures like QR decomposition. Hence the algorithm is a reliable tool for the solution of linear equation systems of any size, even when rank-deficient.

In addition we can take advantage of patterns in the system matrix A such as structural zeros during the calculation of the inner products. This can save a good deal of computer time. This is specially true for bandlimited systems.

The smaller the dimension m the less cycles are required. The reason is that the solution is found in a subspace smaller than $col(A)$.

The algorithm has been implemented in S-Plus and MATLAB.

6.6 Some Illustrative Examples

We present a few examples. The studied equation systems are different in size (square and rectangular) and structural features. Because there is no mathematical theory how to choose an optimal parameter $\hat{\omega}$ we were considering a range of suitable relaxation parameters including $\hat{\omega} = 1$ for no relaxation. The iterations (cycles) required until convergence were studied for consistent and over-determined systems.

Permutations were not applied to the column vectors $\mathbf{a}_i$ of A. Performance was solely controlled through relaxation. Applying equation (3) we took $\omega = 1(0.1)1.9$. The emphasis was on arbitrary system matrices A which are not diagonally dominant to be solved under either *Lemma 1* or *Lemma 2*.

Double precision arithmetic was used throughout. The iterative solutions were calculated up to machine precision on a pentium M platform under Microsoft Windows XP.

Apart from numerical accuracy the number of iterations until convergence (computational expense) was an evaluation criterion. The minimum number was identified and the associated solution compared to the unrelaxed one.

We first consider an example with a 4×4 regular matrix A which does not have diagonal dominance,

$$\begin{pmatrix} 6 & 2 & 4 & 1 \\ 2 & 8 & 1 & -1 \\ 3 & 2 & -11 & 1 \\ -2 & -1 & 1 & 7 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 15 \\ -12 \\ -33 \\ 10 \end{pmatrix}.$$

The minimum iteration number was obtained for $\hat{\omega} = 1.1$. In Table 6.1 the exact and the estimated solution for the vector $\mathbf{x}$ are displayed.

The results are not surprising because A is not ill-posed: this means a very small number of iterations l. The lack of diagonal dominance is irrelevant as pointed out in the theoretical section which can be clearly seen here. As expected a slightly larger value of $\hat{\omega} = 1.1$ reduces the cycles from 28 in the unrelaxed case to 16. The precision of the solution vector $\mathbf{x}$ is extremely high in both instances.

In the next example we are going to see that even for a regular 4×4 matrix the computational burden can sometimes be quite high. Relaxation brings

exact x	estimated $\hat{\mathbf{x}}$	
	$\omega = 1, \quad l = 28$	$\hat{\omega} = 1.1, \quad l = 16$
1	0.9999999999999140	1.000000000000126
-2	-1.9999999999999600	-2.000000000000031
3	2.9999999999999970	3.000000000000004
1	0.9999999999999966	1.000000000000007

Table 6.1: Results for unrelaxed and relaxed iterative projections.

about a substantial reduction of cycles while obtaining the same accuracy.

$$\begin{pmatrix} 1 & 5 & 3 & 2 \\ 1 & -1 & -1 & -1 \\ 2 & 2 & 0 & 7 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ -12 \\ 4 \end{pmatrix}.$$

Table 6.2 shows the necessary iterations and thus obtained solutions for $\hat{\omega} = 1$ and for $\hat{\omega} = 1.7$, the value associated with the minimum of l.

exact x	estimated $\hat{\mathbf{x}}$	
	$\omega = 1, \quad l = 939$	$\hat{\omega} = 1.7, \quad l = 152$
0.5000	0.4999999999997246	0.4999999999996168
-4.3125	-4.3124999999989320	-4.3124999999990860
8.4375	8.4374999999985500	8.4374999999989090
-0.6250	-0.6250000000002203	-0.6250000000001459

Table 6.2: Results for unrelaxed and relaxed iterative projections.

Again we find an excellent approximation to the exact solution vector $\mathbf{x}$. Moreover the unrelaxed and the relaxed results are essentially identical.

The final example concerns a 4×3-dimensional overdetermined equation system (see *Lemma 2*) with $\|A\mathbf{x} - \mathbf{b}\| \rightarrow min$,

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 2 \\ 3 & -1 & 5 \\ 2 & -1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ -3 \\ -2 \\ 4 \end{pmatrix}.$$

Here we have $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$ with

$$\mathbf{b}_1 = \begin{pmatrix} \frac{19}{6} \\ -\frac{8}{3} \\ -\frac{13}{6} \\ 4 \end{pmatrix}, \quad \mathbf{b}_2 = \begin{pmatrix} -\frac{1}{6} \\ -\frac{1}{3} \\ \frac{1}{6} \\ 0 \end{pmatrix}.$$

where $\mathbf{b}_1 \in col(A), \mathbf{b}_2 \perp col(A)$ with $\mathbf{b}_2 \neq 0$ (i.e. $A^T\mathbf{b}_2 = 0$). The solution of the minimization problem, unique in $A\mathbf{x} = \mathbf{b}_1$, is

$$\mathbf{x} = \begin{pmatrix} \frac{43}{18} \\ \frac{119}{54} \\ -\frac{77}{54} \end{pmatrix} = \begin{pmatrix} 2,388888888888889 \\ 2.203703703703704 \\ -1.425925925925926 \end{pmatrix}.$$

In Table 6.3 the obtained results for two selected $\hat{\omega}$ values 1 and 1.7 after 63 respectively 49 cycles are seen. All the solutions are practically identical.

exact x	**estimated $\hat{\mathbf{x}}$**	
	$\omega = 1, \quad l = 1$	$\hat{\omega} = 1.9, \quad l = 1$
2,388888888888889	2,388888888888888	2,388888888888892
2.203703703703704	2.203703703703703	2.203703703703702
-1.425925925925926	-1.425925925925926	-1.425925925925926

Table 6.3: Results for unrelaxed and relaxed iterative projections.

As can be seen in Table 6.3 the RIP solutions – here given for the smallest (i.e. no relaxation) and the largest $\hat{\omega}$ value suggested – perfectly match the fractions of the analytical solution on which the construction of the equation system was based. Apart from computational speed there is no need for relaxation in this example.

6.7 Conclusions

In summary, also based on other evidence we have from real data of the size typical for backfitting applications, we can say that RIP fitting yields reliable results. There are certainly faster algorithms to solve systems of normal equations of full rank. The advantage of RIP fitting is its ability to cope with rank-deficient situations typical for certain non- and semiparametric regression problems. However relaxation can help to substantially reduce the computational burden. Suitable values for $\hat{\omega}$ usually range between 1.5 and 1.8. Hence it is sufficient to consider this much smaller interval compared to the theoretical result. For near-singularity situations $\hat{\omega}$ values up to 1.9, but not larger, are recommended.

For most regular systems the improvement in convergence speed due to relaxation brings about a performance similar to that of basic iterative techniques such as Gauss-Seidel backfitting. Moreover RIP fitting is a powerful tool for the solution of bandlimited linear equation systems because zero elements do not cause any costs (due to the use of inner products).

A major advantage of RIP fitting is that convergence can be established without restrictions on the system matrix. For that reason diagonal dominance

or regularity assumptions as required for basic iterative techniques in least squares problems are not relevant here. As pointed out earlier, rank-deficient linear equation systems should never be solved with classical techniques as regularity of the system matrix is required throughout. For RIP fitting (RIP-GAM) as an alternative to backfitting in generalized additive models (GAM) with substantially correlated predictors (in a medical MRI study) see Schimek (2002).

The proposed RIP algorithm has the potential to bridge the gap between the established iteration techniques for regular systems and the need for more stable numerical methods when the condition of the system matrix is unclear or deteriorating as is typical for black box tasks such as in data mining.

Bibliography

Bjorck, A. (1996) *Numerical Methods for Least Squares Problems.* SIAM. Philadelphia, PA.

Buja, A., Hastie, T. and Tibshirani, R. (1989) Linear smoothers and additive models (with discussion). Annal. Statist. 17, 453–555.

Dominici, F. et al. (2002) On the use of generalized additive models in time-series studies of air pollution and health. Amer. J. Epidemiol., 156, 193–203.

De la Garza, A. (1951) An Iterative Method for Solving Linear Equations. *Oak Ridge Gaseous Diffusion Plant, Rep. K-731*, Oak Ridge, TN.

Demmel, J. W. (1997) *Applied Numerical Linear Algebra.* SIAM, Philadelphia, PA.

Gentle, J. E. (1998) *Numerical Linear Algebra for Applications in Statistics.* Springer, New York.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach.* Chapman & Hall, London.

Golub, G. H. and van Loan, C. F. (1989, 2nd edition) *Matrix Computations.* The Johns Hopkins University Press, Baltimore, MD.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models.* Chapman and Hall, London.

Maess, G. (1988) Projection methods solving rectangular systems of linear equations. *J. Comp. Appl. Math. 24*, 107–119.

Murty, K. G. (1983) *Linear Programming.* Wiley, New York.

Opsomer, J. D. and Ruppert, D. (1997) Fitting a bivariate additive model by local polynomial regression. Annal. Statist., 25, 186–211.

Ramsay T.O., Burnett R.T. and Krewski D. (2003) The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. Epidemiology, 14, 18–23.

Schimek, M. G. (1996). An iterative projection algorithm and some simulation results. In Prat, A. (ed.) *COMPSTAT '96. Proceedings in Computational Statistics.* Physica-Verlag, Heidelberg, 453–458.

Schimek, M. G. (2002) RIP-GAMs and classification trees in quantitative MRI. *J. Jpn. Soc. Comp. Statist., 15*, 123–134.

Schimek, M. G. & Turlach, B. A. (2000) Additive and generalized additive models. In Schimek, M. G. (ed.) *Smoothing and Regression. Approaches, Computation and Application.* Wiley, New York, 277–327.

Wood, S. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. J. Amer. Statist. Assoc., 99, 637–686.

7 About Sense and Nonsense of Non- and Semiparametric Analysis in Applied Economics[1]

Stefan Sperlich

Georg August Universität Göttingen, Institut für Statistik und Ökonometrie, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany

Summary

The discussion about the use of semiparametric analysis in empirical research in economics is as old as the methods are. This article can certainly not be more than a small contribution to the question how useful is non- or semiparametric statistics for applied econometrics. The goal is twofold: to illustrate that also in economics the use of these methods has its justification, and to highlight what might be reasons for the lack of its application in empirical research. We do not give a survey of available methods and procedures. Since we discuss the question of the use of non- or semiparametric methods (in economics) in general, we believe that it is fair enough to stick to kernel smoothing methods. It might be that we will face some deficiencies that are more typical in the context of kernel smoothing than they are for other methods. However, the different smoothing methods share mainly the same advantages and disadvantages we will discuss. Even though many points of this discussion hold also true for other research fields, all our examples are either based on economic data sets or concentrate on models that are typically motivated from economic or econometric theory.

Keywords: Model specification tests, semiparametric econometrics, non- and semiparametric estimation

[1]This research was supported by the "Dirección General de Enseñanza Superior' SEJ2004-04583/ECON. We thank J.Mora and L.Collado for helpful discussion.

7.1 Introduction

When a group of researchers specialized in non- and semiparametric statistics, and working since many years mainly in this filed, meet to a workshop "The art of semiparametrics" (moreover, with an explicit section about econometrics), then this seems to be the right forum for a discussion about the following questions: What are the reasons for the continuing lack of applications of non- and semiparametric methods in the empirical research in economics and applied econometrics? Actually, it is not only the lack of application that should concern; often one can even find a strong rejection of these methods from a significant part of the researchers in economics. It is clear from the beginning that it will be impossible to convince those who insist that "these recent developments in statistics are of no use for a better understanding of economic processes" or that there is no need because "all functional forms found by nonparametric methods could have easily been modeled with more conventional parametric ones". Sometimes it is just insufficient mathematical knowledge that causes the dislikes, when e.g. non- and semiparametric methods are considered as "too technical" or when people justify their dislike with the bias inherent in nonparametrics, the lack of knowledge about the degrees of freedom, etc.. In contrast to those "arguments", there are many good reasons why in empirical research, especially in economics, non- or semiparametric applications are rare and many empirical researcher suspicious of these methods.

In several joint works and discussions with different economists, there usually came up the following criticisms:

- lack of interpretability of the estimates, e.g. causalities remain unclear and the lack of possibilities of modeling
- problems with the choice of smoothing parameters (in future SP), and lack of its interpretability
- economic data sets are usually high dimensional and contain many discrete variables; so they have a structure that is hard to manage for nonparametric methods
- imposing restrictions like monotonicity is rather cumbersome
- the treatment of endogeneity and simultaneous equation systems is rather crucial in economics but neglected in the statistic literature
- often, neither the optimal fit nor the regression function on its own are the target of interest

- lack of automatization; the methods are too complex to be managed by the empirical researcher without support from a specialist in nonparametrics

Further, let us recall what Stone (1985) said about the task of statistical modeling. He states that the three fundamental aspects of statistical models are flexibility, dimensionality and interpretability. "Flexibility is the ability of the model to provide accurate fits in a wide variety of situations, inaccuracy here leading to bias in estimation. Dimensionality can be thought of in terms of the variance in estimation, the curse of dimensionality being that the amount of data required to avoid an unacceptable large variance increases rapidly with increasing dimensionality. In practice there is an inevitable trade-off between flexibility and dimensionality or, as usually put, between bias and variance. Interpretability lies in the potential for shedding light on the underlying structure".

Comparing these criteria with the list of criticism from above, we see very nicely how they are interconnected and related to each other. Moreover, we can say: flexibility is given by the nature of nonparametrics; with respect to dimensionality much has been done in the last ten years, but interpretability and "automatization" (including the SP selection) seem to remain the main obstacles.

Obviously, we neither can comment in detail on all these criticisms nor offer solutions to them here. It is evident that most of them refer mainly to the problem of estimation. Indeed, the use of testing methods is much less polemic, except the discussion about optimality and efficiency among statisticians and econometricians.

Due to all this we have decided for the following organization of this paper: We concentrate on the perspectives of the existing non- and semiparametric methods for (research in) economics, discussing briefly some of the open problems where existent. Further, we will separate the discussion of testing from the one of estimation, giving the main emphasis on the second part. The numerous examples provided form the largest part of this article. They are certainly not closed empirical research projects but shall help for illustration. We always try to consider relatively simple regression models (even in the testing part) to highlight our points. So we exclude e.g. transfomration models, measurement error models, survival functions, etc.. Also, we concentrate on cross sectional data. For an overview of semiparametric estimation methods in econometrics we recommend Horowitz (1998), and Härdle, Müller, Sperlich & Werwatz (2004) for a general introduction into these methods.

The rest of the paper is organized as follows. In Section 7.2 we discuss testing model specification in econometrics, separated in the subsections parametric versus nonparametric, (semi-) parametric versus semiparametric, and non- or

semiparametric versus non- or semiparametric models. Section 7.3 discusses semiparametric estimation, separated in the subsections parametrically specified models with unknown error distribution, structural models with flexible functional forms (with some comments on endogeneity), and unstructured nonparametric models. Note that this separation is by no means motivated by statistical aspects. It moreover tries to reflect the different tools of methods from the empirical researchers point of view.

7.2 Testing Model Specification

7.2.1 Parametric Versus Nonparametric Models

This was probably the first class of nonparametric tests to verify the specification of econometric models. The null hypothesis consists of a parametric specification of the regression function whereas the alternative is not specified at all to yield an omnibus test. To be more specific: Consider a regression problem $E[Y|X] = m(X)$, $X \in \mathbb{R}^d$, $d \geq 1$, and let $m(\bullet)$ be parametrically specified by m_θ , i.e. a function that is known up to the unknown parameter (vector) θ. Then, the question to test is

$$H_0 : \ m = m_\theta \ \text{ versus } \ H_1 : \ m \neq m_\theta \tag{7.2.1}$$

Typical examples are to test the linearity assumption of a simple linear model or to test the link function specification of Probit- and Logit- models.

Even though the following classification is discussable, let us divide the different mathematical approaches for these nonparametric testing problems into the following groups: looking at (integrated) conditional moments, empirical process approaches e.g. combined with Kolmogorov-Smirnov type or Cramer-von-Mises statistics, minimax approaches, and integrated squared differences.

We will not discuss here the differences, advantages and disadvantages of these different approaches but remark one point that could be of interest in practical applications: In the case that the test rejects, the empirical researcher would like to "see" what is this alternative that is considered to be significantly closer to the data generating process (DGP) than its null hypothesis. Many tests of the last mentioned group of tests require an explicit estimation of the alternative. This might be one reason why they are more popular in econometrics. Actually, this last group can be reduced mainly to four different statistics

$$E\left[w_X\{m(X) - m_\theta(X)\}^2\right] \quad , \quad E\left[w_X\{m(X) - m_\theta(X)\}e_X\right] \tag{7.2.2}$$

$$E\left[w_X e_X E[e_X|X]\right] \quad , \quad E\left[w_X\{\sigma^2(X) - \sigma_\theta^2(X)\}\right] \ , \tag{7.2.3}$$

where e_X is the residuum under the null hypothesis H_0, and w_X a weight function. It is interesting to mention that for finite samples non of these tests has been found to dominate the others, see Dette, von Lieres und Wilkau & Sperlich (2003). Note that almost all tests need resampling methods (usually wild bootstrap is applied) to find the critical value in practice, i.e. in finite samples.

A main problem with these tests is the SP selection. To circumvent this, recently there is coming up more and more literature on the so called "adaptive testing". The aim is to find a SP that on the one hand holds the wanted first error level and on the other hand maximizes the power of the test.

7.2.2 Semi or Parametric Versus Semiparametric Models

Since for practical inference the omnibus tests of Section 7.2.1 are much too general, apart from the fact that they usually suffer from the curse of dimensionality, there has been developed a class of tests that consider parametric (or semiparametric) null hypotheses versus semiparametric alternatives. This means, only a part of the model is of interest and made more flexible in the alternative. A good example might be to consider generalized (additive) partial linear models of the form $E[Y|X,T] = G\{\beta^T T + \eta(X)\}$, $T^T = (T_1^T, T_2) \in \mathbb{R}^q$, $q > 1$, $T_2 \in \mathbb{R}^1$, where G is a known link function, β and η unknown. A typical question to test would be

$$\begin{aligned} H_0 &: \quad m(x,t) = G\left\{\beta_1^T t_1 + \beta_2 t_2 + \eta(x)\right\} \quad \text{versus} \\ H_1 &: \quad m(x,t) = G\left\{\beta_1 t_1 + \eta_2(t_2) + \eta(x)\right\} \end{aligned}$$

From a statistical point of view one could just apply (maybe with some minor modifications) the tests statistics introduced in (7.2.2) on $\beta_2 t_2$ versus $\eta_2(t_2)$ but this will be very inefficient in many cases.

Additional problems to the ones discussed in Section 7.2.1 are caused by the nonparametric part in the null hypothesis (in our example $\eta(x)$). Not only that this affects the quality of estimation of both models (null and alternative) and thus the power of the test. Moreover, the necessary resampling methods, in particular wild bootstrap, can be seriously disturbed if the null hypothesis, i.e. the DGP for the bootstrap samples, is poorly estimated.

Example 1.

For 1991, one year after the German unification, we want to investigate the impact of various possible determinants on the intention of East-Germans to migrate to West Germany. The original data set contains 3710 East Germans who were surveyed in 1991 in the Socio-Economic Panel of Germany. Here we consider the data sets from two East German countries: the most northern

country of East Germany, i.e. Mecklenburg-Vorpommern (M-V) with $n = 402$, and the most southern one, Sachsen (Sax) with $n = 955$ observations. We use the following variables: family/friend in West, unemployed/job loss certain, middle sized city (10000-100000 habitants) and female [dummies (= 1 if yes, = 0 if no), age (AGE) and household income (HHINCOME) [studentized continuous variables]. The response is 1 if the person said he is willing to migrate and 0 otherwise.

All methods we use for our study are introduced in Härdle, Huet, Mammen & Sperlich (2004).

In a first step we do a purely parametric logit regression, in a second step we fit a semiparametric generalized additive model for both data sets. Table 7.1 gives the estimates for the parametric part. For the semiparametric model, we give the results for different SPs, ($h = 1.0$ and 1.25 for M-V, $h = 0.75$, 1.0 for Sax for the directions of interest, $1.1 \bullet h$ for the nuisance directions). In Figure 7.1 are plotted the additive components for AGE and HHINCOME.

	M-V			Sax		
	par.	semi.a	semi.b	par.	semi.a	semi.b
family/friends West	.5893	.5920	.5809	.7604	.7137	.7289
unemployed/...	.7799	.7771	.7992	.1354	.1469	.1308
middle sized city	.8216	.7156	.7127	.2596	.3134	.2774
female	-.3884	-.3309	-.3485	-.1868	-.1898	-.1871
age	-0.9227	–	–	-0.5051	–	–
hh.income	0.2318	–	–	0.0936	–	–
constant	-1.367	-1.462	-1.411	-1.092	-1.105	-1.101

Table 7.1: Results of purely parametric estimates (par.) and of the parametric part of a generalized additive partially linear regression model: semi.a (with SP 1.0), semi.b (1.25) for M-V; semi.a (0.75) and semi.b (1.0) for Sax.

The estimates do not depend very much on the chosen bandwidth. Moreover, for the linear part of the model the results are similar to the values of the parametric model. So the qualitative interpretation of the parametric coefficients does not change. In the figure the influence of AGE in M-V does not differ strongly from the influence of AGE in Sax, except that the curve from Sax is more flat in the middle part. In contrast, for HHINCOME the curves from both countries have a totally different shape. On first glance one would guess that AGE could be modeled linearly, at least for M-V. This is less clear for HHINCOME.

In a third step we apply a bootstrap test for linearity to the variables AGE

and HHINCOME. We always use 499 bootstrap resamples to determine the critical value. The bandwidths are chosen as above. For the input AGE, linearity is always rejected for the 1 percent level, for all bandwidths in both countries. For the variable HHINCOME, the observed p-values are for M-V .16 [$h = 1.0$], .14 [$h = 1.25$], and for Sax .02 [$h = 0.75$], .01 [$h = 1.0$]. So the deviations for AGE from linearity are more significant. At a first sight, this seems to be surprising because the plots for HHINCOME differ much more from linearity. Reasons are presumably that the estimates for HHINCOME have large variance and/or the model(s) is (are) misspecified, e.g. the link function $G(\bullet)$ could be misspecified.

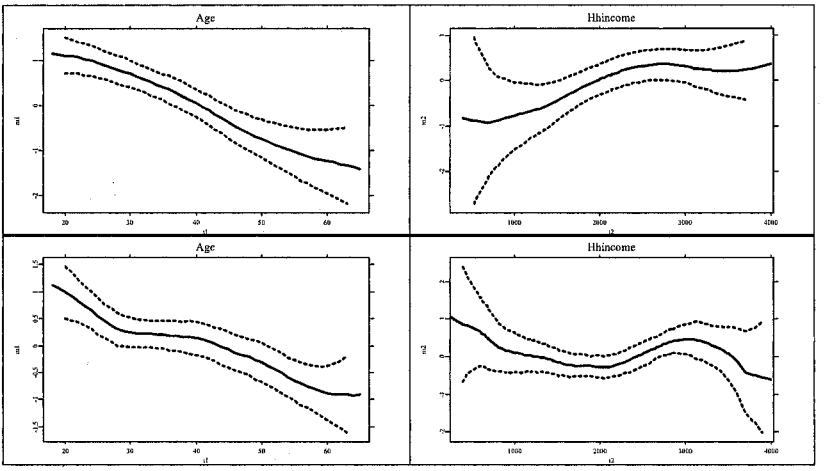


Figure 7.1: Estimates and 95% uniform confidence bands for the impact of AGE (left) and HHINCOME (right) in M-V with bandwidth 1.25 (upper line), in Sax with bandwidth 0.75 (lower line).

To clarify these two points we construct in a next step uniform confidence bands. In Figure 7.1, 95% uniform confidence bands are given for the impact functions for M-V. We use SP $h = 1.25$, and 0.75 respectively, always $1.1h$ for the nuisance directions, and $B = 500$ bootstrap replications. All confidence bands contain a linear fit. Only for HHINCOME in Sax the linear fit lies slightly outside the boundary.

In a last step we test the specification of the link function. For testing we use SP $h = 0.75$, ($1.25h$ for nuisance direction) for M-V and $h = 0.6$ for Sax[1]. With $B = 499$ bootstrap replications we get p-values of about 7% for all SPs for M-V and p-values that are always larger than 15% for Sax. So we

[1]For the test further bandwidths are necessary, see Härdle et al (2004). We tried several over a reasonable range and got always very similar results.

can conclude that the inconsistency we found in the results for AGE in M-V indeed might be caused by a misspecification of the link $G(\bullet)$. □

7.2.3 Non- or Semiparametric Versus Non- or Semiparametric Models

As you can imagine, this title tries to describe something rather general: the check of no parametric specification but of model structures as e.g. additivity, separability, or single index structures. This topic, except additivity testing, can be considered as being still in its infancy.

Here, the maybe most crucial problems are

a) the identification, i.e. the specification of a test statistic that rejects iff H_0 is wrong. E.g. consider weak separable models of the form

$$m(x,t) \;=\; G\{\eta_1(x_1),\ldots,\eta_d(x_d);\theta,t\}\;, \tag{7.2.4}$$

G specified up to an unknown parameter θ. Then, if we reject this model, was it because of the weak separability or because of the specification of G ?

b) the choice of the different SPs, in particular under the null model. Now, often the quality of estimating the null hypothesis has a direct effect on the quality of the test in practice. In most cases, if the null model is not estimated sufficiently well, the bootstrap fails completely even though it is consistent, see Dette, von Lieres und Wilkau & Sperlich (2003). Moreover, the SP of the null easily becomes an inherent part of the hypothesis H_0, see also Rodríguez-Poó, Sperlich & Vieu (2003).

7.3 Non- and Semiparametric Estimation

7.3.1 Parametrically Specified Models with Unknown Error Distribution

A most simple example for estimators in these kind of regression problems are the orthogonal least-squares estimators. For them no new, sophisticated estimation tool is necessary, and they therefore are commonly not mentioned in the context of semiparametric models. But, the hypothesis of unknown error distribution becomes quite a problem when we consider latent variables as response and / or simultaneous equation systems, both rather common in econometrics. Whereas a simple latent variable regression is nothing else than a generalized linear model with unknown link, called single index model

(for that we known a huge amount of estimation literature, see Horowitz (1998) or Härdle et al (2004)), the second problem is much more complex:

Consider the selectivity model

$$y = \{\beta_0 + x^T\beta_1 + u\}\, d \ , \quad d = \mathbb{1}\{g_\theta(t) > \epsilon\} \ ,$$

u, ϵ being error terms, t another vector of explanatory variables, and g_θ a function specified up to θ. Then we can write

$$y = \beta_0 + \beta_1^T x + \lambda\{g_\theta(t)\} + u \ , \qquad \lambda : \ \mathbb{R} \to \mathbb{R} \text{ smooth}$$

and we want to estimate β_1 (maybe also $\lambda\{g_\theta(t)\}$) semiparametrically. So we do not specify $\lambda(\bullet)$, i.e. the joint distribution of (u, ϵ).

You can apply the so called differencing estimator. For the estimation of β_1, function $\lambda(\bullet)$ is an infinite dimensional nuisance parameter. To get rid of it consider the following difference

$$y_i - y_j = (x_i - x_j)^T\beta_1 + \lambda\{g_\theta(t_i)\} - \lambda\{g_\theta(t_j)\} + u_i - u_j \ , \ i \neq j = 1 \ldots, n.$$

With some weights inverse to $|\lambda(\hat{g}_i) - \lambda(\hat{g}_j)|$, i.e. to $|\hat{g}_i - \hat{g}_j|$ you get

$$\hat{p}_{ij} \ = \ \frac{1}{h} L\left(\frac{\hat{g}_i - \hat{g}_j}{h}\right) d_i d_j \ , \qquad L(\bullet) \text{ some kernel function.}$$

Here, $Z = Z(T)$ are some instruments for X (if needed). The final estimator is

$$\hat{\beta}_1 = \hat{S}_{zx}^{-1}\hat{S}_{zy} \quad \text{with} \quad \hat{S}_{zx} = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \hat{p}_{ij}(z_i - z_j)(x_i - x_j)^T \ .$$

This idea and procedure has been suggested by Powell (1987) but without providing proofs. For them we refer e.g. to Rodríguez-Poó, Sperlich & Fernández (2005). Finally, we would like to remark that this approach has become very popular also in the so called (semiparametric) propensity score analysis.

As these models are essentially parametric, there meet almost none of the criticisms on non- and semiparametric methods (mentioned in the introduction), except the choice of SP and its interpretation. It is clear that the optimal SP is the one that minimizes the mean squared error of $\hat{\beta}_1$. However, it is not that clear how to find this in practice.

7.3.2 Structural Models with Flexible Functional Forms

When we speak of structural models we refer to models that are specified in their structure but not (completely) concerning the functional forms. Typical examples are

a) when the empirical researcher wants to specify his model up to some nuisance parameters; e.g. he includes variables in his model to reduce the noise or avoid endogeneity but does not want to specify its functional impact neither is interested in it.

b) models with some pre-specified separability, additive interaction models, multi index models, etc..

The estimation of (generalized) additive models is already well studied, also the one of additive interaction models (for an overview see Sperlich (1998)), whereas the research on semiparametric estimation of weak and latent separable models (compare equation (7.2.4) for its definition) is rather recent, see Rodríguez-Poó et al (2003) and Mammen & Nielsen (2003).

The method of Mammen & Nielsen (2003) is based on smoothed backfitting. On the one hand the identification problems are solved and asymptotic properties developed for a wide range of models, on the other hand the implementation is so far an open problem and it is already clear that the computational expenses will be rather high.

In contrast, Rodríguez-Poó et al (2003) introduce an easy to implement estimation procedure for a wide range of rather general models. However, they could give complete asymptotic theory only for a family fulfilling rather strong (identification) conditions. Their estimation algorithm is based on three-step smoothed likelihood estimation. For identification they need to assume the conditional density of the response as known. They give examples with truncated and censored response variables, in particular the Gronau (1973) model, but allow for flexible functional forms for the η_j, $j = 1, \ldots, d$ (compare equation (7.2.4)).

Example 2.

We estimate a female labor supply model for married woman where labor supply is measured in real hours of work. Note that this variable accounts for the number of hours per week the women had declared to work. Many parametric specifications have been tried to model the hours function in this context. A most famous one is the study about the sensitivity against economic and statistic assumptions by Mroz (1987). In our study we only have to specify the error distribution and how we want to combine the nonparametric components. The hours are assumed to be generated by a Tobit 1 model with truncated variables, i.e.

$$y_i = \begin{cases} h(x_i, t_i) + u_i & \text{if } h(x_i, t_i) + u_i > 0,\ u_i \text{ error term} \\ 0 & \text{otherwise} \end{cases} \tag{7.3.1}$$

We concentrate on a comparison of specifications of possible interactions in h as well as of the behavior of married woman in East and West Germany three years after unification, i.e. in 1993. Those comparisons became quite

popular as, due to completely different political, economic and social systems before 1990, the levels of employment of woman where quite different too: in 1993 in the East still about 65%, in the West only about 54%. Consequently, all the studies in the literature have concentrated on participation at all.

We use data taken from the Social Economic Panel of Germany, wave 1993, cleaned for persons with missing values in the relevant questions and skipping East Germans living in West, West Germans living in East. We have 681 observations for West and 611 for East Germany with a job (i.e. hours > 0). We choose as explaining variables the number of children (Ch1= $\mathbb{1}\{one\ child\}$, Ch2= $\mathbb{1}\{more\ children\}$), education (Edu1= $\mathbb{1}\{high\ school\}$, Edu2= $\mathbb{1}\{academic\ degree\}$) and unemployment rate of the country the person lives in (Urate) for the linear part ($t^T\gamma$). Note that in East Germany there are only 5 countries. For the nonparamertic part $\eta(x)$ we have age of woman (Age), net wage per real hours (Wage), prestige index of their job (PI) and number of years of interruption of professional career (off). For further main income and expenditures we include also the net income of partner per month (Income), and the expenditures for flat minus net income from letting flats (R & L = rent-let). Most probable is an interaction between the determinants of further household income and expenditures apart from the women's one. These are the last two mentioned variables (X_5, X_6). Therefore we study the models of the form

$$h_w(t,x) = t^T\gamma + \eta_1(x_1) + \cdots + \eta_5(x_5) + \eta_5(x_5)\eta_6(x_6) \quad (7.3.2)$$

$$h_s(t,x) = t^T\gamma + \eta_1(x_1) + \cdots + \eta_5(x_5) + \eta_6(x_6) \quad (7.3.3)$$

$$\text{and} \quad h_m(t,x) = t^T\gamma + \eta_1(x_1) + \cdots + \eta_5(x_5, x_6)\ .$$

To make them comparable we set $E[\eta_j(x_j)] = 0$, $j = 1,2,3,4,6$. If by this separability assumption the model is well specified, X_5, X_6 more or less independent, we should get out the same estimates for both specifications, up to a multiplying constant $c = E[\eta_5(x_5)]$ for η_6.

We apply the procedure of Rodríguez-Poó et al (2003). For West Germany we take always SPs $h_j = 1.25\hat{\sigma}_{x_j}$, $j = 1, \ldots, 6$, for East Germany $h_j = 1.5\hat{\sigma}_{x_j}$ as we have less data. Here, $\hat{\sigma}_{x_j}$ indicates the estimated standard deviation of X_j.

Let us first consider the comparison of the different specifications and focus for presentation on the West German data. In Figure 7.2 and Table 7.2 (left side for West Germany) we see the results for the additive case h_s. In the table are given additionally the results for a pure parametric linear model (first two columns), all with standard deviations in brackets. In the parametric model we introduced Age**2. This parametric analysis was only done to compare with the parameter estimates $\hat{\theta} = (\hat{\gamma}^T, \hat{\sigma})$ of the semiparametric model. It can be seen that, apart from Edu2 for East Germans, the coefficient estimates

	West Germany				East Germany			
Ch1	-7.847	(1.087)	-6.913	(.7850)	-2.702	(1.054)	-2.152	(.9910)
Ch2	-11.91	(1.221)	-10.84	(.9549)	-2.313	(1.178)	-2.040	(1.130)
Edu1	-.1027	(1.777)	.5738	(1.383)	1.670	(1.300)	1.318	(1.180)
Edu2	.1403	(2.070)	2.125	(2.084)	1.575	(1.610)	4.868	(1.562)
Urate	.2003	(.2254)	.0925	(.1587)	-.5204	(.3242)	-.4256	(.2934)
Age	1.351	(.4662)	-	(-)	1.460	(.4034)	-	(-)
Age**2	-.0184	(1.E-6)	-	(-)	-.0186	(1.E-6)	-	(-)
ln(Wage)	-7.431	(1.067)	-	(-)	-4.126	(.9695)	-	(-)
PI	.2673	(.0436)	-	(-)	.0820	(.0300)	-	(-)
off	-.3485	(.0616)	-	(-)	-.7367	(.1741)	-	(-)
Income	-.1206	(.0245)	-	(-)	-.1200	(.0316)	-	(-)
R & L	.0188	(.0141)	-	(-)	.1092	(.0469)	-	(-)
σ	10.21	(.2961)	6.955	(.1145)	7.828	(.2241)	6.303	(.1803)
Const	24.06	(9.317)	32.44	(-)	30.16	(9.118)	47.25	(-)

Table 7.2: Results for parametric linear model (columns 1,2 and 5,6) and the semiparametric model (columns 3,4 and 7,8). The standard deviations are given in brackets. In the last line, for the semiparametric model *Const* refers to $\widehat{E}[\eta_5(X_5)] = \frac{1}{N}\sum_i \hat{\eta}_5(x_{i5})$.

do hardly change. But, the error variance (what is not surprising having decreased the degrees of freedom) as well as the variances of the estimates (what is a very good sign) have been reduced a lot using semiparametric methods.

Compare now Figures 7.2 and 7.3. In Figure 7.3 are given the results for the two last component estimates for h_w, $\hat{\eta}_5$ being centered to zero (not for the estimation, only for the presentation). On the bottom of all graphs are given crosses for each observation to indicate the density of the corresponding variable. Up to a multiplying constant c for η_6, they are all the same. For this reason the other components for h_w are not shown as they are exactly the same as we see them in Figure 7.2. Moreover, c is equal to *Const* from Table 7.2. This could be taken as an indicator that the model might be well specified by h_s. The estimation of h_m does thus not add any new information.

Now we look on a comparison between the West and the East Germans. As said in the beginning, they come from completely different political, social and economic systems, and though in 1993 at least the political and the economic systems were the same, there were still differences in the economic and political environments. Let us to mention some specials from the East: the unemployment rate was much higher in the East, a higher willingness and motivation of women to search a job, partly based on the lower salaries

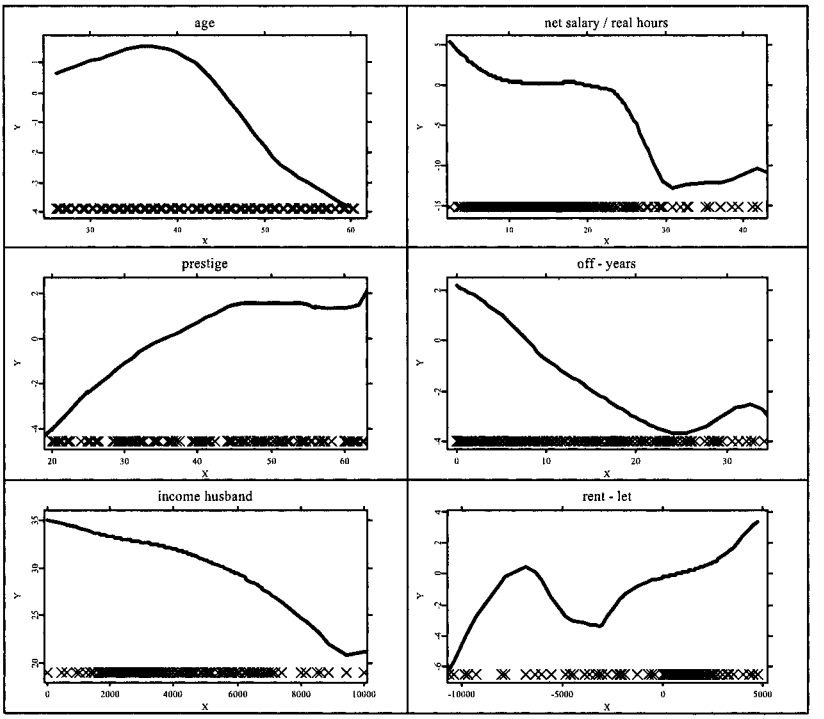


Figure 7.2: West German women. Results for the additive specification (7.3.3). Here, η_5 is centered to zero. Crosses stand for the observations to indicate the density.

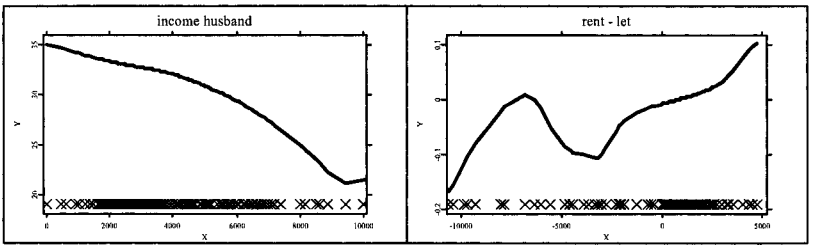


Figure 7.3: West German women. Results for two last components in specification (7.3.2). Here, η_5 is centered to zero.

(compared to the West) of their husbands, a much wider provision of kinder gardens and other possibilities to leave his children. The results are provided in Table 7.2, Figure 7.2 (for the West) and 7.4 (for the East), all based on model h_s. Looking on them, we conclude that behavior for labor sup-

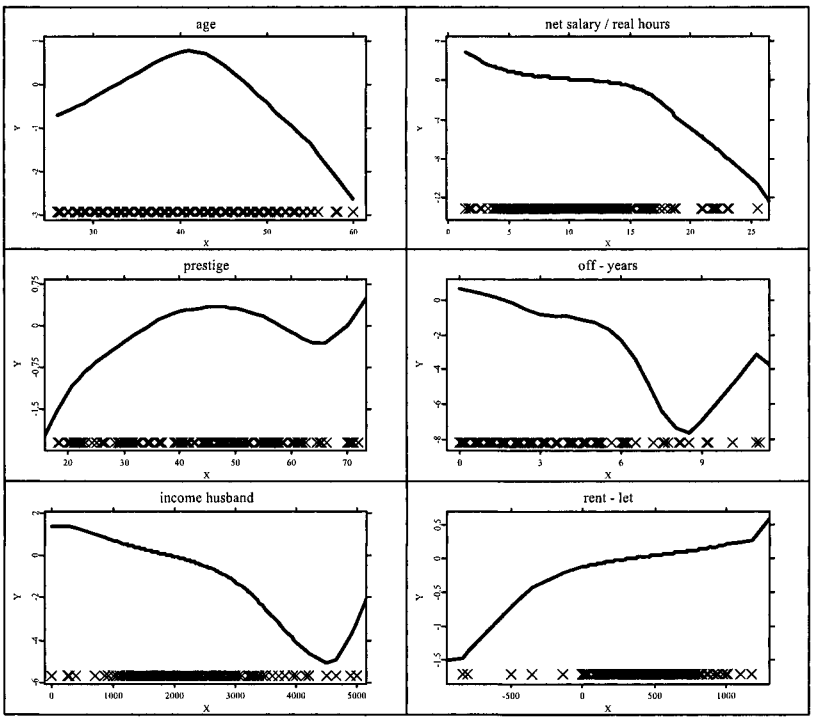


Figure 7.4: East German women. Results for additive specification (7.3.3). Here, η_5 is centered to zero. Crosses stand for the observations to indicate the density.

ply measured in real hours of work is pretty the same in the East and the West, except for education and number of children. The latter outcome was expected for aforementioned reasons. Comparing this with results of other studies which used the same data base, this is a little bit surprising as they found big differences in behavior when looking on participation at all. □

Here now we have faced several of the problems of nonparametric estimation and its solutions (enumerated in the introduction). The lack of the possibility of modeling has been reduced by semiparametric modeling; also imposing restrictions like monotonicity for the nonparametric part is sometimes possible to impose. This improves automatically the interpretability of the estimates, and often enables or facilitates the estimation of parameters or functions of particular interest (e.g. elasticies, rates of substitution, etc.). Additionally, it can reduce the curse of dimensionality, see Stone (1986), Rodríguez-Poó et al (2003). In those models, the choice of the SP should be considered like choosing the degrees of freedom, i.e. the empirical researcher allows for more

flexibility or imposes more smoothness on its functionals. To my opinion, in this context, the "optimality" of the SP has to be defined along the aim of the empirical researcher. Therefore, it is impossible to give here a general rule how to chose it in practice.

Finally let us comment on the problem of endogeneity. To my knowledge, in the context of semiparametric analysis where the regression of interest contains nonparametric functions this problem has been studied first by Fernández, Rodríguez-Poó, and Sperlich [presented 1999 at ESEM, revised in: Rodríguez-Poó, Sperlich & Fernández (2005)] and by Newey, Powell & Vella (1999). Newey, Powell & Vella (1999) did a profound study on identification. They consider nonparametric (and partially linear) models. Rodríguez-Poó, Sperlich & Fernández (2005) allow for separable and generalized models, therefore they apply assumptions on the error distribution. The two articles coincide in several of the main ideas to circumvent the problem of endogeneity, e.g. they use generated regressors and apply two and / or three step estimators.

7.3.3 Unstructured Nonparametric Models

Although "nonparametric" and "unstructured" is essentially the same, we used this title to emphasize the lack of any specification. The only thinkable compromise could be to include partial linear models as long as the linear part serves only to include the impact of dummy variables.

Nonparametric models are useful for optimal prediction (except extrapolation) and explorative data analysis. We might even say: "and for nothing else". The first point is evident because every imposed structure that is not confirmed by the data itself may reduce the quality of the fit. Usually, this approach is interesting whenever we want to predict best whatever the "true model" (if exists) is. Well known examples are financial data problems as predicting stock or bond prices, risks, interest rates, etc.. For a better understanding why and how even totally nonparametric methods can be helpful here see Nielsen & Sperlich (2003).

Less obvious might be the use of nonparametric statistics to explore economic data if the underlying economic process is of interest. To understand this better, let us consider a real data example, taken from Grasshoff, Schwalbach & Sperlich (1999). They do an explorative analysis about the relation of executive pay and corporate financial performance.

Example 3.

Commonly, empirical research concentrates on the pay-performance relationship. Although very different data sets has been adopted the results are

always similar showing rather low pay-for-performance elasticities. Almost all studies assume linear or semi-log linear pay functions without applying a test of the adequate functional form. They do not allow for variations across corporations, industries, countries and time. It it assumed that pay functions are homogeneous across corporations, variations are captured by the fixed effects in the constants and assumption about the errors. So it would be interesting to circumvent these possible misspecifications by adopting an explorative data analysis using nonparametric methods. And indeed, the results of Grasshoff, Schwalbach, and Sperlich (1998) show clearly that all mentioned issues matter, e.g. industry effects are important, assumptions of additivity and linearity are crucial leading to underestimations of the elasticities, etc.. In sum, their results should have far reaching implications for further empirical studies. They also weaken the concern that strong pay-for-performance incentives for executives are missing.

In analyzing executive pay the standard empirical model contains corporate size and financial performance as determinants of pay. Corporate size is a measure of managerial discretion and financial performance is an indicator for managerial incentive compatibility. Both hypotheses are derived from agency theory. Typically, the following regression equation is assumed:

$$\ln C_i^{j,t} = \alpha_{j,t} + \beta^{j,t} P_i^{j,t-1} + \gamma^{j,t} \ln S_i^{j,t-1} + u_i^{j,t} \ , \qquad (7.3.4)$$

where C_{it} stands for executive pay, P_{it} reflects measures of financial performance and S_{it} represents size for firm i at time t. The terms u_{it} are the stochastic error terms whereas the parameters α_i are mostly modeled as firm-specific fixed effects.

The data base is drawn from varies annual executive pay reports by "Kienbaum Vergütungsberatung". The data contain average annual total pay (fixed and variable) by the top executives of German stock companies (Vorstand of Aktiengesellschaften) and 'companies of limited liabilities' (Geschäftsführer of the Gesellschaft mit beschränkter Haftung). In total, we use data of up to 339 manufacturing firms for the period of 1988 to 1994. Company size is measured by the number of employees and corporate financial performance by the rate of return on sales (ROS). Companies are grouped into the following four distinct industry groups: *(1) Basis industries, (2) Capital goods, (3) Consumer goods* and (4) *Food, drinks and tobacco.* For further details see Grasshoff et al (1999).

To get a primary visual impression of the possible functional forms we first applied the multidimensional, in our case two dimensional, Nadaraya-Watson estimator. The model we estimate is of the form

$$\ln C_i^{j,t} = m^{j,t}\left(P_i^{j,t-1}, \ln S_i^{j,t-1}\right) + u_i^{j,t} \ , \ m^{j,t} : \ \mathbb{R}^2 \to \mathbb{R} \text{ unknown.} \quad (7.3.5)$$

We use the quartic kernel with bandwidth $h = 2.5\hat{\sigma}_X$. Notice that since

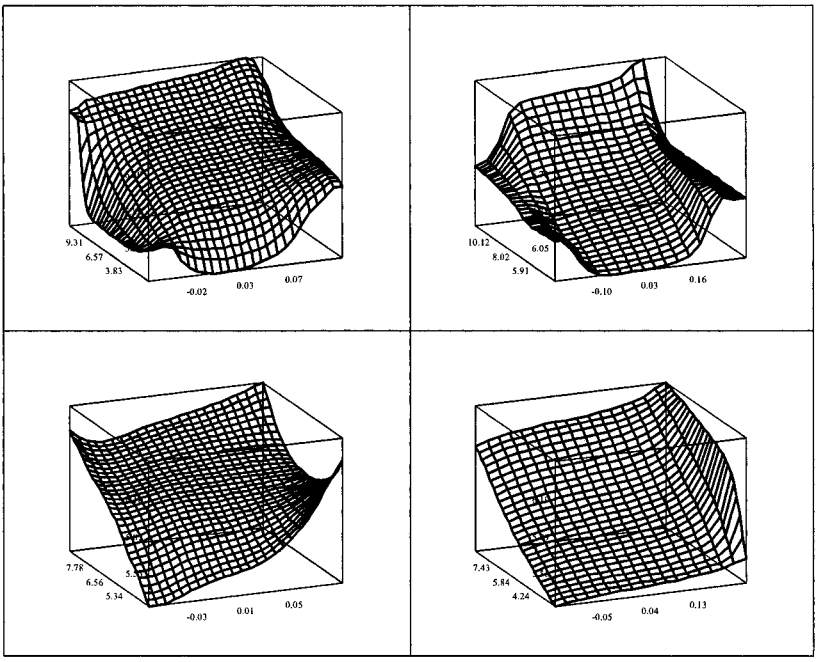


Figure 7.5: The 2-dimensional Nadaraya-Watson estimation for 1989/90. Plotted are the expected executive pay on size (left axes) and ROS (right axes). First row: group1 and 2, Second row: group 3 and 4.

our estimator is a local adaptive one, our results are not effected by possible outliers in the x-direction. For better presentation we show the 3D-plots over trimmed ranges. We have selected the results for two representative years, see Figures 7.5 and 7.6. Considering the plots over the years we can realize strong functional similarities between the industry groups 1 and 2 while the results for the other groups seem not to be homogeneous at all. Regardless the outliers we see a strong positive relation for compensation to firm size at least for group 1 and 2, and a weaker one to the performance measure varying over years and groups.

Further we can recognize some interaction of the independent variables especially in group 3 and 4. This can visually be detected as follows. Imagine you cut slices parallel to the x-axes. If these slices indicate different functional forms within one direction separability of the inputs is not justified. Regarding this procedure we state additivity for group 1 and 2.

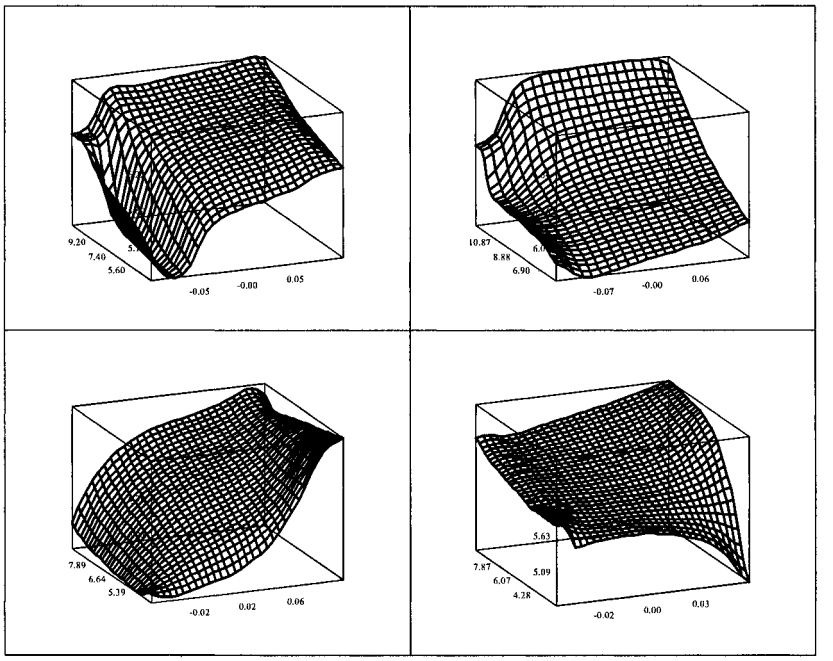


Figure 7.6: The 2-dimensional Nadaraya-Watson estimation for 1990/91. Plotted are the expected executive pay on size (left axes) and ROS (right axes). First row: group 1 and 2, Second row: group 3 and 4.

Next, a study was done where the regression function was modeled additively with backfitting. This study is skipped here as it did not yield much new insight. For more details see Grasshoff et al (1999).

Finally, we estimate the pure marginal effects of the independent variables. The model we estimate can be imagined as general as (7.3.5), but we only estimate the marginal effects, not the joint regression function $m(\bullet)$ (skipping the indices (j,t) above). If the model is of additive form $m(x_1,x_2) = m_1(x_1) + m_2(x_2)$, then the marginal effects correspond to m_1, m_2. We use a local linear kernel smoother with quartic kernel and bandwidths $h = 1.5\hat{\sigma}_{x_k}$, $k = 1,2$, ($2.5\hat{\sigma}_{x_k}$ for the nuisance directions). We present the estimation results together with confidence intervals in forms of $2\hat{\sigma}(\hat{m}_k(x_k))$ -bands, where $\hat{\sigma}(\hat{m}_k(x_k))$ indicates the estimated standard deviation of additive component $\hat{m}_k$ at point x_k.

As a main result we can postulate that these estimation results are consistent with the findings above. First, the nonlinearities of the financial performance influence are strengthened especially for groups 1 and 2. Second, it seems

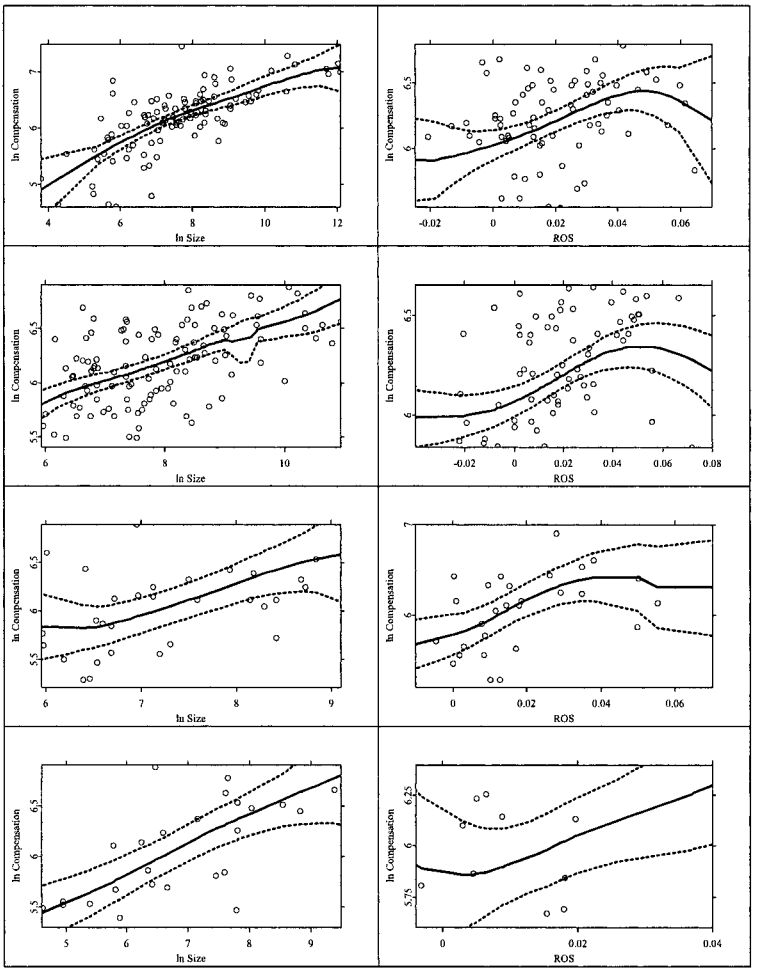


Figure 7.7: Marginal Integration estimates for 1991/92 with $2\hat{\sigma}(\hat{m}_k(x_k))$, $k = 1, 2$ bands for industry groups 1-4 (top to bottom).

that interactions are present, so the assumption of additivity would be wrong what renders an economic interpretation rather difficult. □

Here now meet more ore less all the criticisms against nonparametric methods mentioned in the introduction. Interpretability is hardly given, neither for the estimates nor for the SP choice. In general, the SPs should be chosen data driven (e.g. by cross validation or plug in) to minimize the estimation error or optimize prediction power were prediction is pretended. Certainly, interpretability of the estimates is not necessary if one only wants to predict,

but it is of interest if one wants to make studies as in our example. Further, since the curse of dimensionality kicks in rapidly, the possibilities of these methods are rather limited unless you have really large samples. E.g. in our example, a test for additivity would simply not work, see Dette, von Lieres und Wilkau & Sperlich (2003).

Bibliography

Dette, H., von Lieres und Wilkau, C. & Sperlich, S. (2003), 'A comparison of different nonparametric methods for inference on additive models', *Nonparametric Statistics*, forthcoming

Grasshoff, U., Schwalbach, J. & Sperlich, S. (1999) Executive Pay and Corporate Financial Performance: an Explorative Data Analysis, *Working paper 99-84 (33), Universidad Carlos III de Madrid*

Gronau, R. (1973)The Effects of Children on the Housewife's Value of Time, *Journal of Political Economy* **81**, 168–199.

Härdle, W., Müller, M., Sperlich, S. & Werwatz, A. (2004) *Non - and Semiparametric Models*, Springer Series in Statistics.

Härdle, W., Huet, S., Mammen, E. & Sperlich, S. (2004) Bootstrap Inference in Semiparametric Generalized Additive Models, *Econometric Theory* **20**, 265–300.

Horowitz, J. (1998) *Semiparametric Methods in Econometrics*, Springer.

Mroz, T.A. (1987) The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions, *Econometrica* **55** (4), 765–799.

Newey, W.K., Powell, J.L. & Vella, F. (1999) Nonparametric Estimation of Triangular Simultaneous Equation Models, *Econometrica* **67** (3), 565–604.

Mammen, E. & Nielsen, J.P. (2003) Generalised Structured Models, *Biometrika* **90**, 551–566.

Nielsen, J.P. & Sperlich, S. (2003) Prediction of stocks: A new way to look at it, *Astin Bulletin* **33** (2), 399–417.

Powell, J. L. (1987) Semiparametric Estimation of Bivariate Latent Variable Models, *Working Paper, University of Wisconsin - Madison*

Rodríguez-Poó, J.M., Sperlich, S. & Fernández, A.I. (2005) Semiparametric Three Step Estimation Methods for Simultaneous Equation Systems, *Journal of Applied Econometrics* **20**, 699–721.

Rodríguez-Poó, J.M., Sperlich, S. & Vieu, P. (2003) Semiparametric Estimation of Separable Models with Possibly Limited Dependent Variables, *Econometric Theory* **19**, 1008–1039.

Sperlich, S. (1998) *Additive Modelling and Testing Model Specification*, Shaker Verlag, Aachen.

Stone, C. J. (1985) Additive regression and other nonparametric models, *Annals of Statistics* **13**(2), 689–705.

Stone, C. J. (1986) The dimensionality reduction principle for generalized additive models, *Annals of Statistics* **14**(2), 590–606.

8 Functional Nonparametric Statistics in Action[1]

Frédéric Ferraty[2] and Philippe Vieu[3]

[2] Équipe GRIMM, Université Toulouse Le Mirail, 31058 Toulouse Cedex, France
[3] Laboratoire de Statistique et Probabilités, U.M.R. C5583, Université Paul Sabatier, 31062 Toulouse Cedex, France

Summary

Functional aspects are more and more frequent and varied in modern Statistics so much so that the designation of *Functional Statistics* had emerged recently. From a practical point of view, this is appearing as soon as one has to deal with data which are curves. A symbolical example of this new field of Statistics concerns the problem of nonparametric regression estimation in presence of functional data. This problem is doubly functional: the nature of the observed data (that is, the nature of the curves) is functional and the statistical model is also functional (that is, it is nonparametric). The only goal of this presentation is to show how recent Nonparametric Functional Regression Methods work in different practical situations. Several data sets are chosen to cover different fields of applied statistics (chemiometrics, speech recognition, econometrics) as well as different facets of statistical regression problems. Each example is quickly treated. Complete treatments, theoretical supports and extensive bibliographies are referred to other works.

Keywords: Applied statistics; curves discrimination; functional data; functional statistics; functional regression; nonparametric models; time series.

[1]The authors wish to thank all the participants of the working group STAPH on Statistique Fonctionnelle et Opératorielle of our department. Their continuous support and comments, through the activities of this group, are of great importance in the development of our researches. Complete activities of this group are available on line at *http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html*.

8.1 Introduction

In the few past years, functional aspects had taken an important place in the Statistical Science. In a first side, since the beginning of the sixties, a lot of attention has been paid to free-distribution statitiscal models and/or methods. The functional feature of these methods comes from the nature of the object to be estimated (such as for instance a density function, a regression function, ...) which is not assumed to be parametrizable by a finite number of real quantities. In this setting, one is usually speaking of *Nonparametric Statistics* for which there is an abondant literature. For instance, the reader will find in Härdle (1990) a previous monography for applied nonparametric regression, while Schimek (2000) and Akritas & Politis (2003) present the state of art in these fields. From an other side, there is actually an increasing number of situations coming from different fields of applied statistical (environmetrics, chemometrics, biometrics, medicin, econometrics, ...) in which the collected data are curves. In this situation, the functional feature is linked with the observations. Since the middle of the nineties, this has motivated different statistical developments, that we could quickly name as *Statistics for Functional Variables* (or *Data*). Most often at this time, for seek of simplicity, these functional models were combined with parametric modelling for the object to be estimated, and key references in this respect are those by Ramsay & Silverman (1997), Bosq (2000) and Ramsay & Silverman (2002).

We focus on recent *Functional Nonparametric Statistical Methods*, that is on methods which can capture both the functional (*i.e.* the nonparametric) feature of the statistical target and the functional nature of the statistical samples for which curves (or, more generally, functional objects) are directly observed. We concentrate exclusively on the applicability of these methods, and more precisely on regression type methods. The reader could find the mathematical and methodological supports of the proposed methods in the companion paper (Ferraty & Vieu 2003b), while an extensive bibliographical presentation of the state of art in Functional Nonparametric Regression can be found in Ferraty & Vieu (2003c). Note finally, that in addition to these english references, Ferraty & Vieu (2001), Ferraty & al. (2001) or Ferraty & al. (2002) provides a basic curse on Functional Nonparametric Regression in other languages. To complete this quick bibliographical survey, let us also mention the monographies by Ferraty (2003) and Goia (2003) in which Functional Regression is investigated both through parametric and nonparametric techniques/models, and the one by Niang (2002) which is devoted to the related problem of nonparametric estimation of the density of functional variables with its direct application to diffusion processes. The reader can also look at the activities of the Staph group on Functional and Operatorial Statistics to have an overlook on the state of art all the facets of Functional Statistics. Most recent activities of this group can be found in Staph (2003).

Let us now concentrate ourselves on Nonparametric Regression for Functional Data, and more specifically on applied aspects of this problem. We have arbitrarily selected three applied statistical problems and/or data sets for our purpose. Our selection has been done following a treble goal:

- covering as much as possible different scientific fields involving applied statistics (chemiometrics, speech recognition, econometrics);
- covering as much as possible different facets of regression type methodological statistical problems (regression, time series, supervised curves classification);
- keeping a reasonable short size contribution ...

As a consequence of these choices, it was impossible to present full studies for each problem. In particular, our point of view avoids both for presenting methodological details about the statistical modelizations and for presenting all the details of our applied studies. Indeed, for each situation we only describe the method and give its result on the corresponding data set, but we paid particular attention to give accurate references to be used by the reader to go back to theoretical supports, to computational details or to bibliographical surveys.

Our paper is organized as follows. The three functional data sets are presented in Section 8.2. Before treating these data sets by mean of nonparametric methods for functional data, we quickly discuss in Section 8.3 the question of the dimensionality and we will see how the curse of dimensionality can be overriden by mean of a semi-metric suitably adapted to each problem. Section 8.4 implements a Nonparametric Functional Regression method on chemiometrical data. Section 8.5 focuses on a Supervised Curves Classification problem for some phoneme recognition data set. Section 8.6 concerns some electrical consumption prediction problem by mean of Functional Nonparametric Time Series Prediction techniques. To help the reader that would be interested in only one among the three applications, we have written Sections 8.4, 8.5 and 8.6 independently one of each other. Finally, a short concluding Section 8.7 is presented; its main goal is to discuss some open problems of particular interest.

8.2 Presentation of Some Functional Data Sets

There are several statistical problems from different sources (medicine, biometrics, econometrics, geophysics, ...) in which the data are curves. It is clear that the development of informatic tools (as well in terms of higher

memory capacity as in terms of speeder computations) allows to register and to treat much larger data sets. In the curves setting, even if it is of course always impossible to observe the data in a continuous way, it is now possible to have finely discretized observations. We describe now three sets of curves, selected to cover several different applied statistics fields. For each of these data sets, the discretization is quite enough fine, with the result that one may consider observations as curves. This Section 8.2 is focusing on these data sets, while their treatments by Functional Nonparametric Methods are presented in next sections. More precisely, in each of the following examples, we present continuous curves (obtained by smoothing the discretized observations). The reader will find more extensive presentation on functional data together with a wider set of possible fields of application in Bosq (2000), Ramsay & Silverman (1997) or Ferraty (2003).

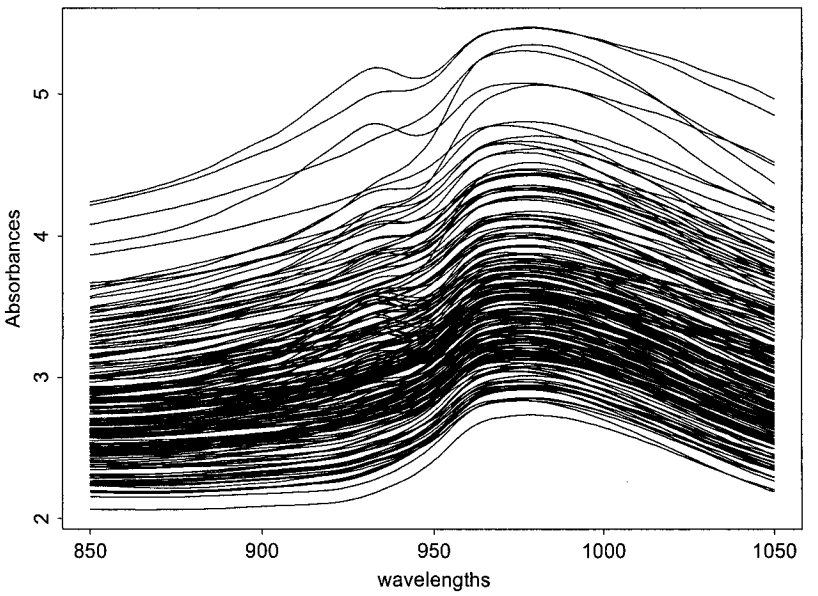


Figure 8.1: The spectrometric curves

Example 1: Spectrometric Data. Spectrometry is a modern tool for analysing the chemical composition of any substance. For instance, here we look at a sample of finely chopped meat. Each analysis can be summarized by some continuous curve giving the observed absorbance as function of the wavelength. Absorbance is the $-log_{10}$ of the transmittance measured by

the spectrometer. As pointed out by Leurgans & al. (1993) we are really in presence of functional continuous data since "the spectra observed are to all intents and purposes functional observations". Our data are recorded on a Tecator Infractec Food and Feed Analyzer working in the wavelength range 850-1050 nm by the near infrared (NIR) transmission principle. More precisely, for each meat sample the data consists of a 100 channel spectrum of absorbances. The data are presented in Figure 8.1 below.

In Section 8.4, we study these data. We will see how these continuous explanatory variables can be used to predict some specific component of the meat (precisely the fat percentage) by using Functional Nonparametric Regression Methods.

Example 2: Phonetic Data. In speech recognition, the observed data are also of functional nature. For instance, look at the following data set, previously introduced and studied by Hastie & al. (1995). The data are log-periodograms corresponding to recording speakers of 32 ms duration. Here also, even if we have to deal with discretized data, the number of observed points is quite large enough to allow for considering the observations to be continuous (as they are, indeed). The study concerns (see Figure 8.2) five speech frames corresponding to five phonemes transcribed as follows:

- "sh" as in "she";
- "dcl" as in "dark";
- "iy" as in "she";
- "aa" as the vowel in "dark";
- "ao" as the first vowel in "water".

Precisely, each speech frame is represented by 400 samples at a 16-kHz sampling rate; only the first 150 frequencies from each subject are retained, and the data consist of 2000 log-periodograms of length 150, with known class phoneme memberships. Indeed, Figure 8.2 only displays 10 log-periodograms for each class phoneme.

In Section 8.5 we attack the natural problem: given a new log-periodogram associated with an unknown phoneme, are we able to predict at which class it belongs? This is typically a curves discrimination problem and we will see how the Nonparametric Supervised Curves Classification Method can provide nice answer to this question.

Example 3: Electrical Consumption Data. Our third example concerns time series prediction. Let us look for instance at the monthly electricity consumption in United States between January 1973 and February 2001.

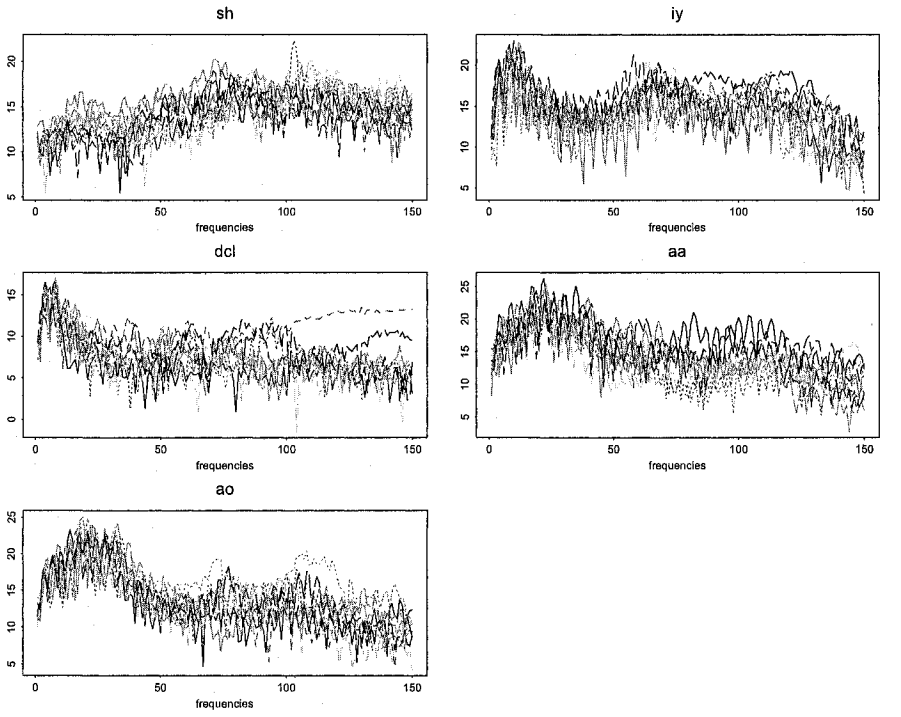


Figure 8.2: A sample of 10 log-periodograms for each of the five phoneme classes

These data, which are presented in Figure 8.3, are obviously exhibiting some linear trend, as well as some heterogeneity in the variance structure.

As usually, in order to avoid for heteroscedasticity or linear trend effects, we work with the differencied log-data. These transformed data are presented in Figure 8.4.

One of the main problems in these situations is to predict future consumption, and usual statistical models (either parametric or nonparametric) achieve that by taking in consideration a finite number of past data. However, one could think that it is more reasonable to take into account as explanatory variable the continuous time series over some period. For our example, we have decided to choose the whole past year as explanatory period. That means that the set of explanatory variables to be included in our statistical method is composed with 28 functional data which are the 28 yearly continuous time series. These functional data are those presented in Figure 8.5.

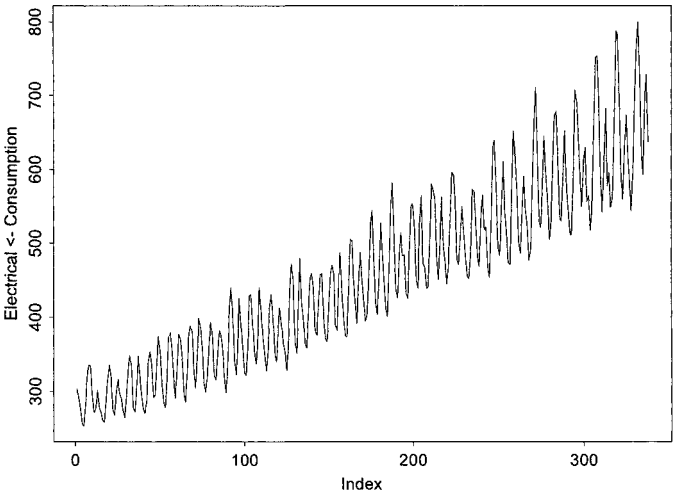


Figure 8.3: The electricity consumption data.

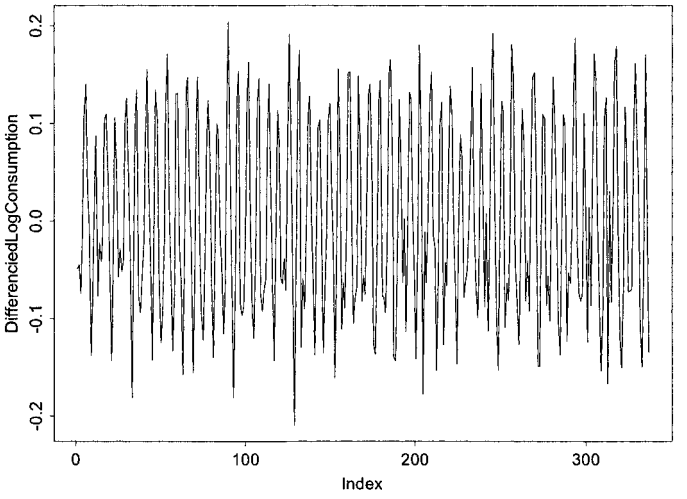


Figure 8.4: Electricity consumption: the differencied log data

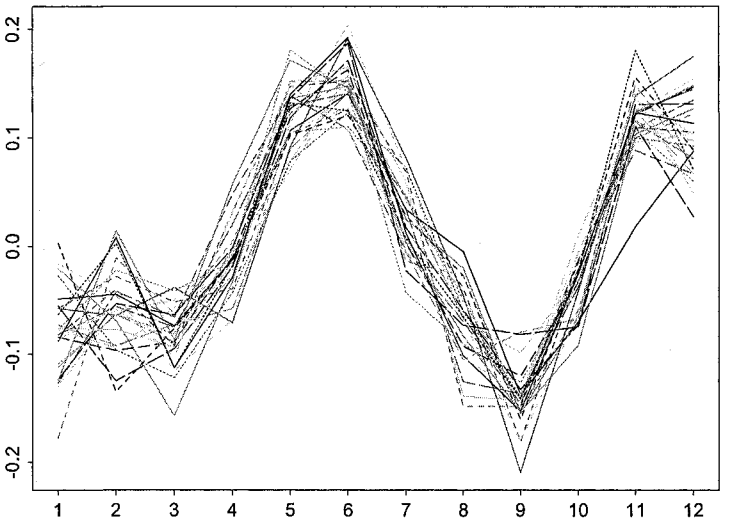


Figure 8.5: Electricity consumption: the 28 yearly differencied log curves

In Section 8.6 we will see how the recent Functional Nonparametric Time Series Prediction Methods can be used to provide good future consumption prediction by mean of these functional explanatory data.

8.3 About Proximity Between Curves and Curse of Dimensionality

Kernel methods are known to be local estimation procedures, in the sense that the data lying in some neighbourhood of x play a major role in the estimation of a function at this point x. When $x \in \mathbb{R}$ or $x \in \mathbb{R}^p$, the notion of neighbourhood is induced by the euclidian metric, but this is of small importance because of the equivalence property of all metrics in finite dimensional spaces. When x is a curve (that is, an object valued in some infinite dimensional space), this equivalence property does not hold anymore. In this framework, we have to define a suitable notion of proximity between curves.

One way to do that could be to introduce some metric, but this can be too much restrictive. For instance, it can be the case in practice that some derivative of a curve is more informative than the curve itself. In particular this is the case for the spectrometric data presented in Figure 8.1 for which there is a vertical shift in the curves that may hide for some interesting feature. So, instead of metric we prefer to introduce some semi-metric. Indeed, most of the nonparametric methodologies existing for curves data (see Ferraty & Vieu 2003c, for a large set of references) consider the curves like objects lying in some semi-metric space. To make ideas clearer, let us note here that in some practical cases we can deal with semi-metrics of the form:

$$d_m(f;g) = \left(\int \left(f^{(m)}(t) - g^{(m)}(t) \right)^2 dt \right)^{1/2}. \qquad (8.3.1)$$

For spectrometric data we will see in Section 4 that this choice of semi-metric is pertinent. For other data sets (see the examples described in Sections 8.5 and 8.6) other choices are possible.

Before to close this section, it is worth to answer to the following natural question: "What about the curse of dimensionality?" Since nonparametric methods are strongly affected by high dimensional problems, one could expect these methods not to be suitable for infinite dimensional framework. Indeed this question, and its answer, is highly correlated with the proximity measure between curves, because the curse of dimensionality comes from the sparseness of the data in high dimensional spaces. One way to reduce this sparseness effect is to construct a semi-metric which insures a good concentration of the curves. This intuitive idea is theoretically supported by asymptotic studies (see Ferraty & Vieu 2003b).

For all these reasons, the semi-metric plays a prominent role in each of the practical cases that we develop in next sections.

8.4 Functional Nonparametric Regression in Action on Spectrometric Data

Let us look at the following problem, linked with the chemiometric data presented in Example 1 before. The task is to predict the fat content of a meat sample on the basis of its NIR absorbance spectrum. More precisely, for each meat sample the data is formed by the spectrum of absorbance (see Figure 8.1) and by the percentage of fat which is determined by analytic chemistry. The question is: "given a new sample of meat, can we predict its corresponding percentage of fat just by looking at its absorbance spectrum? Clearly, unlike the real case, it is difficult to build a simple plot in order to

display the link between fat content and spectrometric curves. In addition, there is no available physical knowledge on the form of the relation between the percentage of fat Y and the absorbance spectrum $X = \{X(t);\ 850 \leq t \leq 1050\}$. So we are typically in front of a Regression Problem which needs to be attacked from a nonparametric point of view. This regression problem is not standard since the explanatory variable is functional (X is a curve). So the model is

$$Y = R(X) + error, \tag{8.4.1}$$

and the data are the following

- $(X_i,\ i = 1, \ldots 215)$ are the spectrometric curves presented in Figure 8.1;
- $(Y_i,\ i = 1, \ldots 215)$ are the corresponding observed percentages of fat.

In order to estimate R we use functional kernel smoothing. For each new curve x, define

$$\hat{R}_n(x) = \sum_{i=1}^{n} K_{n,i}(x) Y_i \text{ where } K_{n,i}(x) = \frac{K\left(h^{-1} d(X_i; x)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(X_i; x)\right)}. \tag{8.4.2}$$

In this definition h is a sequence of positive numbers, K is a weight function that is, all along our spectrometric application, chosen as:

$$K(u) = \frac{3}{2}(1 - u^2),\ \text{if } u \in [0, 1],\ \text{and 0 elsewhere.} \tag{8.4.3}$$

There is a recent mathematical asymptotic support for using such a functional statistical method (Ferraty & Vieu 2003c) but it is not our purpose to describe it here. Also, of course, many other applied statistical problems are open to be treated by such kind of functional regression methods such as the agronomic data presented in Cardot & al. (1999), the environmetric data presented in Aneiros & al. (2003), or the pluviometric data treated in Ramsay & Silverman (1997), but let us only concentrate ourselves on how this procedure works on our chemiometrical example. To show the accuracy of the method to deal with such a functional data set, we have splitted arbitrarily the sample into two parts:

- A training sample (of size 165);
- A testing sample (of size 50).

The training sample has been used to select the parameters of the estimate (that is to select the semi-metric and the bandwidth). Note that the selected

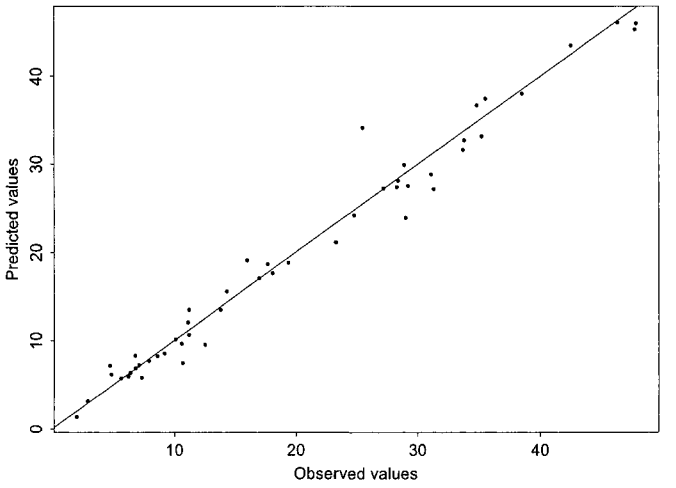


Figure 8.6: Functional nonparametric regression in action on a test sample of 50 spectrometric functional data

semi-metric inside of the family (8.3.1) is the one associated to the second derivative of the curves (i.e. $m = 2$). Once the parameters have been selected, we applied the method to the 50 curves of the testing sample. The results are presented in Figure 8.6.

For seek of shortness, it was out of purpose to present here all the details concerning the application of the method to these spectrometric data (how to choose the bandwidth, how to choose the semi-metric, ...), and the reader who is interested will find answers to these questions in Ferraty & Vieu (2002).

8.5 Nonparametric Curves Classification in Action on Phoneme Data

Let us go now through the supervised classification problem described in our phonetic example 2. Our data are the following:

- $(X_i,\ i = 1, \ldots 2000)$ are the phoneme curves presented in Figure 8.2;
- $(Z_i,\ i = 1, \ldots 2000)$ are the corresponding group.

Z is a categorical response valued in $\overline{G} = \{1, \ldots, G\} = \{sh, dcl, iy, aa, ao\}$. Now, given a curve x, the purpose is to decide at which group it belongs, and for that we look at the following posterior probabilities

$$p_g(x) = P\left(Z = g / \{X(t) = x(t),\ t \in T\}\right), \qquad g \in \overline{G},$$

that are estimated by mean of the following functional version of kernel probability estimates:

$$\hat{p}_{g,n}(x) = \frac{\sum_{i=1}^{n} 1_{[Z_i=g]} K\left(h^{-1} d(X_i, x)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(X_i, x)\right)}.$$

This estimate can be rewritten as

$$\hat{p}_{g,n}(x) = \sum_{\{i/Z_i=g\}} w_{n,i}(x),$$

with

$$w_{n,i}(x) = K(h^{-1} d(X_i, x)) / \sum_{i=1}^{n} K(h^{-1} d(X_i, x)).$$

In all this study, the kernel is choosen to be

$$K(u) = \frac{3}{2}(1 - u^2) 1_{[0,1]}(u),$$

h is the bandwidth, and $d(.;.)$ is a semi-metric which is used to measure the proximity between two curves. We do not enter here in the details of the important role played by the semi-metric, but it is worth to be noted that, as theoretically supported in Ferraty & Vieu (2003b), a good choice of the semi-metric insuring high concentration of the curves data may deal accurately with the curse of dimensionality.

Once these probabilities have been estimated, it is natural to assign the new curve x to the class with highest estimated probability:

$$\hat{Z}(x) = argmax_{\{g \in \overline{G}\}} \hat{p}_{g,n}(x).$$

Extensive discussion of the related bibliography can be found in Ferraty & Vieu (2003c) and theoretical support are in Ferraty & Vieu (2003b). Also, many other data sets are open to be treated by this approach such for instance the Ph evolution data presented in Abraham & al. (2003), but our aim here is just to show how this method can work in practice on functional data sets such as the phoneme one described before. To see that, we have split arbitrarily our set of 2000 phoneme curves into two parts:

- A training sample (of size 750), composed of 150 curves of each group;
- A testing sample (of size 1250), composed of the remaining curves.

The training sample has been used to select the parameters of the estimate (that is to select the semi-metric and the bandwidth). Note that, in this example, the optimal selected semi-metric did not belong to the family (8.3.1) but to some special semi-metric constructed by combining PCA and PLS ideas. The details concerning the application of the method to these phonetic data (how to choose the bandwidth? how to construct precisely the semi-metric? ...) are given in Ferraty & Vieu (2003a), as well as comparisons of the results of the proposed method with classical alternative techniques.

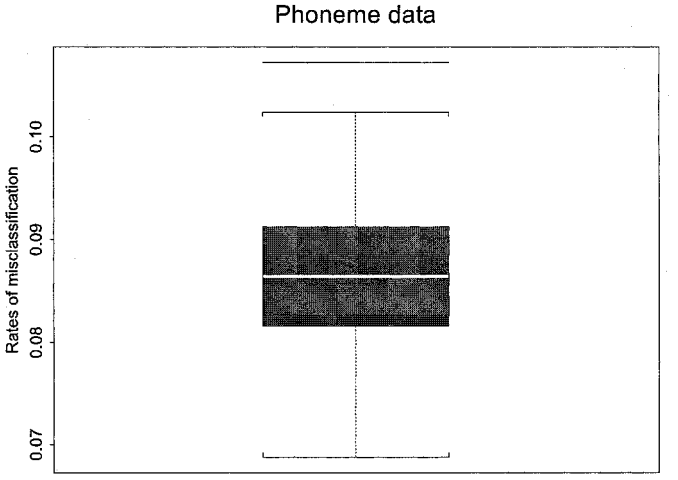


Figure 8.7: Functional nonparametric discrimination in action on 50 test samples of phoneme curves

We have repeated 50 times the random splitting into training and test samples, as described before, and we present the corresponding boxplot for the missclassification rates in Figure 8.7.

8.6 Nonparametric Functional Time Series in Action on Economical Data

Let us now look at the time series described in Example 3 above. The original data (see Figures 8.3 and 8.4) are the differencied logarithms of the monthly electrical consumption

$$(X_i, i = 0, \ldots 337),$$

that can be considered as discrete observations of a continuous time series $X_t, t \in (0, 337)$. The Functional Nonparametric approach consists in using as explanatory variable the whole continuous times series over some fixed period, that has been selected here to be a year. That means, that we consider that we have at hand the sample of functional data:

$$(T_j, j = 1, \ldots 28) \text{ where } T_j = \{X_t, t \in (12j - 11, 12j)\},$$

which has been presented previously in Figure 8.6. The goal is to predict the next value of the time series, by using the previous continuous yearly data. In other words, our model is the following

$$X_{i+1} = R(\{X_t, t \in (i - 11, i)\}) + error,$$

or equivalently

$$X_{12j+1} = R(T_j) + error,$$

and the statistical target is the functional operator R. To estimate this operator, the idea is to use the Functional Nonparametric Prediction technique that consists, if the last observed yearly trajectory is denoted by τ, to estimate $R(\tau)$ by

$$\hat{R}_n(\tau) = \sum_{i=1}^{n} K_{n,i}(\tau) X_{i+1}, \tag{8.6.1}$$

where

$$K_{n,i}(\tau) = \frac{K\left(h^{-1} d(T_i; \tau)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(T_i; \tau)\right)}.$$

In this definition h is a positive number, K is a weight function chosen to be:

$$K(u) = \frac{3}{2}(1 - u^2), \text{ if } u \in [0, 1], \text{ and 0 elsewhere.} \tag{8.6.2}$$

In the estimator $\hat{R}_n$, the notion of proximity is controlled by a semi-metric $d(.;.)$. There is a recent mathematical asymptotic support for using such a functional statistical method that was given by Ferraty & Vieu (2003c) and Ferraty & al. (2003), but it is not our purpose to describe it here. Also, many other data sets can be treated by such kind of methods, and this has been

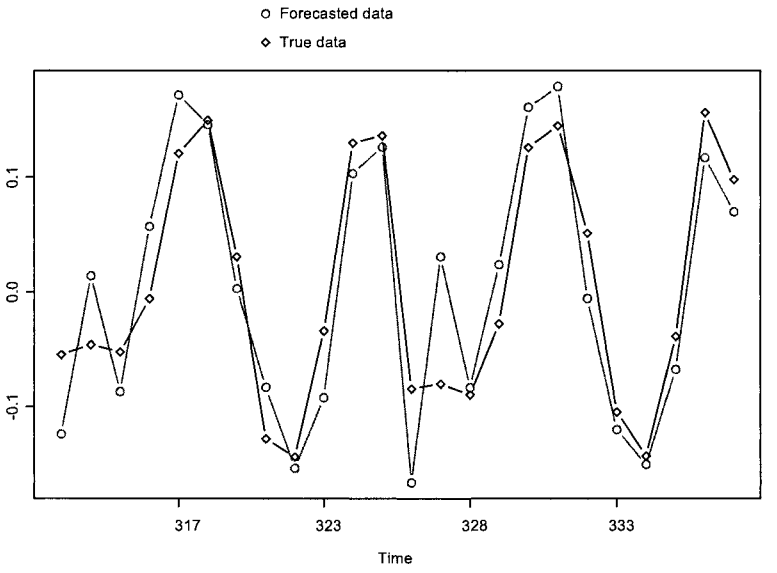


Figure 8.8: Functional nonparametric time series prediction in action on electrical consumption data.

done successfully for instance for climatic data by Cardot & al. (1999), for pollution data by Fernández de Castro & al. (2003) and for environmental data by Damon & Guillas (2002). But, let us concentrate ourselves on how this procedure works on our data. To show the accuracy of the method to deal with such a functional data set, we have left-out of the data set both last periods (that is $X_i, i = 314, \ldots 337$), and we have used the estimate (8.6.1) (with $n = 26$) to get the predicted corresponding values

$$\hat{X}_i = \hat{R}_n(X_i),\ i = 314, \ldots 337.$$

The behaviour of the method is illustrated through Figure 8.8 that presents the predicted values $\hat{X}_i\, i = 314, \ldots 337$ together with the real ones.

Along this study the bandwidth was selected by cross-validation and the semi-metric by Functional PCA techniques introduced by Dauxois & al. (1982). These details are not presented here but the complete study can be found in Ferraty & al. (2003).

8.7 Conclusions

As we have seen along the three examples, the recent functional nonparametric methods can be useful tools in many applied problems for which functional explanatory variables (curves) are involved. Even if for obvious reasons the methods and the examples treated here are very quickly presented, the reader is encouraged in reporting his attention to the companion works cited all along this article and which will allow to have access to methodological supports, to complete bibliography, and to computational issues related with each data set (including comparisons with alternative techniques).

Of course, at this stage of the investigations and because of the novelty of this field of Statistics, many things remain to do. It seems to us that, for a practical point of view, the most appealing opening problems are certainly those related with the choice of the parameters of the estimates. In other words, interesting open problems can be summarized as: can we think in some data-driven bandwidth selection? can we think in some data-driven semi-metric choice? Our guesses is that the first point could certainly be addressed by suitable modification of the usual bandwidth selection rules existing in the un-functional case, but the second point should be much harder to automatize completely.

In any cases, we believe that in the next future this field of Statistics will receive particular attention as well because of the large possibilities of application as for the deep mathematical joint problems.

Bibliography

Abraham, C., Cornillon, P.A., Matzner-Lober, E. & Molinari, N. (2003)Unsupervised curves clustering using B-splines. Preprint.

Akritas, M. & Politis, D. (ed.) (2003)Recent advances and trends in nonparametric statistics. Elsevier, In print.

Aneiros Pérez, G., Cardot, H., Estévez Pérez, G. & Vieu, P. (2003)Maximum ozone concentration forecasting by functional nonparametric approaches. Proceedings of TIES 2002 meeting, Genova Italia.

Besse, P., Cardot, H. & Stephenson, D. (2000)Autoregressive forecasting of some functional climatic variations. *Scandinavian Journal of Statistics* **27**, 673–687.

Bosq, D. (2000)Linear Processes in Function Spaces: Theory and Applications. *Lecture Notes in Statistics* **149**, Springer-Verlag, New-York.

Cardot, H., Ferraty, F. & Sarda, P. (1999)Spline Estimators for the Functional Linear Model. Technical Report **1999–1**, UBIA INRA Toulouse.

Damon, J. & Guillas, S. (2002)The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics* **13**, 759–774.

Dauxois, J., Pousse, A. & Romain, Y. (1982)Asymptotic theory for the principal component analysis of a random vector function: some application to statistical inference. *J. Multivariate Anal.* **12**, 136–154.

Fernández de Castro, B., Guillas, S. & González Manteiga, W. (2003)Functional samples and bootstrap for the prediction of SO2 levels. Preprint.

Ferraty, F. (2003)Modélisation statistique pour variables aléatoires fonctionnelles: théorie et application (in french), HDR, Toulouse III. Free access through http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html.

Ferraty, F., Goia, A. & Vieu, Ph. (2003)Functional model for time series: a fractal approach to dimension reduction. *Test* **11**, 317–344.

Ferraty, F., Goia, A. & Vieu, Ph. (2002)Statistica Funzionale: modelli di regressione non-parametrici (in italian). Collana Statistica, Franco-Angeli, Milano.

Ferraty, F., Núñez Antón, V. & Vieu, Ph. (2001)Regresión no paramétrica: desde la dimensión uno hasta la dimensión infinita (in spanish), Serv. Edit. Univers. País Vasco, Bilbao.

Ferraty, F. & Vieu, Ph. (2001)Statistique Fonctionnelle: Modèles de régression pour variables aléatoires uni, multi et infiniment dimensionées (in french). Coursebook at Univ. P. Sabatier, Toulouse, France, Free access on line at http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html.

Ferraty, F. & Vieu, Ph. (2003)The functional nonparametric model and application to spectrometric data. *Computational Statistics* **17**, 545–564.

Ferraty, F. & Vieu, Ph. (2003)Curves discrimination: a nonparametric functional approach, Computational Statistics & Data Analysis, In print.

Ferraty, F. & Vieu, Ph. (2003)Nonparametric models for functional data, with applications in regression, time series prediction and curves discrimination. *J. Nonparametric Statistics*, In print.

Ferraty, F. & Vieu, Ph. (2003)Functional nonparametric statistics: a double infinite dimensional framework, In *Recent advances and trends in nonparametric statistics, Ed. M. Akritas adn D. Politis.* Elsevier, In print.

Goia, A. (2003)Contribution à l'étude des modèles de régression pour variables aléatoires fonctionnelles (in french), PhD Toulouse III. Accessible on request through http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html.

Härdle, W. (1990)Applied nonparametric regression. Cambridge Univ. Press, UK.

Hastie, T., Buja, A. & Tibshirani, R. (1995)Penalized discriminant analysis. *Ann. Statist.* **13** 435–475.

Leurgans, S.E., Moyeed, R.A. & Silverman, B.W. (1993)Canonical correlation analysis when the data are curves. *J. R. Statist. Soc. B* **55**, 725–740.

Niang, S. Sur l'estimation de la densité en dimension infinie: application aux diffusions (in french), PhD Paris VI.

Ramsay, J. & Silverman, B.W. (1997)Functional Data Analysis. Springer-Verlag, New-York.

Ramsay, J. & Silverman, B. (2002)Applied functional data analysis: Methods and case studies. Springer-Verlag, New York.

Schimek, M. (ed.) (2000)Smoothing and regression: Approaches, Computation, and Application. Wiley Series in Probability and Statistics, Wiley, New-York.

Staph, (group) (2003)Proceedings of the working group in Functional and Operatorial Statistics. **IV**, free access on line at http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html.

9 Productivity Effects of IT-Outsourcing: Semiparametric Evidence for German Companies

Irene Bertschek[1] and Marlene Müller[2]

[1] ZEW, Centre for European Economic Research, P.O. Box 103443, D–68034 Mannheim, Germany
[2] Fraunhofer Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1, D–67663 Kaiserslautern, Germany

Summary

This paper analyzes the relationship between IT-outsourcing and labor productivity of 1142 firms from German manufacturing and service industries surveyed in 2000. An endogenous switching regression model takes into account that firms might follow different productivity regimes depending on whether or not they source out IT-tasks. Two semiparametric approaches are presented and applied to the data. They allow the outsourcing decision to nonlinearly depend on firm size. The empirical results show that firms with IT-outsourcing do not differ significantly from non-outsourcing firms with respect to the partial production elasticities of the input factors labor, IT-investment and non-IT-investment. However, firms without IT-outsourcing turn out to produce more than those sourcing out.

Keywords: Information technology, IT-outsourcing, labor productivity, endogenous switching, semiparametric, partial linear

9.1 Introduction

During the last decade, information technology (IT) has become a well established working tool for many employees. As a consequence, the impact of IT on labor productivity is a broadly discussed topic in management sciences and economics. Several studies find empirical evidence for positive productivity effects of IT at the firm-level, for example Brynjolfsson & Hitt (2000, 1996), Lichtenberg (1995), Greenan & Mairesse (2000), Licht & Moch (1999), and Hempell (2005).

Due to the increasing IT-intensity, especially in the services industries, more and more firms are outsourcing their IT-tasks either completely or at least to some extent. The most common reasons for IT-outsourcing are that firms prefer to concentrate on their core competencies, that they attempt to reduce costs or that they have problems to find qualified personnel for these tasks (see for example Henkel & Kaiser (2002) for further details). Heshmati (2003) gives a comprehensive overview on the effects of outsourcing.

The aim of this paper is to analyze the relationship between IT-outsourcing and firms' labor productivity using a data set of 1142 German firms from the manufacturing industry and from selected service sectors. We take into account the potential simultaneity of labor productivity and IT-outsourcing by estimating an endogenous switching regression model. IT-outsourcing does not only have a unidirectional relationship with productivity, but the decision for IT-outsourcing may also depend on the firms' expectations about the productivity gains from outsourcing. Moreover, a switching regression model allows IT-outsourcing to change the entire set of partial production elasticities (see for example Bertschek & Kaiser, 2004, for an application to organizational changes).

The econometric approach that we use in this paper is a switching regression model that considers two regimes: to source out IT tasks or not. A third equation, the selection equation, models the decision for one of the two regimes. Our interest in studying the impact of firm size on IT-outsourcing does therefore imply to consider a nonparametric component in the selection equation. A computationally simple approach to implement this is a two-step procedure, which first estimates a semiparametric model for the selection equation and uses this to compute correcting factors for the regime equations. We will compare this two-step approach with a full information semiparametric profile likelihood algorithm.

The results of the applied parametric and semiparametric methods imply that IT-outsourcing does not significantly change the partial output elasticities of the production factors labor, IT-investment and non-IT-investment. However, the efficiency parameter measured by the constant term turns out to be significantly larger in the regime without IT-outsourcing. A possible interpretation is that outsourcing may be involved with high coordination costs between the outsourcing firm and the subcontractor and may thus make business processes less efficient. Comparing the two semiparametric approaches, we find that the semiparametric profile likelihood method seems to perform better than the semiparametric two-step procedure.

The structure of the paper is as follows: Section 9.2 outlines the theoretical considerations of the economic model. The data are described in Section 9.3. Section 9.4 introduces the semiparametric modification of the switching re-

gression model. In Section 9.5 we present and discuss the empirical results. Finally, Section 9.6 concludes.

9.2 Theoretical Considerations

Based on the model used in Bertschek & Kaiser (2004), we assume that firm i produces according to a Cobb–Douglas production technology. Output y_i is a function of IT-capital IT_i, non-IT-capital K_i, and labor input L_i:

$$y_i = A_i \, IT_i^{\alpha_1} \, K_i^{\alpha_2} \, L_i^{\alpha_3}. \tag{9.2.1}$$

The scalar A_i represents a parameter of production efficiency that shifts the isoquants of the Cobb–Douglas production function in parallel to the origin. The exponents α_1, α_2 and α_3 denote the output elasticities with respect to IT-capital, non-IT-capital and labor, respectively. Taking logarithms leads to

$$\ln(y_i) = \ln(A_i) + \alpha_1 \ln(IT_i) + \alpha_2 \ln(K_i) + \alpha_3 \ln(L_i). \tag{9.2.2}$$

Labor productivity, i.e. output per worker, is then given by:

$$\ln\left(\frac{y_i}{L_i}\right) = \ln(A_i) + \alpha_1 \ln(IT_i) + \alpha_2 \ln(K_i) + (\alpha_3 - 1) \ln(L_i). \tag{9.2.3}$$

If a firm sources out IT-tasks, its labor productivity is

$$\begin{aligned} \ln\left(\frac{y_i}{L_i}\right)_{out} &= \ln(A_{i,out}) + \alpha_{1,out} \ln(IT_i) \\ &\quad + \alpha_{2,out} \ln(K_i) + (\alpha_{3,out} - 1) \ln(L_i) \\ &= x_i^\top \beta_{out}. \end{aligned} \tag{9.2.4}$$

For firms without IT-outsourcing, labor productivity is

$$\begin{aligned} \ln\left(\frac{y_i}{L_i}\right)_{nout} &= \ln(A_{i,nout}) + \alpha_{1,nout} \ln(ICT_i) \\ &\quad + \alpha_{2,nout} \ln(K_i) + (\alpha_{3,nout} - 1) \ln(L_i) \\ &= x_i^\top \beta_{nout}, \end{aligned} \tag{9.2.5}$$

where the subscripts *out* and *nout* denote the two productivity regimes with and without IT-outsourcing, respectively. Firms decide to source out IT-tasks if the productivity gain from outsourcing is larger than the costs per worker involved with outsourcing C_i. Thus, the latent variable

$$\begin{aligned} I_i^* &= a \left(\ln\left(\frac{y_i}{L_i}\right)_{out} - \ln\left(\frac{y_i}{L_i}\right)_{nout} \right) - C_i \\ &= z_i^\top \beta_{sel} \end{aligned} \tag{9.2.6}$$

represents the difference between the productivity gains and the costs arising from IT-outsourcing, where a represents the effect of the productivity gains from IT-outsourcing on the decision to source out. If $a = 0$, the outsourcing decision is unaffected by the productivity differences. The selection mechanism for observing IT-outsourcing then is

$$I_i = \begin{cases} 1 & \text{if } I_i^* > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{9.2.7}$$

Since direct cost effects from IT-outsourcing generally cannot be identified in a straightforward way, we assume that the following factors are likely to influence the costs of IT-outsourcing and thus might affect a firm's decision to source out IT-tasks: The costs of IT-outsourcing are likely to be lower for exporting firms since these firms are used to adjusting quickly to changes in the international market environment. The same argument holds for firms with a subsidiary in a foreign country. Firms that belong to a group of companies generally have more financial resources than others and thus might have their own IT-division. On the other hand, these firms may have the possibility to source out IT-tasks within the company group. For older firms, the costs of implementing IT and reorganizing the production process is probably more expensive than for younger firms that already started with a high level of IT-intensity. A firm's problems to find appropriate IT-specialists might on the one hand indicate that this firm prefers to do IT tasks inhouse. On the other hand, an IT skill shortage may force these companies to source out certain IT-tasks. Last but not least, the probability of IT-outsourcing is likely to differ across industries.

It is not clear a priori whether small or large firms are more likely to source out. Large firms might source out whole departments as did for example the Deutsche Bank that sourced out its IT-department to IBM (Lamberti 2003). Small firms, in contrast, will probably source out single tasks rather than whole departments. In order to capture the potential nonlinear impact of firm size, a semiparametric estimation procedure will be applied.

9.3 The Data

The data result from a CATI-survey (computer-aided telephone interview) based on a stratified random sample of about 11,000 German firms. The sample was stratified by sector, size class and region, i.e. West and East Germany. Only firms with at least five employees were included in the survey, thereof 50% in the manufacturing industry and 50% in the service sector. The source data set originates from Creditreform, the largest German credit

rating agency.[1] The survey was conducted in the year 2000. About 4400 enterprises participated in the survey, which corresponds to a response rate of approximately 43%. After performing consistency checks and due to item non-response concerning the variables that were included in the empirical model (see below), a sample of 1351 firms forms the basis for the empirical analysis. In order to estimate production functions for the two productivity regimes with and without IT-outsourcing, we have to measure labor productivity, IT-capital and non-IT-capital. Labor productivity is calculated as the ratio of total sales to the total number of employees. Since no information about the two capital variables is available in our survey data, non-IT-capital is measured as investment in physical capital and IT-capital is proxied by IT-investment. Proxying IT-capital by IT-investment does not appear as a severe shortcoming since IT depreciates extremely quickly (Dewan & Min 1997). With regard to the empirical proxy for non-IT-capital, it is important to note that a capital stock could theoretically be calculated using the perpetual inventory method. However, our analysis is based on a cross-sectional data set and thus, we can only observe investment in physical capital for one period.

Table 9.1 displays the descriptive statistics of the variables used in the estimation of labor productivity: IT-investment, non-IT-investment (both in 1,000 DM), total employment and productivity (sales per worker). All quantitative numbers refer to the year 1999. The standard deviations of non-IT-investment and the number of employees are quite large, since small retailers as well as the largest German manufacturing companies are both included in our sample.[2] For the estimations, the quantitative variables L, INV and IT are taken in logarithms.

In the interview, the firms were asked whether or not they source out certain IT-tasks, either partially or completely. From this information a dummy is constructed taking the value one if a firm *completely* sources out IT-tasks and the value zero otherwise. About 68 percent of the firms in the sample have completely sourced out some or all of their IT-tasks.

In the empirical implementation of our model, the variables that are supposed to influence the costs involved with IT-outsourcing are measured as follows: the age of a firm is captured by two dummy-variables, the first one taking the value one if the firm is three years old or younger, the second variable taking the value one if the age is between four and seven years. This categorization is plausible since empirical studies for instance by Prantl (2001) show that hazard functions of young firms reach a first local maximum approximately

[1] As Germany's largest credit rating agency, Creditreform has the most comprehensive database of German firms at its disposal. Creditreform provides data on German firms to the ZEW (Centre for European Economic Research) for research purposes.

[2] For further details on the data, see Bertschek, Fryges & Kaiser (2004).

Table 9.1: Descriptive statistics

Variable		Mean	Std. Dev.	Min.	Max.
PROD†	productivity	0.497	1.321	0.007	23.65
L	number of employees	580.896	7,293.496	5	225,000
INV	non-IT-investment	46,768.77	1,359,455	1.000	49,900,000
IT	IT-investment	1,023.98	6,554.279	1.2	120,001
EXP_VH	export quota	15.052	22.421	0	100
D_OUTV‡	IT-tasks completely sourced out	0.677	0.468	0	1
D_KONZ	firm belongs to group of companies	0.318	0.466	0	1
D_STAND	foreign subsidiary	0.175	0.381	0	1
D_AGE1	age <4 years	0.074	0.262	0	1
D_AGE2	age 4–7 years	0.160	0.367	0	1
D_ITOS	IT skill shortage	0.161	0.367	0	1
D_REG	East Germany	0.251	0.434	0	1
D_BR1	manufacturing industry w/o ICT	0.476	0.500	0	1
D_BR2	distributive services	0.193	0.395	0	1
D_BR3	banking & insurance	0.046	0.209	0	1
D_BR4	technical services	0.076	0.264	0	1
D_BR5	other business-related services	0.057	0.232	0	1
D_BR6	IT w/o retail trade	0.153	0.360	0	1
Obs.		1351			

† Output per employee (total sales per year per employee in 1,000 DM), where sales means balance-sheet total for banks and sum insured for insurance companies.

‡ "D"stands for Dummy.

three years after formation and a second local maximum after approximately seven years. Having survived seven years, the hazard rates stay on a comparably low level such that those firms can be regarded as established or "old" firms. A firm's export activity is captured by the share of sales obtained by exports (export quota). Further dummy variables are constructed measuring whether or not a company has a foreign subsidiary and whether a firm is part of a group of companies. 32% of the firms in our sample belong to a larger group of companies. A further dummy reflects whether a firm has problems to find appropriate IT-specialists which is the case for 16 percent of the firms. Industry-specific characteristics are captured by five industry dummies. A regional dummy taking the value one if a firm is located in Eastern Germany

controls for the fact that East German companies produce with a generally lower productivity than West German firms.

9.4 Switching Regression Model and Nonparametric Components

In what follows we present the model of Section 9.2 in a slightly more formal way. Introduce the shortcuts

$$Y_1^* = \ln\left(\frac{y}{L}\right)_{out} \quad \text{and} \quad Y_2^* = \ln\left(\frac{y}{L}\right)_{nout}.$$

Depending on whether IT-tasks are sourced out or not, one of the following two regression models applies:

$$Y_1^* = x_1^\top \beta_1 + \varepsilon_1 \tag{9.4.1}$$
$$Y_2^* = x_2^\top \beta_2 + \varepsilon_2 \tag{9.4.2}$$

These two equations correspond to equations (9.2.4) and (9.2.5) when in our specific example we set $\beta_1 = \beta_{out}$, $\beta_2 = \beta_{nout}$ and $x_1 = x_2 = x$. We further assume that the decision to source out is explained by a third regression equation

$$I^* = x_3^\top \beta_3 + \varepsilon_3, \tag{9.4.3}$$

where in our example we have $\beta_3 = \beta_{sel}$ and $x_3 = z$. Recall, that we only observe whether or not a firm does completely source out IT-tasks: $I = 1$ if $I^* > 0$ and $I = 0$ otherwise. The model to consider is thus a switching regression model (Maddala 1983, Sections 8.3 and 9.7) which is defined by

$$Y = \begin{cases} Y_1^* & I = 1\,, \\ Y_2^* & I = 0\,. \end{cases} \tag{9.4.4}$$

If we assume that the error terms of (9.4.1)–(9.4.3) have a joint normal distribution, we can write

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_0\sigma_1\sigma_2 & \rho_1\sigma_1\sigma_3 \\ \rho_0\sigma_1\sigma_2 & \sigma_2^2 & \rho_2\sigma_2\sigma_3 \\ \rho_1\sigma_1\sigma_3 & \rho_2\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix} \right). \tag{9.4.5}$$

We remark that, as in a probit model, the variance parameter of ε_3 can only be identified up to a constant, hence we set it to

$$\sigma_3 = 1.$$

We further assume that $\rho_1 \neq 0$ and $\rho_2 \neq 0$ which leads to endogenous switching. Since we cannot observe Y_1^* and Y_2^*, simultaneously, the parameter ρ_0

cannot be estimated (Maddala 1983, Section 9.7). However, ρ_0 is not needed for establishing the likelihood function.

Let Y_i, x_{ij}, and I_i for $i = 1, \ldots, n$ and $j = 1, 2, 3$ denote the individual realizations of the variables Y, x_j and I. We define

$$\eta_{1i} = x_{i1}^\top \beta_1, \quad \eta_{i2} = x_{i2}^\top \beta_2, \quad \eta_{i3} = x_{i3}^\top \beta_3,$$

and

$$\sigma = \begin{pmatrix} \sigma_1 \\ \sigma_2 \end{pmatrix}, \quad \rho = \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix}.$$

This allows to define the log-likelihood of the switching regression model as

$$\ell = \sum_{i=1}^{n} \ell_i(\, \eta_{i1},\, \eta_{i2},\, \eta_{i3},\, \sigma,\, \rho\,), \tag{9.4.6}$$

where we use the notation ℓ_i to indicate that the individual log-likelihood terms do also depend on Y_i and I_i. More precisely, the function ℓ_i is given by

$$\begin{aligned} &\ell_i(\, \eta_{i1},\, \eta_{i2},\, \eta_{i3},\, \sigma,\, \rho\,) \qquad (9.4.7) \\ &= I_i \left\{ \ln \Phi \left(\frac{\eta_{i3} + \rho_1 \frac{(Y_i - \eta_{i1})}{\sigma_1}}{\sqrt{1-\rho_1^2}} \right) - \frac{(Y_i - \eta_{i1})^2}{2{\sigma_1}^2} - \ln\left(\sqrt{2\pi}\sigma_1\right) \right\} \\ &\quad + (1 - I_i) \left\{ \ln \Phi \left(\frac{-\eta_{i3} - \rho_2 \frac{(Y_i - \eta_{i2})}{\sigma_2}}{\sqrt{1-\rho_2^2}} \right) - \frac{(Y_i - \eta_{i2})^2}{2{\sigma_2}^2} - \ln\left(\sqrt{2\pi}\sigma_2\right) \right\}. \end{aligned}$$

Note that here and in the following we denote by Φ the cumulative distribution function of the univariate Gaussian distribution.

In the special case, that the correlations ρ_1 and ρ_2 equal zero (exogenous switching), the third equation (9.4.3) is independent of equations (9.4.1)–(9.4.2). A probit fit for the selection equation (9.4.3) and OLS fits for the level equations (9.4.1)–(9.4.2) could thus be applied to I and Y. In the general case considered here (endogenous switching), we have to take these correlations into account and two approaches are possible:

- A two-step estimation which employs Heckman's (1976) idea of a probit estimation for estimating β_3 and subsequent OLS estimations in the subsamples $I_i = 1$ and $I_i = 0$. In this case, the vectors of explanatory variables need to be supplemented by the respective inverse Mill's ratios.

- A full information maximum-likelihood (FIML) estimator which maximizes the log-likelihood given in (9.4.6).

We will present the two concepts for a switching regression model with a semiparametric modification of the selection equation (9.4.3). Assume that the vector x_3 of explanatory variables can be split into vectors u and t and that the selection condition is explained by the following partial linear model:

$$I^* = u^\top \gamma + m(t) + \varepsilon_3 \,, \tag{9.4.8}$$

where t is a continuous variable or vector of variables and m is a smooth function. We remark that the same partial linear assumption could be made for Y_1^* and Y_2^*. We skip this additional modification since our particular interest is to find out whether labor input $\ln(L)$ as a measure of firm size is nonlinearly related to the decision for IT-outsourcing. Consequently, we will consider

$$t = \ln(L)$$

whereas the remaining factors from the selection equation (9.2.6) are contained in vector u.

9.4.1 Two-Step Estimation

The construction of the semiparametric two-step estimator is a straightforward generalization of the parametric two-step approach. The procedure can be summarized as follows:

Semiparametric Two-Step Switching Regression

- *First step:*
  Estimate the selection equation using a partial linear probit model

  $$P(I = 1) = \Phi\{u^\top \gamma + m(t)\}.$$

- *Second step:*
  Using the results of the first step, compute the Mill's ratios

  $$\lambda_1 = \frac{\Phi'\{u^\top \gamma + m(t)\}}{\Phi\{u^\top \gamma + m(t)\}} \quad \text{and} \quad \lambda_2 = \frac{\Phi'\{-u^\top \gamma - m(t)\}}{\Phi\{-u^\top \gamma - m(t)\}} \,,$$

  and estimate the regression equations

  $$\begin{aligned} E(Y|I=1) &= x_1^\top \beta_1 + \rho_1 \sigma_1 \lambda_1 = x_1^\top \beta_1 + \beta_{\lambda_1} \lambda_1 \,, \\ E(Y|I=0) &= x_2^\top \beta_2 + \rho_2 \sigma_2 \lambda_2 = x_2^\top \beta_2 + \beta_{\lambda_2} \lambda_2 \,. \end{aligned}$$

  The parameters σ_j and ρ_j can be estimated through

  $$\widehat{\sigma}_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(\widehat{\varepsilon}_{ij}^2 + \widehat{\delta}_{ij} \widehat{\beta}_{\lambda_j}^2 \right), \quad \widehat{\rho}_j = \frac{\widehat{\beta}_{\lambda_j}}{\widehat{\sigma}_j} \,,$$

where n_j are the sizes of the subsamples $I_i = 1$ and $I_i = 0$, $\widehat{\varepsilon}_{ij}$ is an estimate of the i-th residual of the corresponding regression equation and $\widehat{\delta}_{ij}$ is the individual estimate of $\delta_j = \lambda_j\{\lambda_j + u^\top\gamma + m(t)\}$.

Using this approach it is not possible to obtain covariance estimates for $\widehat{\sigma}_j$ and $\widehat{\rho}_j$ separately. It is however possible to estimate the covariance matrices

$$\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_{\lambda_j}) = \widehat{\sigma}_j^2(\boldsymbol{X}_j^\top \boldsymbol{X}_j)^{-1}\left\{\boldsymbol{X}_j^\top(\boldsymbol{I} - \widehat{\rho}_j^2\Delta_j)\boldsymbol{X}_j + \boldsymbol{Q}_j\right\}(\boldsymbol{X}_j^\top \boldsymbol{X}_j)^{-1}.$$

Here we use $\boldsymbol{X}_j$ to denote the design matrix consisting of the observations of x_j and λ_j, $\boldsymbol{I}$ for the identity matrix, $\Delta_j = \mathrm{diag}(\delta_{1j}, \ldots, \delta_{nj})$, and

$$\boldsymbol{Q}_j = \widehat{\rho}_j^2\,(\boldsymbol{X}_j^\top \Delta_j\, \boldsymbol{U}_j)\,\mathrm{Cov}(\widehat{\gamma})\,(\boldsymbol{U}_j^\top \Delta_j\, \boldsymbol{X}_j)\,,$$

with $\boldsymbol{U}_j$ the design matrices from the observations of u in the subsamples $I_i = 1$ and $I_i = 0$.

The partial linear probit model in the first step is a special case of the generalized partial linear model (GPLM), see for example Müller (2001) and Severini & Staniswalis (1994) for further details. The GPLM algorithm provides estimates of γ, $\mathrm{Cov}(\widehat{\gamma})$ and the nonparametric function m. For details on the second step we refer to Maddala (1983, Section 8.3), Greene (2000, Chapter 20) and the references therein.

9.4.2 Profile Likelihood Estimator

Let us recall the parametric FIML estimator which aims at solving

$$\mathcal{D}_\ell = \mathrm{vec}\left(\frac{\partial \ell}{\partial \beta_1}, \frac{\partial \ell}{\partial \beta_2}, \frac{\partial \ell}{\partial \beta_3}, \frac{\partial \ell}{\partial \sigma}, \frac{\partial \ell}{\partial \rho}\right) = 0\,. \tag{9.4.9}$$

We want to stress in particular the fact that for $j = 1, 2, 3$,

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \eta_{ij}}(\,\eta_{i1},\, \eta_{i2},\, \eta_{i3},\, \sigma,\, \rho\,)\, x_{ij}\,. \tag{9.4.10}$$

This means, the gradient of ℓ_i with respect to β_j is given by the derivative of ℓ_i with respect to its jth argument multiplied by the vector x_{ij}. This is a property which also holds for simpler models as logit and probit. We will in particular exploit this fact for deriving the semiparametric extension to the parametric switching regression.

Equation (9.4.9) is nonlinear in the parameters $\theta = \mathrm{vec}(\beta_1, \beta_2, \beta_3, \sigma, \rho)$ and therefore must be solved by an iterative procedure. A Newton–Raphson algorithm, for example, uses iteration steps of the form $\theta^{new} = \theta^{old} - \mathcal{H}_\ell^{-1}\mathcal{D}_\ell$ with

$\mathcal{H}_\ell$ denoting the Hessian of ℓ. Alternatively, one can use a BFGS optimization which requires only the gradient $\mathcal{D}_\ell$.

The semiparametric profile likelihood method considered in Severini & Wong (1992), Staniswalis & Thall (2001) is based on the fact, that the conditional distribution of Y given u and t is still parametric. Their approach is to first keep the parameter $\vartheta = \mathrm{vec}(\beta_1, \beta_2, \gamma, \sigma, \rho)$ fix and to estimate the nonparametric function values $m_\vartheta(t)$ in dependence of this fixed ϑ. The resulting estimate $\widehat{m}_\vartheta$ is then used to construct the profile likelihood for ϑ. As a consequence of the profile likelihood, the estimated parameters are typically $\sqrt{n}$-consistent, asymptotically normal and asymptotically efficient. Moreover the estimator $\widehat{m}(\bullet) = \widehat{m}_{\widehat{\vartheta}}(\bullet)$ can be used as an estimator for the nonparametric function m (Severini & Staniswalis 1994).

Due to the nonlinear structure of the log-likelihood function, the algorithm to estimate the parameters and the nonparametric function will be iterative. Recall the parametric log-likelihood (9.4.6). A localized version of this is given by

$$\ell^{\,local}(m_\vartheta(t)) = \sum_{i=1}^{n} \ell_i(\,\eta_{i1},\, \eta_{i2},\, u_i^T\gamma + m_\vartheta(t),\, \sigma,\, \rho\,)\, K_h(t - t_i)\,. \qquad (9.4.11)$$

This function is maximized to estimate the smooth function $m_\vartheta(t)$ at the point t. The local weights $K_h(t - t_i)$ are kernel weights with K denoting a (multidimensional) kernel function and h a bandwidth vector. Using the solution $\widehat{m}_\vartheta$ of (9.4.11) we derive the parametric profile log-likelihood

$$\ell^{\,profile}(\vartheta) = \sum_{i=1}^{n} \ell_i(\,\eta_{i1},\, \eta_{i2},\, u_i^T\gamma + m_\vartheta(t_i),\, \sigma,\, \rho\,)\,. \qquad (9.4.12)$$

This profile log-likelihood will then be optimized in order to obtain an estimate for ϑ.

For the derivation of the algorithm we follow the presentation of the profile likelihood estimator for generalized partial linear models (such as partial linear logit and probit models) in Müller (2001). The maximization of the local likelihood (9.4.11) with respect to the nonparametric component requires to solve

$$\sum_{i=1}^{n} \frac{\partial}{\partial \eta_3} \ell_i(\eta_{i1}, \eta_{i2}, u_i^T\gamma + \mu, \sigma, \rho)\, K_h(t_i - t) = 0 \qquad (9.4.13)$$

with respect to μ for each value of t. In other words, we obtain solutions $m_j = m_\vartheta(t_j)$ which fulfill

$$\sum_{i=1}^{n} \frac{\partial}{\partial \eta_3} \ell_i(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_j, \sigma, \rho)\, K_h(t_i - t_j) = 0 \qquad (9.4.14)$$

for all observations $t_1, \ldots, t_n$. Now, using the profile likelihood (9.4.12) , we have to deal with

$$\sum_{i=1}^{n} \tfrac{\partial}{\partial\beta_1}\ell\left(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_\vartheta(t_i), \sigma, \rho\right) \tfrac{\partial}{\partial\beta_1} m_\vartheta(t_i) = 0\,, \qquad (9.4.15)$$

$$\sum_{i=1}^{n} \tfrac{\partial}{\partial\beta_2}\ell\left(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_\vartheta(t_i), \sigma, \rho\right) \tfrac{\partial}{\partial\beta_2} m_\vartheta(t_i) = 0\,, \qquad (9.4.16)$$

$$\sum_{i=1}^{n} \tfrac{\partial}{\partial\beta_3}\ell\left(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_\vartheta(t_i), \sigma, \rho\right) \left(u_i + \tfrac{\partial}{\partial\beta_3} m_\vartheta(t_i)\right) = 0\,, \qquad (9.4.17)$$

$$\sum_{i=1}^{n} \tfrac{\partial}{\partial\sigma}\ell\left(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_\vartheta(t_i), \sigma, \rho\right) \tfrac{\partial}{\partial\sigma} m_\vartheta(t_i) = 0\,, \qquad (9.4.18)$$

$$\sum_{i=1}^{n} \tfrac{\partial}{\partial\rho}\ell\left(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_\vartheta(t_i), \sigma, \rho\right) \tfrac{\partial}{\partial\rho} m_\vartheta(t_i) = 0\,. \qquad (9.4.19)$$

Taking derivatives of (9.4.13) with respect to the components of ϑ shows that

$$\frac{\partial}{\partial\beta_3} m_\vartheta(t_i) = -\frac{\sum_{i=1}^{n} \frac{\partial^2}{\partial\beta_3^2}\ell_i(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_j, \sigma, \rho)\, K_h(t_i - t_j) u_i}{\sum_{i=1}^{n} \frac{\partial^2}{\partial\beta_3^2}\ell_i(\eta_{i1}, \eta_{i2}, u_i^T\gamma + m_j, \sigma, \rho)\, K_h(t_i - t_j)}\,, \qquad (9.4.20)$$

and $\dfrac{\partial}{\partial\beta_1} m_\vartheta(t) = 0, \quad \dfrac{\partial}{\partial\beta_2} m_\vartheta(t) = 0, \quad \dfrac{\partial}{\partial\sigma} m_\vartheta(t) = 0, \quad \dfrac{\partial}{\partial\rho} m_\vartheta(t) = 0.$

We introduce the shortcuts $\ell_i' = \frac{\partial}{\partial\beta_3}\ell_i\,, \quad \ell_i'' = \frac{\partial^2}{\partial\beta_3^2}\ell_i\,,$ to denote the first and second derivatives of $\ell_i(\bullet)$ with respect to its third argument. (Recall that the third argument of ℓ_i contains the nonparametric component.) Equations (9.4.12), (9.4.14), (9.4.17) and (9.4.20) constitute the following iterative algorithm.

Semiparametric Profile Likelihood

- Calculate the gradient $\mathcal{D}_\ell$ (and the Hessian $\mathcal{H}_\ell$) as if there were explanatory variables x_{i1}, x_{i2} and $x_{i3} = u_i$.
- Compute

$$\widetilde{u}_j = u_j - \frac{\sum_{i=1}^{n} \ell_i''(u_i^T\gamma + m_j)\, K_h(t_i - t_j)\, u_i}{\sum_{i=1}^{n} \ell_i''(u_i^T\gamma + m_j)\, K_h(t_i - t_j)}\,, \text{ for } j = 1, \ldots, n.$$

- Replace in $\mathcal{D}_\ell$ (and $\mathcal{H}_\ell$) the terms η_{i3} by $\widetilde{u}_i^\top\gamma + m_i$ and subsequently the remaining terms x_{i3} by $\widetilde{u}_i$. Denote the resulting gradient by $\widetilde{D}_\ell$

(and the Hessian by $\widetilde{H}_\ell$). The *updating step for* $\vartheta = (\beta_1, \beta_2, \gamma, \sigma, \rho)$ is then implemented by an iterative optimization procedure using $\widetilde{D}_\ell$ (and $\widetilde{H}_\ell$).

- The *updating step for* m_j is given by

$$m_j^{new} = m_j^{old} - \frac{\sum_{i=1}^{n} \ell_i'(u_i^T\gamma + m_j)\, K_h(t_i - t_j)}{\sum_{i=1}^{n} \ell_i''(u_i^T\gamma + m_j)\, K_h(t_i - t_j)}, \quad \text{for } j = 1, \ldots, n.$$

We mention the Hessian in parenthesis as — depending on the chosen optimization for the parametric part — its computation may not be necessary. In practice, optimization routines provided with software packages might be much more efficient at this point. For the nonparametric part however, we use a Newton-Raphson type iteration for reasons of simplicity.

Let us mention a further simplification that we implement in the following. The computation of $\widetilde{u}_i$ and m_j^{new} requires to evaluate terms of the form

$$\sum_{i=1}^{n} \psi_{ij} K_h(t_i - t_j)\,. \tag{9.4.21}$$

Note, that this has to be done at least for all t_j $(j = 1, \ldots, n)$ since the updated values of m at all observation points are required in the updating step for ϑ. Thus, $O(n^2)$ operations are necessary for evaluating the kernel weights. However, the evaluation of (9.4.21) is a standard method if ψ_{ij} only depends on i (as for example in a Nadaraya–Watson and local polynomial kernel regression). The above algorithm requires additionally the computation of $\ell_i'(u_i^T\gamma + m_j)$ and $\ell_i''(u_i^T\gamma + m_j)$ for all $i, j = 1, \ldots, n$. We avoid this additional computational effort by replacing these terms by $\ell_i'(u_i^T\gamma + m_i)$ and $\ell_i''(u_i^T\gamma + m_i)$. Due to the kernel weights $K_h(t_i - t_j)$ becoming negligible for t_i far away from t_j this simplification will change the results only marginally.

9.5 Empirical Results

We apply the semiparametric two-step and profile likelihood methods to the data introduced in Section 9.3. Recall that the variable which is included nonparametrically into the selection equation is $\ln(L)$. Figure 9.1 shows how this variable is related to the decision to source out IT tasks. Since the decision is binary, these relative frequencies are obtained by grouping the

Table 9.2: Descriptive statistics for the restricted data set

Variable		Mean	Std. Dev.	Min.	Max.
PROD†	productivity	0.497	1.381	0.007	23.65
L	number of employees	168.601	243.131	10	1200
INV	non-IT-investment	4661.223	19098.687	1.000	497000.996
IT	IT-investment	505.085	2332.401	1.490	50001.009
EXP_VH	export quota	15.461	22.390	0	100
D_OUTV‡	IT-tasks completely sourced out	69.002	0.463	0	1
D_KONZ	firm belongs to group of companies	32.837	0.470	0	1
D_STAND	foreign subsidiary	16.988	0.376	0	1
D_AGE1	age <4 years	7.268	0.260	0	1
D_AGE2	age 4–7 years	14.361	0.351	0	1
D_ITOS	IT skill shortage	14.799	0.355	0	1
D_REG	East Germany	24.694	0.431	0	1
D_BR1	manufacturing industry w/o IT	50.438	0.500	0	1
D_BR2	distributive services	18.476	0.388	0	1
D_BR3	banking & insurance	4.291	0.203	0	1
D_BR4	technical services	6.743	0.251	0	1
D_BR5	other business-related services	5.254	0.223	0	1
D_BR6	IT w/o retail trade	14.799	0.355	0	1
Obs.		1142			

† Output per employee (total sales per year per employee in 1,000 DM), where sales means balance-sheet total for banks and sum insured for insurance companies.

‡ "D"stands for Dummy.

variable $\ln(L)$ into equi-spaced intervals. The numbers below the points show the number of observations in the corresponding interval.

We see that observations beyond $\ln(L) \approx 7$ are relatively sparse since our data base contains only a small part of large firms. In order to obtain a relatively homogenous set of observations with respect to the nonparametric function, we restrict the data set now to

$$2.3 \leq \ln(L) \leq 7.1 \quad \text{or} \quad 10 \leq L \leq 1200, \quad \text{respectively.}$$

Table 9.2 reports some descriptive statistics for this subset, which covers 1142 out of the originally 1351 observations.

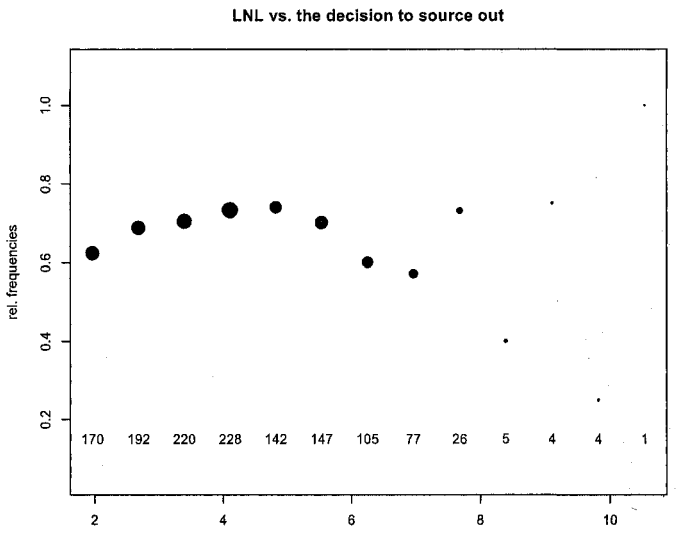


Figure 9.1: $\ln(L)$ (grouped into equi-spaced intervals) versus the relative frequencies of out-sourcing; numbers below the bullets indicate the number of observations in the considered $\ln(L)$-interval.

Table 9.3: Wald tests for identity of the coefficients

	parametric FIML		semiparametric two-step		semiparametric profile likelihood	
	χ^2	p-value	χ^2	p-value	χ^2	p-value
$\ln(IT)$	0.94	0.33	1.10	0.29	1.07	0.30
$\ln(K)$	0.60	0.44	2.39	0.12	0.72	0.40
$\ln(L)$	0.17	0.68	0.93	0.33	0.18	0.67
East Germany	3.64	0.06	1.74	0.19	3.57	0.06
Constant	16.77	0.00	0.21	0.64	20.15	0.00

As starting values for the semiparametric estimates we use the results from a parametric switching regression. The coefficients of this parametric FIML estimation are listed in Table 9.9. The results with respect to the level equations show the expected positive and significant partial output elasticities with respect to the capital variables measured by the logarithms of IT investment and 'normal' investment. Labor has a significantly negative out-

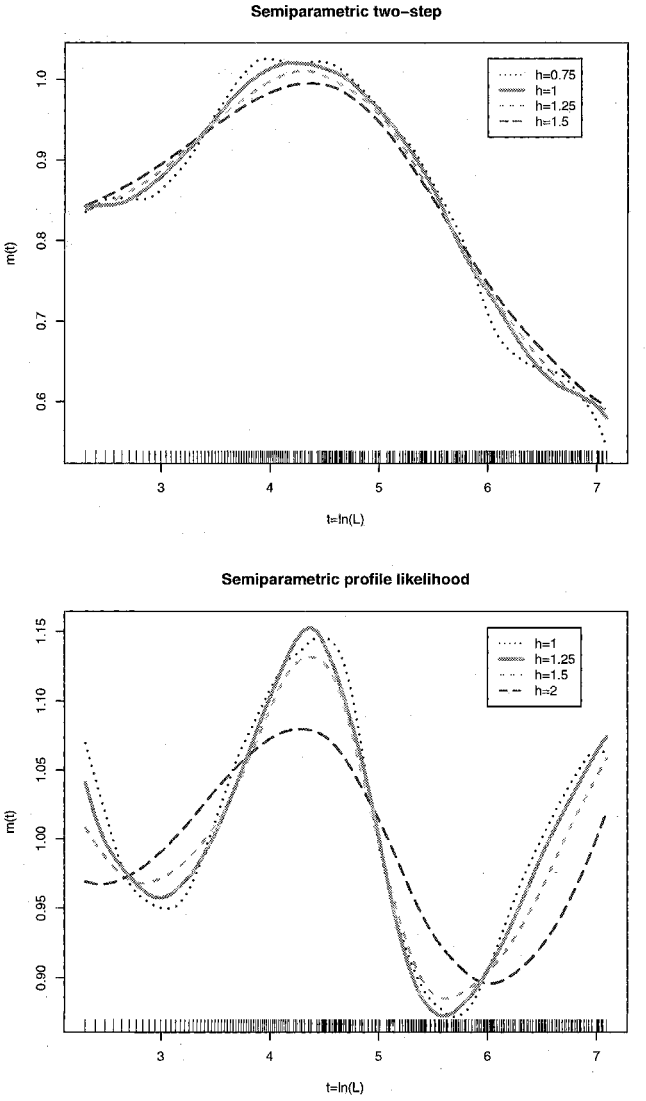


Figure 9.2: Families of nonparametric functions for $\ln(L)$ using the semiparametric two-step and profile likelihood approaches.

Table 9.4: Semiparametric fits in dependence of the bandwidth

	bandwidth	log-likelihood
semiparametric two-step	0.75	-1959.723
	1.00	-1960.422
	1.25	-1966.065
	1.50	-1970.316
semiparametric profile likelihood	1.00	-1932.775
	1.25	-1933.085
	1.50	-1933.411
	2.00	-1971.728

Table 9.5: Estimated coefficients (semiparametric profile likelihood)

	Coeff.	Std.Err.	t-value	p-value
Estimation results for regime with IT-Outsourcing				
CONSTANT	5.308	0.117	45.513	0.000
LNIKT	0.104	0.022	4.682	0.000
LNINV	0.140	0.020	7.077	0.000
LNL	−0.250	0.030	−8.312	0.000
D_REG	−0.364	0.074	−4.905	0.000
D_BR1	−0.389	0.080	−4.842	0.000
D_BR3	0.918	0.107	8.549	0.000
D_BR4	−0.813	0.240	−3.392	0.001
D_BR5	−0.464	0.126	−3.671	0.000
D_BR6	−0.882	0.133	−6.606	0.000
Estimation results for regime without IT-Outsourcing				
CONSTANT	6.606	0.263	25.117	0.000
LNIKT	0.143	0.035	4.032	0.000
LNINV	0.110	0.031	3.541	0.000
LNL	−0.226	0.048	−4.711	0.000
D_REG	−0.132	0.117	−1.121	0.262
D_BR1	−0.499	0.146	−3.415	0.001
D_BR3	1.423	0.211	6.744	0.000
D_BR4	−1.142	0.217	−5.253	0.000
D_BR5	−0.603	0.207	−2.918	0.004
D_BR6	−1.113	0.207	−5.381	0.000

put elasticity since the estimated coefficient corresponds to $\alpha_3 - 1$. Simple Wald test statistics (see Table 9.3) indicate that the elasticities do not differ significantly across the two productivity regimes, thus IT-outsourcing does

Table 9.6: Semiparametric profile likelihood continued

Selection equation				
LNIKT	−0.044	0.028	−1.573	0.116
LNINV	0.034	0.019	1.803	0.071
EXP_VH	−0.002	0.002	−0.890	0.373
D_ITOS	−0.173	0.099	−1.740	0.082
D_KONZ	−0.241	0.071	−3.401	0.001
D_STAND	−0.218	0.103	−2.115	0.034
D_AGE1	0.243	0.114	2.131	0.033
D_AGE2	−0.031	0.085	−0.372	0.710
D_REG	−0.060	0.092	−0.649	0.516
D_BR1	−0.328	0.107	−3.061	0.002
D_BR3	−0.032	0.159	−0.204	0.839
D_BR4	−0.662	0.182	−3.637	0.000
D_BR5	−0.279	0.173	−1.614	0.107
D_BR6	−1.005	0.143	−7.024	0.000
σ_1	0.931	0.029	32.342	0.000
σ_2	1.039	0.072	14.369	0.000
ρ_1	0.845	0.027	31.083	0.000
ρ_2	0.839	0.050	16.659	0.000

not seem to be related to higher partial output elasticities of the input factors. However, the constant term reflecting the firms' production effficiency differs significantly across the two regimes implying that firms without IT-outsourcing produce more efficiently.

All sector dummies included in the level equations and controlling for different measurements of labor productivity across sectors are highly significant. The industries of the base category (trade, transport and postal services) are all at the end of the value added chain, reaching a high value of total sales per employee. The sector of financial intermediation shows a significantly higher labor productivity. All other industries that produce at earlier stages of the value added chain than the base category show a significantly lower value of total sales per employee.

The dummy variable for East Germany has a negative and significant coefficient, reflecting the lower labor productivity especially in the East German manufacturing sector. Identity of these parameters across the two regimes is rejected at the 5-percent level. Thus, the productivity differential for East German firms compared to their West German counterparts is even larger in the regime with IT-outsourcing.

Table 9.7: Estimated coefficients (semiparametric two-step)

	Coeff.	Std.Err.	t-value	p-value
Estimation results for regime with IT-Outsourcing				
CONSTANT	5.416	0.137	39.480	0.000
LNIKT	0.127	0.023	5.632	0.000
LNINV	0.135	0.021	6.496	0.000
LNL	−0.236	0.035	−6.770	0.000
D_REG	−0.337	0.064	−5.256	0.000
D_BR1	−0.357	0.077	−4.628	0.000
D_BR3	1.287	0.156	8.227	0.000
D_BR4	−0.692	0.158	−4.386	0.000
D_BR5	−0.354	0.138	−2.557	0.011
D_BR6	−0.577	0.202	−2.852	0.004
Estimation results for regime without IT-Outsourcing				
CONSTANT	5.118	0.587	8.716	0.000
LNIKT	0.169	0.035	4.862	0.000
LNINV	0.083	0.028	2.931	0.003
LNL	−0.178	0.050	−3.588	0.000
D_REG	−0.172	0.096	−1.789	0.074
D_BR1	−0.209	0.156	−1.345	0.178
D_BR3	1.148	0.253	4.534	0.000
D_BR4	−0.585	0.239	−2.445	0.015
D_BR5	−0.437	0.237	−1.844	0.065
D_BR6	−0.327	0.297	−1.100	0.272

The results for the selection equation show that having a subsidiary in a foreign country or belonging to a group of companies significantly reduces the probability to source out IT-tasks. Maybe, these firms have the possibility to use the resources of the group to do IT-tasks inhouse or within the group. The fact that firms face an IT-skill shortage turns out to reduce the probability of IT-outsourcing probably since these companies have their own IT-departments or IT-specialists. Firms that are three years old or younger have a significantly higher probability for IT-outsourcing. There are no significant impacts with respect to the input factors, the export quota and the dummy indicating whether a firm is between four and seven years old. The manufacturing industry without IT, technical services as well as the IT-sector have significantly lower probabilities to source out IT-tasks than the industries for trade, transport and postal services. This result seems plausible since these firms either do not use IT intensively or they are able to perform IT-tasks by themselves.

The fact that both correlation coefficients ρ_1 and ρ_2 are significantly estimated indicates that it is justified to estimate an endogenous switching re-

Table 9.8: Semiparametric two-step continued

Selection equation				
LNIKT	−0.028	0.032	−0.854	0.393
LNINV	0.030	0.029	1.039	0.299
EXP_VH	0.002	0.002	1.240	0.215
D_ITOS	−0.340	0.120	−2.843	0.004
D_KONZ	−0.105	0.094	−1.118	0.264
D_STAND	−0.056	0.120	−0.470	0.638
D_AGE1	0.016	0.157	0.101	0.919
D_AGE2	−0.067	0.116	−0.578	0.563
D_REG	−0.025	0.097	−0.263	0.792
D_BR1	−0.302	0.121	−2.487	0.013
D_BR3	−0.345	0.218	−1.580	0.114
D_BR4	−0.591	0.183	−3.225	0.001
D_BR5	−0.318	0.202	−1.575	0.115
D_BR6	−1.009	0.151	−6.679	0.000
σ_1	0.765			
σ_2	0.755			
ρ_1	0.159			
ρ_2	0.083			

gression model. If the estimates of the correlation coefficients were insignificant, the appropriate model would be an exogenous switching regression model where the outsourcing decision would be independent on the firms' productivity level.

In the next step, we take into account that the decision whether or not to completely source out IT-tasks may depend nonlinearly on the firm size. As pointed out in Section 9.1, small firms might rather source out single tasks whereas large firms rather source out whole departments. Small and large firms thus might have a higher probability for outsourcing than middle sized firms. The firm size is measured by the logarithm of the number of employees in the firm. As shown in Section 9.4, we apply a semiparametric two-step estimator as well as a full information semiparametric profile likelihood estimator. The estimated coefficients of these semiparametric regressions are presented in Tables 9.7 and 9.5.[3]

The semiparametric profile likelihood results do not differ considerably from the results of the parametric estimation, neither with respect to the coefficients nor with respect to the significance. In particular, also the Wald test

[3]For the implementation of the parametric and semiparametric algorithms we use the statistical programming language R (Ihaka & Gentleman 1996). The R package `optim` provides the BFGS optimization procedure used for the parametric iterations. The nonparametric estimation steps are implemented using a dynamic link library in C.

Table 9.9: Estimated coefficients (parametric FIML)

	Coeff.	Std.Err.	t-value	p-value
Estimation results for regime with IT-Outsourcing				
CONSTANT	5.319	0.137	38.935	0.000
LNIKT	0.104	0.023	4.500	0.000
LNINV	0.138	0.021	6.495	0.000
LNL	−0.250	0.036	−6.875	0.000
D_REG	−0.367	0.075	−4.880	0.000
D_BR1	−0.393	0.082	−4.781	0.000
D_BR3	0.910	0.108	8.395	0.000
D_BR4	−0.822	0.240	−3.430	0.001
D_BR5	−0.466	0.129	−3.620	0.000
D_BR6	−0.893	0.135	−6.637	0.000
Estimation results for regime without IT-Outsourcing				
CONSTANT	6.567	0.302	21.763	0.000
LNIKT	0.141	0.037	3.794	0.000
LNINV	0.109	0.037	2.988	0.003
LNL	−0.226	0.056	−4.039	0.000
D_REG	−0.131	0.118	−1.110	0.267
D_BR1	−0.484	0.151	−3.198	0.001
D_BR3	1.428	0.220	6.482	0.000
D_BR4	−1.113	0.220	−5.067	0.000
D_BR5	−0.607	0.218	−2.790	0.005
D_BR6	−1.076	0.215	−5.003	0.000

statistics (Table 9.3) lead to the same conclusions on the output elasticities. More differences can be found when using the semiparametric two-step estimator. In the level equations, the estimated coefficients differ only quantitatively except of the dummy for the ICT sector which turns to be insignificant in the regime without outsourcing according to the semiparametric two-step estimator. In the selection equation, there are several variables the coefficients of which turn to be insignificant in the two-step estimation.

Since there are no common rules for the choice of the smoothing parameter h, we estimate the semiparametric model for a family of bandwidths. Table 9.4 indicates that the log-likelihood values of the semiparametric profile likelihood estimator seems to be fairly independent of the bandwidth. We observe log-likelihoods around -1933 which are slightly better than the parametric FIML estimate. The semiparametric two-step estimator is more sensitive with respect to the bandwidth with log-likelihood values systematically beyond those of the profile likelihood estimator.

Figure 9.2 shows the estimated nonparametric functions of $\ln(L)$ for a series of bandwidths. The rug plot on the bottom of the figures indicates the obser-

Table 9.10: Parametric FIML continued

Selection equation				
CONSTANT	1.069	0.187	5.731	0.000
LNIKT	−0.043	0.032	−1.335	0.182
LNINV	0.031	0.029	1.055	0.291
LNL	−0.014	0.050	−0.283	0.777
EXP_VH	−0.001	0.002	−0.808	0.419
D_ITOS	−0.169	0.102	−1.664	0.096
D_KONZ	−0.243	0.072	−3.361	0.001
D_STAND	−0.223	0.107	−2.089	0.037
D_AGE1	0.222	0.115	1.921	0.055
D_AGE2	−0.025	0.086	−0.295	0.768
D_REG	−0.057	0.094	−0.605	0.545
D_BR1	−0.310	0.119	−2.607	0.009
D_BR3	−0.033	0.167	−0.196	0.845
D_BR4	−0.666	0.193	−3.453	0.001
D_BR5	−0.260	0.188	−1.382	0.167
D_BR6	−1.003	0.157	−6.382	0.000
σ_1	0.931	0.029	31.925	0.000
σ_2	1.021	0.074	13.854	0.000
ρ_1	0.844	0.027	31.299	0.000
ρ_2	0.822	0.057	14.512	0.000

vations of $\ln(L)$. The semiparametric two-step estimator reveals an inverse U-shape suggesting that the probability to completely source out IT-tasks is first increasing with firm size and then decreasing. By contrast, the semiparametric profile estimator suggests that IT-outsourcing is increasing with firm size for log-labor being larger than 3, then decreasing up to a firm size of about 5.5 and finally increasing again for large firms. According to the log-likelihood values in Table 9.4 it seems that the profile likelihood estimator captures the relationship between IT-outsourcing and firm size more accurately. Moreover, the fact that the results are quite independent of the bandwidth choice gives more confidence into the profile likelihood estimates than into the estimates of the two-step procedure. However, further analyses are needed in order to draw definite conclusions on the reliability of these semiparametric estimation procedures.

9.6 Concluding Remarks

This paper analyzes the relationship between IT-outsourcing and firms' labor productivity. An endogenous switching regression model allows to take into

account simultaneity between labor productivity and IT-outsourcing. Firms may follow different productivity regimes depending on whether or not they completely source out IT-tasks. For this switching regression model a full information maximum likelihood (FIML) or a two-step estimation approach can be applied. We extend these two estimation approaches by a nonparametric component in order to take into account that the outsourcing decision might depend nonlinearly on firm size. Empirical evidence is based on a data set containing 1142 firms of the German manufacturing and the business-related services industries.

The estimation results of the applied methods imply that IT-outsourcing does not significantly change the partial output elasticities of the production factors labor, capital and IT capital. However, firms without IT-outsourcing turn out to produce more efficiently than those with IT-outsourcing. High coordination costs between outsourcing firms and their subcontractors may make business processes less efficient.

The estimation of the nonparametric relationship between IT-outsourcing and firm size captured by the semiparametric approaches show considerable differences. While the semiparametric two-step estimator suggests a rather inverse U-shaped relationship, the corresponding curve resulting from the semiparametric profile likelihood estimator shows more structure in the sense that the probability to source out is first increasing, then decreasing and finally increasing again with firm size.

The advantage of the semiparametric profile likelihood estimator over the semiparametric two-step approach seems to be its robustness with respect to the choice of the bandwidth and its better performance with respect to maximizing the log-likelihood. However, further analyses are needed in order to be able to draw final conclusions about the performance of the estimators applied in this study.

Bibliography

Bertschek, I., Fryges, H. & Kaiser, U. (2004). B2B or not to Be: Does B2B E-commerce increase labour productivity?, *Discussion paper no. 04-45*, ZEW Mannheim.

Bertschek, I. & Kaiser, U. (2004). Productivity effects of organizational change: Microeconometric evidence, *Management Science* **50**(3): 394–404.

Brynjolfsson, E. & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance, *Journal of Economic Perspectives* **14**(4): 23–48.

Dewan, S. & Min, C.-K. (1997). The substitution of information technology for other factors of production: A firm-level analysis, *Management Science* **43**(12): 1660–1675.

Greenan, N. & Mairesse, J. (2000). Computers and productivity in France: Some evidence, *Economics of Innovation and New Technology* **9**(3): 275–315.

Greene, W. H. (2000). *Econometric Analysis*, 4 edn, Prentice Hall, Upper Saddle River, New Jersey.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such model, *Annals of Economic and Social Measurement* **5**: 475–492.

Hempell, T. (2005). What's spurious? What's real? Measuring the productivity of ICT at the firm level, *Empirical Economics* **30**(2): 427–464.

Henkel, J. & Kaiser, U. (2002). Fremdvergabe von IT-Dienstleistungen aus personalwirtschaftlicher Sicht, *Discussion Paper 02-11*, ZEW Mannheim.

Heshmati, A. (2003). Productivity growth, efficiency and outsourcing in manufacturing and service industries, *Journal of Economic Surveys* **17**(1): 79–112.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics* **5**(3): 299–314.

Lamberti, H.-J. (2003). Mit IT-Sourcing zu einer neuen Stufe der Indutrialisierung im Bankbetrieb, *Zeitschrift für das gesamte Kreditwesen* **6**: 307–309.

Licht, G. & Moch, D. (1999). Innovation and information technology in services, *Canadian Journal of Economics* **32**(2): 363–382.

Lichtenberg, F. R. (1995). The output contributions of computer equipment and personnel: A firm-level analysis, *Economics of Innovation and New Technology* **3**: 201–217.

Maddala, G. S. (1983). *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs No. 4, Cambridge University Press.

Müller, M. (2001). Estimation and testing in generalized partial linear models — a comparative study, *Statistics and Computing* **11**: 299–309.

Prantl, S. (2001). Financial distress, liquidations and subsidization of young firms, *Doctoral dissertation*, University of Mannheim, Germany.

Severini, T. A. & Staniswalis, J. G. (1994). Quasi-likelihood estimation in semiparametric models, *Journal of the American Statistical Association* **89**: 501–511.

Severini, T. A. & Wong, W. H. (1992). Generalized profile likelihood and conditionally parametric models, *Annals of Statistics* **20**: 1768–1802.

Staniswalis, J. G. & Thall, P. F. (2001). An explanation of generalized profile likelihoods, *Statistics and Computing* **11**: 293–298.

10 Nonparametric and Semiparametric Estimation of Additive Models with Both Discrete and Continuous Variables under Dependence[1]

Christine Camlong-Viot[2], Juan M. Rodríguez-Póo[2], and Philippe Vieu[4]

[2] Faculté de Pharmacie, Département de Biomathématique, Université Paris-Sud, 92296 Châtenay-Malabry, France
[3] Departamento de Economía, Universidad de Cantabria, 39005 Santander, Spain
[4] Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse, France

Summary

This paper is concerned with the estimation of nonparametric and semiparametric additive models in the presence of discrete variables. The main feature of our work is to deal with possibly dependent variables. Our methodology can be seen as well as an unifying presentation of several different situations, as extensions to dependence structures of several recent advances obtained in the usual i.i.d. case.

Among the different estimation procedures, the method introduced by Linton and Nielsen (1995), based in marginal integration, has became quite popular because both its computational simplicity and the fact that it allows an asymptotic distribution theory. Here, an asymptotic treatment of the marginal integration estimator under different mixtures of continuous-discrete variables is offered, and furthermore, in the semiparametric partially additive setting, an estimator for the parametric part that is consistent and asymptotically efficient is proposed. The estimator is based in minimizing the L_2 distance between the additive nonparametric component and its correspondent linear direction.

Keywords: Additive models, dimension reduction techniques, semiparametric models, strong mixing conditions, marginal integration

10.1 Introduction

This paper addresses an old problem considered here from a rather different-new perspective. The problem of how to treat discrete variables in nonparametric regression problems is already well known in the statistical literature (see among others Hall, 1981; Bierens, 1983; Grund and Hall, 1993; and Ahmad and Cerrito, 1994). When the regressors are discrete no smoothing is required to obtain root-n consistent estimators. Furthermore, if any amount of smoothing is applied, then, the discrete components do not suffer from the curse of dimensionality.

In the econometrics literature, the same problem has been traditionally approached by retreating it as a semiparametric problem. That is, the continuous variables are introduced either in a multivariate or in an additive one-dimensional nonparametric regression setting whereas discrete regressors appear in the form of linear parametric functions. These are the so called partially linear models. In this setting, Delgado and Mora (1995) show that root-n consistency of the parametric part is achieved under much weaker conditions than in the continuous case (see Robinson, 1988).

In many cases (see Horowitz, 1998) the partially linear structure does not appear to be a reasonable restriction. Racine and Li (2000) analyze the case when discrete and continuous variables are mixed within a multivariate nonparametric regression function. They provide the statistical properties of the estimator and a method to choose the different bandwidths. However, the use of multivariate nonparametric regression models presents an important problem: When many explanatory variables are available, the rate at which nonparametric smoothers converge to their true values is very slow, and the the introduction of additive restrictions is recommended (Stone, 1985). In Fan, Härdle and Mammen (1998) the impact of discrete regressors in the estimation of additive models is analyzed. They also consider as a particular case a semiparametric additive partially linear model, and provide root-n consistent estimators of the parameters of interest. Their method is based on local linear regression smoothers, and they allow for components that can be either discrete or continuous. However, their estimation procedure presents some drawbacks. First, they only give the statistical properties of the nonparametric additive components that depend on absolutely continuous regressors, second the resulting estimator for the nonparametric component is created by splitting the sample in several cells. The number of cells depends on the number of categories of the discrete variables, and therefore, if the

[1]This research was financially supported by The Dirección General de Investigación del Ministerio de Educación y Ciencia under research grant SEJ2005-08269/ECON. The authors wish to express their gratitude to the participants of the STAPH (http://www.lsp.ups-tlse.fr/Fp/Ferraty/staph.html) group in Functional Statistics in Toulouse for their many helpful comments and suggestions.

number of cells is high each may not have enough observations to estimate. Finally, the whole analysis is performed under the assumption of independent and identically distributed observations. This assumption, typically rules out regression models that contain lagged endogenous variables as regressors.

This paper addresses the problem of introducing both discrete and continuous explanatory variables into an additive nonparametric (semiparametric) regression setting that accounts for dependent data. In order to estimate the additive components marginal integration techniques (Newey, 1994; Tjostheim and Auestad, 1994 and Linton and Nielsen, 1995) are used. Here, the pilot multivariate nonparametric regression estimator is computed by using kernel methods. Discrete covariates enter in the product kernel although no smoothing is applied to them. We show that estimators of the additive components with discrete covariates exhibit root-n rates and in the mixed case, that is, estimators of the additive components that depend both on continuous and discrete covariates, the rate of convergence is the same as in the continuous case.

Further if we assume that the additive components depending on discrete regressors fall within the class of linear parametric functions, a two step method to estimate the parametric part is proposed. The estimator is based in minimizing the L_2 distance between the additive nonparametric component and its correspondent linear direction. It is root-n consistent and achieves the semiparametric efficiency bound.

An important feature of our work is to consider a strongly dependent model that allows for applications in time series situations. Concretely, we will deal with the quite general α mixing model. Because this notion will be a the center of our paper, we decided to recall its definition right now. This structure was previously introduced by Rosenblatt (1956) and is known to be one of the less restrictive dependence structure. Moreover, this notion turns to be of great interest in nonparametric statistics as can be seen in the monographies by Györfi et al (1989) or Bosq (1998). Recall that the α-mixing coefficient relative to some process $\{U_i\}_{i\geq 1}$ is defined, for any $s \in \mathbb{N}$, by

$$\alpha(s) = \sup\left\{\left|P\left(A \cap B\right) - P\left(A\right)P\left(B\right)\right|, A \in \mathcal{F}_1^r, B \in \mathcal{F}_{r+s}^{\infty}\right\}$$

where $\mathcal{F}_1^r$ and $\mathcal{F}_{r+s}^{\infty}$ are σ-fields generated respectively by $\{U_1, \cdots, U_r\}$ and $\{U_{r+s}, \cdots\}$. A process is strongly mixing if

$$\lim_{s\to\infty} \alpha(s) = 0.$$

The remainder of the paper is organized as follows. The statistical model and the estimator are introduced in Section 10.2. Its asymptotic behavior is also treated in this section. In Section 10.3 we present a two step root-n

consistent semiparametric estimator of the partially additive linear model. Proofs are deferred to the Appendix.

10.2 Additive Nonparametric Regression

Along this section we consider an additive nonparametric regression model where a subset of explanatory variables is discrete and the remaining are continuous. More precisely, let $X^c = (X_1, X_3)$ be a vector of continuous random variables valued in $\mathbb{R}^{p_1+p_2}$ and $X^d = (X_2, X_4)$ be a vector of discrete random variables valued in $\mathbb{R}^{q_1+q_2}$. That is, that there exists $\mathcal{D} \in \mathbb{R}^{q_1+q_2}$ such that

$$P\left(X^d \in \mathcal{D}\right) = 1, \tag{10.2.1}$$

$$\forall x^d \in \mathcal{D}, \quad P\left(X^d = x^d\right) > 0. \tag{10.2.2}$$

Let $X_i = (X_{1i}, X_{2i}, X_{3i}, X_{4i})$. We consider a nonparametric regression model given by

$$Y_i = m(X_i) + \epsilon_i = \omega + m_1(X_{1i}) + m_2(X_{2i}) + m_{34}(X_{3i}, X_{4i}) + \epsilon_i, \tag{10.2.3}$$

where $\{(X_i, Y_i)\}_{i=1}^n$ are observations form a stationary α-mixing process, $E(\epsilon | X_i) = 0$ and $m_1(\bullet)$, $m_2(\bullet)$ and $m_{34}(\bullet, \bullet)$ are of unknown form. For identification purposes $E[m_1(X_{1i})] = E[m_2(X_{2i})] = E[m_{34}(X_{3i}, X_{4i})] = 0$.

Note that this model nests a broad variety of different specifications. If we set $m_1 = m_2 = 0$ then we consider the same model as in Racine and Li (2000). On the other side, if $m_2 = 0$ then we have the model analyzed in Fan, Härdle and Mammen (1998). Of course in both cases all results were obtained for the independent case.

Our interest is to estimate the unknown quantities, that is $m_1(\bullet)$, $m_2(\bullet)$ and $m_{34}(\bullet, \bullet)$ in the dependent regression model. So far, purely additive models have been estimated using the backfitting algorithm and the so called marginal integration techniques. The first method was proposed in Hastie and Tibshirani (1990) and the second was simultaneously developed in Newey (1994); Tjostheim and Auestad (1994) and Linton and Nielsen (1995). From the computational point of view both approaches appear equally feasible. The backfitting has been mostly implemented using splines. Stone, Hansen, Kooperberg and Truong (1997) develop estimation theory using polynomial spline methods and Wahba (1990) uses smoothing splines. Also local polynomial regression has been used as in Opsomer and Ruppert (1994). For the marginal integration techniques, series estimators (Andrews and Whang, 1990; and Newey, 1995), local constant polynomials (Linton and Nielsen, 1995) and local linear polynomials (Fan, Härdle and Mammen, 1998) have

been applied. From the theoretical point of view, although the behavior of the marginal integration estimators is known better, however, important developments have been made in the theory of backfitting (see Mammen, Linton and Nielsen, 1999; Opsomer and Ruppert, 1998; and Opsomer, 2000). In the context of dependent data, to our knowledge, no results are available for the backfitting estimator whereas marginal estimators have been studied in deep by Sperlich, Tjostheim and Yang (2002) and Camlong-Viot, Sarda and Vieu (2000). On these grounds, we opt to estimate the different unknown components by marginal integration techniques.

At this stage it is worth being fixed some notations. In the following, all the integrals related with continuous variables will be taken with respect to Lebesgue measure while all the integrals related with discrete variables will be taken with respect to the counting measure (the counting measure will be denoted by μ). In the following we will also make use of some functions

$$q(x) = q(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2)q_{34}(x_3, x_4),$$

where q_1, q_2 and q_{34} are known density functions respectively defined on $\mathbb{R}^{p_1}$, $\mathbb{R}^{q_1}$ and $\mathbb{R}^{p_2+q_2}$. Moreover, for any $\ell = 1, \cdots, 4$ we will denote by f_ℓ the marginal density of X_ℓ (giving the fact that these marginal densities are either taken with respect to the Lebesgue measure for continuous X_ℓ or with respect to μ for discrete ones). Similarly, for any $\ell = 1, \cdots, 4$ and for any $s > 0$ we will denote by $f_{\ell,s}$ the joint density of $(X_{\ell,j}, X_{\ell,j+s})$. Finally, we will denote by

$$f(x_1, x_2, x_3, x_4) = f_c(x_1, x_3|x_2, x_4)f_D(x_2, x_4),$$

where f_c is the conditional density (with respect to the Lebesgue measure) of (X_1, X_3) given (X_2, X_4) and where f_D is the density (with respect to the counting measure) of (X_2, X_4).

This estimation method consists in integrating the regression function $m(\bullet)$ with respect to a suitable density function. By doing this we obtain

$$\begin{aligned} & \int_{\mathbb{R}^{q_1+p_2+q_2}} m(x)q_2(x_2)q_{34}(x_3, x_4)\mu(dx_2)dx_3\mu(dx_4) \qquad (10.2.4) \\ = \; & \omega + m_1(x_1) + \int_{\mathbb{R}^{q_1}} m_2(x_2)q_2(x_2)\mu(dx_2) \\ & + \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3, x_4)q_{34}(x_3, x_4)dx_3\mu(dx_4). \end{aligned}$$

On the other hand, integrating $m(\bullet)$ with respect to a density function

$q(x_1, x_2, x_3, x_4) = q_1(x_1)q_2(x_2)\, q_{34}(x_3, x_4)$ defined on $\mathbb{R}^{p_1+q_1+p_2+q_2}$ we obtain

$$\begin{aligned} & \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} m(x)q_1(x_1)q_2(x_2)q_{34}(x_3,x_4)dx_1\mu(dx_2)dx_3\mu(dx_4) \\ = \; & \omega + \int_{\mathbb{R}^{p_1}} m_1(x_1)q_1(x_1)dx_1 + \int_{\mathbb{R}^{q_1}} m_2(x_2)q_2(x_2)\mu(dx_2) \\ & + \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3,x_4)q_{34}(x_3,x_4)dx_3\mu(dx_4). \end{aligned} \tag{10.2.5}$$

Then subtracting equation (10.2.5) from (10.2.4) we obtain an expression for the additive component $m_1(x_1)$, up to an additive constant,

$$\begin{aligned} \eta_1(x_1) &= m_1(x_1) - \int_{\mathbb{R}^{p_1}} m_1(x_1)q_1(x_1)dx_1 \\ &= \int_{\mathbb{R}^{q_1+p_2+q_2}} m(x)q_2(x_2)q_{34}(x_3,x_4)dx_3\mu(dx_4)\mu(dx_2) - \\ & \quad \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} m(x)q_1(x_1)q_2(x_2)q_{34}(x_3,x_4)dx_1\mu(dx_2)dx_3\mu(dx_4). \end{aligned}$$

An estimator for $\eta_1(x_1)$, $\hat{\eta}_1(x_1)$, is obtained by replacing in the equation above the unknown quantities by some estimator

$$\begin{aligned} \hat{\eta}_1(x_1) &= \int_{\mathbb{R}^{q_1+p_2+q_2}} \hat{m}_n(x)q_2(x_2)q_{34}(x_3,x_4)dx_3\mu(dx_4)\mu(dx_2) - \\ & \quad \int_{\mathbb{R}^{p_1+q_1+p_2+q_2}} \hat{m}_n(x)q_1(x_1)q_2(x_2)q_{34}(x_3,x_4)dx_1\mu(dx_2)dx_3\mu(dx_4). \end{aligned} \tag{10.2.6}$$

An estimator for $m(x)$ is

$$\hat{m}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) \mathbb{I}(x_2 = X_{2i}) \frac{1}{h_3^{p_2}} L\left(\frac{x_3 - X_{3i}}{h_3}\right) \mathbb{I}(x_4 = X_{4i})}{f\left(X_{1i}, X_{2i}, X_{3i}, X_{4i}\right)},$$

where $\mathbb{I}(A)$ stands for the indicator function (that takes value one if A is true, and zero otherwise), where K and L are kernel functions supposed to satisfy some moment conditions to be specified before in $(H.4)$, and where h_1 and h_3 are positive smoothing parameters for which usual restrictions will be imposed later on (see condition $(H.5)$ below). This estimator is the so-called "internal" estimator of Jones, Davies and Park (1994). In smoothing problems, the indicator function has been proposed in another contexts by Delgado and Mora (1995) and Fan, Härdle and Mammen (1998) to account for discrete variables. Further Racine and Li (2000) propose a kernel function that depends on a smoothing parameter. Delgado and Mora (1995) did not consider the case of a mixture of continuous and discrete variables, Fan, Härdle and Mammen (1998) take the indicator function over a broader set

of values of X^d on its support, and finally Racine and Li (2000) face the additional problem of estimating a control parameter with no theoretical gains in doing so.

Following the marginal integration method, the component

$$\eta_2(x_2) = m_2(x_2) - \int_{\mathbb{R}^{q_1}} m_2(x_2) q_2(x_2) \mu(dx_2),$$

is estimated by

$$\hat{\eta}_2(x_2) = \int_{\mathbb{R}^{p_1+p_2+q_2}} \hat{m}_n(x) q_1(x_1) q_{34}(x_3, x_4) dx_1 dx_3 \mu(dx_4) - \int_{\mathbb{R}^{p_1+p_2+q_1+q_2}} \hat{m}_n(x) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4)$$

and the component

$$\eta_{34}(x_{34}) = m_{34}(x_{34}) - \int_{\mathbb{R}^{p_2+q_2}} m_{34}(x_3, x_4) q_{34}(x_3, x_4) dx_3 \mu(dx_4)$$

is estimated by

$$\hat{\eta}_{34}(x_3, x_4) = \int_{\mathbb{R}^{p_1+q_1}} \hat{m}_n(x) q_1(x_1) q_2(x_2) dx_1 \mu(dx_2) - \int_{\mathbb{R}^{p_1+p_2+q_1+q_2}} \hat{m}_n(x) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4).$$

In what follows we give some results about the asymptotic behavior of the estimators $\hat{\eta}_1$, $\hat{\eta}_2$ and $\hat{\eta}_{34}$. We give first some definitions and assumptions. From now on M and C will denote finite positive generic real constants.

(H.1) $m_1(x_1)$ is k-times continuously differentiable with respect to all its arguments in the support $\mathcal{X}_1$ of X_1. Furthermore, $m_{34}(x_3, x_4)$ is k-times continuously differentiable with respect to $X_3 \in \mathcal{X}_3$ where $\mathcal{X}_3$ is the support of X_3.

(H.2a) $\alpha(s) = O\left(s^{-a}\right)$, with $a > \frac{2\beta}{\beta-2}$, and for $\ell = 1, \cdots, 4$ and $s \geq 1$ assume that we have

$$\forall x, y, |f_{\ell,s}(x, y) - f_\ell(x) f_s(y)| \leq M < \infty.$$

(H.2b) $\alpha(s) = O\left(s^{-a}\right)$, with $a > \frac{2\beta}{\beta-2}\left(\frac{2k}{p_i} + 2\right)$ for $i = 1, 3$, and for $\ell = 1, \cdots, 4$ and $s \geq 1$, assume that we have

$$\forall x, y, |f_{\ell,s}(x, y) - f_\ell(x) f_s(y)| \leq M < \infty.$$

(H.3) $q_1(x_1)$ is bounded and k+1-times continuously differentiable in the support $\mathcal{X}_1$ of X_1. $q_2(x_2)$ is bounded with respect to all its arguments in the support $\mathcal{X}_2$ of X_2. Furthermore, $q_{34}(x_3, x_4)$ is bounded and k+1-times continuously differentiable with respect to X_3 in $\mathcal{X}_3$.

(H.4) The kernel functions $K(\bullet)$ and $L(\bullet)$ are compactly supported, bounded, continuous and they integrate to one. Furthermore[1], $\forall\ (i_1, \cdots, i_{p_1}) \in \mathbb{N}^{*p_1}$,

$$(\forall j, i_j < k) \Rightarrow \int_{\mathbb{R}^{p_1}} u_1^{i_1} \cdots u_{p_1}^{i_{p_1}} K(u_1, \cdots, u_{p_1})\, du_1 \cdots du_{p_1} = 0$$
$$\forall j, \int_{\mathbb{R}^{p_1}} u_j^k K(u_1, \cdots, u_{p_1})\, du_1 \cdots du_{p_1} \in \mathbb{R}^*$$

and $\forall\ (i_1, \cdots, i_{p_2}) \in \mathbb{N}^{*p_2}$,

$$(\forall j, i_j < k) \Rightarrow \int_{\mathbb{R}^{p_2}} u_1^{i_1} \cdots u_{p_2}^{i_{p_2}} L(u_1, \cdots, u_{p_1})\, du_1 \cdots du_{p_2} = 0$$
$$\forall j, \int_{\mathbb{R}^{p_2}} u_j^k L(u_1, \cdots, u_{p_2})\, du_1 \cdots du_{p_2} \in \mathbb{R}^*$$

(H.5) The bandwidth satisfy $h_1 = c_1 n^{-\frac{1}{2k+p_1}}$ and $h_3 = c_1 n^{-\frac{1}{2k+p_3}}$.

(H.6) The functions $f(\bullet)$ and f_ℓ, for $\ell = 1, 2, 3, 4$ are such that

$\exists\ b, B$ such that $0 < b \leq f(x) \leq B < \infty$ and $0 < b \leq f_\ell(x_\ell) \leq B < \infty$.

Let $x_{-\ell} = (x_1, \cdots, x_{\ell-1}, x_{\ell+1}, \cdots, x_4)$. Then the conditional density $f(x_{-\ell}|x_\ell)$ exists and it is bounded away from zero on the support of $f(\bullet)$.

(H.7) The conditional variance $\sigma_0^2(x) = \mathrm{Var}(Y|X = x)$ is continuous.

(H.8) $\forall i, j,\ E\left[|Y_i Y_j|^{\beta/2} \middle| X\right] \leq M < \infty,\ \beta > 2.$

Assumptions (H.1), (H.4), (H.5) and (H.6) are standard in nonparametric regression techniques. In fact (H.4) assumes higher order kernels (see Vieu, 1991). Note that as expected the number of derivatives allowed in (H.1) matches the order of the kernels in (H.4). This is needed to control the bias in the multivariate estimator. The bandwidth rates in (H.5) are chosen according the previous conditions on kernels and densities. (H.6) introduces a strong assumption: The densities must be compactly supported. This is done without loss of generality. In fact we could allow for unbounded support using trimming techniques (Robinson, 1988), but this would complicate the analysis unnecessarily. (H.2a) and (H.2b) are mixing conditions. Note that we have considered separately the discrete and the continuous covariates case.

[1] In these definitions, and further on this paper, we denote by $\mathbb{N}^*$ (resp. by $\mathbb{R}^*$) the set of all the positive integers (resp. the set of all real numbers) different from 0.

In this condition it is assumed that mixing coefficients decay at a algebraic rate. This is the weakest condition it can be imposed for the rate of decay of the mixing coefficients (see Bosq, 1998).

Now with the previous assumptions in hand we provide two results that characterize the asymptotic properties of the different components. The proofs are relegated to the Appendix. We start by the estimators of the component that depend respectively on continuous explanatory variables, $\hat{\eta}_1(x_1)$, and a mixture of continuous and discrete regressors, $\hat{\eta}_{34}(x_3, x_4)$.

Theorem 1 *i) Consider assumptions (H.1), (H.2b), (H.3), (H.4), (H.5), (H.6), (H.7) and (H.8) hold, then as* $n \to \infty$, *we have*

$$\sqrt{nh_1^{p_1}}\left(\hat{\eta}_1(x_1) - \eta_1(x_1)\right) \to_d \mathcal{N}\left(b(x_1), v^2(x_1)\right) \tag{10.2.7}$$

$$b(x_1) = \frac{1}{k!}\sum_{j=1}^{p_1}\int u_j^k K(u)du\left[(-1)^k\frac{\partial^k m_1}{\partial x_{1j}^k}(x_1) + \int m_1(z_1)\frac{\partial^k q_1}{\partial z_{1j}^k}(z_1)dz_1\right],$$

$$\begin{aligned} v^2(x_1) &= \int K^2(u)du\int\int\int\left[\sigma_0^2(x_1,x_2,x_3,x_4) + m^2(x_1,x_2,x_3,x_4)\right] \\ &\quad\times\frac{\left[q_2(x_2)q_{34}(x_3,x_4)\right]^2}{f(x_1,x_2,x_3,x_4)}\mu(dx_2)dx_3\mu(dx_4). \end{aligned}$$

ii) Furthermore, as n grows up to infinity, we have

$$\sqrt{nh_3^{p_2}}\left(\hat{\eta}_{34}(x_3,x_4) - \eta_{34}(x_3,x_4)\right) \to_d \mathcal{N}\left(b(x_3,x_4), v^2(x_3,x_4)\right) \tag{10.2.8}$$

with

$$\begin{aligned} b(x_3,x_4) &= \frac{1}{k!}\sum_{j=1}^{p_2}\int u_j^k L(u)du\left[(-1)^k\frac{\partial^k m_{34}}{\partial x_{3j}^k}(x_3,x_4)\right. \\ &\quad\left.+\int m_{34}(z_3,z_4)\frac{\partial^k q_{34}}{\partial z_{3j}^k}(z_3,z_4)dz_3\mu(dz_4)\right], \end{aligned}$$

and

$$\begin{aligned} v^2(x_3,x_4) &= f_4(x_4)\int L^2(u)du\int\int\left[\sigma_0^2(x_1,x_2,x_3,x_4) + m^2(x_1,x_2,x_3,x_4)\right] \\ &\quad\times\frac{\left[q_1(x_1)q_2(x_2)\right]^2}{f(x_1,x_2,x_3,x_4)}dx_1\mu(dx_2). \end{aligned}$$

Our result in (10.2.7) is a generalization of the one obtained in Theorem 1 from Fan, Härdle and Mammen (1998) to dependent observations. Furthermore, the result in (10.2.8) remarks that in the case of mixture between

continuous and discrete variables, the asymptotic variance of the marginal integration estimator suffers only from the dimensionality of the continuous variables. That is, the dimension of the discrete variables does not affect the rate of convergence of the estimator. Finally, we provide also an interesting result for the marginal integration estimator with all discrete covariables, $\hat{\eta}_2(x_2)$. The statistical properties of this estimator are given in the next result:

Theorem 2 *Consider assumptions (H.1), (H.2a), (H.3), (H.4), (H.5), (H.6), (H.7) and (H.8) hold, then*

$$\sqrt{n}\left(\hat{\eta}_2(x_2) - \eta_2(x_2)\right) \to_d \mathcal{N}\left(0, v^2(x_2)\right)$$

$$\begin{aligned} v^2(x_2) &= \left(f_2(x_2) - q_2(x_2)\right)^2 \int\int\int \left[\sigma_0^2(z_1,z_2,z_3,z_4) + m^2(z_1,z_2,z_3,z_4)\right] \\ &\quad \times \frac{\left[q_1(z_1)q_{34}(z_3,z_4)\right]^2}{f(z_1,z_2,z_3,z_4)} dz_1 dz_3 \mu(dz_4) - \eta_2^2(x_2), \end{aligned}$$

as n tends to infinity.

Note that although the multivariate nonparametric estimator contains some smoothing, the bias of $\hat{\eta}_2(x_2)$ is exactly equal to zero. This is because the marginal integration estimator of $\eta_2(x_2)$ is obtained by integrating out all directions that contain some smoothness.

10.3 A Semiparametric Estimator of an Additive Partially Linear Model

As already indicated in the Introduction, the presence of discrete explanatory in nonparametric regression problems can be approached by rewriting the model as a semiparametric one. This semiparametric model combines a linear parametric part (with discrete covariates) plus a nonparametric term that contains the continuous variables. The partially linear model has long tradition in the econometrics literature and it was fully analyzed in an i.i.d. context in Robinson (1988) among others, while recent advances in the dependent setting can be found in Aneiros et al. (2003). Furthermore, if an additional restriction of additivity in the nonparametric part is added, then we obtain the so called additive partially linear model. Examples of this model have been considered in Opsomer (1999) and Li (2000). Although, as explained in Section 10.2, many econometric problems of interest do not admit the partially additive linear decomposition in this Section we adopt it and we obtain a root-n consistent semiparametric estimator of the parametric part. This estimator can be compared with other previous in the literature.

If in the econometric model introduced in Section 10.2 we impose the additional restrictions $m_2(x_2) = \sum_{l=1}^{q_1} m_{2l}(x_{2l})$ and, without loss of generality, $m_{2l}(x_{2l}) = \theta_l + \gamma_l x_{2l}$, then (10.2.3) has the following expression

$$Y_i = \omega + \sum_{l=1}^{q_1} \theta_l + m_1(X_{1i}) + \sum_{l=1}^{q_1} \gamma_l X_{2li} + m_{34}(X_{3i}, X_{4i}) + \epsilon_i. \qquad (10.3.1)$$

Note that in this context, the identification restriction $E[m_2(x_2)] = 0$ implies that $\theta_l = -\gamma_l E(X_{2l})$, for $l = 1, \cdots, q_1$. If we rewrite (10.3.1) under the previous restriction we obtain

$$Y_i = \omega + m_1(X_{1i}) + \sum_{l=1}^{q_1} \gamma_l (X_{2li} - E(X_{2l})) + m_{34}(X_{3i}, X_{4i}) + \epsilon_i. \quad (10.3.2)$$

In this model, it is of interest to estimate the components $\gamma_1, \gamma_2, \cdots, \gamma_{q_1}$ at root-n rate. Furthermore, in order to make inference it is interesting to obtain its asymptotic distribution. One problem is that the previous identification restriction introduces in the estimating equation quantities that are unknown for the researcher as the expected values $E(X_{2l}), \cdots, E(X_{2q_1})$. One way to solve this problem is to introduce the following assumption

(H.9) $q_2(x_2) = C$ in the support of X_2.

Note that other identification strategies are possible. For example, in Fan, Härdle and Mammen (1998), p. 952, for the sake of identification they make $\theta = \sum \omega + \sum_{l=1}^{q_1} \theta_l$ and they overestimate the quantities $m_1(\bullet)$ and $m_{34}(\bullet, \bullet)$ by an amount of θ.

Let $\{x_{2lj}\}_{j=1}^{J}$ be the set of all possible values that X_{2l} can take such that $f_{2l}(x_{2lj}) = P(X_{2l} = x_{2lj}) > 0$ for $j = 1, \cdots, J$. Then, the easiest way to define an estimator seems to us to choose the value of γ_l that minimizes the L_2 distance between the model estimated nonparametrically, $\hat{\eta}_{2l}(x_{2l})$, and its corresponding linear direction, $\gamma_l(x_{2l} - \bar{X}_{2l})$, i. e.

$$\hat{\gamma}_l = \operatorname{argmin} \sum_{j=1}^{J} \left(\gamma_l(x_{2lj} - \bar{X}_{2l}) - \hat{\eta}_{2l}(x_{2l})\right)^2,$$

where $\bar{X}_{2l} = \frac{1}{J}\sum_{j=1}^{J} x_{2lj}$. This idea was already explored in another context by Cristobal, Faraldo and Gonzalez-Manteiga (1987). Compared to others our estimator presents some advantages. First, its asymptotic properties are obtained under much weaker conditions. Mainly, lagged endogenous variables may appear as regressors. Second, the estimator is unique, and it does not depend on cells or predetermined sets of values that can take the discrete variable. The following result is shown in the Appendix

Theorem 3 *Consider assumptions (H.1), (H.2b), (H.3),(H.4), (H.5), (H.6), (H.7), (H.8) and (H.9) hold, then*

$$\sqrt{n}\left(\hat{\gamma}_l - \gamma_l\right) \rightarrow_d \mathcal{N}\left(0, v_l^2\right)$$

with

$$\begin{aligned} v_l^2 &= \frac{1}{J^2}\sum_{j=1}^{J}\left\{ f_{2l}(x_{2lj})\left(x_{2lj} - \bar{X}_{2l}\right)^2 \int\int\int \left[\sigma_0^2(z_1, x_{2lj}, z_3, z_4)\right.\right. \\ &\left.\left. + m^2(z_1, x_{2lj}, z_3, z_4)\right] \frac{\left[q_1(z_1)q_{34}(z_3, z_4)\right]^2}{f(z_1, x_{2lj}, z_3, z_4)} dz_1 dz_3 \mu(dz_4) \right\} - \eta_2^2(x_2), \end{aligned}$$

Appendix: Proofs

The main difficulty that we will meet along our proofs is to deal with possibly dependent variables. That means that, in addition to the usual bias and variance terms appearing in the classical i.i.d. setting, a third terms (which will be basically written as a sum of covariance terms) will systematically appear at each step of our calculous. For each of these additional covariance components we will make use of recent probabilistic tools for *alpha* mixing random variable. For instance, after suitable preliminary calculous, the asymptotic normality results will be obtained from a Central Limit Theorem stated by Rio (2000) while minor terms in our developments will be treated by mean of some covariance inequality for mixing variables such as those stated in Bosq (1998). In addition to thesetechnical difficulties coming from the dependence between the data, the second difficulty appearing in our proofs comes from the large variety of possibilities that we are allowing in our methodology (see the model (10.2.3)), since we wish to include all the situations mixing discrete and continuous explanatory variables with possible interactions between both of them.

Proof of Theorem 1.i.

We first state some notations. Let

$$\begin{aligned} \alpha_1(x_1) &= \int\int\int m(x_1, x_2, x_3, x_4) q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4); \\ \hat{\alpha}_1(x_1) &= \int\int\int \hat{m}_n(x_1, x_2, x_3, x_4) q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4); \end{aligned}$$

$$
\begin{aligned}
C_n &= \omega + \int\int\int \left(m_2(z_2) + m_{34}(z_3, z_4)\right) g_n\left(z_2, z_3, z_4\right) \mu(dz_2) dz_3 \mu(dz_4); \\
\hat{C}_n &= \int\int\int\int \hat{m}_n(x_1, x_2, x_3, x_4) q_1(x_1) q_2(x_2) q_{34}(x_3, x_4) dx_1 \mu(dx_2) dx_3 \mu(dx_4) \\
C &= \int m_1(x_1) q_1(x_1) dx_1;
\end{aligned}
$$

$$
\begin{aligned}
g_n\left(z_2, z_3, z_4\right) = & \int\int\int \mathbb{I}(x_2 = z_2) \frac{1}{h_3^{p_3}} L\left(\frac{x_3 - z_3}{h_3}\right) \mathbb{I}(x_4 = z_4) \\
& q_2(x_2) q_{34}(x_3, x_4) \mu(dx_2) dx_3 \mu(dx_4).
\end{aligned}
$$

Remark: By (H.3) we have $g_n(z_2, z_3, z_4) = q_2(z_2) q_{34}(z_3, z_4) + o(1)$.
We can also write

$$
\hat{\alpha}_1(x_1) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) \frac{\widetilde{Y}_{ni}}{f_1(X_{1i})}
$$

with

$$
\widetilde{Y}_{ni} = \frac{Y_i f_1(X_{1i})}{f\left(X_{1i}, X_{2i}, X_{3i}, X_{4i}\right)} g_n\left(X_{2i}, X_{3i}, X_{4i}\right).
$$

Then, we have written $\hat{\alpha}_1$ as a nonparametric estimator of $\widetilde{m}_n(\bullet) = E\left(\widetilde{Y}_{ni} | X_{1i} = \bullet\right)$, and we have

$$
\begin{aligned}
\widetilde{m}_n(x_1) &= m_1(x_1) + C_n, \\
\eta_1(x_1) &= m_1(x_1) - C, \\
\hat{\eta}_1(x_1) &= \hat{\alpha}_1(x_1) - \hat{C}_n.
\end{aligned}
$$

The proof of the asymptotic normality of $\hat{\eta}_1 - \eta_1$ is obtained by the proof of the three following points:

$$
\sqrt{n h_1^{p_1}} \left(\hat{\alpha}_1(x_1) - \widetilde{m}_n(x_1)\right) \to_d \mathcal{N}\left(b_1(x_1), v^2(x_1)\right), \tag{10.3.3}
$$

$$
E\left(\hat{C}_n - C_n - C\right) = h_1^k b_1 + o\left(h_1^k\right), \tag{10.3.4}
$$

$$
\mathrm{Var}\left(\hat{C}_n\right) = o\left(\frac{1}{n h_1^{p_1}}\right), \tag{10.3.5}
$$

where

$$
b_1 = \frac{1}{k!} \sum_{j=1}^{p_1} \int u_j^k K(u) du \int m_1(z_1) \frac{\partial^k q_1}{\partial z_{1j}^k}(z_1) dz_1,
$$

and where $b_1(x_1) = b(x_1) - b_1$.

Proof of (10.3.3)

For the bias part, integrating by substitution and using a Taylor expansion of m_1, we have

$$E\hat{\alpha}_1(x_1) - \widetilde{m}_n(x_1) = E\left(\frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_1}{h_1}\right) \frac{\widetilde{Y}_{n1}}{f_1(X_1)}\right) - \widetilde{m}_n(x_1)$$

$$= h_1^k \frac{(-1)^k}{k!} \sum_{j=1}^{p_1} \int u_j^k K(u) du \frac{\partial^k m_1}{\partial x_{1j}^k}(x_1) + o(h_1^k). \tag{10.3.6}$$

Now we have to compute the variance of $\hat{\alpha}_1(x_1)$.

$$\mathrm{Var}\left(\hat{\alpha}_1(x_1)\right) = \frac{1}{nh_1^{2p_1}} \mathrm{Var}\left(\Delta_i\right) + \frac{2}{(nh_1)^{2p_1}} \sum \sum_{1 \le i < j \le n} \mathrm{Cov}\left(\Delta_i, \Delta_j\right), \tag{10.3.7}$$

where

$$\Delta_i = K\left(\frac{x_1 - X_{1i}}{h_1}\right) \frac{\widetilde{Y}_{ni}}{f_1(X_{1i})} - E\left\{K\left(\frac{x_1 - X_{1i}}{h_1}\right) \frac{\widetilde{Y}_{ni}}{f_1(X_{1i})}\right\}.$$

Integrating by substitution and by (H.3) and (H.4), we have $E\Delta_i = \mathcal{O}\left(h_1^{p_1}\right)$ and then

$$\lim_{n \to \infty} nh_1^{p_1} \left[\frac{1}{nh_1^{2p_1}} E\Delta_i^2\right] = 0. \tag{10.3.8}$$

Integrating by substitution and using (H.3), (H.4) and (H.5), we obtain that

$$E\Delta_i^2 = v^2(x_1) + o\left(\frac{1}{nh_1^{p_1}}\right), \tag{10.3.9}$$

with

$$v^2(x_1) = \int K^2(u) du \int\int\int \left[\sigma_0^2(x_1, x_2, x_3, x_4) + m^2(x_1, x_2, x_3, x_4)\right] \times \frac{\left[q_2(x_2) q_{34}(x_3, x_4)\right]^2}{f(x_1, x_2, x_3, x_4)} \mu(dx_2) dx_3 \mu(dx_4).$$

Now we will look at the covariance terms. Integrating by substitution and using (H.3), (H.4) and (H.5), we have

$$\mathrm{Cov}\left(\Delta_i, \Delta_j\right) = \mathcal{O}\left(h_1^{2p_1}\right). \tag{10.3.10}$$

On the other hand, by (H.8), $E\left|\widetilde{Y}_{ni}\right|^\beta \le M < \infty$, and then $E\left|\Delta_i\right|^\beta \le M < \infty$, that allows us to use the covariance inequality for strongly mixing

processes (see e.g. Bosq, 1998, Corollary 1.1, p. 21). Then we have

$$|\mathrm{Cov}\,(\Delta_i, \Delta_j)| \leq M\alpha^{\frac{\beta-2}{\beta}}\,(|i-j|)\,.$$

Now we proceed as in Bosq (1998, p. 43) and we introduce a sequence u_n of integers that allows to write

$$\begin{aligned}\sum_{1\leq i<j\leq n} \mathrm{Cov}\,(\Delta_i, \Delta_j) &= \sum_{|i-j|\leq u_n} \mathrm{Cov}\,(\Delta_i, \Delta_j) + \sum_{|i-j|>u_n} \mathrm{Cov}\,(\Delta_i, \Delta_j) \\ &= \mathcal{O}\left(h_1^{2p_1} n u_n + n^2 \alpha^{\frac{\beta-2}{\beta}}(u_n)\right).\end{aligned}$$

Choosing $u_n = (h_1^{p_1} \log n)^{-1}$ gives with (H.2b)

$$\lim_{n\to\infty} n h_1^{p_1} \left[\frac{1}{n h_1^{2p_1}} \sum \sum_{1\leq i<j\leq n} \mathrm{Cov}\,(\Delta_i, \Delta_j)\right] = 0. \tag{10.3.11}$$

Because of (H.2b) we can now apply a CLT for mixing random variables (see e. g. Rio, 2000, Theorem 4.2., p. 64). So, the relations (10.3.6), (10.3.7), (10.3.8), (10.3.9), (10.3.10), and (10.3.11) lead directly to (10.3.3).

Proof of (10.3.4)

Computing $E\hat{m}_n(x_1, x_2, x_3, x_4)$ in a standard way, and since the regression function m is additive, we arrive at

$$E\left(\hat{C}_n - C_n\right) = \int m_1(z_1) \int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - z_1}{h_1}\right) q_1(x_1) dx_1 dz_1.$$

A Taylor expansion of q_1 leads directly to (10.3.4).

Proof of (10.3.5)

We have to compute

$$\mathrm{Var}\left(\hat{C}_n\right) = \frac{1}{n}\mathrm{Var}\,(U_1) + \frac{2}{n^2} \sum \sum_{1\leq i<j\leq n} \mathrm{Cov}\,(U_i, U_j)\,,$$

where

$$U_i = \frac{Y_i}{f\,(X_{1i}, X_{2i}, X_{3i}, X_{4i})} p_n\,(X_{1i})\, g_n\,(X_{2i}, X_{3i}, X_{4i})$$

and

$$p_n\,(X_{1i}) = \int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) q_1(x_1) dx_1.$$

By (H.3), (H.4), (H.5) and integrating by substitution, we can see that $\mathrm{Var}(U_1) = \mathcal{O}(1)$ and $E|U_i|^\beta \leq M < \infty$. Then, the covariance terms can be treated exactly as we did before for getting (10.3.11) by Rio's inequality and by condition (H.2b). This is enough to see that the relation (10.3.5) is proved.

Proof of Theorem 1.ii.

It remains now to prove the second part of Theorem 1, namely the equation (10.2.8). The proof follows the same lines as for the estimation of the additive component m_1, because m_{34} depends on some continuous random variable. So we will just give the main steps. Introduce the notations:

$$\begin{aligned}
\alpha_{34}(x_3,x_4) &= \int\int m(x_1,x_2,x_3,x_4)\, q_1(x_1)q_2(x_2)dx_1\mu(dx_2);\\
\hat{\alpha}_{34}(x_3,x_4) &= \int\int \hat{m}_n(x_1,x_2,x_3,x_4)\, q_1(x_1)q_2(x_2)dx_1\mu(dx_2);\\
D_n &= \omega + \int\int (m_1(z_1)+m_2(z_2))\, g_n(z_1,z_2)\, dz_1\mu(dz_2);\\
\hat{D}_n &= \int\int\int\int \hat{m}_n(x_1,x_2,x_3,x_4)\, q_1(x_1)q_2(x_2)\\
&\quad \times q_{34}(x_3,x_4)dx_1\mu(dx_2)dx_3\mu(dx_4);\\
D &= \int\int m_{34}(x_3,x_4)q_{34}(x_3,x_4)dx_3\mu(dx_4);\\
g_n(z_1,z_2) &= \int\int \frac{1}{h_1^{p_1}} K\left(\frac{x_1-z_1}{h_1}\right)\mathbb{I}(x_2=z_2)q_1(x_1)q_2(x_2)dx_1\mu(dx_2).
\end{aligned}$$

Remark: By (H.3), we have $g_n(z_1,z_2) = q_1(z_1)q_2(z_2) + o(1)$.

We can also write

$$\hat{\alpha}_{34}(x_3,x_4) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_3^{p_2}} L\left(\frac{x_3-X_{3i}}{h_3}\right)\mathbb{I}(x_4=z_4)\frac{\widetilde{Y}_{ni}}{f_4(X_{4i})f_c(X_{3i}|X_{4i})}$$

with

$$\widetilde{Y}_{ni} = \frac{Y_i f_3(X_{3i})}{f_c(X_{1i},X_{3i}|X_{2i},X_{4i})f_2(X_{2i})} g_n(X_{1i},X_{2i}).$$

Then, we have rewritten $\hat{\alpha}_{34}$ as a nonparametric estimator of $\widetilde{m}_n(\bullet,\bullet) = E\left(\widetilde{Y}_{ni}\middle|(X_{3i},X_{4i})=(\bullet,\bullet)\right)$, and we have

$$\begin{aligned}
\widetilde{m}_n(x_3,x_4) &= m_{34}(x_3,x_4) + D_n,\\
\eta_{34}(x_3,x_4) &= m_{34}(x_3,x_4) - D,\\
\hat{\eta}_{34}(x_3,x_4) &= \hat{\alpha}_{34}(x_3,x_4) - \hat{D}_n.
\end{aligned}$$

The proof of the asymptotic normality of $\hat{\eta}_{34} - \eta_{34}$ will be obtained from the three following points that can be proved exactly as results (10.3.3), (10.3.4) and (10.3.5):

$$\sqrt{nh_3^{p_2}}\left(\hat{\alpha}_{34}(x_3,x_4) - \widetilde{m}_n(x_3,x_4)\right) \to_d \mathcal{N}\left(b_{34}(x_3,x_4), v^2(x_3,x_4)\right), \quad (10.3.12)$$

$$E\left(\hat{D}_n - D_n - D\right) = h_3^k b_3 + o\left(h_3^k\right), \quad (10.3.13)$$

$$\mathrm{Var}\left(\hat{D}_n\right) = o\left(\frac{1}{nh_3^{p_2}}\right), \quad (10.3.14)$$

with

$$b_{34}(x_3,x_4) = h_3^k \frac{(-1)^k}{k!} \sum_{j=1}^{p_2} \int u_j^k L(u)du \frac{\partial^k m_{34}}{\partial x_{3j}^k}(x_3,x_4) + o(h_3^k)$$

$$b_3 = \frac{1}{k!} \sum_{j=1}^{p_2} \int u_j^k L(u)du \int m_{34}(z_3,z_4) \frac{\partial^k q_{34}}{\partial x_{3j}^k}(x_3,x_4) dz_3 \mu(dz_4).$$

Proof of Theorem 2

To prove the asymptotic normality of $\eta_2 - \hat{\eta}_2$ we have to show the following relationships

$$\sqrt{n}\left(\hat{\eta}_2(x_2) - E\hat{\eta}_2(x_2)\right) \to_d \mathcal{N}\left(0, v^2(x_2)\right) \quad (10.3.15)$$

$$E\hat{\eta}_2(x_2) = \eta_2(x_2). \quad (10.3.16)$$

Proof of (10.3.15)

We write

$$\hat{\eta}_2(x_2) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{I}(x_2 = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})}\right)\widetilde{Z}_{ni} \equiv \frac{1}{n}\sum_{i=1}^{n}\Delta_i$$

where

$$\begin{aligned}\widetilde{Z}_{ni} &= Y_i \int\int\int \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - X_{1i}}{h_1}\right) \frac{1}{h_3^{p_2}} L\left(\frac{x_3 - X_{3i}}{h_3}\right) \mathbb{I}(x_4 = X_{4i}) \\ &\quad \times \frac{q_1(x_1) q_{34}(x_3,x_4)}{f_c(X_{1i}, X_{3i} | X_{2i}, X_{4i}) f_4(X_{4i})} dx_1 dx_3 \mu(dx_4).\end{aligned}$$

Let us first compute the variance term

$$\mathrm{Var}\left(\hat{\alpha}_2(x_2)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(\Delta_i\right) + \frac{2}{n^2}\sum\sum_{1\le i<j\le n}\mathrm{Cov}\left(\Delta_i, \Delta_j\right), \quad (10.3.17)$$

Using (10.3.16) we directly have

$$E\Delta_i = E\hat{\eta}_2(x_2) = \eta_2(x_2). \tag{10.3.18}$$

Integrating by substitution and using (H.3), we obtain

$$\begin{aligned} E\Delta_i^2 &= (f_2(x_2) - q_2(x_2))^2 \int\int\int \left[\sigma_0^2(z_1, x_2, z_3, z_4) + m^2(z_1, x_2, z_3, z_4)\right] \\ &\quad \times \frac{[q_1(z_1)q_{34}(z_3, z_4)]^2}{f(z_1, x_2, z_3, z_4)} dz_1 dz_3 \mu(dz_4) + o(1). \end{aligned} \tag{10.3.19}$$

Now, for the computation of the covariance terms, by using (H.8) we obtain that $E\left|\widetilde{Z}_{ni}\right|^\beta \leq M < \infty$ and then $E\left|\Delta_i\right|^\beta \leq M < \infty$, that allows us to use the covariance inequality for strongly mixing processes (see e.g. Bosq, 1998, Corollary 1.1, p. 21). Then we have

$$\left|\mathrm{Cov}\left(\Delta_i, \Delta_j\right)\right| \leq M\alpha^{\frac{\beta-2}{\beta}}\left(|i-j|\right).$$

By a simple computation, and using (H.2a), we obtain

$$\left|\frac{1}{n^2}\sum\sum_{1\leq i<j\leq n} \mathrm{Cov}\left(\Delta_i, \Delta_j\right)\right| \leq Mn^{1-\alpha\frac{\beta-2}{\beta}}. \tag{10.3.20}$$

Finally, using a Central Limit Theorem for strongly mixing processes (Rio, 2000, Theorem 4.2., p. 64) with relations (10.3.17), (10.3.18), (10.3.19), (10.3.20) and with (H.2a) we get directly (10.3.15).

Proof of (10.3.16)

We first compute the expectation of $\hat{m}_n(x_1, x_2, x_3, x_4)$: $E\left\{\hat{m}_n(x_1, x_2, x_3, x_4)\right\}$

$$= m\left(z_1, x_2, z_3, x_4\right) \frac{1}{h_1^{p_1}} K\left(\frac{x_1 - z_1}{h_1}\right) \frac{1}{h_3^{p_2}} L\left(\frac{x_3 - z_3}{h_3}\right) dz_1 dz_3,$$

and then, since the regression function is additive we easily obtain that

$$E\left\{\hat{\eta}_2(x_2)\right\} = m_2(x_2) - \int m_2(x_2)q_2(x_2)\mu(dx_2) = \eta_2(x_2),$$

and (10.3.16) is proved.

Proof of Theorem 3

Note that just to simplify notations we have removed the index l form X_{2l}. That is, along the proof we will use X_{2i} instead of X_{2li} and f_2 instead of f_{2l}. This is done just for notational convenience and without loss of generality. Let us define

$$\bar{X}_2 = \frac{1}{J}\sum_{j=1}^{J} x_{2j},$$

and

$$\sigma^2_{X_2} = \frac{1}{J}\sum_{j=1}^{J}\left(x_{2j} - \bar{X}_2\right)^2.$$

The estimator $\hat{\gamma}_l$ of γ_l is defined as follows:

$$\hat{\gamma}_l = \frac{\sum_{j=1}^{J}\hat{\eta}_2\left(x_{2j}\right)\left(x_{2j} - \bar{X}_2\right)}{\sum_{j=1}^{J}\left(x_{2j} - \bar{X}_2\right)^2}$$

The bias term is not difficult to compute. Because of Theorem 2, we have

$$\forall x_{2j},\ E\left\{\hat{\eta}_2(x_{2j})\right\} = \eta_2(x_{2j}),$$

while, by assumption (H.9), the choice made for q_2 allows to see that

$$\eta_2(x_{2j}) = \gamma_l\left(x_{2j} - \bar{X}_2\right).$$

Clearly, this implies that we have

$$E\hat{\gamma}_l = \gamma_l.$$

So the only remaining question is to compute the variance term, $\mathrm{Var}\left(\hat{\gamma}_l\right)$. It can be written as

$$\sigma^4_{X_2}\mathrm{Var}\left(\hat{\gamma}_l\right) = \frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\mathrm{Cov}\left(\hat{\eta}_2\left(x_{2j}\right)\left(x_{2j} - \bar{X}_2\right), \hat{\eta}_2\left(x_{2j'}\right)\left(x_{2j'} - \bar{X}_2\right)\right).$$

Let, as in the proof of (10.3.15), introduce the quantity

$$\Delta_i\left(x_{2j}\right) = \left(\frac{\mathbb{I}(x_{2j} = X_{2i}) - q_2(X_{2i})}{f_2(X_{2i})}\right)\tilde{Y}_{ni}.$$

Then,

$$\begin{aligned}
&\operatorname{Cov}\left(\hat{\eta}_2\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right), \hat{\eta}_2\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right) \\
&= \operatorname{Cov}\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right), \frac{1}{n}\sum_{k=1}^{n}\Delta_k\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Cov}\left(\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right), \Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right) \quad (10.3.21) \\
&+\frac{1}{n^2}\sum\sum_{i\neq k}\operatorname{Cov}\left(\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right), \Delta_k\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right). \quad (10.3.22)
\end{aligned}$$

Let us now look at the computation of (10.3.21). Note first that we have

$$\begin{aligned}
&\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\frac{1}{n}\operatorname{Cov}\left(\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right), \Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right) \\
&= \frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\left[E\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right)\Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right. \\
&\left.-E\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right)E\Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\right]
\end{aligned}$$

Using the calculations of the proof of (10.3.15), we easily obtain

$$\begin{aligned}
&\frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}E\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right)E\Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right) \\
&= \frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}m_2(x_{2j})\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right)m_2\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right) \\
&= \frac{1}{n}\gamma_l^2\left[\frac{1}{J}\sum_{j=1}^{J}m_2(x_{2j})(x_{2j}-\bar{X}_2)\right]^2 \\
&= \frac{1}{n}\gamma_l^2\sigma_{X_2}^4. \quad (10.3.23)
\end{aligned}$$

On the other hand we have

$$\begin{aligned}
&\frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J} E\Delta_i\left(x_{2j}\right)\left(x_{2j}-\bar{X}_2\right)\Delta_i\left(x_{2j'}\right)\left(x_{2j'}-\bar{X}_2\right)\\
=\;&\frac{1}{n}E\left[\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\left(\frac{\mathbb{I}(x_{2j}=X_{2i})-q_2(X_{2i})}{f_2(X_{2i}}\right)\tilde{Y}_{ni}\left(x_{2j}-\bar{X}_2\right)\right.\\
&\left.\times\left(\frac{\mathbb{I}(x_{2j'}=X_{2i})-q_2(X_{2i})}{f_2(X_{2i}}\right)\tilde{Y}_{ni}\left(x_{2j'}-\bar{X}_2\right)\right]\\
=\;&\frac{1}{n}E\left[\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\frac{\widetilde{Y}_{ni}^2}{f_2^2(X_{2i})}\left(x_{2j}-\bar{X}_2\right)\left(x_{2j'}-\bar{X}_2\right)\left\{\mathbb{I}(x_{2j}=X_{2i})\mathbb{I}(x_{2j'}\right.\right.\\
&\left.\left.=X_{2i})-\mathbb{I}(x_{2j}=X_{2i})q_2(X_{2i})-\mathbb{I}(x_{2j'}=X_{2i})q_2(X_{2i})+q_2^2(X_{2i})\right\}\right]\\
=\;&\frac{1}{n}E\left[\frac{1}{J^2}\sum_{j=1}^{J}\sum_{j'=1}^{J}\frac{\widetilde{Y}_{ni}^2}{f_2^2(X_{2i})}\left(x_{2j}-\bar{X}_2\right)\left(x_{2j'}-\bar{X}_2\right)\right.\\
&\left.\mathbb{I}(x_{2j}=X_{2i})\mathbb{I}(x_{2j'}=X_{2i})\right]\\
=\;&\frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}E\left[\frac{\widetilde{Y}_{ni}^2}{f_2^2(X_{2i})}\left(x_{2j}-\bar{X}_2\right)^2\mathbb{I}(x_{2j}=X_{2i})\right].
\end{aligned}$$

Integrating by substitution and using (H.3), (H.4) and (H.5) give

$$\begin{aligned}
&\frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}E\left[\frac{\widetilde{Y}_{ni}^2}{f_2^2(X_{2i})}\left(x_{2j}-\bar{X}_2\right)^2\mathbb{I}(x_{2j}=X_{2i})\right]\\
=\;&\frac{1}{n}\frac{1}{J^2}\sum_{j=1}^{J}\left\{f_2(x_{2j})\left(x_{2j}-\bar{X}_2\right)^2\int\int\int\left[\sigma_0^2\left(z_1,x_{2j},z_3,z_4\right)+\right.\right.\\
&\left.\left.m^2\left(z_1,x_{2j},z_3,z_4\right)\right]\frac{\left[q_1(z_1)q_{34}(z_3,z_4)\right]^2}{f(z_1,x_{2j},z_3,z_4)}dz_1dz_3\mu(dz_4)+o(1)\right\}.
\end{aligned}$$

Finally,

$$\begin{aligned}
\lim_{n\to\infty}\mathrm{Var}\left(\hat{\gamma}_l\right)\;=\;&\tfrac{1}{J^2}\textstyle\sum_{j=1}^{J}\left\{f_2(x_{2j})\left(x_{2j}-\bar{X}_2\right)^2\int\int\int\left[\sigma_0^2\left(z_1,x_{2j},z_3,z_4\right)\right.\right.\\
&\left.\left.+m^2\left(z_1,x_{2j},z_3,z_4\right)\right]\tfrac{\left[q_1(z_1)q_{34}(z_3,z_4)\right]^2}{f(z_1,x_{2j},z_3,z_4)}dz_1dz_3\mu(dz_4)\right\}-\gamma_l^2.
\end{aligned}$$

It remains just to look at the computation of (10.3.22). Proceeding as in (10.3.20), we have

$$\frac{1}{n^2}\sum\sum_{i\neq k}\mathrm{Cov}\left(\Delta_i(x_{2j})(x_{2j}-\bar{X}_2),\Delta_k(x_{2j'})(x_{2j'}-\bar{X}_2)\right)=o\left(\frac{1}{n}\right).$$

We can write $\hat{\gamma}_l$ as

$$\hat{\gamma}_l = \frac{1}{n}\sum_{i=1}^{n} \delta_i,$$

where

$$\delta_i = \frac{1}{J\sigma^2_{X_2}} \Delta_i(x_{2j})(x_{2j} - \bar{X}_2),$$

and then, applying the central limit theorem for strongly mixing processes (Rio, 2000, Theorem 4.2., p. 64), our result is proved.

Bibliography

Ahmad, I. A. & Cerrito, P. B. (1994) Nonparametric estimation of joint discrete-continuous probability densities with applications. *Journal of Statistical Planning and Inference* **41**, 349–364.

Andrews, D. W. K. & Whang, Y.-J. (1990) Additive interactive regression models: circumvention of the curse of dimensionality. *Econometric Theory* **6**, 466–479.

Aneiros Perez, G., Gonzalez Manteiga, W. & Vieu, P. (2003). Estimation and testing in a partial linear regression model with long-memory errors. *Bernoulli*, in print.

Bierens, H. J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 699–707.

Bosq, D. (1998) *Nonparametric statistics for stochastic processes: Estimation and prediction.* Lecture Notes in Statistics, 110. Springer-Verlag, New York.

Camlong-Viot, Ch., Sarda, P. & Vieu, Ph. (2000) Additive time series: the kernel integration method. *Mathematical Methods of Statistics* **9**, 358–375.

Cristóbal, J. A., Faraldo, P. & Gonzalez-Manteiga, W. (1987) A class of linear regression parameter estimators constructed by nonparametric estimation. *Annals of Statistics* **15**, 603–609.

Delgado, M. A. & Mora, J. (1995) Nonparametric and semiparametric estimation with discrete regressors. *Econometrica* **63**, 1477–1484.

Fan, J., Härdle, W. & Mammen, E. (1998) Direct estimation of low - dimensional components in additive models. *The Annals of Statistics* **26**, 943–971.

Grund, B. & Hall, P. (1993) On the performance of kernel estimators for high-dimensional, sparse binary data, *Journal of Multivariate Analysis* **44**, 321–344.

Györfi, L., Härdle, W., Sarda, P. & Vieu, P. (1989). *Nonparametric curve estimation from time series*, Lecture Notes in Statistics, 60. Springer-Verlag, New York.

Hall, P. (1981) On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287–294.

Hastie, T. J. & Tibshirani, R. (1990) *Generalized Additive Models.* Chapman and Hall, London.

Horowitz, J. (1998) *Semiparametric methods in Econometrics.* Lecture Notes in Statistics, Springer Verlag, New York.

Jones, M. C., Davies, S. J. & Park, B. U. (1994) Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.* **89**, 825–832.

Li, Q. (2000) Efficient estimation of partially linear models. *International Economic Review* **41**, 1073–1091.

Linton, O. & Nielsen, J. P. (1995) A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93–100.

Mammen, E. Linton, O. & Nielsen, J. P. (1999) The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Annals of Statistics* **27**, 1443–1490.

Newey, W. K. (1994) Kernel estimation of partial means. *Econometric Theory* **10**, 233–253.

Newey, W. K. (1995) Convergence for series estimators. In G. S. Maddala, P. C. B. Phillips & T. N. Srinavsan (eds.), *Statistical methods of Economics and Quantitative Economics: Essays in Honor of C. R. Rao*, 254–275.

Opsomer, J. D. & Ruppert, D. (1994) Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics* **25**, 186–211.

Opsomer, J. D. & Ruppert, D. (1998) A fully automated bandwidth selection method for fitting additive models. *J. Amer. Statist. Assoc.* **93**, 605–619.

Opsomer, J. D. (2000) Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis* **73**, 166–179.

Racine, J. & Li, Q. (2000) Nonparametric estimation of regression functions with both categorical and continuous data. Preprint.

Rio, E. (2000) *Théorie asymptotique des processus aléatoires faiblement dépendants.* Mathematiques et Applications. Springer Verlag, New York.

Robinson, P. M. (1988) Root-n consistent semiparametric regression. Econometrica **56**, 931–954.

Rosenblatt, M. (1956) A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. U.S.A.* **42**, 43–47.

Sperlich, S., Tjostheim, D. & Yang, L. (2002) Nonparametric estimation and testing of interactions in additive models. *Econometric Theory* **18**, 197–251.

Stone, C. J. (1985) Additive regression and other nonparametric models. *Annals of Statistics* **14**, 592–606.

Stone, C. J., Hansen, M. H., Kooperberg, C. & Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* **25**, 1371–1470.

Tjostheim, D. & Auestad, B. (1994) Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89**, 1398–1409.

Vieu, Ph. (1991) Quadratic errors for nonparametric estimators under dependence. *Journal of Multivariate Analysis* **39**, 324–347.

Wahba, G. (1990) *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics, **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.