Springer Series in Statistics

Smoothing Spline ANOVA Models

Second Edition

Springer

Springer Series in Statistics 297

For further volumes:
http://www.springer.com/series/692

Chong Gu

Smoothing Spline ANOVA Models

Second Edition

Springer

Chong Gu
Department of Statistics
Purdue University
West Lafayette, IN 47907
USA

ISSN 0172-7397
ISBN 978-1-4614-5368-0 ISBN 978-1-4614-5369-7 (eBook)
DOI 10.1007/978-1-4614-5369-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012950795

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my father
For the books and the bookcases

Preface to the First Edition

Thirty years have passed since the pioneering work of Kimeldorf and Wahba (1970a, 1970b, 1971) and Good and Gaskins (1971), and during this time, a rich body of literature has been developed on smoothing methods with roughness penalties. There have been two books solely devoted to the subject prior to this one, of which Wahba (1990) compiled an excellent synthesis for work up to that date, and Green and Silverman (1994) provided a mathematically gentler introduction to the field through regression models that are largely univariate.

Much has happened in the past decade, and more has been done with the penalty method than just regression. In this book, I have tried to assemble a comprehensive treatment of penalty smoothing under a unified framework. Treated are (i) regression with Gaussian and non-Gaussian responses as well as with censored lifetime data, (ii) density and conditional density estimation under a variety of sampling schemes, and (iii) hazard rate estimation with censored lifetime data and covariates. The unifying themes are the general penalized likelihood method and the construction of multivariate models with certain ANOVA decompositions built in. Extensive discussions are devoted to model (penalty) construction, smoothing parameter selection, computation, and asymptotic convergence. There are, however, many omissions, and the selection and treatment of topics solely reflect my personal preferences and views. Most of the materials have appeared in the literature, but a few items are new, as noted in the bibliographic notes at the end of the chapters.

An adequate treatment of model construction in the context requires some elementary knowledge of reproducing kernel Hilbert spaces, of which a self-contained introduction is included early in the book; the materials should be accessible to a second-year graduate student with a good training in calculus and linear algebra. Also assumed is a working knowledge of basic statistical inference such as linear models, maximum likelihood estimates, etc. To better understand materials on hazard estimation, prior knowledge of basic survival analysis would also help.

Most of the computational and data analytical tools discussed in the book are implemented in R, an open-source clone of the popular S/Splus language. Code for regression is reasonably polished and user-friendly and has been distributed in the R package `gss` available through CRAN, the Comprehensive R Archive Network, with the master site at

http://cran.r-project.org

The use of `gss` facilities is illustrated in the book through simulated and real-data examples.

Remaining on my wish list are (i) polished, user-friendly software tools for density estimation and hazard estimation, (ii) fast computation via approximate solutions of penalized likelihood problems, and (iii) handling of parametric random effects such as those appearing in longitudinal models and hazard models with frailty. All of the above are under active development and could be addressed in a later edition of the book or, sooner than that, in later releases of `gss`.

The book was conceived in Spring 1996 when I was on leave at the Department of Statistics, University of Michigan, which offered me the opportunity to teach a course on the subject. Work on the book has been on and off since then, with much of the progress being made in the 1997–1998 academic year during my visit at the National Institute of Statistical Sciences, and in Fall 2000 when I was teaching a course on the subject at Purdue.

I am indebted to Grace Wahba, who taught me smoothing splines, and to Doug Bates, who taught me statistical computing. Bill Studden carefully read various drafts of Chaps. 1, 2, and 4; his questions alerted me to numerous accounts of mathematical sloppiness in the text and his suggestions led to much improved presentations. Detailed comments and suggestions by Nancy Heckman on a late draft helped me to fix numerous problems throughout the first five chapters and to shape the final organization of the book (e.g., the inclusion of §1.4). For various ways in which they helped, I would also like to thank Mary Ellen Bock, Jerry Davis, Nels Grevstad, Wensheng Guo, Alan Karr, Youngju Kim, Ping Ma, Jerry Sacks, Jingyuan Wang, Yuedong Wang, Jeff Wu, Dong Xiang, Liqing Yan, and the classes at Michigan and Purdue. Last but not least, I would like to thank the R Core Team, for creating a most enjoyable platform for statistical computing.

West Lafayette, Indiana Chong Gu
July 2001

Preface

When the first edition was published a decade ago, I wrote in the Preface:

> Remaining on my wish list are (i) polished, user-friendly software tools for density estimation and hazard estimation, (ii) fast computation via approximate solutions of penalized likelihood problems, and (iii) handling of parametric random effects such as those appearing in longitudinal models and hazard models with frailty.

I am happy to report that the wishes have been fulfilled, plus some more, and it is time to present an updated treatise on smoothing methods with roughness penalties.

The developments of software tools embodied in an R package `gss` have gone a long way in the past decade, with the user-interface polished, functionality expanded, and/or numerical efficiency improved from release to release. The primary objective of this new edition is to introduce extensive software illustrations to complement the theoretical and methodological discussions, so the reader not only can read about the methods but also can use them in everyday data analysis.

Newly developed theoretical, methodological, and computational techniques are integrated in a few new chapters and new sections, along with some previously omitted entries; due modifications are made in related chapters and sections to maintain coherence. Empirical studies are expanded, reorganized, and mostly rerun using the latest software.

Two appendices are also added. One appendix outlines the overall design of the R package `gss`. The other presents some conceptual critiques on a few issues concerning smoothing methods at large, which are potentially controversial.

Much of the new materials that went into this edition were taken from or inspired by collaborations or communications with Pang Du, Anouschka Foltz, Chun Han, Young-Ju Kim, Yi Lin, Ping Ma, Christophe Pouzat, Jingyuan Wang, and Tonglin Zhang, to whom I owe thanks. I can not thank enough the R Core Team, for creating and maintaining a most enjoyable platform for statistical computing.

West Lafayette, Indiana Chong Gu
August 2011

Contents

Preface to the First Edition **vii**

Preface **ix**

1 Introduction **1**
1.1 Estimation Problem and Method 2
1.1.1 Cubic Smoothing Spline 2
1.1.2 Penalized Likelihood Method 4
1.2 Notation 5
1.3 Decomposition of Multivariate Functions 6
1.3.1 ANOVA Decomposition and Averaging Operator . . 6
1.3.2 Multiway ANOVA Decomposition 7
1.3.3 Multivariate Statistical Models 10
1.4 Case Studies 12
1.4.1 Water Acidity in Lakes 12
1.4.2 AIDS Incubation 14
1.4.3 Survival After Heart Transplant 15
1.5 Scope 17
1.6 Bibliographic Notes 19
1.7 Problems 20

2 Model Construction **23**
2.1 Reproducing Kernel Hilbert Spaces 24
2.1.1 Hilbert Spaces and Linear Subspaces 24
2.1.2 Riesz Representation Theorem 29
2.1.3 Reproducing Kernel and Non-Negative Definite Function 29
2.2 Smoothing Splines on $\{1, \ldots, K\}$ 32
2.3 Polynomial Smoothing Splines on $[0, 1]$ 34
2.3.1 A Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$ 34
2.3.2 Computation of Polynomial Smoothing Splines . . . 36
2.3.3 Another Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$ 37
2.4 Smoothing Splines on Product Domains 40
2.4.1 Tensor Product Reproducing Kernel Hilbert Spaces 40
2.4.2 Reproducing Kernel Hilbert Spaces on $\{1, \ldots, K\}^2$ 41
2.4.3 Reproducing Kernel Hilbert Spaces on $[0, 1]^2$ 42
2.4.4 Reproducing Kernel Hilbert Spaces on $\{1, \ldots, K\} \times [0, 1]$ 44
2.4.5 Multiple-Term Reproducing Kernel Hilbert Spaces: General Form 45
2.5 Bayes Model 48
2.5.1 Shrinkage Estimates as Bayes Estimates 48
2.5.2 Polynomial Splines as Bayes Estimates 49
2.5.3 Smoothing Splines as Bayes Estimates 51
2.6 Minimization of Penalized Functional 51
2.6.1 Existence of Minimizer 52
2.6.2 Penalized and Constrained Optimization 53
2.7 Bibliographic Notes 54
2.8 Problems 56

3 Regression with Gaussian-Type Responses **61**
3.1 Preliminaries 62
3.2 Smoothing Parameter Selection 64
3.2.1 Unbiased Estimate of Relative Loss 65
3.2.2 Cross-Validation and Generalized Cross-Validation 67
3.2.3 Restricted Maximum Likelihood Under Bayes Model 70
3.2.4 Weighted and Replicated Data 72
3.2.5 Empirical Performance 74
3.3 Bayesian Confidence Intervals 75
3.3.1 Posterior Distribution 76
3.3.2 Confidence Intervals on Sampling Points 78
3.3.3 Across-the-Function Coverage 78

3.4 Computation: Generic Algorithms . . . 79
3.4.1 Algorithm for Fixed Smoothing Parameters . . . 80
3.4.2 Algorithm for Single Smoothing Parameter . . . 80
3.4.3 Algorithm for Multiple Smoothing Parameters . . . 82
3.4.4 Calculation of Posterior Variances . . . 84
3.5 Efficient Approximation . . . 85
3.5.1 Preliminaries . . . 85
3.5.2 Bayes Model . . . 86
3.5.3 Computation . . . 88
3.5.4 Empirical Choice of q . . . 90
3.5.5 Numerical Accuracy . . . 92
3.6 Software . . . 93
3.6.1 RKPACK . . . 93
3.6.2 R Package `gss`: `ssanova` and `ssanova0` Suites . . . 94
3.7 Model Checking Tools . . . 98
3.7.1 Cosine Diagnostics . . . 98
3.7.2 Examples . . . 99
3.7.3 Concepts and Heuristics . . . 103
3.8 Square Error Projection . . . 104
3.9 Case Studies . . . 106
3.9.1 Nitrogen Oxides in Engine Exhaust . . . 106
3.9.2 Ozone Concentration in Los Angeles Basin . . . 107
3.10 Computation: Special Algorithms . . . 111
3.10.1 Fast Algorithm for Polynomial Splines . . . 112
3.10.2 Iterative Algorithms and Monte Carlo Cross-Validation . . . 114
3.11 Bibliographic Notes . . . 115
3.12 Problems . . . 118

4 More Splines **125**
4.1 Partial Splines . . . 126
4.2 Splines on the Circle . . . 127
4.2.1 Periodic Polynomial Splines . . . 127
4.2.2 Splines as Low-Pass Filters . . . 128
4.2.3 More on Asymptotics of §3.2 . . . 130
4.3 Thin-Plate Splines . . . 134
4.3.1 Semi-Kernels for Thin-Plate Splines . . . 135
4.3.2 Reproducing Kernels for Thin-Plate Splines . . . 136
4.3.3 Tensor Product Splines with Thin-Plate Marginals . . . 139
4.3.4 Case Study: Water Acidity in Lakes . . . 140
4.4 Splines on the Sphere . . . 143
4.4.1 Spherical Harmonics . . . 143
4.4.2 Laplacian on the Sphere and Spherical Splines . . . 144

4.4.3 Reproducing Kernels in Closed Forms 146
4.4.4 Case Study: Global Temperature Map 147
4.5 L-Splines 149
4.5.1 Trigonometric Splines 150
4.5.2 Chebyshev Splines 153
4.5.3 General Construction 157
4.5.4 Case Study: Weight Loss of Obese Patient 161
4.5.5 Fast Algorithm 165
4.6 Bibliographic Notes 166
4.7 Problems 167

5 Regression with Responses from Exponential Families 175
5.1 Preliminaries 176
5.2 Smoothing Parameter Selection 177
5.2.1 Performance-Oriented Iteration 178
5.2.2 Direct Cross-Validation 181
5.3 Inferential Tools 184
5.3.1 Approximate Bayesian Confidence Intervals 185
5.3.2 Kullback-Leibler Projection 186
5.4 Software, Customization, and Empirical Performance 187
5.4.1 R Package `gss`: `gssanova`, `gssanova0`, and `gssanova1` Suites 187
5.4.2 Binomial Family 188
5.4.3 Poisson Family 191
5.4.4 Gamma Family 193
5.4.5 Inverse Gaussian Family 196
5.4.6 Negative Binomial Family 199
5.5 Case Studies 202
5.5.1 Eruption Time of Old Faithful 202
5.5.2 Spectrum of Yearly Sunspots 203
5.5.3 Progression of Diabetic Retinopathy 205
5.5.4 Colorectal Cancer Mortality Rate 208
5.6 Bibliographic Notes 210
5.7 Problems 212

6 Regression with Correlated Responses 215
6.1 Models for Correlated Data 216
6.1.1 Random Effects 216
6.1.2 Stationary Time Series 216
6.2 Mixed-Effect Models and Penalized Joint Likelihood 217
6.2.1 Smoothing Matrices 218
6.2.2 Bayes Model 219
6.2.3 Optimality of Generalized Cross-Validation 219
6.2.4 Empirical Performance 221

6.2.5 Non-Gaussian Regression . . . 222
6.2.6 R Package gss: Optional Argument random . . . 222
6.3 Penalized Likelihood with Correlated Data . . . 223
6.3.1 Bayes Model . . . 223
6.3.2 Extension of Cross-Validation . . . 225
6.3.3 Optimality of Cross-Validation . . . 226
6.3.4 Empirical Performance . . . 228
6.3.5 R Package gss: ssanova9 Suite . . . 230
6.4 Case Studies . . . 231
6.4.1 Treatment of Bacteriuria . . . 231
6.4.2 Ozone Concentration in Los Angeles Basin . . . 232
6.5 Bibliographic Notes . . . 233
6.6 Problems . . . 235

7 Probability Density Estimation 237
7.1 Preliminaries . . . 238
7.2 Poisson Intensity . . . 242
7.3 Smoothing Parameter Selection . . . 243
7.3.1 Kullback-Leibler Loss . . . 243
7.3.2 Cross-Validation . . . 244
7.3.3 Empirical Performance . . . 246
7.4 Computation, Inference, and Software . . . 247
7.4.1 Newton Iteration . . . 247
7.4.2 Numerical Integration . . . 248
7.4.3 Kullback-Leibler Projection . . . 250
7.4.4 R Package gss: ssden Suite . . . 250
7.5 Case Studies . . . 253
7.5.1 Buffalo Snowfall . . . 253
7.5.2 Eruption Time of Old Faithful . . . 254
7.5.3 AIDS Incubation . . . 255
7.6 Biased Sampling and Random Truncation . . . 257
7.6.1 Biased and Truncated Samples . . . 257
7.6.2 Penalized Likelihood Estimation . . . 258
7.6.3 Empirical Performance . . . 260
7.6.4 R Package gss: ssden Suite . . . 260
7.6.5 Case Study: AIDS Incubation . . . 262
7.7 Conditional Densities . . . 263
7.7.1 Penalized Likelihood Estimation . . . 263
7.7.2 Empirical Performance of Cross-validation . . . 265
7.7.3 Kullback-Leibler Projection . . . 266
7.7.4 R Package gss: sscden Suite . . . 266
7.7.5 Case Study: Penny Thickness . . . 268
7.8 Regression with Cross-Classified Responses . . . 269
7.8.1 Logistic Regression . . . 269
7.8.2 Log-Linear Regression Models . . . 271

7.8.3 Bayesian Confidence Intervals for y-Contrasts 271
7.8.4 Mixed-Effect Models for Correlated Data 272
7.8.5 Empirical Performance of Cross-Validation 273
7.8.6 R Package `gss`: `ssllrm` Suite 274
7.8.7 Case Study: Eyetracking Experiments 275
7.9 Response-Based Sampling 278
7.9.1 Response-Based Samples 278
7.9.2 Penalized Likelihood Estimation 279
7.10 Bibliographic Notes 280
7.11 Problems 282

8 Hazard Rate Estimation **285**
8.1 Preliminaries 286
8.2 Smoothing Parameter Selection 288
8.2.1 Kullback-Leibler Loss and Cross-Validation 289
8.2.2 Empirical Performance 291
8.3 Inference and Software 292
8.3.1 Bayesian Confidence Intervals 292
8.3.2 Kullback-Leibler Projection 293
8.3.3 Frailty Models for Correlated Data 293
8.3.4 R Package `gss`: `sshzd` Suite 293
8.4 Case Studies 295
8.4.1 Treatments of Gastric Cancer 295
8.4.2 Survival After Heart Transplant 297
8.5 Penalized Partial Likelihood 299
8.5.1 Partial Likelihood and Biased Sampling 299
8.5.2 Inference 300
8.5.3 R Package `gss`: `sscox` Suite 300
8.5.4 Case Study: Survival After Heart Transplant 302
8.6 Models Parametric in Time 303
8.6.1 Location-Scale Families and Accelerated Life Models 303
8.6.2 Kullback-Leibler and Cross-Validation 305
8.6.3 Weibull Family 305
8.6.4 Log Normal Family 309
8.6.5 Log Logistic Family 311
8.6.6 Case Study: Survival After Heart Transplant 314
8.7 Bibliographic Notes 316
8.8 Problems 317

9 Asymptotic Convergence **319**
9.1 Preliminaries 319
9.2 Rates for Density Estimates 322
9.2.1 Linear Approximation 323
9.2.2 Approximation Error and Main Results 325

9.2.3 Efficient Approximation . . . 327
9.2.4 Convergence Under Incorrect Model . . . 330
9.2.5 Estimation Under Biased Sampling . . . 331
9.2.6 Estimation of Conditional Density . . . 332
9.2.7 Estimation Under Response-Based Sampling . . . 332
9.3 Rates for Hazard Estimates . . . 333
9.3.1 Martingale Structure . . . 333
9.3.2 Linear Approximation . . . 334
9.3.3 Approximation Error and Main Results . . . 335
9.3.4 Efficient Approximation . . . 338
9.3.5 Convergence Under Incorrect Model . . . 341
9.4 Rates for Regression Estimates . . . 341
9.4.1 General Formulation . . . 341
9.4.2 Linear Approximation . . . 342
9.4.3 Approximation Error and Main Result . . . 343
9.4.4 Efficient Approximation . . . 345
9.4.5 Convergence Under Incorrect Model . . . 347
9.5 Bibliographic Notes . . . 348
9.6 Problems . . . 349

10 Penalized Pseudo Likelihood **351**
10.1 Density Estimation on Product Domains . . . 352
10.1.1 Pseudo and Genuine Likelihoods . . . 352
10.1.2 Preliminaries . . . 353
10.1.3 Smoothing Parameter Selection . . . 354
10.1.4 Square Error Projection . . . 357
10.1.5 R Package `gss`: `ssden1` Suite . . . 358
10.1.6 Case Study: Transcription Factor Association . . . 359
10.2 Density Estimation: Asymptotic Convergence . . . 360
10.2.1 Linear Approximation . . . 361
10.2.2 Approximation Error and Main Results . . . 361
10.2.3 Efficient Approximation . . . 363
10.3 Conditional Density Estimation . . . 364
10.3.1 Preliminaries . . . 365
10.3.2 Smoothing Parameter Selection . . . 366
10.3.3 Square Error Projection . . . 369
10.3.4 R Package `gss`: `sscden1` Suite . . . 369
10.3.5 Case Study: Penny Thickness . . . 371
10.3.6 Asymptotic Convergence . . . 372
10.4 Hazard Estimation . . . 372
10.4.1 Preliminaries . . . 373
10.4.2 Smoothing Parameter Selection . . . 374
10.4.3 Inference . . . 375
10.4.4 R Package `gss`: `sshzd1` Suite . . . 376
10.4.5 Case Study: Survival After Heart Transplant . . . 378

10.5 Hazard Estimation: Asymptotic Convergence 378
10.5.1 Linear Approximation 379
10.5.2 Approximation Error and Main Results 380
10.5.3 Efficient Approximation 381
10.6 Bibliographic Notes 383
10.7 Problems 384

A R Package gss **387**
A.1 Model Construction 387
A.1.1 Marginal Configurations 388
A.1.2 Construction of Interaction Terms 389
A.1.3 Custom Types 390
A.2 Modeling and Data Analytical Tools 391
A.3 Numerical Engines 393

B Conceptual Critiques **395**
B.1 Model Indexing 395
B.2 Optimal and Cross-Validation Indices 397
B.3 Loss, Risk, and Smoothing Parameter Selection 398
B.4 Degrees of Freedom 400

References **403**

Author Index **417**

Subject Index **421**

1
Introduction

Data and models are two sources of information in a statistical analysis. Data carry noise but are "unbiased," whereas models, effectively a set of constraints, help to reduce noise but are responsible for "biases." Representing the two extremes on the spectrum of "bias-variance" trade-off are standard parametric models and constraint-free nonparametric "models" such as the empirical distribution for a probability density. In between the two extremes, there exist scores of nonparametric or semiparametric models, of which most are also known as smoothing methods. A family of such nonparametric models in a variety of stochastic settings can be derived through the penalized likelihood method, forming the subject of this book.

The general penalized likelihood method can be readily abstracted from the cubic smoothing spline as the solution to a minimization problem, and its applications in regression, density estimation, and hazard estimation set out the subject of study (§1.1). Some general notation is set in §1.2. Multivariate statistical models can often be characterized through function decompositions similar to the classical analysis of variance (ANOVA) decomposition, which we discuss in §1.3. To illustrate the potential applications of the methodology, previews of selected case studies are presented in §1.4. Brief summaries of the chapters to follow are given in §1.5.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_1,

1.1 Estimation Problem and Method

The problem to be addressed in this book is flexible function estimation based on stochastic data. To allow for flexibility in the estimation of η, say, soft constraints of the form $J(\eta) \leq \rho$ are used in lieu of the rigid constraints of parametric models, where $J(\eta)$ quantifies the roughness of η and ρ sets the allowance; an example of $J(\eta)$ for η on $[0,1]$ is $\int_0^1 (d^2\eta/dx^2)^2 dx$. Solving the constrained maximum likelihood problem by the Lagrange method, one is led to the penalized likelihood method.

In what follows, a brief discussion of the cubic smoothing spline helps to motivate the idea, and a simple simulation illustrates the role of ρ through the Lagrange multiplier, better known as the smoothing parameter in the context. Following a straightforward abstraction, the penalized likelihood method is exemplified in regression, density estimation, and hazard estimation.

1.1.1 Cubic Smoothing Spline

Consider a regression problem $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $x_i \in [0,1]$ and $\epsilon_i \sim N(0, \sigma^2)$. In a classical parametric regression analysis, η is assumed to be of form $\eta(x, \boldsymbol{\beta})$, known up to the parameters $\boldsymbol{\beta}$, which are to be estimated from the data. When $\eta(x, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, one has a standard linear model. A parametric model characterizes a set of rigid constraints on η. The dimension of the model space (i.e., the number of unknown parameters) is typically much smaller than the sample size n.

To avoid possible model misspecification in a parametric analysis, otherwise known as bias, an alternative approach to estimation is to allow η to vary in a high-dimensional (possibly infinite) function space, leading to various nonparametric or semiparametric estimation methods. A popular approach to the nonparametric estimation of η is via the minimization of a penalized least squares score,

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda \int_0^1 \ddot{\eta}^2 dx, \tag{1.1}$$

with $\ddot{\eta} = d^2\eta/dx^2$, where the first term discourages the lack of fit of η to the data, the second term penalizes the roughness of η, and the smoothing parameter λ controls the trade-off between the two conflicting goals. The minimization of (1.1) is implicitly over functions with square integrable second derivatives. The minimizer η_λ of (1.1) is called a cubic smoothing spline. As $\lambda \to 0$, η_λ approaches the minimum curvature interpolant. As $\lambda \to \infty$, η_λ approaches the simple linear regression line. Note that the linear polynomials $\{f : f = \beta_0 + \beta_1 x\}$ form the so-called null space of the roughness penalty $\int_0^1 \ddot{f}^2 dx$, $\{f : \int_0^1 \ddot{f}^2 dx = 0\}$.

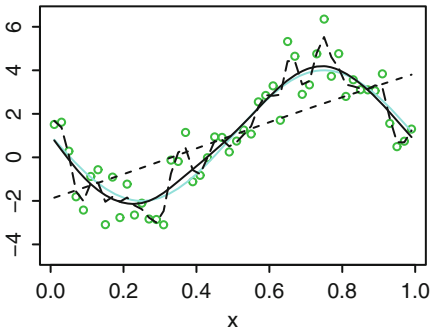


FIGURE 1.1. Cubic smoothing splines. The test function is in the *faded line* and the estimates are in the *solid*, *dashed*, and *long-dashed lines*. The data are superimposed as *circles*.

To illustrate, consider a simple simulation with $x_i = (i - 0.5)/50$, $i = 1, \ldots, 50$, $\eta(x) = 1 + 3\sin(2\pi x - \pi)$, and $\sigma^2 = 1$. The estimate η_λ was calculated at $\log_{10} n\lambda = 0, -3, -6$. Plotted in Fig. 1.1 are the test function (faded line), the estimates (solid, dashed, and long-dashed lines), and the data (circles). The rough fit corresponds to $\log_{10} n\lambda = -6$, the near straight line to $\log_{10} n\lambda = 0$, and the close fit to $\log_{10} n\lambda = -3$.

An alternative derivation of the cubic smoothing spline is through a constrained least squares problem, which solves

$$\min \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2, \quad \text{subject to} \quad \int_0^1 \ddot{\eta}^2 dx \le \rho, \tag{1.2}$$

for some $\rho \ge 0$. The solution to (1.2) usually falls on the boundary of the permissible region, $\int_0^1 \ddot{\eta}^2 dx = \rho$, and by the Lagrange method, it can be calculated as the minimizer of (1.1) with an appropriate Lagrange multiplier λ. Thus, up to the choices of λ and ρ, a penalized least squares problem with a penalty proportional to $\int_0^1 \ddot{\eta}^2 dx$ is equivalent to a constrained least squares problem subject to a soft constraint of the form $\int_0^1 \ddot{\eta}^2 dx \le \rho$; see, e.g., Schoenberg (1964). See also §2.6.2.

Defined as the solution to a penalized optimization problem, a smoothing spline is also known as a natural spline in the numerical analysis literature. The minimizer η_λ of (1.1) is called a cubic spline because it is a piecewise cubic polynomial. It is three times differentiable, with the third derivative jumping at the knots $\xi_1 < \xi_2 < \cdots < \xi_q$, the ordered distinctive sampling points x_i, and it is linear beyond the first knot ξ_1 and the last knot ξ_q. See Schumaker (1981, Chap. 8) for a comprehensive treatment of smoothing splines from a numerical analytical perspective. See also de Boor (1978).

1.1.2 Penalized Likelihood Method

The cubic smoothing spline of (1.1) is a specialization of the general penalized likelihood method in univariate Gaussian regression. To estimate a function of interest η on a generic domain $\mathcal{X}$ using stochastic data, one may use the minimizer of

$$L(\eta|\text{data}) + \frac{\lambda}{2} J(\eta), \tag{1.3}$$

where $L(\eta|\text{data})$ is usually taken as the minus log likelihood of the data and $J(f)$ is a quadratic roughness functional with a null space $\mathcal{N}_J = \{f : J(f) = 0\}$ of low dimension; see §2.1.1 for the definition of quadratic functional. The solution of (1.3) is the maximum likelihood estimate in a model space $\mathcal{M}_\rho = \{f : J(f) \le \rho\}$ for some $\rho \ge 0$, and the smoothing parameter λ in (1.3) is the Lagrange multiplier. See §2.6.2 for a detailed discussion of the role of λ as a Lagrange multiplier.

A few examples of penalized likelihood estimation follow.

Example 1.1 (Response data regression) Assume

$$Y|x \sim \exp\big\{\big(y\eta(x) - b(\eta(x))\big)/a(\phi) + c(y,\phi)\big\},$$

an exponential family density with a modeling parameter η and a possibly unknown nuisance parameter ϕ. Observing independent data (x_i, Y_i), $i = 1, \ldots, n$, the method estimates η via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta(x_i) - b(\eta(x_i))\big\} + \frac{\lambda}{2}J(\eta). \tag{1.4}$$

When the density is Gaussian, (1.4) reduces to a penalized least squares problem; see Problem 1.1. Penalized least squares regression for Gaussian-type responses is the subject of Chap. 3. Penalized likelihood regression for non-Gaussian responses will be studied in Chap. 5. □

Example 1.2 (Density estimation) Observing independent and identically distributed samples X_i, $i = 1, \ldots, n$ from a probability density $f(x)$ supported on a bounded domain $\mathcal{X}$, the method estimates f by $e^\eta / \int_{\mathcal{X}} e^\eta dx$, where η minimizes

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(X_i) - \log\int_{\mathcal{X}} e^{\eta(x)}dx\right\} + \frac{\lambda}{2}J(\eta). \tag{1.5}$$

A side condition, say $\int_{\mathcal{X}} \eta\, dx = 0$, shall be imposed on η for a one-to-one transform $f \leftrightarrow e^\eta / \int_{\mathcal{X}} e^\eta dx$. Penalized likelihood density estimation is the subject of Chap. 7. □

Example 1.3 (Hazard estimation) Let T be the lifetime of an item with survival function $S(t|u) = P(T > t|u)$, possibly dependent on a covariate U. The hazard function is defined as $e^{\eta(t,u)} = -\partial \log S(t|u)/\partial t$. Let Z be the left-truncation time and C be the right-censoring time, independent of T and of each other. Observing $(U_i, Z_i, X_i, \delta_i)$, $i = 1, \ldots, n$, where $X = \min(T, C)$, $\delta = I_{[T \le C]}$, and $Z < X$, the method estimates the log hazard η via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)}dt\right\} + \frac{\lambda}{2}J(\eta); \tag{1.6}$$

see Problem 1.2 for the derivation of the likelihood. Penalized likelihood hazard estimation will be studied in Chap. 8. □

The two basic components of a statistical model, the deterministic part and the stochastic part, are well separated in (1.3). The structure of the deterministic part is determined by the construction of $J(\eta)$ for η on a domain $\mathcal{X}$, of which a comprehensive treatment is presented in Chap. 2. The stochastic part is reflected in the likelihood $L(\eta|\text{data})$ and determines, among other things, the natural measures with which the performance of the estimate is to be assessed. The minimizer of (1.3) with a varying λ defines a family of estimates, and from the cubic spline simulation shown in Fig. 1.1, we have seen how differently the family members may behave. Data-driven procedures for the proper selection of the smoothing parameter are crucial to the practicability of penalized likelihood estimation, to which extensive discussion will be devoted in the settings of regression, density estimation, and hazard estimation in their respective chapters.

1.2 Notation

Listed below is some general notation used in this book. Context-specific or subject-specific notation may differ from that listed here, in which case every effort will be made to avoid possible confusion.

Domains are usually denoted by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, etc., or subscripted as $\mathcal{X}_1$, $\mathcal{X}_2$, etc. Points on domains are usually denoted by $x \in \mathcal{X}$, $y \in \mathcal{Y}$, or $x_1, x_2, y \in \mathcal{X}$. Points on product domains are denoted by $x_1, x_2, y \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, with $x_{1\langle 1\rangle}, x_{2\langle 1\rangle}, y_{\langle 1\rangle} \in \mathcal{X}_1$ and $x_{1\langle 2\rangle}, x_{2\langle 2\rangle}, y_{\langle 2\rangle} \in \mathcal{X}_2$, or by $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Ordinary subscripts are used to denote multiple points on a domain, but *not* coordinates of a point on a product domain.

Function spaces are usually denoted by $\mathcal{H}$, $\mathcal{G}$, etc. Functions in function spaces are usually denoted by $f, g, h \in \mathcal{H}$, $\eta, \phi, \xi \in \mathcal{H}$, etc. Derivatives of a univariate function $f(x)$ are denoted by $\dot{f} = df/dx$, $\ddot{f} = d^2 f/dx^2$,

or by the general notation $f^{(m)} = d^m f/dx^m$. Derivatives of multivariate functions $f(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ on $\mathcal{X}_1 \times \mathcal{X}_2$ or $g(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ are denoted by $f^{(3)}_{\langle 112\rangle} = \partial^3 f/\partial x^2_{\langle 1\rangle} \partial x_{\langle 2\rangle}$, $\ddot{g}_{\langle xy\rangle} = \partial^2 g/\partial x \partial y$, etc.

Matrices are denoted by the standard notation of uppercase letters. Vectors, however, are often *not* denoted by boldface letters in this book. For a point on a product domain $\mathcal{X} = \prod_{\gamma=1}^{\Gamma} \mathcal{X}_\gamma$, we write $x = (x_{\langle 1\rangle}, \ldots, x_{\langle \Gamma\rangle})$. For a function on domain $\mathcal{X} = \{1, \ldots, K\}$, we write $f = \big(f(1), \ldots, f(K)\big)^T$, which may be used as a vector in standard matrix arithmetic. Boldface vectors are used where confusion may result otherwise. For example, $\mathbf{1} = (1, \ldots, 1)^T$ is used to denote a vector of all one's, and $\mathbf{c} = (c_1, \ldots, c_n)^T$ is used to encapsulate subscripted coefficients. In formulas concerning matrix computation, vectors are always set in boldface.

The standard O_p, o_p notation is used in the asymptotic analyses of §§3.2, 4.2.3, 5.2, 6.2, 6.3, Chap. 9, §§10.2, and 10.5. If $P\big(|X| > KY\big) \to 0$ for some constant $K < \infty$, we write $X = O_p(Y)$, and when $P\big(|X| > \epsilon Y\big) \to 0$, $\forall \epsilon > 0$, we denote $X = o_p(Y)$.

1.3 Decomposition of Multivariate Functions

An important aspect of statistical modeling, which distinguishes it from mere function approximation, is the interpretability of the results. Of great utility are decomposition of multivariate functions similar to the classical analysis of variance (ANOVA) decomposition and the associated notions of main effect and interaction. Higher-order interactions are often excluded in practical estimation to control model complexity; the exclusion of all interactions yields the popular additive models. Selective exclusion of certain interactions also characterizes many interesting statistical models in a variety of stochastic settings.

Casting the classical one-way ANOVA decomposition as the decomposition of functions on a discrete domain, a simple averaging operator is introduced to facilitate the generalization of the notion to arbitrary domains. Multiway ANOVA decomposition is then defined, with the identifiability of the terms assured by side conditions specified through the averaging operators. Examples are given and a proposition is proved concerning certain intrinsic structures that are independent of the side conditions. The utility and implication of selective term trimming in an ANOVA decomposition are then briefly discussed in the context of regression, density estimation, and hazard estimation.

1.3.1 ANOVA Decomposition and Averaging Operator

Consider a standard one-way ANOVA model, $Y_{ij} = \mu_i + \epsilon_{ij}$, where μ_i are the treatment means at treatment levels $i = 1, \ldots, K$ and ϵ_{ij} are

independent normal errors. Writing $\mu_i = \mu + \alpha_i$, one has the "overall mean" μ and the treatment effect α_i. The identifiability of μ and α_i are assured through a side condition, of which common choices include $\alpha_1 = 0$ with level 1 treated as the control and $\sum_{i=1}^{K} \alpha_i = 0$ with all levels treated symmetrically.

The one-way ANOVA model can be recast as $Y_j = f(x_j) + \epsilon_j$, where $f(x)$ is defined on the discrete domain $\mathcal{X} = \{1, \ldots, K\}$; the treatment levels are now coded by x and the subscript j labels the observations. The ANOVA decomposition $\mu_i = \mu + \alpha_i$ in the standard ANOVA model notation can be written as

$$f(x) = Af + (I - A)f = f_\emptyset + f_x,$$

where A is an averaging operator that "averages out" the argument x to return a constant function and I is the identity operator. For example, with $Af = f(1)$, one has $f(x) = f(1) + \{f(x) - f(1)\}$, corresponding to $\alpha_1 = 0$. With $Af = \sum_{x=1}^{K} f(x)/K = \bar{f}$, one has $f(x) = \bar{f} + \big(f(x) - \bar{f}\big)$, corresponding to $\sum_{i=1}^{K} \alpha_i = 0$. Note that applying A to a constant function returns that constant, hence the name "averaging." It follows that $A(Af) = Af$, $\forall f$, or, simply, $A^2 = A$. The constant term $f_\emptyset = Af$ is the "overall mean" and the term $f_x = (I - A)f$ is the treatment effect, or "contrast," that satisfies the side condition $Af_x = 0$.

On a continuous domain, say $\mathcal{X} = [a, b]$, one may similarly define an ANOVA decomposition $f(x) = Af + (I - A)f = f_\emptyset + f_x$ through an appropriately defined averaging operator A, where f_x satisfies the side condition $Af_x = 0$. For example, with $Af = f(a)$, one has $f(x) = f(a) + \{f(x) - f(a)\}$. Similarly, with $Af = \int_a^b f dx/(b - a)$, one has $f(x) = \int_a^b f dx/(b - a) + \{f(x) - \int_a^b f dx/(b - a)\}$.

1.3.2 Multiway ANOVA Decomposition

Now consider a function $f(x) = f(x_{\langle 1 \rangle}, \ldots, x_{\langle \Gamma \rangle})$ on a product domain $\mathcal{X} = \prod_{\gamma=1}^{\Gamma} \mathcal{X}_\gamma$, where $x_{\langle \gamma \rangle} \in \mathcal{X}_\gamma$ denotes the γth coordinate of $x \in \mathcal{X}$. Let A_γ be an averaging operator on $\mathcal{X}_\gamma$ that averages out $x_{\langle \gamma \rangle}$ from the active argument list and satisfies $A_\gamma^2 = A_\gamma$; $A_\gamma f$ is constant on the $\mathcal{X}_\gamma$ axis but not necessarily an overall constant function. An ANOVA decomposition of f can be defined as

$$f = \left\{ \prod_{\gamma=1}^{\Gamma} (I - A_\gamma + A_\gamma) \right\} f = \sum_{\mathcal{S}} \left\{ \prod_{\gamma \in \mathcal{S}} (I - A_\gamma) \prod_{\gamma \notin \mathcal{S}} A_\gamma \right\} f = \sum_{\mathcal{S}} f_{\mathcal{S}}, \quad (1.7)$$

where $\mathcal{S} \subseteq \{1, \ldots, \Gamma\}$ enlists the active arguments in $f_{\mathcal{S}}$ and the summation is over all of the 2^Γ subsets of $\{1, \ldots, \Gamma\}$. The term $f_\emptyset = \prod A_\gamma f$ is a constant, the term $f_\gamma = f_{\{\gamma\}} = (I - A_\gamma) \prod_{\alpha \neq \gamma} A_\alpha f$ is the $x_{\langle \gamma \rangle}$ main effect,

the term $f_{\gamma,\delta} = f_{\{\gamma,\delta\}} = (I - A_\gamma)(I - A_\delta)\prod_{\alpha\neq\gamma,\delta} A_\alpha f$ is the $x_{\langle\gamma\rangle}$-$x_{\langle\delta\rangle}$ interaction, and so forth. The terms of such a decomposition satisfy the side conditions $A_\gamma f_{\mathcal{S}} = 0, \forall \mathcal{S} \ni \gamma$. The choices of A_γ, or the side conditions on each axes, are open to specification.

The domains $\mathcal{X}_\gamma$ are generic in the above discussion; in particular, they can be product domains themselves. As a matter of fact, the ANOVA decomposition of (1.7) can also be defined recursively through a series of nested constructions with $\Gamma = 2$; see, e.g., Problem 1.3.

The ANOVA decomposition can be built into penalized likelihood estimation through the proper construction of the roughness functional $J(f)$; details are to be found in §2.4.

Example 1.4 When $\Gamma = 2$, $\mathcal{X}_1 = \{1, \ldots, K_1\}$, and $\mathcal{X}_2 = \{1, \ldots, K_2\}$, the decomposition reduces to a standard two-way ANOVA decomposition. With averaging operators $A_1 f = f(1, x_{\langle 2\rangle})$ and $A_2 f = f(x_{\langle 1\rangle}, 1)$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f(1,1), \\ f_1 &= (I - A_1)A_2 f = f(x_{\langle 1\rangle}, 1) - f(1,1), \\ f_2 &= A_1(I - A_2)f = f(1, x_{\langle 2\rangle}) - f(1,1), \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f(x_{\langle 1\rangle}, 1) - f(1, x_{\langle 2\rangle}) + f(1,1). \end{aligned}$$

With $A_\gamma f = \sum_{x_{\langle\gamma\rangle}=1}^{K_\gamma} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_\gamma$, $\gamma = 1, 2$, one similarly has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f_{\cdot\cdot}, \\ f_1 &= (I - A_1)A_2 f = f_{x_{\langle 1\rangle}\cdot} - f_{\cdot\cdot}, \\ f_2 &= A_1(I - A_2)f = f_{\cdot x_{\langle 2\rangle}} - f_{\cdot\cdot}, \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f_{x_{\langle 1\rangle}\cdot} - f_{\cdot x_{\langle 2\rangle}} + f_{\cdot\cdot}, \end{aligned}$$

where $f_{\cdot\cdot} = \sum_{x_{\langle 1\rangle}, x_{\langle 2\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_1 K_2$, $f_{x_{\langle 1\rangle}\cdot} = \sum_{x_{\langle 2\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_2$, and $f_{\cdot x_{\langle 2\rangle}} = \sum_{x_{\langle 1\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_1$. One may also use different averaging operators on different axes; see Problem 1.4. □

Example 1.5 Consider $\Gamma = 2$ and $\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$. With $A_1 f = f(0, x_{\langle 2\rangle})$ and $A_2 f = f(x_{\langle 1\rangle}, 0)$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f(0,0), \\ f_1 &= (I - A_1)A_2 f = f(x_{\langle 1\rangle}, 0) - f(0,0), \\ f_2 &= A_1(I - A_2)f = f(0, x_{\langle 2\rangle}) - f(0,0), \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f(x_{\langle 1\rangle}, 0) - f(0, x_{\langle 2\rangle}) + f(0,0). \end{aligned}$$

With $A_\gamma f = \int_0^1 f dx_{\langle\gamma\rangle}$, $\gamma = 1, 2$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = \textstyle\int_0^1 \int_0^1 f dx_{\langle 1\rangle} dx_{\langle 2\rangle}, \\ f_1 &= (I - A_1) A_2 f = \textstyle\int_0^1 (f - \int_0^1 f dx_{\langle 1\rangle}) dx_{\langle 2\rangle}, \\ f_2 &= A_1 (I - A_2) f = \textstyle\int_0^1 (f - \int_0^1 f dx_{\langle 2\rangle}) dx_{\langle 1\rangle}, \\ f_{1,2} &= (I - A_1)(I - A_2) f \\ &= f - \textstyle\int_0^1 f dx_{\langle 2\rangle} - \int_0^1 f dx_{\langle 1\rangle} + \int_0^1 \int_0^1 f dx_{\langle 1\rangle} dx_{\langle 2\rangle}. \end{aligned}$$

Similar results with different averaging operators on different axes are also straightforward; see Problem 1.5. □

In standard ANOVA models, higher-order terms are frequently eliminated, whereas main effects and lower-order interactions are estimated from the data. One learns not to drop the $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$ main effects if the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$ interaction is considered, however, and not to drop the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$ interaction when the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$-$x_{\langle 3\rangle}$ interaction is included. Although the ANOVA decomposition as defined in (1.7) obviously depends on the averaging operators A_γ, certain structures are independent of the particular choices of A_γ. Specifically, for any index set $\mathcal{I}$, if $f_\mathcal{S} = 0$, $\forall \mathcal{S} \supseteq \mathcal{I}$ with a particular set of A_γ, then the structure also holds for any other choices of A_γ, as the following proposition asserts.

Proposition 1.1 *For any two sets of averaging operators A_γ and $\tilde{A}_\gamma$ satisfying $A_\gamma^2 = A_\gamma$ and $\tilde{A}_\gamma^2 = \tilde{A}_\gamma$, $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ if and only if $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma) f = 0$, where $\mathcal{I}$ is any index set.*

Note that the condition $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ means that $f_\mathcal{S} = 0$, $\forall \mathcal{S} \supseteq \mathcal{I}$. For example, $(I - A_1) f = 0$ implies that all terms involving $x_{\langle 1\rangle}$ vanish, and $(I - A_1)(I - A_2) f = 0$ means that all terms involving both $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$ disappear. Model structures that can be characterized through constraints of the form $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ permit a term $f_\mathcal{S}$ only when all of its "subset terms," $f_{\mathcal{S}'}$ for $\mathcal{S}' \subset \mathcal{S}$, are permitted. A simple corollary of the proposition is the obvious fact that an additive model remains an additive model regardless of the side conditions.

Proof of Proposition 1.1: It is easy to see that $(I - \tilde{A}_\gamma) A_\gamma = 0$. Suppose $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ and define the ANOVA decomposition in (1.7) using A_γ. Now, for any nonzero term $f_\mathcal{S}$ in (1.7), one has $\mathcal{S} \not\supseteq \mathcal{I}$, so there exists $\gamma \in \mathcal{I}$ but $\gamma \notin \mathcal{S}$, hence $f_\mathcal{S} = [\cdots A_\gamma \cdots] f$. The corresponding $(I - \tilde{A}_\gamma)$ in $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma)$ then annihilates the term. It follows that all nonzero ANOVA terms in (1.7) are annihilated by $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma)$, so $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma) f = 0$. The converse is true by symmetry. □

1.3.3 Multivariate Statistical Models

Many multivariate statistical models can be characterized by selective term elimination in an ANOVA decomposition. Some of such models are discussed below.

Curse of Dimensionality and Additive Models

Recall the classical ANOVA models with $\mathcal{X}_\gamma$ discrete. In practical data analysis, one usually includes only the main effects, with the possible addition of a few lower-order interactions. Higher-order interactions are less interpretable yet more difficult to estimate, as they usually consume many more degrees of freedom than the lower-order terms. Models with only main effects included are called additive models.

The difficulty associated with function estimation in high-dimensional spaces may be perceived through the sparsity of the space. Take $\mathcal{X}_\gamma = [0, 1]$, for example, a k-dimensional cube with each side of length 0.5 has volume 0.5^k. Assume a uniform distribution of the data and consider a piecewise constant function with jumps only possible at $x_{\langle\gamma\rangle} = 0.5$. To estimate such a function in 1 dimension with two pieces, one has information from 50 % of the data per piece, in 2 dimensions with four pieces, 25 % per piece, in 3 dimensions with eight pieces, 12.5 % per piece, etc. The lack of data due to the sparsity of high-dimensional spaces is often referred to as the curse of dimensionality. Alternatively, the curse of dimensionality may also be characterized by the explosive increase in the number of parameters, or the degrees of freedom, that one would need to approximate a function well in a high-dimensional space. To achieve the flexibility of a five-piece piecewise polynomial in 1 dimension, for example, one would end up with 125 pieces in 3 dimensions by taking products of the pieces in 1 dimension.

To combat the curse of dimensionality in multivariate function estimation, one needs to eliminate higher-order interactions to control model complexity. As with classical ANOVA models, additive models with the possible addition of second-order interactions are among the most popular models used in practice.

Conditional Independence and Graphical Models

To simplify notation, the marginal domains will be denoted by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, etc., in the rest of the section instead of the subscripted $\mathcal{X}$ used in (1.7).

Consider a probability density $f(x)$ of a random variable X on a domain $\mathcal{X}$. Writing

$$f(x) = \frac{e^{\eta(x)}}{\int_{\mathcal{X}} e^{\eta(x)} dx}, \tag{1.8}$$

known as the logistic density transform, the log density $\eta(x)$ is free of the positivity and unity constraints, $f(x) > 0$ and $\int_{\mathcal{X}} f(x)dx = 1$, that $f(x)$

must satisfy. The transform is not one-to-one, though, as $e^{\eta(x)}/\int_{\mathcal{X}} e^{\eta(x)}dx = e^{C+\eta(x)}/\int_{\mathcal{X}} e^{C+\eta(x)}dx$ for any constant C. The transform can be made one-to-one, however, by imposing a side condition $A_x\eta = 0$ for some averaging operator A_x on $\mathcal{X}$; this can be achieved by eliminating the constant term in a one-way ANOVA decomposition $\eta = A_x\eta + (I - A_x)\eta = \eta_\emptyset + \eta_x$.

For a joint density $f(x, y)$ of random variables (X, Y) on a product domain $\mathcal{X} \times \mathcal{Y}$, one may write

$$f(x,y) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta(x,y)}dy} = \frac{e^{\eta_x+\eta_y+\eta_{x,y}}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta_x+\eta_y+\eta_{x,y}}dy},$$

where η_x, η_y, and $\eta_{x,y}$ are the main effects and interaction of $\eta(x, y)$ in an ANOVA decomposition; the constant is eliminated in the rightmost expression for a one-to-one transform. The conditional distribution of Y given X has a density

$$f(y|x) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{Y}} e^{\eta(x,y)}dy} = \frac{e^{\eta_y+\eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_y+\eta_{x,y}}dy}, \tag{1.9}$$

where the logistic conditional density transform is one-to-one only for the rightmost expression with the side conditions $A_y(\eta_y + \eta_{x,y}) = 0$, $\forall x \in \mathcal{X}$, where A_y is the averaging operator on $\mathcal{Y}$ that help to define the ANOVA decomposition. The independence of X and Y, denoted by $X \perp Y$, is characterized by $\eta_{x,y} = 0$, or $(I - A_x)(I - A_y)\eta = 0$.

The domains $\mathcal{X}$ and $\mathcal{Y}$ are generic in (1.9); in particular, they can be product domains themselves. Substituting (y, z) for y in (1.9), one has

$$f(y,z|x) = \frac{e^{\eta_y+\eta_z+\eta_{y,z}+\eta_{x,y}+\eta_{x,z}+\eta_{x,y,z}}}{\int_{\mathcal{Y}} dy \int_{\mathcal{Z}} e^{\eta_y+\eta_z+\eta_{y,z}+\eta_{x,y}+\eta_{x,z}+\eta_{x,y,z}}dz},$$

where $\eta_{(y,z)}$ is expanded out as $\eta_y + \eta_z + \eta_{y,z}$ and $\eta_{x,(y,z)}$ is expanded out as $\eta_{x,y} + \eta_{x,z} + \eta_{x,y,z}$; see Problem 1.3. The conditional independence of Y and Z given X, denoted by $(Y \perp Z)\big|X$, is characterized by $\eta_{y,z} + \eta_{x,y,z} = 0$, or $(I - A_y)(I - A_z)\eta = 0$.

Now, consider the joint density of four random variables (U, V, Y, Z), with $(U \perp V)\big|(Y, Z)$ and $(Y \perp Z)\big|(U, V)$. It can be shown that such a structure is characterized by $\eta_{u,v} + \eta_{y,z} + \eta_{u,v,y} + \eta_{u,v,z} + \eta_{u,y,z} + \eta_{v,y,z} + \eta_{u,v,y,z} = 0$ in an ANOVA decomposition, or $(I - A_u)(I - A_v)\eta = (I - A_y)(I - A_z)\eta = 0$; see Problem 1.7.

As noted above, the ANOVA decompositions in the log density η that characterize conditional independence structures are all of the type covered in Proposition 1.1. The elimination of lower-order terms in (1.8) and (1.9) for one-to-one transforms only serve to remove technical redundancies introduced by the "overparameterization" of $f(x)$ or $f(y|x)$ by the corresponding unrestricted η.

Conditional independence structures can be represented as graphs, and models for multivariate densities with specified conditional independence structures built in are called graphical models; see, e.g., Whittaker (1990) for some general discussion and for the parametric estimation of graphical models.

Proportional Hazard Models and Beyond

For $\eta(t, u)$ a log hazard on the product of a time domain $\mathcal{T}$ and a covariate domain $\mathcal{U}$, an additive model $\eta(t, u) = \eta_\emptyset + \eta_t + \eta_u$ characterizes a proportional hazard model, with $e^{\eta_\emptyset + \eta_t}$ being the base hazard and e^{η_u} being the relative risk. When the interaction $\eta_{t,u}$ is included in the model, one has something beyond the proportional hazard model. The covariate domain can be a product domain itself, on which nested ANOVA decompositions can be introduced.

1.4 Case Studies

To illustrate potential applications of the techniques to be developed in this book, we shall now present previews of a few selected case studies. Full accounts of these studies are to be found in later chapters.

1.4.1 Water Acidity in Lakes

From the Eastern Lake Survey of 1984 conducted by the United States Environmental Protection Agency (EPA), Douglas and Delampady (1990) derived a data set containing geographic information, water acidity measurements, and main ion concentrations in 1,798 lakes in three regions, northeast, upper midwest, and southeast, in the eastern United States. Of interest is the dependence of the water acidity on the geographic locations and other information concerning the lakes.

Preliminary analysis and consultation with a water chemist suggest that a model for the surface pH in terms of the geographic location and the calcium concentration is appropriate. A model of the following form is considered:

$$\text{pH} = \eta_\emptyset + \eta_c(\text{calcium}) + \eta_g(\text{geography}) + \eta_{c,g}(\text{calcium}, \text{geography}) + \epsilon.$$

The model can be fitted to the data using tensor product splines with a thin-plate marginal, to be discussed in §4.3, with the geographic location treated in an isotropically invariant manner. The isotropic invariance is in the following sense: After converting the longitude and latitude of the geographic location to the x-y coordinates (in distance) with respect to a local origin, the fitting of the model is invariant to arbitrary shift and

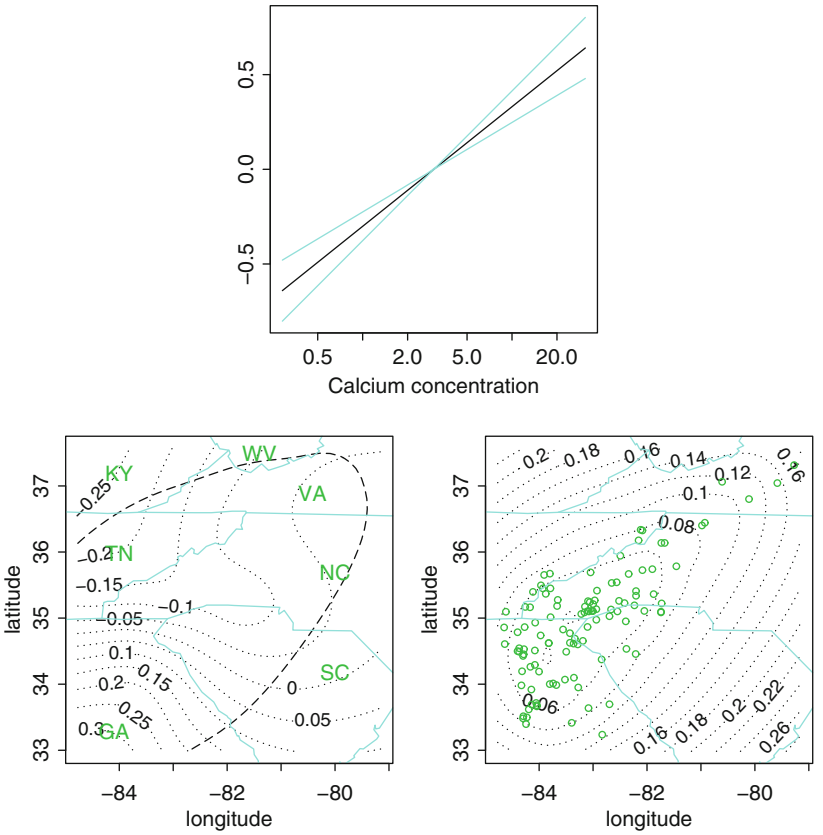


FIGURE 1.2. Water acidity fit for lakes in the Blue Ridge. *Top*: Calcium effect with 95 % Bayesian confidence intervals. *Left*: Geography effect. *Right*: Standard errors of geography effect with the lakes superimposed.

rotation of the x-y coordinates. The geographic location is mathematically two dimensional, but, conceptually, it makes little sense to talk about north-south effect or east-west effect, or any other directional decomposition of the geographic location, in the context. The isotropically invariant treatment preserves the integrity of the geographic location as an inseparable entity.

For illustration, consider the fitting of the model to 112 lakes in the Blue Ridge. As inputs to the fitting algorithm, the longitude and latitude were converted to x-y coordinates in distance, and a log transform was applied to the calcium concentration. The interaction $\eta_{c,g}$ was negligible as assessed by the model selection devices of §§3.7 and 3.8, so an additive model was fitted. Plotted in Fig. 1.2 are the fitted calcium effect with 95 % confidence intervals, the estimated geography effect, and the standard errors of the estimated geography effect; see §3.3 for the definition and interpretation of the standard errors and confidence intervals. The 0.14 contour of the

geography standard errors, which encloses all but one lake, is superimposed as the dashed line in the plot of the geography effect. The lakes are superimposed in the plot of geography standard errors. The fit has an R^2 of 0.53 and the "explained" variation in pH are roughly 70 % by calcium concentration and 30 % by geography.

A full account of the analysis is to be found in §4.3.4.

1.4.2 AIDS Incubation

To study the AIDS incubation time, a valuable source of information is in the records of patients who were infected with the HIV virus through blood transfusion, of which the date can be ascertained retrospectively. A data set collected by the Centers for Disease Control and Prevention (CDC) is listed in Wang (1989), which includes the time X from transfusion to the diagnosis of AIDS, the time Y from transfusion to the end of study (July 1986), both in months, and the age of the individual at the time of transfusion, for 295 individuals. It is clear that $X \le Y$ (i.e., the data are truncated).

Assuming the independence of X and Y in the absence of truncation, and conditioning on the truncation mechanism, the density of (X, Y) is given by

$$f(x,y) = \frac{e^{\eta_x(x)+\eta_y(y)}}{\int_0^a dy \int_0^y e^{\eta_x(x)+\eta_y(y)} dx},$$

where $[0, a]$ is a finite interval covering the data. The penalized likelihood score (1.5) can be specified as

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta_x(X_i)+\eta_y(Y_i)-\log\int_0^a dy\int_0^y e^{\eta_x(x)+\eta_y(y)}dx\right\} + \frac{\lambda_x}{2}\int_0^a \ddot{\eta}_x^2 dx + \frac{\lambda_y}{2}\int_0^a \ddot{\eta}_y^2 dx, \quad (1.10)$$

where η_x and η_y satisfy certain side conditions such as $\int_0^a \eta_x dx = 0$ and $\int_0^a \eta_y dy = 0$.

Grouping the individuals by age, one has 141 "elderly patients" of age 60 or above. Estimating $f(x, y)$ for this age group through the minimization of (1.10), with $a = 100$ and λ_x and λ_y selected through a device introduced in §7.3, one obtains the estimate contoured in Fig. 1.3, where the data are superimposed and the marginal densities $f(x) = e^{\eta_x} / \int_0^{100} e^{\eta_x} dx$ and $f(y) = e^{\eta_y} / \int_0^{100} e^{\eta_y} dy$ are plotted in the empty space on their respective axes.

Further discussions concerning the analysis of this data set will be presented in §§7.5.3 and 7.6.5.

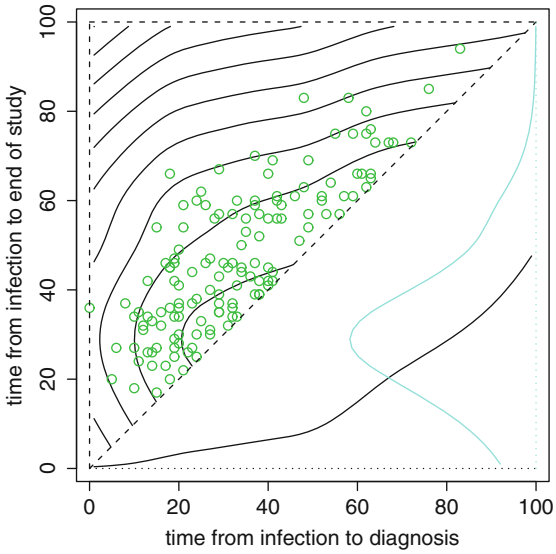


FIGURE 1.3. AIDS incubation and HIV infection of the elderly. Contours are estimated density on the observable region surrounded by *dashed lines*. *Circles* are the observations. *Curves* over the *dotted lines* in the empty space are the estimated marginal densities.

1.4.3 Survival After Heart Transplant

One of the most demonstrated survival data is the Stanford heart transplant data. In this study, we consider the data listed in Miller and Halpern (1982). Recorded were survival or censoring times of 184 patients after (first) heart transplant, in days, their ages at transplant, and a certain tissue-type mismatch score for 157 of the patients. There were 113 recorded deaths and 71 censorings. From the analysis by Miller and Halpern (1982) and others, the tissue-type mismatch score did not have significant impact on survival, so we will try to estimate the hazard as a function of time after transplant and the age of patient at transplant.

In the notation of Example 1.3, $Z = 0$ and U is the age at transplant. With a proportional hazard model $\eta(t, u) = \eta_\emptyset + \eta_t + \eta_u$, the penalized likelihood score (1.6) can be specified as

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\big(\eta_\emptyset + \eta_t(X_i) + \eta_u(U_i)\big) - e^{\eta_\emptyset+\eta_u(U_i)}\int_0^{X_i} e^{\eta_t(t)}dt\right\} + \frac{\lambda_t}{2}\int_0^{T^*}\ddot{\eta}_t^2 dt + \frac{\lambda_u}{2}\int_a^b \ddot{\eta}_u^2 du, \quad (1.11)$$

where $X_i \le T^*$ and $U_i \in [a, b]$.

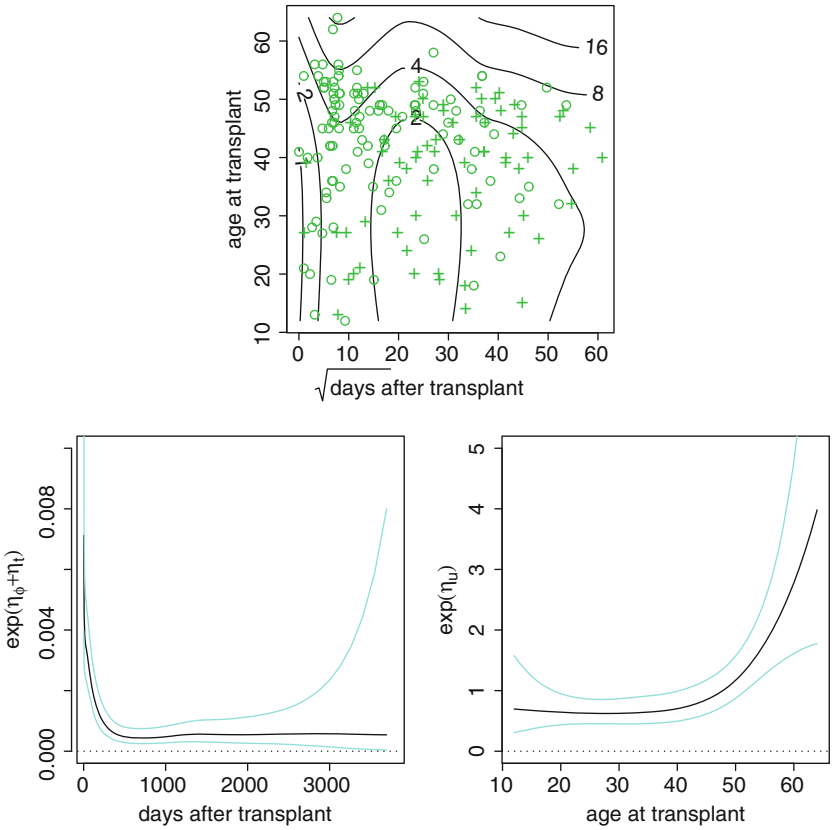


FIGURE 1.4. Hazard after heart transplant. *Top*: Contours of $100e^{\tilde{\eta}(t^*,u)}$, where $t^* = \sqrt{t}$, with deceased (*circles*) and censored (*pluses*) patients superimposed. *Left*: Base hazard $e^{\eta_\emptyset + \eta_t}$ with 95 % Bayesian confidence intervals, on the original time scale. *Right*: Age effect e^{η_u} with 95 % Bayesian confidence intervals.

Before fitting the model to the data, the time axis was rescaled by a square root transform $t^* = \sqrt{t}$ to make X_i more evenly scattered. Once $e^{\tilde{\eta}(t^*,u)} = -d \log S(t^*, u)/dt^*$ is estimated, the hazard on the original time scale is simply

$$e^{\eta(t,u)} = e^{\tilde{\eta}(t^*,u)}(dt^*/dt) = e^{\tilde{\eta}(\sqrt{t},u)}/\big(2\sqrt{t}\big).$$

Fitting the proportional hazard model through the minimization of (1.11), with λ_t and λ_u selected via a device introduced in §8.2, one obtains the fit plotted in Fig. 1.4: In the top frame, $e^{\tilde{\eta}(t^*,u)}$ is contoured with the data superimposed, and in the left and right frames, the base hazard $e^{\eta_\emptyset + \eta_t}$ (on the original time scale) and the age effect e^{η_u} are plotted along with 95 % confidence intervals.

Further details concerning the analysis of this data set can be found in §§8.4.2, 8.5.4, 8.6.6, and 10.4.5.

1.5 Scope

This book presents a systematic treatment of function estimation on generic domains using the penalized likelihood method. Main topics to be covered include model construction, smoothing parameter selection, computation, and asymptotic convergence.

Chapter 2 is devoted to the construction of $J(\eta)$ for use in (1.3) on generic domains; of particular interest is that on product domains with ANOVA decompositions built in. Among examples used to illustrate the construction are shrinkage estimates, polynomial smoothing splines, and their tensor products. Other issues that do not involve the stochastic structure of $L(\eta|\text{data})$ are also discussed in the chapter, which include the empirical Bayes model associated with (1.3) and the existence of the minimizer of (1.3).

Chapter 3 discusses penalized least squares regression with Gaussian-type responses. Effective methods for smoothing parameter selection and generic algorithms for computation are the main focus of the discussion. Data analytical tools are presented, which include interval estimates and diagnostics for model selection. Also discussed are fast algorithms in settings with certain special structures.

Chapter 4 enlists some generalizations and variations of the polynomial smoothing splines. Among subjects under discussion are the partial splines, the periodic splines, the thin-plate splines, the spherical splines, and the L-splines. Conceptually, these are simply further examples of the general construction presented in Chap. 2, but some of the mathematical details are more involved.

Chapter 5 studies penalized likelihood regression with non-Gaussian responses. The central issue is, again, the effective selection of smoothing parameters and the related computation. Computational and data analytical tools developed in Chap. 3 are extended to non-Gaussian regression.

Chapter 6 develops methods to accommodate correlated data. Using random effects to model correlations in the likes of longitudinal and clustered data, mixed-effect models can be fitted with tuning parameters selected by devices developed for independent data. When the covariance matrix differs from diagonal by more than a low-rank matrix update, methods are also derived for tuning parameter selection in Gaussian regression.

Chapter 7 deals with penalized likelihood density estimation under a variety of sampling schemes. Beside the standard method of Example 1.2 for independent and identically distributed samples, variation is also

discussed for data subject to biased sampling and random truncation. Further variations include conditional density estimation, of which regression with cross-classified responses is a special case, and density estimation with data from response-based sampling. Methods for effective smoothing parameter selection are developed and the related computation is outlined.

Chapter 8 handles penalized likelihood hazard estimation. Under discussion are (i) the method of Example 1.3, (ii) the estimation of relative risk in a proportional hazard model via penalized partial likelihood, and (iii) the estimation of models parametric in time. The numerical structure of Example 1.3 parallels that of Example 1.2, and the partial likelihood is isomorphic to the likelihood for density estimation under biased sampling, so the smoothing parameters in (i) and (ii) can be selected using the methods developed in Chap. 7. For (iii), the smoothing parameters are selected by the methods developed in Chap. 5.

Chapter 9 investigates the asymptotic convergence of penalized likelihood estimates. Convergence rates are calculated in terms of problem-specific losses derived from the respective stochastic settings, and the notion of efficient approximation provides the theoretical basis for much of the computational developments in earlier chapters. Also noted are the mode and rates of convergence of the estimates when the models are incorrect.

Chapter 10 explores a variant of penalized likelihood estimation that trades statistical performance for numerical efficiency. The computational benefit comes from the avoidance of costly numerical integrations, making density estimation feasible in high dimensions and reducing substantially the execution time for the estimation of conditional density $f(y|x)$ with continuous y and for the estimation of hazard with continuous covariates.

Throughout Chaps. 3–8 and 10, open-source software is illustrated that implements the computational and data analytical tools developed; the code is collected in an R package `gss` with a friendly user-interface. The overall design of `gss` is outlined in **Appendix A**.

Parametric statistical models such as $J(\eta) = 0$ resides in some low dimensional model spaces regardless of the sample size, whereas nonparametric models such as $J(\eta) \leq \rho$ have expanding model spaces (with $\rho \uparrow \infty$) as the sample size increases. The philosophical difference between the two approaches is often overlooked, however, and attempts to extend familiar notions in parametric inference to nonparametric estimation can easily fall victim to conceptual pitfalls. **Appendix B** presents a few conceptual critiques that scrutinize some widely publicized notions concerning nonparametric statistical models.

1.6 Bibliographic Notes

Section 1.1

A discrete version of (1.1) for data smoothing dated back to Whittaker (1923). Early results on the modern theory of smoothing spline interpolation with exact data (i.e., with $\lambda = 0$ in (1.1) for $Y_i = f(x_i)$) can be found in, e.g., Schoenberg (1964) and de Boor and Lynch (1966), among others; see the Foreword of Wahba (1990) for further historical notes. A comprehensive treatment of smoothing splines from a numerical analytical perspective can be found in Schumaker (1981, Chap. 8). A popular reference on splines, especially on the popular B-splines, is de Boor (1978). B-splines, however, are *not* smoothing splines.

Pioneered by the work of Kimeldorf and Wahba (1970a, 1970b, 1971), the study of (1.1) and generalizations thereof in a statistical context has over the years produced a vast literature on penalized least squares regression. Historical breakthroughs can be found in Craven and Wahba (1979) and Wahba (1983), among others. Wahba (1990) compiled an excellent synthesis for work up to that date. See §3.11 for further notes on penalized least squares regression.

The penalized likelihood method was introduced by Good and Gaskins (1971) in the context of density estimation; the formulation of Example 1.2 by Gu and Qiu (1993) evolved from the work of Leonard (1978) and Silverman (1982). The penalized likelihood regression of Example 1.1 was formulated by O'Sullivan, Yandell, and Raynor (1986); see also Silverman (1978). The penalized likelihood hazard estimation of Example 1.3, which was formulated by Gu (1996), evolved from the work of Anderson and Senthilselvan (1980), O'Sullivan (1988a, 1988b), and Zucker and Karr (1990).

Section 1.3

Classical ANOVA models can be found in statistics textbooks of almost all levels. The definition (1.7) on generic domains can be found in Gu and Wahba (1991a, 1993b). The result of Proposition 1.1 on discrete domains can be found in standard graduate-level textbooks on linear models. See, e.g., Scheffe (1959, §4.1) and Seber (1977, p. 277).

Additive models are routinely used in standard linear model analysis. Their use in nonparametric regression was popularized by the work of Stone (1985) and Hastie and Tibshirani (1986, 1990), among others. Graphical models have their roots in the classical log linear models for categorical data; comprehensive modern treatments with a mixture of continuous and categorical data can be found in, e.g., Whittaker (1990) and Lauritzen (1996). The proportional hazard models, especially the so-called

Cox models proposed by Cox (1972), are among standard tools found in most textbooks on survival analysis; see, e.g., Kalbfleisch and Prentice (1980) and Fleming and Harrington (1991).

Section 1.4

The EPA lake acidity data of §1.4.1 was used in Gu and Wahba (1993a) to illustrate tensor product thin-plate splines and in Gu and Wahba (1993b) to illustrate componentwise Bayesian confidence intervals.

The CDC blood transfusion data was used by Kalbfleisch and Lawless (1989) to motivate and illustrate methods for nonparametric (in the sense of empirical distribution) and parametric inference based on retrospective ascertainment. Wang (1989) analyzed the data using a semiparametric maximum likelihood method designed for truncated data. The analysis illustrated in §1.4.2 is similar to the one presented in Gu (1998c).

The Stanford heart transplant data has become a benchmark example for many researchers to showcase innovations in survival analysis. Early references on the analysis of the data include Turnbull, Brown, and Hu (1974), Miller (1976) and Crowley and Hu (1977). The analysis illustrated in §1.4.3 is similar to the one presented in Gu (1998c).

1.7 Problems

Section 1.1

1.1 Consider univariate regression on $\mathcal{X} = [0,1]$. Take $J(\eta) = \int \ddot{\eta}^2 dx$ in (1.4).

(a) For $Y|x \sim N\big(\mu(x), \sigma^2\big)$, verify that (1.4) with $\eta = \mu$ reduces to (1.1).

(b) For $Y|x \sim \text{Binomial}\big(1, p(x)\big)$, specialize (1.4) with $\eta = \log\big\{p/(1-p)\big\}$ to obtain a score for penalized likelihood logistic regression.

(c) For $Y|x \sim \text{Poisson}\big(\lambda(x)\big)$, specialize (1.4) with $\eta = \log\lambda$ to obtain a score for penalized likelihood Poisson regression.

1.2 Consider the hazard estimation problem in Example 1.3.

(a) Verify that $S(t|u) = \exp\big\{-\int_0^t e^{\eta(s,u)}ds\big\}$.

(b) The likelihood of exact lifetime T is simply its density $f(t)$ evaluated at T. The likelihood of right-censored lifetime $T > C$ is the survival probability $P(T > C) = S(C)$. Verify that the likelihood of (Z, X, δ) is $e^{\delta\eta(X)}S(X)/S(Z)$, where the dependence on the covariate U is suppressed from the notation.

(c) Verify that the first term in (1.6) is indeed the minus log likelihood of $(U_i, Z_i, X_i, \delta_i)$, $i = 1, \ldots, n$.

Section 1.3

1.3 For averaging operators A_γ on $\mathcal{X}_\gamma$, verify that

$$I - A_1 A_2 = (I - A_1)A_2 + A_1(I - A_2) + (I - A_1)(I - A_2).$$

Use the result to construct the ANOVA decomposition of (1.7) with $\Gamma = 3$ through two nested constructions with $\Gamma = 2$.

1.4 For the discrete domains of Example 1.4, obtain $f_\emptyset$, f_1, f_2, and $f_{1,2}$ for $A_1 f = f(1, x_{\langle 2\rangle})$ and $A_2 f = \sum_{x_{\langle 2\rangle}=1}^{K_2} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_2$.

1.5 For the continuous domains of Example 1.5, obtain $f_\emptyset$, f_1, f_2, and $f_{1,2}$ for $A_1 f = f(0, x_{\langle 2\rangle})$ and $A_2 = \int_0^1 f dx_{\langle 2\rangle}$.

1.6 The domains $\mathcal{X}_\gamma$ in (1.7) can be a mixture of different types. As a simple example, consider $\Gamma = 2$, $\mathcal{X}_1 = \{1, \ldots, K\}$, and $\mathcal{X}_2 = [0, 1]$, with $A_1 f = \sum_{x_{\langle 1\rangle}=1}^{K} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K$ and $A_2 f = \int_0^1 f dx_{\langle 2\rangle}$. Obtain $f_\emptyset$, f_1, f_2, and $f_{1,2}$ in an ANOVA decomposition.

1.7 Prove that if the joint density of (U, V, Y, Z) has the expression

$$f(u, v, y, z) = \frac{e^{\eta_u + \eta_v + \eta_y + \eta_z + \eta_{u,y} + \eta_{u,z} + \eta_{v,y} + \eta_{v,z}}}{\int_{\mathcal{U}} \int_{\mathcal{V}} \int_{\mathcal{Y}} \int_{\mathcal{Z}} e^{\eta_u + \eta_v + \eta_y + \eta_z + \eta_{u,y} + \eta_{u,z} + \eta_{v,y} + \eta_{v,z}}},$$

then $(U \perp V) \big| (Y, Z)$ and $(Y \perp Z) \big| (U, V)$.

2
Model Construction

The two basic components of a statistical model, the deterministic part and the stochastic part, are well separated in the penalized likelihood score $L(f)+(\lambda/2)J(f)$ of (1.3). The deterministic part is specified via $J(f)$, which defines the notion of smoothness for functions on domain $\mathcal{X}$. The stochastic part is characterized by $L(f)$, which reflects the sampling structure of the data.

In this chapter, we are mainly concerned with the construction of $J(f)$ for use in $L(f)+(\lambda/2)J(f)$. At the foundation of the construction is some elementary theory of reproducing kernel Hilbert spaces, of which a brief self-contained introduction is given in §2.1. Illustrations of the construction are presented on the domain $\{1,\ldots,K\}$ through shrinkage estimates (§2.2) and on the domain $[0,1]$ through polynomial smoothing splines (§2.3); the discrete case also provides insights into the entities in a reproducing kernel Hilbert space through those in a standard vector space. The construction of models on product domains with the ANOVA structure of §1.3.2 built in is discussed in §2.4, with detailed examples on domains $\{1,\ldots,K_1\}\times\{1,\ldots,K_2\}$, $[0,1]^2$, and $\{1,\ldots,K\}\times[0,1]$.

Also included in this chapter are some general properties of the penalized likelihood score $L(f)+(\lambda/2)J(f)$ that are largely independent of $L(f)$. One such property is the fact that a quadratic functional $J(f)$ acts like the minus log likelihood of a Gaussian process prior for f, which leads to the Bayes model discussed in §2.5. Other important properties include the existence of the minimizer of $L(f)+(\lambda/2)J(f)$ and the equivalence of penalized minimization and constrained minimization (§2.6).

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_2,

The definitions of numerous technical terms are embedded in the text. For convenient back reference, the terms are set in boldface at the point of definition.

Mathematically more sophisticated constructions, such as the thin-plate splines on $(-\infty, \infty)^d$, are deferred to Chap. 4.

2.1 Reproducing Kernel Hilbert Spaces

By adding a roughness penalty $J(f)$ to the minus log likelihood $L(f)$, one considers only smooth functions in the space $\{f : J(f) < \infty\}$ or a subspace therein. To assist analysis and computation, one needs a metric and a geometry in the space, and the score $L(f) + (\lambda/2)J(f)$ to be continuous in f under the metric. The so-called reproducing kernel Hilbert space, of which a brief introduction is presented here, is adequately equipped for the purpose.

We start with the definition of Hilbert space and some of its elementary properties. The discussion is followed by the Riesz representation theorem, which provides the technical foundation for the notion of a reproducing kernel. The definition of reproducing kernel Hilbert space comes next and it is shown that a reproducing kernel Hilbert space is uniquely determined by its reproducing kernel, for which any non-negative definite function qualifies.

2.1.1 Hilbert Spaces and Linear Subspaces

As abstract generalizations of the familiar vector spaces, Hilbert spaces inherit many of the structures of the vector spaces. To provide insights into the technical concepts introduced here, abstract materials are followed by vector space examples set in italic.

For elements $f, g, h, \ldots$, define the operation of **addition** satisfying the following properties: (i) $f+g = g+f$, (ii) $(f+g)+h = f+(g+h)$, and (iii) for any two elements f and g, there exists an element h such that $f + h = g$. The third property implies the existence of an element 0 satisfying $f + 0 = f$. Further, define the operation of **scalar multiplication** satisfying $\alpha(f + g) = \alpha f + \alpha g$, $(\alpha + \beta)f = \alpha f + \beta f$, $1f = f$, and $0f = 0$, where α and β are real numbers. A set $\mathcal{L}$ of such elements form a **linear space** if $f, g \in \mathcal{L}$ implies that $f + g \in \mathcal{L}$ and $\alpha f \in \mathcal{L}$. A set of elements $f_i \in \mathcal{L}$ are said to be **linearly independent** if $\sum_i \alpha_i f_i = 0$ holds only for $\alpha_i = 0, \forall i$. The maximum number of elements in $\mathcal{L}$ that can be linearly independent defines its **dimension**.

Take real vectors of a given length as the elements; the standard vector addition and scalar-vector multiplication satisfy the conditions specified for

the operations of addition and scalar multiplication. The notions of linear space, linear independence, and dimension reduce to those in standard vector spaces.

A **functional** in a linear space $\mathcal{L}$ operates on an element $f \in \mathcal{L}$ and returns a real number as its value. A **linear functional** L in $\mathcal{L}$ satisfies $L(f+g) = Lf + Lg$, $L(\alpha f) = \alpha Lf$, $f, g \in \mathcal{L}$, α real. A **bilinear form** $J(f,g)$ in a linear space $\mathcal{L}$ takes $f, g \in \mathcal{L}$ as arguments and returns a real value and satisfies $J(\alpha f + \beta g, h) = \alpha J(f,h) + \beta J(g,h)$, $J(f, \alpha g + \beta h) = \alpha J(f,g) + \beta J(f,h)$, $f, g, h \in \mathcal{L}$, α, β real. Fixing one argument in a bilinear form, one gets a linear functional in the other argument. A bilinear form $J(\cdot,\cdot)$ is **symmetric** if $J(f,g) = J(g,f)$. A symmetric bilinear form is **non-negative definite** if $J(f,f) \geq 0$, $\forall f \in \mathcal{L}$, and it is **positive definite** if the equality holds only for $f = 0$. For $J(\cdot,\cdot)$ non-negative definite, $J(f) = J(f,f)$ is called a **quadratic functional**.

Consider the linear space of all real vectors of a given length. A functional in such a space is simply a multivariate function with the coordinates of the vector as its arguments. A linear functional in such a space can be written as a dot product, $Lf = g_L^T f$, where g_L is a vector "representing" L. A bilinear form can be written as $J(f,g) = f^T B_J g$ with B_J a square matrix, and $J(f,g)$ is symmetric, non-negative definite, or positive definite when B_J is symmetric, non-negative definite, or positive definite. A quadratic functional $J(f) = f^T B_J f$ is better known as a quadratic form in the classical linear model theory.

A linear space is often equipped with an **inner product**, a positive definite bilinear form with a notation $(\cdot,\cdot)$. An inner product defines a **norm** in the linear space, $\|f\| = \sqrt{(f,f)}$, which induces a metric to measure the **distance** between elements in the space, $D[f,g] = \|f - g\|$. The Cauchy-Schwarz inequality,

$$|(f,g)| \leq \|f\| \, \|g\|, \tag{2.1}$$

with equality if and only if $f = \alpha g$, and the triangle inequality,

$$\|f + g\| \leq \|f\| + \|g\|, \tag{2.2}$$

with equality if and only if $f = \alpha g$ for some $\alpha > 0$, hold in such a linear space; see Problems 2.1 and 2.2.

Equip the linear space of all real vectors of a given length with an inner product $(f,g) = f^T g$; one obtains the Euclidean space. The Euclidean norm $\|f\| = \sqrt{f^T f}$ induces the familiar Euclidean distance between vectors. The Cauchy-Schwarz inequality and the triangle inequality are familiar results in a Euclidean space.

When $\lim_{n\to\infty} \|f_n - f\| = 0$ for a sequence of elements f_n, the sequence is said to **converge** to its **limit point** f, with a notation $\lim_{n\to\infty} f_n = f$ or $f_n \to f$. A functional L is **continuous** if $\lim_{n\to\infty} Lf_n = Lf$ whenever $\lim_{n\to\infty} f_n = f$. By the Cauchy-Schwarz inequality, (f,g) is continuous in f or g when the other argument is fixed.

In the Euclidean space, a functional is a multivariate function in the coordinates of the vector, and the definition of continuity reduces to the definition found in standard multivariate calculus.

A sequence satisfying $\lim_{n,m\to\infty} \|f_n - f_m\| = 0$ is called a **Cauchy sequence**. A linear space $\mathcal{L}$ is **complete** if every Cauchy sequence in $\mathcal{L}$ converges to an element in $\mathcal{L}$. An element is a **limit point of a set** A if it is the limit point of a sequence in A. A set A is **closed** if it contains all of its own limit points.

The Euclidean space is complete. In the two-dimensional Euclidean space, $(-\infty, \infty) \times \{0\}$ *is a closed set, so is* $[a_1, b_1] \times [a_2, b_2]$*, where* $-\infty < a_i \leq b_i < \infty$*,* $i = 1, 2$.

A **Hilbert space** $\mathcal{H}$ is a complete inner product linear space. A closed linear subspace of $\mathcal{H}$ is itself a Hilbert space. The **distance** between a point $f \in \mathcal{H}$ and a closed linear subspace $\mathcal{G} \subset \mathcal{H}$ is defined by $D[f, \mathcal{G}] = \inf_{g\in\mathcal{G}} \|f - g\|$. By the closedness of $\mathcal{G}$, there exists an $f_\mathcal{G} \in \mathcal{G}$, called the **projection** of f in $\mathcal{G}$, such that $\|f - f_\mathcal{G}\| = D[f, \mathcal{G}]$. Such an $f_\mathcal{G}$ is unique by the triangle inequality. See Problem 2.3.

In the two-dimensional Euclidean space, $\mathcal{G} = \{f : f = (a, 0)^T, a \text{ real}\}$ *is a closed linear subspace. The distance between* $f = (a_f, b_f)^T$ *and* $\mathcal{G}$ *is* $D[f, \mathcal{G}] = |b_f|$*, and the projection of* f *in* $\mathcal{G}$ *is* $f_\mathcal{G} = (a_f, 0)^T$.

Proposition 2.1 *Let* $f_\mathcal{G}$ *be the projection of* $f \in \mathcal{H}$ *in a closed linear subspace* $\mathcal{G} \subset \mathcal{H}$*. Then,* $(f - f_\mathcal{G}, g) = 0$*,* $\forall g \in \mathcal{G}$.

Proof: We prove by negation. Suppose $(f - f_\mathcal{G}, h) = \alpha \neq 0$, $h \in \mathcal{G}$. Write $\beta = (h, h)$ and take $g = f_\mathcal{G} + (\alpha/\beta)h \in \mathcal{G}$. It is easy to compute

$$\|f - g\|^2 = \|f - f_\mathcal{G}\|^2 - \alpha^2/\beta < \|f - f_\mathcal{G}\|^2,$$

a contradiction. □

The linear subspace $\mathcal{G}^c = \{f : (f, g) = 0, \forall g \in \mathcal{G}\}$ is called the **orthogonal complement** of $\mathcal{G}$. By the continuity of (f, g), $\mathcal{G}^c$ is closed. Using Proposition 2.1, it is easy to verify that

$$\begin{aligned}\|f - f_\mathcal{G} - f_{\mathcal{G}^c}\|^2 &= (f - f_\mathcal{G} - f_{\mathcal{G}^c}, f - f_{\mathcal{G}^c} - f_\mathcal{G}) \\ &= (f - f_\mathcal{G}, f - f_{\mathcal{G}^c}) - (f - f_\mathcal{G}, f_\mathcal{G}) \\ &\quad - (f_{\mathcal{G}^c}, f - f_{\mathcal{G}^c}) + (f_{\mathcal{G}^c}, f_\mathcal{G}) \\ &= 0,\end{aligned}$$

where $f_\mathcal{G} \in \mathcal{G}$ and $f_{\mathcal{G}^c} \in \mathcal{G}^c$ are the projections of f in $\mathcal{G}$ and $\mathcal{G}^c$, respectively. Hence, there exists a unique decomposition $f = f_\mathcal{G} + f_{\mathcal{G}^c}$ for every $f \in \mathcal{H}$. It is clear now that $(\mathcal{G}^c)^c = \mathcal{G}$. The decomposition $f = f_\mathcal{G} + f_{\mathcal{G}^c}$ is called a **tensor sum decomposition** and is denoted by $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^c$, $\mathcal{G}^c = \mathcal{H} \ominus \mathcal{G}$, or $\mathcal{G} = \mathcal{H} \ominus \mathcal{G}^c$. Multiple-term tensor sum decompositions can be defined recursively.

In the two-dimensional Euclidean space, the orthogonal complement of $\mathcal{G} = \{f : f = (a, 0)^T, a \text{ real}\}$ *is* $\mathcal{G}^c = \{f : f = (0, b)^T, b \text{ real}\}$.

Consider linear subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$ of a linear space $\mathcal{L}$, equipped with inner products $(\cdot,\cdot)_0$ and $(\cdot,\cdot)_1$, respectively. Assume the completeness of $\mathcal{H}_0$ and $\mathcal{H}_1$ so that they are Hilbert spaces. If $\mathcal{H}_0$ and $\mathcal{H}_1$ have only one common element 0, then one may define a tensor sum Hilbert space $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ with elements $f = f_0 + f_1$ and $g = g_0 + g_1$, where $f_0, g_0 \in \mathcal{H}_0$ and $f_1, g_1 \in \mathcal{H}_1$, and an inner product $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1$. It is easy to verify that such a bottom-up pasting is consistent with the aforementioned top-down decomposition; see Problem 2.4.

Consider the two-dimensional vector space. Equip the space $\mathcal{H}_0 = \{f : f = (a, 0)^T, a \text{ real}\}$ *with the inner product* $(f, g)_0 = a_f a_g$, *where* $f = (a_f, 0)^T$ *and* $g = (a_g, 0)^T$, *and equip* $\mathcal{H}_1 = \{f : f = (0, b)^T, b \text{ real}\}$ *with the inner product* $(f, g)_1 = b_f b_g$, *where* $f = (0, b_f)^T$ *and* $g = (0, b_g)^T$. $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ *has elements of the form* $f = f_0 + f_1 = (a_f, 0)^T + (0, b_f)^T = (a_f, b_f)^T$ *and* $g = (a_g, 0)^T + (0, b_g)^T = (a_g, b_g)^T$, *and an inner product* $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1 = a_f a_g + b_f b_g$.

A non-negative definite bilinear form $J(f, g)$ in a linear space $\mathcal{H}$ defines a **semi-inner-product** in $\mathcal{H}$ which induces a square **seminorm** $J(f) = J(f, f)$. Unless $J(f, g)$ is positive definite, the **null space** $\mathcal{N}_J = \{f : J(f, f) = 0, f \in \mathcal{H}\}$ is a linear subspace of $\mathcal{H}$ containing more elements than just 0. With a nondegenerate $\mathcal{N}_J$, one typically can define another non-negative definite bilinear form $\tilde{J}(f, g)$ in $\mathcal{H}$ satisfying the following conditions: (i) it is positive definite when restricted to $\mathcal{N}_J$, so $\tilde{J}(f) = \tilde{J}(f, f)$ defines a square full norm in $\mathcal{N}_J$ and (ii) for every $f \in \mathcal{H}$, there exists $g \in \mathcal{N}_J$ such that $\tilde{J}(f - g) = 0$. With such an $\tilde{J}(f, g)$, it is easy to verify that $J(f, g)$ is positive definite in the linear subspace $\mathcal{N}_{\tilde{J}} = \{f : \tilde{J}(f, f) = 0, f \in \mathcal{H}\}$ and that $(J + \tilde{J})(f, g)$ is positive definite in $\mathcal{H}$. Hence, a semi-inner-product can be made a full inner product either via restriction to a subspace or via augmentation by an extra term, both through the definition of an inner product in its null space. If $\mathcal{H}$ is complete under the norm induced by $(J + \tilde{J})(f, g)$, then it is easy to see that $\mathcal{N}_J$ and $\mathcal{N}_{\tilde{J}}$ form a tensor sum decomposition of $\mathcal{H}$.

In the two-dimensional vector space $\mathcal{H}$ *with elements* $f = (a_f, b_f)^T$ *and* $g = (a_g, b_g)^T$, $J(f, g) = b_f b_g$ *defines a semi-inner-product with the null space* $\mathcal{N}_J = \{f : f = (a, 0)^T, a \text{ real}\}$. *Define* $\tilde{J}(f, g) = a_f a_g$, *which satisfies the two conditions specified above. It follows that* $\mathcal{N}_{\tilde{J}} = \{f : f = (0, b)^T, b \text{ real}\}$, *in which* $J(f, g) = b_f b_g$ *is positive definite. Clearly,* $(J + \tilde{J})(f, g) = b_f b_g + a_f a_g$ *is positive definite in* $\mathcal{H}$.

Example 2.1 (L_2 space) All square integrable functions on $[0, 1]$ form a Hilbert space

$$\mathcal{L}_2[0, 1] = \{f : \textstyle\int_0^1 f^2 dx < \infty\}$$

with an inner product $(f,g)=\int_0^1 fg\,dx$. The space

$$\mathcal{G}=\{f: f=gI_{[x\le 0.5]}, g\in\mathcal{L}_2[0,1]\}$$

is a closed linear subspace with an orthogonal complement

$$\mathcal{G}^c=\{f: f=gI_{[x\ge 0.5]}, g\in\mathcal{L}_2[0,1]\}.$$

Note that elements in $\mathcal{L}_2[0,1]$ are defined not by individual functions but by equivalent classes.

The bilinear form $J(f,g)=\int_0^{0.5} fg\,dx$ defines a semi-inner-product in $\mathcal{L}_2[0,1]$, with a null space

$$\mathcal{N}_J=\mathcal{G}^c=\{f: f=gI_{[x\ge 0.5]}, g\in\mathcal{L}_2[0,1]\}.$$

Define $\tilde{J}(f,g)=C\int_{0.5}^1 fg\,dx$, with $C>0$ a constant; one has an inner product $(f,g)=(J+\tilde{J})(f,g)=\int_0^{0.5} fg\,dx+C\int_{0.5}^1 fg\,dx$ on $\mathcal{L}_2[0,1]$. On $\mathcal{G}=\mathcal{L}_2\ominus\mathcal{N}_J$, $J(f,g)$ is a full inner product. □

Example 2.2 (Euclidean space) Functions on $\{1,\ldots,K\}$ are vectors of length K. Consider the Euclidean K-space with an inner product

$$(f,g)=\sum_{x=1}^K f(x)g(x)=f^Tg.$$

The space $\mathcal{G}=\{f: f(1)=\cdots=f(K)\}$ is a closed linear subspace with an orthogonal complement $\mathcal{G}^c=\{f:\sum_{x=1}^K f(x)=0\}$.

Write $\bar{f}=\sum_{x=1}^K f(x)/K$. The bilinear form

$$J(f,g)=\sum_{x=1}^K\big(f(x)-\bar{f}\big)\big(g(x)-\bar{g}\big)=f^T\left(I-\frac{1}{K}\mathbf{1}\mathbf{1}^T\right)g$$

defines a semi-inner-product in the vector space with a null space

$$\mathcal{N}_J=\mathcal{G}=\{f: f(1)=\cdots=f(K)\}.$$

Define $\tilde{J}(f,g)=C\bar{f}\bar{g}=Cf^T(\mathbf{1}\mathbf{1}^T/K)g$, with $C>0$ a constant; one has an inner product in the vector space,

$$(f,g)=(J+\tilde{J})(f,g)=f^T\left(I+\frac{C-1}{K}\mathbf{1}\mathbf{1}^T\right)g,$$

which reduces to the Euclidean inner product when $C=1$. On $\mathcal{G}^c=\{f:\sum_{x=1}^K f(x)=0\}$, $J(f,g)$ is a full inner product. □

2.1.2 Riesz Representation Theorem

For every g in a Hilbert space $\mathcal{H}$, $L_g f = (g, f)$ defines a continuous linear functional L_g. Conversely, every continuous linear functional L in $\mathcal{H}$ has a representation $Lf = (g_L, f)$ for some $g_L \in \mathcal{H}$, called the **representer** of L, as the following theorem asserts.

Theorem 2.2 (Riesz representation) *For every continuous linear functional L in a Hilbert space $\mathcal{H}$, there exists a unique $g_L \in \mathcal{H}$ such that $Lf = (g_L, f)$, $\forall f \in \mathcal{H}$.*

Proof: Let $\mathcal{N}_L = \{f : Lf = 0\}$ be the null space of L. Since L is continuous, $\mathcal{N}_L$ is a closed linear subspace. If $\mathcal{N}_L = \mathcal{H}$, take $g_L = 0$. When $\mathcal{N}_L \subset \mathcal{H}$, there exists a nonzero element $g_0 \in \mathcal{H} \ominus \mathcal{N}_L$. Since $(Lf)g_0 - (Lg_0)f \in \mathcal{N}_L$, $\big((Lf)g_0 - (Lg_0)f, g_0\big) = 0$. Some algebra yields

$$Lf = \left(\frac{Lg_0}{(g_0, g_0)} g_0, f\right).$$

Hence, one can take $g_L = (Lg_0)g_0/(g_0, g_0)$. The uniqueness is trivial. □

The continuity of L is necessary for the theorem to hold, or otherwise $\mathcal{N}_L$ is no longer closed and the proof breaks down.

All linear functionals in a finite-dimensional Hilbert space are continuous. Actually, there is an isomorphism between any K-dimensional Hilbert space and the Euclidean K-space. See Problems 2.5 and 2.6.

2.1.3 Reproducing Kernel and Non-Negative Definite Function

The likelihood part $L(f)$ of the penalized likelihood functional $L(f) + (\lambda/2)J(f)$ usually involves evaluations; thus, for it to be continuous in f, one needs the continuity of the **evaluation functional** $[x]f = f(x)$. Consider a Hilbert space $\mathcal{H}$ of functions on domain $\mathcal{X}$. If the evaluation functional $[x]f = f(x)$ is continuous in $\mathcal{H}$, $\forall x \in \mathcal{X}$, then $\mathcal{H}$ is called a **reproducing kernel Hilbert space**.

By the Riesz representation theorem, there exists $R_x \in \mathcal{H}$, the representer of the evaluation functional $[x](\cdot)$, such that $(R_x, f) = f(x)$, $\forall f \in \mathcal{H}$. The symmetric bivariate function $R(x, y) = R_x(y) = (R_x, R_y)$ has the reproducing property $\big(R(x, \cdot), f(\cdot)\big) = f(x)$ and is called the **reproducing kernel** of the space $\mathcal{H}$. The reproducing kernel is unique when it exists (Problem 2.7).

The $\mathcal{L}_2[0, 1]$ space of Example 2.1 is not a reproducing kernel Hilbert space. In fact, since the elements in $\mathcal{L}_2[0, 1]$ are defined by equivalent classes but not individual functions, evaluation is not even well defined. A finite-dimensional Hilbert space is always a reproducing kernel Hilbert space since all linear functionals are continuous.

Example 2.3 (Euclidean space) Consider again the Euclidean K-space with $(f, g) = f^T g$, with vectors perceived as functions on $\mathcal{X} = \{1, \ldots, K\}$. The evaluation functional $[x]f = f(x)$ is simply coordinate extraction. Since $f(x) = e_x^T f$, where e_x is the xth unit vector, one has $R_x(y) = I_{[x=y]}$. A bivariate function on $\{1, \ldots, K\}$ can be written as a square matrix, and the reproducing kernel in the Euclidean space is simply the identity matrix. □

A bivariate function $F(x, y)$ on $\mathcal{X}$ is said to be a **non-negative definite function** if $\sum_{i,j} \alpha_i \alpha_j F(x_i, x_j) \geq 0, \forall x_i \in \mathcal{X}, \forall \alpha_i$ real. For $R(x, y) = R_x(y)$ a reproducing kernel, it is easy to verify that

$$\Big\| \sum_i \alpha_i R_{x_i} \Big\|^2 = \sum_{i,j} \alpha_i \alpha_j R(x_i, x_j) \geq 0,$$

so $R(x, y)$ is non-negative definite. As a matter of fact, there exists a one-to-one correspondence between reproducing kernel Hilbert spaces and non-negative definite functions, as the following theorem asserts.

Theorem 2.3 *For every reproducing kernel Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$, there corresponds an unique reproducing kernel $R(x, y)$, which is non-negative definite. Conversely, for every non-negative definite function $R(x, y)$ on $\mathcal{X}$, there corresponds a unique reproducing kernel Hilbert space $\mathcal{H}$ that has $R(x, y)$ as its reproducing kernel.*

By Theorem 2.3, one may construct a reproducing kernel Hilbert space simply by specifying its reproducing kernel. The following lemma is needed in the proof of the theorem.

Lemma 2.4 *Let $R(x, y)$ be any non-negative definite function on $\mathcal{X}$. If*

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j R(x_i, x_j) = 0,$$

then $\sum_{i=1}^n \alpha_i R(x_i, x) = 0$, $\forall x \in \mathcal{X}$.

Proof: Augment the (x_i, α_i) sequence by adding (x_0, α_0), where $x_0 \in \mathcal{X}$ and α_0 real are arbitrary. Since

$$0 \leq \sum_{i=0}^n \sum_{j=0}^n \alpha_i \alpha_j R(x_i, x_j) = 2\alpha_0 \sum_{i=1}^n \alpha_i R(x_i, x_0) + \alpha_0^2 R(x_0, x_0)$$

and $R(x_0, x_0) \geq 0$, it is necessary that $\sum_{i=1}^n \alpha_i R(x_i, x_0) = 0$. □

Proof of Theorem 2.3: Only the converse needs a proof. Given $R(x, y)$, write $R_x = R(x, \cdot)$; one starts with the linear space

$$\mathcal{H}^* = \big\{ f : f = \textstyle\sum_i \alpha_i R_{x_i}, x_i \in \mathcal{X}, \alpha_i \text{ real} \big\},$$

and defines in $\mathcal{H}^*$ an inner product

$$\left(\sum_i \alpha_i R_{x_i}, \sum_j \beta_j R_{y_j}\right) = \sum_{i,j} \alpha_i \beta_j R(x_i, y_j).$$

It is trivial to verify the properties of inner product for such a (f, g), except that $(f, f) = 0$ holds only for $f = 0$, which is proved in Lemma 2.4. It is also easy to verify that $(R_x, f) = f(x)$, $\forall f \in \mathcal{H}^*$.

By the Cauchy-Schwarz inequality,

$$\big|f(x)\big| = \big|(R_x, f)\big| \le \sqrt{R(x,x)}\,\|f\|,$$

so convergence in norm implies pointwise convergence. For every Cauchy sequence $\{f_n\}$ in $\mathcal{H}^*$, $\big\{f_n(x)\big\}$ is a Cauchy sequence on the real line converging to a limit. Note also that $\big|\,\|f_n\| - \|f_m\|\,\big| \le \|f_n - f_m\|$, so $\big\{\|f_n\|\big\}$ has a limit as well. The limit point of $\{f_n\}$ can then be defined by $f(x) = \lim_{n\to\infty} f_n(x)$, $\forall x \in \mathcal{X}$, with $\|f\| = \lim_{n\to\infty} \|f_n\|$. It will be shown shortly that $\|f\|$, thus defined, is unique; that is, for two Cauchy sequences $\{f_n\}$ and $\{g_n\}$ satisfying $\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} g_n(x)$, $\forall x \in \mathcal{X}$, it is necessary that $\lim_{n\to\infty} \|f_n\| = \lim_{n\to\infty} \|g_n\|$. Adjoining all these limit points of Cauchy sequences to $\mathcal{H}^*$, one obtains a complete linear space $\mathcal{H}$ with the norm $\|f\|$. It is easy to verify that $(f, g) = \big(\|f+g\|^2 - \|f\|^2 - \|g\|^2\big)/2$ extends the inner product from $\mathcal{H}^*$ to $\mathcal{H}$ and that $(R_x, f) = f(x)$ holds in $\mathcal{H}$, so $\mathcal{H}$ is a reproducing kernel Hilbert space with $R(x, y)$ as its reproducing kernel.

We now verify the uniqueness of the definition of $\|f\|$ in the completed space, and it suffices to show that for every Cauchy sequence $\{f_n\}$ in $\mathcal{H}^*$ satisfying $\lim_{n\to\infty} f_n(x) = 0$, $\forall x \in \mathcal{X}$, it necessarily holds that $\lim_{n\to\infty} \|f_n\| = 0$. We prove the assertion by negation. Suppose $f_n(x) \to 0$, $\forall x \in \mathcal{X}$, but $\|f_n\|^2 \to 3\delta > 0$. Take $\epsilon \in (0, \delta)$. For n and m sufficiently large, one has $\|f_n\|^2, \|f_m\|^2 > 2\delta$ and $\|f_n - f_m\|^2 < \epsilon$. Fix such an m and write $f_m = \sum_i \alpha_i R_{x_i}$ a finite sum. Since $f_n(x) \to 0$, $\forall x \in \mathcal{X}$, it follows that $\sum_i \alpha_i f_n(x_i) \to 0$. Hence, for n sufficiently large,

$$\big|(f_n, f_m)\big| = \big|(f_n, \textstyle\sum_i \alpha_i R_{x_i})\big| = \big|\textstyle\sum_i \alpha_i f_n(x_i)\big| < \epsilon.$$

Now,

$$\epsilon > \|f_n - f_m\|^2 = \|f_n\|^2 + \|f_m\|^2 - 2(f_n, f_m) > 4\delta - 2\epsilon > 2\delta,$$

a contradiction.

It remains to be shown that if a space $\tilde{\mathcal{H}}$ has $R(x, y)$ as its reproducing kernel, then $\tilde{\mathcal{H}}$ must be identical to the space $\mathcal{H}$ constructed above. Since $R_x = R(x, \cdot) \in \tilde{\mathcal{H}}$, $\forall x \in \mathcal{X}$, so $\mathcal{H} \subseteq \tilde{\mathcal{H}}$. Now, for any $h \in \tilde{\mathcal{H}} \ominus \mathcal{H}$, by orthogonality, $h(x) = (R_x, h) = 0$, $\forall x \in \mathcal{X}$, so $\tilde{\mathcal{H}} = \mathcal{H}$. The proof is now complete. □

From the construction in the proof, one can see that the space $\mathcal{H}$ corresponding to R is generated from the "columns" $R_x = R(\cdot, x)$ of R, very much like a vector space generated from the columns of a matrix.

In the sections to follow, we will be constantly decomposing reproducing kernel Hilbert spaces into tensor sums or pasting up larger spaces by taking tensor sums of smaller ones. The following theorem spells out some of the rules in such operations.

Theorem 2.5 *If the reproducing kernel R of a space $\mathcal{H}$ on domain $\mathcal{X}$ can be decomposed into $R = R_0 + R_1$, where R_0 and R_1 are both non-negative definite, $R_0(x, \cdot), R_1(x, \cdot) \in \mathcal{H}$, $\forall x \in \mathcal{X}$, and $\big(R_0(x, \cdot), R_1(y, \cdot)\big) = 0$, $\forall x, y \in \mathcal{X}$, then the spaces $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding respectively to R_0 and R_1 form a tensor sum decomposition of $\mathcal{H}$. Conversely, if R_0 and R_1 are both non-negative definite and $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, then $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ has a reproducing kernel $R = R_0 + R_1$.*

Proof: By the orthogonality between $R_0(x, \cdot)$ and $R_1(y, \cdot)$,

$$R_0(x, y) = \big(R_0(x, \cdot), R(y, \cdot)\big) = \big(R_0(x, \cdot), R_0(y, \cdot)\big),$$

so the inner product in $\mathcal{H}_0$ is consistent with that in $\mathcal{H}$; hence, $\mathcal{H}_0$ is a closed linear subspace of $\mathcal{H}$. Now, for every $f \in \mathcal{H}$, let f_0 be the projection of f in $\mathcal{H}_0$ and write $f = f_0 + f_0^c$. Straightforward calculation yields

$$\begin{aligned} f(x) &= \big(R(x, \cdot), f\big) \\ &= \big(R_0(x, \cdot), f_0\big) + \big(R_0(x, \cdot), f_0^c\big) + \big(R_1(x, \cdot), f_0\big) + \big(R_1(x, \cdot), f_0^c\big) \\ &= f_0(x) + \big(R_1(x, \cdot), f_0^c\big), \end{aligned}$$

so $\big(R_1(x, \cdot), f_0^c\big) = f(x) - f_0(x) = f_0^c(x)$. This shows that R_1 is the reproducing kernel of $\mathcal{H} \ominus \mathcal{H}_0$; hence, $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$.

For the converse, it is trivial to verify that

$$\big(R(x, \cdot), f\big) = \big(R_0(x, \cdot), f_0\big)_0 + \big(R_1(x, \cdot), f_1\big)_1 = f_0(x) + f_1(x) = f(x),$$

where $f = f_0 + f_1 \in \mathcal{H}$ with $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$, and $(\cdot, \cdot)_0$ and $(\cdot, \cdot)_1$ are the inner products in $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. □

2.2 Smoothing Splines on $\{1, \ldots, K\}$

As discussed in Example 2.3, a function on the discrete domain $\mathcal{X} = \{1, \ldots, K\}$ is a vector of length K, evaluation is coordinate extraction, and a reproducing kernel can be written as a non-negative definite matrix. A linear functional in a finite-dimensional space is always continuous, so a vector space equipped with an inner product is a reproducing kernel Hilbert space.

Let B be any $K \times K$ non-negative definite matrix. Consider the column space of B, $\mathcal{H}_B = \{f : f = B\mathbf{c} = \sum_j c_j B(\cdot, j)\}$, equipped with the inner product $(f, g) = f^T B g$. The standard eigenvalue decomposition gives

$$B = UDU^T = (U_1, U_2) \begin{pmatrix} D_1 & O \\ O & O \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} = U_1 D_1 U_1^T,$$

where the diagonal of D_1 contains the positive eigenvalues of B and the columns of U_1 are the associated eigenvectors. The Moore-Penrose inverse of B has an expression $B^+ = U_1 D_1^{-1} U_1^T$. It is clear that $\mathcal{H}_B = \mathcal{H}_{B^+} = \{f : f = U_1 \mathbf{c}\}$. Now, $B^+ B = U_1 U_1^T$ is the projection matrix onto $\mathcal{H}_B$, so $B^+ B f = f$, $\forall f \in \mathcal{H}_B$. It then follows that

$$[x]f = f(x) = e_x^T f = e_x^T B^+ B f = (B^+ e_x)^T B f,$$

$\forall f \in \mathcal{H}_B$ (i.e., the representer of $[x](\cdot)$ is the xth column of B^+). Hence, the reproducing kernel is given by $R(x, y) = B^+(x, y)$, where $B^+(x, y)$ is the (x, y)th entry of B^+. The result of Example 2.3 is a trivial special case with $B = I$.

The duality between $(f, g) = f^T B g$ and $R = B^+$ provides a useful insight into the relation between the inner product in a space and the corresponding reproducing kernel: *In a sense, the inner product and the reproducing kernel are inverses of each other.*

Now, consider a decomposition of the reproducing kernel in the Euclidean K-space, $R(x, y) = I_{[x=y]} = 1/K + (I_{[x=y]} - 1/K)$, or in matrix terms, $I = (\mathbf{1}\mathbf{1}^T/K) + (I - \mathbf{1}\mathbf{1}^T/K)$. Since $(\mathbf{1}\mathbf{1}^T/K)(I - \mathbf{1}\mathbf{1}^T/K) = O$, $\big(R_0(x, \cdot), R_1(y, \cdot)\big) = 0$, $\forall x, y$. By Theorem 2.5, the decomposition defines a tensor sum decomposition of the space $R^K = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{f : f(1) = \cdots = f(K)\}$ and $\mathcal{H}_1 = \{f : \sum_{x=1}^K f(x) = 0\}$. The inner products in $\mathcal{H}_0$ and $\mathcal{H}_1$ have expressions $(f, g)_0 = f^T g = f^T(\mathbf{1}\mathbf{1}^T/K)g$ and $(f, g)_1 = f^T g = f^T(I - \mathbf{1}\mathbf{1}^T/K)g$, respectively, where $\mathbf{1}\mathbf{1}^T/K$ is the Moore-Penrose inverse of $R_0 = \mathbf{1}\mathbf{1}^T/K$ and $I - \mathbf{1}\mathbf{1}^T/K$ is the Moore-Penrose inverse of $R_1 = I - \mathbf{1}\mathbf{1}^T/K$. The decomposition defines a one-way ANOVA decomposition with an averaging operator $Af = \sum_{x=1}^K f(x)/K$. See Problem 2.8 for a construction yielding a one-way ANOVA decomposition with an averaging operator $Af = f(1)$.

Regression on $\mathcal{X} = \{1, \ldots, K\}$ yields the classical one-way ANOVA model. Consider a roughness penalty

$$J(f) = \sum_{x=1}^K \big(f(x) - \bar{f}\big)^2 = f^T \left(I - \frac{\mathbf{1}\mathbf{1}^T}{K}\right) f,$$

where $\bar{f} = \sum_{x=1}^K f(x)/K$. The minimizer of

$$\frac{1}{n} \sum_{i=1}^n \big(Y_i - \eta(x_i)\big)^2 + \lambda \sum_{x=1}^K \big(\eta(x) - \bar{\eta}\big)^2 \tag{2.3}$$

defines a shrinkage estimate being shrunk toward a constant. Similarly, if one sets $J(f) = f^T f$, then the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\sum_{x=1}^{K}\eta^2(x) \tag{2.4}$$

defines a shrinkage estimate being shrunk toward zero. Hence, smoothing splines on a discrete domain reduce to shrinkage estimates.

The roughness penalty $\sum_{x=1}^{K}\big(f(x)-\bar{f}\big)^2$ appears natural for x nominal. For x ordinal, however, one may consider alternatives such as

$$\sum_{x=2}^{K}\big(f(x)-f(x-1)\big)^2,$$

which have the same null space but use different "scaling" in the penalized contrast space $\mathcal{H}_1 = \big\{f : \sum_{x=1}^{K} f(x) = 0\big\}$.

2.3 Polynomial Smoothing Splines on $[0,1]$

The cubic smoothing spline of §1.1.1 is a special case of the polynomial smoothing splines, the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1\big(\eta^{(m)}\big)^2 dx, \tag{2.5}$$

in the space $\mathcal{C}^{(m)}[0,1] = \big\{f : f^{(m)} \in \mathcal{L}_2[0,1]\big\}$. Equipped with appropriate inner products, the space $\mathcal{C}^{(m)}[0,1]$ can be made a reproducing kernel Hilbert space.

We will present two such constructions and outline an approach to the computation of polynomial smoothing splines. The two constructions yield identical results for univariate smoothing, but provide building blocks satisfying different side conditions for multivariate smoothing with built-in ANOVA decompositions.

2.3.1 A Reproducing Kernel in $\mathcal{C}^{(m)}[0,1]$

For $f \in \mathcal{C}^{(m)}[0,1]$, the standard Taylor expansion gives

$$f(x) = \sum_{\nu=0}^{m-1}\frac{x^\nu}{\nu!}f^{(\nu)}(0) + \int_0^1\frac{(x-u)_+^{m-1}}{(m-1)!}f^{(m)}(u)du, \tag{2.6}$$

where $(\cdot)_+ = \max(0,\cdot)$. With an inner product

$$(f,g) = \sum_{\nu=0}^{m-1}f^{(\nu)}(0)g^{(\nu)}(0) + \int_0^1 f^{(m)}g^{(m)}dx, \tag{2.7}$$

it can be shown that the representer of evaluation $[x](\cdot)$ is

$$R_x(y) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!} + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du. \tag{2.8}$$

To see this, note that $R_x^{(\nu)}(0) = x^\nu/\nu!$, $\nu = 0, \ldots, m-1$, and that $R_x^{(m)}(y) = (x-y)_+^{m-1}/(m-1)!$. Plugging these into (2.7) with $g = R_x$, one obtains the right-hand side of (2.6), so $(R_x, f) = f(x)$.

The two terms of the reproducing kernel $R(x, y) = R_x(y)$,

$$R_0(x, y) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!}, \tag{2.9}$$

and

$$R_1(x, y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du, \tag{2.10}$$

are both non-negative definite themselves, and it is also easy to verify the other conditions of Theorem 2.5. To R_0 there corresponds the space of polynomials $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$ with an inner product $(f, g)_0 = \sum_{\nu=0}^{m-1} f^{(\nu)}(0)g^{(\nu)}(0)$, and to R_1 there corresponds the orthogonal complement of $\mathcal{H}_0$,

$$\mathcal{H}_1 = \{f : f^{(\nu)}(0) = 0, \nu = 0, \ldots, m-1, \textstyle\int_0^1 \left(f^{(m)}\right)^2 dx < \infty\}, \tag{2.11}$$

with an inner product $(f, g)_1 = \int_0^1 f^{(m)}g^{(m)}dx$. The space $\mathcal{H}_0$ can be further decomposed into the tensor sum of m subspaces of monomials $\{f : f \propto (\cdot)^\nu\}$ with inner products $f^{(\nu)}(0)g^{(\nu)}(0)$ and reproducing kernels $(x^\nu/\nu!)(y^\nu/\nu!)$, $\nu = 0, \ldots, m-1$.

Setting $m = 1$, one has $R_0(x, y) = 1$ and

$$R_1(x, y) = \int_0^1 I_{[u<x]}I_{[u<y]}du = x \wedge y, \tag{2.12}$$

where $x \wedge y = \min(x, y)$. This setting is useful for the computation of a linear smoothing spline, the minimizer of

$$\frac{1}{n}\sum_{i=1}^n \left(Y_i - \eta(x_i)\right)^2 + \lambda \int_0^1 \dot{\eta}^2 dx. \tag{2.13}$$

Setting $m = 2$, one has $R_0(x, y) = 1 + xy$ and

$$\begin{aligned} R_1(x, y) &= \int_0^1 (x-u)_+(y-u)_+du \\ &= (x \wedge y)^2\big(3(x \vee y) - (x \wedge y)\big)/6, \end{aligned} \tag{2.14}$$

where $x \vee y = \max(x, y)$. The latter formula can be used in the computation of a cubic smoothing spline.

For $m = 1$, the tensor sum decomposition characterized by $R = R_0 + R_1 = [1] + [x \wedge y]$ naturally defines a one-way ANOVA decomposition with an averaging operator $Af = f(0)$, where the corresponding $\mathcal{H}_0$ spans the "mean" space and $\mathcal{H}_1$ spans the "contrast" space; see §1.3.1 for discussions on ANOVA decomposition and averaging operator.

For $m = 2$, the same ANOVA decomposition is characterized by the kernel decomposition

$$R = R_{00} + [R_{01} + R_1] = [1] + \big[xy + \big\{(x \wedge y)^2\big(3(x \vee y) - (x \wedge y)\big)/6\big\}\big],$$

where $R_0 = 1 + xy$ is further decomposed into the sum of $R_{00} = 1$ and $R_{01} = xy$. The kernel R_{00} generates the "mean" space and the kernels R_{01} and R_1 together generate the "contrast" space, with R_{01} contributing to the "parametric contrast" and R_1 to the "nonparametric contrast."

2.3.2 Computation of Polynomial Smoothing Splines

Given the sampling points x_i, $i = 1, \ldots, n$ in (2.5) and noting that the space $\big\{f : f = \sum_{i=1}^n \alpha_i R_1(x_i, \cdot)\big\}$ is a closed linear subspace of $\mathcal{H}_1$ given in (2.11), one may write $\eta \in \mathcal{C}^{(m)}[0, 1]$ as

$$\eta(x) = \sum_{\nu=0}^{m-1} d_\nu \frac{x^\nu}{\nu!} + \sum_{i=1}^n c_i R_1(x_i, x) + \rho(x), \tag{2.15}$$

where c_i and d_ν are real coefficients, R_1 is given in (2.10), and

$$\rho \in \mathcal{H}_1 \ominus \big\{f : f = \textstyle\sum_{i=1}^n c_i R_1(x_i, \cdot)\big\}.$$

By orthogonality, $\rho(x_i) = \big(R_1(x_i, \cdot), \rho\big) = 0$, $i = 1, \ldots, n$. Denoting by S the $n \times m$ matrix with the (i, ν)th entry $x_i^\nu/\nu!$ and by Q the $n \times n$ matrix with the (i, j)th entry $R_1(x_i, x_j)$, (2.5) can be written as

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda\, \mathbf{c}^T Q\mathbf{c} + n\lambda\, (\rho, \rho), \tag{2.16}$$

where the fact that $\int_0^1 R_1^{(m)}(x_i, x) R_1^{(m)}(x_j, x) dx = R_1(x_i, x_j)$ is used. Note that ρ only appears in the third term in (2.16), which is minimized at $\rho = 0$. Hence, a polynomial smoothing spline resides in a space

$$\mathcal{H}_0 \oplus \big\{f : f = \textstyle\sum_{i=1}^n c_i R_1(x_i, \cdot)\big\},$$

of finite dimension, and so can be computed via the minimization of the first two terms of (2.16) with respect to $\mathbf{c}$ and $\mathbf{d}$.

In this approach to the computation of polynomial smoothing splines, one needs the reproducing kernel R_1 that corresponds to a space $\mathcal{H}_1$ in which the roughness penalty $\int_0^1 \big(f^{(m)}\big)^2 dx$ is a full square norm, plus a basis that spans the null space of the penalty.

2.3.3 Another Reproducing Kernel in $\mathcal{C}^{(m)}[0,1]$

The bilinear form $\int_0^1 f^{(m)}g^{(m)}dx$ is a semi-inner-product in $\mathcal{C}^{(m)}[0,1]$, which can be augmented to a full inner product by the addition of an inner product in its null space, the space $\{f : f^{(m)} = 0\}$ of polynomials up to order $m-1$. In §2.3.1, we used $\sum_{\nu=0}^{m-1} f^{(\nu)}(0)g^{(\nu)}(0)$ as the inner product in $\{f : f^{(m)} = 0\}$. In this section, we will use a different inner product, $\sum_{\nu=0}^{m-1}\big(\int_0^1 f^{(\nu)}dx\big)\big(\int_0^1 g^{(\nu)}dx\big)$, in $\{f : f^{(m)} = 0\}$, and derive the reproducing kernel associated with

$$(f,g) = \sum_{\nu=0}^{m-1}\Big(\int_0^1 f^{(\nu)}dx\Big)\Big(\int_0^1 g^{(\nu)}dx\Big) + \int_0^1 f^{(m)}g^{(m)}dx, \tag{2.17}$$

which defines an inner product different from that in (2.7).

The sought-after reproducing kernel can most conveniently be expressed in terms of the functions

$$k_r(x) = -\Bigg(\sum_{\mu=-\infty}^{-1} + \sum_{\mu=1}^{\infty}\Bigg)\frac{\exp(2\pi\mathbf{i}\mu x)}{(2\pi\mathbf{i}\mu)^r}, \quad r = 1, 2, \ldots, \tag{2.18}$$

where $\mathbf{i} = \sqrt{-1}$. It is easy to verify that for $r > 1$, k_r is well defined and continuous on the real line, and for $r = 1$, it is well defined and continuous at noninteger points; see Problem 2.9(a). It is also easy to verify that $k_r(x)$ is real-valued and is periodic with period 1; see Problem 2.9(b). It can be seen that $k_r^{(p)} = k_{r-p}$, $p = 1, \ldots, r-2$ and that $k_r^{(r-1)}(x) = k_1(x)$ for x not an integer. It is known that $k_1(x) = x - 0.5$ on $(0,1)$ (Problem 2.9(c)), and we define $k_0 = 1$. The k_r functions are actually scaled Bernoulli polynomials, $k_r(x) = B_r(x)/r!$; see Abramowitz and Stegun (1964, Chap. 23) for a comprehensive list of results concerning the Bernoulli polynomials $B_r(x)$.

From the properties listed above, it is easy to verify that $\int_0^1 k_\mu^{(\nu)}dx = \delta_{\mu,\nu}$, $\mu, \nu = 0, \ldots, m-1$, where $\delta_{\mu,\nu}$ is the Kronecker delta. It then follows that k_ν, $\nu = 0, \ldots, m-1$ form an orthonormal basis of $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$ under the inner product $(f,g)_0 = \sum_{\nu=0}^{m-1}\big(\int_0^1 f^{(\nu)}dx\big)\big(\int_0^1 g^{(\nu)}dx\big)$ and that

$$R_0(x,y) = \sum_{\nu=0}^{m-1} k_\nu(x)k_\nu(y) \tag{2.19}$$

is the reproducing kernel in $\mathcal{H}_0$; see Problem 2.5(c) for the definition of orthonormal basis. In fact, $\mathcal{H}_0$ can be further decomposed into the tensor sum of m subspaces $\{f : f \propto k_\nu\}$ with inner products $\big(\int_0^1 f^{(\nu)}dx\big)\big(\int_0^1 g^{(\nu)}dx\big)$ and reproducing kernels $k_\nu(x)k_\nu(y)$, $\nu = 0, \ldots, m-1$, respectively.

We now show that in the space

$$\mathcal{H}_1 = \Big\{f : \int_0^1 f^{(\nu)}dx = 0, \nu = 0, \ldots, m-1, f^{(m)} \in \mathcal{L}_2[0,1]\Big\} \tag{2.20}$$

with a square norm $(f,g)_1 = \int_0^1 f^{(m)} g^{(m)} dx$, the function

$$R_x(y) = k_m(x)k_m(y) + (-1)^{m-1} k_{2m}(x-y) \tag{2.21}$$

is the representer of evaluation $[x](\cdot)$. From the properties of k_r, it is easy to verify that $\int_0^1 R_x^{(\nu)}(y)dy = 0$, $\nu = 0, \ldots, m-1$, and that $R_x^{(m)}(y) = k_m(x) - k_m(x-y) \in \mathcal{L}_2[0,1]$, so $R_x \in \mathcal{H}_1$ for $\mathcal{H}_1$ given in (2.20). Integrating by parts, and using the periodicity of k_r, $r > 1$, and the fact that $\int_0^1 f^{(\nu)} dx = 0$, $\nu = 0, \ldots, m-1$, one can show that, for $m > 1$,

$$\begin{aligned}(R_x, f)_1 &= \int_0^1 \big(k_m(x) - k_m(x-y)\big) f^{(m)}(y) dy \\ &= -\int_0^1 k_{m-1}(x-y) f^{(m-1)}(y) dy \\ &= \cdots = -\int_0^1 k_1(x-y) \dot{f}(y) dy; \end{aligned} \tag{2.22}$$

see Problem 2.10. Now, since

$$k_1(x-y) = \begin{cases} x - y - 0.5 = k_1(x) - y, & y \in (0,x), \\ (1+x-y) - 0.5 = k_1(x) - y + 1, & y \in (x,1), \end{cases}$$

straightforward calculation yields

$$\begin{aligned} -\int_0^1 & k_1(x-y)\dot{f}(y)dy \\ &= -\int_0^1 k_1(x)\dot{f}(y)dy + \int_0^1 y\dot{f}(y)dy - \int_x^1 \dot{f}(y)dy \\ &= 0 + f(1) - \big(f(1) - f(x)\big) = f(x). \end{aligned}$$

The result holds for $m = 1$ via direct calculation. This proves that

$$R_1(x,y) = k_m(x)k_m(y) + (-1)^{m-1} k_{2m}(x-y) \tag{2.23}$$

is the reproducing kernel of $\mathcal{H}_1$ given in (2.20).

Obviously, $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, so by the converse of Theorem 2.5, $\mathcal{C}^{(m)}[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_1$ has the reproducing kernel $R = R_0 + R_1$. The identity

$$f(x) = \sum_{\nu=0}^{m-1} k_\nu(x) \int_0^1 f^{(\nu)}(y)dy + \int_0^1 \big(k_m(x) - k_m(x-y)\big) f^{(m)}(y)dy, \tag{2.24}$$

$\forall f \in \mathcal{C}^{(m)}[0,1]$, may be called a generalized Taylor expansion, where the scaled Bernoulli polynomials $k_\nu(x)$ play the role of the scaled monomials $x^\nu/\nu!$ in the standard Taylor expansion of (2.6). The standard Taylor expansion is asymmetric with respect to the domain $[0,1]$, in the sense that

a swapping of the two ends 0 and 1 would change its composition entirely, whereas the generalized Taylor expansion of (2.24) is symmetric with respect to the domain.

The computation of polynomial smoothing splines as outlined in §2.3.2 can also be performed by using the R_1 of (2.23) instead of that of (2.10). Also, one may use any basis $\{\phi_\nu\}_{\nu=0}^{m-1}$ of the subspace $\mathcal{H}_0$ in the place of $\{x^\nu/\nu!\}_{\nu=0}^{m-1}$ in the expression of η given in (2.15). The coefficients c_i and d_ν will be different when different ϕ_ν and R_1 are used, but the function estimate

$$\eta(x) = \sum_{\nu=0}^{m-1} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R_1(x_i, x)$$

will remain the same regardless of the choices of ϕ_ν and R_1.

When $m = 1$, $R_0(x, y) = 1$ and

$$R_1(x, y) = k_1(x)k_1(y) + k_2(x - y). \tag{2.25}$$

When $m = 2$, $R_0(x, y) = 1 + k_1(x)k_1(y)$ and

$$R_1(x, y) = k_2(x)k_2(y) - k_4(x - y). \tag{2.26}$$

The R_1 in (2.25) and (2.26) can be used in the computation of linear and cubic smoothing splines in lieu of those in (2.12) and (2.14). To calculate R_1 in (2.25) and (2.26), one has, on $x \in [0, 1]$,

$$\begin{aligned} k_2(x) &= \frac{1}{2}\Big(k_1^2(x) - \frac{1}{12}\Big), \\ k_4(x) &= \frac{1}{24}\Big(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240}\Big), \end{aligned} \tag{2.27}$$

where $k_1(x) = x - 0.5$; see Problem 2.11. Note that k_2 and k_4 are symmetric with respect to 0.5 on $[0, 1]$, so for $x \in [-1, 0]$,

$$k_2(x) = k_2(x + 1) = k_2\big(0.5 + (x + 0.5)\big) = k_2\big(0.5 - (x + 0.5)\big) = k_2(-x),$$

and likewise, $k_4(x) = k_4(-x)$. It then follows that $k_2(x - y) = k_2\big(|x - y|\big)$ and $k_4(x - y) = k_4\big(|x - y|\big)$, for $x, y \in [0, 1]$.

For $m = 1$, the tensor sum decomposition characterized by $R = R_0 + R_1 = [1] + \big[k_1(x)k_1(y) + k_2(x - y)\big]$ defines a one-way ANOVA decomposition with an averaging operator $Af = \int_0^1 f dx$, where the corresponding $\mathcal{H}_0$ spans the "mean" space and $\mathcal{H}_1$ spans the "contrast" space.

For $m = 2$, the same ANOVA decomposition is characterized by the kernel decomposition

$$R = R_{00} + [R_{01} + R_1] = [1] + \big[k_1(x)k_1(y) + \big\{k_2(x)k_2(y) - k_4(x - y)\big\}\big],$$

where $R_0 = 1 + k_1(x)k_1(y)$ is further decomposed into the sum of $R_{00} = 1$ and $R_{01} = k_1(x)k_1(y)$. The kernel R_{00} generates the "mean" space and

the kernels R_{01} and R_1 together generate the "contrast" space, with R_{01} contributing to the "parametric contrast" and R_1 to the "nonparametric contrast."

2.4 Smoothing Splines on Product Domains

To incorporate the ANOVA decomposition introduced in §1.3.2 for the estimation of a multivariate function, one may construct a tensor product reproducing kernel Hilbert space. Given Theorem 2.3, the construction of the space can be done through the construction of the reproducing kernel, for which one uses reproducing kernels on the marginal domains. One-way ANOVA decompositions on the marginal domains naturally induce an ANOVA decomposition on the product domain.

We begin with some general discussion of tensor product reproducing kernel Hilbert spaces, where it is shown that the products of reproducing kernels on the marginal domains form reproducing kernels on the product domain. The construction is then illustrated with marginal domains $\{1, \ldots, K\}$ and $[0, 1]$, using the (marginal) reproducing kernels introduced in §§2.2 and 2.3.

2.4.1 Tensor Product Reproducing Kernel Hilbert Spaces

A convenient approach to the construction of reproducing kernel Hilbert spaces on a product domain $\prod_{\gamma=1}^{\Gamma} \mathcal{X}_\gamma$ is by taking the tensor product of spaces constructed on the marginal domains $\mathcal{X}_\gamma$. The construction builds on the following theorem.

Theorem 2.6 *For $R_{\langle 1\rangle}(x_{\langle 1\rangle}, y_{\langle 1\rangle})$ non-negative definite on $\mathcal{X}_1$ and $R_{\langle 2\rangle}(x_{\langle 2\rangle}, y_{\langle 2\rangle})$ non-negative definite on $\mathcal{X}_2$, $R(x, y) = R_{\langle 1\rangle}(x_{\langle 1\rangle}, y_{\langle 1\rangle})R_{\langle 2\rangle}(x_{\langle 2\rangle}, y_{\langle 2\rangle})$ is non-negative definite on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$.*

Proof: It suffices to show that, for two non-negative definite matrices A and B of the same size, their entrywise product, $A \circ B$, is necessarily non-negative definite. By elementary matrix theory, A and B are non-negative definite if and only if there exist vectors a_i and b_j such that $A = \sum_i a_i a_i^T$ and $B = \sum_j b_j b_j^T$. Now,

$$\begin{aligned} A \circ B &= \left(\sum_i a_i a_i^T\right) \circ \left(\sum_j b_j b_j^T\right) \\ &= \sum_{i,j} (a_i a_i^T) \circ (b_j b_j^T) = \sum_{i,j} (a_i \circ b_j)(a_i \circ b_j)^T, \end{aligned}$$

so $A \circ B$ is non-negative definite. □

By Theorem 2.3, every non-negative definite function R on domain $\mathcal{X}$ corresponds to a reproducing kernel Hilbert space with R as its reproducing

kernel. Given $\mathcal{H}_{\langle 1\rangle}$ on $\mathcal{X}_1$ with reproducing kernel $R_{\langle 1\rangle}$ and $\mathcal{H}_{\langle 2\rangle}$ on $\mathcal{X}_2$ with reproducing kernel $R_{\langle 2\rangle}$, $R = R_{\langle 1\rangle}R_{\langle 2\rangle}$ is non-negative definite on $\mathcal{X}_1 \times \mathcal{X}_2$ by Theorem 2.6. The reproducing kernel Hilbert space corresponding to such an R is called the **tensor product space** of $\mathcal{H}_{\langle 1\rangle}$ and $\mathcal{H}_{\langle 2\rangle}$, and is denoted by $\mathcal{H}_{\langle 1\rangle} \otimes \mathcal{H}_{\langle 2\rangle}$. The operation extends to multiple-term products recursively.

Suppose one has reproducing kernel Hilbert spaces $\mathcal{H}_{\langle\gamma\rangle}$ on domains $\mathcal{X}_\gamma$, $\gamma = 1, \ldots, \Gamma$, respectively. Further, assume that the spaces have one-way ANOVA decompositions built in via the tensor sum decompositions $\mathcal{H}_{\langle\gamma\rangle} = \mathcal{H}_{0\langle\gamma\rangle} \oplus \mathcal{H}_{1\langle\gamma\rangle}$, where $\mathcal{H}_{0\langle\gamma\rangle} = \{f : f \propto 1\}$ has a reproducing kernel $R_{0\langle\gamma\rangle} \propto 1$ and $\mathcal{H}_{1\langle\gamma\rangle}$ has a reproducing kernel $R_{1\langle\gamma\rangle}$ satisfying side conditions $A_\gamma R_{1\langle\gamma\rangle}(x_{\langle\gamma\rangle}, \cdot) = 0$, $\forall x_{\langle\gamma\rangle} \in \mathcal{X}_\gamma$, where A_γ are the averaging operators defining the one-way ANOVA decompositions on $\mathcal{X}_\gamma$. The tensor product space $\mathcal{H} = \otimes_{\gamma=1}^{\Gamma}\mathcal{H}_{\langle\gamma\rangle}$ has a tensor sum decomposition

$$\mathcal{H} = \bigotimes_{\gamma=1}^{\Gamma}(\mathcal{H}_{0\langle\gamma\rangle} \oplus \mathcal{H}_{1\langle\gamma\rangle}) = \bigoplus_{\mathcal{S}}\Big\{\Big(\bigotimes_{\gamma\in\mathcal{S}}\mathcal{H}_{1\langle\gamma\rangle}\Big) \otimes \Big(\bigotimes_{\gamma\notin\mathcal{S}}\mathcal{H}_{0\langle\gamma\rangle}\Big)\Big\} = \bigoplus_{\mathcal{S}}\mathcal{H}_{\mathcal{S}}, \quad (2.28)$$

which parallels (1.7) on page 7, where the summation is over all subsets $\mathcal{S} \subseteq \{1, \ldots, \Gamma\}$. The term $\mathcal{H}_{\mathcal{S}}$ has a reproducing kernel $R_{\mathcal{S}} \propto \prod_{\gamma\in\mathcal{S}} R_{1\langle\gamma\rangle}$, and the projection of $f \in \mathcal{H}$ in $\mathcal{H}_{\mathcal{S}}$ is the $f_{\mathcal{S}}$ appearing in (1.7). The minimizer of $L(f) + (\lambda/2)J(f)$ in a tensor product reproducing kernel Hilbert space is called a **tensor product smoothing spline**. Examples of the construction follow.

2.4.2 Reproducing Kernel Hilbert Spaces on $\{1, \ldots, K\}^2$

Set $A_\gamma f = \sum_{x_{\langle\gamma\rangle}=1}^{K_\gamma} f(x)/K_\gamma$ on discrete domains $\mathcal{X}_\gamma = \{1, \ldots, K_\gamma\}$, $\gamma = 1, 2$. The marginal reproducing kernels that define the one-way ANOVA decomposition on $\mathcal{X}_\gamma$ can be taken as $R_{0\langle\gamma\rangle}(x_{\langle\gamma\rangle}, y_{\langle\gamma\rangle}) = 1/K_\gamma$ and

$$R_{1\langle\gamma\rangle}(x_{\langle\gamma\rangle}, y_{\langle\gamma\rangle}) = I_{[x_{\langle\gamma\rangle}=y_{\langle\gamma\rangle}]} - 1/K_\gamma,$$

$\gamma = 1, 2$, as given in §2.2.

A function on $\{1, \ldots, K_1\} \times \{1, \ldots, K_2\}$ can be written as a vector of length K_1K_2,

$$f = \big(f(1,1), \ldots, f(1,K_2), \ldots, f(K_1,1), \ldots, f(K_1,K_2)\big)^T,$$

and a reproducing kernel as a $(K_1K_2) \times (K_1K_2)$ matrix. Using matrix notation, the products of the marginal reproducing kernels $R_{0\langle\gamma\rangle}$ and $R_{1\langle\gamma\rangle}$ given above and the subspaces they correspond to are listed in Table 2.1, where $\mathbf{1}_K$ is of length K, I_K is of size $K \times K$, and, as a matrix operator, $\otimes$ denotes the Kronecker product of matrices. The corresponding inner products are defined by the Moore-Penrose inverses of these matrices, which

TABLE 2.1. Product reproducing kernels on $\{1,\ldots,K_1\}\times\{1,\ldots,K_2\}$.

Subspace	Reproducing kernel
$\mathcal{H}_{0\langle 1\rangle}\otimes\mathcal{H}_{0\langle 2\rangle}$	$(\mathbf{1}_{K_1}\mathbf{1}_{K_1}^T/K_1)\otimes(\mathbf{1}_{K_2}\mathbf{1}_{K_2}^T/K_2)$
$\mathcal{H}_{0\langle 1\rangle}\otimes\mathcal{H}_{1\langle 2\rangle}$	$(\mathbf{1}_{K_1}\mathbf{1}_{K_1}^T/K_1)\otimes(I_{K_2}-\mathbf{1}_{K_2}\mathbf{1}_{K_2}^T/K_2)$
$\mathcal{H}_{1\langle 1\rangle}\otimes\mathcal{H}_{0\langle 2\rangle}$	$(I_{K_1}-\mathbf{1}_{K_1}\mathbf{1}_{K_1}^T/K_1)\otimes(\mathbf{1}_{K_2}\mathbf{1}_{K_2}^T/K_2)$
$\mathcal{H}_{1\langle 1\rangle}\otimes\mathcal{H}_{1\langle 2\rangle}$	$(I_{K_1}-\mathbf{1}_{K_1}\mathbf{1}_{K_1}^T/K_1)\otimes(I_{K_2}-\mathbf{1}_{K_2}\mathbf{1}_{K_2}^T/K_2)$

are themselves because they are idempotent. The decomposition of (2.28) is seen to be

$$\begin{aligned}\mathcal{H} &= (\mathcal{H}_{0\langle 1\rangle}\oplus\mathcal{H}_{1\langle 1\rangle})\otimes(\mathcal{H}_{0\langle 2\rangle}\oplus\mathcal{H}_{1\langle 2\rangle})\\ &= (\mathcal{H}_{0\langle 1\rangle}\otimes\mathcal{H}_{0\langle 2\rangle})\oplus(\mathcal{H}_{1\langle 1\rangle}\otimes\mathcal{H}_{0\langle 2\rangle})\\ &\quad\oplus(\mathcal{H}_{0\langle 1\rangle}\otimes\mathcal{H}_{1\langle 2\rangle})\oplus(\mathcal{H}_{1\langle 1\rangle}\otimes\mathcal{H}_{1\langle 2\rangle})\\ &= \mathcal{H}_{\{\}}\oplus\mathcal{H}_{\{1\}}\oplus\mathcal{H}_{\{2\}}\oplus\mathcal{H}_{\{1,2\}},\end{aligned}\tag{2.29}$$

where $\mathcal{H}_{\{\}}$ spans the constant, $\mathcal{H}_{\{1\}}$ spans the $x_{\langle 1\rangle}$ main effect, $\mathcal{H}_{\{2\}}$ spans the $x_{\langle 2\rangle}$ main effect, and $\mathcal{H}_{\{1,2\}}$ spans the interaction.

If one would like to use the averaging operator $Af=f(1)$ on a marginal domain $\{1,\ldots,K\}$, the K-dimensional vector space may be decomposed alternatively as

$$\mathcal{H}_0\oplus\mathcal{H}_1=\{f:f(1)=\cdots=f(K)\}\oplus\{f:f(1)=0\},$$

with the reproducing kernels given by $R_0=1$ and $R_1(x,y)=I_{[x=y\neq 1]}$; see Problem 2.8.

2.4.3 Reproducing Kernel Hilbert Spaces on $[0,1]^2$

Set $Af=\int_0^1 f dx$ on $[0,1]$. The tensor product reproducing kernel Hilbert spaces on $[0,1]^2$ can be constructed using the reproducing kernels (2.19) and (2.23) derived in §2.3.3.

Example 2.4 (Tensor product linear spline) Setting $m=1$ in §2.3.3, one has

$$\begin{aligned}\{f:\dot{f}\in\mathcal{L}_2[0,1]\} &= \{f:f\propto 1\}\oplus\{f:\textstyle\int_0^1 f dx=0,\dot{f}\in\mathcal{L}_2[0,1]\}\\ &= \mathcal{H}_0\oplus\mathcal{H}_1,\end{aligned}$$

with reproducing kernels $R_0(x,y)=1$ and $R_1(x,y)=k_1(x)k_1(y)+k_2(x-y)$. This marginal space can be used on both axes to construct a tensor product reproducing kernel Hilbert space with the structure of (2.28), with averaging operators $A_\gamma f=\int_0^1 f dx_{\langle\gamma\rangle}$, $\gamma=1,2$. The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.2. □

TABLE 2.2. Reproducing Kernels and Inner Products in Example 2.4.

Subspace	Reproducing Kernel	Inner Product
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{0\langle 2\rangle}$	1	$(\int_0^1 \int_0^1 f)(\int_0^1 \int_0^1 g)$
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + k_2(x_{\langle 2\rangle} - y_{\langle 2\rangle})$	$\int_0^1(\int_0^1 \dot{f}_{\langle 2\rangle}dx_{\langle 1\rangle})(\int_0^1 \dot{g}_{\langle 2\rangle}dx_{\langle 1\rangle})dx_{\langle 2\rangle}$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{0\langle 2\rangle}$	$k_1(x_{\langle 1\rangle})k_1(y_{\langle 1\rangle}) + k_2(x_{\langle 1\rangle} - y_{\langle 1\rangle})$	$\int_0^1(\int_0^1 \dot{f}_{\langle 1\rangle}dx_{\langle 2\rangle})(\int_0^1 \dot{g}_{\langle 1\rangle}dx_{\langle 2\rangle})dx_{\langle 1\rangle}$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$\big[k_1(x_{\langle 1\rangle})k_1(y_{\langle 1\rangle}) + k_2(x_{\langle 1\rangle} - y_{\langle 1\rangle})\big]\big[k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + k_2(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]$	$\int_0^1 \int_0^1 \ddot{f}_{\langle 12\rangle}\ddot{g}_{\langle 12\rangle}$

TABLE 2.3. Reproducing Kernels and Inner Products in Example 2.5.

Subspace	Reproducing Kernel	Inner Product
$\mathcal{H}_{00\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$	1	$(\int_0^1 \int_0^1 f)(\int_0^1 \int_0^1 g)$
$\mathcal{H}_{01\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$	$k_1(x_{\langle 1\rangle})k_1(y_{\langle 1\rangle})$	$(\int_0^1 \int_0^1 \dot{f}_{\langle 1\rangle})(\int_0^1 \int_0^1 \dot{g}_{\langle 1\rangle})$
$\mathcal{H}_{01\langle 1\rangle} \otimes \mathcal{H}_{01\langle 2\rangle}$	$k_1(x_{\langle 1\rangle})k_1(y_{\langle 1\rangle})k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle})$	$(\int_0^1 \int_0^1 \ddot{f}_{\langle 12\rangle})(\int_0^1 \int_0^1 \ddot{g}_{\langle 12\rangle})$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$	$k_2(x_{\langle 1\rangle})k_2(y_{\langle 1\rangle}) - k_4(x_{\langle 1\rangle} - y_{\langle 1\rangle})$	$\int_0^1(\int_0^1 \ddot{f}_{\langle 11\rangle}dx_{\langle 2\rangle})(\int_0^1 \ddot{g}_{\langle 11\rangle}dx_{\langle 2\rangle})dx_{\langle 1\rangle}$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{01\langle 2\rangle}$	$\big[k_2(x_{\langle 1\rangle})k_2(y_{\langle 1\rangle}) - k_4(x_{\langle 1\rangle} - y_{\langle 1\rangle})\big]k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle})$	$\int_0^1(\int_0^1 f^{(3)}_{\langle 112\rangle}dx_{\langle 2\rangle})(\int_0^1 g^{(3)}_{\langle 112\rangle}dx_{\langle 2\rangle})dx_{\langle 1\rangle}$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$\big[k_2(x_{\langle 1\rangle})k_2(y_{\langle 1\rangle}) - k_4(x_{\langle 1\rangle} - y_{\langle 1\rangle})\big]\big[k_2(x_{\langle 2\rangle})k_2(y_{\langle 2\rangle}) - k_4(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]$	$\int_0^1 \int_0^1 f^{(4)}_{\langle 1122\rangle}g^{(4)}_{\langle 1122\rangle}$

Example 2.5 (Tensor product cubic spline) Setting $m = 2$ in §2.3.3, one has

$$\begin{aligned}\{f : \ddot{f} \in \mathcal{L}_2[0,1]\} &= \{f : f \propto 1\} \oplus \{f : f \propto k_1\} \\ &\quad \oplus \{f : \textstyle\int_0^1 f dx = \int_0^1 \dot{f} dx = 0, \ddot{f} \in \mathcal{L}_2[0,1]\} \\ &= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1,\end{aligned}$$

where $\mathcal{H}_{01} \oplus \mathcal{H}_1$ forms the contrast in a one-way ANOVA decomposition with an averaging operator $Af = \int_0^1 f dx$. The corresponding reproducing kernels are $R_{00}(x,y) = 1$, $R_{01}(x,y) = k_1(x)k_1(y)$, and $R_1(x,y) = k_2(x)k_2(y) - k_4(x-y)$. Note that $\int_0^1 R_{01}(x,y)dy = \int_0^1 R_1(x,y)dy = 0$, $\forall x \in [0,1]$. Using this space on both marginal domains, one can construct a tensor product space with nine tensor sum terms. The subspace $\mathcal{H}_{00\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$ spans the constant term in (1.7) on page 7, the subspaces $\mathcal{H}_{00\langle 1\rangle} \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ and $(\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle}) \otimes \mathcal{H}_{00\langle 2\rangle}$ span the main effects, and the subspace $(\mathcal{H}_{01\langle 1\rangle} \oplus \mathcal{H}_{1\langle 1\rangle}) \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ spans the interaction. The reproducing kernels and the corresponding inner products in some of the subspaces are listed in Table 2.3. The separation of $\mathcal{H}_{01}$ and $\mathcal{H}_1$ is intended to facilitate adequate numerical treatment of the different components; it is not needed for the characterization of the ANOVA decomposition in (2.28). □

For the averaging operator $Af = f(0)$, similar tensor product reproducing kernel Hilbert spaces can be constructed using the marginal spaces described in §2.3.1; details are to be worked out in Problem 2.13. Note that it is not necessary to use the same marginal space on both axes. Actually, the choice of the order m and that of the averaging operator Af on different axes are unrelated to each other. Although the reproducing kernels of §§2.3.1 and 2.3.3 lead to identical polynomial smoothing splines for univariate smoothing on $[0,1]$, they do yield different tensor product smoothing splines on $[0,1]^2$, as their respective roughness penalties are different.

2.4.4 Reproducing Kernel Hilbert Spaces on $\{1,\ldots,K\} \times [0,1]$

Setting $A_1 f = \sum_{x\langle 1\rangle=1}^{K} f(x)/K$ on $\mathcal{X}_1 = \{1,\ldots,K\}$ and $A_2 f = \int_0^1 f dx_{\langle 2\rangle}$ on $\mathcal{X}_2 = [0,1]$, tensor product spaces with the structure of (2.28) built in can be constructed using the marginal spaces used in §§2.4.2 and 2.4.3.

Example 2.6 One construction of a tensor product space is by using $R_{0\langle 1\rangle}(x_{\langle 1\rangle}, y_{\langle 1\rangle}) = 1/K$ and $R_{1\langle 1\rangle}(x_{\langle 1\rangle}, y_{\langle 1\rangle}) = I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K$ on $\mathcal{X}_1$ and $R_{0\langle 2\rangle}(x_{\langle 2\rangle}, y_{\langle 2\rangle}) = 1$ and $R_{1\langle 2\rangle}(x_{\langle 2\rangle}, y_{\langle 2\rangle}) = k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + k_2(x_{\langle 2\rangle} - y_{\langle 2\rangle})$ on $\mathcal{X}_2$. The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.4. □

Example 2.7 Using $R_{0\langle 1\rangle} = 1/K$ and $R_{1\langle 1\rangle} = I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K$ on $\mathcal{X}_1$ and $R_{00\langle 2\rangle} = 1$, $R_{01\langle 2\rangle} = k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle})$, and $R_{1\langle 2\rangle} = k_2(x_{\langle 2\rangle})k_2(y_{\langle 2\rangle}) - k_4(x_{\langle 2\rangle} - y_{\langle 2\rangle})$ on $\mathcal{X}_2$, one can construct a tensor product space with six tensor sum terms. The subspace $\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$ spans the constant, $\mathcal{H}_{0\langle 1\rangle} \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ and $\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$ span the main effects, and $\mathcal{H}_{1\langle 1\rangle} \otimes (\mathcal{H}_{01\langle 2\rangle} \oplus \mathcal{H}_{1\langle 2\rangle})$ spans the interaction. The reproducing kernels and the corresponding inner products in the subspaces are listed in Table 2.5. □

2.4.5 Multiple-Term Reproducing Kernel Hilbert Spaces: General Form

The examples of tensor product reproducing kernel Hilbert spaces on product domains presented above all contain multiple tensor sum terms. In general, a multiple-term reproducing kernel Hilbert space can be written as $\mathcal{H} = \oplus_\beta \mathcal{H}_\beta$, where β is a generic index, with subspaces $\mathcal{H}_\beta$ having inner products $(f_\beta, g_\beta)_\beta$ and reproducing kernels R_β, where f_β is the projection of f in $\mathcal{H}_\beta$. It is often convenient to write $(f, g)_\beta$ for $(f_\beta, g_\beta)_\beta$, which can be formally defined as a semi-inner-product in $\mathcal{H}$ satisfying $(f - f_\beta, f - f_\beta)_\beta = 0$.

The subspaces $\mathcal{H}_\beta$ are independent modules, and the within-module metrics implied by the inner products $(f_\beta, g_\beta)_\beta$ are not necessarily comparable between the modules. Allowing for intermodule rescaling of the metrics, an inner product in $\mathcal{H}$ can be specified via

$$J(f, g) = \sum_\beta \theta_\beta^{-1} (f_\beta, g_\beta)_\beta, \tag{2.30}$$

where $\theta_\beta \in (0, \infty)$ are tunable parameters. The reproducing kernel associated with (2.30) is $R_J = \sum_\beta \theta_\beta R_\beta$, as

$$J\big(R_J(x, \cdot), f\big) = \sum_\beta \theta_\beta^{-1} \big(\theta_\beta R_\beta(x, \cdot), f_\beta\big)_\beta = \sum_\beta f_\beta(x) = f(x).$$

When some of the θ_β are set to ∞ in (2.30), $J(f, g)$ defines a semi-inner-product in $\mathcal{H} = \oplus_\beta \mathcal{H}_\beta$. Such a semi-inner-product may be used to specify $J(f) = J(f, f)$ for use in $L(f) + (\lambda/2)J(f)$. Subspaces not contributing to $J(f)$ form the null space of $J(f)$, $\mathcal{N}_J = \big\{f : J(f) = 0\big\}$. Subspaces contributing to $J(f)$ form the space $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$, in which $J(f, g)$ is a full inner product.

Observing $Y_i = \eta(x_i) + \epsilon_i$, where $x_i \in \mathcal{X}$ is a product domain and $\epsilon_i \sim N(0, \sigma^2)$, one may estimate η via the minimization of

$$\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \eta(x_i)\big)^2 + \lambda J(\eta), \tag{2.31}$$

TABLE 2.4. Reproducing kernels and inner products in Example 2.6.

Subspace	Reproducing kernel	Inner product
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{0\langle 2\rangle}$	$1/\mathrm{K}$	$(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 f)(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 g)/K$
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$\big[k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + k_2(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]/K$	$\int_0^1 (\sum_{x_{\langle 1\rangle}=1}^{K} \dot{f}_{\langle 2\rangle})(\sum_{x_{\langle 1\rangle}=1}^{K} \dot{g}_{\langle 2\rangle})/K$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{0\langle 2\rangle}$	$I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K$	$\sum_{x_{\langle 1\rangle}=1}^{K} (\int_0^1 (I - A_1)f)(\int_0^1 (I - A_1)g)$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$(I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K)\big[k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle}) + k_2(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]$	$\int_0^1 \sum_{x_{\langle 1\rangle}=1}^{K} (I - A_1)\dot{f}_{\langle 2\rangle}(I - A_1)\dot{g}_{\langle 2\rangle}$

TABLE 2.5. Reproducing kernels and inner products in Example 2.7.

Subspace	Reproducing kernel	Inner roduct
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$	$1/\mathrm{K}$	$(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 f)(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 g)/K$
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{01\langle 2\rangle}$	$k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle})/K$	$(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 \dot{f}_{\langle 2\rangle})(\sum_{x_{\langle 1\rangle}=1}^{K} \int_0^1 \dot{g}_{\langle 2\rangle})/K$
$\mathcal{H}_{0\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$\big[k_2(x_{\langle 2\rangle})k_2(y_{\langle 2\rangle}) - k_4(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]/K$	$\int_0^1 (\sum_{x_{\langle 1\rangle}=1}^{K} \ddot{f}_{\langle 22\rangle})(\sum_{x_{\langle 1\rangle}=1}^{K} \ddot{g}_{\langle 22\rangle})/K$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{00\langle 2\rangle}$	$I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K$	$\sum_{x_{\langle 1\rangle}=1}^{K} (\int_0^1 (I - A_1)f)(\int_0^1 (I - A_1)g)$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{01\langle 2\rangle}$	$(I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K)k_1(x_{\langle 2\rangle})k_1(y_{\langle 2\rangle})$	$\sum_{x_{\langle 1\rangle}=1}^{K} (\int_0^1 (I - A_1)\dot{f}_{\langle 2\rangle})(\int_0^1 (I - A_1)\dot{g}_{\langle 2\rangle})$
$\mathcal{H}_{1\langle 1\rangle} \otimes \mathcal{H}_{1\langle 2\rangle}$	$(I_{[x_{\langle 1\rangle}=y_{\langle 1\rangle}]} - 1/K)\big[k_2(x_{\langle 2\rangle})k_2(y_{\langle 2\rangle}) + k_4(x_{\langle 2\rangle} - y_{\langle 2\rangle})\big]$	$\int_0^1 \sum_{x_{\langle 1\rangle}=1}^{K} (I - A_1)\ddot{f}_{\langle 22\rangle}(I - A_1)\ddot{g}_{\langle 22\rangle}$

where $J(f) = J(f, f)$ is as given above. The minimizer of (2.31) defines a smoothing spline on $\mathcal{X}$. The computation strategy outlined in §2.3.2 readily applies here, with the subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$ in §2.3.2 replaced by $\mathcal{N}_J$ and $\mathcal{H}_J$, respectively.

When some of the θ_β are set to 0 in $J(f) = J(f, f)$, the corresponding f_β are not allowed in the estimate. One simply eliminates the corresponding $\mathcal{H}_\beta$ from the tensor sum.

Note that for the computation of a smoothing spline, all that one needs are a basis of $\mathcal{N}_J$ and the reproducing kernel R_J associated with $J(f)$ in $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$. In particular, the explicit form of $J(f)$ is *not* needed.

Example 2.8 Consider the construction of Example 2.5 on $\mathcal{X} = [0, 1]^2$. Denote $\mathcal{H}_{\nu,\mu} = \mathcal{H}_{\nu\langle 1\rangle} \otimes \mathcal{H}_{\mu\langle 2\rangle}$, $\nu, \mu = 00, 01, 1$, with inner products $(f, g)_{\nu,\mu}$ and reproducing kernels $R_{\nu,\mu} = R_{\nu\langle 1\rangle} R_{\mu\langle 2\rangle}$. One may set

$$\begin{aligned} J(f, g) = {} & \theta_{1,00}^{-1}(f, g)_{1,00} + \theta_{1,01}^{-1}(f, g)_{1,01} \\ & + \theta_{00,1}^{-1}(f, g)_{00,1} + \theta_{01,1}^{-1}(f, g)_{01,1} + \theta_{1,1}^{-1}(f, g)_{1,1} \end{aligned}$$

and minimize (2.31) in $\mathcal{H} = \oplus_{\nu,\mu} \mathcal{H}_{\nu,\mu}$. The null space of $J(f) = J(f, f)$ is

$$\begin{aligned} \mathcal{N}_J &= \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{01,01} \\ &= \text{span}\{\phi_{00,00}, \phi_{01,00}, \phi_{00,01}, \phi_{01,01}\} \\ &= \text{span}\big\{1, k_1(x_{\langle 1\rangle}), k_1(x_{\langle 2\rangle}), k_1(x_{\langle 1\rangle}) k_1(x_{\langle 2\rangle})\big\}, \end{aligned}$$

where the basis functions $\phi_{\nu,\mu}$ are explicitly specified. The minimizer of (2.31) in $\mathcal{H} = \oplus_{\nu,\mu} \mathcal{H}_{\nu,\mu}$ has an expression

$$\eta(x) = \sum_{\nu,\mu=00,01} d_{\nu,\mu} \phi_{\nu,\mu}(x) + \sum_{i=1}^{n} c_i R_J(x_i, x),$$

where

$$R_J = \theta_{1,00} R_{1,00} + \theta_{1,01} R_{1,01} + \theta_{00,1} R_{00,1} + \theta_{01,1} R_{01,1} + \theta_{1,1} R_{1,1}.$$

The projections of η in $\mathcal{H}_{\nu,\mu}$ are readily available from the expression. For example, $\eta_{01,00} = d_{01,00} \phi_{01,00}(x)$ and $\eta_{01,1} = \sum_{i=1}^{n} c_i \theta_{01,1} R_{01,1}(x_i, x)$.

To fit an additive model, one may set

$$J(f, g) = \theta_{1,00}^{-1}(f, g)_{1,00} + \theta_{00,1}^{-1}(f, g)_{00,1}$$

and minimize (2.31) in $\mathcal{H}_a = \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{1,00} \oplus \mathcal{H}_{00,01} \oplus \mathcal{H}_{00,1}$. The null space is now

$$\mathcal{N}_J = \mathcal{H}_{00,00} \oplus \mathcal{H}_{01,00} \oplus \mathcal{H}_{00,01} = \text{span}\{\phi_{00,00}, \phi_{01,00}, \phi_{00,01}\},$$

and $\mathcal{H}_J = \mathcal{H}_{1,00} \oplus \mathcal{H}_{00,1}$ with a reproducing kernel

$$R_J = \theta_{1,00} R_{1,00} + \theta_{00,1} R_{00,1}.$$

The spaces $\mathcal{H}_{01,01}$, $\mathcal{H}_{1,01}$, $\mathcal{H}_{01,1}$, and $\mathcal{H}_{1,1}$ are eliminated from $\mathcal{H}_a$. □

2.5 Bayes Model

Penalized likelihood estimation in a reproducing kernel Hilbert space $\mathcal{H}$ with the penalty $J(f)$ a square (semi) norm is equivalent to a certain empirical Bayes model with a Gaussian prior. The prior has a diffuse component in the null space $\mathcal{N}_J$ of $J(f)$ and a proper component in $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$ with mean zero and a covariance function proportional to the reproducing kernel R_J in $\mathcal{H}_J$. The Bayes model may also be perceived as a mixed-effect model, with the fixed effects residing in $\mathcal{N}_J$ and the random effects residing in $\mathcal{H}_J$.

We start the discussion with the familiar shrinkage estimates on discrete domains, followed by the polynomial smoothing splines on $[0, 1]$. The calculus is seen to depend only on the null space $\mathcal{N}_J$ of $J(f)$ and the reproducing kernel R_J in its orthogonal complement $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$, hence applies to smoothing splines in general. The general results are noted concerning the general multiple-term smoothing splines of §2.4.5.

2.5.1 Shrinkage Estimates as Bayes Estimates

Consider the classical one-way ANOVA model with independent observations $Y_i \sim N\big(\eta(x_i), \sigma^2\big)$, $i = 1, \ldots, n$, where $x_i \in \big\{1, \ldots, K\big\}$. With a prior $\eta \sim N(0, bI)$, it is easy to see that the posterior mean of η is given by the minimizer of

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \big(Y_i - \eta(x_i)\big)^2 + \frac{1}{b} \sum_{x=1}^{K} \eta^2(x). \tag{2.32}$$

Setting $b = \sigma^2/n\lambda$, (2.32) is equivalent to (2.4) of §2.2.

Now, consider $\eta = \alpha \mathbf{1} + \eta_1$, with independent priors $\alpha \sim N(0, \tau^2)$ for the mean and $\eta_1 \sim N\big(0, b(I - \mathbf{1}\mathbf{1}^T/K)\big)$ for the contrast. Note that $\eta_1^T \mathbf{1} = 0$ almost surely and that $\bar{\eta} = \sum_{x=1}^{K} \eta(x)/K = \alpha$. The posterior mean of η is given by the minimizer of

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} \big(Y_i - \eta(x_i)\big)^2 + \frac{1}{\tau^2} \bar{\eta}^2 + \frac{1}{b} \sum_{x=1}^{K} \big(\eta(x) - \bar{\eta}\big)^2. \tag{2.33}$$

Letting $\tau^2 \to \infty$ and setting $b = \sigma^2/n\lambda$, (2.33) reduces to (2.3) of §2.2. In the limit, α is said to have a diffuse prior. This setting may also be considered as a mixed-effect model, with $\alpha\mathbf{1}$ being the fixed effect and η_1 being the random effect.

Next we look at a two-way ANOVA model on $\{1,\ldots,K_1\}\times\{1,\ldots,K_2\}$ using the notation of §2.4.2. Assume that $\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_{1,2}$ has four independent components, with priors $\eta_\emptyset \sim N(0, b\theta_\emptyset R_\emptyset)$, $\eta_1 \sim N(0, b\theta_1 R_1)$, $\eta_2 \sim N(0, b\theta_2 R_2)$, and $\eta_{1,2} \sim N(0, b\theta_{1,2} R_{1,2})$, where $R_\emptyset = R_{0\langle 1\rangle} R_{0\langle 2\rangle}$, $R_1 = R_{1\langle 1\rangle} R_{0\langle 2\rangle}$, $R_2 = R_{0\langle 1\rangle} R_{1\langle 2\rangle}$, and $R_{1,2} = R_{1\langle 1\rangle} R_{1\langle 2\rangle}$, as given in Table 2.1. Note that R_β's are orthogonal to each other and that an η_β resides in the column space of R_β almost surely. The posterior mean of η is given by the minimizer of

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \frac{1}{b}\sum_{\beta}\theta_\beta^{-1}\eta^T R_\beta^+ \eta. \tag{2.34}$$

Setting $b = \sigma^2/n\lambda$ and $J(f) = \sum_\beta \theta_\beta^{-1} f^T R_\beta^+ f$, (2.34) reduces to (2.31) of §2.4.5, which defines a bivariate smoothing spline on a discrete product domain. A $\theta_\beta = \infty$ in $J(f)$ puts η_β in $\mathcal{N}_J$, which is equivalent to a diffuse prior, or a fixed effect in a mixed-effect model. To obtain the additive model, one simply eliminates $\eta_{1,2}$ by setting $\theta_{1,2} = 0$.

2.5.2 Polynomial Smoothing Splines as Bayes Estimates

Consider $\eta = \eta_0 + \eta_1$ on $[0,1]$, with η_0 and η_1 having independent Gaussian priors with mean zero and covariance functions,

$$\begin{aligned} E\big[\eta_0(x)\eta_0(y)\big] &= \tau^2 R_0(x,y) = \tau^2 \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!}, \\ E\big[\eta_1(x)\eta_1(y)\big] &= bR_1(x,y) = b\int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du, \end{aligned}$$

where R_0 and R_1 are taken from (2.9) and (2.10) of §2.3.1. Observing $Y_i \sim N\big(\eta(x_i), \sigma^2\big)$, the joint distribution of $\mathbf{Y}$ and $\eta(x)$ is normal with mean zero and a covariance matrix

$$\begin{pmatrix} bQ + \tau^2 SS^T + \sigma^2 I & b\boldsymbol{\xi} + \tau^2 S\boldsymbol{\phi} \\ b\boldsymbol{\xi}^T + \tau^2\boldsymbol{\phi}^T S^T & bR_1(x,x) + \tau^2\boldsymbol{\phi}^T\boldsymbol{\phi} \end{pmatrix}, \tag{2.35}$$

where Q is $n\times n$ with the (i,j)th entry $R_1(x_i,x_j)$, S is $n\times m$ with the (i,ν)th entry $x_i^{\nu-1}/(\nu-1)!$, $\boldsymbol{\xi}$ is $n\times 1$ with the ith entry $R_1(x_i,x)$, and $\boldsymbol{\phi}$ is $m\times 1$ with the νth entry $x^{\nu-1}/(\nu-1)!$. Using a standard result on multivariate normal distribution (see, e.g., Johnson and Wichern (1992, Result 4.6)), the posterior mean of $\eta(x)$ is seen to be

$$
\begin{aligned}
E\big[\eta(x)|\mathbf{Y}\big] &= (b\boldsymbol{\xi}^T + \tau^2\boldsymbol{\phi}^T S^T)(bQ + \tau^2 SS^T + \sigma^2 I)^{-1}\mathbf{Y} \\
&= \boldsymbol{\xi}^T(Q + \rho SS^T + n\lambda I)^{-1}\mathbf{Y} \\
&\quad + \boldsymbol{\phi}^T\rho S^T(Q + \rho SS^T + n\lambda I)^{-1}\mathbf{Y}, \qquad (2.36)
\end{aligned}
$$

where $\rho = \tau^2/b$ and $n\lambda = \sigma^2/b$.

Lemma 2.7 *Suppose M is symmetric and nonsingular and S is of full column rank.*

$$
\lim_{\rho\to\infty}(\rho SS^T + M)^{-1} = M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1}, \qquad (2.37)
$$

$$
\lim_{\rho\to\infty}\rho S^T(\rho SS^T + M)^{-1} = (S^T M^{-1}S)^{-1}S^T M^{-1}. \qquad (2.38)
$$

Proof: It can be verified that (Problem 2.17)

$$
\begin{aligned}
&(\rho SS^T + M)^{-1} = \\
&\quad M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}(I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}S^T M^{-1}. \qquad (2.39)
\end{aligned}
$$

Equation (2.37) follows trivially from (2.39). Substituting (2.39) into the left-hand side of (2.38), some algebra leads to

$$
\begin{aligned}
\rho S^T(\rho SS^T + M)^{-1} &= \rho(I - (I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1})S^T M^{-1} \\
&= (S^T M^{-1}S)^{-1}(I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}S^T M^{-1}.
\end{aligned}
$$

Letting $\rho \to \infty$ yields (2.38). □

Setting $\rho \to \infty$ in (2.36) and applying Lemma 2.7, the posterior mean $E\big[\eta(x)|\mathbf{Y}\big]$ is of the form $\boldsymbol{\xi}^T\mathbf{c} + \boldsymbol{\phi}^T\mathbf{d}$, with the coefficients given by

$$
\begin{aligned}
\mathbf{c} &= (M^{-1} - M^{-1}S(S^T M^{-1}S)^{-1}S^T M^{-1})\mathbf{Y}, \\
\mathbf{d} &= (S^T M^{-1}S)^{-1}S^T M^{-1}\mathbf{Y},
\end{aligned} \qquad (2.40)
$$

where $M = Q + n\lambda I$.

Theorem 2.8 *The polynomial smoothing spline of (2.5) is the posterior mean of $\eta = \eta_0 + \eta_1$, where η_0 diffuses in span$\{x^{\nu-1}, \nu = 1, \ldots, m\}$ and η_1 has a Gaussian process prior with mean zero and a covariance function*

$$
bR_1(x, y) = b\int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du,
$$

for $b = \sigma^2/n\lambda$.

Proof: The only thing that remains to be verified is that $\mathbf{c}$ and $\mathbf{d}$ in (2.40) minimize (2.16) on page 36. Differentiating (2.16) with respect to $\mathbf{c}$ and $\mathbf{d}$ and setting the derivatives to 0, one gets

$$\begin{aligned} Q\{(Q+n\lambda I)\mathbf{c}+S\mathbf{d}-\mathbf{Y}\} &= 0, \\ S^T\{Q\mathbf{c}+S\mathbf{d}-\mathbf{Y}\} &= 0. \end{aligned} \tag{2.41}$$

It is easy to verify that $\mathbf{c}$ and $\mathbf{d}$ given in (2.40) satisfy (2.41). □

2.5.3 Smoothing Splines as Bayes Estimates: General Form

Besides the choices of covariance functions R_0 and R_1, nothing is specific to polynomial smoothing splines in the derivation of §2.5.2. In general, consider a reproducing kernel Hilbert space $\mathcal{H} = \oplus_{\beta=0}^{p}\mathcal{H}_\beta$ on a domain $\mathcal{X}$ with an inner product

$$(f,g) = \sum_{\beta=0}^{p}\theta_\beta^{-1}(f,g)_\beta = \sum_{\beta=0}^{p}\theta_\beta^{-1}(f_\beta,g_\beta)_\beta$$

and a reproducing kernel

$$R(x,y) = \sum_{\beta=0}^{p}\theta_\beta R_\beta(x,y),$$

where $(f,g)_\beta$ is an inner product in $\mathcal{H}_\beta$ with a reproducing kernel R_β, f_β is the projection of f in $\mathcal{H}_\beta$, and $\mathcal{H}_0$ is finite dimensional. Observing $Y_i \sim N\big(\eta(x_i),\sigma^2\big)$, a smoothing spline on $\mathcal{X}$ can be defined as the minimizer of the functional

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i-\eta(x_i)\big)^2 + \lambda\sum_{\beta=1}^{p}\theta_\beta^{-1}(\eta,\eta)_\beta \tag{2.42}$$

in $\mathcal{H}$; see also (2.31) of §2.4.5. A smoothing spline thus defined is a Bayes estimate of $\eta = \sum_{\beta=0}^{p}\eta_\beta$, where η_0 has a diffuse prior in $\mathcal{H}_0$ and η_β, $\beta = 1,\ldots,p$, have mean zero Gaussian process priors on $\mathcal{X}$ with covariance functions $E\big[\eta_\beta(x)\eta_\beta(y)\big] = b\theta_\beta R_\beta(x,y)$, independent of each other, where $b = \sigma^2/n\lambda$. Treated as a mixed-effect model, η_0 contains the fixed effects and η_β, $\beta = 1,\ldots,p$, are the random effects.

2.6 Minimization of Penalized Functional

As an optimization object, analytical properties of the penalized likelihood functional $L(f)+(\lambda/2)J(f)$ can be studied under general functional analytical conditions such as the continuity, convexity, and differentiability of $L(f)$ and $J(f)$. Among such properties are the existence of the minimizer and the equivalence of penalized optimization and constrained optimization.

We first show that the penalized likelihood estimate exists as long as the maximum likelihood estimate uniquely exists in the null space $\mathcal{N}_J$ of $J(f)$.

We then prove that the minimization of $L(f) + (\lambda/2)J(f)$ is equivalent to the minimization of $L(f)$ subject to a constraint of the form $J(f) \leq \rho$ for some $\rho \geq 0$, and quantify the relation between ρ and λ.

2.6.1 Existence of Minimizer

A functional $A(f)$ in a linear space $\mathcal{L}$ is said to be **convex** if for $f, g \in \mathcal{L}$, $A\big(\alpha f + (1-\alpha)g\big) \leq \alpha A(f) + (1-\alpha)A(g)$, $\forall \alpha \in (0,1)$; the convexity is strict if the equality holds only for $f = g$.

Theorem 2.9 (Existence) *Suppose $L(f)$ is a continuous and convex functional in a Hilbert space $\mathcal{H}$ and $J(f)$ is a square (semi) norm in $\mathcal{H}$ with a null space $\mathcal{N}_J$, of finite dimension. If $L(f)$ has a unique minimizer in $\mathcal{N}_J$, then $L(f) + (\lambda/2)J(f)$ has a minimizer in $\mathcal{H}$.*

The minus log likelihood $L(f|\text{data})$ in (1.3) is usually convex in f, as will be verified on a case-by-case basis in later chapters. The quadratic functional $J(f)$ is convex; see Problem 2.18. A minimizer of $L(f)$ is unique in $\mathcal{N}_J$ if the convexity is strict in it, which is often the case.

Without loss of generality, one may set $\lambda = 2$ in the theorem. The proof of the theorem builds on the following two lemmas, with $L(f)$ and $J(f)$ in the lemmas being the same as those in Theorem 2.9.

Lemma 2.10 *If a continuous and convex functional $A(f)$ has a unique minimizer in $\mathcal{N}_J$, then it has a minimizer in the cylinder area $C_\rho = \big\{f : f \in \mathcal{H}, J(f) \leq \rho\big\}$, $\forall \rho \in (0, \infty)$.*

Lemma 2.11 *If $L(f) + J(f)$ has a minimizer in $C_\rho = \big\{f : f \in \mathcal{H}, J(f) \leq \rho\big\}$, $\forall \rho \in (0, \infty)$, then it has a minimizer in $\mathcal{H}$.*

The rest of the section are the proofs.

Proof of Lemma 2.10: Let $\|\cdot\|_0$ be the norm in $\mathcal{N}_J$, and f_0 be the unique minimizer of $A(f)$ in $\mathcal{N}_J$. By Theorem 4 of Tapia and Thompson (1978, p. 162), $A(f)$ has a minimizer in a "rectangle"

$$R_{\rho,\gamma} = \big\{f : f \in \mathcal{H}, J(f) \leq \rho, \|f - f_0\|_0 \leq \gamma\big\}.$$

Now, if the lemma is not true (i.e., that $A(f)$ has no minimizer in C_ρ for some ρ), then a minimizer f_γ of $A(f)$ in $R_{\rho,\gamma}$ must satisfy $\|f_\gamma - f_0\|_0 = \gamma$. By the convexity of $A(f)$ and the fact that $A(f_\gamma) \leq A(f_0)$,

$$A\big(\alpha f_\gamma + (1-\alpha)f_0\big) \leq \alpha A(f_\gamma) + (1-\alpha)A(f_0) \leq A(f_0), \tag{2.43}$$

for $\alpha \in (0,1)$. Now, take a sequence $\gamma_i \to \infty$ and set $\alpha_i = \gamma_i^{-1}$, and write $\alpha_i f_{\gamma_i} + (1-\alpha_i)f_0 = f_i^o + f_i^*$, where $f_i^o \in \mathcal{N}_J$ and $f_i^* \in \mathcal{H} \ominus \mathcal{N}_J$. It is

easy to check that $\|f_i^o - f_0\|_0 = 1$ and that $J(f_i^*) \leq \alpha_i^2 \rho$. Since $\mathcal{N}_J$ is finite dimensional, $\{f_i^o\}$ has a convergent subsequence converging to, say, $f_1 \in \mathcal{N}_J$, and $\|f_1 - f_0\|_0 = 1$. It is apparent that $f_i^* \to 0$. By the continuity of $A(f)$ and (2.43), $A(f_1) \leq A(f_0)$, which contradicts the fact that f_0 uniquely minimizes $A(f)$ in $\mathcal{N}_J$. Hence, $\|f_\gamma - f_0\|_0 = \gamma$ cannot hold for all $\gamma \in (0, \infty)$. This completes the proof. □

Proof of Lemma 2.11: Without loss of generality we assume $L(0) = 0$. If the lemma is not true, then a minimizer f_ρ of $L(f) + J(f)$ in C_ρ must fall on the boundary of C_ρ for every ρ (i.e., $J(f_\rho) = \rho$, $\forall \rho \in (0, \infty)$). By the convexity of $L(f)$,

$$L(\alpha f_\rho) \leq \alpha L(f_\rho), \tag{2.44}$$

for $\alpha \in (0, 1)$. By the definition of f_ρ,

$$L(f_\rho) + J(f_\rho) \leq L(\alpha f_\rho) + J(\alpha f_\rho). \tag{2.45}$$

Combining (2.44) and (2.45) and substituting $J(f_\rho) = \rho$, one obtains

$$L(\alpha f_\rho)/\alpha + \rho \leq L(\alpha f_\rho) + \alpha^2 \rho,$$

which, after some algebra, yields

$$L(\alpha f_\rho) \leq -\alpha(1 + \alpha)\rho. \tag{2.46}$$

Now, choose $\alpha = \rho^{-1/2}$. Since $J(\alpha f_\rho) = 1$, (2.46) leads to

$$L(f_1) \leq -(\rho^{1/2} + 1),$$

which is impossible for large enough ρ. This proves the lemma. □

Proof of Theorem 2.9: Applying Lemma 2.10 on $A(f) = L(f) + J(f)$ leads to the condition of Lemma 2.11, and the lemma, in turn, yields the theorem. □

2.6.2 Penalized and Constrained Optimization

For a functional $A(f)$ in a linear space $\mathcal{L}$, define $A_{f,g}(\alpha) = A(f + \alpha g)$ as functions of α real indexed by $f, g \in \mathcal{L}$. If $\dot{A}_{f,g}(0)$ exists and is linear in g, $\forall f, g \in \mathcal{L}$, $A(f)$ is said to be **Fréchet differentiable** in $\mathcal{L}$, and $\dot{A}_{f,g}(0)$ is the **Fréchet derivative** of A at f in the direction of g.

Theorem 2.12 *Suppose $L(f)$ is continuous, convex, and Fréchet differentiable in a Hilbert space $\mathcal{H}$, and $J(f)$ is a square (semi) norm in $\mathcal{H}$. If f^* minimizes $L(f)$ in $C_\rho = \{f : f \in \mathcal{H}, J(f) \leq \rho\}$, then f^* minimizes $L(f) + (\lambda/2)J(f)$ in $\mathcal{H}$, where the Lagrange multiplier relates to ρ via $\lambda = -\rho^{-1}\dot{L}_{f^*, f_1^*}(0) \geq 0$, with f_1^* being the projection of f^* in $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$. Conversely, if f^o minimizes $L(f) + (\lambda/2)J(f)$ in $\mathcal{H}$, where $\lambda > 0$, then f^o minimizes $L(f)$ in $\{f : f \in \mathcal{H}, J(f) \leq J(f^o)\}$.*

The minus log likelihood $L(f|\text{data})$ in (1.3) is usually Fréchet differentiable, as will be verified on a case-by-case basis in later chapters.

Proof of Theorem 2.12: If $J(f^*) < \rho$, then by the convexity of $L(f)$, f^* is a global minimizer of $L(f)$, so the result holds with $\lambda = \dot{L}_{f^*, f_1^*}(0) = 0$.

In general, $J(f^*) = \rho$; thus, f^* minimizes $L(f)$ on the boundary contour $C_\rho^o = \{f : f \in \mathcal{H}, J(f) = \rho\}$. It is easy to verify that $\dot{J}_{f,g}(0) = 2J(f, g)$, where $J(f, g)$ is the (semi) inner product associated with $J(f)$. The space tangent to the contour C_ρ^o at f^* is thus $\mathcal{G} = \{g : J(f^*, g) = J(f_1^*, g) = 0\}$.

Pick an arbitrary $g \in \mathcal{G}$. When $J(g) = 0$, $f^* + \alpha g \in C_\rho^o$. Since

$$0 \le L(f^* + \alpha g) - L(f^*) = \alpha \dot{L}_{f^*, g}(0) + o(\alpha),$$

one has $\dot{L}_{f^*, g}(0) = 0$. When $J(g) \neq 0$, without loss of generality one may scale g so that $J(g) = \rho$; then, $\sqrt{1 - \alpha^2} f^* + \alpha g \in C_\rho^o$. Now, write $\gamma = \left(\sqrt{1 - \alpha^2} - 1\right)/\alpha$. By the linearity of $\dot{L}_{f,g}(0)$ in g, one has

$$\begin{aligned} 0 &\le L\left(\sqrt{1 - \alpha^2} f^* + \alpha g\right) - L(f^*) \\ &= L\left(f^* + \alpha(\gamma f^* + g)\right) - L(f^*) \\ &= \alpha\gamma \dot{L}_{f^*, f^*}(0) + \alpha \dot{L}_{f^*, g}(0) + o(\alpha) \\ &= \alpha \dot{L}_{f^*, g}(0) + o(\alpha), \end{aligned}$$

where $\alpha\gamma = \sqrt{1 - \alpha^2} - 1 = O(\alpha^2) = o(\alpha)$; so, again, $\dot{L}_{f^*, g}(0) = 0$.

It is easy to see that $J(f_1^*) = \rho$ and that $\mathcal{G}^c = \text{span}\{f_1^*\}$. Now, every $f \in \mathcal{H}$ has an unique decomposition $f = \beta f_1^* + g$, with β real and $g \in \mathcal{G}$; hence,

$$\begin{aligned} \dot{L}_{f^*, f}(0) + \frac{\lambda}{2} \dot{J}_{f^*, f}(0) =& \dot{L}_{f^*, \beta f_1^*}(0) + \dot{L}_{f^*, g}(0) + \lambda J(f^*, \beta f_1^* + g) \\ =& \beta \dot{L}_{f^*, f_1^*}(0) + \beta\lambda\rho. \end{aligned} \tag{2.47}$$

With $\lambda = -\rho^{-1} \dot{L}_{f^*, f_1^*}(0)$, (2.47) is annihilated for all $f \in \mathcal{H}$; thus, f^* minimizes $L(f) + (\lambda/2) J(f)$. Finally, note that $L(f^* - \alpha f_1^*) \ge L(f^*)$ for $\alpha \in (0, 1)$, so $\dot{L}_{f^*, f_1^*}(0) \le 0$. The converse is straightforward and is left as an exercise (Problem 2.21). □

2.7 Bibliographic Notes

Section 2.1

The theory of Hilbert space is at the core of many advanced analysis courses. The elementary materials presented in §2.1.1 provide a minimal exposition for our need. An excellent treatment of vector spaces can be found in Rao (1973, Chap. 1). Proofs of the Riesz representation theorem

can be found in many references, of different levels of abstraction; the one given in §2.1.2 was taken from Akhiezer and Glazman (1961). The theory of reproducing kernel Hilbert space was developed by Aronszajn (1950), which remains the primary reference on the subject. The exposition in §2.1.3 is minimally sufficient to serve our need.

Section 2.2

Shrinkage estimates are among basic techniques in classical decision theory and Bayesian statistics; see, e.g., Lehmann and Casella (1998, §5.5). The interpretation of shrinkage estimates as smoothing splines on discrete domains has not appeared elsewhere. Vector spaces are much more familiar to statisticians than reproducing kernel Hilbert spaces, and this section is intended to help the reader to gain further insights into entities in a reproducing kernel Hilbert space.

Section 2.3

The space $\mathcal{C}^{(m)}[0,1]$ with the inner product (2.7) and the representer of evaluation (2.8) derived from the standard Taylor expansion are standard results found in numerical analysis literature; see, e.g., Schumaker (1981, Chap. 8). The reproducing kernel (2.21) of $\mathcal{C}^{(m)}[0,1]$ associated with the inner product (2.17) was derived by Craven and Wahba (1979), and was used more often than (2.8) as marginal kernels in tensor product smoothing splines. Results concerning Bernoulli polynomials can be found in Abramowitz and Stegun (1964, Chap. 23).

The computational strategy outlined in §2.3.2 was derived by Kimeldorf and Wahba (1971) in the setting of Chebyshev splines, of which the polynomial smoothing splines of (2.5) are special cases; see §4.5.2 for Chebyshev splines. For many years, however, the device was not used much in actual numerical computation. The reasons were multifold. First, algorithms based on (2.16) are of order $O(n^3)$, whereas $O(n)$ algorithms exist for polynomial smoothing splines; see §§3.4 and 3.10. Second, portable numerical linear algebra software and powerful desktop computing were not available until much later. Since the late 1980s, generic algorithms and software have been developed based on (2.16) for the computation of smoothing splines, univariate and multivariate alike; see §3.4 for details.

Section 2.4

A comprehensive treatment of tensor product reproducing kernel Hilbert spaces can be found in Aronszajn (1950), where Theorem 2.6 was quoted as a classical result of I. Schur. The proof given here was suggested by Liqing Yan.

The idea of tensor product smoothing splines was conceived by Barry (1986) and Wahba (1986). Dozens of references appeared in the literature since then, among which Chen (1991), Gu and Wahba (1991b, 1993a, 1993b), Gu (1992b, 1995a 1996, 2004), Wahba, Wang, Gu, Klein, and Klein (1995) and Gu and Ma (2011) registered notable innovations in the theory and practice of the tensor product spline technique. The materials of §§2.4.3–2.4.5 are scattered in these references. The materials of §2.4.2, however, had not appeared in the smoothing literature prior to the first edition of this book.

Section 2.5

The Bayes model of polynomial smoothing splines was first observed by Kimeldorf and Wahba (1970a, 1970b). The materials of §§2.5.2 and 2.5.3 are mainly taken from Wahba (1978, 1983). The elementary materials of §2.5.1 in the familiar discrete setting provide insights into the general results. In Bayesian statistics, such models are more specifically referred to as empirical Bayes models; see, e.g., Berger (1985, §4.5).

Section 2.6

The existence of penalized likelihood estimates has been discussed by many authors in various settings; see, e.g., Tapia and Thompson (1978, Chap. 4) and Silverman (1982). The general result of Theorem 2.9 and the elementary proof are taken from Gu and Qiu (1993).

The relation between penalized optimization and constrained optimization in the context of natural polynomial splines was noted by Schoenberg (1964), where $L(f)$ was a least squares functional. The general result of Theorem 2.12 was adapted from the discussion of Gill, Murray, and Wright (1981, §3.4) on constrained nonlinear optimization.

2.8 Problems

Section 2.1

2.1 Prove the Cauchy-Schwarz inequality of (2.1).

2.2 Prove the triangle inequality of (2.2).

2.3 Let $\mathcal{H}$ be a Hilbert space and $\mathcal{G} \subset \mathcal{H}$ a closed linear subspace. For every $f \in \mathcal{H}$, prove that the projection of f in $\mathcal{G}$, $f_{\mathcal{G}} \in \mathcal{G}$, that satisfies

$$\|f - f_{\mathcal{G}}\| = \inf_{g \in \mathcal{G}} \|f - g\|$$

uniquely exists.

(a) Show that there exists a sequence $\{g_n\} \subset \mathcal{G}$ such that

$$\lim_{n\to\infty} \|f - g_n\| = \delta = \inf_{g\in\mathcal{G}} \|f - g\|.$$

(b) Show that

$$\|g_m - g_n\|^2 = 2\|f - g_m\|^2 + 2\|f - g_n\|^2 - 4\|f - \frac{g_m + g_n}{2}\|^2.$$

Since $\lim_{m,n\to\infty} \|f - \frac{g_m+g_n}{2}\| = \delta$, $\{g_n\}$ is a Cauchy sequence.

(c) Show the uniqueness of $f_{\mathcal{G}}$ using the triangle inequality.

2.4 Given Hilbert spaces $\mathcal{H}_0$ and $\mathcal{H}_1$ satisfying $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, prove that the space $\mathcal{H} = \{f : f = f_0 + f_1, f_0 \in \mathcal{H}_0, f_1 \in \mathcal{H}_1\}$ with an inner product $(f, g) = (f_0, g_0)_0 + (f_1, g_1)_1$ is a Hilbert space, where $f = f_0 + f_1$, $g = g_0 + g_1$, $f_0, g_0 \in \mathcal{H}_0$, $f_1, g_1 \in \mathcal{H}_1$, and $(\cdot,\cdot)_0$ and $(\cdot,\cdot)_1$ are the inner products in $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively. Prove that $\mathcal{H}_0$ and $\mathcal{H}_1$ are the orthogonal complements of each other as closed linear subspaces of $\mathcal{H}$.

2.5 The isomorphism between a K-dimensional Hilbert space $\mathcal{H}$ and the Euclidean K-space is outlined in the following steps:

(a) Take any $\phi \in \mathcal{H}^0 = \mathcal{H}$ nonzero, denote $\phi_1 = \phi/\|\phi\|$, and obtain

$$\mathcal{H}^1 = \mathcal{H}^0 \ominus \{f : f = \alpha\phi_1, \alpha \text{ real}\}.$$

Prove that $\mathcal{H}^1$ contains nonzero elements if $K > 1$.

(b) Repeat step (a) for $\mathcal{H}^{i-1}$, $i = 2, \ldots, K$, to obtain ϕ_i and

$$\mathcal{H}^i = \mathcal{H}^{i-1} \ominus \{f : f = \alpha\phi_i, \alpha \text{ real}\}.$$

Prove that $\mathcal{H}^{K-1} = \{f : f = \alpha\phi_K, \alpha \text{ real}\}$, so $\mathcal{H}^K = \{0\}$.

(c) Verify that $(\phi_i, \phi_j) = \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta. The elements ϕ_i, $i = 1, \ldots, K$, are said to form an orthonormal basis of $\mathcal{H}$. For every $f \in \mathcal{H}$, there is a unique representation $f = \sum_{i=1}^K \alpha_i\phi_i$, where α_i are real coefficients.

(d) Prove that the mapping $f \leftrightarrow \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ are the coefficients of f, defines an isomorphism between $\mathcal{H}$ and the Euclidean space.

2.6 Prove that in an Euclidean space, every linear functional is continuous.

2.7 Prove that the reproducing kernel of a Hilbert space, when it exists, is unique.

Section 2.2

2.8 On $\mathcal{X} = \{1, \ldots, K\}$, the constructions of reproducing kernel Hilbert spaces outlined below yield a one-way ANOVA decomposition with an averaging operator $Af = f(1)$.

(a) Verify that the reproducing kernel $R_0 = 1 = \mathbf{1}\mathbf{1}^T$ generates the space $\mathcal{H}_0 = \{f : f(1) = \cdots = f(K)\}$ with an inner product $(f, g)_0 = f^T(\mathbf{1}\mathbf{1}^T/K^2)g$.

(b) Verify that the reproducing kernel $R_1 = I_{[x=y\neq 1]} = (I - e_1 e_1^T)$ generates the space $\mathcal{H}_1 = \{f : f(1) = 0\}$ with an inner product $(f, g)_1 = f^T(I - e_1 e_1^T)g$, where e_1 is the first unit vector.

(c) Note that $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, so $\mathcal{H}_0 \oplus \mathcal{H}_1$ is well defined and has the reproducing kernel $R_0 + R_1$. With the expressions given in (a) and (b), however, one in general has $(f_1, f_1)_0 \neq 0$ for $f_1 \in \mathcal{H}_1$ and $(f_0, f_0)_1 \neq 0$ for $f_0 \in \mathcal{H}_0$. Nevertheless, $f = \mathbf{1}e_1^T f$ for $f \in \mathcal{H}_0$, so one may write $(f, g)_0 = f^T(e_1 e_1^T)g$. Similarly, as $f = (I - \mathbf{1}e_1^T)f$ for $f \in \mathcal{H}_1$, one may write $(f, g)_1 = f^T(I - e_1\mathbf{1}^T)(I - \mathbf{1}e_1^T)g$. Verify the new expressions of $(f, g)_0$ and $(f, g)_1$. Check that with the new expressions, $(f_1, f_1)_0 = 0$, $\forall f_1 \in \mathcal{H}_1$, and that $(f_0, f_0)_1 = 0$, $\forall f_0 \in \mathcal{H}_0$, so the inner product in $\mathcal{H}_0 \oplus \mathcal{H}_1$ can be written as $(f, g) = (f, g)_0 + (f, g)_1$ with the new expressions.

(d) Verify that $(\mathbf{1}\mathbf{1}^T + I - e_1 e_1^T)^{-1} = e_1 e_1^T + (I - e_1\mathbf{1}^T)(I - \mathbf{1}e_1^T)$ (i.e., the reproducing kernel $R_0 + R_1$ and the inner product $(f, g)_0 + (f, g)_1$ are inverses of each other).

Section 2.3

2.9 Consider the function $k_r(x)$ of (2.18).

(a) Prove that the infinite series converges for $r > 1$ on the real line and for $r = 1$ at noninteger points.

(b) Prove that $k_r(x)$ is real-valued.

(c) Prove that $k_1(x) = x - 0.5$ on $x \in (0, 1)$.

2.10 Prove (2.22) through integration by parts, for $m > 1$. Note that k_r, $r > 1$, are periodic with period 1 and that $\int_0^1 f^{(\nu)} dx = 0$, $\nu = 0, \ldots, m-1$.

2.11 Derive the expressions of $k_2(x)$ and $k_4(x)$ on $[0, 1]$ as given in (2.27) by successive integration from $k_1(x) = x - .5$. Note that for $r > 1$, $dk_r/dx = k_{r-1}$ and $k_r(0) = k_r(1)$.

Section 2.4

2.12 On $\mathcal{X} = \{1, \ldots, K_1\} \times \{1, \ldots, K_2\}$, construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

(a) With $A_1 f = f(1, x_{\langle 2 \rangle})$ and $A_2 f = f(x_{\langle 1 \rangle}, 1)$.

(b) With $A_1 f = f(1, x_{\langle 2 \rangle})$ and $A_2 f = \sum_{x_{\langle 2 \rangle}=1}^{K_2} f(x)/K_2$.

2.13 On $\mathcal{X} = [0, 1]^2$, construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

(a) With $A_1 f = f(0, x_{\langle 2 \rangle})$ and $A_2 f = f(x_{\langle 1 \rangle}, 0)$, using (2.9) and (2.10) with $m = 1, 2$.

(b) With $A_1 f = f(0, x_{\langle 2 \rangle})$ and $A_2 f = \int_0^1 f dx_{\langle 2 \rangle}$, using (2.9), (2.10), (2.19) and (2.23), with $m = 1, 2$.

2.14 On $\mathcal{X} = \{1, \ldots, K\} \times [0, 1]$, construct tensor product reproducing kernel Hilbert spaces with the structure of (2.28).

(a) With $A_1 f = f(1, x_{\langle 2 \rangle})$ and $A_2 f = f(x_{\langle 1 \rangle}, 0)$.

(b) With $A_1 f = f(1, x_{\langle 2 \rangle})$ and $A_2 f = \int_0^1 f dx_{\langle 2 \rangle}$.

(c) With $A_1 f = \sum_{x_{\langle 1 \rangle}=1}^{K} f(x)/K$ and $A_2 f = f(x_{\langle 1 \rangle}, 0)$.

2.15 To compute the tensor product smoothing splines of Example 2.8, one may use the strategy outlined in §2.3.2.

(a) Specify the matrices S and Q in (2.16), for both the full model and the additive model.

(b) Decompose the expression of $\eta(x)$ into those of the constant, the main effects, and the interaction.

2.16 In parallel to Example 2.8 and Problem 2.15, work out the corresponding details for the computation of tensor product smoothing splines on $\{1, \ldots, K\} \times [0, 1]$, using the construction of Example 2.7.

Section 2.5

2.17 Verify (2.39).

Section 2.6

2.18 Prove that a quadratic functional $J(f)$ is convex.

2.19 Let $A(f)$ be a strictly convex functional in a Hilbert space $\mathcal{H}$. Prove that if the minimizer of $A(f)$ exists in $\mathcal{H}$, then it is also unique.

2.20 Consider a strictly convex continuous function $f(x)$ on $(-\infty, \infty)^2$. Prove that if $f_1(x_{\langle 1 \rangle}) = f(x_{\langle 1 \rangle}, 0)$ has a minimizer, then $f(x) + x_{\langle 2 \rangle}^2$ has a unique minimizer.

2.21 Prove that if f^o minimizes $L(f) + \lambda J(f)$, where $\lambda > 0$, then f^o minimizes $L(f)$ subject to $J(f) \leq J(f^o)$.

3

Regression with Gaussian-Type Responses

For regression with Gaussian responses, $L(f) + (\lambda/2)J(f)$ reduces to the familiar penalized least squares functional. Among topics of primary interest are the selection of smoothing parameters, the computation of the estimates, the asymptotic convergence of the estimates, and various data analytical tools.

The main focus of this chapter is on the development of generic computational and data analytical tools for the general multiple-term smoothing splines as formulated in §2.4.5. After a brief review of elementary facts in §3.1, we discuss (§3.2) three popular scores for smoothing parameter selection in detail, namely an unbiased estimate of relative loss, the generalized cross-validation, and the restricted maximum likelihood under the Bayes model of §2.5. In §3.3, we derive the Bayesian confidence intervals of Wahba (1983) and briefly discuss their across-the-function coverage property. Generic algorithms implementing these tools are described in §3.4. Minimizers of $L(f) + (\lambda/2)J(f)$ in certain low dimensional function spaces can deliver as efficient statistical performances, and the theory and practice of such approximations are explored in §3.5. Open-source software implementing the modeling tools are illustrated in §3.6. Heuristic diagnostics are introduced in §§3.7 and 3.8 for the identifiability and practical significance of terms in multiple-term models. Real-data examples are presented in §3.9. Also presented (§3.10) are selected fast algorithms for problems admitting structures through alternative formulations, such as the $O(n)$ algorithm for univariate polynomial splines.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_3,

The asymptotic convergence of penalized least squares estimates will be discussed in Chap. 9, along with that of penalized likelihood estimates in other settings.

3.1 Preliminaries

Observing $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, with $\epsilon_i \sim N(0, \sigma^2)$, the minus log likelihood functional $L(f)$ in $L(f) + (\lambda/2)J(f)$ of (1.3) reduces to the least squares functional proportional to $\sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2$. As discussed in §§2.4.5 and 2.5.3, the general form of penalized least squares functional in a reproducing kernel Hilbert space $\mathcal{H} = \oplus_{\beta=0}^{p} \mathcal{H}_\beta$ can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \big(Y_i - \eta(x_i)\big)^2 + \lambda J(\eta), \tag{3.1}$$

where $J(f) = J(f, f) = \sum_{\beta=1}^{p} \theta_\beta^{-1} (f, f)_\beta$ and $(f, g)_\beta$ are inner products in $\mathcal{H}_\beta$ with reproducing kernels $R_\beta(x, y)$. The penalty is seen to be

$$\lambda J(f) = \lambda \sum_{\beta=1}^{p} \theta_\beta^{-1} (f, f)_\beta,$$

with λ and θ_β as smoothing parameters. This is an overparameterization, as what really matter are the ratios λ/θ_β. One may choose to fix one of the θ_β, but we opt to preserve the symmetry and we do want to keep a λ up front. The bilinear form $J(f, g) = \sum_{\beta=1}^{p} \theta_\beta^{-1} (f, g)_\beta$ is an inner product in $\oplus_{\beta=1}^{p} \mathcal{H}_\beta$, with a reproducing kernel $R_J(x, y) = \sum_{\beta=1}^{p} \theta_\beta R_\beta(x, y)$ and a null space $\mathcal{N}_J = \mathcal{H}_0$ of finite dimension, say m. By the arguments of §2.3.2, the minimizer η_λ of (3.1) has an expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i R_J(x_i, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{3.2}$$

where $\{\phi_\nu\}_{\nu=1}^{m}$ is a basis of $\mathcal{N}_J = \mathcal{H}_0$, $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions, and $\mathbf{c}$ and $\mathbf{d}$ are vectors of real coefficients. The estimation then reduces to the minimization of

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T (\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c} \tag{3.3}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$ and Q is $n \times n$ with the (i, j)th entry $R_J(x_i, x_j)$. See also (2.16) on page 36.

The least squares functional $\sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2$ is continuous and convex in $\mathcal{H}$, and when S is of full column rank, the convexity is strict in $\mathcal{N}_J$. Also, (3.1) is strictly convex in $\mathcal{H}$ when S is of full column rank. See Problem 3.1.

By Theorem 2.9, the minimizer η_λ of (3.1) uniquely exists as long as it uniquely exists in $\mathcal{N}_J$, which requires S to be of full column rank. When Q is singular, (3.3) may have multiple solutions for $\mathbf{c}$ and $\mathbf{d}$, all that satisfy (2.41) on page 51. All the solutions, however, yield the same function estimate η_λ through (3.2). For definiteness in the numerical calculation, we shall compute a particular solution of (3.3) by solving the linear system

$$\begin{aligned} (Q + n\lambda I)\mathbf{c} + S\mathbf{d} &= \mathbf{Y}, \\ S^T\mathbf{c} &= 0. \end{aligned} \tag{3.4}$$

It is easy to verify that (3.4) has a unique solution that satisfies (2.41) (Problem 3.2).

Suppose S is of full column rank. Let

$$S = FR^* = (F_1, F_2)\begin{pmatrix} \tilde{R} \\ O \end{pmatrix} = F_1\tilde{R} \tag{3.5}$$

be the QR-decomposition of S with F orthogonal and $\tilde{R}$ upper-triangular; see, e.g., Golub and Van Loan (1989, §5.2) for QR-decomposition. From $S^T\mathbf{c} = 0$, one has $F_1^T\mathbf{c} = 0$, so $\mathbf{c} = F_2F_2^T\mathbf{c}$. Premultiplying the first equation of (3.4) by F_2^T and F_1^T, simple algebra leads to

$$\begin{aligned} \mathbf{c} &= F_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T\mathbf{Y}, \\ \mathbf{d} &= \tilde{R}^{-1}(F_1^T\mathbf{Y} - F_1^TQ\mathbf{c}). \end{aligned} \tag{3.6}$$

Denote the fitted values by $\hat{\mathbf{Y}} = \big(\eta_\lambda(x_1), \ldots, \eta_\lambda(x_n)\big)^T$ and the residuals by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Some algebra yields

$$\begin{aligned} \hat{\mathbf{Y}} &= Q\mathbf{c} + S\mathbf{d} \\ &= (F_1F_1^T + F_2F_2^TQF_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T)\mathbf{Y} \\ &= (I - F_2(I - F_2^TQF_2(F_2^TQF_2 + n\lambda I)^{-1})F_2^T)\mathbf{Y} \\ &= (I - n\lambda F_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T)\mathbf{Y}. \end{aligned}$$

The symmetric matrix

$$A(\lambda) = I - n\lambda F_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T \tag{3.7}$$

is known as the smoothing matrix associated with (3.1), which has all its eigenvalues in the range $[0, 1]$ (Problem 3.3). It is easy to see from (3.4) that $\mathbf{e} = \big(I - A(\lambda)\big)\mathbf{Y} = n\lambda\mathbf{c}$. Using formula (2.40) on page 50 for $\mathbf{c}$ and $\mathbf{d}$, the smoothing matrix can alternatively be written as

$$A(\lambda) = I - n\lambda(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}), \tag{3.8}$$

where $M = Q + n\lambda I$.

When $\epsilon_i \sim N(0, \sigma^2/w_i)$ with w_i known, $L(f)+(\lambda/2)J(f)$ of (1.3) reduces to a penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n} w_i\big(Y_i - \eta(x_i)\big)^2 + \lambda J(\eta). \tag{3.9}$$

The counter part of (3.4) is

$$\begin{aligned}(Q_w + n\lambda I)\mathbf{c}_w + S_w\mathbf{d} &= \mathbf{Y}_w,\\ S_w^T\mathbf{c}_w &= 0,\end{aligned} \tag{3.10}$$

where $Q_w = W^{1/2}QW^{1/2}$, $\mathbf{c}_w = W^{-1/2}\mathbf{c}$, $S_w = W^{1/2}S$, and $\mathbf{Y}_w = W^{1/2}\mathbf{Y}$, for $W = \mathrm{diag}(w_i)$; see Problem 3.4. Write $\hat{\mathbf{Y}}_w = W^{1/2}\hat{\mathbf{Y}} = A_w(\lambda)\mathbf{Y}_w$ and $\mathbf{e}_w = \mathbf{Y}_w - \hat{\mathbf{Y}}_w$; it is easy to see that $\mathbf{e}_w = n\lambda\mathbf{c}_w$ and that

$$A_w(\lambda) = I - n\lambda F_2(F_2^T Q_w F_2 + n\lambda I)^{-1}F_2^T, \tag{3.11}$$

where $F_2^T F_2 = I$ and $F_2^T S_w = 0$. Parallel to (3.8), one also has

$$A_w(\lambda) = I - n\lambda(M_w^{-1} - M_w^{-1}S_w(S_w^T M_w^{-1} S_w)^{-1}S_w^T M_w^{-1}), \tag{3.12}$$

where $M_w = Q_w + n\lambda I$.

Other than the claim that the least squares functional is proportional to the log likelihood, the normality of ϵ_i has not been used so far. Indeed, many of the results to be presented in this chapter only require moment conditions of ϵ_i. This is reflected in the title of the chapter, where we advertise Gaussian-*type* responses instead of strict Gaussian responses.

3.2 Smoothing Parameter Selection

With varying smoothing parameters λ and θ_β, the minimizer η_λ of (3.1) defines a family of possible estimates. In practice, one has to choose some specific estimate from the family, which calls for effective methods for smoothing parameter selection.

We introduce three scores that are in popular use for smoothing parameter selection in the context. The first score, which assumes a known variance σ^2, is an unbiased estimate of a relative loss. The second score, the generalized cross-validation of Craven and Wahba (1979), targets the same loss without assuming a known σ^2. These scores are presented along with their asymptotic justifications. The third score is derived from the Bayes model of §2.5 through restricted maximum likelihood, which is of appeal to some but is not designed to minimize any particular loss. Parallel scores for weighted and replicated data are also presented. The empirical performance of the three methods is illustrated through simple simulations.

To keep the notation simple, we only make the dependence of various entities on the smoothing parameter λ explicit and suppress their dependence on θ_β. The derivations and proofs apply without change to the general case, with both λ and θ_β tunable.

3.2.1 Unbiased Estimate of Relative Loss

As an estimate of η based on data collected from the sampling points x_i, $i = 1, \ldots, n$, the performance of η_λ can be assessed via the loss function

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \big(\eta_\lambda(x_i) - \eta(x_i)\big)^2. \tag{3.13}$$

This is not to be confused with the log likelihood functional $L(f)$, which will not appear again in this chapter except in Problem 3.1. The λ that minimizes $L(\lambda)$ represents the ideal choice one would like to make given the data, and will be referred to as the optimal smoothing parameter.

Write $\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\eta} = \big(\eta(x_1), \ldots, \eta(x_n)\big)^T$. It is easy to verify that

$$\begin{aligned} L(\lambda) &= \frac{1}{n}\big(A(\lambda)\mathbf{Y} - \boldsymbol{\eta}\big)^T \big(A(\lambda)\mathbf{Y} - \boldsymbol{\eta}\big) \\ &= \frac{1}{n}\boldsymbol{\eta}^T \big(I - A(\lambda)\big)^2 \boldsymbol{\eta} - \frac{2}{n}\boldsymbol{\eta}^T \big(I - A(\lambda)\big) A(\lambda)\boldsymbol{\epsilon} + \frac{1}{n}\boldsymbol{\epsilon}^T A^2(\lambda)\boldsymbol{\epsilon}. \end{aligned}$$

Define

$$U(\lambda) = \frac{1}{n}\mathbf{Y}^T \big(I - A(\lambda)\big)^2 \mathbf{Y} + 2\frac{\sigma^2}{n}\mathrm{tr}A(\lambda). \tag{3.14}$$

Simple algebra yields

$$\begin{aligned} U(\lambda) = &\frac{1}{n}\big(A(\lambda)\mathbf{Y} - \boldsymbol{\eta}\big)^T \big(A(\lambda)\mathbf{Y} - \boldsymbol{\eta}\big) + \frac{1}{n}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ &+ \frac{2}{n}\boldsymbol{\eta}^T \big(I - A(\lambda)\big)\boldsymbol{\epsilon} - \frac{2}{n}\big(\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon} - \sigma^2 \mathrm{tr}A(\lambda)\big). \end{aligned}$$

It follows that

$$\begin{aligned} U(\lambda) - L(\lambda) - &n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} \\ &= \frac{2}{n}\boldsymbol{\eta}^T \big(I - A(\lambda)\big)\boldsymbol{\epsilon} - \frac{2}{n}\big(\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon} - \sigma^2 \mathrm{tr}A(\lambda)\big). \end{aligned} \tag{3.15}$$

It is easy to see that $U(\lambda)$ is an unbiased estimate of the relative loss $L(\lambda) + n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$.

Denote the risk function by

$$R(\lambda) = E\big[L(\lambda)\big] = \frac{1}{n}\boldsymbol{\eta}^T \big(I - A(\lambda)\big)^2 \boldsymbol{\eta} + \frac{\sigma^2}{n}\mathrm{tr}A^2(\lambda), \tag{3.16}$$

where the first term represents the "bias" in the estimation and the second term represents the "variance." Under a condition

Condition 3.2.1 $nR(\lambda) \to \infty$ as $n \to \infty$ and $\lambda \to 0$,

one can establish the consistency of $U(\lambda)$. Condition 3.2.1 is a mild one, as one would not expect nonparametric estimation to deliver a parametric convergence rate of $O(n^{-1})$. See §4.2.3 and Chap. 9.

Theorem 3.1 *Assume independent ϵ_i with mean zero, a common variance, and uniformly bounded fourth moments. Under Condition 3.2.1, as $n \to \infty$ and $\lambda \to 0$,*

$$U(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big).$$

Note that $n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ does not depend on λ, so $U(\lambda)$ traces $L(\lambda)$ closely. The theorem falls short of fully justifying the use of $U(\lambda)$, however, as the λ here is deterministic but the minimizers λ_o of $L(\lambda)$ and λ_u of $U(\lambda)$ are stochastic. It was shown by Li (1986), using much more sophisticated machinery, that the result holds uniformly over a set of λ, yielding $L(\lambda_u)/L(\lambda_o) = 1 + o_p(1)$.

Proof of Theorem 3.1: From (3.15), it suffices to show that

$$L(\lambda) - R(\lambda) = o_p\big(R(\lambda)\big), \tag{3.17}$$

$$\frac{1}{n}\boldsymbol{\eta}^T\big(I - A(\lambda)\big)\boldsymbol{\epsilon} = o_p\big(R(\lambda)\big), \tag{3.18}$$

$$\frac{1}{n}\big(\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon} - \sigma^2 \mathrm{tr} A(\lambda)\big) = o_p\big(R(\lambda)\big). \tag{3.19}$$

We will show (3.17), (3.18), and (3.19) only for the case with ϵ_i normal here, leaving the more tedious general case to Problem 3.5. Let $A(\lambda) = PDP^T$ be the eigenvalue decomposition of $A(\lambda)$, where P is orthogonal and D is diagonal with diagonal entries d_i, $i = 1, \ldots, n$. It is seen that the eigenvalues d_i are in the range $[0, 1]$; see Problem 3.3. Write $\tilde{\boldsymbol{\eta}} = P^T\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\epsilon}} = P^T\boldsymbol{\epsilon}$. It follows that

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{(1-d_i)^2\tilde{\eta}_i^2 - 2d_i(1-d_i)\tilde{\eta}_i\tilde{\epsilon}_i + d_i^2\tilde{\epsilon}_i^2\big\},$$

$$R(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{(1-d_i)^2\tilde{\eta}_i^2 + d_i^2\sigma^2\big\}.$$

To see (3.17), note that

$$\mathrm{Var}\big[L(\lambda)\big] = \frac{1}{n^2}\sum_{i=1}^{n}\big\{4d_i^2(1-d_i)^2\tilde{\eta}_i^2\sigma^2 + 2d_i^4\sigma^4\big\} \le \frac{4\sigma^2}{n}R(\lambda) = o\big(R^2(\lambda)\big).$$

Similarly, (3.18) follows from

$$\mathrm{Var}\left[\frac{1}{n}\boldsymbol{\eta}^T(I - A(\lambda))\boldsymbol{\epsilon}\right] = \frac{1}{n^2}\sum_{i=1}^{n}(1-d_i)^2\tilde{\eta}_i^2\sigma^2 = o\big(R^2(\lambda)\big),$$

and (3.19) follows from $E\big[\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}\big] = \sigma^2 \mathrm{tr}A(\lambda)$ and

$$\mathrm{Var}\left[\frac{1}{n}\boldsymbol{\epsilon}^T A(\lambda)\boldsymbol{\epsilon}\right] = \frac{2}{n^2}\sum_{i=1}^{n} d_i^2\sigma^4 = o\big(R^2(\lambda)\big).$$

The proof is thus complete for the case with $\epsilon_i \sim N(0, \sigma^2)$. □

3.2.2 Cross-Validation and Generalized Cross-Validation

To use $U(\lambda)$ as defined in (3.14), one needs to know the sampling variance σ^2, which is impractical in many applications. The problem can be circumvented, however, by using the method of cross-validation.

The method of cross-validation aims at the prediction error at the sampling points. If an independent validation data set were available with $Y_i^* = \eta(x_i) + \epsilon_i^*$, then an intuitive strategy for the selection of λ would be to minimize $n^{-1}\sum_{i=1}^{n}\big(\eta_\lambda(x_i) - Y_i^*\big)^2$. Lacking an independent validation data set, an alternative strategy is to cross-validate, that is, to minimize

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big(\eta_\lambda^{[i]}(x_i) - Y_i\big)^2, \tag{3.20}$$

where $\eta_\lambda^{[k]}$ is the minimizer of the "delete-one" functional

$$\frac{1}{n}\sum_{i\neq k}\big(Y_i - \eta(x_i)\big)^2 + \lambda J(\eta). \tag{3.21}$$

Instead of solving (3.21) n times, one can perform the delete-one operation analytically with the assistance of the following lemma.

Lemma 3.2 *The minimizer $\eta_\lambda^{[k]}$ of the "delete-one" functional (3.21) minimizes the full data functional (3.1) with $\tilde{Y}_k = \eta_\lambda^{[k]}(x_k)$ replacing Y_k.*

Proof: For all $\eta \neq \eta_\lambda^{[k]}$,

$$\begin{aligned}\frac{1}{n}\Big(\big(\tilde{Y}_k - \eta_\lambda^{[k]}(x_k)\big)^2 + \sum_{i\neq k}\big(Y_i - \eta_\lambda^{[k]}(x_i)\big)^2\Big) + \lambda J(\eta_\lambda^{[k]}) \\ = \frac{1}{n}\sum_{i\neq k}\big(Y_i - \eta_\lambda^{[k]}(x_i)\big)^2 + \lambda J(\eta_\lambda^{[k]})\end{aligned}$$

$$< \frac{1}{n}\sum_{i\neq k}\big(Y_i-\eta(x_i)\big)^2+\lambda J(\eta)$$
$$\le \frac{1}{n}\Big(\big(\tilde{Y}_k-\eta(x_k)\big)^2+\sum_{i\neq k}\big(Y_i-\eta(x_i)\big)^2\Big)+\lambda J(\eta).$$

The lemma follows. □

The fitted values $\hat{\mathbf{Y}}=A(\lambda)\mathbf{Y}$ are linear in $\mathbf{Y}$. By Lemma 3.2, it is easy to see that

$$\eta_\lambda(x_i)-\eta_\lambda^{[i]}(x_i)=a_{i,i}\big(Y_i-\eta_\lambda^{[i]}(x_i)\big),$$

where $a_{i,i}$ is the (i,i)th entry of $A(\lambda)$. Solving for $\eta_\lambda^{[i]}(x_i)$, one has

$$\eta_\lambda^{[i]}(x_i)=\frac{\eta_\lambda(x_i)-a_{i,i}Y_i}{1-a_{i,i}}.$$

It then follows that

$$\eta_\lambda^{[i]}(x_i)-Y_i=\frac{\eta_\lambda(x_i)-Y_i}{1-a_{i,i}}.$$

Hence,

$$V_0(\lambda)=\frac{1}{n}\sum_{i=1}^{n}\frac{\big(Y_i-\eta_\lambda(x_i)\big)^2}{(1-a_{i,i})^2}. \tag{3.22}$$

It is rarely the case that all sampling points contribute equally to the estimation of $\eta(x)$. To adjust for such an imbalance, it might pay to consider alternative scores with unequal weights,

$$\tilde{V}(\lambda)=\frac{1}{n}\sum_{i=1}^{n}w_i\frac{\big(Y_i-\eta_\lambda(x_i)\big)^2}{(1-a_{i,i})^2}.$$

With the choice of $w_i=(1-a_{i,i})^2/\big\{n^{-1}\mathrm{tr}\big(I-A(\lambda)\big)\big\}^2$ [i.e., substituting $a_{i,i}$ in (3.22) by its average $n^{-1}\sum_{i=1}^{n}a_{i,i}$], one obtains the generalized cross-validation (GCV) score of Craven and Wahba (1979),

$$V(\lambda)=\frac{n^{-1}\mathbf{Y}^T\big(I-A(\lambda)\big)^2\mathbf{Y}}{\big\{n^{-1}\mathrm{tr}\big(I-A(\lambda)\big)\big\}^2}. \tag{3.23}$$

A desirable property of the GCV score $V(\lambda)$ is its invariance to an orthogonal transform of $\mathbf{Y}$. Under an extra condition

Condition 3.2.2 $\big\{n^{-1}\mathrm{tr}A(\lambda)\big\}^2/n^{-1}\mathrm{tr}A^2(\lambda)\to 0$ as $n\to\infty$ and $\lambda\to 0$,

$V(\lambda)$ can be shown to be a consistent estimate of the relative loss. Condition 3.2.2 generally holds in most settings of interest; see Craven and Wahba (1979) and Li (1986) for details. See also §4.2.3.

Theorem 3.3 *Assume independent ϵ_i with mean zero, a common variance, and uniformly bounded fourth moments. Under Conditions 3.2.1 and 3.2.2, as $n \to \infty$ and $\lambda \to 0$,*

$$V(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big).$$

Similar to Theorem 3.1, this is poor man's justification for the use of $V(\lambda)$. The ultimate justification can be found in Li (1986), where it was shown that $L(\lambda_v)/L(\lambda_o) = 1 + o_p(1)$, with λ_v minimizing $V(\lambda)$.

Proof of Theorem 3.3: Write $\mu = n^{-1}\mathrm{tr}A(\lambda)$ and $\tilde{\sigma}^2 = n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$. Note that $n^{-1}\mathrm{tr}A^2(\lambda) < 1$, so Condition 3.2.2 implies that $\mu \to 0$. Straightforward algebra yields

$$\begin{aligned} V(\lambda) - L(\lambda) - \tilde{\sigma}^2 &= \frac{1}{(1-\mu)^2}\big\{U(\lambda) - 2\sigma^2\mu - \big(L(\lambda) + \tilde{\sigma}^2\big)(1-\mu)^2\big\} \\ &= \frac{U(\lambda) - L(\lambda) - \tilde{\sigma}^2}{(1-\mu)^2} + \frac{(2-\mu)\mu L(\lambda)}{(1-\mu)^2} \\ &\quad - \frac{\mu^2\tilde{\sigma}^2}{(1-\mu)^2} + \frac{2\mu(\tilde{\sigma}^2 - \sigma^2)}{(1-\mu)^2}. \end{aligned}$$

The first term is $o_p\big(L(\lambda)\big)$ by Theorem 3.1. The second term is $o_p\big(L(\lambda)\big)$ since $\mu \to 0$. By Condition 3.2.2, $\mu^2 = o_p\big(L(\lambda)\big)$, so the third term is $o_p\big(L(\lambda)\big)$. Combining this with $\tilde{\sigma}^2 - \sigma^2 = O_p(n^{-1/2}) = o_p\big(L^{1/2}(\lambda)\big)$, one obtains $o_p\big(L(\lambda)\big)$ for the fourth term. □

When the conditions of Theorem 3.3 hold uniformly in a neighborhood of the optimal λ, the minimizers λ_u of $U(\lambda)$ and λ_v of $V(\lambda)$ should be close to each other. Differentiating $U(\lambda)$ and setting the derivative to zero, one gets

$$\frac{d}{d\lambda}\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y} = -2\sigma^2\frac{d}{d\lambda}\mathrm{tr}A(\lambda). \tag{3.24}$$

Differentiating $V(\lambda)$ and setting the derivative to zero, one similarly has

$$\frac{d}{d\lambda}\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y} = -2\frac{\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y}}{\mathrm{tr}\big(I - A(\lambda)\big)}\frac{d}{d\lambda}\mathrm{tr}A(\lambda). \tag{3.25}$$

Setting $\lambda_u = \lambda_v$ by equating (3.24) and (3.25) and solving for σ^2, one obtains a variance estimate

$$\hat{\sigma}_v^2 = \frac{\mathbf{Y}^T\big(I - A(\lambda_v)\big)^2\mathbf{Y}}{\mathrm{tr}\big(I - A(\lambda_v)\big)}. \tag{3.26}$$

The consistency of the variance estimate $\hat{\sigma}_v^2$ is established below.

Theorem 3.4 *If Conditions 3.2.1 and 3.2.2 hold uniformly in a neighborhood of the optimal λ, then the variance estimate $\hat{\sigma}_v^2$ of (3.26) is consistent.*

Proof: By Theorems 3.1 and 3.3 and (3.17),

$$o_p\big(R(\lambda_v)\big) = V(\lambda_v) - U(\lambda_v) = \hat{\sigma}_v^2/(1-\mu) - \hat{\sigma}_v^2(1-\mu) - 2\sigma^2\mu,$$

where $\mu = n^{-1}\mathrm{tr}A(\lambda_v)$, as in the proof of Theorem 3.3. Solving for σ^2, one has

$$\sigma^2 = \hat{\sigma}_v^2 \frac{1-\mu/2}{1-\mu} + o_p\big(\mu^{-1}R(\lambda_v)\big) = \hat{\sigma}_v^2\big(1+o(1)\big) + o_p\big(\mu^{-1}R(\lambda_v)\big).$$

It remains to show that $\mu^{-1}R(\lambda_v) = O(1)$. In the neighborhood of the optimal λ, the "bias" term and the "variance" term of $R(\lambda)$ should be of the same order, so $R(\lambda) = O\big(n^{-1}\mathrm{tr}A^2(\lambda)\big)$. Since the eigenvalues of $A(\lambda)$ are in the range of $[0,1]$, $\mathrm{tr}A^2(\lambda)/\mathrm{tr}A(\lambda) \le 1$. Now,

$$\mu^{-1}R(\lambda) = \mu^{-1}n^{-1}\mathrm{tr}A^2(\lambda)\big\{R(\lambda)/n^{-1}\mathrm{tr}A^2(\lambda)\big\} = O(1).$$

This completes the proof. □

It is easy to see that any estimate of the form $\hat{\sigma}_v^2\big(1+o_p(1)\big)$ is also consistent. The consistency of $\hat{\sigma}_v^2(1+o_p(1))$ may also be obtained directly from Theorem 3.3 and the fact that $L(\lambda) = o_p(1)$.

Despite its asymptotic optimality, the GCV score $V(\lambda)$ of (3.23) is known to occasionally deliver severe undersmoothing. A modified version,

$$V(\lambda) = \frac{n^{-1}\mathbf{Y}^T\big(I-A(\lambda)\big)^2\mathbf{Y}}{\big\{n^{-1}\mathrm{tr}\big(I-\alpha A(\lambda)\big)\big\}^2}, \tag{3.27}$$

with a fudge factor $\alpha > 1$ proves rather effective in curbing undersmoothing while maintaining the otherwise good performance of GCV; $\alpha = 1.4$ was found to be adequate in the simulation studies of Kim and Gu (2004).

3.2.3 Restricted Maximum Likelihood Under Bayes Model

As an alternative to cross-validation, one may select the smoothing parameters in the context via the restricted maximum likelihood (REML) under the Bayes model of §2.5. The method may be of appeal to some, but it is not designed to minimize any specific loss.

Under the Bayes model, one observes $Y_i = \eta(x_i) + \epsilon_i$ with $\epsilon_i \sim N(0,\sigma^2)$ and $\eta(x) = \sum_{\nu=1}^m d_\nu\phi_\nu(x) + \eta_1(x)$, where $\eta_1(x)$ is a mean zero Gaussian process with a covariance function $E\big[\eta_1(x)\eta_1(y)\big] = bR_J(x,y)$. To eliminate the nuisance parameters d_ν, a common practice is to consider the likelihood

of the contrasts $\mathbf{Z} = F_2^T \mathbf{Y}$, where F_2 is as in (3.5) on page 63. The minus log (restricted) likelihood of σ^2 and b based on the restricted data $\mathbf{Z}$ is seen to be

$$\begin{aligned}&\frac{1}{2}\mathbf{Z}^T(bQ^* + \sigma^2 I)^{-1}\mathbf{Z} + \frac{1}{2}\log\left|bQ^* + \sigma^2 I\right| \\ &\quad = \frac{1}{2b}\mathbf{Z}^T(Q^* + n\lambda I)^{-1}\mathbf{Z} + \frac{1}{2}\log\left|Q^* + n\lambda I\right| + \frac{n-m}{2}\log b, \quad (3.28)\end{aligned}$$

where $Q^* = F_2^T Q F_2$ and $n\lambda = \sigma^2/b$; see Problem 3.6. Minimizing (3.28) with respect to b, one gets

$$\hat{b} = \frac{\mathbf{Z}^T(Q^* + n\lambda I)^{-1}\mathbf{Z}}{n-m},$$

with λ to be estimated by the minimizer of the profile minus log likelihood,

$$\frac{1}{2}\log\left|Q^* + n\lambda I\right| + \frac{n-m}{2}\log(\hat{b}). \quad (3.29)$$

From (3.7), one has

$$\mathbf{Z}^T(Q^* + n\lambda I)^{-1}\mathbf{Z} = (n\lambda)^{-1}\mathbf{Y}^T\big(I - A(\lambda)\big)\mathbf{Y}$$

and

$$\left|Q^* + n\lambda I\right| = (n\lambda)^{n-m}\left|I - A(\lambda)\right|_+^{-1},$$

where $|B|_+$ denotes the product of positive eigenvalues of B. With some algebra, a monotone transform of (3.29) gives

$$M(\lambda) = \frac{n^{-1}\mathbf{Y}^T\big(I - A(\lambda)\big)\mathbf{Y}}{\left|I - A(\lambda)\right|_+^{1/(n-m)}}, \quad (3.30)$$

whose minimizer λ_m is called the generalized maximum likelihood (GML) estimate of λ by Wahba (1985). The corresponding variance estimate is then

$$\hat{\sigma}_m^2 = \frac{\mathbf{Y}^T\big(I - A(\lambda_m)\big)\mathbf{Y}}{n-m}. \quad (3.31)$$

As $n \to \infty$, it was shown by Wahba (1985) that $\lambda_m = o_p(\lambda_v)$ for η "super-smooth" (in the sense that η satisfies smoothness conditions more stringent than $J(\eta) < \infty$) and that $\lambda_m \asymp \lambda_v$ otherwise; see §4.2.3. Hence, asymptotically, GML tends to deliver rougher estimates than GCV.

3.2.4 Weighted and Replicated Data

For weighted data with $E[\epsilon_i^2] = \sigma^2/w_i$, it is appropriate to replace the loss function $L(\lambda)$ of (3.13) by its weighted version

$$L_w(\lambda) = \frac{1}{n}\sum_{i=1}^{n} w_i\big(\eta_\lambda(x_i) - \eta(x_i)\big)^2. \tag{3.32}$$

The unbiased estimate of relative loss is now

$$U_w(\lambda) = \frac{1}{n}\mathbf{Y}_w^T\big(I - A_w(\lambda)\big)^2\mathbf{Y}_w + 2\frac{\sigma^2}{n}\mathrm{tr}A_w(\lambda), \tag{3.33}$$

where $\mathbf{Y}_w = W^{1/2}\mathbf{Y}$ for $W = \mathrm{diag}(w_i)$ and $A_w(\lambda)$ is as given in (3.12). The corresponding GCV score is

$$V_w(\lambda) = \frac{n^{-1}\mathbf{Y}_w^T\big(I - A_w(\lambda)\big)^2\mathbf{Y}_w}{\big\{n^{-1}\mathrm{tr}\big(I - A_w(\lambda)\big)\big\}^2}. \tag{3.34}$$

The following theorem establishes the consistency of $U_w(\lambda)$ and $V_w(\lambda)$ as estimates of the relative loss $L_w(\lambda)+n^{-1}\boldsymbol{\epsilon}^T W\boldsymbol{\epsilon}$, with the proof easily adapted from the proofs of Theorems 3.1 and 3.3; see Problem 3.7.

Theorem 3.5 *Suppose the scaled noise $\sqrt{w_i}\epsilon_i$ are independent with mean zero, a common variance σ^2, and uniformly bounded fourth moments. Denote $R_w(\lambda) = E\big[L_w(\lambda)\big]$. If $nR_w(\lambda) \to \infty$ and $\big\{n^{-1}trA_w(\lambda)\big\}^2/n^{-1}tr A_w^2(\lambda) \to 0$ as $n \to \infty$ and $\lambda \to 0$, then*

$$U_w(\lambda) - L_w(\lambda) - n^{-1}\boldsymbol{\epsilon}^T W\boldsymbol{\epsilon} = o_p\big(L_w(\lambda)\big),$$
$$V_w(\lambda) - L_w(\lambda) - n^{-1}\boldsymbol{\epsilon}^T W\boldsymbol{\epsilon} = o_p\big(L_w(\lambda)\big).$$

For the restricted maximum likelihood under the Bayes model, one can start with the contrasts of $\mathbf{Y}_w$ and derive the corresponding GML score

$$M_w(\lambda) = \frac{n^{-1}\mathbf{Y}_w^T\big(I - A_w(\lambda)\big)\mathbf{Y}_w}{\big|I - A_w(\lambda)\big|_+^{1/(n-m)}}. \tag{3.35}$$

Now, suppose one observes replicated data $Y_{i,j} = \eta(x_i) + \epsilon_{i,j}$, where $j = 1, \ldots, w_i$, $i = 1, \ldots, n$, and $\epsilon_{i,j} \sim N(0, \sigma^2)$. The penalized unweighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{w_i}\big(Y_{i,j} - \eta(x_i)\big)^2 + \lambda J(\eta) \tag{3.36}$$

is equivalent to the penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n} w_i\big(\bar{Y}_i - \eta(x_i)\big)^2 + \lambda J(\eta), \tag{3.37}$$

where $\bar{Y}_i = \sum_{j=1}^{w_i} Y_{i,j}/w_i$; see Problem 3.8(a). Let $\tilde{\mathbf{Y}}$ be the response vector in (3.36) of length $N = \sum_{i=1}^{n} w_i$ and $\tilde{A}(\lambda)$ be the corresponding smoothing matrix, and let $\mathbf{Y}_w$ be the weighted response vector in (3.37) of length n with the ith entry $\sqrt{w_i}\bar{Y}_i$ and $A_w(\lambda)$ be the corresponding smoothing matrix as given in (3.11). It can be shown that $\mathbf{Y}_w = W^{-1/2}P^T\tilde{\mathbf{Y}}$ and

$$I - \tilde{A}(\lambda) = PW^{-1/2}\big(I - A_w(\lambda)\big)W^{-1/2}P^T + F_3F_3^T,$$

where $P = \mathrm{diag}(\mathbf{1}_{w_i})$ is of size $N\times n$ and F_3 is orthogonal of size $N\times(N-n)$ satisfying $F_3^TP = O$; see Problem 3.8. It follows that

$$\begin{aligned}\tilde{\mathbf{Y}}^T\big(I_N - \tilde{A}(\lambda)\big)^p\tilde{\mathbf{Y}} &= \mathbf{Y}_w^T\big(I_n - A_w(\lambda)\big)^p\mathbf{Y}_w + (N-n)\tilde{\sigma}^2, \qquad p = 1, 2,\\ \mathrm{tr}\big(I_N - \tilde{A}(\lambda)\big) &= \mathrm{tr}\big(I_n - A_w(\lambda)\big) + (N-n),\end{aligned}$$

where the sizes of the identity matrices are marked by the subscripts N and n and $\tilde{\sigma}^2 = \sum_{i=1}^{n}\sum_{j=1}^{w_i}(Y_{i,j} - \bar{Y}_i)^2/(N-n)$. It is easy to see that $\mathrm{tr}\tilde{A}(\lambda) = \mathrm{tr}A_w(\lambda)$ and $\big|I_N - \tilde{A}(\lambda)\big|_+ = \big|I_n - A_w(\lambda)\big|_+$. Hence, the $U(\lambda)$, $V(\lambda)$, and $M(\lambda)$ scores associated with (3.36) can be expressed in terms of $\mathbf{Y}_w$ and $A_w(\lambda)$ as

$$U(\lambda) = \frac{1}{N}\mathbf{Y}_w^T\big(I_n - A_w(\lambda)\big)^2\mathbf{Y}_w + 2\frac{\sigma^2}{N}\mathrm{tr}A_w(\lambda) + \frac{N-n}{N}\tilde{\sigma}^2, \tag{3.38}$$

$$V(\lambda) = \frac{N^{-1}\big\{\mathbf{Y}_w^T\big(I_n - A_w(\lambda)\big)^2\mathbf{Y}_w + (N-n)\tilde{\sigma}^2\big\}}{\big\{1 - N^{-1}\mathrm{tr}A_w(\lambda)\big\}^2}, \tag{3.39}$$

$$M(\lambda) = \frac{N^{-1}\big\{\mathbf{Y}_w^T\big(I_n - A_w(\lambda)\big)\mathbf{Y}_w + (N-n)\tilde{\sigma}^2\big\}}{\big|I_n - A_w(\lambda)\big|_+^{1/(N-m)}}. \tag{3.40}$$

It is clear that $U(\lambda)$ of (3.38) is equivalent to $U_w(\lambda)$ of (3.33), but $V(\lambda)$ of (3.39) and $V_w(\lambda)$ of (3.34) are different, so are $M(\lambda)$ of (3.40) and $M_w(\lambda)$ of (3.35). Note that the information concerning σ^2 contained in $\tilde{\sigma}^2$ is ignored in $V_w(\lambda)$ and $M_w(\lambda)$.

The numerical treatment through (3.4) on page 63 is immune to possible singularity of Q, so one usually can ignore the presence of replicated data. When n is substantially smaller than N, however, the computation via (3.37) can result in substantial savings; see §3.4 for the cost of computation. Also, a fast algorithm for the computation of L-splines of §4.5 assumes distinctive x_i's; see §4.5.5.

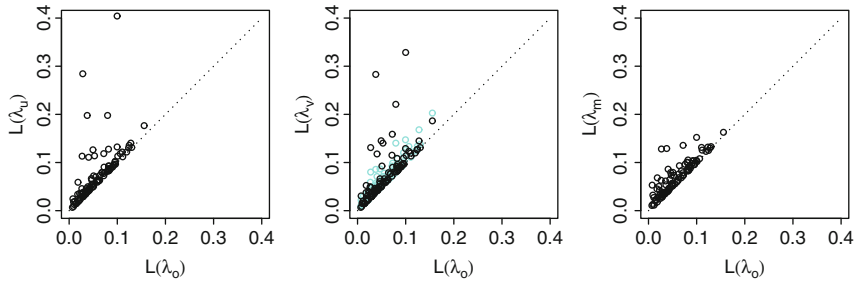


FIGURE 3.1. Performance of $U(\lambda)$, $V(\lambda)$, and $M(\lambda)$ in simulation: $n = 100$. *Left*: Loss achieved by $U(\lambda)$ of (3.14). *Center*: Loss achieved by $V(\lambda)$ of (3.27) with $\alpha = 1$ (*solid*) and $\alpha = 1.4$ (*faded*). *Right*: Loss achieved by $M(\lambda)$ of (3.30).

3.2.5 Empirical Performance

We now illustrate the practical performance of the methods discussed above through some simple simulation. One hundred replicates of samples of size $n = 100$ were generated from $Y_i = \eta(x_i) + \epsilon_i$, $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$, where

$$\eta(x) = 1 + 3\sin(2\pi x - \pi)$$

and $\epsilon_i \sim N(0, 1)$. Cubic smoothing splines were calculated with λ minimizing $U(\lambda)$, $V(\lambda)$, and $M(\lambda)$, and with λ on the grid $\log_{10} n\lambda = (-6)(0.1)(0)$. The mean square error $L(\lambda) = n^{-1} \sum_{i=1}^{n} \big(\eta_\lambda(x_i) - \eta(x_i)\big)^2$ was calculated for all the estimates, from which the optimal λ_o was located. The losses $L(\lambda_u)$, $L(\lambda_v)$, and $L(\lambda_m)$ are plotted against $L(\lambda_o)$ for all the replicates in Fig. 3.1, where a point on the dotted line indicates a perfect selection by the empirical method. All of the methods appeared to perform well most of the time, with occasional wild failures found in $L(\lambda_u)$ and $L(\lambda_v)$ but not in $L(\lambda_m)$. The modified GCV score $V(\lambda)$ of (3.27) was also minimized on the grid, for $\alpha = 1.4$, with the resulting $L(\lambda_v)$ superimposed in the center frame of Fig. 3.1 in faded circles; the wild failures of the unmodified $V(\lambda)$ were effectively curtailed by the fudge factor $\alpha = 1.4$.

To empirically investigate the asymptotic behavior of $V(\lambda)$ versus that of $M(\lambda)$, part of the simulation was repeated for sample sizes $n = 200$ and $n = 500$, each with one hundred replicates. Plotted in Fig. 3.2 are the relative efficacy $L(\lambda_m)/L(\lambda_v)$ of λ_v over λ_m, the comparison of the magnitudes of λ_v versus λ_m, and the performance of the variance estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_m^2$; results for unmodified $V(\lambda)$ are in solid and those with $\alpha = 1.4$ are in faded, and the two sets of $\hat{\sigma}_v^2$ were numerically duplicates of each other. It appeared that $L(\lambda_v)$ came ahead of $L(\lambda_m)$ more often than the other way around, and the frequency of such increased as n increased. The magnitude of λ_m indeed came below that of λ_v in general, as predicted by the asymptotic analysis of Wahba (1985), but λ_v from the unmodified $V(\lambda)$ was severely undersmoothing in a few cases, which actually were responsible

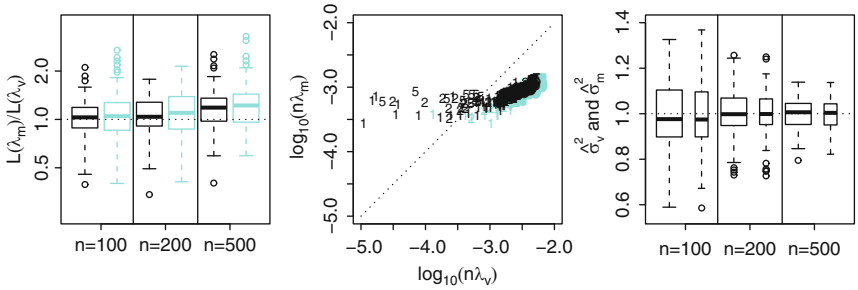


FIGURE 3.2. Comparison of $V(\lambda)$ versus $M(\lambda)$ in simulation. Results for $\alpha = 1$ in (3.27) are in *solid* and those for $\alpha = 1.4$ in *faded*. *Center*: Symbols "1," "2," and "5" indicate replicates with $n = 100$, 200, and 500, respectively. *Right*: $\hat{\sigma}_v^2$ are in *wider boxes*, $\hat{\sigma}_m^2$ are in *thinner boxes*, $\sigma^2 = 1$.

for its occasional wild failures seen in Fig. 3.1. The performances of the variance estimates were reasonably good and did improve as n increased. The variance estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_m^2$ were actually within 1.5 % of each other in all but eight $n = 100$ replicates, three $n = 200$ replicates, and two $n = 500$ replicates.

3.3 Bayesian Confidence Intervals

Point estimate alone is often insufficient in practical applications, as it lacks an assessment of the estimation precision. Lacking parametric sampling distributions, however, an adequately justified interval estimate is a rarity in nonparametric function estimation. An exception to this is the Bayesian confidence intervals of Wahba (1983), which are derived from the Bayes model of §2.5.

We derive the posterior mean and the posterior variance of $\eta(x)$ and those of its components under the Bayes model, which form the basis for the construction of the interval estimates. The posterior variance permits a somewhat simpler expression on the sampling points, which we will also explore. Despite their derivation from the Bayes model, the interval estimates demonstrate a certain across-the-function coverage property for η fixed and smooth, which makes them comparable to the standard parametric confidence intervals. The practical performance of the interval estimates is illustrated through simple simulation. Parallel results for weighted data are also briefly noted.

3.3.1 Posterior Distribution

Consider $\eta = \eta_0 + \eta_1$, where η_0 and η_1 have independent mean zero Gaussian process priors with covariances $E\big[\eta_0(x)\eta_0(y)\big] = \tau^2 \sum_{\nu=1}^{m} \phi_\nu(x)\phi_\nu(y)$ and $E\big[\eta_1(x)\eta_1(y)\big] = bR_J(x,y)$, respectively. From (2.35) on page 49 and a standard result on multivariate normal distribution (see, e.g., Johnson and Wichern (1992, Result 4.6)), the conditional variance of $\eta(x)$ given $Y_i = \eta(x_i) + \epsilon_i$ is seen to be

$$\begin{aligned} &bR_J(x,x) + \tau^2 \boldsymbol{\phi}^T \boldsymbol{\phi} - (b\boldsymbol{\xi}^T + \tau^2 \boldsymbol{\phi}^T S^T) \\ &\qquad \times (bQ + \tau^2 SS^T + \sigma^2 I)^{-1} (b\boldsymbol{\xi} + \tau^2 S\boldsymbol{\phi}) \\ &= b\big\{R_J(x,x) + \rho \boldsymbol{\phi}^T \boldsymbol{\phi} \\ &\quad - (\boldsymbol{\xi}^T + \rho \boldsymbol{\phi}^T S^T)(Q + \rho SS^T + n\lambda I)^{-1} (\boldsymbol{\xi} + \rho S\boldsymbol{\phi})\big\} \\ &= b\big\{R_J(x,x) + \boldsymbol{\phi}^T (\rho I - \rho^2 S^T (\rho SS^T + M)^{-1} S)\boldsymbol{\phi} \\ &\quad - 2\boldsymbol{\phi}^T (\rho S^T (\rho SS^T + M)^{-1})\boldsymbol{\xi} - \boldsymbol{\xi}^T (\rho SS^T + M)^{-1} \boldsymbol{\xi}\big\}, \end{aligned} \tag{3.41}$$

where $\boldsymbol{\xi}$ is $n \times 1$ with the ith entry $R_J(x_i, x)$, Q is $n \times n$ with the (i,j)th entry $R_J(x_i, x_j)$, $\boldsymbol{\phi}$ is $m \times 1$ with the νth entry $\phi_\nu(x)$, S is $n \times m$ with the (i,ν)th entry $\phi_\nu(x_i)$, $\rho = \tau^2/b$, $n\lambda = \sigma^2/b$, and $M = Q + n\lambda I$. Setting $\rho \to \infty$ in (3.41), one obtains the following theorem.

Theorem 3.6 *Let $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in* $\mathrm{span}\{\phi_\nu, \nu = 1, \ldots, \mathrm{m}\}$ *and η_1 has a mean zero Gaussian process prior with covariance function $E\big[\eta_1(x)\eta_1(y)\big] = bR_J(x,y)$. Observing $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$, the posterior variance of $\eta(x)$ satisfies*

$$b^{-1}\mathrm{Var}\big[\eta(x)|\mathbf{Y}\big] = R_J(x,x) + \boldsymbol{\phi}^T (S^T M^{-1} S)^{-1} \boldsymbol{\phi} - 2\boldsymbol{\phi}^T \tilde{\mathbf{d}} - \boldsymbol{\xi}^T \tilde{\mathbf{c}}, \tag{3.42}$$

where

$$\begin{aligned} \tilde{\mathbf{c}} &= (M^{-1} - M^{-1} S (S^T M^{-1} S)^{-1} S^T M^{-1}) \boldsymbol{\xi}, \\ \tilde{\mathbf{d}} &= (S^T M^{-1} S)^{-1} S^T M^{-1} \boldsymbol{\xi}. \end{aligned} \tag{3.43}$$

The proof of Theorem 3.6 follows readily from Lemma 2.7 of §2.5.2 and the following lemma.

Lemma 3.7 *Suppose M is symmetric and nonsingular and S is of full column rank.*

$$\lim_{\rho \to \infty} \rho I - \rho^2 S^T (\rho SS^T + M)^{-1} S = (S^T M^{-1} S)^{-1}. \tag{3.44}$$

Proof: From (2.39) on page 50, one has

$$\begin{aligned} S^T (\rho SS^T + M)^{-1} S &= (I - (I + \rho^{-1} (S^T M^{-1} S)^{-1})^{-1}) S^T M^{-1} S \\ &= \rho^{-1} (I + \rho^{-1} (S^T M^{-1} S)^{-1})^{-1}, \end{aligned}$$

so

$$\begin{aligned}\rho I - \rho^2 S^T(\rho SS^T + M)^{-1}S &= \rho(I - (I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}) \\ &= (I + \rho^{-1}(S^T M^{-1}S)^{-1})^{-1}(S^T M^{-1}S)^{-1}.\end{aligned}$$

The lemma follows. □

Now, consider the multiple-term model of §2.5.3 in $\mathcal{H} = \oplus_{\beta=0}^{p}\mathcal{H}_\beta$,

$$\eta(x) = \sum_{\nu=1}^{m}\psi_\nu(x) + \sum_{\beta=1}^{p}\eta_\beta(x),$$

where ψ_ν have diffuse priors in $\text{span}\{\phi_\nu\}$ with $\{\phi_\nu\}_{\nu=1}^m$ a basis of $\mathcal{H}_0$ and $\eta_\beta(x)$ have independent Gaussian process priors with mean zero and covariance functions $b\theta_\beta R_\beta(x,y)$. Remember that the model may also be perceived as a mixed-effect model, with ψ_ν, $\nu = 1,\ldots,m$, being the fixed effects and η_β, $\beta = 1,\ldots,p$, being the random effects.

Theorem 3.8 *Under the multiple-term model specified above, observing* $Y_i = \eta(x_i) + \epsilon_i$, $\epsilon_i \sim N(0,\sigma^2)$, $i = 1,\ldots,n$, *the posterior means and covariances of the fixed effects* ψ_ν *and the random effects* η_β *are as follows:*

$$E\big[\psi_\nu(x)|\mathbf{Y}\big] = \phi_\nu(x)\mathbf{e}_\nu^T\mathbf{d}, \tag{3.45}$$

$$E\big[\eta_\beta(x)|\mathbf{Y}\big] = \boldsymbol{\xi}_\beta^T\mathbf{c}, \tag{3.46}$$

$$b^{-1}\text{Cov}\big[\psi_\nu(x),\psi_\mu(x)|\mathbf{Y}\big] = \phi_\nu(x)\phi_\mu(x)\mathbf{e}_\nu^T(S^TM^{-1}S)^{-1}\mathbf{e}_\mu, \tag{3.47}$$

$$b^{-1}\text{Cov}\big[\psi_\nu(x),\eta_\beta(x)|\mathbf{Y}\big] = -\phi_\nu(x)\mathbf{e}_\nu^T\tilde{\mathbf{d}}_\beta, \tag{3.48}$$

$$b^{-1}\text{Cov}\big[\eta_\beta(x),\eta_\gamma(x)|\mathbf{Y}\big] = \theta_\beta R_\beta(x,x)\delta_{\beta,\gamma} - \tilde{\mathbf{c}}_\beta^T\boldsymbol{\xi}_\gamma, \tag{3.49}$$

where $\mathbf{c}$ *and* $\mathbf{d}$ *are as given in (2.40),* $\mathbf{e}_\nu$ *is the* ν*th unit vector of size* $m\times 1$, $\boldsymbol{\xi}_\beta$ *is* $n\times 1$ *with the* i*th entry* $\theta_\beta R_\beta(x_i,x)$, *and*

$$\begin{aligned}\tilde{\mathbf{c}}_\beta &= (M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\boldsymbol{\xi}_\beta, \\ \tilde{\mathbf{d}}_\beta &= (S^TM^{-1}S)^{-1}S^TM^{-1}\boldsymbol{\xi}_\beta.\end{aligned} \tag{3.50}$$

The proof of the theorem is straightforward but tedious following the lines of the proofs of Theorems 2.8 and 3.6; see Problem 3.9.

The results of Theorems 2.8, 3.6, and 3.8 can be used to construct interval estimates of $\eta(x)$, of its components $\psi_\nu(x)$ and $\eta_\beta(x)$, and of their linear combinations. See Problem 3.10.

For weighted data with weights w_i, one simply replaces, in the formulas appearing in Theorems 3.6 and 3.8, S by $W^{1/2}S$, $M = Q + n\lambda I$ by $M_w = W^{1/2}QW^{1/2} + n\lambda I$, $\boldsymbol{\xi}_\beta$ by $W^{1/2}\boldsymbol{\xi}_\beta$, and $\mathbf{c}$, $\tilde{\mathbf{c}}$, and $\tilde{\mathbf{c}}_\beta$ by $W^{-1/2}\mathbf{c}$, $W^{-1/2}\tilde{\mathbf{c}}$, and $W^{-1/2}\tilde{\mathbf{c}}_\beta$, respectively, where $W = \text{diag}(w_i)$; see Problem 3.11.

3.3.2 Confidence Intervals on Sampling Points

At a sampling point x_i, $\boldsymbol{\phi}^T$ is the ith row of S and $\boldsymbol{\xi}$ is the ith column of Q. Write $B = S(S^T M^{-1} S)^{-1} S^T$. It is easy to check that $b^{-1}\mathrm{Var}\big[\eta(x_i)|\mathbf{Y}\big]$ as given in Theorem 3.6 is the (i,i)th entry of the matrix

$$Q + B - BM^{-1}Q - QM^{-1}B - Q(M^{-1} - M^{-1}BM^{-1})Q. \tag{3.51}$$

Note that $QM^{-1} = M^{-1}Q = I - n\lambda M^{-1}$. Following straightforward but tedious algebra, (3.51) simplifies to

$$n\lambda\big(I - n\lambda(M^{-1} - M^{-1}BM^{-1})\big) = n\lambda A(\lambda),$$

where the last equation is from (3.8); see Problem 3.12. With b and $\sigma^2 = (n\lambda)b$ known, the $100(1-\alpha)\,\%$ confidence interval of $\eta(x_i)$ based on the posterior distribution is thus

$$\eta_\lambda(x_i) \pm z_{\alpha/2}\,\sigma\sqrt{a_{i,i}}, \tag{3.52}$$

where η_λ is the minimizer of (3.1) and $a_{i,i}$ is the (i,i)th entry of the smoothing matrix $A(\lambda)$ given in (3.8).

For weighted data with weights w_i, it can be shown that $b^{-1}\mathrm{Var}\big[\eta(x_i)|\mathbf{Y}\big]$ is the (i,i)th entry of $n\lambda W^{-1/2}A_w(\lambda)W^{-1/2}$, where $A_w(\lambda)$ is given in (3.12); see Problem 3.13.

3.3.3 Across-the-Function Coverage

Despite its derivation from the Bayes model, the interval estimates of (3.52), when used with the GCV smoothing parameter λ_v and the corresponding variance estimate $\hat{\sigma}_v^2$, demonstrate a certain across-the-function coverage property for η fixed and smooth, as was illustrated by Wahba (1983).

Over the sampling points, define the average coverage proportion

$$\mathrm{ACP}(\alpha) = \frac{1}{n}\#\big\{i : \big|\eta_{\lambda_v}(x_i) - \eta(x_i)\big| < z_{\alpha/2}\,\hat{\sigma}_v\,\sqrt{a_{i,i}}\big\}.$$

Simulation results in Wahba (1983) suggest that for n large,

$$E\big[\mathrm{ACP}(\alpha)\big] \approx 1 - \alpha, \tag{3.53}$$

where the expectation is with respect to ϵ_i in $Y_i = \eta(x_i) + \epsilon_i$ with $\eta(x)$ fixed and smooth. Note that the construction of the intervals is pointwise but the coverage property is across-the-function. Heuristic arguments in support of (3.53) can be found in Wahba (1983). A more rigorous treatment for smoothing splines on $[0,1]$ was given by Nychka (1988), but it is unclear whether a general treatment is possible.

For the components $\psi_\nu(x)$ and $\eta_\beta(x)$ and their linear combinations, one may likewise define the corresponding average coverage proportion.

TABLE 3.1. Empirical ACP in simulation.

α	$n = 100$	$n = 200$	$n = 500$
0.05	0.943	0.958	0.962
0.10	0.897	0.915	0.911

The counterpart of (3.53) for componentwise intervals appears less plausible, however, as the simulations of Gu and Wahba (1993b) suggest.

To put (3.53) in perspective, consider some parametric model $\eta(x) = f(x, \boldsymbol{\beta})$ with $f(x, \boldsymbol{\beta})$ known up to the parameters $\boldsymbol{\beta}$. The standard large sample confidence interval for $\eta(x)$, $f(x, \hat{\boldsymbol{\beta}}) \pm z_{\alpha/2}\, \hat{\sigma}_{f(x,\hat{\beta})}$, has the pointwise coverage property

$$P\big(\big|f(x, \hat{\boldsymbol{\beta}}) - \eta(x)\big| < z_{\alpha/2}\, \hat{\sigma}_{f(x,\hat{\beta})}\big) \approx 1 - \alpha. \tag{3.54}$$

The property (3.53) is weaker than (3.54), but (3.54) does imply (3.53). Hence, the intervals satisfying (3.53) can be compared with the standard confidence intervals in parametric models on the basis of the across-the-function coverage property.

For the replicates in the simulation of §3.2.5, ACP(α) was also calculated for $\alpha = 0.05, 0.10$. The results are summarized in Table 3.1.

3.4 Computation: Generic Algorithms

For the estimation tools developed in §§3.2 and 3.3 to be practical, one needs efficient algorithms for the minimization of $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$ with respect to the smoothing parameters. Generic algorithms based on the linear system (3.4) are the topic of this section. From discussions in §§3.1–3.3 concerning weighted data, it is clear that the same algorithms are applicable to the penalized weighted least squares problem of (3.9) through the linear system (3.10). Special algorithms for problems with certain structures are to be found in §3.10.

Fixing the smoothing parameters, one needs $n^3/3 + O(n^2)$ floating-point operations, or flops, to calculate η_λ. This serves as a benchmark to measure the relative efficiency of the practical algorithms to follow. With only λ tunable, one needs about four times as many flops to execute the algorithm of §3.4.2 to minimize $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$. With λ and θ_β, $\beta = 1, \ldots, p$, all tunable, the iterative algorithm of §3.4.3 takes $4pn^3/3 + O(n^2)$ flops per iteration and needs about 5–10 iterations to converge on most problems. The algorithms are largely based on standard numerical linear algebra procedures, of which details, including the flop counts, can be found in Golub and Van Loan (1989).

As in previous sections, we suppress from the notation the dependence of entities on θ_β, except in §3.4.3.

3.4.1 Algorithm for Fixed Smoothing Parameters

Fixing the smoothing parameters λ and θ_β hidden in Q, the calculation of $\mathbf{c}$ and $\mathbf{d}$ in (3.6) is straightforward using standard numerical linear algebra procedures.

For $\mathbf{c}$, one calculates the Cholesky decomposition $(F_2^T Q F_2 + n\lambda I) = G^T G$, where G is upper-triangular, solves for $\mathbf{u}$ from $G\mathbf{u} = F_2^T \mathbf{Y}$ by back substitution and for $\mathbf{v}$ from $G^T \mathbf{v} = \mathbf{u}$ by forward substitution, then $\mathbf{c} = F_2 \mathbf{v}$; for $\mathbf{d}$, one simply solves $\tilde{R}\mathbf{d} = (F_1^T \mathbf{Y} - F_1^T Q F_2 \mathbf{v})$ by back substitution. See, e.g., Golub and Van Loan (1989, §§4.2 and 3.1) for Cholesky decomposition and forward and back substitutions.

The calculation of the Cholesky decomposition takes $n^3/3 + O(n^2)$ flops, and the rest of the computation, including the QR-decomposition $S = FR^* = (F_1, F_2)\begin{pmatrix} \tilde{R} \\ O \end{pmatrix}$ and the formation of $F^T Q F$, takes $O(n^2)$ flops. This algorithm is rarely used in practice, since it is inadequate to fix the smoothing parameters, but its flop count serves as a benchmark to measure the relative efficiency of the practical algorithms to follow.

3.4.2 Algorithm for Single Smoothing Parameter

We now present an algorithm for the minimization of $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$ as functions of a single smoothing parameter λ. The algorithm employs a one-time $O(n^3)$ matrix decomposition to introduce a certain banded structure, with which the evaluations of $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$ become negligible $O(n)$ operations. The algorithm also serves as a building block in the algorithm for multiple smoothing parameters, to be discussed in §3.4.3.

Algorithm 3.1 Given S, Q, $\mathbf{Y}$, and possibly σ^2 as inputs, perform the following steps to minimize $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$, and return the associated coefficients $\mathbf{c}$, $\mathbf{d}$:

1. Initialization:
   (a) Compute the QR-decomposition $S = FR^* = (F_1, F_2)\begin{pmatrix} \tilde{R} \\ O \end{pmatrix}$.
   (b) Compute $F^T \mathbf{Y}$, $F^T Q F$, from which $\mathbf{z} = F_2^T \mathbf{Y}$, $Q^* = F_2^T Q F_2$, $F_1^T \mathbf{Y}$, and $F_1^T Q F_2$ can be extracted.
2. Tridiagonalization and minimization:
   (a) Compute $Q^* = UTU^T$, where U is orthogonal and T is tridiagonal.
   (b) Compute $\mathbf{x} = U^T \mathbf{z}$.

(c) Minimize one of the following scores:

$$U^*(\lambda) = \frac{1}{n}\mathbf{x}^T(T+n\lambda I)^{-2}\mathbf{x} - \frac{2\sigma^2}{n}(n\lambda)\mathrm{tr}(T+n\lambda I)^{-1}, \quad (3.55)$$

$$V(\lambda) = \frac{n^{-1}\mathbf{x}^T(T+n\lambda I)^{-2}\mathbf{x}}{[n^{-1}\mathrm{tr}(T+n\lambda I)^{-1}]^2}, \quad (3.56)$$

$$M(\lambda) = \frac{n^{-1}\mathbf{x}^T(T+n\lambda I)^{-1}\mathbf{x}}{|T+n\lambda I|^{-1/(n-M)}}, \quad (3.57)$$

with respect to λ.

3. Compute return values:

(a) Compute $\mathbf{v} = U(T+n\lambda I)^{-1}\mathbf{x}$ at the selected λ.

(b) Return $\mathbf{c} = F_2\mathbf{v}$ and $\mathbf{d} = \tilde{R}^{-1}(F_1^T\mathbf{Y} - F_1^TQF_2\mathbf{v})$.

Note that $U^*(\lambda) = U(\lambda) - 2\sigma^2$ and that

$$I - A(\lambda) = (n\lambda)F_2(F_2^TQF_2 + n\lambda I)^{-1}F_2^T = (n\lambda)F_2U(T+n\lambda I)^{-1}U^TF_2^T.$$

Step 1(a) and $F^T\mathbf{Y}$ in step 1(b) are implemented in the LINPACK routines `dqrdc` and `dqrsl`; see Dongarra et al. (1979). An implementation of $Q = F^TQF$ in step 1(b), which uses the output of `dqrdc` in a similar manner as `dqrsl` does, is implemented in RKPACK; see Gu (1989). Golub and Van Loan (1989, §§5.1–5.2) and Dongarra et al. (1979) are good places to read about the details of these calculations. The execution of step 1 takes $O(n^2)$ flops.

Step 2(a) via Householder tridiagonalization is the most time-consuming step in Algorithm 3.1, which usually takes $4n^3/3$ flops; see, e.g., Golub and Van Loan (1989, §8.2.1). With a numerically singular Q^*, however, it is possible to speed up the process by employing a certain truncation scheme in the algorithm; see Gu et al. (1989). Step 2(b) is simply another application of the LINPACK routine `dqrsl`.

The crux of Algorithm 3.1 is in step 2(c), where one has to evaluate $U(\lambda)$, $V(\lambda)$, or $M(\lambda)$ at multiple λ values. The band Cholesky decomposition $T + n\lambda I = C^TC$ for T tridiagonal can be computed in $O(n)$ flops, where

$$C = \begin{pmatrix} a_1 & b_1 & & \\ & \ddots & \ddots & \\ & & a_{n_1-1} & b_{n_1-1} \\ & & & a_{n_1} \end{pmatrix}$$

for $n_1 = n - m$; see Golub and Van Loan (1989, §4.3.6). Through a band back substitution followed by a band forward substitution, $(T+n\lambda I)^{-1}\mathbf{x}$ is now available in $O(n)$ flops; see Golub and Van Loan (1989, §4.3.2).

For $M(\lambda)$ in (3.57), $|T+n\lambda I| = \prod_{i=1}^{n_1} a_i^2$ is straightforward. The nontrivial part of this step is the efficient evaluation of the term $\mathrm{tr}(T+n\lambda I)^{-1} = \mathrm{tr}(C^{-1}C^{-T})$ in $U^*(\lambda)$ of (3.55) and $V(\lambda)$ of (3.56).

Write $C^{-T} = (\mathbf{c}_1, \ldots, \mathbf{c}_{n_1})$; it is clear that $\mathrm{tr}(C^{-1}C^{-T}) = \sum_{i=1}^{n_1} \mathbf{c}_i^T \mathbf{c}_i$. From

$$C^{-T}C^T = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{n_1}) \begin{pmatrix} a_1 & & & \\ b_1 & \ddots & & \\ & \ddots & a_{n_1-1} & \\ & & b_{n_1-1} & a_{n_1} \end{pmatrix} = I,$$

one has

$$\begin{aligned} a_{n_1}\mathbf{c}_{n_1} &= \mathbf{e}_{n_1}, \\ a_i\mathbf{c}_i &= \mathbf{e}_i - b_i\mathbf{c}_{i+1}, \qquad i = n_1-1, \ldots, 1, \end{aligned}$$

where $\mathbf{e}_i$ is the ith unit vector. Because C^{-T} is lower-triangular (Problem 3.14), $\mathbf{c}_{i+1}$ is orthogonal to $\mathbf{e}_i$. Thus, one has recursive formulas

$$\begin{aligned} \mathbf{c}_{n_1}^T\mathbf{c}_{n_1} &= a_{n_1}^{-2}, \\ \mathbf{c}_i^T\mathbf{c}_i &= (1 + b_i^2\mathbf{c}_{i+1}^T\mathbf{c}_{i+1})a_i^{-2}, \qquad i = n_1-1, \ldots, 1. \end{aligned} \tag{3.58}$$

The calculation in (3.58) is clearly of order $O(n)$. This technique for the efficient calculation of $\mathrm{tr}\big(I - A(\lambda)\big)$ is due to Elden (1984).

At the selected λ, one has

$$\begin{aligned} \mathbf{c} &= F_2U(T+n\lambda I)^{-1}\mathbf{x}, \\ \mathbf{d} &= \tilde{R}^{-1}\left(F_1^T\mathbf{Y} - (F_1^TQF_2)U(T+n\lambda I)^{-1}\mathbf{x}\right), \end{aligned}$$

which are available in $O(n)$ flops. Also available in $O(n)$ flops are

$$\begin{aligned} \hat{\sigma}_v^2 &= \frac{(n\lambda_v)\mathbf{x}(T+n\lambda_v I)^{-2}\mathbf{x}}{\mathrm{tr}(T+n\lambda_v I)^{-1}}, \\ \hat{\sigma}_m^2 &= \frac{(n\lambda_m)\mathbf{x}(T+n\lambda_m I)^{-1}\mathbf{x}}{n-M}. \end{aligned}$$

Overall, Algorithm 3.1 takes $4n^3/3 + O(n^2)$ flops to execute, about four times what is needed for the calculation of $\mathbf{c}$ and $\mathbf{d}$ with a fixed λ.

3.4.3 Algorithm for Multiple Smoothing Parameters

We now briefly describe an algorithm for the minimization of $U(\lambda;\boldsymbol{\theta})$, $V(\lambda;\boldsymbol{\theta})$, or $M(\lambda;\boldsymbol{\theta})$ as functions of smoothing parameters λ and θ_β hidden in $Q = \sum_{\beta=1}^{p}\theta_\beta Q_\beta$, where Q_β has the (i,j)th entry $R_\beta(x_i, x_j)$. The algorithm operates on λ and $\vartheta_\beta = \log\theta_\beta$. We state the algorithm in terms of $V(\lambda;\boldsymbol{\theta})$, but the same procedures readily apply to $U(\lambda;\boldsymbol{\theta})$ and $M(\lambda;\boldsymbol{\theta})$.

Algorithm 3.2 Given S, Q_β, $\beta = 1, \ldots, p$, $\mathbf{Y}$, starting values $\boldsymbol{\vartheta}_0$, and possibly σ^2 as inputs, perform the following steps to minimize $V(\lambda; \boldsymbol{\theta})$ and return the associated coefficients $\mathbf{c}$, $\mathbf{d}$:

1. Initialization:
   (a) Compute the QR-decomposition $S = FR^* = (F_1, F_2)\left(\begin{smallmatrix}\tilde{R}\\ O\end{smallmatrix}\right)$.
   (b) Compute $F^T\mathbf{Y}$ and $F^T Q_\beta F$, from which $\mathbf{z} = F_2^T\mathbf{Y}$, $Q_\beta^* = F_2^T Q F_2$, $F_1^T\mathbf{Y}$, and $F_1^T Q_\beta F_2$ can be extracted.
   (c) Set $\Delta\boldsymbol{\vartheta} = 0$, $\boldsymbol{\vartheta}_- = \boldsymbol{\vartheta}_0$, and $V_- = \infty$.
2. Iteration:
   (a) For the trial value $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_- + \Delta\boldsymbol{\vartheta}$, collect $Q^* = \sum_{\beta=1}^p \theta_\beta Q_\beta^*$ and scale it to have a fixed trace.
   (b) Compute $Q^* = UTU^T$, where U is orthogonal and T is tridiagonal. Compute $\mathbf{x} = U^T\mathbf{z}$.
   (c) Minimize $V(\lambda; \boldsymbol{\theta})$ with respect to λ. If $V > V_-$, set $\Delta\boldsymbol{\vartheta} = \Delta\boldsymbol{\vartheta}/2$, go to (a); else proceed.
   (d) Evaluate the gradient $\mathbf{g} = (\partial/\partial\boldsymbol{\vartheta})V(\lambda; \boldsymbol{\theta})$ and the Hessian $H = (\partial^2/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^T)V(\lambda; \boldsymbol{\theta})$.
   (e) Calculate the increment $\Delta\boldsymbol{\vartheta} = -\tilde{H}^{-1}\mathbf{g}$, where $\tilde{H} = H + \mathrm{diag}(\mathbf{e})$ is positive definite. If H itself is positive definite "enough," $\mathbf{e}$ is simply set to 0.
   (f) Check convergence conditions. If the conditions fail, set $\boldsymbol{\vartheta}_- = \boldsymbol{\vartheta}$, $V_- = V$, go to (a).
3. Compute return values:
   (a) Compute $\mathbf{v} = U(T + n\lambda I)^{-1}\mathbf{x}$ at the converged λ and $\boldsymbol{\vartheta}$.
   (b) Return $\mathbf{c} = F_2\mathbf{v}$ and $\mathbf{d} = \tilde{R}^{-1}(F_1^T\mathbf{Y} - F_1^T Q F_2\mathbf{v})$, with $Q = \sum_{\beta=1}^p Q_\beta$.

The calculations in step 1 of Algorithm 3.2 are the same as those in step 1 of Algorithm 3.1 and can be executed in $O(n^2)$ flops. Steps 2(a) through 2(c) with fixed θ_β virtually duplicate step 2 of Algorithm 3.1, which takes $4n^3/3 + O(n^2)$ flops to execute. The calculation of gradient and Hessian in step 2(d) takes an extra $4(p-1)n^3/3 + O(n^2)$ flops; see Gu and Wahba (1991b). Each iteration of step 2 takes altogether $4pn^3/3 + O(n^2)$ flops.

The scores $U(\lambda; \boldsymbol{\theta})$, $V(\lambda; \boldsymbol{\theta})$, or $M(\lambda; \boldsymbol{\theta})$ are fully parameterized by

$$(\lambda_1, \ldots, \lambda_p) = (\lambda\theta_1^{-1}, \ldots, \lambda\theta_p^{-1}),$$

so $(\lambda, \theta_1, \ldots, \theta_p)$ form an overparameterization. This is the reason for the scaling in step 2(a). One may directly employ the Newton iteration with

respect to the parameters $\log \lambda_\beta$ to minimize the scores, but the calculation of the gradient and the Hessian would take $4pn^3/3 + O(n^2)$ flops anyway. In this sense, the extra gain through step 2(c) is virtually free.

Step 2(e) returns a descent direction even when the Hessian H is not positive definite. The algorithm to use here is the modified Cholesky decomposition as described in Gill et al. (1981, §4.4.2.2), which adds positive mass to the diagonal elements of H, if necessary, to produce a factorization $\tilde{H} = G^T G$, where G is upper-triangular.

Algorithm 3.3 To obtain a set of starting values $(\lambda_0, \theta_{10}, \ldots, \theta_{p0})$ for use in Algorithm 3.2, perform the following steps.

1. Set $\tilde{\theta}_\beta = \big(\mathrm{tr}(Q_\beta)\big)^{-1}$ and $Q = \sum_{\beta=1}^p \tilde{\theta}_\beta Q_\beta$, then use Algorithm 3.1 to obtain an initial fit $\tilde{\eta} = \sum_{\beta=0}^p \tilde{\eta}_\beta$, where $\tilde{\eta}_0 = \boldsymbol{\phi}^T \mathbf{d}$ and $\tilde{\eta}_\beta = \boldsymbol{\xi}_\beta^T \mathbf{c}$, $\beta = 1, \ldots, p$, with $\boldsymbol{\xi}_\beta$ having entries $\tilde{\theta}_\beta R_\beta(x_i, x)$.

2. Set $\theta_{\beta 0} \propto (\tilde{\eta}, \tilde{\eta})_\beta = \tilde{\theta}_\beta^2 \mathbf{c}^T Q_\beta \mathbf{c}$ and $Q = \sum_{\beta=1}^p \theta_{\beta 0} Q_\beta$, then use Algorithm 3.1 again to obtain λ_0.

The choice of $\tilde{\theta}_\beta$ in Step 1 of Algorithm 3.3 is arbitrary but invariant to the relative scaling of $(f, f)_\beta$. The initial fit $\tilde{\eta}$ reveals where structures in the true η rest and one should apply less penalty where signal is strong; remember that $J(f) = \sum_{\beta=1}^p \theta_\beta^{-1} (f, f)_\beta$. Using starting values from Algorithm 3.3, Algorithm 3.2 typically converges in five to ten iterations.

3.4.4 Calculation of Posterior Variances

From (3.47) to (3.49) in Theorem 3.8, one needs $(S^T M^{-1} S)^{-1}$, $\tilde{\mathbf{c}}_\beta$, and $\tilde{\mathbf{d}}_\beta$ to construct the Bayesian confidence intervals. At the converged λ and θ_β, it is easy to calculate

$$\begin{aligned} \tilde{\mathbf{c}}_\beta &= F_2 U (T + n\lambda I)^{-1} U^T F_2^T \boldsymbol{\xi}_\beta, \\ \tilde{\mathbf{d}}_\beta &= \tilde{R}^{-1} (F_1^T \boldsymbol{\xi}_\beta - (F_1^T Q F_2) U (T + n\lambda I)^{-1} U^T F_2^T \boldsymbol{\xi}_\beta) \end{aligned} \tag{3.59}$$

in $O(n)$ extra flops. The remaining task is the calculation of $(S^T M^{-1} S)^{-1}$. Using an elementary matrix identity (Problem 3.15), one has

$$\begin{aligned} S^T M^{-1} S &= \tilde{R}^T F_1^T (Q + n\lambda I)^{-1} F_1 \tilde{R} \\ &= \tilde{R}^T (I, O) F^T (Q + n\lambda I)^{-1} F \begin{pmatrix} I \\ O \end{pmatrix} \tilde{R} \\ &= \tilde{R}^T (I, O) (F^T Q F + n\lambda I)^{-1} \begin{pmatrix} I \\ O \end{pmatrix} \tilde{R} \\ &= \tilde{R}^T \big((F_1^T Q F_1 + n\lambda I) \\ &\qquad - (F_1^T Q F_2)(Q^* + n\lambda I)^{-1} (F_2^T Q F_1) \big)^{-1} \tilde{R} \\ &= \tilde{R}^T \big((F_1^T Q F_1 + n\lambda I) \\ &\qquad - (F_1^T Q F_2) U (T + n\lambda I)^{-1} U^T (F_2^T Q F_1) \big)^{-1} \tilde{R}; \end{aligned}$$

hence,

$$(S^T M^{-1} S)^{-1} = \tilde{R}^{-1}\big((F_1^T Q F_1 + n\lambda I) - (F_1^T Q F_2) U (T + n\lambda I)^{-1} U^T (F_2^T Q F_1)\big)\tilde{R}^{-T}, \quad (3.60)$$

which is available in $O(n)$ extra flops.

3.5 Efficient Approximation

The penalty $\lambda J(f)$ effectively enforces a low dimensional model space (see, e.g., §4.2.2), so an infinite dimensional $\mathcal{H}$ is not really necessary. It is shown in §9.4.4 that the minimizer of (3.1) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\big\{R_J(z_j, \cdot), j = 1, \ldots, q\big\}$$

shares the same asymptotic convergence rates as the minimizer in $\mathcal{H}$, and hence is statistically as efficient, where $\{z_j\}$ is a random subset of $\{x_i\}$ and $q \to \infty$ can be at a rate much slower than n. This allows for algorithms of order $O(nq^2)$, more scalable than $O(n^3)$ for $q = o(n)$.

The minimizer of (3.1) in $\mathcal{H}^*$ can also be cast as a Bayes estimate, and the results of §§2.5, 3.3, and 3.2.3 remain valid after minor modifications. The algorithms of §3.4 no longer apply, so alternative numerical approaches will be explored. A small q is preferred for numerical efficiency but too small a q may impair statistical performance; the practical choice of q will be guided by asymptotic analysis and empirical simulations. Also assessed is the numerical accuracy of quantities associated with the minimizer in $\mathcal{H}^*$ as approximations to those associated with the minimizer in $\mathcal{H}$.

3.5.1 Preliminaries

Functions in $\mathcal{H}^*$ can be written as

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_J(z_j, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \quad (3.61)$$

with (3.2) on page 62 as a special case at $q = n$. Plugging (3.61) into (3.1), one minimizes

$$(\mathbf{Y} - S\mathbf{d} - R\mathbf{c})^T (\mathbf{Y} - S\mathbf{d} - R\mathbf{c}) + n\lambda \mathbf{c}^T Q \mathbf{c} \quad (3.62)$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is as in (3.3), R is $n \times q$ with the (i, j)th entry $R_J(x_i, z_j)$, and Q is $q \times q$ with the (j, k)th entry $R_J(z_j, z_k)$; note that Q is part of R, and (3.3) is a special case of (3.62) with $R = Q$. We assume a full column rank for S as in §3.1, which ensures a unique minimizer of (3.1) even though the coefficients $\mathbf{c}$ and $\mathbf{d}$ may not be unique.

Differentiating (3.62) with respect to $\mathbf{c}$ and $\mathbf{d}$ and setting the derivatives to 0, some algebra yields the linear system

$$\begin{pmatrix} S^TS & S^TR \\ R^TS & R^TR + n\lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T\mathbf{Y} \\ R^T\mathbf{Y} \end{pmatrix}. \tag{3.63}$$

For the weighted data as in §3.2.4, one may simply replace $(\mathbf{Y}, S, R)$ in (3.63) by $(\mathbf{Y}_w, S_w, R_w) = W^{1/2}(\mathbf{Y}, S, R)$, and all the derivations in the rest of the section hold for weighted data with these substitutions.

On $\mathcal{X} = [0,1]$ with $J(f) = \int_0^1 \ddot{f}^2 dx$, the minimizer of (3.1) in $\mathcal{H}$ is a piecewise cubic polynomial as noted in §1.1.1. The basis functions $R_J(x_i, x)$ can involve x^4, see (2.26) and (2.27) on page 39, but the constraint $S^T\mathbf{c} = 0$ in (3.4) ensures that the coefficients of x^4 cancel out. Such a constraint does not apply to the solution of (3.63), however, so the minimizer in $\mathcal{H}^*$ may no longer be a piecewise cubic polynomial. Still, despite the technical inaccuracy, we will keep referring to such estimates as cubic splines.

3.5.2 Bayes Model

Consider $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in $\mathcal{N}_J$ and η_1 has a mean zero Gaussian process prior with a covariance function

$$E\big[\eta_1(x)\eta_1(y)\big] = bR_J(x, \mathbf{z}^T)Q^+R_J(\mathbf{z}, y),$$

where Q^+ is the Moore-Penrose inverse of $Q = R_J(\mathbf{z}, \mathbf{z}^T)$. The counterpart of (2.35) on page 49 is given by

$$\begin{pmatrix} bRQ^+R^T + \tau^2SS^T + \sigma^2I & bRQ^+\boldsymbol{\xi} + \tau^2S\boldsymbol{\phi} \\ b\boldsymbol{\xi}^TQ^+R^T + \tau^2\boldsymbol{\phi}^TS^T & b\boldsymbol{\xi}^TQ^+\boldsymbol{\xi} + \tau^2\boldsymbol{\phi}^T\boldsymbol{\phi}, \end{pmatrix} \tag{3.64}$$

and that of (2.36) by

$$\begin{aligned} E\big[\eta(x)|Y\big] &= (b\boldsymbol{\xi}^TQ^+R^T + \tau^2\boldsymbol{\phi}^TS^T)(bRQ^+R^T + \tau^2SS^T + \sigma^2I)^{-1}\mathbf{Y} \\ &= \rho\boldsymbol{\phi}^TS^T(M + \rho SS^T)^{-1}\mathbf{Y} + \boldsymbol{\xi}^TQ^+R^T(M + \rho SS^T)^{-1}\mathbf{Y}, \end{aligned}$$

where $M = RQ^+R^T + n\lambda I$, $n\lambda = \sigma^2/b$, and $\rho = \tau^2/b$. Setting $\rho \to \infty$ and applying Lemma 2.7, one has

$$E\big[\eta(x)|Y\big] = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{3.65}$$

where

$$\begin{aligned} \mathbf{d} &= (S^TM^{-1}S)^{-1}S^TM^{-1}\mathbf{Y}, \\ \mathbf{c} &= Q^+R^T(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})\mathbf{Y}. \end{aligned} \tag{3.66}$$

Since $J(f)$ is a square norm in $\text{span}\{\xi_j\} = \mathcal{H}^* \ominus \mathcal{N}_J$, $J(\boldsymbol{\xi}^T\mathbf{c}) = \mathbf{c}^TQ\mathbf{c} = 0$ implies $\boldsymbol{\xi}^T\mathbf{c} = 0$, so $\boldsymbol{\xi}(x)$ is in the column space of Q, $\forall x$, and hence

$QQ^+R^T = R^T$, where QQ^+ is the projection matrix in the column space of Q. It is then easy to verify that the $\mathbf{c}$ and $\mathbf{d}$ in (3.66) solve (3.63) (Problem 3.16). Parallel to (3.42) on page 76, one also has

$$b^{-1}\mathrm{var}\big[\eta(x)|Y\big] = \boldsymbol{\xi}^T Q^+ \boldsymbol{\xi} + \boldsymbol{\phi}^T (S^T M^{-1} S)^{-1} \boldsymbol{\phi} - 2\boldsymbol{\phi}^T \tilde{\mathbf{d}} - \boldsymbol{\xi}^T \tilde{\mathbf{c}}, \quad (3.67)$$

where

$$\begin{aligned} \tilde{\mathbf{d}} &= (S^T M^{-1} S)^{-1} S^T M^{-1} R Q^+ \boldsymbol{\xi}, \\ \tilde{\mathbf{c}} &= Q^+ R^T (M^{-1} - M^{-1} S (S^T M^{-1} S)^{-1} S^T M^{-1}) R Q^+ \boldsymbol{\xi}. \end{aligned} \quad (3.68)$$

From (3.66), it is easy to verify that

$$A(\lambda) = I - n\lambda (M^{-1} - M^{-1} S (S^T M^{-1} S)^{-1} S^T M^{-1}), \quad (3.69)$$

which appears identical to (3.8) on page 63 but with an alternatively defined $M = RQ^+R^T + n\lambda I$. Evaluating (3.67) at a sampling point x_i yields the (i,i)th entry of $n\lambda A(\lambda)$; (3.51) on page 78 holds with RQ^+R^T replacing Q and the same algebra carries through.

For $R_J(x,y) = \sum_{\beta=1}^p \theta_\beta R_\beta(x,y)$, replace η_1 above by a sum $\sum_{\beta=1}^p \eta_\beta$ with prior covariance functions given by

$$E\big[\eta_\beta(x)\eta_\gamma(y)\big] = b\,\theta_\beta\,\theta_\gamma\, R_\beta(x, \mathbf{z}^T) Q^+ R_\gamma(\mathbf{z}, y), \quad \beta, \gamma = 1, \ldots, p.$$

Also decompose the diffuse terms $\eta_0 = \sum_{\nu=1}^m \psi_\nu$, where $\psi_\nu \in \mathrm{span}\{\phi_\nu\}$. The counterpart of Theorem 3.8 is tedious to state, but the posterior means and variances of arbitrary partial sums of ψ_ν and η_β can be obtained by simple modifications of (3.65), (3.67), and (3.68). For example, for the partial sum $\psi_1 + \eta_1 + \eta_2$, one simply replaces $\boldsymbol{\phi}$ in (3.65), (3.67), and (3.68) by $(\phi_1(x), 0, \ldots, 0)^T$ and $\boldsymbol{\xi}$ by $\theta_1 R_1(\mathbf{z}, x) + \theta_2 R_2(\mathbf{z}, x)$.

The derivation of REML in §3.2.3 remains largely intact after replacing Q by RQ^+R^T, yielding

$$M(\lambda) = \frac{n^{-1}\mathbf{Y}^T F_2 (F_2^T M F_2)^{-1} F_2^T \mathbf{Y}}{|F_2^T M F_2|^{-1/(n-m)}}, \quad (3.70)$$

where F_2 (and F_1 below) is from (3.5) on page 63 and $M = RQ^+R^T + n\lambda I$ as in (3.69). Partition

$$(F^T M F)^{-1} = F^T M^{-1} F = \begin{pmatrix} F_1^T M^{-1} F_1 & F_1^T M^{-1} F_2 \\ F_2^T M^{-1} F_1 & F_2^T M^{-1} F_2 \end{pmatrix}.$$

Using Problem 3.15, the bottom-right block of $F^T M F$ is seen to be

$$F_2^T M F_2 = (F_2^T M^{-1} F_2 - F_2^T M^{-1} F_1 (F_1^T M^{-1} F_1)^{-1} F_1^T M^{-1} F_2)^{-1}.$$

Note that (3.69) holds with F_1 replacing S, so one has

$$(F_2^T M F_2)^{-1} = (n\lambda)^{-1} F_2^T \big(I - A(\lambda)\big) F_2.$$

$I = F_1F_1^T + F_2F_2^T$, and from (3.69), $S^T\big(I - A(\lambda)\big) = O = F_1^T\big(I - A(\lambda)\big)$, so one has

$$F_2(F_2^T M F_2)^{-1}F_2^T = (n\lambda)^{-1}\big(I - A(\lambda)\big), \tag{3.71}$$

thus (3.70) can again be written as (3.30) on page 71 but with $A(\lambda)$ in (3.69) defined via $M = RQ^+R^T + n\lambda I$; (3.71) is the counterpart of (3.7).

3.5.3 Computation

The algorithms of §3.4 rely on a special structure in (3.3) not shared by (3.62) in general, that $R = Q$, so alternative numerical treatments are needed here.

With multiple smoothing parameters, analytical gradient and Hessian of $V(\lambda)$ (or $U(\lambda)$, $M(\lambda)$) used in Algorithm 3.2 are no longer available, and one has to employ quasi-Newton iterations with numerical derivatives, such as those developed in Dennis and Schnabel (1996), for smoothing parameter selection; (3.63) has to be updated and solved for each evaluation of $V(\lambda)$. When the number of θ_β's is large, quasi-Newton iterations can be slow to converge, but one may choose to skip the process as the starting values from Algorithm 3.3 often deliver "80 % or more" of the achievable performance.

Fixing the smoothing parameters λ and θ_β hidden in R and Q, and assuming a full column rank of R, the linear system (3.63) can be easily solved by a Cholesky decomposition of the $(m+q)\times(m+q)$ matrix followed by forward and back substitutions; see, e.g., Golub and Van Loan (1989, §§4.2 and 3.1). The formation of (3.63) takes $O(nq^2)$ flops, which, for $q = o(n)$, dominates the $O(q^3)$ Cholesky decomposition.

Care must be taken when R is not of full column rank. Write the Cholesky decomposition

$$\begin{pmatrix} S^TS & S^TR \\ R^TS & R^TR + n\lambda Q \end{pmatrix} = \begin{pmatrix} G_1^T & O \\ G_2^T & G_3^T \end{pmatrix}\begin{pmatrix} G_1 & G_2 \\ O & G_3 \end{pmatrix}, \tag{3.72}$$

where $S^TS = G_1^TG_1$, $G_2 = G_1^{-T}S^TR$, and

$$G_3^TG_3 = R^T(I - S(S^TS)^{-1}S^T)R + n\lambda Q.$$

Possibly with an permutation of indices known as pivoting, one may write

$$G_3 = \begin{pmatrix} J_1 & J_2 \\ O & O \end{pmatrix} = \begin{pmatrix} J \\ O \end{pmatrix},$$

where J_1 is nonsingular. Now define

$$\tilde{G}_3 = \begin{pmatrix} J_1 & J_2 \\ O & \delta I \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} G_1 & G_2 \\ O & \tilde{G}_3 \end{pmatrix};$$

one has

$$\tilde{G}^{-1} = \begin{pmatrix} G_1^{-1} & -G_1^{-1}G_2\tilde{G}_3^{-1} \\ O & \tilde{G}_3{}^{-1} \end{pmatrix}.$$

Premultiplying (3.63) by $\tilde{G}^{-T}$, some algebra yields

$$\begin{pmatrix} I & O \\ O & \tilde{G}_3^{-T}G_3^TG_3\tilde{G}_3{}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{d}^* \\ \mathbf{c}^* \end{pmatrix} = \begin{pmatrix} G_1^{-T}S^T\mathbf{Y} \\ \tilde{G}_3^{-T}R^T(I - S(S^TS)^{-1}S^T)\mathbf{Y} \end{pmatrix}, \tag{3.73}$$

where $\left(\begin{smallmatrix} \mathbf{d}^* \\ \mathbf{c}^* \end{smallmatrix}\right) = \tilde{G}\left(\begin{smallmatrix} \mathbf{d} \\ \mathbf{c} \end{smallmatrix}\right)$. Partitioning $\tilde{G}_3^{-1} = (K, L)$ such that $JK = I$ and $JL = O$, so

$$\tilde{G}_3^{-T}G_3^TG_3\tilde{G}_3{}^{-1} = \begin{pmatrix} K^T \\ L^T \end{pmatrix} G_3^TG_3(K, L) = \begin{pmatrix} K^T \\ L^T \end{pmatrix} J^TJ(K, L) = \begin{pmatrix} I & O \\ O & O \end{pmatrix}.$$

$L^TG_3^TG_3L = O$ implies $L^TR^T(I - S(S^TS)^{-1}S^T)RL = O$, so one has

$$L^TR^T(I - S(S^TS)^{-1}S^T)\mathbf{Y} = 0.$$

The linear system (3.73) is thus of the form

$$\begin{pmatrix} I & O & O \\ O & I & O \\ O & O & O \end{pmatrix} \begin{pmatrix} \mathbf{d}^* \\ \mathbf{c}_1^* \\ \mathbf{c}_2^* \end{pmatrix} = \begin{pmatrix} * \\ * \\ 0 \end{pmatrix}, \tag{3.74}$$

which is a solvable system but $\mathbf{c}_2^*$ can be arbitrary. Replacing the lower-right block O in the matrix on the left-hand side by I, which amounts to replacing G_3 in (3.72) by $\tilde{G}_3$, one sets $\mathbf{c}_2^* = 0$ in (3.74). In practice, one may simply perform the Cholesky decomposition of (3.72) with pivoting, replace the trailing O (if present) by δI for an appropriate value of δ, then proceed as if R were of full column rank.

The calculation of GCV scores is straightforward given that

$$\hat{\mathbf{Y}} = S\mathbf{d} + R\mathbf{c} = (S, R)\tilde{G}^{-1}\tilde{G}^{-T}\begin{pmatrix} S^T \\ R^T \end{pmatrix}\mathbf{Y} = A(\lambda)\mathbf{Y},$$

noting that $\mathrm{tr}A(\lambda)$ is the square norm of $(S, R)\tilde{G}^{-1}$ when it is treated as a long vector; this is an $O(nq^2)$ operation. The numerical accuracy of such trace evaluation is adequate unless $n\lambda$ is very small, a case one could prevent by using a fudge factor in (3.27). A stable, much more accurate algorithm for trace evaluation also exists but is of order $O(n^2q)$; see Kim and Gu (2004).

For the denominator of (3.70), as $|I - AB| = |I - BA|$ (Problem 3.17),

$$\begin{aligned} \left|(n\lambda)^{-1}F_2^TMF_2\right| &= \left|(n\lambda)^{-1}F_2^TRQ^+R^TF_2 + I\right| \\ &= \left|(n\lambda)^{-1}Q^+R^TF_2F_2^TR + I\right|. \end{aligned} \tag{3.75}$$

Consider the eigenvalue decomposition

$$Q^+ = (P_1, P_2)\begin{pmatrix} D_Q^{-1} & O \\ O & O \end{pmatrix}\begin{pmatrix} P_1^T \\ P_2^T \end{pmatrix} = P_1 D_Q^{-1} P_1^T,$$

where D_Q is diagonal with the positive eigenvalues of Q. As $P_2^T R^T = O$,

$$\begin{aligned} \left|(n\lambda)^{-1} Q^+ R^T F_2 F_2^T R + I\right| &= \left|D_Q^{-1}(n\lambda)^{-1} P_1^T R^T F_2 F_2^T R P_1 + I\right| \\ &= \left|Q + (n\lambda)^{-1} R^T F_2 F_2^T R\right|_+ / |Q|_+. \end{aligned}$$

The formation of $R^T F_2 F_2^T R$ is $O(nq^2)$ and the eigenvalue problem is $O(q^3)$.

For the evaluation of (3.67), $\tilde{\mathbf{d}}$ and $\tilde{\mathbf{c}}$ are available from $RQ^+\boldsymbol{\xi}$ and the Cholesky factor $\tilde{G}$, and $\boldsymbol{\xi}^T Q^+ \boldsymbol{\xi} = \boldsymbol{\xi}^T P_1 D_Q^{-1} P_1^T \boldsymbol{\xi}$. We now show that

$$\begin{aligned} (S^T M^{-1} S)^{-1} &= (n\lambda)(G_1^{-1} G_1^{-T} + G_1^{-1} G_2 \tilde{G}_3^{-1} \tilde{G}_3^{-T} G_2^T G_1^{-T}) \\ &= (n\lambda)\{(S^T S)^{-1} + (S^T S)^{-1} S^T R \tilde{G}_3^{-1} \tilde{G}_3^{-T} R^T S (S^T S)^{-1}\}, \end{aligned} \tag{3.76}$$

which is $n\lambda$ times the upper-left block of $\tilde{G}^{-1}\tilde{G}^{-T}$. First note that

$$M^{-1} = (n\lambda)^{-1}(I - R(n\lambda Q + R^T R)^+ R^T) \tag{3.77}$$

(Problem 3.18); multiply with M and simplify using the fact that

$$QQ^+ R^T = (n\lambda Q + R^T R)(n\lambda Q + R^T R)^+ R^T = R^T.$$

Substituting (3.77) in $S^T M^{-1} S$ and multiplying with the right-hand side of (3.76), straightforward algebra yields identity (Problem 3.19); remember that $G_3^T G_3 = R^T(I - S(S^T S)^{-1} S)R + n\lambda Q$ and note that

$$G_3^T G_3 \tilde{G}_3^{-1} \tilde{G}_3^{-T} R^T = R^T,$$

where $G_3^T G_3 = J^T J$ so J^T shares the same column space with Q, and $G_3^T G_3 \tilde{G}_3^{-1} \tilde{G}_3^{-T} = J^T K^T$ acts like a projection matrix as $JK = I$.

3.5.4 Empirical Choice of q

A small q is preferred computationally, but too small a q could make the fit overly dependent on the choice of $\{z_j\} \subset \{x_i\}$ or even introduce model bias. The empirical choice of q is to be guided by the theory of Chap. 9.

As $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$, the minimizer of (3.1) in $\mathcal{H}$ converges to the true η at a rate $O_p(n^{-1}\lambda^{1/r} + \lambda^p)$, for some $r > 1$ and $p \in [1, 2]$, with the optimal rate achieved at $\lambda \asymp n^{-r/(pr+1)}$; see Theorem 9.17. For the minimizer in $\mathcal{H}^*$ to share the same convergence rate, one needs $q\lambda^{2/r} \to \infty$ (Theorem 9.20), hence it suffices to have $q \asymp n^{2/(pr+1)+\epsilon}$, $\forall \epsilon > 0$. For $J(f) = \int_0^1 \ddot{f}^2 dx$ on $\mathcal{X} = [0, 1]$, $r = 4$ (Example 9.1), $p = 1$ when $\ddot{\eta}^2$

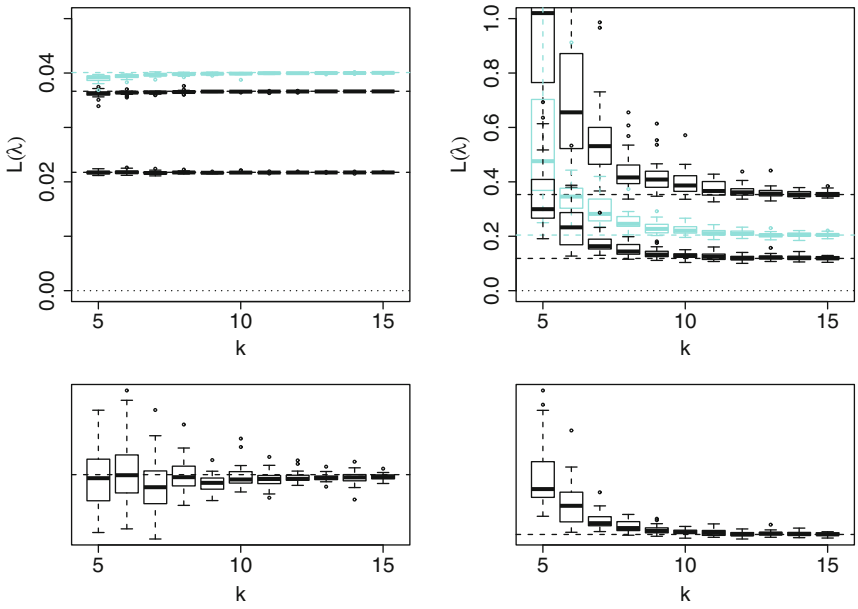


FIGURE 3.3. Effect of q on estimation consistency. Boxplots of $L(\lambda)$ with 30 different random subsets $\{z_j\} \subset \{x_i\}$ of size $q = kn^{2/9}$. *Left*: Cubic spline fits to three simulated samples. *Right*: Tensor product cubic spline fits to three simulated samples. *Top*: $n = 100, 500$ in *solid*, from high to low, and $n = 300$ in *faded*. *Bottom*: $n = 500$ with better resolution. The *dashed lines* correspond to $q = n$.

is "barely" integrable, and $p = 2$ if $\eta^{(4)}$ is square integrable; for tensor product cubic splines, the rate holds for any $r < 4$ (Example 9.2). Setting $r = 4$, $p = 2$, and $\epsilon = 0$, one may use $q \propto n^{2/9}$ in practice.

Samples of sizes $n = 100, 300, 500$ were generated from $Y_i = \eta(x_i) + \epsilon_i$, $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$, where

$$\eta(x) = 1 + 3\sin(2\pi x - \pi)$$

and $\epsilon_i \sim N(0, 1)$. For each of the three samples and every k on the grid $k = 5(1)15$, thirty different random subsets $\{z_j\} \subset \{x_i\}$ of size $q = kn^{2/9}$ were generated, and cubic splines were fitted with the smoothing parameter minimizing $V(\lambda)$ of (3.27) with $\alpha = 1.4$. The fits with $q = n$ were also calculated. The loss $L(\lambda)$ of (3.13) was recorded for all the fits and the results are summarized in the left frames of Fig. 3.3 in box plots. The experiments were repeated on $\mathcal{X} = [0, 1]^2$ using tensor product cubic splines, with

$$\begin{aligned}\eta(x) = 5 &+ \exp(3x_{\langle 1\rangle}) + 10^6 x_{\langle 2\rangle}^{11}(1 - x_{\langle 2\rangle})^6 \\ &+ 10^4 x_{\langle 2\rangle}^3(1 - x_{\langle 2\rangle})^{10} + 5\cos\big(2\pi(x_{\langle 1\rangle} - x_{\langle 2\rangle})\big),\end{aligned}$$

TABLE 3.2. Quantiles of $\left|\tilde{\eta}(x_i)-\hat{\eta}(x_i)\right|/\sqrt{L}$ and $\left|\log\left(s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)\right)\right|$ in univariate simulation: $n = 100, 300$.

		50 %	75 %	90 %	95 %	99 %	100 %
$\lvert\tilde{\eta}-\hat{\eta}\rvert/e$:	$n = 100$	0.005	0.011	0.021	0.031	0.079	1.551
	$n = 300$	0.005	0.010	0.018	0.025	0.047	0.212
$\lvert\log(s_{\tilde{\eta}}/s_{\hat{\eta}})\rvert$:	$n = 100$	0.002	0.004	0.011	0.016	0.028	0.088
	$n = 300$	0.001	0.004	0.010	0.016	0.028	0.063

$x_i \sim U(0,1)^2$, and $\epsilon_i \sim N(0, 3^2)$; corresponding results are summarized in the right frames of Fig. 3.3. The bivariate results demonstrate much more variability, likely due to the five smoothing parameters involved; also note that the same loss $L(\lambda)$ could be achieved by different sets of θ_β's, so the variability in the actual fits could be greater. The fact that the box width gradually decreases as k increases indicates that $q \asymp n^{2/9}$ is the "correct" scale, and a k around 10 appears to deliver stable enough results for practical use.

3.5.5 Numerical Accuracy

For $q = n$, $RQ^+R^T = Q$, so all the formulas in §3.5.2 reduce to their respective counterparts in §§2.5 and 3.3. We now assess the numerical accuracy of quantities calculated with $q = 10n^{2/9}$ as approximations to those calculated with $q = n$.

Consider again the univariate simulation of §3.5.4 using cubic splines. For sample size $n = 100$, one hundred replicates were generated and cross-validated fits were calculated using $q = n$ and $V(\lambda)$ with $\alpha = 1.4$; posterior means $\hat{\eta}(x_i)$ and posterior standard deviations $s_{\hat{\eta}}(x_i)$ were calculated on the sampling points. For each of the replicates, ten different random subsets $\{z_j\} \subset \{x_i\}$ of size $q = 10n^{2/9}$ were used to calculate ten more cross-validated fits, with posterior means $\tilde{\eta}(x_i)$ and posterior standard deviations $s_{\tilde{\eta}}(x_i)$. The standardized differences $\left|\tilde{\eta}(x_i)-\hat{\eta}(x_i)\right|/\sqrt{L}$ in posterior mean and the log ratios $\left|\log\left(s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)\right)\right|$ in posterior standard deviation were recorded, where $L = e^2 = n^{-1}\sum_{i=1}^n \left(\hat{\eta}(x_i)-\eta(x_i)\right)^2$ was the mean square error loss of the fit with $q = n$. This yielded $100(10)(100) = 10^5$ entries of differences and log ratios. The experiment was repeated for sample size $n = 300$ on fifty replicates, yielding $50(10)(300) = 1.5 \times 10^5$ entries of differences and log ratios. These results are summarized in Table 3.2.

Fifty samples of size $n = 300$ were also generated from the bivariate simulation of §3.5.4 and sets of cross-validated tensor product cubic splines were fitted to the data. The differences $\left|\tilde{\eta}(x_i)-\hat{\eta}(x_i)\right|/\sqrt{L}$ and log ratios $\left|\log\left(s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)\right)\right|$ were calculated for the overall function

$$\eta(x) = \eta_\emptyset + \eta_1(x_{\langle 1\rangle}) + \eta_2(x_{\langle 2\rangle}) + \eta_{1,2}(x_{\langle 1\rangle}, x_{\langle 2\rangle})$$

TABLE 3.3. Quantiles of $\left|\tilde{\eta}(x_i)-\hat{\eta}(x_i)\right|/\sqrt{L}$ and $\left|\log\left(s_{\tilde{\eta}}(x_i)/s_{\hat{\eta}}(x_i)\right)\right|$ in bivariate simulation: $n = 300$.

		50 %	75 %	90 %	95 %	99 %	100 %
$\lvert\tilde{\eta}-\hat{\eta}\rvert/e$:	η	0.133	0.238	0.370	0.475	0.771	2.962
	η_1	0.053	0.098	0.161	0.213	0.351	1.267
	η_2	0.077	0.139	0.217	0.282	0.437	1.804
	$\eta_{1,2}$	0.111	0.198	0.307	0.397	0.674	2.745
$\lvert\log(s_{\tilde{\eta}}/s_{\hat{\eta}})\rvert$:	η	0.047	0.081	0.115	0.137	0.182	0.462
	η_1	0.068	0.111	0.160	0.199	0.315	0.591
	η_2	0.044	0.074	0.108	0.134	0.184	0.358
	$\eta_{1,2}$	0.079	0.118	0.159	0.188	0.262	0.735

as well as its ANOVA components η_1, η_2, and $\eta_{1,2}$; the mean square error L of $\hat{\eta}$ was calculated only for the overall function and the same divisor $\sqrt{L}$ was used to standardize the differences $\left|\tilde{\eta}(x_i) - \hat{\eta}(x_i)\right|$ in both the overall function and the ANOVA components. The results are summarized in Table 3.3. Were the same θ_β's used in the $\hat{\eta}$ and $\tilde{\eta}$ being compared, the numbers in Table 3.3 could be more in line with those in Table 3.2, but cross-validated smoothing parameters are part of the whole package. The overall consistency appears to be reasonable.

3.6 Software

To facilitate data analysis by practitioners, most of the techniques presented throughout this book have been implemented in open-source software. Code for regression is available in collections of FORTRAN compatible routines and in suites of functions in an R package.

3.6.1 RKPACK

The algorithms of §3.4 have been implemented in a collection of public domain RATFOR (Rational FORTRAN (Kernighan 1975)) routines collectively known as RKPACK, first released in 1989 (Gu 1989). Routines from public domain linear algebra libraries BLAS and LINPACK have been used extensively in RKPACK routines as building blocks; see Dongarra et al. (1979) for descriptions of BLAS and LINPACK. The user interface of RKPACK is through four routines, `dsidr`, `dmudr`, `dsms`, and `dcrdr`, which implement Algorithms 3.1 and 3.2, (3.60) and (3.59), respectively. A few sample application programs in RATFOR are also included in the package. RKPACK has been deposited to Netlib and StatLib. The latest version can be found at

`http://www.stat.purdue.edu/~chong/software.html`

RATFOR is a dialect of FORTRAN with a structural syntax similar to that of the S language (Becker et al. 1988). Most UNIX systems understand RATFOR. In compilation, RATFOR routines are translated by a RATFOR preprocessor into standard FORTRAN routines, transparent to the user, which are then sent to the compiler. For those without access to a RATFOR preprocessor, the FORTRAN translation of the routines are included in the package, but in-line comments are lost in the translation.

3.6.2 R Package `gss`: `ssanova` and `ssanova0` Suites

R, an open-source environment for data analysis and graphics not unlike the S/Splus language (Becker et al. 1988, Chambers and Hastie 1992), has emerged in the past decade as the de facto standard platform for statistical computing. R was originally created by Ihaka and Gentleman (1996), and is currently being developed and maintained by a core group of more than a dozen prominent statisticians/programmers stationed over several continents. R resources are archived on the Comprehensive R Archive Network (CRAN), with the master site at

`http://cran.r-project.org`

Add-on modules in R are known as packages, as in S/Splus, and at this writing, more than four thousands of R packages can be found on CRAN. The installations of R and add-on packages on all major operating systems are clearly explained in the R FAQ (Frequently Asked Questions on R) by Kurt Hornik (Hornik 2010), to be found on CRAN.

Suites of R functions implementing the methods presented in this book are collected in the R package `gss`, with the name abbreviated from *general smoothing splines*. The overall design of `gss` is outlined in Appendix A at the end of the book, and the basic usage of the suites is illustrated using simulated and real-data examples in the chapters and sections where the respective methods are developed.

For regression with Gaussian-type responses, one may use the `ssanova` or the `ssanova0` suites. The `ssanova0` suite is virtually the original `ssanova` suite referred to in the first edition of this book, serving as a front end to RKPACK which implements the algorithms of §3.4. The current `ssanova` suite implements the algorithms of §3.5.3 for the efficient approximation.

Some working knowledge is assumed of the modeling facilities in R, which have syntax nearly identical to those in S/Splus; a good reference on the subject is Venables and Ripley (2002). The syntax of the `ssanova0` and `ssanova` suites is similar to that of the `lm` suite for linear models, as can be seen in the following examples.

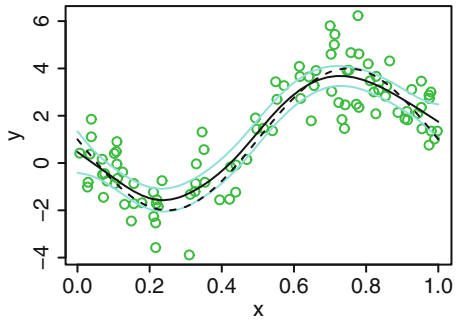


FIGURE 3.4. A cross-validated cubic spline fit. The fit is in the *solid line* and the 95 % Bayesian confidence intervals are in *faded lines*, with the test function superimposed in *dashed line* and the data in *circles*.

Example 3.1 (Cubic spline) Assume that the `gss` package is installed. At the R prompt, the following command loads the `gss` package into R:

```
library(gss)
```

The following sequence generates some synthetic data and fits a cubic spline to the data, with the smoothing parameter minimizing the GCV score $V(\lambda)$ of (3.27) with $\alpha = 1.4$:

```
set.seed(5732)
x <- runif(100)
y <- 1+3*sin(2*pi*x-pi)+rnorm(x)
fit.cubic <- ssanova(y~x,method="v",alpha=1.4)
```

The `set.seed` command resets the pseudo-random number generator so the reader could reproduce the reported results including figures. The default options `method="v"` and `alpha=1.4` are usually omitted. The results assigned to `fit.cubic` is a list object of class `"ssanova"`. To evaluate the fit on a grid for plotting purposes, one may try the following:

```
grid <- seq(0,1,len=51)
est <- predict(fit.cubic,data.frame(x=grid),se.fit=TRUE)
```

The flag `se.fit=TRUE` requests the calculation of the posterior standard deviation corresponding to the evaluated posterior mean; `est` is a list object consisting of elements `fit` (posterior mean) and `se.fit` (posterior standard deviation). Figure 3.4 displays a plot with the data, the test function, the cross-validated fit, and the 95 % Bayesian confidence intervals, which can be produced by the following commands:

```
plot(x,y,col=3); lines(grid,est$fit)
lines(grid,est$fit+1.96*est$se.fit,col=5)
lines(grid,est$fit-1.96*est$se.fit,col=5)
lines(grid,1+3*sin(2*pi*grid-pi),lty=2)
```

By default, `ssanova` uses a random subset $\{z_j\} \subset \{x_i\}$ of size $q \approx 10n^{2/9}$, so multiple calls with the same `x` and `y` would return slightly different fits barring resettings of the seed in between calls. One may reset the seed within `ssanova` via an optional argument `seed`, and one may pass the same selection of $\{z_j\}$ from `fit0` to `fit1` through

```
fit1 <- ssanova(...,id.basis=fit0$id.basis)
```

To override the default $q \approx 10n^{2/9}$, one may use `nbasis=q`. □

Example 3.2 (Tensor product cubic spline) The following sequence generates some synthetic data and fits a tensor product cubic spline to the data, with the smoothing parameters minimizing the unmodified GCV score $V(\lambda)$ of (3.23):

```
set.seed(5732)
x1 <- runif(100); x2 <- runif(100)
y <- 5 + exp(3*x1)+10^6*x2^11*(1-x2)^6+
     10^4*x2^3*(1-x2)^10+5*cos(2*pi*(x1-x2))+
     3*rnorm(x1)
mtype=list("cubic",c(0,1))
fit.tpcubic <- ssanova0(y~x1*x2,type=list(x1=mtype,
                                          x2=mtype))
```

The default `method="v"` is omitted in the call and `alpha` is not an option for `ssanova0` as it only implements unmodified GCV. The marginal domains are explicitly specified here as $\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$ via the `type` argument, overriding the default which would be the data range extended by 5 % on both ends. The model has four terms, labeled `1`, `x1`, `x2`, and `x1:x2` representing $\eta_\emptyset$, η_1, η_2, and $\eta_{1,2}$, respectively. To evaluate the fit on a grid, one may try the following:

```
grid1 <- seq(0,1,length=51)
grid2 <- seq(0,1,length=51)
new <- data.frame(x1=rep(grid1,51),
                  x2=rep(grid2,rep(51,51)))
est <- predict(fit.tpcubic,newdata=new,se.fit=TRUE)
post.mean <- matrix(est$fit,51,51)
post.stdev <- matrix(est$se.fit,51,51)
```

Now, let us plot the contours of the posterior mean and the posterior standard deviation, with the data superimposed:

```
contour(grid1,grid2,post.mean,sub="GCV Fit")
points(x1,x2)
contour(grid1,grid2,post.stdev,sub="Standard Deviation")
points(x1,x2)
```

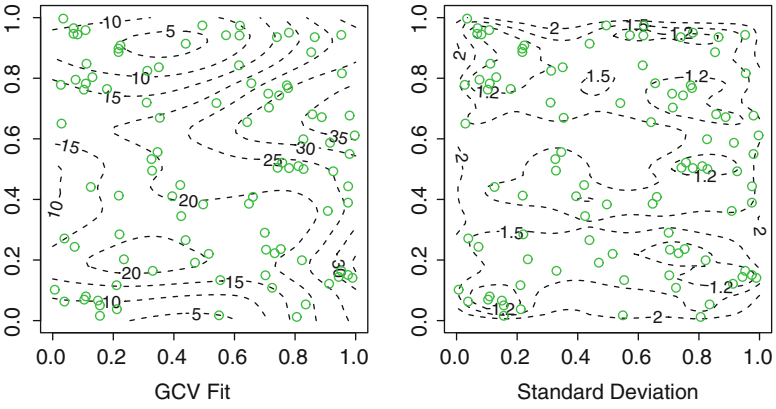


FIGURE 3.5. A cross-validated tensor product cubic spline fit. *Left*: Contours of the fit. *Right*: Contours of standard error. The data are superimposed as *circles*.

The plots are given in Fig. 3.5. The posterior standard deviation is rather flat away from the edges, slightly smaller where data are dense and larger where data are sparse. When sitting in front of a color monitor, one may want to replace `contour` by `filled.contour`.

By default, `predict` evaluates the overall function, but a partial sum of selected model terms can also be obtained via the specification of an optional argument `include`. For example, the following command returns the interaction on the grid:

```
est.int <- predict(fit.tpcubic,new,
                   se=TRUE,include="x1:x2")
```

One can now plot the contours of the interaction and compare with those of the overall function. □

As hinted by the `type` argument in the `ssanova0` call in Example 3.2, the margins of tensor product splines can be configured individually. The marginal domains for cubic splines can be arbitrary, either specified via `type` or extended from the data range by default, but they must contain all the observed data; the marginal domains are mapped onto $[0, 1]$ internally and the formulas of §§2.3.3 and 2.4.3 are used to calculate the reproducing kernels. As a consequence of such numerical treatment, any attempt to evaluate the fit beyond the domain will result in an error.

For n up to a thousand and probably beyond, the $O(n^3)$ algorithms of `ssanova0` often execute faster than the $O(nq^2) = O(n^{13/9})$ algorithms of `ssanova` for the default $q \approx 10n^{2/9}$, especially when multiple smoothing parameters are involved; Newton iterations using analytical derivatives are far more efficient than quasi-Newton iterations using numerical derivatives. The numerical efficiency of the algorithms in §3.4 rests with the special structure $R = Q$, which on the other hand severely restricts the scope of

their applicability. The algorithms of §3.5.3 however can be readily adapted to handle further modeling tools, such as the square error projection of §3.8 and the mixed-effect models of §6.2.

Incorporating the modified GCV score in `ssanova0` means opening up the legacy RKPACK routines for nontrivial modifications, an endeavor we chose not to pursue given the limited benefit. In theory, an `ssanova0` fit can be reproduced by `ssanova` with `id.basis=1:n` (so $\{z_j\} = \{x_i\}$) and `alpha=1`, which is indeed the case for the data of Example 3.2, but the different optimization algorithms used in `ssanova0` and `ssanova` may return different solutions when $V(\lambda)$ has a flat bottom. Also, `ssanova` has safeguards built in that automatically invoke some $\alpha \in (1, 3]$ to override $\alpha = 1$ when very small values of $n\lambda$ are searched upon, whereas `ssanova0` faithfully minimizes $V(\lambda)$ of (3.23) as defined; a quick check reveals that the `ssanova0` fit to the data of Example 3.1 is a case of severe undersmoothing, but the `ssanova` fits with various configurations of $\{z_j\}$ and α all look good.

3.7 Model Checking Tools

Two phases of statistical modeling are model fitting and model checking. For parametric models, model checking tools include diagnostics for the lack of fit, diagnostics for the identifiability of model terms such as the collinearity in linear models, and diagnostics for the practical significance of model terms through various tests. For nonparametric models, the lack of fit is no longer a main concern, but the danger of overfitting and overinterpreting makes the other two issues ever more important.

With respect to function decompositions such as the ANOVA decomposition of §1.3.2, we introduce some geometric diagnostics for the identifiability and the practical significance of the fitted terms. The use and effectiveness of the diagnostics are illustrated through simple simulations. Also presented are some heuristic arguments and related conceptual discussion concerning the diagnostics.

3.7.1 Cosine Diagnostics

Consider $\eta = \sum_{\beta=0}^{p} f_\beta$, where $f_0 \propto 1$ and f_β, $\beta > 0$ are terms in a function decomposition such as the ANOVA decomposition of §1.3.2. Evaluating a fit at the sampling points x_i, one obtains a retrospective linear model

$$\mathbf{Y} = \mathbf{f}_0 + \mathbf{f}_1 + \cdots + \mathbf{f}_p + \mathbf{e}, \tag{3.78}$$

where $\mathbf{f}_\beta = \big(f_\beta(x_1), \ldots, f_\beta(x_n)\big)^T$. Projecting (3.78) onto $\{\mathbf{1}\}^\perp = \big\{\mathbf{f} : \mathbf{f}^T\mathbf{1} = 0\big\}$ to remove the constant term, one gets

$$\mathbf{Y}^* = \mathbf{f}_1^* + \cdots + \mathbf{f}_p^* + \mathbf{e}^*. \tag{3.79}$$

The collinearity indices κ_β of $\left(\mathbf{f}_1^*, \ldots, \mathbf{f}_p^*\right)$ (Stewart 1987), which equal the square roots of the variance inflation factors, measure the identifiability of the f_β's in the fit. Denoting by C the $p \times p$ matrix with the (β, γ)th entry $\cos(\mathbf{f}_\beta^*, \mathbf{f}_\gamma^*)$, the κ_β^2's are given by the diagonals of C^{-1}. Write $\hat{\mathbf{Y}}^* = \mathbf{f}_1^* + \cdots + \mathbf{f}_p^*$. The scaled dot products $\pi_\beta = (\mathbf{f}_\beta^*)^T \hat{\mathbf{Y}}^* / \|\hat{\mathbf{Y}}^*\|^2$ provide a "decomposition" of unity, $\sum_{\beta=1}^p \pi_\beta = 1$, although π_β can be negative. When $\mathbf{f}_\beta^*$ are nearly orthogonal to each other, the π_β's come close to form a percentage decomposition of the sum of squares of $\hat{\mathbf{Y}}^*$ into those of its components.

The $\mathbf{f}_\beta^*$'s are supposed to predict the response $\mathbf{Y}^*$, so a near-orthogonal angle between an $\mathbf{f}_\beta^*$ and $\mathbf{Y}^*$ indicates a noise term. Signal terms should be reasonably orthogonal to the residuals, so a large cosine between an $\mathbf{f}_\beta^*$ and $\mathbf{e}^*$ makes a term suspect. Among informative measures for the signal-to-noise ratio are $\cos(\mathbf{Y}^*, \mathbf{e}^*)$ and $R^2 = \|\mathbf{Y}^* - \mathbf{e}^*\|^2 / \|\mathbf{Y}^*\|^2$. Finally, a very small Euclidean norm of an $\mathbf{f}_\beta^*$ as compared to that of $\mathbf{Y}^*$ also indicates a negligible term.

These geometric diagnostics will be collectively referred to as the cosine diagnostics, as they are largely based on the cosines among the vectors appearing in (3.79).

For weighted data, one may simply premultiply (3.78) by $W^{1/2}$, project the terms onto $\{W^{1/2}\mathbf{1}\}^\perp$, and operate from the resulting vectors. For replicated data, κ_β and π_β remain the same regardless of whether the retrospective linear model is based on (3.36) (unweighted) or (3.37) (weighted), but entities involving $\mathbf{Y}^*$ or $\mathbf{e}^*$ do vary; see Problem 3.20.

3.7.2 Examples

As illustrations of the use and effectiveness of the cosine diagnostics, we now analyze a few simple synthetic examples on $[0, 1]^3$ using the `ssanova0` and `ssanova` suites in `gss`.

Example 3.3 (Independent design) First, generate some synthetic data and fit a tensor product cubic spline:

```
set.seed(5732)
x1 <- runif(100); x2 <- runif(100); x3 <- runif(100)
y <- 10*sin(pi*x2)+exp(3*x3)+
     5*cos(2*pi*(x1-x2))+3*rnorm(x1)
fit <- ssanova(y~x1*x2*x3-x1:x2:x3)
```

The diagnostics for the fit can be obtained using the method `summary`:

```
sum.fit <- summary(fit,diagnostics=TRUE)
```

A look at the κ_β's confirms that there is no identifiability problem with this fit; the pound sign `#` is added in front of each line of the computer printout to distinguish it from the command one types in:

```
round(sum.fit$kappa,2)
#   x1    x2    x3 x1:x2 x1:x3 x2:x3
# 1.12  1.09  1.05  1.04  1.06  1.06
```

Given below are the π_β's, the cosines between $\mathbf{Y}^*$, $\mathbf{e}^*$ and the $\mathbf{f}^*_\beta$'s, and the norms of the vectors, where the `cos.y` line gives $\cos(\mathbf{Y}^*, \cdot)$ and the `cos.e` line gives $\cos(\mathbf{e}^*, \cdot)$:

```
round(sum.fit$pi,2)
#   x1    x2    x3 x1:x2 x1:x3 x2:x3
# 0.00  0.15  0.63  0.22  0.01 -0.01
round(sum.fit$cosines,2)
#         x1    x2    x3 x1:x2 x1:x3 x2:x3
# cos.y 0.03  0.43  0.79  0.46  0.14 -0.15
# cos.e 0.04  0.03  0.02  0.10  0.08  0.09
# norm  1.41 22.17 50.44 32.00  5.31  5.23
#        yhat     y     e
# cos.y  0.96  1.00  0.37
# cos.e  0.08  0.37  1.00
# norm  67.23 72.06 20.86
```

The terms `x1`, `x1:x3`, and `x2:x3` appear weak, both from the π_β's and from their weak correlations with the response. Eliminating `x1:x3` and `x2:x3` but keeping `x1` due to the presence of `x1:x2`, a new model is fitted to the data:

```
fit.new <- ssanova(y~x1*x2+x3,id.basis=fit$id.basis)
```

where for a more direct comparison we took care to specify via `id.basis` the same $\{z_j\}$ used in `fit`. A quick check shows that there is little meaningful change in the diagnostics associated with the remaining terms:

```
sum.new<-summary(fit.new,TRUE)
round(sum.new$pi,2)
#   x1    x2    x3 x1:x2
# 0.00  0.13  0.66  0.21
round(sum.new$cos,2)
#          x1    x2    x3 x1:x2  yhat     y    e
# cos.y -0.06  0.43  0.79  0.45  0.95  1.00  0.4
# cos.e  0.11  0.07  0.02  0.11  0.08  0.40  1.0
# norm   0.28 19.66 51.08 30.22 66.28 72.06 23.4
```

Results using `ssanova0` are similar. □

Example 3.4 (Simple aliasing design) Instead of an independent design, we now put $x_{i\langle 1\rangle}$ and $x_{i\langle 2\rangle}$ on a curve to create some identifiability problem:

```
set.seed(5732)
```

```
x2 <- runif(100); x3 <- runif(100)
x1 <- sqrt(x2)
y <- 10*sin(pi*x2)+exp(3*x3)+
     5*cos(2*pi*(x1-x2))+3*rnorm(x1)
```

Fitting a tensor product cubic spline using `ssanova0` and obtaining the diagnostics, one has:

```
fit <- ssanova0(y~x1*x2*x3-x1:x2:x3)
sum.fit <- summary(fit,TRUE)
round(sum.fit$kappa,2)
#    x1    x2    x3 x1:x2 x1:x3 x2:x3
# 27.31 28.92  3.33  5.20  7.91  7.88
round(sum.fit$pi,2)
#    x1    x2    x3 x1:x2 x1:x3 x2:x3
# -0.68  0.50  1.21  0.35 -0.21 -0.17
round(sum.fit$cos,2)
#          x1     x2    x3 x1:x2 x1:x3 x2:x3
# cos.y -0.14   0.09  0.85  0.27 -0.17 -0.15
# cos.e  0.00   0.00  0.02  0.00  0.01  0.00
# norm 293.38 332.23 87.53 79.07 72.77 70.03
#        yhat     y     e
# cos.y  0.95  1.00  0.34
# cos.e  0.04  0.34  1.00
# norm  64.60 68.63 20.48
```

The κ_β's indicate severe collinearity among the $\mathbf{f}_\beta^*$'s, and the large magnitude of `x1` coupled with its negative correlation with $\mathbf{Y}^*$ suggest that it provides no help in predicting the response but is merely offsetting other terms. Removing all terms involving `x1`, one has:

```
fit.new <- ssanova0(y~x2*x3)
sum.new <- summary(fit.new,TRUE)
round(sum.new$kappa,2)
#    x2    x3 x2:x3
#  1.02  1.01  1.02
round(sum.new$pi,2)
#    x2    x3 x2:x3
#  0.16  0.83  0.01
round(sum.new$cos,2)
#          x2    x3 x2:x3  yhat     y     e
# cos.y  0.38  0.85  0.27  0.95  1.00  0.38
# cos.e  0.06  0.03  0.17  0.06  0.38  1.00
# norm  26.25 58.37  3.81 63.73 68.63 21.69
```

The results are cleaned out, though the term `x2:x3` could also be removed due to the high $\cos(\mathbf{e}^*, \mathbf{f}_\beta^*)$ relative to $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*)$ and the very small κ_β. □

Example 3.5 (Complex aliasing design) We now change the aliasing pattern to $x_{i\langle 1\rangle} = (x^2_{i\langle 2\rangle} + x^2_{i\langle 3\rangle})/2$ and obtain a tensor product cubic spline fit and its diagnostics:

```
set.seed(5732)
x2 <- runif(100); x3 <- runif(100)
x1 <- (x2^2+x3^2)/2
y <- 10*sin(pi*x2)+exp(3*x3)+
     5*cos(2*pi*(x1-x2))+3*rnorm(x1)
fit <- ssanova(y~x1*x2*x3-x1:x2:x3)
sum.fit <- summary(fit,TRUE)
round(sum.fit$kappa,2)
#    x1    x2    x3 x1:x2 x1:x3 x2:x3
# 10.68  8.65  8.47  2.95  2.46  3.69
round(sum.fit$pi,2)
#    x1    x2    x3 x1:x2 x1:x3 x2:x3
# -1.43 -0.84  3.43 -0.02 -0.04 -0.11
round(sum.fit$cosines,2)
#           x1      x2     x3 x1:x2 x1:x3 x2:x3
# cos.y  -0.32  -0.26   0.78 -0.03 -0.04 -0.11
# cos.e   0.00   0.00   0.01  0.00  0.00  0.03
# norm  359.73 263.35 356.97 49.78 78.50 76.65
#         yhat     y     e
# cos.y   0.96  1.00  0.32
# cos.e   0.05  0.32  1.00
# norm   85.09 89.87 25.13
```

The situation is similar to that in Example 3.4 but we now have both `x1` and `x2` offending. The `x1` term plays a bigger role and it could be twisting perceptions concerning other terms, so we first take out terms involving `x1` and check the results:

```
fit.new <- ssanova(y~x2*x3,id.basis=fit$id.basis)
sum.new <- summary(fit.new,TRUE)
round(sum.new$kappa,2)
#   x2    x3 x2:x3
# 1.01  1.01  1.02
round(sum.new$pi,2)
#   x2    x3 x2:x3
# 0.22  0.71  0.07
round(sum.new$cosines,2)
#          x2    x3 x2:x3  yhat     y     e
# cos.y  0.50  0.79  0.26  0.96  1.00  0.35
# cos.e  0.07  0.02  0.09  0.07  0.35  1.00
# norm  36.44 71.44 24.42 84.28 89.87 25.70
```

The results are now clean, so no further action is needed. □

3.7.3 Concepts and Heuristics

We now briefly discuss the heuristics behind the cosine diagnostics and some related concepts. The primary issues are the identifiability of the f_β's, which the κ_β's are designed to diagnose, and the practical significance of individual terms, which the $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*)$'s are designed to diagnose.

First consider the identifiability. By construction, the decomposition $\eta = \sum_{\beta=0}^{p} f_\beta$ is well defined on its domain, say $\mathcal{X}$. When the function is being estimated from the data, however, information only comes from the sampling points $\mathcal{X}_0 = \{x_i\}_{i=1}^n$, and the identifiability of the terms in the decomposition depends on how well the decomposition is supported on the restricted domain $\mathcal{X}_0$. Parallel to collinearity, such an identifiability problem is called concurvity by Buja et al. (1989).

There exist two kinds of concurvity: the retrospective, or observed, concurvity, and the prospective concurvity. The observed concurvity can be defined as the collinearity of the restrictions of the estimated f_β's to $\mathcal{X}_0$, which the κ_β's are designed to diagnose. Prospective concurvity, the same in spirit as what was under discussion in Buja et al. (1989), is a (undesirable) property of the model and the design $\mathcal{X}_0$ based on preobservation analysis. For a parametric linear model, concurvity reduces to collinearity, the form of the fit is fully predictable from the model and the design, so there is no distinction between prospective and retrospective collinearity.

What is so bad about concurvity? One calculates an estimate $f = \sum_{\beta=0}^{p} f_\beta$ based on information from $\mathcal{X}_0$, but its restriction to $\mathcal{X}_0$, say $f^0 = \sum_{\beta=0}^{p} f_\beta^0$, is not well defined. If there is an alternative breakup $f^0 = \sum_{\beta=0}^{p} \alpha_\beta f_\beta^0$, then one could have used an alternative estimate $g = \sum_{\beta=0}^{p} \alpha_\beta f_\beta$ instead of $f = \sum_{\beta=0}^{p} f_\beta$. For this to be of serious concern to us, however, the difference $(\alpha_\beta - 1) f_\beta$ would have to be practically meaningful, and $J(f-g) = \sum_\beta (\alpha_\beta - 1)^2 J_\beta(f_\beta)$ would have to be negligible, where $J_\beta(f_\beta)$ is the roughness contribution of f_β to $J(f)$. This pretty much rules out the participation of "nonparametric" components in serious concurvity: For $(\alpha_\beta - 1) f_\beta$ to be practically significant, one must have negligible $J_\beta(f_\beta)$; hence, f_β would be primarily a parametric component in $\mathcal{N}_J$. The main concern of Buja et al. (1989), the numerical instability caused by concurvity to their back-fitting algorithm, is, however, not an issue here, as all terms are estimated simultaneously via the linear systems (3.4) or (3.63).

Now, consider the practical significance of individual terms. Recall that in a parametric regression model, insignificant terms are often detected using various F-statistics. Consider a linear model $\mathbf{Y} = \alpha \mathbf{1} + \beta \mathbf{x} + \boldsymbol{\epsilon}$, where $\mathbf{1}^T \mathbf{x} = 0$; if $\mathbf{1}^T \mathbf{x} \neq 0$, replace $\mathbf{x}$ by $(I - \mathbf{1}\mathbf{1}^T/n)\mathbf{x}$. Write $\mathbf{f}_0 = \hat{\alpha}\mathbf{1}$ and $\mathbf{f}_1 = \hat{\beta}\mathbf{x} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}$. The standard F-statistic for testing $\beta = 0$, or $\mathbf{f}_1 = 0$, is

$$F = \frac{\mathbf{Y}^T \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}}{\mathbf{Y}^T(I - \mathbf{1}\mathbf{1}^T/n - \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T)\mathbf{Y}} = \frac{\cos^2(\mathbf{Y}^*, \mathbf{f}_1^*)}{1 - \cos^2(\mathbf{Y}^*, \mathbf{f}_1^*)}, \tag{3.80}$$

which is monotone in

$$\cos^2(\mathbf{Y}^*, \mathbf{f}_1^*) = \frac{\mathbf{Y}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y}}{\mathbf{Y}^T(I - \mathbf{1}\mathbf{1}^T/n)\mathbf{Y}}; \tag{3.81}$$

see Problem 3.21. Hence, $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*)$ coincide with the classical measures in a specific simple parametric setting.

We suggest that $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*)$ be taken as absolute measures when the smoothing parameters are selected using a data-adaptive criterion such as $V(\lambda)$, for, in such a circumstance, different terms are allowed to compete with each other and with the residual term for shares of resources based on their qualifications as predictors of $\mathbf{Y}$. These diagnostics are objective quantities, but their calibration has to be subjective in lack of sampling distributions. Our limited experience seems to suggest that a term with $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*) > 0.4$ shall not be overlooked and a term with $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*) < 0.25$ may be safely suppressed. The calibration of $\|\mathbf{f}_\beta\|$ (an analog of χ^2-statistics) is much more difficult, so their use is limited and is of secondary importance. The π_β's provide reasonable measures for the relative strengths of the fitted terms, especially when the terms $\mathbf{f}_\beta^*$ are nearly orthogonal.

3.8 Square Error Projection

Consider a testing problem $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, where the notation is not to be confused with that in §3.1. For an example, $\mathcal{H}_0$ could be an additive model in an ANOVA decomposition involving only main effects, with $\mathcal{H}_1$ containing interaction terms. Lacking sampling distributions with an infinite dimensional $\mathcal{H}_0$, we now develop a geometric diagnostic for the practical significance of $\mathcal{H}_1$.

Denote by $\hat{\eta}$ an estimate of η in $\mathcal{H}_0 \oplus \mathcal{H}_1$. Minimizing

$$\mathrm{SE}(\hat{\eta}, \eta) = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{\eta}(x_i) - \eta(x_i)\big)^2 \tag{3.82}$$

with respect to $\eta \in \mathcal{H}_0$, one obtains a square error projection of $\hat{\eta}$ in $\mathcal{H}_0$, to be denoted by $\tilde{\eta}$. Suppose $\mathrm{span}\{1\} \subset \mathcal{H}_0$ and write $\eta_c = \bar{Y}$ the constant fit. One has a square error decomposition (Problem 3.22)

$$\mathrm{SE}(\hat{\eta}, \eta_c) = \mathrm{SE}(\hat{\eta}, \tilde{\eta}) + \mathrm{SE}(\tilde{\eta}, \eta_c). \tag{3.83}$$

When the ratio $\rho = \mathrm{SE}(\hat{\eta}, \tilde{\eta})/\mathrm{SE}(\hat{\eta}, \eta_c)$ is small, one loses little by cutting out $\mathcal{H}_1$. Note that this process does not involve the estimation of η in $\mathcal{H}_0$, which shall take place after H_0 is concluded.

The minimization of (3.82) in an infinite dimensional space is ill-posed, so the above procedure has to be regulated. Calculating $\hat{\eta}$ following the

approach of §3.5, $\tilde{\eta}$ can be set in a form similar to (3.61) but with basis $\phi_\nu(x) \in \mathcal{H}_1$ removed and with components $\theta_\beta R_\beta(z_j, x) \in \mathcal{H}_1$ trimmed from $R_J(z_j, x)$. The computation can be done via a modified (3.63), with a possibly skinnier S, fewer hidden components in R, $\mathbf{Y}$ replaced by $\hat{\eta}(\mathbf{x})$, and $n\lambda = 0$; such projection is well-posed for $q = o(n)$, but as a safeguard we use a small but positive $n\lambda$. One may also allow the remaining θ_β's in $R_J(z_j, x)$ to vary to bring $\mathrm{SE}(\hat{\eta}, \tilde{\eta})$ further down, though iterations for this process often stalls. Such square error projection is implemented in the `ssanova` suite.

For the data in Example 3.3, one may try:

```
fit <- ssanova(y~x1*x2*x3-x1:x2:x3)
project(fit,include=c("x1","x2","x1:x2","x3"))
```

where `project` returns a list object with elements `ratio` ($\rho = 0.0072$), `kl` ($\mathrm{SE}(\hat{\eta}, \tilde{\eta}) = 0.33$), and `check` ($\rho + \mathrm{SE}(\tilde{\eta}, \eta_c)/\mathrm{SE}(\hat{\eta}, \eta_c) = 0.999995$); the use of a positive $n\lambda$ breaks (3.83) and `check` monitors by how much it is off.

For the data in Example 3.5, one may similarly perform:

```
fit <- ssanova(y~x1*x2*x3-x1:x2:x3)
project(fit,include=c("x2","x3","x2:x3"))
```

This returns $\rho = 0.055$ and a `check` value 0.99998. The procedure is designed to diagnose the practical significance of $\mathcal{H}_1$ assuming $\mathcal{H}_0 \oplus \mathcal{H}_1$ is well defined, but the concurvity in the given data threw things off a bit.

To perceive such a geometric inferential tool in contrast to the classical hypothesis testing, consider a standard linear model

$$\mathbf{Y} = \mathbf{1}\beta_0 + X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

with a null $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. One has

$$\mathrm{SE}(\tilde{\eta}, \eta_c) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - \bar{Y})^2, \quad \mathrm{SE}(\hat{\eta}, \eta_c) = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2,$$

where $\tilde{\mathbf{Y}} = \tilde{X}_1(\tilde{X}_1^T\tilde{X}_1)^{-1}\tilde{X}_1^T\mathbf{Y}$, $\hat{\mathbf{Y}} = \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\mathbf{Y}$ for $\tilde{X}_1 = (\mathbf{1}, X_1)$, $\tilde{X} = (\mathbf{1}, X_1, X_2)$. It follows that

$$\rho = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \tilde{Y}_i)^2}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2} = \frac{\mathrm{SSR}(X_2|X_1)}{\mathrm{SSR}(X_1, X_2)}$$

with X_1 and X_2 indicating groups of predictors; note that neither the variance of ϵ nor the sample size is referenced here. If $\rho = 0.02$, one may well feel comfortable to settle with $\boldsymbol{\beta}_2 = \mathbf{0}$, although $\boldsymbol{\beta}_2$ could be statistically significant due to a small error variance or a large sample size. On the other hand, a $\rho = 0.10$ as the sole clue would likely keep $\boldsymbol{\beta}_2$ in the model, but $\boldsymbol{\beta}_2$ could be statistically insignificant with a large error variance or a small sample size.

3.9 Case Studies

We now apply the techniques developed so far to analyze a few real data sets. As with all data analysis exercises, subjective choices will have to be made along the way, and the author's preferences by no means represent the only "correct" solutions.

3.9.1 Nitrogen Oxides in Engine Exhaust

In an experiment reported by Brinkman (1981), a single-cylinder engine was run with ethanol to see how the NO_x concentration in the exhaust depended on the compression ratio and the equivalence ratio. There were 88 measurements made, and the data were analyzed by Cleveland and Devlin (1988) and Breiman (1991), among others, using other smoothing methods.

The data are included in `gss` as a data frame `nox` with elements `nox`, `comp`, and `equi`. A tensor product cubic spline was fitted to the data and the diagnostics obtained:

```
data(nox); set.seed(5732)
fit.nox <- ssanova(log(nox)~comp*equi,data=nox)
sum.nox <- summary(fit.nox,TRUE)
round(sum.nox$kappa,2)
#     comp      equi comp:equi
#     1.08      1.05      1.04
round(sum.nox$pi,2)
#     comp      equi comp:equi
#    -0.02      1.01      0.01
round(sum.nox$cos,2)
#        comp  equi comp:equi  yhat     y    e
# cos.y -0.08  0.95      0.07  0.98  1.00 0.23
# cos.e  0.02  0.04      0.03  0.06  0.23 1.00
# norm   4.23 19.09      3.48 18.36 18.83 3.29
project(fit.nox,"equi")$ratio
# 0.02151077
```

The `set.seed` command ensures a reproducible $\{z_j\}$. The NO_x concentrations are positive with some near-zero readings, so a log transform was applied. The effect of equivalence ratio was dominant, but the compression ratio had little impact. Eliminating terms involving `comp`, one can fit a cubic spline in `equi` and plot the data, the fit, and the 95% Bayesian confidence intervals, as in Fig. 3.6:

```
set.seed(5732)
fit.nox <- ssanova(log(nox)~equi,data=nox)
grid <- sort(nox$equi)
```

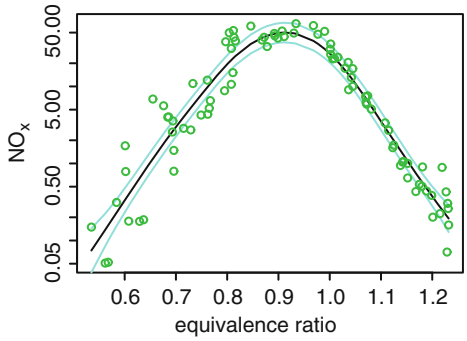


FIGURE 3.6. A cubic spline fit to NO_x data. The fit is in the *solid line*, the 95 % Bayesian confidence intervals in *faded lines*, and the data in *circles*.

```
est <- predict(fit.nox,data.frame(equi=grid),se=TRUE)
plot(nox$equi,nox$nox,log="y",xlab="equivalence ratio",
     ylab=expression(NO[x]),col=3)
lines(grid,exp(est$fit))
lines(grid,exp(est$fit+1.96*est$se),col=5)
lines(grid,exp(est$fit-1.96*est$se),col=5)
```

The compression ratio had only five distinctive values, so it could have been treated as an ordinal discrete variable; it would not make a difference though, as `nox` is plain flat on the `comp` axis. Cleveland and Devlin (1988) and Breiman (1991) both used the cubic root transform for `nox` instead of the log transform; parallel analysis using the cubic root transform yields essentially the same results.

3.9.2 Ozone Concentration in Los Angeles Basin

Daily measurements of ozone concentration and eight meteorological quantities in the Los Angeles basin were recorded for 330 days of 1976. The data were used by Breiman and Friedman (1985) to illustrate their ACE algorithm (alternating conditional expectation) and by Buja et al. (1989) to illustrate nonparametric additive models through the back-fitting algorithm. The data are included in `gss` as a data frame `ozone` with the following elements:

`upo3` Upland ozone concentration (ppm).

`vdht` Vandenberg 500 millibar height (m).

`wdsp` Wind speed (mph).

`hmdt` Humidity (%).

`sbtp` Sandburg Air Base temperature (^oC).

`ibht` Inversion base height (ft).

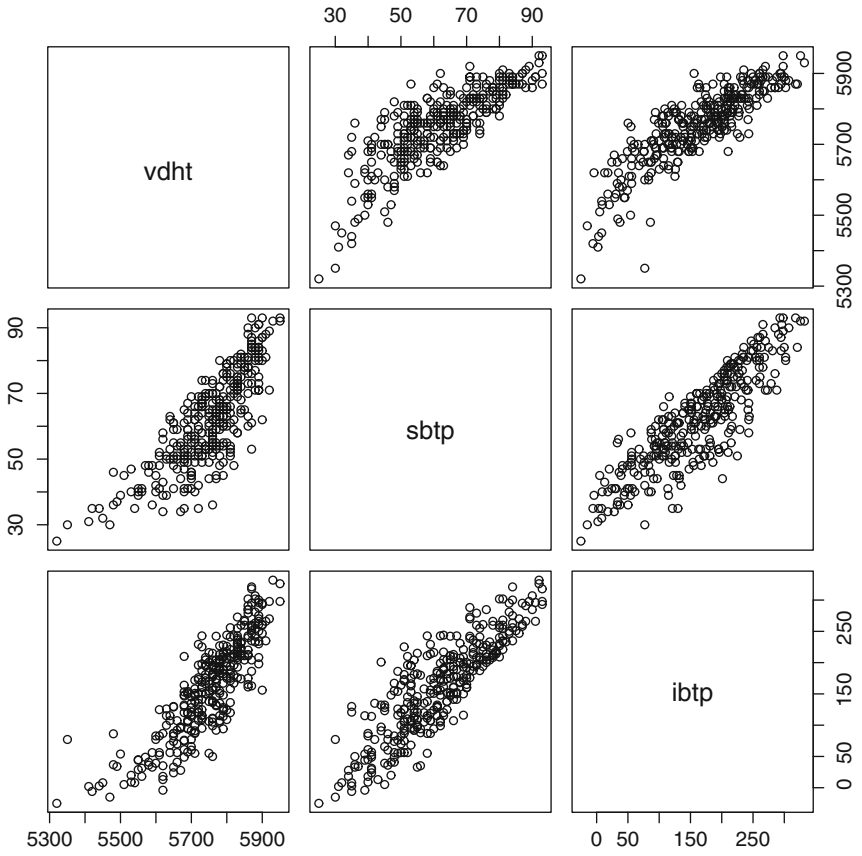


FIGURE 3.7. Scatter plot matrix of ozone data: A correlated group.

`dgpg` Dagget pressure gradient (mmHg).

`ibtp` Inversion base temperature (^oF).

`vsty` Visibility (miles).

From the scatter plot matrix, the three variables `vdht`, `sbtp`, and `ibtp` appeared to be highly correlated; see Fig. 3.7. We decided not to include these variables simultaneously in our preliminary analysis. We also decided not to include the variable `wdsp`, which showed little relation with any of the other variables.

Our first attempt was to fit tensor product cubic splines on five variables: one of `vdht`, `sbtp`, or `ibtp`, plus four others, `hmdt`, `ibht`, `dgpg`, and `vsty`. Included in the models were five main effects and ten pairwise interactions. The log transform was applied to the response since it is positive with some

readings near zero. The measure $R^2 = \|\mathbf{Y}^* - \mathbf{e}^*\|^2/\|\mathbf{Y}^*\|^2$ was calculated to be 0.750, 0.776, and 0.770 for the three fits. We now proceed with fits involving `sbtp`:

```
data(ozone); set.seed(5732)
fit.oz0 <- ssanova(log10(upo3)~
                     (sbtp+hmdt+ibht+dgpg+vsty)^2,
                   data=ozone)
sum.oz0 <- summary(fit.oz0,TRUE)
round(sum.oz0$kappa,2)
round(sum.oz0$cos,2)
```

The largest κ_β was 2.09, indicating modest concurvity. The interaction terms `sbtp:ibht`, `sbtp:dgpg`, `hmdt:ibht`, `hmdt:vsty`, and `ibht:dgpg` had $\cos(\mathbf{Y}^*, \mathbf{f}^*_\beta) \leq 0.02$, so we refit the model without these terms:

```
fit.oz1 <- ssanova(log10(upo3)~
                     (sbtp+hmdt+ibht+dgpg+vsty)^2
                     -(sbtp:ibht+sbtp:dgpg+hmdt:ibht
                       +hmdt:vsty+ibht:dgpg),
                   id.basis=fit.oz0$id,data=ozone)
sum.oz1 <- summary(fit.oz1,TRUE)
round(sum.oz1$pi,2)
round(sum.oz1$cos,2)
```

The terms `sbtp:hmdt`, `sbtp:vsty`, and `dgpg:vsty` had $\cos(\mathbf{Y}^*, \mathbf{f}^*_\beta) \leq 0.22$, and the main effect `hmdt` had $\cos(\mathbf{Y}^*, \mathbf{f}^*_\beta) = -0.43$. Eliminating the three interactions listed but keeping `hmdt` for now, we inspect the next fit:

```
fit.oz2 <- ssanova(log10(upo3)~sbtp+hmdt+ibht+dgpg+vsty
                                 +hmdt:dgpg+ibht:vsty,
                   id.basis=fit.oz0$id,data=ozone)
sum.oz2 <- summary(fit.oz2,TRUE)
round(sum.oz2$pi,2)
round(sum.oz2$cos,2)
```

The terms `hmdt` and `hmdt:dgpg` had $\cos(\mathbf{Y}^*, \mathbf{f}^*_\beta) = -0.43, 0.43$ and similar norms, apparently offsetting each other. Removing these two terms and adding back as main effects the previously excluded `vdht`, `ibtp`, and `wdsp` to double check their effects, one has:

```
fit.oz3 <- ssanova(log10(upo3)~sbtp+ibht+dgpg+vsty
                               +vdht+ibtp+wdsp+ibht:vsty,
                   id.basis=fit.oz0$id,data=ozone)
sum.oz3 <- summary(fit.oz3,TRUE)
round(sum.oz3$pi,2)
round(sum.oz3$cos,2)
```

The terms `vdht`, `wdsp`, and `ibht:vsty` had small $\cos(\mathbf{Y}^*, \mathbf{f}_\beta^*)$ or small norm or both. The square error projection into a five-term additive model yields:

```
project(fit.oz3,c("sbtp","ibht",
                  "dgpg","vsty","ibtp"))$ratio
# 0.009726355
```

We now fit the five-term additive model and check its diagnostics:

```
fit.oz4 <- ssanova(log10(upo3)~ibtp+sbtp+ibht+dgpg+vsty,
                   id.basis=fit.oz0$id,data=ozone)
sum.oz4 <- summary(fit.oz4,TRUE)
round(sum.oz4$kappa,2)
# ibtp sbtp ibht dgpg vsty
# 3.06 2.41 1.78 1.23 1.12
round(sum.oz4$pi,2)
# ibtp sbtp ibht dgpg vsty
# 0.10 0.52 0.20 0.11 0.07
round(sum.oz4$cos,2)
#       ibtp sbtp ibht dgpg vsty yhat    y    e
# cos.y 0.74 0.79 0.67 0.42 0.45 0.86 1.00 0.52
# cos.e 0.00 0.01 0.01 0.03 0.03 0.02 0.52 1.00
# norm  0.58 2.83 1.33 1.19 0.64 5.03 5.90 2.98
```

The concurvity between `ibtp` and `sbtp` is evident, and `vsty` appears weak. In fact, one has:

```
project(fit.oz4,c("sbtp","ibht","dgpg"))$ratio
# 0.01210439
project(fit.oz3,c("sbtp","ibht","dgpg"))$ratio
# 0.03033862
```

So one may also consider a three-term additive model:

```
fit.oz5 <- ssanova(log10(upo3)~sbtp+ibht+dgpg,
                   id.basis=fit.oz0$id,data=ozone)
sum.oz5 <- summary(fit.oz5,TRUE)
round(sum.oz5$kappa,2)
# sbtp ibht dgpg
# 1.22 1.21 1.05
round(sum.oz5$pi,2)
# sbtp ibht dgpg
# 0.62 0.27 0.10
round(sum.oz5$cos,2)
#       sbtp ibht dgpg yhat    y    e
# cos.y 0.79 0.66 0.43 0.86 1.00 0.53
# cos.e 0.01 0.01 0.04 0.02 0.53 1.00
# norm  3.36 1.75 1.08 5.00 5.90 3.05
```

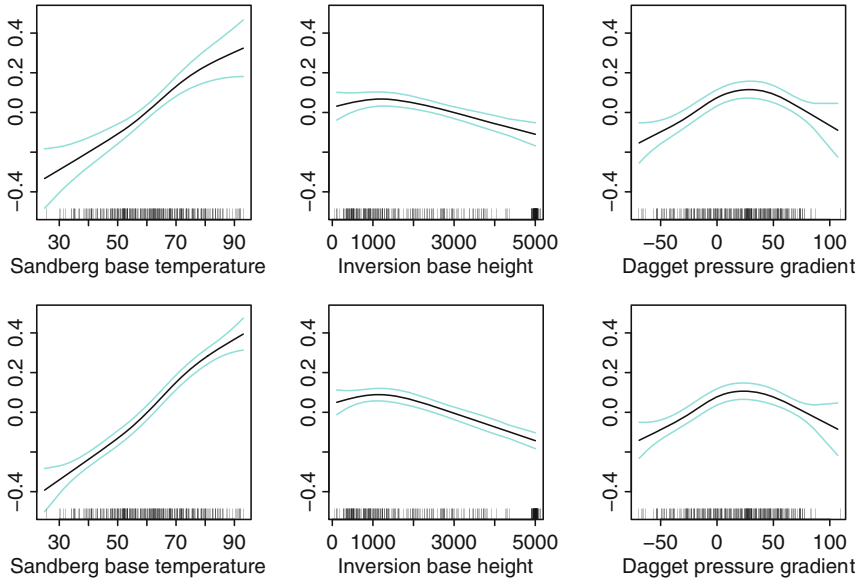


FIGURE 3.8. Three terms in additive cubic spline fits to ozone data. The fits are in *solid lines* and the 95 % Bayesian confidence intervals in *faded*. *Top*: `fit.oz4` with concurvity. *Bottom*: `fit.oz5` without concurvity. The rugs on the bottom in each frame mark the data points, slightly jittered.

The fits `fit.oz3`, `fit.oz4`, and `fit.oz5` have $R^2 = 0.749$, 0.729, and 0.719, respectively. To obtain a fitted term with standard errors on the data points, say the term `sbtp` in `fit.oz4`, one may use:

```
est4.sbtp <- predict(fit.oz4,ozone,inc="sbtp",se=TRUE)
```

Plotted in Fig. 3.8 are the terms `sbtp`, `ibht`, and `dgpg` in `fit.oz4` and `fit.oz5`, with the rugs on the bottom in each frame marking jittered data points. It is easily seen that `fit.oz4` has a slightly weaker `sbtp` effect with larger standard errors. The `sbtp` effect in `fit.oz5` is split between `sbtp` and `ibtp` in `fit.oz4`, with the concurvity causing identifiability problems.

3.10 Computation: Special Algorithms

The generic algorithms of §3.4 are of order $O(n^3)$ and those of §3.5.3 are of order $O(nq^2) = O(n^{13/9})$ with the default $q \asymp n^{2/9}$. For some problems, however, structures can be introduced through alternative formulations, yielding more scalable algorithms for calculations with fixed smoothing parameter. To select the smoothing parameter using $U(\lambda)$ or $V(\lambda)$, one needs algorithms of comparable speed for the evaluation of $\mathrm{tr}A(\lambda)$, which is

the focus of this section. According to current knowledge, the score $M(\lambda)$ is largely beyond reach with the alternative formulations, so are the posterior variances which one would need for the construction of Bayesian confidence intervals.

For polynomial splines on $[0,1]$, bandedness can be introduced into the matrices involved through the use of ordered local-support basis, and $O(n)$ algorithms are available for both $\hat{\mathbf{Y}}$ and $\mathrm{tr}A(\lambda)$ (§3.10.1). For problems such as tomographical reconstruction and the smoothing of digital images, one usually solves sparse or highly structured linear systems through iterative procedures, and the term $\mathrm{tr}A(\lambda)$ can be estimated through a parallel run with some $\mathbf{w} \sim N(0, I)$ replacing $\mathbf{Y}$ (§3.10.2).

3.10.1 Fast Algorithm for Polynomial Splines

A polynomial smoothing spline on $[0,1]$, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1 \big(\eta^{(m)}\big)^2 dx, \tag{3.84}$$

is called a natural spline in the numerical analysis literature. It is a piecewise polynomial of order $2m-1$, with up to the $(2m-2)$nd derivatives continuous and the $(2m-1)$st derivative jumping at the knots $\xi_1 < \cdots < \xi_q$, the ordered distinctive sampling points x_i. On $[0,\xi_1]$ and $[\xi_q,1]$, it is a polynomial of order $m-1$. See, e.g., de Boor (1978).

The natural splines with a given set of knots $\xi_1 < \cdots < \xi_q$ form a linear space of dimension q; see Problem 3.27. There exists a local-support basis $\{B_j(x), j = 1,\ldots,q\}$ for these natural splines, with each of the B_j's supported on at most $2m$ of the adjacent intervals $[0,\xi_1]$, $[\xi_1,\xi_2]$, $\ldots$, $[\xi_q,1]$, and at most $2m$ of the B_j's are nonzero at any $x \in [0,1]$; see Schumaker (1981, §8.2). Plugging the expression $\eta(x) = \sum_{j=1}^{q} c_j B_j(x)$ into (3.84), one has

$$(\mathbf{Y} - X\mathbf{c})^T(\mathbf{Y} - X\mathbf{c}) + n\lambda\mathbf{c}^T J\mathbf{c}, \tag{3.85}$$

where X is $n \times q$ with the (i,j)th entry $B_j(x_i)$ and J is $q \times q$ with the (i,j)th entry $\int_0^1 B_i^{(m)}B_j^{(m)}dx$. Minimizing (3.85) with respect to $\mathbf{c}$, one gets $\mathbf{c} = (X^TX + n\lambda J)^{-1}X^T\mathbf{Y}$ and $\hat{\mathbf{Y}} = X(X^TX + n\lambda J)^{-1}X^T\mathbf{Y}$.

Ordering the basis functions B_j increasingly by their supports, one has

$$B_i(x)B_j(x) = B_i^{(m)}(x)B_j^{(m)}(x) = 0$$

for $|i-j| \geq 2m$. It is clear that X^TX and J are both banded with bandwidth $4m-1$. The band Cholesky decomposition $(X^TX + n\lambda J) = C^TC$ takes $O(q)$ flops, with the upper-triangular C banded with bandwidth $2m$; see Golub and Van Loan (1989, §4.3.6). The coefficients $\mathbf{c}$ then are available in $O(q)$

extra flops through a band back substitution followed by a band forward substitution; see Golub and Van Loan (1989, §4.3.2).

The nontrivial part of the algorithm is the fast evaluation of

$$\mathrm{tr}A(\lambda) = \mathrm{tr}\{(X^TX + n\lambda J)^{-1}(X^TX)\}.$$

For $B = X^TX$ and $C^{-T} = (\mathbf{c}_1, \ldots, \mathbf{c}_q)$, one has $\mathrm{tr}A(\lambda) = \sum_{i,j} b_{i,j}\, \mathbf{c}_i^T\mathbf{c}_j$. Since B is symmetric and banded with bandwidth $4m-1$, only $\mathbf{c}_i^T\mathbf{c}_j$ for $0 \le i - j < 2m$ need to be computed. From $C^{-T}C^T = I$, one has

$$\mathbf{e}_i = \sum_{j=1}^{q} d_{i,j}\mathbf{c}_j = \sum_{j=i}^{q\wedge(i+2m-1)} d_{i,j}\mathbf{c}_j,$$

where $\mathbf{e}_i$ is the ith unit vector and $d_{i,j}$ is the (i,j)th entry of C with $d_{i,j} = 0$ for $j < i$ and $j \ge i + 2m$. Write $n(i) = q \wedge (i + 2m - 1)$. From

$$d_{i,i}\mathbf{c}_i = \mathbf{e}_i - \sum_{j=i+1}^{n(i)} d_{i.j}\mathbf{c}_j,$$

one has, recursively,

$$\begin{aligned}
\mathbf{c}_q^T\mathbf{c}_q &= d_q^{-2},\\
\mathbf{c}_i^T\mathbf{c}_k &= -d_{i,i}^{-1}\sum_{j=i+1}^{n(i)} d_{i,j}\mathbf{c}_j^T\mathbf{c}_k, \quad i < k,\\
\mathbf{c}_i^T\mathbf{c}_i &= d_{i,i}^{-2}\left(1 + \sum_{j=i+1}^{n(i)}\sum_{l=i+1}^{n(i)} d_{i,j}d_{i,l}\mathbf{c}_j^T\mathbf{c}_l\right),
\end{aligned} \tag{3.86}$$

where the fact that $\mathbf{e}_i^T\mathbf{c}_k = 0$ for $i < k$ is used. These formulas are immediate extensions of (3.58) on page 82. Using (3.86), one can fill $\mathbf{c}_i^T\mathbf{c}_j$ in the band $0 \le i - j < 2m$, from the bottom row up, backward within each row, without any reference to entries outside the band. The calculations take $O(q)$ flops.

The key to this algorithm is the band structure made available by the ordered local-support basis. Many authors use the popular B-spline basis as the $B_j(x)$ in the above formulation, which makes no difference in computation and performance, but, technically, B-splines are not natural splines, as they have different boundary conditions; see de Boor (1978) and Schumaker (1981) for details. The algorithm has been implemented for B-splines independently by Finbarr O'Sullivan and by H. J. Woltring, with code available from the NETLIB at `http://www.netlib.org/gcv`.

3.10.2 Iterative Algorithms and Monte Carlo Cross-Validation

Smoothing with a quadratic penalty is a special case of generalized ridge regression and can often be formulated in the form of (3.85) for some X. Fixing the smoothing parameter, one solves the linear system

$$(X^TX + n\lambda J)\mathbf{c} = X^T\mathbf{Y} \tag{3.87}$$

for $\mathbf{c}$ and calculates $\hat{\mathbf{Y}} = X\mathbf{c} = A(\lambda)\mathbf{Y}$ and $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \big(I - A(\lambda)\big)\mathbf{Y}$. In many applications, the matrix $X^TX + n\lambda J$ is sparse or highly structured, although not necessarily banded as in §3.10.1, which allows for the fast calculation of the matrix-vector multiplication $(X^TX + n\lambda J)\mathbf{c}$. Iterative procedures such as the conjugate gradient method are often the most efficient for solving such linear systems; see, e.g., Golub and Van Loan (1989, Chap. 10).

Examples of such formulation can be found in, e.g., Girard (1989). Detailed algorithmic specifications, which vary from problem to problem, are not directly relevant to our discussion. Our primary concern here is the implementation of automatic smoothing parameter selection through scores like $U(\lambda)$ or $V(\lambda)$ when iterative procedures are used to solve (3.87).

When the linear system (3.87) is solved iteratively, one has no direct access to the structure of the smoothing matrix $A(\lambda)$ and its trace. To use $U(\lambda)$ or $V(\lambda)$ for the selection of the smoothing parameter in such a circumstance, a Monte Carlo approximation of $\mathrm{tr}A(\lambda)$ was proposed by Girard (1989). The idea is simple and easy to implement. Let $\mathbf{w}$ be a vector of n independent standard normal deviates. Passing $\mathbf{w}$ through the same iterative procedures that produce $\hat{\mathbf{Y}} = A(\lambda)\mathbf{Y}$, one obtains $A(\lambda)\mathbf{w}$. One then can use $\mathbf{w}^TA(\lambda)\mathbf{w}$ to approximate $\mathrm{tr}A(\lambda)$ and select the smoothing parameter by minimizing

$$\tilde{U}(\lambda) = \frac{1}{n}\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y} + 2\frac{\sigma^2}{n}\mathbf{w}^TA(\lambda)\mathbf{w}$$

for σ^2 known, or by minimizing

$$\tilde{V}(\lambda) = \frac{n^{-1}\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y}}{\big\{1 - n^{-1}\mathbf{w}^TA(\lambda)\mathbf{w}\big\}^2}$$

for σ^2 unknown. The justification of the approximation is through the following theorem.

Theorem 3.9 *Assume independent noise ϵ_i with mean zero, a common variance σ^2, and uniformly bounded fourth moments. If Condition 3.2.1 of §3.2.1 holds, then*

$$\tilde{U}(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big). \tag{3.88}$$

If, in addition, Condition 3.2.2 of §3.2.2 also holds, then

$$\tilde{V}(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big). \tag{3.89}$$

Parallel to Theorems 3.1 and 3.3, this is poor man's justification. Results parallel to those of Li (1986) can be found in Girard (1991).

Proof of Theorem 3.9: Recalling (3.17) and (3.19) in the proof of Theorem 3.1, one has

$$\tilde{U}(\lambda) - U(\lambda) = 2\frac{\sigma^2}{n}\big(\mathbf{w}^T A(\lambda)\mathbf{w} - \mathrm{tr}A(\lambda)\big) = o_p\big(L(\lambda)\big),$$

which together with Theorem 3.1 yields (3.88). To prove (3.89) with Theorem 3.3 in mind, it suffices to show that $\tilde{V}(\lambda) - V(\lambda) = o_p\big(L(\lambda)\big)$. Write $\mu = n^{-1}\mathrm{tr}A(\lambda)$ and $\tilde{\mu} = n^{-1}\mathbf{w}^T A(\lambda)\mathbf{w}$. Simple algebra yields

$$\tilde{V}(\lambda) - V(\lambda) = V(\lambda)\left\{\frac{(1-\mu)^2}{(1-\tilde{\mu})^2} - 1\right\} = V(\lambda)\frac{2-\mu-\tilde{\mu}}{(1-\tilde{\mu})^2}(\tilde{\mu}-\mu),$$

which is $o_p\big(L(\lambda)\big)$ since $\tilde{\mu} - \mu = o_p\big(L(\lambda)\big)$, $V(\lambda) = O_p(1)$, and $\mu = o(1)$. This completes the proof. □

It is clear that each evaluation of $\tilde{U}(\lambda)$ or $\tilde{V}(\lambda)$ takes about twice as many flops as the calculation of $\hat{\mathbf{Y}}$ alone. In practice, it is advisable to generate a single $\mathbf{w}$ for use in $\tilde{U}(\lambda)$ or $\tilde{V}(\lambda)$ for all evaluations. One benefit of this is the continuity of the resulting score, and the other benefit is possible faster convergence of the iteration when $A(\lambda)\mathbf{w}$ at some nearby λ is used as the starting value. The approximation may be improved a little by averaging $\mathbf{w}^T A(\lambda)\mathbf{w}$ over a few replicates of $\mathbf{w}$ at further computational cost. Since n is usually very large when $\tilde{U}(\lambda)$ or $\tilde{V}(\lambda)$ is used, however, any benefit from such practice, if any, may not be worth the extra cost.

Compared to $\tilde{\mu} = n^{-1}\mathbf{w}^T A(\lambda)\mathbf{w}$, $\mu^* = \mathbf{w}^T A(\lambda)\mathbf{w}/\mathbf{w}^T\mathbf{w}$ provides a better estimator of $\mu = n^{-1}\mathrm{tr}A(\lambda)$ that one may use in practice; see Problem 3.28. Theorem 3.9 remains valid when $\tilde{\mu}$ is replaced by μ^*.

3.11 Bibliographic Notes

Section 3.1

The general problem of penalized least squares regression with multiple penalty terms was formulated by Wahba (1986) and studied numerically in Gu, Bates, Chen, and Wahba (1989) and Gu and Wahba (1991b). The linear system (3.4) as the basis for computation first appeared in Wahba and Wendelberger (1980). The smoothing matrix in the form of (3.7) was given by Wahba (1978).

Section 3.2

The score $U(\lambda)$, originally proposed by Mallows (1973) for use in ridge regression, is usually referred to as Mallows' C_L. Cross-validation is a classical technique for model selection in a variety of parametric and nonparametric problems. The generalized cross-validation score $V(\lambda)$ was due to Craven and Wahba (1979). Theorems 3.1 and 3.3 represent a step up from versions in the literature concerning expectations $R(\lambda)$, $E[U(\lambda)]$, and $E[V(\lambda)]$ but remain primitive compared to the results by Li (1986). The simple, direct proof of Theorem 3.3 is largely adapted from related arguments in Craven and Wahba (1979). The modified cross-validation score $V(\lambda)$ of (3.27) was explored in Kim and Gu (2004), following parallel development in density estimation (Gu and Wang 2003).

The score $M(\lambda)$ was proposed and studied in the context by Wahba (1985). Restricted maximum likelihood (REML) has been widely used in the literature on variance components and mixed-effect models; see, e.g., Harville (1977) and Robinson (1991). In Bayesian statistics, such an approach to the estimation of prior parameters is known as the type-II maximum likelihood; see, e.g., Berger (1985, §3.5.4).

The variance estimate $\hat{\sigma}_v^2$ was proposed by Wahba (1983) based on heuristic arguments and excellent simulation results. The motivation by equating λ_u and λ_v represents an alternative interpretation of the arguments developed in Gu, Heckman, and Wahba (1992) for smoothing parameter selection with replicated data. The primary result of Gu, Heckman, and Wahba (1992) was the calculus leading to (3.38)—(3.40).

Section 3.3

The Bayesian confidence intervals were proposed by Wahba (1983), with the across-the-function coverage property suggested through heuristic arguments and demonstrated via empirical simulations. A more rigorous treatment of the across-the-function coverage property for univariate polynomial splines can be found in Nychka (1988). The componentwise intervals derived through Theorem 3.8 were explored in Gu and Wahba (1993a).

Section 3.4

The developments in this section draw heavily on some standard numerical linear algebra results, for which Golub and Van Loan (1989) and Dongarra, Moler, Bunch, and Stewart (1979) are excellent references. Algorithm 3.1 was proposed by Gu, Bates, Chen, and Wahba (1989), with important ideas borrowed from earlier work by Elden (1984) and Bates, Lindstrom, Wahba, and Yandell (1987). Algorithms 3.2 and 3.3 were developed by Gu and Wahba (1991b), where further details are to be found.

Section 3.5

The materials in this section are largely taken from Kim and Gu (2004). The simulations are rerun, however, as the underlying code has gone through several updates since the original publication. The idea of efficient approximation first appeared in Gu and Kim (2002) and Gu and Wang (2003).

Section 3.6

RKPACK was first released to the public in 1989, with the two drivers `dsidr` and `dmudr` each having two options for smoothing parameter selection, $V(\lambda)$ or $M(\lambda)$. The option $U(\lambda)$ and the two utility routines `dcrdr` and `dsms` were added in 1992.

The R package `gss` was first released to the public in 1999. It was originally designed as a front end to RKPACK, but has since taken a life of its own with the addition of numerous suites implementing modeling tools beyond regression with independent data.

Section 3.7

An excellent review of diagnostics for collinearity can be found in Stewart (1987), where the collinearity indices are introduced. Earlier discussion of concurvity and its numerical ramifications can be found in Buja, Hastie, and Tibshirani (1989). This section draws heavily on materials from Gu (1992b), where more examples and further discussion are to be found. The values of κ_β^2 were mistakenly reported as κ_β in the examples of Gu (1992b), although the mistake was inconsequential.

Section 3.8

The materials in this section are largely taken from Gu (2004), where the more general Kullback-Leibler projection was proposed; the square error projection in Gaussian regression is a special case.

Section 3.9

In earlier analyses of the NO_x data, Cleveland and Devlin (1988) used multivariate local weighted regression and Breiman (1991) used his $\prod$ method, and both concluded that the interaction between the compression ratio and the equivalence ratio was significant. The analysis presented in §3.9.1 concludes otherwise.

In Breiman and Friedman (1985), an additive model in `sbtp`, `ibht`, `dgpg`, and `vsty` was fitted to the Los Angeles ozone data using alternating conditional expectation (ACE). Buja, Hastie, and Tibshirani (1989) used the data as a running example in the discussion of additive models and backfitting algorithm. A slew of analyses of the ozone data using a variety of

techniques were compared in Hastie and Tibshirani (1990, §10.3), where a scatter plot matrix of all the variables can be found.

Section 3.10

A comprehensive treatment of natural splines can be found in Schumaker (1981, Chap. 8). The $O(n)$ evaluation of $\mathrm{tr}A(\lambda)$ was proposed by Hutchinson and de Hoog (1985); see also O'Sullivan (1985). The distinction between the B-splines and the natural splines is discussed in de Boor (1978) and Schumaker (1981).

The Monte Carlo approximation of the trace term $\mathrm{tr}A(\lambda)$ was proposed by Girard (1989); see also Hutchinson (1989).

3.12 Problems

Section 3.1

3.1 Consider the least squares functional $L(f) = \sum_{i=1}^{n} \big(Y_i - f(x_i)\big)^2$ in a reproducing kernel Hilbert space $\mathcal{H}$ with a square seminorm $J(f)$.

(a) Prove that $L(f)$ is continuous, convex, and Fréchet differentiable.

(b) Let $\{\phi_\nu, \nu = 1, \ldots, m\}$ be a basis of $\mathcal{N}_J = \big\{f : J(f) = 0\big\}$ and S be $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$. Prove that if S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$.

(c) Prove that if S is of full column rank, then $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$.

3.2 Prove that the linear system

$$\begin{aligned}(Q + n\lambda I)\mathbf{c} + S\mathbf{d} &= \mathbf{Y},\\ S^T\mathbf{c} &= 0,\end{aligned}$$

where S is of full column rank, Q non-negative definite, and $\lambda > 0$, has a unique solution that satisfies

$$\begin{aligned}Q\big\{(Q + n\lambda I)\mathbf{c} + S\mathbf{d} - \mathbf{Y}\big\} &= 0,\\ S^T\{Q\mathbf{c} + S\mathbf{d} - \mathbf{Y}\} &= 0.\end{aligned}$$

3.3 Prove that the eigenvalues of the smoothing matrix $A(\lambda)$ as defined in (3.7) are all in the range $[0, 1]$.

3.4 Show that the solution of (3.10) minimizes

$$(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c})^T W(\mathbf{Y} - S\mathbf{d} - Q\mathbf{c}) + n\lambda\mathbf{c}^T Q\mathbf{c}.$$

Section 3.2

3.5 Prove Theorem 3.1 under the general moment conditions on ϵ_i as stated in the theorem.

(a) Let B and C be $n \times n$ matrices, where B is symmetric. Show that

$$\mathrm{Var}[\boldsymbol{\epsilon}^T B\boldsymbol{\epsilon}] \le 2\sigma^4 \mathrm{tr}B^2 + \sum_{i=1}^{n} b_{ii}^2(K - 3\sigma^4), \tag{3.90}$$

$$\mathrm{Var}[\boldsymbol{\eta}^T C\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{\eta}^T CC^T \boldsymbol{\eta}, \tag{3.91}$$

where K bounds $E[\epsilon_i^4]$ uniformly.

(b) Prove (3.17) by applying (3.90) with $B = A^2(\lambda)$ and applying (3.91) with $C = \big(I - A(\lambda)\big)A(\lambda)$. Note that the Cauchy-Schwarz inequality can be used to bound $\mathrm{Cov}[\boldsymbol{\epsilon}^T B\boldsymbol{\epsilon}, \boldsymbol{\eta}^T C\boldsymbol{\epsilon}]$.

(c) Prove (3.18) by applying (3.91) with $C = I - A(\lambda)$.

(d) Prove (3.19) by applying (3.90) with $B = A(\lambda)$.

3.6 Show that (3.28) is the minus log likelihood of $\mathbf{Z} = F_2^T \mathbf{Y}$.

3.7 Prove Theorem 3.5.

3.8 Consider replicated data $Y_{i,j} = \eta(x_i) + \epsilon_{i,j}$, where $j = 1, \ldots, w_i$, $i = 1, \ldots, n$. Denote the total sample size by $N = \sum_{i=1}^n w_i$ and the response vector of length N by $\tilde{\mathbf{Y}}$. Let S be $n \times m$ with entries $\phi_\nu(x_i)$, Q be $n \times n$ with entries $R_J(x_i, x_j)$, and $P = \mathrm{diag}(\mathbf{1}_{w_i})$ of size $N \times n$.

(a) Write $\bar{Y}_i = \sum_{j=1}^{w_i} Y_{i,j}/w_i$. Show that

$$\sum_{i=1}^{n}\sum_{j=1}^{w_i} \big(Y_{i,j} - \eta(x_i)\big)^2 = \sum_{i=1}^{n} w_i \big(\bar{Y}_i - \eta(x_i)\big)^2 + \sum_{i=1}^{n}\sum_{j=1}^{w_i} (Y_{i,j} - \bar{Y}_i)^2.$$

(b) Solving (3.36) directly through (3.3) with Y, S, Q replaced by $\tilde{Y}$, $\tilde{S}$, $\tilde{Q}$, respectively, verify that $\tilde{S} = PS$ and $\tilde{Q} = PQP^T$.

(c) Let $\mathbf{Y}_w$ be of length n with the ith entry $\sqrt{w_i}\bar{Y}_i$. Verify that $\mathbf{Y}_w = W^{-1/2}P^T\tilde{Y}$, where $W = P^TP = \mathrm{diag}(w_i)$.

(d) Consider F_2 orthogonal of size $n \times (n-m)$ satisfying $F_2^T W^{1/2} S = O$, and F_3 orthogonal of size $N \times (N-n)$ satisfying $F_3^T P = O$. Verify that $\tilde{F}_2 = (PW^{-1/2}F_2, F_3)$ is orthogonal and satisfies $\tilde{F}_2^T \tilde{S} = O$.

(e) The smoothing matrix for (3.36) is given by

$$\tilde{A}(\lambda) = I - n\lambda\tilde{F}_2(\tilde{F}_2^T\tilde{Q}\tilde{F}_2 + n\lambda I)^{-1}\tilde{F}_2^T,$$

and that for (3.37) is given by

$$A_w(\lambda) = I - n\lambda F_2(F_2^T W^{1/2} Q W^{1/2} F_2 + n\lambda I)^{-1} F_2^T;$$

see (3.7) and (3.11). Show that

$$I - \tilde{A}(\lambda) = PW^{-1/2}\big(I - A_w(\lambda)\big)W^{-1/2}P^T + F_3F_3^T.$$

Section 3.3

3.9 Prove Theorem 3.8. Similar to the proofs of Theorems 2.8 and 3.6, first consider independent proper priors for $\psi_\nu = d_\nu\phi_\nu$, $d_\nu \sim N(0, \tau^2)$, then let $\tau^2 \to \infty$.

(a) Find the covariance matrix of $\mathbf{Y}$, $\psi_\nu(x)$, and $\psi_\mu(x)$ and use it to prove (3.45) and (3.47).

(b) Find the covariance matrix of $\mathbf{Y}$, $\eta_\beta(x)$, and $\eta_\gamma(x)$ and use it to prove (3.46) and (3.49).

(c) Find the covariance matrix of $\mathbf{Y}$, ψ_ν and $\eta_\beta(x)$ and use it to prove (3.48).

3.10 Suppose $Y_i = \eta(x_i) + \epsilon_i$, where $\eta = \sum_{\nu=1}^4 \psi_\nu + \sum_{\beta=1}^5 f_\beta$ with fixed effects ψ_ν and random effects f_β, as in Theorem 3.8.

(a) Derive $E\big[\psi_3(x) + f_2(x)\big|\mathbf{Y}\big]$ and $b^{-1}\mathrm{Var}\big[\psi_3(x) + f_2(x)\big|\mathbf{Y}\big]$.

(b) Derive $E\big[\psi_4(x)+f_3(x)+f_4(x)+f_5(x)\big|\mathbf{Y}\big]$ and $b^{-1}\mathrm{Var}\big[\psi_4(x)+f_3(x)+f_4(x)+f_5(x)\big|\mathbf{Y}\big]$.

3.11 Derive the results of Theorems 3.6 and 3.8 for weighted data with $\epsilon_i \sim N(0, \sigma^2/w_i)$.

3.12 Verify that (3.51) simplifies to $n\lambda\, A(\lambda)$.

3.13 Show that for weighted data with weights w_i, $b^{-1}\mathrm{Var}\big[\eta(x_i)|\mathbf{Y}\big]$ is the (i,i)th entry of $n\lambda W^{-1/2}A_w(\lambda)W^{-1/2}$.

Section 3.4

3.14 For L lower-triangular, prove that L^{-1} is also lower-triangular.

3.15 For an invertible block matrix $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$, show that

$$M^{-1} = \begin{pmatrix} E^{-1} & -E^{-1}BD^{-1} \\ -D^{-1}CE^{-1} & D^{-1} + D^{-1}CE^{-1}BD^{-1} \end{pmatrix},$$

where $E = A - BD^{-1}C$.

Section 3.5

3.16 Verify that the $\mathbf{c}$ and $\mathbf{d}$ given in (3.66) solve (3.63).

3.17 For a square block matrix $M = \left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)$ with A invertible, show that $|M| = |A|\,|D - CA^{-1}B|$; premultiply M by $\left(\begin{smallmatrix} I & O \\ -CA^{-1} & I \end{smallmatrix}\right)$.

3.18 Verify (3.77).

3.19 Verify (3.76).

Section 3.7

3.20 Consider the replicated data of Problem 3.8 and keep all the notation and definitions. Write the retrospective linear model corresponding to (3.36) as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{f}}_0 + \tilde{\mathbf{f}}_1 + \cdots + \tilde{\mathbf{f}}_p + \tilde{\mathbf{e}} \tag{3.92}$$

and that corresponding to (3.37) as

$$\bar{\mathbf{Y}} = \mathbf{f}_0 + \mathbf{f}_1 + \cdots + \mathbf{f}_p + \mathbf{e}, \tag{3.93}$$

where $\bar{\mathbf{Y}} = W^{-1}P^T\tilde{\mathbf{Y}}$ has the ith entry $\bar{Y}_i$. It is easy to see that $\tilde{\mathbf{f}}_\beta = P\mathbf{f}_\beta$.

(a) Verify that $PW^{-1}P^T$ is a projection matrix and $I - PW^{-1}P^T = F_3F_3^T$.

(b) Show that $\tilde{\mathbf{Y}} = F_3F_3^T\tilde{\mathbf{Y}} + P\bar{\mathbf{Y}}$, $F_3F_3^T\tilde{\mathbf{Y}} = F_3F_3^T\tilde{\mathbf{e}}$, and $W^{-1}P^T\tilde{\mathbf{e}} = \mathbf{e}$.

(c) Projecting (3.92) onto $\{\mathbf{1}_N\}^\perp$, where the subscript N indicates the length of the vector, one gets $\tilde{\mathbf{Y}}^* = \tilde{\mathbf{f}}_1^* + \cdots + \tilde{\mathbf{f}}_p^* + \tilde{\mathbf{e}}^*$. Show that

$$\begin{aligned} \tilde{\mathbf{f}}_\beta^* &= P(I - \mathbf{1}_n\mathbf{1}_n^TW/N)\mathbf{f}_\beta, \\ \tilde{\mathbf{Y}}^* &= P(I - \mathbf{1}_n\mathbf{1}_n^TW/N)\bar{\mathbf{Y}} + F_3F_3^T\tilde{\mathbf{Y}}, \\ \tilde{\mathbf{e}}^* &= P(I - \mathbf{1}_n\mathbf{1}_n^TW/N)\mathbf{e} + F_3F_3^T\tilde{\mathbf{Y}}. \end{aligned}$$

(d) Verify that $I - W^{1/2}\mathbf{1}_n\mathbf{1}_n^T W^{1/2}/N$ is the projection matrix onto $\{W^{1/2}\mathbf{1}\}^\perp$.

(e) For $(\tilde{\mathbf{a}}, \mathbf{a}) = (\tilde{\mathbf{f}}_\gamma^*, \mathbf{f}_\gamma), (\tilde{\mathbf{Y}}^*, \bar{\mathbf{Y}}), (\tilde{\mathbf{e}}^*, \mathbf{e})$, show that

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{f}}_\beta^* = (W^{1/2}\mathbf{a})^T(I - W^{1/2}\mathbf{1}_n\mathbf{1}_n^T W^{1/2}/N)(W^{1/2}\mathbf{f}_\beta)$$

(f) For $(\tilde{\mathbf{a}}, \mathbf{a}), (\tilde{\mathbf{b}}, \mathbf{b}) = (\tilde{\mathbf{Y}}^*, \bar{\mathbf{Y}}), (\tilde{\mathbf{e}}^*, \mathbf{e})$, show that

$$\tilde{\mathbf{a}}^T\tilde{\mathbf{b}} = (W^{1/2}\mathbf{a})^T(I - W^{1/2}\mathbf{1}_n\mathbf{1}_n^T W^{1/2}/N)(W^{1/2}\mathbf{b}) + \tilde{\mathbf{Y}}^T F_3 F_3^T \tilde{\mathbf{Y}}.$$

3.21 Verify (3.80) and (3.81).

Section 3.8

3.22 Verify (3.83).

Section 3.9

3.23 Analyze the NO_x data of §3.9.1, with the cubic root of NO_x concentration as the response.

3.24 Analyze the NO_x data of §3.9.1, with the compression ratio treated as an ordinal factor; replace `comp` by `ordered(comp)` in `nox`.

3.25 Consider the ozone data of §3.9.2.

(a) Fit a tensor product cubic spline in the variables `vdht`, `hmdt`, `ibht`, `dgpg`, and `vsty`, with all pairwise interactions included.

(b) Simplify the model with the help of cosine diagnostics and/or square error projection; iterate the process if necessary.

(c) Obtain selected main effects from the final model and compare with those illustrated in Fig. 3.8.

3.26 Consider the ozone data of §3.9.2.

(a) Fit a cubic spline additive model in all variables.

(b) Simplify the model with the help of cosine diagnostics and/or square error projection; iterate the process if necessary.

(c) Obtain selected main effects from the final model and compare with those illustrated in Fig. 3.8.

Section 3.10

3.27 Given a set of knots $0 < \xi_1 < \cdots < \xi_q < 1$, a natural spline is a piecewise polynomial of order $2m-1$ on $[\xi_1, \xi_q]$, $m-1$ on $[0, \xi_1]$ and $[\xi_q, 1]$, with up to the $(2m-2)$nd derivatives continuous and the $(2m-1)$st derivative jumping at the knots. Verify that a natural spline has q free parameters.

3.28 Prove the inequality $E[\mu^* - \mu]^2 < E[\tilde{\mu} - \mu]^2$, where $\mu = n^{-1}\mathrm{tr}A(\lambda)$, $\mu^* = \mathbf{w}^T A(\lambda)\mathbf{w}/\mathbf{w}^T\mathbf{w}$, and $\tilde{\mu} = n^{-1}\mathbf{w}^T A(\lambda)\mathbf{w}$, for $\mathbf{w} \sim N(0, I)$.

(a) Show that without loss of generality, one may assume $A(\lambda)$ to be diagonal.

(b) Show that $\mathbf{w}/\sqrt{\mathbf{w}^T\mathbf{w}}$ and $\mathbf{w}^T\mathbf{w}$ are independent.

(c) For $A(\lambda) = \mathrm{diag}(d_i)$, calculate

$$E[\mu^* - \mu]^2 = \frac{E\left[n^{-1}\sum d_i w_i^2 - (n^{-1}\sum d_i)(n^{-1}\sum w_i^2)\right]^2}{E\left[n^{-1}\sum w_i^2\right]^2},$$

and compare with $E[\tilde{\mu} - \mu]^2 = E\left[n^{-1}\sum d_i w_i^2 - n^{-1}\sum d_i\right]^2$.

4
More Splines

The framework for model construction as laid out in Chap. 2 takes as building blocks any reproducing kernel. The polynomial splines of §2.3 are the standard choices on continuous domains, but generalizations or restrictions are sometimes called for by the nature of the applications. The technical underpinnings of the variants are generally different from that of polynomial splines, but once the reproducing kernels are specified, everything else remains largely intact.

In this chapter, we present several variants of polynomial splines that have a broad range of applications. Discussed in §4.2 are splines on the circle, or periodic polynomial splines, which are often used to model periodic phenomena as well as to showcase asymptotic calculations. To model spatial data in a natural manner, one has at his disposal the isotropically invariant thin-plate splines on the domain $\mathcal{X} = (-\infty, \infty)^d$ (§4.3) and spherical splines on the sphere $\mathcal{X} = \mathcal{S}$ (§4.4). L-Splines are discussed in §4.5, where the null space $\mathcal{N}_J$ of the roughness penalty $J(f)$ is not restricted to lower-order polynomials. The derivation of the reproducing kernels is the main focus of the discussion, although some advanced mathematical background is relegated to the literature.

The simple but useful idea of partial splines is also briefly discussed and illustrated (§4.1).

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_4,

4.1 Partial Splines

In some applications, one may want to use a semiparametric model,

$$Y = \mathbf{z}^T\boldsymbol{\beta} + \eta(x) + \epsilon,$$

where $\mathbf{z}$ comprises the parametric covariate with coefficient $\boldsymbol{\beta}$ and x is the nonparametric covariate. The minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \mathbf{z}_i^T\boldsymbol{\beta} - \eta(x_i)\right)^2 + \lambda J(\eta) \tag{4.1}$$

with respect to $\boldsymbol{\beta}$ and $\eta \in \mathcal{H} = \{f : J(f) < \infty\}$ is called a partial spline.

To compute a partial spline, one simply augments the matrix S in (3.4) or (3.63) by $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$, $\tilde{S} = (Z, S)$, augments $\mathbf{d}$ by $\boldsymbol{\beta}$, $\tilde{\mathbf{d}} = (\boldsymbol{\beta}^T, \mathbf{d}^T)^T$, and replaces $(S, \mathbf{d})$ by $(\tilde{S}, \tilde{\mathbf{d}})$ in the algorithms of §§3.4 and 3.5.3. The minimizer of (4.1) uniquely exists when $\tilde{S}$ is of full column rank.

The `ssanova` and `ssanova0` suites have partial spline utilities built in.

Example 4.1 (Cubic spline with jump) To estimate a function that has a possible jump at a known location $x = 0.7$ but otherwise believed to be smooth on $[0, 1]$, one may minimize

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \beta I_{[x_i>0.7]} - \eta(x_i)\right)^2 + \lambda\int_0^1 \ddot{\eta}^2 dx$$

with respect to β and $\eta \in \{f : \int_0^1 \ddot{f}^2 dx < \infty\}$.

The following sequence generates some synthetic data and fits a cubic spline with a jump:

```
set.seed(5732)
x <- runif(100); z <- as.numeric(x>.7)
y <- 1+3*sin(2*pi*x-pi)-2*z+rnorm(x)
fit.part <- ssanova(y~x,partial=~z)
```

Linear parametric terms are to be generated by `partial` as in `lm` but each term here will be standardized internally to have mean 0 and variance 1. One can then evaluate the fit and plot as shown in Fig. 4.1:

```
grid <- seq(0,1,len=51)
new <- data.frame(x=grid,z=as.numeric(grid>.7))
est <- predict(fit.part,new,se=TRUE)
plot(x,y,col=3); lines(grid,est$fit)
lines(grid,est$fit+1.96*est$se,col=5)
lines(grid,est$fit-1.96*est$se,col=5)
lines(grid,1+3*sin(2*pi*grid-pi)-2*(grid>.7),lty=2)
```

Obviously, the same variable should not appear in both formulas as that will create identifiability problems. □

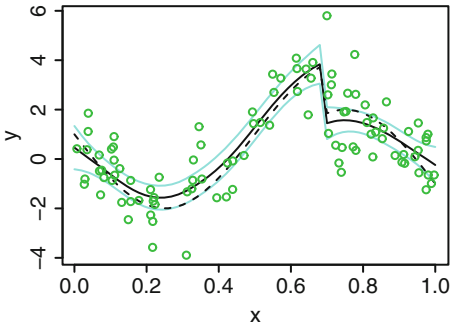


FIGURE 4.1. A cross-validated cubic spline fit with jump. The fit is in the *solid line* and the 95 % Bayesian confidence intervals are in *faded lines*, with the test function superimposed in *dashed line* and the data in *circles*.

4.2 Splines on the Circle

Functions on the circle are isomorphic to periodic functions on $[0,1]$. A periodic function $f(x)$ on $[0,1]$ can usually be expressed in the form of a Fourier series expansion

$$f(x) = a_0 + \sum_{\mu=1}^{\infty} \big(a_\mu \cos 2\pi\mu x + b_\mu \sin 2\pi\mu x\big), \tag{4.2}$$

where $\sum_{\mu=1}^{\infty}(a_\mu^2 + b_\mu^2) < \infty$. Denote by $\mathcal{P}[0,1]$ the linear space of all functions on $[0,1]$ permitting the Fourier series expansion (4.2); all continuous periodic functions belong to $\mathcal{P}[0,1]$.

In parallel to §2.3.3, we present a family of reproducing kernels on $[0,1]$ for periodic polynomial splines. With equally spaced data, a periodic polynomial spline is shown to be equivalent to a low-pass filter through an analytical spectral decomposition of the matrix Q appearing in (3.4). Assisted by such an analytical spectral decomposition, it is also possible to illustrate further details of the asymptotics of §3.2 concerning smoothing parameter selection.

4.2.1 Periodic Polynomial Splines

Consider the space $\mathcal{H} = \big\{f : f \in \mathcal{P}[0,1], f^{(m)} \in \mathcal{L}_2[0,1]\big\}$. By the orthogonality of the trigonometric basis, it is easy to calculate

$$\int_0^1 \big(f^{(m)}\big)^2 dx = \frac{1}{2}\sum_{\mu=1}^{\infty}(a_\mu^2 + b_\mu^2)(2\pi\mu)^{2m} \tag{4.3}$$

for $f \in \mathcal{P}[0,1]$, noting that $\int_0^1 \sin^2 2\pi\mu x\,dx = \int_0^1 \cos^2 2\pi\mu x\,dx = 1/2$; see Problem 4.1. Hence, $\mathcal{H} = \{f : f \in \mathcal{P}[0,1], \sum_{\mu=1}^{\infty}(a_\mu^2 + b_\mu^2)\mu^{2m} < \infty\}$. With an inner product

$$(f,g) = \left(\int_0^1 f dx\right)\left(\int_0^1 g dx\right) + \int_0^1 f^{(m)} g^{(m)} dx,$$

the reproducing kernel is seen to be

$$\begin{aligned} R(x,y) &= 1 + \sum_{\mu=1}^{\infty} \frac{2}{(2\pi\mu)^{2m}}(\cos 2\pi\mu x \cos 2\pi\mu y + \sin 2\pi\mu x \sin 2\pi\mu y) \\ &= 1 + \sum_{\mu=1}^{\infty} \frac{2\cos 2\pi\mu(x-y)}{(2\pi\mu)^{2m}}; \end{aligned} \tag{4.4}$$

see Problem 4.2. Comparing this with (2.18) on page 37, it is easy to verify that $R(x,y) = 1 + (-1)^{m-1}k_{2m}(x-y)$; see Problem 4.3. A one-way ANOVA decomposition with the averaging operator $Af = \int_0^1 f dx$ is built in, with $R_0 = 1$ generating the "mean" space and $R_1 = (-1)^{m-1}k_{2m}(x-y)$ generating the "contrast" space.

Consider $Y_i = \eta(x_i) + \epsilon_i$, where $\epsilon_i \sim N(0,\sigma^2)$. The minimizer η_λ of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1 \big(\eta^{(m)}\big)^2 dx, \tag{4.5}$$

for $\eta \in \mathcal{H} \subset \mathcal{P}[0,1]$, is a periodic polynomial spline.

To fit a periodic cubic spline to the data of Example 3.1, one may use

```
ssanova(y~x,type=list(x=list("per",c(0,1))))
```

where the domain, which is $[0,1]$ here, must be specified; one may specify any domain, which will be mapped to $[0,1]$. The same sequence used in Example 3.1 for the evaluation and the plotting of the fit yields Fig. 4.2; the Bayesian confidence intervals here do not grow wider towards 0 and 1, which are now the same point. One may also configure selected margins in tensor product splines as periodic polynomial splines.

4.2.2 Splines as Low-Pass Filters

In the notation of §3.1, $\mathcal{N}_J = \text{span}\{1\}$ and $R_J(x,y) = (-1)^{m-1}k_{2m}(x-y)$. To compute the minimizer η_λ of (4.5) via (3.4) on page 63, one has $S = \mathbf{1}$ and Q with the (i,j)th entry $(-1)^{m-1}k_{2m}(x_i - x_j)$.

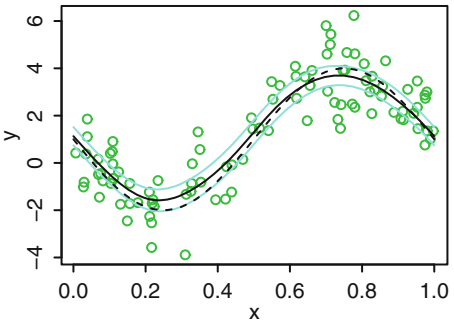


FIGURE 4.2. A cross-validated periodic cubic spline fit. The fit is in the *solid line* and the 95 % Bayesian confidence intervals are in *faded lines*, with the test function superimposed in *dashed line* and the data in *circles*.

Consider equally spaced data with $x_i = (i-1)/n$. The (i,j)th entry of Q is then $(-1)^{m-1}k_{2m}\big((i-j)/n\big)$. Substituting in the expression (2.18), straightforward algebra yields

$$\begin{aligned}(-1)^{m-1}k_{2m}\big((i-j)/n\big) &= \left(\sum_{\mu=-\infty}^{-1} + \sum_{\mu=1}^{\infty}\right)\frac{\exp\big(2\pi\mathbf{i}\mu(i-j)/n\big)}{(2\pi\mu)^{2m}}\\ &= \left(\sum_{\xi=-\infty}^{-1} + \sum_{\xi=1}^{\infty}\right)\frac{\exp\big(2\pi\mathbf{i}(n\xi)(i-j)/n\big)}{(2\pi n\xi)^{2m}}\\ &\quad + \sum_{\nu=1}^{n-1}\sum_{\xi=-\infty}^{\infty}\frac{\exp\big(2\pi\mathbf{i}(\nu+n\xi)(i-j)/n\big)}{\big(2\pi(\nu+n\xi)\big)^{2m}}\\ &= \sum_{\nu=0}^{n-1}\lambda_\nu\frac{\exp\big(2\pi\mathbf{i}\nu(i-j)/n\big)}{n}, \end{aligned} \tag{4.6}$$

where $\mathbf{i} = \sqrt{-1}$ and

$$\begin{aligned}\lambda_0 &= 2n(2\pi)^{-2m}\sum_{\xi=1}^{\infty}(n\xi)^{-2m},\\ \lambda_\nu &= n(2\pi)^{-2m}\sum_{\xi=-\infty}^{\infty}(\nu+n\xi)^{-2m}, \quad \nu = 1,\dots,n-1.\end{aligned} \tag{4.7}$$

Hence, one has the spectral decomposition $Q = \Gamma\Lambda\Gamma^H$, where Γ is the Fourier matrix with the (i,j)th entry $n^{-1/2}\exp\big(2\pi\mathbf{i}(i-1)(j-1)/n\big)$, Γ^H the conjugate transpose of Γ, $\Gamma^H\Gamma = \Gamma\Gamma^H = I$, and $\Lambda = \mathrm{diag}(\lambda_0,\dots,\lambda_{n-1})$; see Problem 4.4. Note that $\lambda_\nu = \lambda_{n-\nu}$, $\nu = 1,\dots,n-1$.

The operation $\tilde{\mathbf{z}} = \Gamma^H\mathbf{z}$ defines the discrete Fourier transform of $\mathbf{z}$ and $\mathbf{z} = \Gamma\tilde{\mathbf{z}}$ defines its inverse. It is easy to see that $\Gamma^H\mathbf{1} = \sqrt{n}\mathbf{e}_1$, where $\mathbf{e}_1$ is

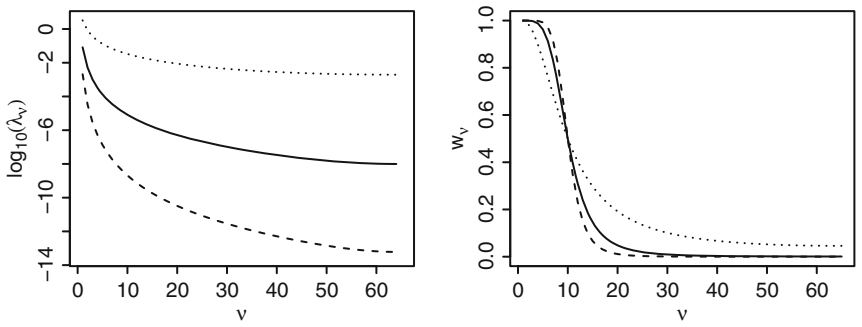


FIGURE 4.3. Splines as low-pass filters. *Left*: Eigenvalues λ_ν of Q. *Right*: Damping factors w_ν with $w_{10} = .5$. The *dotted lines* are for $m = 1$, the *solid lines* for $m = 2$, and the *dashed lines* for $m = 3$. The sample size is $n = 128$.

the first unit vector. Let $\tilde{Y}$ be the discrete Fourier transform of $\mathbf{Y}$ and $\tilde{\mathbf{c}}$ be that of $\mathbf{c}$. The linear system (3.4) reduces to

$$\begin{aligned}(\Lambda + n\lambda I)\tilde{\mathbf{c}} + \sqrt{n}\mathbf{e}_1 d &= \tilde{\mathbf{Y}},\\ \tilde{c}_1 &= 0.\end{aligned}$$

Hence, one has $\tilde{c}_\nu = \tilde{Y}_\nu/(\lambda_{\nu-1} + n\lambda)$, $\nu = 2, \ldots, n$. Remember that $\hat{\mathbf{Y}} = \mathbf{Y} - n\lambda\mathbf{c}$, so $\tilde{\hat{Y}}_\nu = w_\nu \tilde{Y}_\nu$, where $w_1 = 1$, $w_\nu = \lambda_{\nu-1}/(\lambda_{\nu-1} + n\lambda)$, $\nu = 2, \ldots, n$. The eigenvalues λ_ν of Q monotonically decreases up to $\nu = n/2$, so a periodic spline with equally spaced data is virtually a low-pass filter.

For $n = 128$ and $m = 1, 2, 3$, $\log_{10}\lambda_\nu$, $\nu = 1, \ldots, 64$, are plotted in the left frame of Fig. 4.3, and w_ν with $n\lambda = \lambda_9$, $\nu = 1, \ldots, 65$, are plotted in the right frame. The order m controls the shape of the filter, and the smoothing parameter λ determines the half-power frequency.

4.2.3 More on Asymptotics of §3.2

Assisted by the analytical spectral decomposition $Q = \Gamma\Lambda\Gamma^H$ for periodic splines with equally spaced data, we can now look into further details of the asymptotics of §3.2 concerning smoothing parameter selection.

Write $W = \mathrm{diag}(w_1, \ldots, w_n)$, where $w_1 = 1$ and $w_\nu = \lambda_{\nu-1}/(\lambda_{\nu-1} + n\lambda)$, $\nu = 2, \ldots, n$. From $\hat{\tilde{\mathbf{Y}}} = W\tilde{\mathbf{Y}}$, one has $A(\lambda) = \Gamma W \Gamma^H$. It follows that

$$\mathrm{tr}A(\lambda) = 1 + \sum_{\nu=1}^{n-1} \frac{\lambda_\nu}{\lambda_\nu + n\lambda} = 1 + \sum_{\nu=1}^{n-1} \frac{1}{1 + \lambda\rho_\nu}, \tag{4.8}$$

where $\rho_\nu = n/\lambda_\nu$, and

$$\mathrm{tr}A^2(\lambda) = 1 + \sum_{\nu=1}^{n-1} \frac{1}{(1 + \lambda\rho_\nu)^2}. \tag{4.9}$$

For $\nu \le n/2$, it follows from (4.7) that $\rho_\nu = n/\lambda_\nu \asymp \nu^{2m}$. As $\lambda \to 0$ and $n\lambda^{1/2m} \to \infty$,

$$\begin{aligned}
\mathrm{tr}A(\lambda) &= K_1 + 2\Big(\sum_{\nu \le \lambda^{-1/2m}} + \sum_{\lambda^{-1/2m} < \nu < n/2}\Big)\frac{1}{1+\lambda\rho_\nu} \\
&= K_1 + K_2\lambda^{-1/2m} + K_3\int_{\lambda^{-1/2m}}^{n/2} \frac{1}{1+\lambda x^{2m}}dx \\
&= K_1 + K_2\lambda^{-1/2m} + K_3\lambda^{-1/2m}\int_1^{\infty} \frac{1}{1+x^{2m}}dx \\
&\asymp \lambda^{-1/2m},
\end{aligned}$$

where the K_i's are constants bounded away from 0 and ∞. Similarly, one has $\mathrm{tr}A^2(\lambda) \asymp \lambda^{-1/2m}$. Condition 3.2.2 of §3.2.2, that

$$\big\{n^{-1}\mathrm{tr}A(\lambda)\big\}^2/n^{-1}\mathrm{tr}A^2(\lambda) \to 0,$$

follows when $\lambda \to 0$ and $n\lambda^{1/2m} \to \infty$.

We now calculate the risk $R(\lambda) = E\big[n^{-1}\sum_{i=1}^n \big(\eta_\lambda(x_i) - \eta(x_i)\big)^2\big]$ and verify Condition 3.2.1 of §3.2.1, that $nR(\lambda) \to \infty$. From (3.16) on page 65, one has

$$R(\lambda) = \frac{1}{n}\boldsymbol{\eta}^T\big(I - A(\lambda)\big)^2\boldsymbol{\eta} + \frac{\sigma^2}{n}\mathrm{tr}A^2(\lambda) = B(\lambda) + O(n^{-1}\lambda^{-1/2m}), \quad (4.10)$$

where $\boldsymbol{\eta} = \big(\eta(0), \eta(1/n), \ldots, \eta((n-1)/n)\big)^T$, with the bias term

$$\begin{aligned}
B(\lambda) = \frac{1}{n}\boldsymbol{\eta}^T\big(I - A(\lambda)\big)^2\boldsymbol{\eta} &= \frac{1}{n}\sum_{\nu=1}^{n-1}\frac{(n\lambda)^2}{(\lambda_\nu + n\lambda)^2}|\tilde{\eta}_{\nu+1}|^2 \\
&= \lambda\sum_{\nu=1}^{n-1}\frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\frac{\rho_\nu}{n}|\tilde{\eta}_{\nu+1}|^2, \qquad (4.11)
\end{aligned}$$

where $\tilde{\eta}_{\nu+1}$ is the $(\nu+1)$st entry of $\Gamma^H\boldsymbol{\eta}$. It will be shown that $B(\lambda) = O(\lambda^p)$ for some $p \in [1,2]$, so Condition 3.2.1 follows when $n\lambda^p \to \infty$ and $\lambda \to 0$. The optimal $\lambda \asymp n^{-2m/(2pm+1)}$ satisfies these conditions.

We now show that $B(\lambda) = O(\lambda^p)$ for some $p \in [1,2]$. For $\eta \in \mathcal{P}[0,1]$,

$$\eta(i/n) = \tilde{a}_0 + \sum_{\mu=1}^{n-1}\big(\tilde{a}_\mu\cos(2\pi\mu i/n) + \tilde{b}_\mu\sin(2\pi\mu i/n)\big),$$

where $\tilde{a}_0 = \sum_{\xi=0}^{\infty} a_{n\xi}$, $\tilde{a}_\nu = \sum_{\xi=0}^{\infty} a_{\nu+n\xi}$, and $\tilde{b}_\nu = \sum_{\xi=0}^{\infty} b_{\nu+n\xi}$, $\nu = 1, \dots, n-1$. For integers ν and μ, one has the orthogonality relations

$$\begin{aligned}
&\sum_{i=1}^{n} \cos(2\pi\nu i/n)\cos(2\pi\mu i/n) = \frac{n}{2}\delta_{\nu,\mu}, && \nu,\mu \in [1, n/2),\\
&\sum_{i=1}^{n} \cos^2(2\pi\nu i/n) = n, && \nu = n/2,\\
&\sum_{i=1}^{n} \sin(2\pi\nu i/n)\sin(2\pi\mu i/n) = \frac{n}{2}\delta_{\nu,\mu}, && \nu,\mu \in [1, n/2),\\
&\sum_{i=1}^{n} \cos(2\pi\nu i/n)\sin(2\pi\mu i/n) = 0,
\end{aligned} \tag{4.12}$$

where $\delta_{\nu,\mu}$ is the Kronecker delta. It follows that

$$\tilde{\eta}_{\nu+1} = \frac{\sqrt{n}}{2}\big\{(\tilde{a}_\nu + \tilde{a}_{n-\nu}) + \mathbf{i}(b_\nu - b_{n-\nu})\big\}, \quad \nu = 1, \dots, n-1, \tag{4.13}$$

so

$$|\tilde{\eta}_{\nu+1}|^2 = \frac{n}{4}\big\{(\tilde{a}_\nu + \tilde{a}_{n-\nu})^2 + (\tilde{b}_\nu - \tilde{b}_{n-\nu})^2\big\}, \quad \nu = 1, \dots, n-1;$$

see Problem 4.5. For $\nu > 0$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\tilde{a}_\nu^2 &\le \left(\sum_{\xi=0}^{\infty} a_{\nu+n\xi}^2(\nu+n\xi)^{2m}\right)\left(\sum_{\xi=0}^{\infty}(\nu+n\xi)^{-2m}\right),\\
\tilde{b}_\nu^2 &\le \left(\sum_{\xi=0}^{\infty} b_{\nu+n\xi}^2(\nu+n\xi)^{2m}\right)\left(\sum_{\xi=0}^{\infty}(\nu+n\xi)^{-2m}\right),
\end{aligned}$$

where $\sum_{\xi=0}^{\infty}(\nu+n\xi)^{-2m} \propto \lambda_\nu/n = \rho_\nu^{-1}$. Since $\sum_{\mu=1}^{\infty}(a_\mu^2 + b_\mu^2)\mu^{2m} < \infty$, one has $\sum_{\nu=1}^{n-1}\rho_\nu\tilde{a}_\nu^2 < \infty$ and $\sum_{\nu=1}^{n-1}\rho_\nu\tilde{b}_\nu^2 < \infty$. It follows that

$$\sum_{\nu=1}^{n-1}\frac{\rho_\nu}{n}|\tilde{\eta}_{\nu+1}|^2 \le \frac{1}{2}\sum_{\nu=1}^{n/2}\rho_\nu\big\{(\tilde{a}_\nu + \tilde{a}_{n-\nu})^2 + (\tilde{b}_\nu - \tilde{b}_{n-\nu})^2\big\} = O(1).$$

Plugging this into (4.11) and noting that $\lambda\rho_\nu/(1+\lambda\rho_\nu)^2 < 1$, one has $B(\lambda) = O(\lambda)$. When η is "supersmooth," in the sense that

$$\sum_{\mu=1}^{\infty}(a_\mu^2 + b_\mu^2)\mu^{2pm} < \infty \tag{4.14}$$

holds for some $p > 1$, similar calculation yields $B(\lambda) = O(\lambda^p)$, for p up to 2. When (4.14) holds for $p > 2$ but $B_2 = \lambda^{-2}B(\lambda)|_{\lambda=0} > 0$, it can be shown that $\lambda^{-2}B(\lambda) - B_2 = o(1)$ for $\lambda = o(1)$ (Problem 4.6), so $O(\lambda^2)$ is the best attainable rate for $B(\lambda)$.

Finally, let us see how the minimizer λ_m of the score $M(\lambda)$ undersmoothes when η is "supersmooth." Plugging the spectral decomposition $A(\lambda) = \Gamma W \Gamma^H$ into (3.30) on page 71, after some algebra one gets

$$M(\lambda) = \frac{\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}|\tilde{Y}_{\nu+1}|^2}{\left(\prod_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}\right)^{1/(n-1)}}.$$

Straightforward calculation yields

$$\begin{aligned}\frac{d\log M(\lambda)}{d\log\lambda} &= \frac{\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}|\tilde{Y}_{\nu+1}|^2}{\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}|\tilde{Y}_{\nu+1}|^2} - \frac{1}{n-1}\sum_{\nu=1}^{n-1}\frac{1}{1+\lambda\rho_\nu} \\ &= \frac{N(\lambda)}{D(\lambda)} - \frac{1}{n-1}\mathrm{tr}A(\lambda), \end{aligned} \tag{4.15}$$

say; see Problem 4.7. As shown earlier, $(n-1)^{-1}\mathrm{tr}A(\lambda) \asymp n^{-1}\lambda^{-1/2m}$. Now

$$|\tilde{Y}_{\nu+1}|^2 = |\tilde{\eta}_{\nu+1}|^2 + |\tilde{\epsilon}_{\nu+1}|^2 + (\tilde{\eta}_{\nu+1}\bar{\tilde{\epsilon}}_{\nu+1} + \bar{\tilde{\eta}}_{\nu+1}\tilde{\epsilon}_{\nu+1}),$$

where $\bar{z}$ denotes the conjugate of complex number z, and, correspondingly, $N(\lambda)$ and $D(\lambda)$ can each be decomposed into three terms. We shall calculate the rates for the terms corresponding to $|\tilde{\eta}_{\nu+1}|^2$ and $|\tilde{\epsilon}_{\nu+1}|^2$, which control the rate of the cross-term through the Cauchy-Schwarz inequality; see Problem 4.8.

It is easy to verify that

$$\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}|\tilde{\eta}_{\nu+1}|^2 = O(\lambda)$$

and that

$$\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{1+\lambda\rho_\nu}|\tilde{\epsilon}_{\nu+1}|^2 = \frac{1}{n}\boldsymbol{\epsilon}^T\big(I - A(\lambda)\big)\boldsymbol{\epsilon} = \sigma^2(1-\mu_1) + o_p\big(R(\lambda) + n^{-1}\big),$$

where $\mu_1 = n^{-1}\mathrm{tr}A(\lambda)$ and (3.19) on page 66 is used. It follows that $D(\lambda) = \sigma^2\big(1 + o_p(1)\big)$. Similarly,

$$N_1(\lambda) = \frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}|\tilde{\eta}_{\nu+1}|^2 = O(\lambda)$$

and

$$\frac{1}{n}\sum_{\nu=1}^{n-1} \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}|\tilde{\epsilon}_{\nu+1}|^2 = \frac{1}{n}\boldsymbol{\epsilon}^T\big(A(\lambda) - A^2(\lambda)\big)\boldsymbol{\epsilon} = O_p(n^{-1}\lambda^{-1/2m}),$$

so $N(\lambda) = O_p(\lambda + n^{-1}\lambda^{-1/2m})$.

When η is "supersmooth" but $\lambda^{-1}N_1(\lambda)|_{\lambda=0} > 0$, one has $N_1(\lambda) \asymp \lambda$; the proof is similar to Problem 4.6. Hence, λ is the best attainable rate for $N_1(\lambda)$. Putting things together, it follows that λ cannot exceed the order of $n^{-1}\lambda^{-1/2m}$ for (4.15) to evaluate to zero. This leads to $\lambda_m = O(n^{-2m/(2m+1)})$, which is smaller than the optimal $\lambda \asymp n^{-2m/(2pm+1)}$ when $p > 1$.

4.3 Thin-Plate Splines

A thin-plate spline is the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda J_m^d(\eta) \tag{4.16}$$

on the d-dimensional domain $\mathcal{X} = (-\infty, \infty)^d$, where

$$J_m^d(f) = \sum_{\alpha_1+\cdots+\alpha_d=m} \frac{m!}{\alpha_1!\cdots\alpha_d!} \times \int\cdots\int \left(\frac{\partial^m f}{\partial x_{\langle 1\rangle}^{\alpha_1}\cdots\partial x_{\langle d\rangle}^{\alpha_d}}\right)^2 dx_{\langle 1\rangle}\cdots dx_{\langle d\rangle}. \tag{4.17}$$

The null space of $J_m^d(f)$ consists of polynomials of up to $(m-1)$ total order, which is of dimension $M = \binom{d+m-1}{d}$; see Problem 4.9. The quadratic functional $J_m^d(f)$ is invariant under a rotation of the coordinates; see Problem 4.10. In the space $\mathcal{H} = \big\{f : J_m^d(f) < \infty\big\}$ with $J_m^d(f)$ as a square semi norm, it is necessary that $2m - d > 0$ for the evaluation functional $[x]f = f(x)$ to be continuous; see Duchon (1977), Meinguet (1979) and Wahba and Wendelberger (1980).

The derivation of reproducing kernels for thin-plate splines requires some advanced knowledge of differential equation theory; details can be found

in Duchon (1977), Meinguet (1979) and references cited therein. In the sections to follow, we try to keep the exposition to an elementary level, leaving the technically more advanced discussion to the literature. For the fitting of a thin-plate spline alone using Algorithm 3.1 of §3.4.2, an easy-to-evaluate, conditionally non-negative definite semi-kernel is all that one would need (§4.3.1), but to compute the Bayesian confidence intervals or to use thin-plate marginals to construct tensor product splines, genuine reproducing kernels have to be constructed (§4.3.2). Tensor product splines with thin-plate marginals are briefly discussed in §4.3.3, and the case study previewed in §1.4.1 is developed in full in §4.3.4.

4.3.1 Semi-Kernels for Thin-Plate Splines

When the parametric least squares estimate in the null space of $J_m^d(f)$ uniquely exists, the minimizer η_λ of (4.16) uniquely exists; see Theorem 2.9. From Duchon (1977, Theorem 4 bis), η_λ has an expression

$$\eta_\lambda(x) = \sum_{\nu=1}^{M} d_\nu \phi_\nu(x) + \sum_{i=1}^{n} c_i E\big(|x_i - x|\big), \tag{4.18}$$

where $\{\phi_\nu\}_{\nu=1}^M$ span the null space of $J_m^d(f)$, c_i's are subject to the constraints $S^T\mathbf{c} = 0$ with S the $n \times M$ matrix with the (i,ν)th entry $\phi_\nu(x_i)$, $|x-y|$ is the Euclidean distance, and

$$E(x) = \begin{cases} \theta_{m,d}\, x^{2m-d} \log x, & d \text{ even, for} \\ & \theta_{m,d} = \frac{(-1)^{d/2+m+1}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, \\ \theta_{m,d}\, x^{2m-d}, & d \text{ odd, for} \\ & \theta_{m,d} = \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}. \end{cases} \tag{4.19}$$

The constant $\theta_{m,d}$ in (4.19) is not really needed for (4.18), as it is readily absorbed into c_i's. The reproducing kernels, however, are expressed in terms of $E(x)$ with $\theta_{m,d}$ attached, as will be seen shortly.

For c_i's satisfying $S^T\mathbf{c} = 0$, it can be shown that

$$J_m^d\big(\textstyle\sum_{i=1}^n c_i E(|x_i - x|)\big) = \displaystyle\sum_i \sum_j c_i c_j E\big(|x_i - x_j|\big); \tag{4.20}$$

see Meinguet (1979) and Wahba and Wendelberger (1980). Plugging (4.18) and (4.20) into (4.16), the estimation reduces to the minimization of

$$(\mathbf{Y} - S\mathbf{d} - K\mathbf{c})^T(\mathbf{Y} - S\mathbf{d} - K\mathbf{c}) + n\lambda\, \mathbf{c}^T K\mathbf{c} \tag{4.21}$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, subject to the constraints $S^T\mathbf{c} = 0$, where K is $n \times n$ with the (i,j)th entry $E\big(|x_i - x_j|\big)$.

Compare (4.21) with (3.3) on page 62 and (3.4) on page 63. It is easily seen that the solution of the linear system

$$\begin{aligned}(K + n\lambda I)\mathbf{c} + S\mathbf{d} &= \mathbf{Y},\\ S^T\mathbf{c} &= 0,\end{aligned} \tag{4.22}$$

is a solution of the constrained minimization problem (4.21).

To compute a thin-plate spline, one may use Algorithm 3.1 of §3.4.2, which was designed for the linear system (3.4). The only difference between (3.4) and (4.22) is that Q in (3.4) is non-negative definite, whereas K in (4.22) is not. It is easy to verify, however, that one only needs $F_2^T Q F_2$ to be non-negative definite for Algorithm 3.1 to work, and indeed it is the case; check (4.20). The matrix K satisfying

$$S^T\mathbf{c} = 0 \quad \Longrightarrow \quad \mathbf{c}^T K \mathbf{c} \geq 0$$

is said to be conditionally non-negative definite.

The bivariate function $E\big(|x-y|\big)$ acts like a reproducing kernel in this approach to the computation of thin-plate splines, and hence is called a semi-kernel. Note that only the sign of $\theta_{m,d}$ matters for the calculation, as the magnitude can be absorbed into c_i's and λ.

Example 4.2 (Cubic spline) With $d = 1$ and $m = 2$, one has $J_2^1(f) = \int_{-\infty}^{\infty} \ddot{f}^2 dx$, yielding a cubic spline on the real line. Since $\Gamma(1/2 - 2) > 0$, $E\big(|x-y|\big) \propto |x-y|^3$. One has

$$\eta_\lambda(x) = d_1 + d_2 x + \sum_{i=1}^{n} c_i |x_i - x|^3,$$

with $\mathbf{c}$ and $\mathbf{d}$ solving (4.22) for K with the (i,j)th entry $|x_i - x_j|^3$. Under this formulation, one does not need to map the data into $[0,1]$. □

Example 4.3 With $d = 2$ and $m = 2$, one has $J_2^2(f) = \int\int \big(\ddot{f}^2_{\langle 11\rangle} + 2\ddot{f}^2_{\langle 12\rangle} + \ddot{f}^2_{\langle 22\rangle}\big) dx_{\langle 1\rangle} dx_{\langle 2\rangle}$. Obviously, $d/2 + m + 1$ is even, so $E\big(|x-y|\big) \propto |x-y|^2 \log|x-y|$. It follows that

$$\eta_\lambda(x) = d_1 + d_2 x_{i\langle 1\rangle} + d_3 x_{i\langle 2\rangle} + \sum_{i=1}^{n} c_i |x_i - x|^2 \log|x_i - x|,$$

with $\mathbf{c}$ and $\mathbf{d}$ the solution of (4.22), where the matrix K has the (i,j)th entry $|x_i - x_j|^2 \log|x_i - x_j|$. □

4.3.2 Reproducing Kernels for Thin-Plate Splines

For the calculation of the fit alone, it is sufficient to know the semi-kernel. To evaluate the posterior variance for the Bayesian confidence intervals of

§3.3 or to construct tensor product splines of §2.4 with thin-plate splines as building blocks on the marginal domains, one will have to calculate the genuine reproducing kernel, which is the subject of this section.

Denote by ψ_ν a set of polynomials that span $\mathcal{N}_J$, the null space of $J_m^d(f)$. Define

$$(f, g)_0 = \sum_{i=1}^{N} p_i f(u_i) g(u_i), \tag{4.23}$$

where $u_i \in (-\infty, \infty)^d$, $p_i > 0$, $\sum_{i=1}^{N} p_i = 1$ are specified such that the Gram matrix with the (ν, μ)th entry $(\psi_\nu, \psi_\mu)_0$ is nonsingular. Following some standard orthogonalization procedure, one can find an orthonormal basis ϕ_ν, $\nu = 1, \ldots, M$, for $\mathcal{N}_J$ with $\phi_1(x) = 1$ and $(\phi_\nu, \phi_\mu)_0 = \delta_{\nu,\mu}$, where $\delta_{\nu,\mu}$ is the Kronecker delta. The reproducing kernel in $\mathcal{N}_J$ is seen to be

$$R_0(x, y) = \sum_{\nu=1}^{M} \phi_\nu(x) \phi_\nu(y). \tag{4.24}$$

The projection of f onto $\mathcal{N}_J$ is defined by the operator P through

$$(Pf)(x) = \sum_{\nu=1}^{M} (f, \phi_\nu)_0 \phi_\nu(x). \tag{4.25}$$

Define

$$R_1(x, y) = (I - P_{(x)})(I - P_{(y)}) E\big(|x - y|\big), \tag{4.26}$$

where I is the identity operator and $P_{(x)}$ and $P_{(y)}$ are the projection operator of (4.25) applied to the arguments x and y, respectively.

Plugging (4.25) into (4.26), one has, for fixed x,

$$\begin{aligned} R_1(x, u) &= E\big(|x - u|\big) - \sum_{\nu=1}^{M} \phi_\nu(x) \sum_{i=1}^{N} p_i \phi_\nu(u_i) E\big(|u_i - u|\big) + \pi(u) \\ &= E(|x - u|) + \sum_{i=1}^{N} c_i(x) E\big(|u_i - u|\big) + \pi(u), \end{aligned}$$

where $\pi(u) \in \mathcal{N}_J$ and $c_i(x) = -\sum_{\nu=1}^{M} p_i \phi_\nu(u_i) \phi_\nu(x)$. From (4.20), it is easy to show that (Problem 4.11)

$$J_m^d\big(\textstyle\sum_{i=1}^{n} c_i E(|x_i - \cdot|), \sum_{j=1}^{p} \tilde{c}_i E(|y_j - \cdot|)\big) = \sum_{i,j} c_i \tilde{c}_j E\big(|x_i - y_j|\big), \tag{4.27}$$

for c_i and $\tilde{c}_j$ satisfying $\sum_{i=1}^n c_i\phi_\nu(x_i) = \sum_{j=1}^p \tilde{c}_j\phi_\nu(y_j) = 0$, $\nu = 1, \ldots, M$, where $J_m^d(f, g)$ denotes the (semi) inner product associated with the square (semi) norm $J_m^d(f)$. It is easy to check that

$$\phi_\nu(x) + \sum_{i=1}^N c_i(x)\phi_\nu(u_i) = \phi_\nu(x) - \sum_{\mu=1}^M (\phi_\mu, \phi_\nu)_0\phi_\mu(x) = 0$$

for $\nu = 1, \ldots, M$. Taking $n = p = N + 1$, $x_i = y_i = u_i$, $c_i = c_i(x)$, $\tilde{c}_i = c_i(y)$, $i = 1, \ldots, N$, $x_{N+1} = x$, $y_{N+1} = y$, and $c_{N+1} = \tilde{c}_{N+1} = 1$ in (4.27), one has

$$\begin{aligned}
&J_m^d\big(R_1(x,\cdot), R_1(y,\cdot)\big) \\
&\quad = E\big(|x-y|\big) - \sum_{\nu=1}^M \phi_\nu(x) \sum_{i=1}^N p_i\phi_\nu(u_i) E\big(|u_i - y|\big) \\
&\qquad - \sum_{\nu=1}^M \phi_\nu(y) \sum_{i=1}^N p_i\phi_\nu(u_i) E\big(|u_i - x|\big) \\
&\qquad + \sum_{\nu,\mu=1}^M \phi_\nu(x)\phi_\mu(y) \sum_{i,j=1}^N p_i p_j \phi_\nu(u_i)\phi_\mu(u_j) E\big(|u_i - u_j|\big) \\
&\quad = (I - P_{(x)})(I - P_{(y)}) E\big(|x-y|\big) = R_1(x, y); \qquad (4.28)
\end{aligned}$$

see Problem 4.12. It follows from (4.28) that $R_1(x, y)$ is non-negative definite, hence a reproducing kernel (by Theorem 2.3), and that in the corresponding reproducing kernel Hilbert space, $J_m^d(f, g)$ is the inner product. Actually, for all $f \in \mathcal{H} = \big\{f : J_m^d(f) < \infty\big\}$, one has

$$J_m^d\big((I - P)f, R_1(x,\cdot)\big) = (I - P)f(x),$$

so $R_1(x, y)$ is indeed the reproducing kernel of $\mathcal{H} \ominus \mathcal{N}_J$; further details can be found in Meinguet (1979) and Wahba and Wendelberger (1980). Write $R_{00}(x, y) = \phi_1(x)\phi_1(y) = 1$ and $R_{01}(x, y) = \sum_{\nu=2}^M \phi_\nu(x)\phi_\nu(y)$. The kernel decomposition $R = R_{00} + [R_{01} + R_1]$ defines a one-way ANOVA decomposition on the domain $\mathcal{X} = (-\infty, \infty)^d$ with an averaging operator $Af = \sum_{i=1}^N p_i f(u_i)$.

Example 4.4 (Cubic spline) Consider a cubic spline on the real line with $d = 1$, $m = 2$, and $E\big(|x-y|\big) \propto |x-y|^3$. Take $N = 2$, $u_1 = -1$, $u_2 = 1$, $p_1 = p_2 = 0.5$, and $\phi_2 = x$. It is easy to calculate that

$$\begin{aligned}
R_1(x, y) \propto\ & |x-y|^3 - 0.5\big\{(1-x)|1+y|^3 + (1+x)|1-y|^3\big\} \\
& - 0.5\big\{(1-y)|1+x|^3 + (1+y)|1-x|^3\big\} \\
& + 2\big\{(1+x)(1-y) + (1-x)(1+y)\big\}; \qquad (4.29)
\end{aligned}$$

see Problem 4.13. □

Whereas the semi-kernel $E\big(|x-y|\big)$ is rather convenient to work with, the reproducing kernel $R_1(x,y)$ can be a bit cumbersome to evaluate. With the choices $N=n$, $u_i=x_i$, and $p_i=1/n$, $i=1,\ldots,n$, however, efficient algorithms do exist for the calculation of the $n\times n$ matrix Q with the (i,j)th entry $R_1(x_i,x_j)$, and for the calculation of the $n\times 1$ vector $\boldsymbol{\xi}(x)$ with the ith entry $R_1(x_i,x)$. The matrix Q is used in the computation of the fit, and the vector $\boldsymbol{\xi}(x)$ is used in the evaluation of the estimate.

Set $N=n$, $u_i=x_i$, and $p_i=1/n$. To derive an orthonormal basis ϕ_ν from a set of polynomials ψ_ν that span $\mathcal{N}_J$, one forms the $n\times M$ matrix $\tilde{S}$ with the (i,ν)th entry $\psi_\nu(x_i)$ and calculates a QR-decomposition $\tilde{S}=(F_1,F_2)\binom{R}{O}=F_1R$. It follows that $\boldsymbol{\phi}=\sqrt{n}R^{-T}\boldsymbol{\psi}$ forms an orthonormal basis in $\mathcal{N}_J$ with the inner product $(f,g)_0=\sum_{i=1}^n f(x_i)g(x_i)/n$ and that F_1 has the (i,ν)th entry $\phi_\nu(x_i)/\sqrt{n}$ (Problem 4.14). From the expression in (4.28), it is easy to see that

$$Q=(I-F_1F_1^T)K(I-F_1F_1^T)=F_2F_2^TKF_2F_2^T, \tag{4.30}$$

where K is $n\times n$ with the (i,j)th entry $E\big(|x_i-x_j|\big)$ (Problem 4.15). To make sure that $\phi_1=1$, one needs to set $\psi_1=1$ and to exclude the first column of $\tilde{S}$ from pivoting when calculating the QR-decomposition. Similar to (4.30), one has

$$\begin{aligned}\boldsymbol{\xi}(x)&=(I-F_1F_1^T)\big(\boldsymbol{\kappa}(x)-KF_1\boldsymbol{\phi}(x)/\sqrt{n}\big)\\&=F_2F_2^T\big(\boldsymbol{\kappa}(x)-KF_1R^{-T}\boldsymbol{\psi}(x)\big),\end{aligned} \tag{4.31}$$

where $\boldsymbol{\kappa}(x)$ is $n\times 1$ with the ith entry $E\big(|x_i-x|\big)$ (Problem 4.16).

4.3.3 Tensor Product Splines with Thin-Plate Marginals

Using $R_0(x,y)$ of (4.24) and $R_1(x,y)$ of (4.26) in Theorem 2.6, one can construct tensor product splines with thin-plate marginals. Aside from the complication in the evaluation of the reproducing kernels, there is nothing special technically or computationally about thin-plate marginals.

Tensor product splines with thin-plate marginals do offer something conceptually novel, however, albeit technically trivial. The novel feature is the notion of multivariate main effect in an ANOVA decomposition, in a genuine sense. Consider spatial modeling with geography as one of the covariates. Using a $d=2$ thin-plate marginal on the geography domain, one is able to construct an isotropic geography main effect and interactions involving geography that are rotation invariant in the geography domain. This is often a more natural treatment as compared to breaking the geography into, say, the longitude and the latitude, which would lead to a longitude effect, a latitude effect, plus a longitude-latitude interaction, that may not make much practical sense.

4.3.4 Case Study: Water Acidity in Lakes

We now fill in details concerning the analysis of the EPA lake acidity data discussed in §1.4.1. A subset of the data concerning 112 lakes in the Blue Ridge is included in `gss` as a data frame `LakeAcidity` with elements `ph`, `cal`, `lon`, `lat`, and `geog`, where `geog` contains the x-y coordinates (in distance) of the lakes with respect to a local origin; for (ϕ, θ) the longitude-latitude of lakes around a local origin (ϕ_0, θ_0), the x-y coordinates are obtained through

$$\begin{aligned} x &= \cos\big(\pi\theta/180\big)\sin\big(\pi(\phi-\phi_0)/180\big), \\ y &= \sin\big(\pi(\theta-\theta_0)/180\big), \end{aligned} \tag{4.32}$$

with the Earth's radius as the unit distance. Such mapping and its inverse can be done in R using the following functions.

```
ltln2xy <- function(latlon,latlon0) {
  lat <- latlon[,1]*pi/180; lon <- latlon[,2]*pi/180
  lt0 <- latlon0[1]*pi/180; ln0 <- latlon0[2]*pi/180
  x <- cos(lt0)*sin(lon-ln0); y <- sin(lat-lt0)
  cbind(x,y)
}
xy2ltln <- function(xy,latlon0) {
  x <- xy[,1]; y <- xy[,2]
  lt0 <- latlon0[1]*pi/180
  lat <- asin(y)/pi*180+latlon0[1]
  lon <- asin(x/cos(lt0))/pi*180+latlon0[2]
  data.frame(lat=lat,lon=lon)
}
```

A tensor product spline can be fitted to the data using `ssanova`:

```
data(LakeAcidity); set.seed(5732)
fit.lake <- ssanova(ph~log(cal)*geog,data=LakeAcidity)
```

The variable `geog` in the data frame `LakeAcidity` is a matrix with its integrity preserved by the as-is function `I(...)`:

```
LakeAcidity <- data.frame(...,geog=I(geog),...)
```

By default, a thin-plate spline is configured for a matrix variable, with $m = 2$ in $J^d_m(f)$, $\{u_i\} = \{\tilde{x}_i\}$ and $p_i = 1/n$ in (4.23), where $\tilde{x}_i$ are the marginal sampling points; a cubic spline is the default for the vector variable `log(cal)`. Checking the diagnostics:

```
sum.lake <- summary(fit.lake,diag=TRUE)
round(sum.lake$kappa,2)
#     log(cal)          geog log(cal):geog
```

```
#         1.06          1.03          1.04
round(sum.lake$cos,2)
#       log(cal) geog log(cal):geog yhat    y    e
# cos.y     0.65 0.53          -0.1 0.76 1.00 0.68
# cos.e     0.00 0.09           0.0 0.04 0.68 1.00
# norm      2.37 1.40           0.1 2.99 4.10 2.68
project(fit.lake,c("log(cal)","geog"))$ratio
# 0.0005530675
```

it is seen that the interaction can be eliminated. An additive model is now fitted to the data, which was plotted in Fig. 1.2:

```
fit.lake.a <- ssanova(ph~log(cal)+geog,data=LakeAcidity,
                      id.basis=1:112)
```

where `id.basis=1:112` sets $q = n$; `project` could mislead on fits with $q = n$ so $q < n$ was used earlier in `fit.lake`. The plots are reproduced in Fig. 4.4 for convenient reference.

To obtain the `log(cal)` effect as plotted in the top frame of Fig. 4.4, which is virtually a linear function, one may use:

```
est.cal <- predict(fit.lake.a,fit.lake.a$mf,
                   se=TRUE,inc="log(cal)")
```

To evaluate the `geog` effect on a grid, try:

```
grid0 <- seq(-.04,.04,len=31)
grid <- cbind(rep(grid0,31),rep(grid0,rep(31,31)))
est.geog <- predict(fit.lake.a,data.frame(geog=I(grid)),
                    se=TRUE,inc="geog")
```

The fitted values are contoured in the left frame and the standard errors in the right frame in dotted lines, with the x-y grid mapped back to longitude-latitude:

```
library(maps)
m.lat <- (min(LakeAcidity$lat)+max(LakeAcidity$lat))/2
m.lon <- (min(LakeAcidity$lon)+max(LakeAcidity$lon))/2
ltln.grid <- xy2ltln(cbind(grid0,grid0),c(m.lat,m.lon))
lon.gd <- ltln.grid[,2]; lat.gd <- ltln.grid[,1];
contour(lon.gd,lat.gd,matrix(est.geog$fit,31,31))
map("state",add=TRUE,col=5)
contour(lon.gd,lat.gd,matrix(est.geog$se,31,31))
map("state",add=TRUE,col=5)
points(LakeAcidity$lon,LakeAcidity$lat,col=3)
```

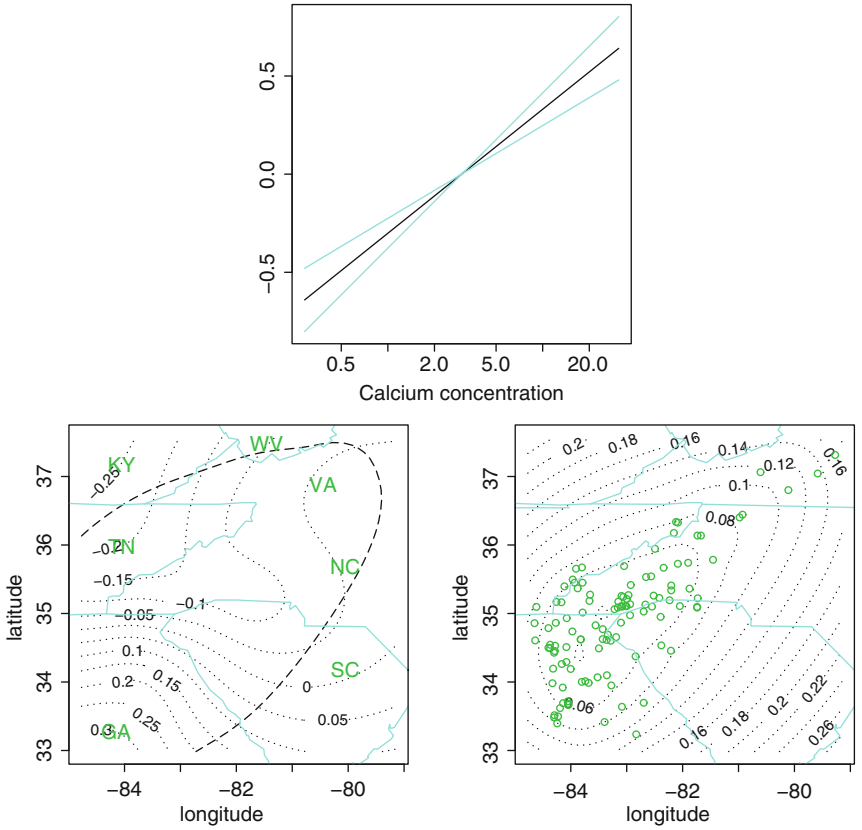


FIGURE 4.4. Water acidity fit for lakes in the Blue Ridge. *Top*: Calcium effect with 95 % Bayesian confidence intervals. *Left*: Geography effect. *Right*: Standard errors of geography effect with the lakes superimposed.

where one needs the R package `maps` pre-installed for the `map` command to work. The R^2 and the decomposition π_β of the "explained" variation in pH can be obtained from the summaries of the fit:

```
sum.lake.a <- summary(fit.lake.a,diag=TRUE)
sum.lake.a$r.squared
# 0.5300598
round(sum.lake.a$pi,3)
# log(cal)     geog
#    0.702    0.298
```

see §3.7 for the definitions of R^2 and π_β.

4.4 Splines on the Sphere

To estimate functions on small geographic regions, one may use thin-plate splines on $(-\infty, \infty)^2$, but surface curvature can not be ignored on larger geographic regions or for global mapping. Using the spherical coordinates (r, θ, ϕ) in $(-\infty, \infty)^3$, where

$$x_{\langle 1\rangle} = r \sin\theta \cos\phi, \quad x_{\langle 2\rangle} = r \sin\theta \sin\phi, \quad x_{\langle 3\rangle} = r \cos\theta$$

for $r \in [0, \infty)$, $\theta \in [0, \pi]$, $\phi \in [0, 2\pi]$, and setting $r = 1$, we consider the unit sphere $\mathcal{X} = \mathcal{S}$ in this section; θ is the angle from the north pole, off by $\pi/2$ from the latitude, and ϕ is the longitude.

The infinitesimal rectangle with corners at points (θ, ϕ), $(\theta + d\theta, \phi)$, $(\theta, \phi + d\phi)$, and $(\theta + d\theta, \phi + d\phi)$ on the unit sphere $\mathcal{S}$ has area $\sin\theta \, d\theta \, d\phi$ (Problem 4.17), so integrals on $\mathcal{S}$ are given by

$$\int_{\mathcal{S}} f(x) dx = \int_0^{2\pi} \int_0^{\pi} f(\theta, \phi) \sin\theta \, d\theta \, d\phi.$$

Much like the standard Fourier expansion (4.2) for functions on the circle, square integrable functions on $\mathcal{S}$ can be expressed as

$$f(x) = f(\theta, \phi) = \sum_{\mu=0}^{\infty} \sum_{k=-\mu}^{\mu} f_{\mu,k} H_{\mu,k}(\theta, \phi), \tag{4.33}$$

where $H_{\mu,k}(\theta, \phi)$ are the spherical harmonics.

After a brief review of pertinent facts concerning the spherical harmonics (§4.4.1), we discuss the Laplacian on the sphere and introduce the spherical splines of Wahba (1981) (§4.4.2). The reproducing kernels under standard Laplacian penalties are inconvenient to compute as sums of infinite series, but closed form formulas are available under slightly modified penalties (§4.4.3). As an illustration, a global temperature map is estimated in §4.4.4 using a spherical spline.

4.4.1 Spherical Harmonics

Spherical harmonics is widely used in mathematical physics. Treatments of the classical subject can be found in numerous sources, such as Byerly (1959, Chap. 6), where the results quoted below are developed.

The spherical harmonics of degree μ, order k are given by

$$H_{\mu,k}(\theta, \phi) = \begin{cases} \kappa_{\mu,k} P_\mu^k(\cos\theta) \cos(k\phi), & k \geq 0, \\ \kappa_{\mu,k} P_\mu^{-k}(\cos\theta) \sin(k\phi), & k < 0, \end{cases} \tag{4.34}$$

where $\kappa_{\mu,k} = \kappa_{\mu,-k}$ are normalizing constants to be specified below and $P_\mu^k(z)$, $k \geq 0$ are the associated Legendre functions on $z \in [-1,1]$ that solve differential equations

$$\frac{d}{dz}\Big((1-z^2)\frac{df}{dz}\Big) + \Big(\mu(\mu+1) - \frac{k^2}{1-z^2}\Big)f = 0. \tag{4.35}$$

It is known that for $k \geq 0$,

$$\int_{-1}^{1} P_\mu^k(z) P_\nu^k(z) dz = \delta_{\mu,\nu} \frac{2\,(\mu+k)!}{(2\mu+1)(\mu-k)!},$$

where $\delta_{\mu,\nu}$ is the Kronecker delta, so to make $\{H_{\mu,k}\}$ an orthonormal basis, one needs

$$\kappa_{\mu,k}^2 = \begin{cases} \frac{2\mu+1}{2\pi}\frac{(\mu-k)!}{(\mu+k)!}, & k > 0, \\ \frac{2\mu+1}{4\pi}, & k = 0. \end{cases}$$

For $x, y \in \mathcal{S}$, one has

$$\sum_{k=-\mu}^{\mu} H_{\mu,k}(x) H_{\mu,k}(y) = \frac{2\mu+1}{4\pi} P_\mu(x \cdot y), \tag{4.36}$$

where $P_\mu(z) = P_\mu^0(z)$ is the μth Legendre polynomial and $x \cdot y$ is the cosine of the angle between x and y. Also of interest is the expansion

$$\frac{1}{\sqrt{1+h^2-2hz}} = \sum_{\mu=0}^{\infty} h^\mu P_\mu(z), \tag{4.37}$$

where the left-hand side is known as the generating function of $P_\mu(z)$.

4.4.2 Laplacian on the Sphere and Spherical Splines

Consider the Laplacian operator on $(-\infty,\infty)^3$,

$$\Delta = \frac{\partial^2}{\partial x_{\langle 1\rangle}^2} + \frac{\partial^2}{\partial x_{\langle 2\rangle}^2} + \frac{\partial^2}{\partial x_{\langle 3\rangle}^2}, \tag{4.38}$$

which is rotation invariant (Problem 4.18). Under the spherical coordinates (r,θ,ϕ), (4.38) transforms into (Problem 4.19)

$$\Delta = \frac{1}{r^2}\frac{\partial}{\partial r}\Big(r^2\frac{\partial}{\partial r}\Big) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\Big(\sin\theta\frac{\partial}{\partial\theta}\Big) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2}{\partial\phi^2}, \tag{4.39}$$

and upon setting $r = 1$, one has the Laplace-Beltrami operator on $\mathcal{S}$,

$$\Delta = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\Big(\sin\theta\frac{\partial}{\partial\theta}\Big) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\phi^2}. \tag{4.40}$$

Noting that

$$\frac{1}{\sin\theta}\frac{\partial}{\partial\theta} = -\frac{\partial}{\partial\cos\theta},$$

(4.40) can be written as

$$\Delta = \frac{\partial}{\partial\cos\theta}\Big((1-\cos^2\theta)\frac{\partial}{\partial\cos\theta}\Big) + \frac{1}{1-\cos^2\theta}\frac{\partial^2}{\partial\phi^2}.$$

It then follows, as $P_\mu^k(z)$ solve (4.35), that

$$\Delta H_{\mu,k}(\theta,\phi) = -\mu(\mu+1)H_{\mu.k}(\theta,\phi). \tag{4.41}$$

For $m > 0$ an even integer, define

$$J_m(f) = \int_{\mathcal{S}} \big\{\Delta^{m/2} f(x)\big\}^2 dx = \int_0^{2\pi}\int_0^{\pi} \big\{\Delta^{m/2} f(\theta,\phi)\big\}^2 \sin\theta\, d\theta\, d\phi.$$

A spherical spline on $\mathcal{X} = \mathcal{S}$ minimizes over $\eta \in \big\{f : J_m(f) < \infty\big\}$ the penalized least squares functional

$$\frac{1}{n}\sum_{i=1}^n \big(Y_i - \eta(x_i)\big)^2 + \lambda J_m(\eta). \tag{4.42}$$

By (4.41), the spherical harmonics $H_{\mu,k}(\theta,\phi)$ are the eigenfunctions of $J_m(f)$ with eigenvalues $\big\{\mu(\mu+1)\big\}^m$ (see §9.1 for discussion of eigenfunctions and eigenvalues); the eigenvalues, when put in an increasing order $\rho_\nu \uparrow \infty$, grow at a rate $\rho_\nu \asymp \nu^m$. To the inner product

$$(f,g) = \frac{1}{(4\pi)^2}\Big(\int_{\mathcal{S}} f dx\Big)\Big(\int_{\mathcal{S}} g dx\Big) + \int_{\mathcal{S}} \big(\Delta^{m/2} f\big)\big(\Delta^{m/2} g\big) dx$$

in $\mathcal{H} = \big\{f : J_m(f) < \infty\big\}$ corresponds the reproducing kernel

$$\begin{aligned} R(x,y) &= 1 + \sum_{\mu=1}^{\infty}\sum_{k=-\mu}^{\mu} \frac{1}{\mu^m(\mu+1)^m} H_{\mu,k}(x)H_{\mu,k}(y) \\ &= 1 + \frac{1}{4\pi}\sum_{\mu=1}^{\infty} \frac{2\mu+1}{\mu^m(\mu+1)^m} P_\mu(x\cdot y), \end{aligned} \tag{4.43}$$

where (4.36) is plugged in; note that for $f(x)$ as given in (4.33),

$$J_m(f) = \sum_{\mu=1}^{\infty}\sum_{k=-\mu}^{\mu} \big\{\mu^m(\mu+1)^m\big\} f_{\mu,k}^2. \tag{4.44}$$

where $f_{\mu.k} = \int_{\mathcal{S}} f(x)H_{\mu,k}(x)dx$ are the Fourier coefficients. The formulation also extends to m odd via (4.44).

4.4.3 Reproducing Kernels in Closed Forms

The infinite sum in (4.43) is inconvenient to compute, but a slight modification of $J_m(f)$ solves the problem. Combining (4.37) with the identity

$$\frac{1}{r!}\int_0^1 (1-h)^r h^\mu dh = \frac{1}{(\mu+1)\cdots(\mu+r+1)},$$

one has

$$\frac{q_r(z)}{r!} = \frac{1}{r!}\int_0^1 \frac{(1-h)^r}{\sqrt{1+h^2-2hz}}dh = \sum_{\mu=0}^{\infty}\frac{P_\mu(z)}{(\mu+1)\cdots(\mu+r+1)},$$

where $q_r(z)$ can be obtained analytically through recursive formulas; see Problem 4.20. One thus may use the reproducing kernel in closed form,

$$\begin{aligned}\tilde{R}(x,y) &= 1+\sum_{\mu=1}^{\infty}\sum_{k=-\mu}^{\mu}\frac{H_{\mu,k}(x)H_{\mu,k}(y)}{(\mu+1/2)(\mu+1)\cdots(\mu+2m-1)}\\ &= 1+\frac{1}{2\pi}\sum_{\mu=1}^{\infty}\frac{P_\mu(x\cdot y)}{(\mu+1)\cdots(\mu+2m-1)}\\ &= 1+\frac{q_{2m-2}(x\cdot y)-1/(2m-1)}{2\pi(2m-2)!},\end{aligned} \tag{4.45}$$

which is associated with the penalty

$$\tilde{J}_m(f) = \sum_{\mu=1}^{\infty}\sum_{k=-\mu}^{\mu}\{(\mu+1/2)(\mu+1)\cdots(\mu+2m-1)\}f_{\mu,k}^2. \tag{4.46}$$

$\tilde{J}_m(f)$ and $J_m(f)$ are equivalent penalties with the ratios of their respective eigenvalues satisfying $\tilde{\rho}_\nu/\rho_\nu \to 1$, where $\tilde{\rho}_\nu$ and ρ_ν are in increasing order.

The expressions of $q_r(z)$ for $r = 0,\ldots,10$ were listed in Wahba (1981) with an erratum in Wahba (1982). For $m = 2,3,4$, one needs

$$2\,q_2(z) = a(12w^2-4w)-6cw+6w+1, \tag{4.47}$$

$$\begin{aligned}12\,q_4(z) = {}& a(840w^4-720w^3+72w^2)+420w^3\\ &+c(-420w^3+220w^2)-150w^2-4w+3,\end{aligned} \tag{4.48}$$

$$\begin{aligned}30\,q_6(z) = {}& a(27720w^6-37800w^5+12600w^4-600w^3)\\ &+13860w^5+c(-13860w^5+14280w^4-2772w^3)\\ &-11970w^4+1470w^3+15w^2-3w+5,\end{aligned} \tag{4.49}$$

where $w = (1-z)/2$, $a = \log(1+1/\sqrt{w})$, and $c = 2\sqrt{w}$.

4.4.4 Case Study: Global Temperature Map

Maps of meteorological quantities constructed from records registered at weather stations are valuable tools in numerous applications such as climate change studies. A data frame `climate` involving 690 weather stations worldwide can be found in the R package `assist` by Yuedong Wang and Chunlei Ke. The data were repackaged in a data frame `clim` in `gss`, with elements `temp` (average temperatures from December 1980 to February 1981) and `geog` (geographic locations of weather stations); `geog` is a matrix with the latitude in the first column and the longitude in the second, in degrees. The range of latitude is $[-90, 90]$ and that of longitude is $[-180, 180]$, shifted from the ranges of (θ, ϕ) in the proceeding mathematical treatments.

To fit a temperature map to the data, one may use:

```
data(clim)
fit.clim <- ssanova(temp~geog,type=list(geog="sphere"),
                    data=clim,id.basis=1:dim(clim)[1])
```

$\tilde{J}_m(f)$ of (4.46) is used in (4.42) in the place of $J_m(f)$ and the default order is $m = 2$; the order could be alternatively specified via something like `type=list(geog=list("sphere",3))`, but only $m = 2, 3, 4$ are implemented, with $\tilde{R}(x, y)$ of (4.45) constructed using the formulas given in (4.47)–(4.49). To evaluate the fit on a regular grid, try:

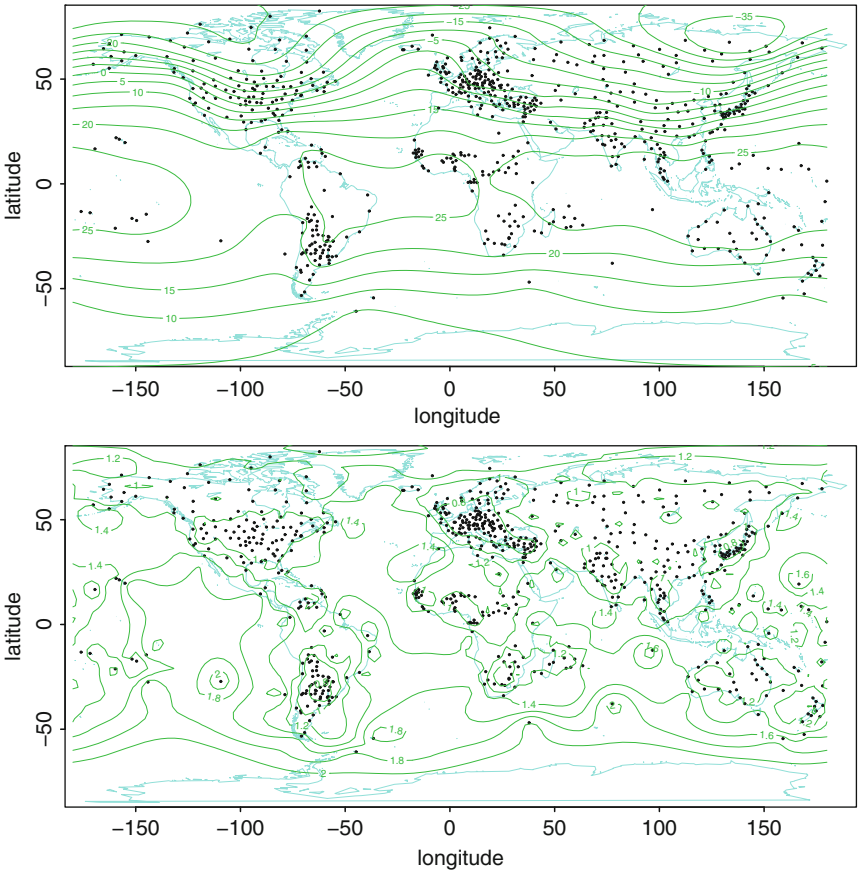
```
lat <- seq(-90,90,length=61)
lon <- seq(-180,180,length=121)
new <- cbind(rep(lat,rep(121,61)),rep(lon,61))
est <- predict(fit.clim,data.frame(geog=I(new)),se=TRUE)
```

We can now plot the estimated temperature on the world map as shown in the top frame of Fig. 4.5:

```
library(maps)
map("world",interior=FALSE,col=5); box()
points(clim$geog[,2:1],pch=19,cex=.2)
contour(lon,lat,matrix(est$fit,121,61),
        col=3,lwd=.5,labcex=.4,add=TRUE)
```

Replacing `est$fit` in the `contour` command above by `est$se`, one gets the bottom frame of Fig. 4.5.

To keep things simple, we used $q = n$ in the fit above, but the execution of `ssanova` and `predict` was slow; the execution would be much faster with the `ssanova0` suite but the fit was a bit rough. Due to the uneven distribution of the weather stations, a simple random subset $\{z_j\} \subset \{x_i\}$ is likely to have over-representations in Japan and Europe while missing out areas with sparsely scattering stations. One however could try to "fill" the space by prohibiting the selected z_j's to be too close to each other, a strategy implemented in the following R function `subset.sphere`.



FIGUR E 4.5. Global temperature map. *Top*: Estimated temperature. *Bottom*: Standard errors. The weather stations are superimposed as *dots* and the *shore lines* are on the background.

```
subset.sphere <- function(x,size,tol) {
  nobs <- dim(x)[1]; x <- x/180*pi
  pick <- samp <- sample(1:nobs,1)
  while(length(samp)<size) {
    if (!(length(pick)-nobs)) stop("list exhausted")
    wk <- sample((1:nobs)[-pick],1)
    pick <- c(pick,wk); okey <- TRUE
      for (j in samp) {
        if (cos.angle(x[wk,],x[j,])>tol) {
          okey <- FALSE; break
        }
      }
      if (okey) samp <- c(samp,wk)
```

```
    }
  samp
}
cos.angle <- function(x,y) {
  cos(x[1])*cos(y[1])*cos(x[2]-y[2])+sin(x[1])*sin(y[1])
}
```

Note also that the default $q \asymp n^{2/9}$ for cubic splines assumes $\rho_\nu \asymp \nu^4$ for the eigenvalues ρ_ν of $J(f)$, but one has $\rho_\nu \asymp \nu^m$ for spherical splines and we used $m = 2$, so the choice of q here would be *ad hoc*.

To select a "space-filling" random subset $\{z_j\} \subset \{x_i\}$, say of size $q = 200$ and with z_j's at least 3 angular degrees apart from each other, and fit the model, one may use:

```
id.select <- subset.sphere(clim$geog,200,cos(3/180*pi))
fit0.clim <- ssanova(temp~geog,type=list(geog="sphere"),
                     data=clim,id.basis=id.select)
```

The commands `ssanova` and `predict` execute much faster now. The z_j's could be identified on the map via

```
points(clim$geog[id.select,2:1],pch=19,cex=.2,col=2)
```

One may check on the consistency by comparing the fits on the grid, numerically or graphically.

4.5 L-Splines

Consider functions on $[0, 1]$. Given a general differential operator L and a weight function $h(x) > 0$, the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1 (L\eta)^2(x)h(x)dx \tag{4.50}$$

is called an L-spline. The polynomial splines of §2.3 are special cases of L-splines. In applications where $\mathcal{N}_L = \{f : Lf = 0\}$ provides a more natural parametric model than a low-order polynomial, an L-spline other than a polynomial spline often provides a better estimate.

Popular examples of L-splines include trigonometric splines and Chebyshev splines, which we will discuss in §§4.5.1 and 4.5.2, respectively; of interest are the characterization of the null space of L and the derivation of the reproducing kernels. A general approach to the construction of reproducing kernels for L-splines is described next (§4.5.3), and data analysis with L-splines is illustrated through a real-data example (§4.5.4). Based on a special structure in the reproducing kernel from the general construction of §4.5.3, a fast algorithm similar to that of §3.10.1 is also described for the computation of L-splines (§4.5.5).

4.5.1 Trigonometric Splines

Consider $f \in \mathcal{P}[0,1]$ periodic with $\int_0^1 f dx = a_0 = 0$. The differential operator

$$L_2 = D^2 + (2\pi)^2 \tag{4.51}$$

has a null space $\mathcal{N}_L = \text{span}\{\cos 2\pi x, \sin 2\pi x\}$. To the inner product

$$2\left(\int_0^1 f(x)\cos 2\pi x dx\right)\left(\int_0^1 g(x)\cos 2\pi x dx\right) + 2\left(\int_0^1 f(x)\sin 2\pi x dx\right)\left(\int_0^1 g(x)\sin 2\pi x dx\right)$$

in $\mathcal{N}_L$ corresponds the reproducing kernel

$$2\cos 2\pi x \cos 2\pi y + 2\sin 2\pi x \sin 2\pi y = 2\cos 2\pi(x-y). \tag{4.52}$$

Take $h(x) = 1$ and define $\mathcal{H} = \left\{f : f \in \mathcal{P}[0,1], a_0 = 0, \int_0^2 (L_2 f)^2 dx < \infty\right\}$, and consider $\mathcal{H}_L = \mathcal{H} \ominus \mathcal{N}_L$ with the inner product $\int_0^1 (L_2 f)(L_2 g) dx$. Since

$$f(x) = \sum_{\mu=2}^{\infty} (a_\mu \cos 2\pi\mu x + b_\mu \sin 2\pi\mu x)$$

for $f \in \mathcal{H}_L$, the reproducing kernel of $\mathcal{H}_L$ is easily seen to be

$$\begin{aligned} R_2(x,y) &= \sum_{\mu=2}^{\infty} \frac{2}{(2\pi)^4(\mu^2-1)^2}(\cos 2\pi\mu x \cos 2\pi\mu y + \sin 2\pi\mu x \sin 2\pi\mu y) \\ &= \sum_{\mu=2}^{\infty} \frac{2\cos 2\pi\mu(x-y)}{(2\pi)^4(\mu^2-1)^2}; \end{aligned} \tag{4.53}$$

see Problem 4.21. Note that for $f \in \mathcal{P}[0,1]$,

$$\int_0^1 (L_2 f)^2 dx = (2\pi)^4 a_0^2 + \frac{(2\pi)^4}{2} \sum_{\mu=2}^{\infty} (a_\mu^2 + b_\mu^2)(\mu^2-1)^2, \tag{4.54}$$

so $\int_0^1 (L_2 f)^2 dx < \infty$ is equivalent to $\int_0^1 \ddot{f}^2 dx < \infty$; compare (4.54) with (4.3) of §4.2 for $m = 2$. Naturally, one would like to add the constant term a_0 back in as an unpenalized term, which can be achieved by using $\lambda \sum_{\mu=2}^{\infty} (a_\mu^2 + b_\mu^2)(\mu^2-1)^2$ as the penalty term instead of $\lambda \int_0^1 (L_2 f)^2 dx$. This procedure is technically an application of the partial spline technique discussed in §4.1.

More generally, the differential operator

$$L_{2r} = \left(D^2 + (2\pi)^2\right) \cdots \left(D^2 + (2\pi r)^2\right) \tag{4.55}$$

has a null space $\mathcal{N}_L = \text{span}\{\cos 2\pi\nu x, \sin 2\pi\nu x, \nu = 1, \ldots, r\}$. In the space

$$\mathcal{H}_L = \left\{f : f = \sum_{\mu=r+1}^{\infty}(a_\mu \cos 2\pi\mu x + b_\mu \sin 2\pi\mu x), \int_0^1 (L_{2r}f)^2 dx < \infty\right\}$$

with the inner product $\int_0^1 (L_{2r}f)(L_{2r}g)dx$, the reproducing kernel is seen to be

$$R_{2r}(x, y) = \sum_{\mu=r+1}^{\infty} \frac{2\cos 2\pi\mu(x-y)}{(2\pi)^{4r}(\mu^2-1)^2\cdots(\mu^2-r^2)^2}; \tag{4.56}$$

see Problem 4.22.

With the differential operator

$$L_3 = D\big(D^2 + (2\pi)^2\big), \tag{4.57}$$

the null space $\mathcal{N}_L = \text{span}\{1, \cos 2\pi x, \sin 2\pi x\}$ automatically contains the constant term. To the inner product

$$\left(\int_0^1 f dx\right)\left(\int_0^1 g dx\right) + 2\left(\int_0^1 f(x)\cos 2\pi x dx\right)\left(\int_0^1 g(x)\cos 2\pi x dx\right) + 2\left(\int_0^1 f(x)\sin 2\pi x dx\right)\left(\int_0^1 g(x)\sin 2\pi x dx\right)$$

in $\mathcal{N}_L$ corresponds the reproducing kernel

$$1 + 2\cos 2\pi x \cos 2\pi y + 2\sin 2\pi x \sin 2\pi y.$$

Take $h(x) = 1$ and define $\mathcal{H} = \left\{f : f \in \mathcal{P}[0, 1], \int_0^1 (L_3 f)^2 dx < \infty\right\}$. Corresponding to the inner product $\int_0^1 (L_3 f)(L_3 g)dx$, the reproducing kernel of $\mathcal{H}_L = \mathcal{H} \ominus \mathcal{N}_L$ is seen to be

$$R_3(x, y) = \sum_{\mu=2}^{\infty} \frac{2\cos 2\pi\mu(x-y)}{(2\pi)^6\mu^2(\mu^2-1)^2}; \tag{4.58}$$

see Problem 4.23. For $f \in \mathcal{P}[0,1]$,

$$\int_0^1 (L_3 f)^2 dx = \frac{(2\pi)^6}{2} \sum_{\mu=2}^{\infty} (a_\mu^2 + b_\mu^2)\mu^2(\mu^2-1)^2, \tag{4.59}$$

so $\int_0^1 (L_3 f)^2 dx < \infty$ is equivalent to $\int_0^1 \left(f^{(3)}\right)^2 dx < \infty$; compare (4.59) with (4.3) of §4.2 for $m = 3$.

In general, the differential operator

$$L_{2r+1} = D\big(D^2 + (2\pi)^2\big) \cdots \big(D^2 + (2\pi r)^2\big) \tag{4.60}$$

has a null space $\mathcal{N}_L = \text{span}\{1, \cos 2\pi\nu x, \sin 2\pi\nu x, \nu = 1, \ldots, r\}$. In the space

$$\mathcal{H}_L = \big\{f : f = \textstyle\sum_{\mu=r+1}^{\infty}(a_\mu \cos 2\pi\mu x + b_\mu \sin 2\pi\mu x), \int_0^1 (L_{2r+1} f)^2 dx < \infty\big\}$$

with the inner product $\int_0^1 (L_{2r+1} f)(L_{2r+1} g)dx$, the reproducing kernel is given by

$$R_{2r+1}(x,y) = \sum_{\mu=r+1}^{\infty} \frac{2\cos 2\pi\mu(x-y)}{(2\pi)^{4r+2}\mu^2(\mu^2-1)^2 \cdots (\mu^2-r^2)^2}; \tag{4.61}$$

see Problem 4.24.

The infinite sums in (4.56) and (4.61) are inconvenient to compute, but similar to the treatment in §4.4.3, one may obtain closed form reproducing kernels under slightly modified, indirectly defined $J(f)$. For example, $R_2(x,y)$ in (4.53) may be replaced by

$$\tilde{R}_2(x,y) = -k_4(x-y) - 2\cos 2\pi(x-y)/(2\pi)^4, \tag{4.62}$$

and $R_3(x,y)$ in (4.58) by

$$\tilde{R}_3(x,y) = k_6(x-y) - 2\cos 2\pi(x-y)/(2\pi)^6;$$

recall (4.4) and Problem 4.3. Pasting (4.52) and (4.62) together, one has a kernel decomposition in $\mathcal{H} = \big\{f : f \in \mathcal{P}[0,1], \tilde{J}_2(f) < \infty\big\}$,

$$R(x,y) = 1 + 2\cos 2\pi(x-y) + \tilde{R}_2(x,y), \tag{4.63}$$

where $\tilde{J}_2(f) = (2\pi)^4 \sum_{\mu=2}^{\infty}(a_\mu^2 + b_\mu^2)\mu^4/2$ is equivalent to $\int_0^2 (L_2 f)^2 dx$ in $\mathcal{H} \ominus$ span$\{1, \cos 2\pi x, \sin 2\pi x\}$; (4.63) defines a one-way ANOVA decomposition for periodic functions on $[0,1]$, with $2\cos 2\pi(x-y)$ representing a two-dimensional "parametric contrast" and $\tilde{R}_2(x,y)$ representing the "nonparametric contrast." This differs only slightly from a cubic periodic spline discussed in §4.2, just with the base frequency pulled out of the penalty. To specify (4.63) for a variable `x` in `ssanova`, say, one may use something like

```
ssanova(y~x,type=list(x=list("trig",c(0,1))))
```

where the domain does not have to be $[0,1]$; the syntax parallels that for periodic splines as seen in §4.2.1.

4.5.2 Chebyshev Splines

Let $w_i(x) \in \mathcal{C}^{(m-i+1)}[0,1]$, $i = 1, \ldots, m$, be strictly positive functions with $w_i(0) = 1$. Consider the differential operator

$$L_m = D_m \cdots D_1, \tag{4.64}$$

where $D_i f = D(f/w_i)$.

The null space $\mathcal{N}_L$ of L_m is spanned by

$$\begin{aligned} \phi_1(x) &= w_1(x) \\ \phi_2(x) &= w_1(x) \int_0^x w_2(t_2) dt_2 \\ &\vdots \\ \phi_m(x) &= w_1(x) \int_0^x w_2(t_2) dt_2 \cdots \int_0^{t_{m-1}} w_m(t_m) dt_m, \end{aligned} \tag{4.65}$$

which form a so-called Chebyshev system on $[0,1]$, in the sense that

$$\det\big[\phi_j(x_i)\big]_{i,j=1}^m > 0 \quad \text{for all } x_1 < x_2 < \cdots < x_m,\ [x_1, x_m] \subseteq [0,1];$$

see Schumaker (1981, §2.5, Theorem 9.2). The functions ϕ_ν in (4.65) also form an extended Chebyshev system, in the sense that

$$\det\big[\phi_j^{(i-1)}(x)\big]_{i,j=1}^m > 0, \quad \forall x \in [0,1];$$

see Karlin and Studden (1966, §1.2, Theorem 1.2 on page 379). The matrix

$$\big[\phi_j^{(i-1)}(x)\big]_{i,j=1}^m = \begin{pmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_m(x) \\ \dot{\phi}_1(x) & \dot{\phi}_2(x) & \cdots & \dot{\phi}_m(x) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(m-1)}(x) & \phi_2^{(m-1)}(x) & \cdots & \phi_m^{(m-1)}(x) \end{pmatrix},$$

is known as the Wronskian matrix of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^T$. Write $L_0 = I$, $L_1 = D_1$, …, $L_{m-1} = D_{m-1} \cdots D_1$. One has $(L_\mu \phi_\nu)(0) = \delta_{\mu+1,\nu}$, $\mu = 0, \ldots, m-1$, $\nu = 1, \ldots, m$, where $\delta_{\mu,\nu}$ is the Kronecker delta. It follows that $\sum_{\nu=1}^m \phi_\nu(x)\phi_\nu(y)$ is the reproducing kernel of $\mathcal{N}_L$ corresponding to the inner product

$$\sum_{\nu=1}^m (L_{\nu-1} f)(0)(L_{\nu-1} g)(0).$$

Actually, $\{\phi_\nu\}_{\nu=1}^m$ form an orthonormal basis of $\mathcal{N}_L$ under the given inner product.

Define $\mathcal{H} = \left\{f : \int_0^1 (L_m f)^2 h dx < \infty\right\}$ and denote $\mathcal{H}_L = \mathcal{H} \ominus \mathcal{N}_L$. For $f \in \mathcal{H}_L$, noting that $(L_\nu f)(0) = 0$, $\nu = 0, \ldots, m-1$, it is straightforward to verify that

$$\begin{aligned} f(x) &= w_1(x) \int_0^x w_2(t_2) dt_2 \cdots \int_0^{t_{m-1}} w_m(t_m) dt_m \int_0^{t_m} (L_m f)(u) du \\ &= \int_0^x (L_m f)(u) du \left\{ w_1(x) \int_u^x w_2(t_2) dt_2 \cdots \int_u^{t_{m-1}} w_m(t_m) dt_m \right\} \\ &= \int_0^x G(x;u)(L_m f)(u) du, \end{aligned} \tag{4.66}$$

where

$$G(x;u) = \begin{cases} w_1(x) \int_u^x w_2(t_2) dt_2 \cdots \int_u^{t_{m-1}} w_m(t_m) dt_m, & u \le x, \\ 0, & u > x; \end{cases} \tag{4.67}$$

see Problem 4.25. The function $G(x;u)$ is called a Green's function associated with the differential operator L_m. After some algebra, one has the expression

$$G(x;u) = \begin{cases} \sum_{\nu=1}^m \phi_\nu(x)\psi_\nu(u), & u \le x, \\ 0, & u > x, \end{cases} \tag{4.68}$$

where

$$\begin{aligned} \psi_\nu(u) = &- \int_0^u w_{\nu+1}(t_{\nu+1}) dt_{\nu+1} \\ &\times \int_u^{t_{\nu+1}} w_{\nu+2}(t_{\nu+2}) dt_{\nu+2} \cdots \int_u^{t_{m-1}} w_m(t_m) dt_m, \end{aligned}$$

$\nu = 1, \ldots, m-2$, $\psi_{m-1}(u) = -\int_0^u w_m(t_m) dt_m$, and $\psi_m(u) = 1$ (Problem 4.26). Write

$$R_x(y) = \int_0^1 G(x;u) G(y;u) \big(h(u)\big)^{-1} du.$$

It is straightforward to verify that $(L_\nu R_x)(0) = 0, \nu = 0, \ldots, m-1$, and that $(L_m R_x)(y) = G(x;y)/h(y)$; see Problem 4.27. Hence, by (4.66), the reproducing kernel in $\mathcal{H}_L$ corresponding to an inner product $\int_0^1 (L_m f)(L_m g) h dx$ is given by

$$R_L(x,y) = \int_0^1 G(x;u) G(y;u) \big(h(u)\big)^{-1} du. \tag{4.69}$$

By Theorem 2.5, the reproducing kernel of $\mathcal{H}$ under the inner product

$$\sum_{\nu=1}^{m}(L_{\nu-1}f)(0)(L_{\nu-1}g)(0)+\int_0^1(L_mf)(L_mg)h dx$$

is seen to be

$$R(x,y)=\sum_{\nu=1}^{m}\phi_\nu(x)\phi_\nu(y)+\int_0^1 G(x;u)G(y;u)\big(h(u)\big)^{-1}du.$$

Parallel to (2.6) on page 34, one has, for $f\in\mathcal{H}$, the generalized Taylor expansion,

$$f(x)=\sum_{\nu=1}^{m}(L_{\nu-1}f)(0)\phi_\nu(x)+\int_0^x G(x;u)(L_mf)(u)du.$$

Since $G(x;u)=0$, $u>x$, one may rewrite (4.69) as

$$R_L(x,y)=\int_0^{x\wedge y}G(x;u)G(y;u)\big(h(u)\big)^{-1}du.$$

It is easy to see that the calculus of this section applies on any domain of the form $[0,a]$, where a is not necessarily scaled to 1.

Example 4.5 (Polynomial splines) Setting $w_i(x)=1$, $i=1,\ldots,m$, and $h(x)=1$, one gets the polynomial splines of §2.3.1; see Problem 4.28. □

Example 4.6 (Exponential splines) Setting $w_i(x)=e^{\beta_i x}$, $i=1,\ldots,m$, where $\beta_i\geq 0$ with the strict inequality holding for $i>1$, one gets the so-called exponential splines; see, e.g., Schumaker (1981, §9.9). Denote $\alpha_i=\sum_{j=1}^{i}\beta_j$. It is easy to verify that

$$L_\nu=e^{-\alpha_\nu x}(D-\alpha_\nu)\cdots(D-\alpha_1),\quad \nu=1,\ldots,m,$$

and that L_m has the null space $\mathcal{N}_L=\text{span}\{e^{\alpha_i x},i=1,\ldots,m\}$.

As a specific case, consider $m=2$, $\beta_1=0$, and $\beta_2=\theta$. One has $L_2=e^{-\theta x}(D-\theta)D$ with the null space $\mathcal{N}_L=\text{span}\{1,e^{\theta x}\}$. The orthonormal basis of $\mathcal{N}_L$ consists of $\phi_1=1$ and $\phi_2=(e^{\theta x}-1)/\theta$. Now,

$$G(x;u)=\int_u^x e^{\theta t}dt=\theta^{-1}\big(e^{\theta x}-e^{\theta u}\big)=\phi_2(x)-\phi_2(u),\quad u\leq x,$$

so

$$R_L(x,y) = \int_0^{x\wedge y} G(x;u)G(y;u)\big(h(u)\big)^{-1}du$$
$$= \int_0^{x\wedge y} \big(\phi_2(x)-\phi_2(u)\big)\big(\phi_2(y)-\phi_2(u)\big)\big(h(u)\big)^{-1}du.$$

The generalized Taylor expansion is seen to be

$$f(x) = f(0)+\dot f(0)\phi_2(x)+\int_0^x \big(\phi_2(x)-\phi_2(u)\big)e^{-\theta u}\big(\ddot f(u)-\theta\dot f(u)\big)du, \quad (4.70)$$

which, after a change of variable $\tilde x = \phi_2(x)$, reduces to

$$g(\tilde x) = g(0) + \dot g(0)\tilde x + \int_0^{\tilde x} (\tilde x - \tilde u)\ddot g(\tilde u)d\tilde u, \quad (4.71)$$

where $g(\tilde x) = f\big(\phi_2^{-1}(\tilde x)\big)$ for ϕ_2^{-1} the inverse of ϕ_2; see Problem 4.29. With $1/h(x) = d\phi_2/dx = e^{\theta x}$,

$$R_L(x,y) = \int_0^{x\wedge y} G(x;u)G(y;u)\frac{d\phi_2(u)}{du}du$$
$$= \int_0^{x\wedge y} \big(\phi_2(x)-\phi_2(u)\big)\big(\phi_2(y)-\phi_2(u)\big)d\phi_2(u), \quad (4.72)$$

so the formulation virtually yields a cubic spline in $\tilde x = \phi_2(x)$; compare (4.72) with (2.10) on page 35 for $m = 2$.

More generally, an exponential spline on $[0,a]$ with $\beta_1 = 0$, $\beta_i = \theta$, $i = 2,\dots,m$, and $h(x) = e^{-\theta x}$ reduces to a polynomial spline in $\tilde x = \phi_2(x)$ with a penalty proportional to $\int_0^{\phi_2(a)} \big(g^{(m)}(\tilde x)\big)^2 d\tilde x$; see Problem 4.30. □

Example 4.7 (Hyperbolic splines) For $m = 2r$, let $\beta_1 = 0$, $\beta_i > 0$, $i = 2,\dots,r$, and denote $\alpha_i = \sum_{j=1}^i \beta_j$, $i = 1,\dots,r$. Setting $w_i(x) = e^{\beta_i x}$, $i = 1,\dots,r$, $w_{r+1}(x) = e^{-2\alpha_r x}$, and $w_{r+i}(x) = w_{r-i+2}(x)$, $i = 2,\dots,r$, one gets the so-called hyperbolic splines; see Schumaker (1981, §9.9). It is straightforward to verify that

$$L_\nu = e^{-\alpha_\nu x}(D-\alpha_\nu)\cdots(D-\alpha_1), \qquad \nu = 1,\dots,r,$$
$$L_{2r-\nu+1} = e^{\alpha_\nu x}(D+\alpha_\nu)\cdots(D+\alpha_r)$$
$$\times (D-\alpha_r)\cdots(D-\alpha_1), \qquad \nu = r,\dots,1.$$

The differential operator

$$L_m = D(D+\alpha_2)\cdots(D+\alpha_r)(D-\alpha_r)\cdots(D-\alpha_2)D$$

has the null space $\mathcal{N}_L = \text{span}\{1, x, e^{-\alpha_\nu x}, e^{\alpha_\nu x}, \nu = 2,\dots,r\}$.

Consider the case with $r = 2$ and $\beta_2 = \theta$. One has $L_4 = D(D+\theta)(D-\theta)D$ with the null space $\mathcal{N}_L = \text{span}\{1, x, e^{-\theta x}, e^{\theta x}\}$. The orthonormal basis of $\mathcal{N}_L$ consists of $\phi_1 = 1$, $\phi_2 = (e^{\theta x} - 1)/\theta$, $\phi_3 = (\cosh\theta x - 1)/\theta^2$, and $\phi_4 = (\sinh\theta x - \theta x)/\theta^3$. The Green's function is

$$G(x,u) = \big(\sinh\theta(x-u) - \theta(x-u)\big)/\theta^3,$$

for $u \le x$; see Problem 4.31.

More generally, with $\beta_i = \theta$, $i = 2, \ldots, r$, one can show that, for $\phi_2 = (e^{\theta x} - 1)/\theta$,

$$\phi_\nu = \frac{\phi_2^{\nu-1}(x)}{(\nu-1)!},$$

$$\phi_{r+\nu} = \int_0^x \frac{\phi_2^{\nu-1}(v)}{(\nu-1)!} \frac{\big(\phi_2(x) - \phi_2(v)\big)^{r-1}}{(r-1)!} \frac{d\phi_2(v)}{\big(1+\theta\phi_2(v)\big)^{2r-1}}, \tag{4.73}$$

$\nu = 1, \ldots, r$, and that

$$G(x;u) = \int_u^x \frac{\big(\phi_2(v)-\phi_2(u)\big)^{r-1}}{(r-1)!} \frac{\big(\phi_2(x)-\phi_2(v)\big)^{r-1}}{(r-1)!} \frac{d\phi_2(v)}{\big(1+\theta\phi_2(v)\big)^{2r-1}}, \tag{4.74}$$

for $u \le x$; see Problem 4.32. □

4.5.3 General Construction

Consider a differential operator of the form

$$L = D^m + \sum_{j=0}^{m-1} a_j(x) D^j. \tag{4.75}$$

This effectively covers the operator L_m of (4.64) as a special case, which can be written as

$$L_m = \Big\{\prod_{i=1}^m w_i(x)\Big\}^{-1}\big(D^m + \textstyle\sum_{j=0}^{m-1} a_j(x)D^j\big),$$

since the factor $\big\{\prod_{i=1}^m w_i(x)\big\}^{-1}$ can be absorbed into the weight function $h(x)$. When $a_j \in \mathcal{C}^{(m-j)}[0,1]$, it is known that the null space of L, $\mathcal{N}_L = \big\{f : Lf = 0\big\}$, is an m-dimensional linear subspace of infinitely differentiable functions; see Schumaker (1981, §10.1). Let ϕ_ν, $\nu = 1, \ldots, m$, be a basis of such an $\mathcal{N}_L$. The Wronskian matrix of $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^T$,

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} \phi_1(x) & \phi_2(x) & \cdots & \phi_m(x) \\ \dot\phi_1(x) & \dot\phi_2(x) & \cdots & \dot\phi_m(x) \\ \vdots & \vdots & & \vdots \\ \phi_1^{(m-1)}(x) & \phi_2^{(m-1)}(x) & \cdots & \phi_m^{(m-1)}(x) \end{pmatrix},$$

is known to be nonsingular, $\forall x \in [0,1]$; see Schumaker (1981, §10.1). Since $W(\boldsymbol{\phi})(0)$ is invertible, $\sum_{\nu=1}^{m} f^{(\nu-1)}(0)g^{(\nu-1)}(0)$ forms an inner product in $\mathcal{N}_L$ (Problem 4.33). Define $\tilde{\boldsymbol{\phi}} = \big[W(\boldsymbol{\phi})(0)\big]^{-T}\boldsymbol{\phi}$. It is easy to verify that $\tilde{\phi}_\nu^{(\mu-1)}(0) = \delta_{\mu,\nu}$, $\mu,\nu = 1,\ldots,m$, so $\tilde{\phi}_\nu$, $\nu = 1,\ldots,m$, form an orthonormal basis of $\mathcal{N}_L$ and $\sum_{\nu=1}^{m} \tilde{\phi}_\nu(x)\tilde{\phi}_\nu(y)$ is its reproducing kernel.

An m-dimensional function space on an interval is called a Chebyshev space if it has a basis that is a Chebyshev system on the interval; see Schumaker (1981, §2.5). A function in an m-dimensional Chebyshev space is uniquely determined by its values on m distinctive points on the interval. The space $\mathcal{N}_L$ may not be a Chebyshev space on $[0,1]$, but for some $\delta > 0$, it is always a Chebyshev space on intervals shorter than δ; see Schumaker (1981, Theorem 10.5).

Define $\mathcal{H} = \big\{f : \int_0^1 (Lf)^2 h dx < \infty\big\}$ and $\mathcal{H}_L = \mathcal{H} \ominus \mathcal{N}_L$. Let $\psi_\nu(x)$, $\nu = 1,\ldots,m$, be the entries of the last column of $\big[W(\boldsymbol{\phi})(x)\big]^{-1}$. It is easy to see that

$$\begin{aligned} &\sum_{\nu=1}^{m} \phi_\nu^{(j)}(x)\psi_\nu(x) = 0, \quad j = 0,\ldots,m-2, \\ &\sum_{\nu=1}^{m} \phi_\nu^{(m-1)}(x)\psi_\nu(x) = 1. \end{aligned} \tag{4.76}$$

Write

$$G(x;u) = \begin{cases} \sum_{\nu=1}^{m} \phi_\nu(x)\psi_\nu(u), & u \le x, \\ 0, & u > x; \end{cases} \tag{4.77}$$

we show that $G(x;u)$ is a Green's function associated with L in (4.75). For $g \in \mathcal{L}_2[0,1]$, define

$$\tilde{g}(x) = \int_0^1 G(x;u)g(u)du.$$

Using (4.76), it is easy to calculate

$$\begin{aligned} \tilde{g}^{(j)}(x) &= \sum_{\nu=1}^{m} \phi_\nu^{(j)}(x) \int_0^x \psi_\nu(u)g(u)du, \quad j = 0,\ldots,m-1, \\ \tilde{g}^{(m)}(x) &= \sum_{\nu=1}^{m} \phi_\nu^{(m)}(x) \int_0^x \psi_\nu(u)g(u)du + g(x); \end{aligned} \tag{4.78}$$

see Problem 4.34. Hence, $\tilde{g}^{(j)}(0) = 0$, $j = 0,\ldots,m-1$, and since $\phi_\nu^{(m)}(x) + \sum_{j=0}^{m-1} a_j(x)\phi_\nu^{(j)}(x) = 0$ as $\phi_\nu \in \mathcal{N}_L$, $L\tilde{g} = g$. It follows that for $f \in \mathcal{H}_L$,

$$f(x) = \int_0^x G(x;u)(Lf)(u)du,$$

and corresponding to the inner product $\int_0^1 (Lf)(Lg)h dx$, one has the reproducing kernel

$$R_L(x,y) = \int_0^{x\wedge y} G(x;u)G(y;u)\big(h(u)\big)^{-1} du. \tag{4.79}$$

For $f \in \mathcal{H}$, one has the generalized Taylor expansion

$$f(x) = \sum_{\nu=1}^{m} f^{(\nu-1)}(0)\tilde{\phi}_\nu(x) + \int_0^x G(x;u)(Lf)(u)du.$$

Example 4.8 (Cubic spline) Consider $L = D^2$ with $\phi_1(x) = 1$ and $\phi_2(x) = x$. The Wronskian matrix and its inverse are respectively

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \big[W(\boldsymbol{\phi})(x)\big]^{-1} = \begin{pmatrix} 1 & -x \\ 0 & 1 \end{pmatrix}.$$

One has $\tilde{\phi}_1 = \phi_1$, $\tilde{\phi}_2 = \phi_2$, and $G(x;u) = x - u$ for $u \le x$. The results coincide with those derived in §2.3.1 and Example 4.5. □

Example 4.9 (Exponential spline) Consider $L = (D-\theta)D$ for $\theta > 0$ with $\phi_1(x) = 1$ and $\phi_2(x) = e^{\theta x}$. The Wronskian matrix and its inverse are respectively

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} 1 & e^{\theta x} \\ 0 & \theta e^{\theta x} \end{pmatrix} \quad \text{and} \quad \big[W(\boldsymbol{\phi})(x)\big]^{-1} = \begin{pmatrix} 1 & -\theta^{-1} \\ 0 & \theta^{-1}e^{-\theta x} \end{pmatrix}.$$

One has $\tilde{\phi}_1(x) = 1$, $\tilde{\phi}_2(x) = (e^{\theta x} - 1)/\theta$, and

$$G(x;u) = e^{-\theta u}\big(\tilde{\phi}_2(x) - \tilde{\phi}_2(u)\big)$$

for $u \le x$. The results agree with those of Example 4.6 for $m = 2$, after adjusting for the factor $e^{-\theta x}$ appearing in the operator $L_2 = e^{-\theta x}(D-\theta)D$ of Example 4.6. □

Example 4.10 Consider $L = (D+\theta)D$ for $\theta > 0$ with $\phi_1(x) = 1$ and $\phi_2(x) = e^{-\theta x}$. Substituting $-\theta$ for θ in Example 4.9, one has $\tilde{\phi}_1(x) = 1$, $\tilde{\phi}_2(x) = (1 - e^{-\theta x})/\theta$, and

$$G(x;u) = e^{\theta u}\big(\tilde{\phi}_2(x) - \tilde{\phi}_2(u)\big)$$

for $u \le x$. With a weight function $h(x) = e^{3\theta x}$, one obtains a cubic spline in $\tilde{\phi}_2(x)$. □

Example 4.11 (Trigonometric splines) Consider $L = D^2 + (2\pi)^2$ with $\phi_1(x) = \sin 2\pi x$ and $\phi_2(x) = \cos 2\pi x$. The Wronskian matrix and its inverse are respectively

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} \sin 2\pi x & \cos 2\pi x \\ (2\pi)\cos 2\pi x & -(2\pi)\sin 2\pi x \end{pmatrix}$$

and

$$\big[W(\boldsymbol{\phi})(x)\big]^{-1} = \begin{pmatrix} \sin 2\pi x & (2\pi)^{-1}\cos 2\pi x \\ \cos 2\pi x & -(2\pi)^{-1}\sin 2\pi x \end{pmatrix}.$$

One has $\tilde{\phi}_1 = \cos 2\pi x$, $\tilde{\phi}_2(x) = (2\pi)^{-1}\sin 2\pi x$, and

$$G(x;u) = \frac{1}{2\pi}\sin 2\pi(x-u)$$

for $u \le x$. The reproducing kernel of $\mathcal{H}_L$ corresponding to the inner product $\int_0^1 (Lf)(Lg)dx$ is thus

$$\begin{aligned} R_L &= \frac{1}{(2\pi)^2}\int_0^{x\wedge y} \sin 2\pi(x-u)\sin 2\pi(y-u)du \\ &= \frac{(x\wedge y)\cos 2\pi(x-y)}{2(2\pi)^2} - \frac{\sin 2\pi(x+y) - \sin 2\pi|x-y|}{4(2\pi)^3}. \end{aligned} \tag{4.80}$$

This reproducing kernel is different from the one given in (4.53) of §4.5.1, where the constant and the nonperiodic functions are excluded.

Now, consider $L = D(D^2 + (2\pi)^2)$ with $\phi_1(x) = 1$, $\phi_2(x) = \sin 2\pi x$, and $\phi_3(x) = \cos 2\pi x$. The Wronskian matrix and its inverse are respectively

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} 1 & \sin 2\pi x & \cos 2\pi x \\ 0 & (2\pi)\cos 2\pi x & -(2\pi)\sin 2\pi x \\ 0 & -(2\pi)^2\sin 2\pi x & -(2\pi)^2\cos 2\pi x \end{pmatrix}$$

and

$$\big[W(\boldsymbol{\phi})(x)\big]^{-1} = \begin{pmatrix} 1 & 0 & (2\pi)^{-2} \\ 0 & (2\pi)^{-1}\cos 2\pi x & -(2\pi)^{-2}\sin 2\pi x \\ 0 & -(2\pi)^{-1}\sin 2\pi x & -(2\pi)^{-2}\cos 2\pi x \end{pmatrix}.$$

One has $\tilde{\phi}_1(x) = 1$, $\tilde{\phi}_2(x) = (2\pi)^{-1}\sin 2\pi x$, $\tilde{\phi}_3(x) = (2\pi)^{-2}(1 - \cos 2\pi x)$, and

$$G(x;u) = \frac{1}{(2\pi)^2}\big(1 - \cos 2\pi(x-u)\big)$$

for $u \le x$. The reproducing kernel of $\mathcal{H}_L$ corresponding to the inner product $\int_0^1 (Lf)(Lg)dx$ is thus

$$\begin{aligned} R_L &= \frac{1}{(2\pi)^4} \int_0^{x\wedge y} \big(1-\cos 2\pi(u-x)\big)\big(1-\cos 2\pi(u-y)\big)du \\ &= \frac{x\wedge y}{(2\pi)^4} - \frac{\sin 2\pi x + \sin 2\pi y - \sin 2\pi|x-y|}{(2\pi)^5} \\ &\quad + \frac{(x\wedge y)\cos 2\pi(x-y)}{2(2\pi)^4} + \frac{\sin 2\pi(x+y) - \sin 2\pi|x-y|}{4(2\pi)^5}. \end{aligned} \tag{4.81}$$

This reproducing kernel is different from the one given in (4.58) of §4.5.1, where the nonperiodic functions are excluded. □

Example 4.12 (Logistic spline) Consider $D\big(D - \gamma\theta e^{-\theta x}/(1+\gamma e^{-\theta x})\big)$ for $\theta, \gamma > 0$, with $\phi_1(x) = 1$ and $\phi_2(x) = 1/(1+\gamma e^{-\theta x})$. The Wronskian matrix and its inverse are respectively

$$W(\boldsymbol{\phi})(x) = \begin{pmatrix} 1 & (1+\gamma e^{-\theta x})^{-1} \\ 0 & \gamma\theta e^{-\theta x}(1+\gamma e^{-\theta x})^{-2} \end{pmatrix}$$

and

$$\big[W(\boldsymbol{\phi})(x)\big]^{-1} = \begin{pmatrix} 1 & -(\gamma\theta)^{-1}e^{\theta x}(1+\gamma e^{-\theta x}) \\ 0 & (\gamma\theta)^{-1}e^{\theta x}(1+\gamma e^{-\theta x})^2 \end{pmatrix}.$$

One has $\tilde{\phi}_1(x) = 1$,

$$\tilde{\phi}_2(x) = \frac{(1+\gamma)^2}{\gamma\theta}\Big(\frac{1}{1+\gamma e^{-\theta x}} - \frac{1}{1+\gamma}\Big),$$

and

$$G(x;u) = \frac{e^{\theta u}(1+\gamma e^{-\theta u})^2}{(1+\gamma)^2}\big(\tilde{\phi}_2(x) - \tilde{\phi}_2(u)\big)$$

for $u \le x$. With a weight function $h(x) \propto e^{3\theta x}(1+\gamma e^{-\theta x})^6$, one gets a cubic spline in $\tilde{\phi}_2(x)$. □

4.5.4 Case Study: Weight Loss of Obese Patient

Obese patients on a weight rehabilitation program tend to lose adipose tissue at a diminishing rate as the treatment progresses. A data set concerning the weight loss of a male patient can be found in the R package `MASS`, as a data frame `wtloss` with two elements, `Weight` and `Days`. A nonlinear regression model was considered in Venables and Ripley (2002, Chap. 8),

$$Y = \beta_0 + \beta_1 2^{-x/\theta} + \epsilon, \tag{4.82}$$

where Y was the weight at x days after the start of the rehabilitation program. The least squares estimates of the parameters were given by $\hat{\beta}_0 = 81.374$, $\hat{\beta}_1 = 102.68$, and $\hat{\theta} = 141.91$. The parameter β_0 may be interpreted as the ultimate lean weight, β_1 the total amount to be lost, and θ the "half-decay" time.

Note that $2^{-x/\theta} = e^{-\tilde{\theta}x}$ with $\tilde{\theta} = \log 2/\theta$. The nonlinear model (4.82) is in the null space of the differential operator $L = (D + \theta)D$ considered in Example 4.10. To allow for possible departures from the parametric model, we consider a cubic spline in $e^{-\tilde{\theta}x}$, which is an L-spline with $L = (D + \tilde{\theta})D$ and $h(x) = e^{3\tilde{\theta}x}$. Fixing $\tilde{\theta}$, the smoothing parameter can be selected using the GCV score $V(\lambda)$ of (3.23), and to choose $\tilde{\theta}$, one may compare the minimum $V(\lambda)$ scores obtained with different $\tilde{\theta}$. Note that Theorem 3.3 is still useful in this situation. The R code below finds the GCV estimate of the parameter $\tilde{\theta}$:

```
library(MASS); data(wtloss)
tmp.fun <- function(theta) {
    theta <- theta/100
    ssanova0(Weight~exp(-theta*Days),data=wtloss)$score
}
nlm(tmp.fun,1)$estimate
# 0.4884628
```

The `tmp.fun` function returns the minimum $V(\lambda)$ score for fixed $\tilde{\theta}$. The `nlm` function finds the minimal point of `tmp.fun` using a quasi-Newton algorithm with numerical derivatives; see Dennis and Schnabel (1996) for algorithmic details. The scaling of `theta` in `tmp.fun` was introduced so that `nlm` would use appropriate differencing steps for the calculation of numerical derivatives. The solution corresponds to $\theta = \log(2)/0.004884628 = 141.9038$, matching the least squares estimate in the parametric model. The minimum $V(\lambda)$ for $\tilde{\theta} = 0.004885$ is 0.8166.

The fit with $\tilde{\theta} = 0.004885$ can now be calculated and plotted as the solid line in the left frame of Fig. 4.6, which is indistinguishable from the parametric fit plotted as the dashed line; the data are superimposed as circles. A cubic spline in x is also calculated and superimposed as the long dashed line, which is nearly indistinguishable from the other two fits; the minimum $V(\lambda)$ for the cubic spline is 0.9283.

```
# calculate the L-spline and cubic spline fits
wtloss$dd <- exp(-.004885*wtloss$Days)
wtloss.fit1 <- ssanova0(Weight~dd,data=wtloss)
wtloss.fit2 <- ssanova0(Weight~Days,data=wtloss)
tt <- seq(0,250,length=101)
est1 <- predict(wtloss.fit1,
                data.frame(dd=exp(-.004885*tt)),se=TRUE)
est2 <- predict(wtloss.fit2,data.frame(Days=tt),se=TRUE)
```

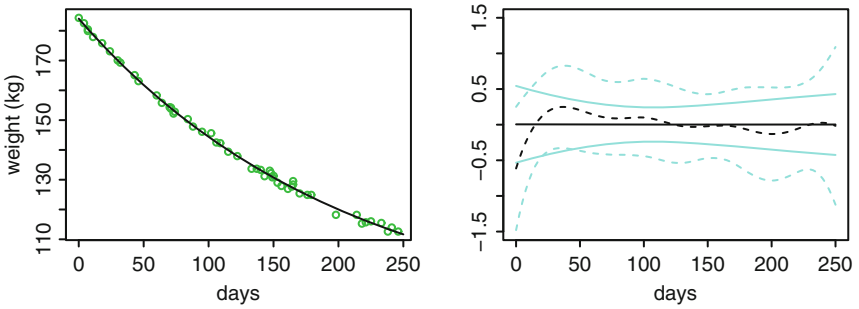


FIGURE 4.6. Weight loss of obese patient. *Left*: The L-spline fit, the cubic spline fit, and the nonlinear parametric fit are visually indistinguishable; the data are superimposed in *circles*. *Right*: Spline fits and Bayesian confidence intervals minus the parametric fit; the L-spline fit is in *solid lines* and the cubic spline fit in *dashed lines*.

```
est0 <- 81.374+102.68*2^(-tt/141.91)
# plot the fits
plot(wtloss$Days,wtloss$Weight,col=3)
lines(tt,est1$fit)
lines(tt,est0,lty=2)
lines(tt,est2$fit,lty=5)
```

In the right frame of Fig. 4.6, the L-spline and cubic spline fits and their corresponding Bayesian confidence intervals are plotted after the parametric fit is subtracted from each curve.

```
plot(tt,est1$fit-est0,type="l",ylim=c(-1.5,1.5))
lines(tt,est2$fit-est0,lty=3)
lines(tt,est1$fit-est0-1.96*est1$se,col=5)
lines(tt,est1$fit-est0+1.96*est1$se,col=5)
lines(tt,est2$fit-est0-1.96*est2$se,lty=3,col=5)
lines(tt,est2$fit-est0+1.96*est2$se,lty=3,col=5)
```

It is clear that the L-spline fit has smaller standard errors than the cubic spline fit.

Admittedly, the relative noise level in the weight measurements is way below what one usually sees in stochastic data, although the displayed nonlinearity might not be detectable at a higher noise level. To confirm the usefulness of the demonstrated techniques on "ordinary" data, a simple simulation is conducted below. On $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$, responses are generated according to $Y_i = 5 + 3e^{-4x_i} + 2e^{-8x_i} + \epsilon_i$, where $\epsilon_i \sim N(0, 0.5^2)$:

```
set.seed(5732)
tt <- ((1:100)-.5)/100
yy <- 5+3*exp(-4*tt)+2*exp(-8*tt)+.5*rnorm(tt)
```

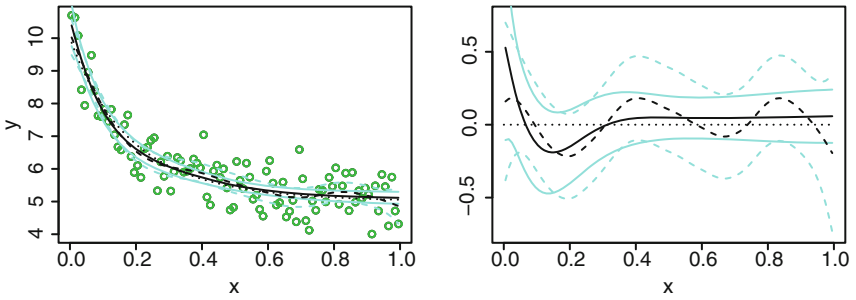


FIGURE 4.7. L-spline simulation. *Left*: The L-spline fit and the corresponding Bayesian confidence intervals are in *solid lines*, the cubic spline fit in *dashed lines*, the test function in *dotted line*, and the data are superimposed as *circles*. *Right*: The *left frame curves* minus the test function.

L-Splines with $L = (D + \tilde{\theta})D$ and $h(x) = e^{3\tilde{\theta}x}$ are tried, and the $\tilde{\theta}$ that minimizes the minimum $V(\lambda)$ is obtained:

```
tmp.fun <- function(theta) {
    ssanova0(yy~exp(-theta*tt))$score
}
nlm(tmp.fun,4)$estimate
# 4.790263
```

The minimum $V(\lambda)$ for $\tilde{\theta} = 4.7903$ is 0.3375, and that for a cubic spline in x is 0.3556:

```
ssanova0(yy~exp(-4.7903*tt))$score
# 0.3374706
ssanova0(yy~tt)$score
# 0.3555772
```

One can now calculate and plot the fits as in the left frame of Fig. 4.7, where the L-spline fit and the corresponding Bayesian confidence intervals are drawn in solid and faded solid lines, the cubic spline in dashed and faded dashed lines, and the test function in the dotted line. The data are superimposed as circles.

```
ttt <- exp(-4.7903*tt)
fit.L <- ssanova0(yy~ttt)
est.L <- predict(fit.L,data.frame(ttt=ttt),se=TRUE)
fit.c <- ssanova0(yy~tt)
est.c <- predict(fit.c,data.frame(tt=tt),se=TRUE)
#
plot(tt,yy,col=3)
lines(tt,est.L$fit)
lines(tt,est.L$fit-1.96*est.L$se,col=5)
```

```
lines(tt,est.L$fit+1.96*est.L$se,col=5)
lines(tt,est.c$fit,lty=2)
lines(tt,est.c$fit-1.96*est.c$se,col=5,lty=2)
lines(tt,est.c$fit+1.96*est.c$se,col=5,lty=2)
lines(tt,5+3*exp(-4*tt)+2*exp(-8*tt),lty=3)
```

Subtracting the test function from each of the lines, one gets the right frame of Fig. 4.7.

4.5.5 Fast Algorithm

We now describe a fast algorithm for the computation of L-splines due to Heckman and Ramsay (2000). The algorithm assumes that $x_1 < x_2 < \cdots < x_n$, that the space $\mathcal{N}_L = \text{span}\{\phi_\nu, \nu = 1, \ldots, m\}$ is Chebyshev on the intervals $[x_{i+1}, x_{i+m}]$, $i = 1, \ldots, n-m$, and that

$$R_L(x,y) = \int_0^1 G(x;u)G(y;u)\big(h(u)\big)^{-1}du,$$

where $G(x;u)$ is of the form $\sum_{\nu=1}^m \phi_\nu(x)\psi_\nu(u)$ for $u \le x$. For replicated data, one may work with (3.37) on page 73 and select λ using $U(\lambda)$ of (3.38) or $V(\lambda)$ of (3.39). As with the algorithms of §3.10, the score $M(\lambda)$ and the posterior variances of §3.3 are not available through the fast algorithm, according to current knowledge.

Without loss of generality, consider (3.10) on page 64. From $S_w^T\mathbf{c}_w = 0$, $\mathbf{c}_w = T\boldsymbol{\gamma}$ for some $n \times (n-m)$ matrix T of full column rank satisfying $S_w^T T = O$. Premultiplying the first equation of (3.10) by T^T and plugging in $T\boldsymbol{\gamma}$ for $\mathbf{c}_w$, one has

$$\big(T^TQ_wT + (n\lambda)\,T^TT\big)\boldsymbol{\gamma} = T^T\mathbf{Y}_w.$$

Now, since $\mathbf{Y}_w - \hat{\mathbf{Y}}_w = \big(I - A_w(\lambda)\big)\mathbf{Y}_w = (n\lambda)\mathbf{c}_w$, one has

$$I - A_w(\lambda) = (n\lambda)\,T\big(T^TQ_wT + (n\lambda)\,T^TT\big)^{-1}T^T.$$

If T can be chosen such that both T^TT and T^TQ_wT are banded, then the $O(n)$ algorithm of §3.10.1 can be readily applied to calculate L-splines with λ selected by $U(\lambda)$ or $V(\lambda)$.

Let $\mathbf{t}_i$ be an n-vector with $i-1$ leading zeros, $n-m-i$ trailing zeros, and the middle $m+1$ entries $t_{j,i}$ satisfying conditions $t_{i,i} \neq 0$ and

$$\textstyle\sum_{j=i}^{i+m} t_{j,i}\sqrt{w_j}\mathbf{s}_j = 0,$$

where $\mathbf{s}_j^T = \big(\phi_1(x_j), \ldots, \phi_m(x_j)\big)$ is the jth row of S; the latter condition is possible because $\mathbf{s}_j$, $j = i, \ldots, i+m$, are linearly dependent, and the former condition is possible because $\mathbf{s}_j$, $j = i+1, \ldots, i+m$, are linearly

independent since x_j's are distinctive and $\mathcal{N}_L$ is Chebyshev on $[x_{i+1}, x_{i+m}]$. Set $T = (\mathbf{t}_1, \ldots, \mathbf{t}_{n-m})$. It is obvious that $S_w^T T = O$ and that T is of full column rank. It is also clear that $T^T T$ is banded with bandwidth $2m+1$. Plugging in the expression $G(x;u) = \sum_{\nu=1}^m \phi_\nu(x)\psi_\nu(u)$ for $u \le x$, the (k,l)th entry of Q_w can be written as

$$\begin{aligned} q_{k,l} = \sqrt{w_k}\sqrt{w_l} R_L(x_k, x_l) &= \int_0^{x_k \wedge x_l} G(x_k;u) G(x_l;u) \big(h(u)\big)^{-1} du \\ &= (\sqrt{w_k}\mathbf{s}_k)^T P(x_k \wedge x_l)(\sqrt{w_l}\mathbf{s}_l), \end{aligned}$$

where $P(v)$ is $m \times m$ with the (μ,ν)th entry $\int_0^v \psi_\mu(u)\psi_\nu(u)\big(h(u)\big)^{-1} du$. Now, for $i < j$, consider the (i,j)th entry of $T^T Q_w T$,

$$\begin{aligned} r_{i,j} &= \sum_{k,l} t_{k,i} (\sqrt{w_k}\mathbf{s}_k)^T P(x_k \wedge x_l)(\sqrt{w_l}\mathbf{s}_l) t_{l,j} \\ &= \sum_{k \le l} t_{k,i} (\sqrt{w_k}\mathbf{s}_k)^T P(x_k)(\sqrt{w_l}\mathbf{s}_l) t_{l,j} \\ &\quad + \sum_{k > l} t_{k,i} (\sqrt{w_k}\mathbf{s}_k)^T P(x_l)(\sqrt{w_l}\mathbf{s}_l) t_{l,j} \\ &= r'_{i,j} + r''_{i,j}, \end{aligned}$$

say. By the construction of T, $\sum_{l=k}^n t_{l,j}(\sqrt{w_l}\mathbf{s}_l) = 0$ unless $j < k \le j+m$, and $t_{k,i} = 0$ unless $i \le k \le i+m$, so one must have $j < i+m$, or $j-i < m$, for $r'_{i,j} \neq 0$. Similarly, one must have $j - i < m$ for $r''_{i,j} \neq 0$. Hence, $T^T Q_w T$ is banded with bandwidth $2m-1$.

The algorithm relies on the particular form $\int_0^1 G(x;u)G(y;u)\big(h(u)\big)^{-1} du$ of reproducing kernels with $G(x;u) = \sum_{\nu=1}^m \phi_\nu(x)\psi_\nu(u)$, $u \le x$, so it does not work with the reproducing kernels of §§2.3.3 and 4.5.1.

4.6 Bibliographic Notes

Section 4.1

The idea of partial splines appeared in the literature since the early 1980s in various forms. Extensive discussion on the subject can be found in Wahba (1990, Chap. 6) and Green and Silverman (1994, Chap. 4).

Section 4.2

Fourier series expansion and discrete Fourier transform are among elementary tools in the spectral analysis of time series; see, e.g., Priestley (1981, §§4.2, 6.1 and 7.6) for comprehensive treatments of related subjects.

The spectral decomposition of (4.6) was found in Craven and Wahba (1979), where it was used to analyze the behavior of generalized cross-validation. Some other uses of this decomposition can be found in Gu (1993a) and Stein (1993). The materials of §4.2.3 are largely repackaged arguments found in Craven and Wahba (1979) and Wahba (1985).

Section 4.3

Standard references on thin-plate splines are Duchon (1977), Meinguet (1979) and Wahba and Wendelberger (1980), upon which much of the materials were drawn. See also Wahba (1990, §§2.4 and 2.5). Tensor product splines with thin-plate marginals were proposed and illustrated by Gu and Wahba (1993b).

Section 4.4

The materials of this section, sans §4.4.4, are largely drawn from Wahba (1981). The mathematics concerning spherical harmonics, Laplacian, and Legendre functions is widely used in mathematical physics; results concerning Legendre functions can be found in Abramowitz and Stegun (1964, Chap. 8). Further discussions concerning the fitting of the temperature map in §4.4.4 can be found in Kim and Gu (2004).

Section 4.5

A comprehensive treatment of L-splines from a numerical analytical perspective can be found in Schumaker (1981, Chaps. 9 and 10), upon which a large portion of the technical materials presented here were drawn. The Chebyshev splines of §4.5.2 were found in Kimeldorf and Wahba (1971); see also Wahba (1990, §1.2). Further results on L-splines and their statistical applications can be found in Ramsay and Dalzell (1991), Ansley, Kohn, and Wong (1993), Dalzell and Ramsay (1993), Wang and Brown (1996) and Heckman and Ramsay (2000).

4.7 Problems

Section 4.2

4.1 Verify (4.3) for $f \in \mathcal{P}[0,1]$.

4.2 For $f \in \mathcal{P}[0,1]$ and $R_x(y) = R(x,y)$ with $R(x,y)$ as given in (4.4), prove that

$$\left(\int_0^1 f dy\right)\left(\int_0^1 R_x dy\right) + \int_0^1 f^{(m)} R_x^{(m)} dy = f(x).$$

4.3 Compare (4.4) with (2.18) on page 37 to verify that $R(x,y) = 1 + (-1)^{m-1}k_{2m}(x-y)$.

4.4 Let Γ be the Fourier matrix with the (i,j)th entry

$$\frac{1}{\sqrt{n}} \exp\left\{2\pi \mathbf{i} \frac{(i-1)(j-1)}{n}\right\}.$$

(a) Verify that $\Gamma^H\Gamma = \Gamma\Gamma^H = I$.

(b) Verify that (4.6) implies $Q = \Gamma\Lambda\Gamma^H$.

4.5 Verify (4.13) using the orthogonality conditions in (4.12).

4.6 Prove that when (4.14) holds for some $p > 2$ and $B_2 = \lambda^{-2}B(\lambda)|_{\lambda=0} > 0$, then $\lambda^{-2}B(\lambda) - B_2 = o(1)$ for $\lambda = o(1)$.

4.7 Verify (4.15).

4.8 For $c_\nu > 0$ and z_ν and y_ν complex, show that

$$\frac{1}{2}\left|\sum_\nu c_\nu(\bar{z}_\nu y_\nu + z_\nu \bar{y}_\nu)\right| \le \left\{\sum_\nu c_\nu |z_\nu|^2\right\}^{1/2}\left\{\sum_\nu c_\nu |y_\nu|^2\right\}^{1/2},$$

where $\bar{z}$ denotes the conjugate of z.

Section 4.3

4.9 On a d-dimensional real domain, the space of polynomials of up to $(m-1)$ total order is of dimension $M = \binom{d+m-1}{d}$.

(a) Show that the number of polynomials of up to $(m-1)$ total order is the same as the number of ways to choose $m-1$ objects from a set of $d+1$ objects *allowing repeats*.

(b) Show that the number of ways to choose $m-1$ objects from a set of $d+1$ objects *allowing repeats* is the same as the number of ways to choose $m-1$ objects from a set of $(d+1)+(m-1)-1 = d+m-1$ objects *disallowing repeats*, hence is $\binom{d+m-1}{m-1} = \binom{d+m-1}{d}$.

4.10 The quadratic functional $J_m^d(f)$ of (4.17) is rotation invariant.

(a) Write $D_i = \partial/\partial x_{\langle i\rangle}$. Show that

$$J_m^d(f) = \int\cdots\int\left\{\sum_{i_1=1}^d \cdots \sum_{i_m=1}^d \left(D_{i_1}\cdots D_{i_m} f\right)^2\right\} dx_{\langle 1\rangle}\cdots dx_{\langle d\rangle}.$$

(b) Let P be a $d \times d$ orthogonal matrix with the (i,j)th entry $p_{i,j}$ and let $y = P^T x$. Note that the Jacobian of the orthogonal transform $y = P^T x$ is 1. Write $\tilde{D}_j = \partial/\partial y_{\langle j \rangle}$. Verify that $\tilde{D}_j = \sum_{i=1}^{d} p_{i,j} D_i$.

(c) Calculating $J_m^d(f)$ with respect to y, the integrand is given by

$$\sum_{j_1} \cdots \sum_{j_m} (\tilde{D}_{j_1} \cdots \tilde{D}_{j_m} f)^2 = \sum_{j_1} \cdots \sum_{j_m} \left\{ \prod_{k=1}^{m} \left(\sum_{i=1}^{d} p_{i,j_k} D_i \right) f \right\}^2$$
$$= \sum_{j_1} \cdots \sum_{j_m} \left\{ \sum_{i_1} \cdots \sum_{i_m} (p_{i_1,j_1} \cdots p_{i_m,j_m})(D_{i_1} \cdots D_{i_m} f) \right\}^2 .$$

Expanding $\left\{ \sum_{i_1} \cdots \sum_{i_m} (p_{i_1,j_1} \cdots p_{i_m,j_m}) \big(D_{i_1} \cdots D_{i_m} f\big) \right\}^2$, one gets d^m square terms and $\binom{d^m}{2}$ cross-terms. Summing over $(j_1, \ldots, j_m)$, show that the square terms add up to $\sum_{i_1} \cdots \sum_{i_m} \big(D_{i_1} \cdots D_{i_m} f\big)^2$ and the cross-terms all vanish.

4.11 Given (4.20), prove (4.27).

4.12 Verify (4.28).

4.13 Verify (4.29).

4.14 Let ψ_ν, $\nu = 1, \ldots, M$, be a set of polynomials that span $\mathcal{N}_J$ and $\tilde{S}$ an $n \times M$ matrix with the (i,ν)th entry $\psi_\nu(x_i)$. Write $\tilde{S} = F_1 R$ for the QR-decomposition of $\tilde{S}$. Verify that $\boldsymbol{\phi} = \sqrt{n} R^{-T} \boldsymbol{\psi}$ forms an orthonormal basis in $\mathcal{N}_J$ with the inner product $(f,g)_0 = \sum_{i=1}^{n} f(x_i) g(x_i)/n$ and that F_1 has the (i,ν)th entry $\phi_\nu(x_i)/\sqrt{n}$.

4.15 Verify (4.30).

4.16 Verify (4.31).

Section 4.4

4.17 Show that the infinitesimal parallelogram on the unit sphere with corners at (θ,ϕ), $(\theta+d\theta,\phi)$, $(\theta,\phi+d\phi)$, and $(\theta+d\theta,\phi+d\phi)$ is a rectangle and has area $\sin\theta\, d\theta\, d\phi$.

(a) The line segment from (θ,ϕ) to $(\theta+d\theta,\phi)$ has Cartesian coordinates $(\cos\theta\cos\phi, \cos\theta\sin\phi, -\sin\theta)\, d\theta$, and the segment from (θ,ϕ) to $(\theta,\phi+d\phi)$ has coordinates $(-\sin\theta\sin\phi, \sin\theta\cos\phi, 0)\, d\phi$.

(b) The line segments in (a) are perpendicular and are of lengths $d\theta$ and $\sin\theta\, d\phi$, respectively.

4.18 Show that the Laplacian of (4.38) is rotation invariant using the technique of Problem 4.10.

4.19 Following Williamson (1899, Chap. 22), verify (4.39).

(a) For $x = \rho\cos\phi$, $y = \rho\sin\phi$, show that

$$\frac{\partial(\rho,\phi)}{\partial(x,y)^T} = \left(\frac{\partial(x,y)}{\partial(\rho,\phi)^T}\right)^{-1} = \begin{pmatrix}\cos\phi & -\sin\phi/\rho \\ \sin\phi & \cos\phi/\rho\end{pmatrix},$$

so by the chain rule,

$$\frac{\partial}{\partial x} = \cos\phi\,\frac{\partial}{\partial\rho} - \frac{\sin\phi}{\rho}\,\frac{\partial}{\partial\phi},$$
$$\frac{\partial}{\partial y} = \sin\phi\,\frac{\partial}{\partial\rho} + \frac{\cos\phi}{\rho}\,\frac{\partial}{\partial\phi}.$$

(b) Verify that

$$\frac{\partial^2}{\partial x^2} = \cos^2\phi\,\frac{\partial^2}{\partial\rho^2} + \frac{\sin 2\phi}{\rho}\Big(\frac{1}{\rho}\frac{\partial}{\partial\phi} - \frac{\partial^2}{\partial\rho\partial\phi}\Big) + \frac{\sin^2\phi}{\rho}\Big(\frac{\partial}{\partial\rho} + \frac{1}{\rho}\frac{\partial^2}{\partial\phi^2}\Big)$$
$$\frac{\partial^2}{\partial y^2} = \sin^2\phi\,\frac{\partial^2}{\partial\rho^2} - \frac{\sin 2\phi}{\rho}\Big(\frac{1}{\rho}\frac{\partial}{\partial\phi} - \frac{\partial^2}{\partial\rho\partial\phi}\Big) + \frac{\cos^2\phi}{\rho}\Big(\frac{\partial}{\partial\rho} + \frac{1}{\rho}\frac{\partial^2}{\partial\phi^2}\Big),$$

so

$$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial\rho^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2}.$$

(c) With $z = r\cos\theta$, $\rho = r\sin\theta$, and (x, y) given above,

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \frac{1}{\rho}\frac{\partial}{\partial\rho} + \frac{1}{\rho^2}\frac{\partial^2}{\partial\phi^2}.$$

Substituting $\rho = r\sin\theta$ and

$$\frac{\partial}{\partial\rho} = \sin\theta\,\frac{\partial}{\partial r} + \frac{\cos\theta}{r}\,\frac{\partial}{\partial\theta},$$

some algebra yields (4.39).

4.20 Derive recursive formulas for $q_r(z) = \int_0^1 (1-h)^r (1+h^2-2hz)^{-1/2} dh$, $r = 0, 1, 2, \ldots$.

(a) Define $g(u; a) = \log\big(u + \sqrt{u^2+a}\big)$. Verify that $dg/du = (u^2+a)^{-1/2}$. Hence, $q_0(z) = g\big(1-z; 1-z^2\big) - g\big(-z; 1-z^2\big) = \log\big(1 + 1/\sqrt{w}\big)$, where $w = (1-z)/2$.

(b) Verify that $q_1(z) = 2wq_0(z) - \left(2\sqrt{w} - 1\right)$.

(c) Write $q_r(z) = \int_{-z}^{1-z}(1-z-u)^r(u^2+1-z^2)^{-1/2}du$, where $u = h - z$. Expanding $(1-z-u)^r$, one has

$$q_r(z) = \sum_{i=0}^{r}\binom{r}{i}(1-z)^{r-i}(-1)^i\int_{-z}^{1-z}\frac{u^i}{\sqrt{u^2+1-z^2}}du.$$

Integrating by parts, one has for $i > 1$,

$$\begin{aligned}\int_{-z}^{1-z}\frac{u^i du}{\sqrt{u^2+1-z^2}} &= u^{i-1}\sqrt{u^2+1-z^2}\Big|_{-z}^{1-z} - \int_{-z}^{1-z}\frac{(i-1)u^{i-2}du}{\sqrt{u^2+1-z^2}}\\ &= 2^i w^{i-1/2} - (-z)^{i-1} - \int_{-z}^{1-z}\frac{(i-1)u^{i-2}du}{\sqrt{u^2+1-z^2}};\end{aligned}$$

for i even, the integral recursively reduces to $q_0(z)$, and for i odd, it reduces to $q_1(z) - 2wq_0(z) = 2\sqrt{w} - 1$.

Section 4.5

4.21 Write $R_x(y) = R_2(x,y)$, where R_2 is given in (4.53). Prove that for $f(x) = \sum_{\mu=2}^{\infty}(a_\mu\cos 2\pi\mu x + b_\mu\sin\pi\mu x)$,

$$\int_0^1 (L_2 f)(y)(L_2 R_x)(y)dy = f(x),$$

where L_2 is given in (4.51).

4.22 Write $R_x(y) = R_{2r}(x,y)$, where R_{2r} is given in (4.56). Prove that for $f(x) = \sum_{\mu=r+1}^{\infty}(a_\mu\cos 2\pi\mu x + b_\mu\sin\pi\mu x)$,

$$\int_0^1 (L_{2r} f)(y)(L_{2r} R_x)(y)dy = f(x),$$

where L_{2r} is given in (4.55).

4.23 Write $R_x(y) = R_3(x,y)$, where R_3 is given in (4.58). Prove that for $f(x) = \sum_{\mu=2}^{\infty}(a_\mu\cos 2\pi\mu x + b_\mu\sin\pi\mu x)$,

$$\int_0^1 (L_3 f)(y)(L_3 R_x)(y)dy = f(x),$$

where L_3 is given in (4.57).

4.24 Write $R_x(y) = R_{2r+1}(x,y)$, where R_{2r+1} is given in (4.61). Prove that for $f(x) = \sum_{\mu=r+1}^{\infty}(a_\mu \cos 2\pi\mu x + b_\mu \sin \pi\mu x)$,

$$\int_0^1 (L_{2r+1}f)(y)(L_{2r+1}R_x)(y)dy = f(x),$$

where L_{2r+1} is given in (4.60).

4.25 Verify (4.66).

4.26 Verify (4.68).

4.27 Consider $R_x(y) = \int_0^1 G(x;u)G(y;u)\big(h(u)\big)^{-1}du$, with $G(x;u)$ given in (4.67). For L_ν as defined in §4.5.2, verify that $(L_\nu R_x)(0) = 0$, $\nu = 0,\dots,m-1$, and that $(L_m R_x)(y) = G(x;y)/h(y)$.

4.28 In the setting of §4.5.2, set $w_i(x) = 1$, $i = 1,\dots,m$. Verify that $\phi_\nu(x) = x^{\nu-1}/(\nu-1)!$ in (4.65), $\nu = 1,\dots,m$, and that for $u \le x$, $G(x;u) = (x-u)_+^{m-1}/(m-1)!$ in (4.67).

4.29 With $g(\tilde{x}) = f\big(\phi_2^{-1}(\tilde{x})\big)$, where ϕ_2^{-1} is the inverse of $\phi_2 = (e^{\theta x}-1)/\theta$, prove that (4.70) reduces to (4.71).

4.30 In the setting of §4.5.2, set $w_1 = 1$ and $w_i = e^{\theta x}$, $i = 2,\dots,m$.

(a) Show that $\phi_\nu(x) = \phi_2^{\nu-1}(x)/(\nu-1)!$ in (4.65), $\nu = 1,\dots,m$, where $\phi_2(x) = (e^{\theta x}-1)/\theta$.

(b) Show that for $u \le x$, $G(x;u) = \big(\phi_2(x)-\phi_2(u)\big)_+^{m-1}/(m-1)!$ in (4.67).

(c) Given $d\tilde{x}/dx = e^{\theta x}$, show that

$$D_{\tilde{x}}^\nu f = e^{-\nu\theta x}\big(D_x - (\nu-1)\theta\big)\cdots D_x f = \big(L_{\nu(x)}f\big)(dx/d\tilde{x}),$$

$\nu = 1,\dots,m$, where $D_{\tilde{x}}f = df/d\tilde{x}$, $D_x f = df/dx$, and $L_{\nu(x)}$ is the operator L_ν applied to the variable x.

4.31 In the setting of §4.5.2, set $m = 4$, $w_1 = 1$, $w_2 = w_4 = e^{\theta x}$, and $w_3 = e^{-2\theta x}$.

(a) Show that $\phi_1 = 1$, $\phi_2 = (e^{\theta x}-1)/\theta$, $\phi_3 = (\cosh\theta x - 1)/\theta^2$, and $\phi_4 = (\sinh\theta x - \theta x)/\theta^3$ in (4.65).

(b) Show that for $u \le x$, $G(x,u) = \big(\sinh\theta(x-u) - \theta(x-u)\big)/\theta^3$ in (4.67).

4.32 Prove Eqs. (4.73) and (4.74) by a change of variable, $\tilde{x} = \phi_2(x) = (e^{\theta x} - 1)/\theta$.

4.33 Consider the setting of §4.5.3. For $W(\boldsymbol{\phi})(0)$ invertible, show that $\sum_{\nu=1}^{m} f^{(\nu-1)}(0) g^{(\nu-1)}(0)$ forms an inner product in $\text{span}\{\phi_\nu, \nu = 1, \ldots, m\}$.

4.34 Verify (4.78).

5

Regression with Responses from Exponential Families

For responses from exponential family distributions, (1.4) of Example 1.1 defines penalized likelihood regression. Among topics of primary interest are the selection of smoothing parameters, the computation of the estimates, the asymptotic behavior of the estimates, and various data analytical tools.

With a nonquadratic log likelihood, iterations are needed to calculate penalized likelihood regression fit even for fixed smoothing parameters. Elementary properties concerning the penalized likelihood functional are given in §5.1, followed by discussions in §5.2 of two approaches to smoothing parameter selection. One of the approaches makes use of the scores $U_w(\lambda)$, $V_w(\lambda)$, and $M_w(\lambda)$ of §3.2.4 and the algorithms of §§3.4 or 3.5.3 via iterated reweighted (penalized) least squares, whereas the other implements a version of direct cross-validation. Approximate Bayesian confidence intervals can be calculated through the penalized weighted least squares that approximates the penalized likelihood functional at the converged fit (§5.3.1), and the "testing" of the practical significance of model terms can be performed via Kullback-Leibler projection (§5.3.2). The customizations of the general methods in specific distribution families are detailed in §5.4, along with the exploration of the empirical performances of methods and the illustration of software tools. Real-data examples are given in §5.5, where it is also shown how the techniques of this chapter can be used to estimate the spectral density of a stationary time series or to estimate a disease map.

The asymptotic convergence of penalized likelihood regression estimates will be discussed in Chap. 9.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_5,

5.1 Preliminaries

Consider exponential family distributions with densities of the form

$$f(y|x) = \exp\big\{\big(y\vartheta(x) - b(\vartheta(x))\big)/a(\phi) + c(y,\phi)\big\},$$

where $a > 0$, b, and c are known functions, $\vartheta(x)$ is the canonical parameter dependent on a covariate x, and ϕ is either known or considered as a nuisance parameter that is independent of x. Observing $Y_i|x_i \sim f(y|x_i)$, $i = 1\ldots,n$, one is to estimate the regression function $\vartheta(x) = \vartheta\big(\eta(x)\big)$ via a link η. Much of the general developments in this chapter are presented under the canonical link $\eta = \vartheta$, which covers the cases of logistic regression for binary data and Poisson regression for count data. Ramifications of the use of non-canonical links in other families will be noted in §5.4.

Parallel to (3.1) on page 62, one has the penalized likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta(x_i) - b\big(\eta(x_i)\big)\big\} + \frac{\lambda}{2}J(\eta) \tag{5.1}$$

for $\eta \in \mathcal{H} = \oplus_{\beta=0}^{p}\mathcal{H}_\beta$, where $J(f) = J(f,f) = \sum_{\beta=1}^{p}\theta_\beta^{-1}(f,f)_\beta$ and $(f,g)_\beta$ are inner products in $\mathcal{H}_\beta$ with reproducing kernels $R_\beta(x,y)$. The terms $c(Y_i,\phi)$ are independent of $\eta(x)$ and, hence, are dropped from (5.1), and the dispersion parameter $a(\phi)$ is absorbed into λ. The bilinear form $J(f,g)$ is an inner product in $\oplus_{\beta=1}^{p}\mathcal{H}_\beta$ with a reproducing kernel $R_J(x,y) = \sum_{\beta=1}^{p}\theta_\beta R_\beta(x,y)$ and a null space $\mathcal{N}_J = \mathcal{H}_0$. The first term of (5.1) depends on η only through the evaluations $[x_i]\eta = \eta(x_i)$, so the argument of §2.3.2 applies and the minimizer η_λ of (5.1) has an expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu\phi_\nu(x) + \sum_{i=1}^{n} c_i R_J(x_i,x) = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{5.2}$$

where $\{\phi_\nu\}_{\nu=1}^{m}$ is a basis of $\mathcal{N}_J = \mathcal{H}_0$, $\boldsymbol{\xi}$ and $\boldsymbol{\phi}$ are vectors of functions, and $\mathbf{c}$ and $\mathbf{d}$ are vectors of coefficients. The efficient approximation of §3.5 can also be used here, and for general purposes we shall replace $\sum_{i=1}^{n} c_i R_J(x_i,x)$ in (5.2) by $\sum_{j=1}^{q} c_j R_J(z_j,x)$; the former is a special case with $\{z_j\} = \{x_i\}$.

Example 5.1 (Gaussian regression) Consider Gaussian responses with $Y|x \sim N\big(\eta(x),\sigma^2\big)$. One has $a(\phi) = \sigma^2$ and $b(\eta) = \eta^2/2$. This reduces to the penalized least squares problem treated in Chap. 3. □

Example 5.2 (Logistic regression) Consider binary responses with $P(Y=1|x) = p(x)$ and $P(Y=0|x) = 1-p(x)$. The density is

$$f(y|x) = p(x)^y\big(1-p(x)\big)^{1-y} = \exp\big\{y\eta(x) - \log(1+e^{\eta(x)})\big\},$$

where $\eta(x) = \log\{p(x)/(1-p(x))\}$ is the logit function. One has $a(\phi) = 1$ and $b(\eta) = \log(1+e^{\eta})$. This is a special case of penalized likelihood logistic regression with binomial data. □

Example 5.3 (Poisson regression) Consider Poisson responses with $P(Y = y|x) = \{\lambda(x)\}^y e^{-\lambda(x)}/y!$, $y = 0, 1, \ldots$. The density can be written as

$$f(y|x) = \big(\lambda(x)\big)^y e^{-\lambda(x)}/y! = \exp\big\{y\eta(x) - e^{\eta(x)} - \log(y!)\big\},$$

where $\eta(x) = \log\lambda(x)$ is the log intensity. One has $a(\phi) = 1$ and $b(\eta) = e^{\eta}$. This defines penalized likelihood Poisson regression for count data. □

By standard exponential family theory, $E\big[Y\big|x\big] = \dot{b}\big(\eta(x)\big) = \mu(x)$ and $\mathrm{Var}\big[Y\big|x\big] = \ddot{b}\big(\eta(x)\big)a(\phi) = v(x)a(\phi)$; see, e.g., McCullagh and Nelder (1989, §2.2.2). The functional $L(f) = -\sum_{i=1}^n \big\{Y_i f(x_i) - b\big(f(x_i)\big)\big\}$ is thus continuous and convex in $f \in \mathcal{H}$. When the matrix S as given in (3.3) on page 62 is of full column rank, one can show that $L(f)$ is strictly convex in $\mathcal{N}_J$, and that (5.1) is strictly convex in $\mathcal{H}$; see Problem 5.1. By Theorem 2.9, the minimizer η_λ of (5.1) uniquely exists when S is of full column rank, which we will assume throughout this chapter.

Fixing the smoothing parameters λ and θ_β hidden in $J(\eta)$, (5.1) is strictly convex in η, of which the minimizer η_λ may be computed via Newton iteration. Write $\tilde{u}_i = -Y_i + \dot{b}\big(\tilde{\eta}(x_i)\big) = -Y_i + \tilde{\mu}(x_i)$ and $\tilde{w}_i = \ddot{b}\big(\tilde{\eta}(x_i)\big) = \tilde{v}(x_i)$. The quadratic approximation of $-Y_i\eta(x_i) + b\big(\eta(x_i)\big)$ at $\tilde{\eta}(x_i)$ is

$$\begin{aligned} &-Y_i\tilde{\eta}(x_i) + b\big(\tilde{\eta}(x_i)\big) + \tilde{u}_i\big\{\eta(x_i) - \tilde{\eta}(x_i)\big\} + \frac{1}{2}\tilde{w}_i\big\{\eta(x_i) - \tilde{\eta}(x_i)\big\}^2 \\ &\qquad = \frac{1}{2}\tilde{w}_i\left\{\eta(x_i) - \tilde{\eta}(x_i) + \frac{\tilde{u}_i}{\tilde{w}_i}\right\}^2 + C_i, \end{aligned}$$

where C_i is independent of $\eta(x_i)$. The Newton iteration updates $\tilde{\eta}$ by the minimizer of the penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^n \tilde{w}_i\big(\tilde{Y}_i - \eta(x_i)\big)^2 + \lambda J(\eta), \tag{5.3}$$

where $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$. Compare (5.3) with (3.9) on page 64.

5.2 Smoothing Parameter Selection

Smoothing parameter selection remains the most important practical issue for penalized likelihood regression. With (5.1) nonquadratic, one needs iterations to compute η_λ even for fixed smoothing parameters, which adds

to the complexity of the problem. Our task here is to devise efficient and effective algorithms to locate good estimates from among the η_λ's with varying smoothing parameters.

The first approach under discussion makes use of the scores $U_w(\lambda)$, $V_w(\lambda)$, and $M_w(\lambda)$ of §3.2.4 through (5.3) in a so-called performance-oriented iteration. The method tracks an appropriate loss in an indirect manner and, hence, may not be the most effective, but the simultaneous updating of (λ, θ_β) and η_λ makes it numerically efficient. Alternatively, one may employ the generalized approximate cross-validation of Xiang and Wahba (1996) or its variants, which could improve performance but at the cost of numerical efficiency. The empirical performances of the methods will be explored in §5.4 for commonly used distributions, case by case, along with possible customizations.

As in §3.2, we only make the dependence of various entities on the smoothing parameter λ explicit and suppress their dependence on θ_β in the notation.

5.2.1 Performance-Oriented Iteration

Within an exponential family, the discrepancy between distributions parameterized by (η, ϕ) and (η_λ, ϕ) can be measured by the Kullback-Leibler distance

$$\begin{aligned} \mathrm{KL}(\eta, \eta_\lambda) &= E_\eta\big[Y(\eta - \eta_\lambda) - \big(b(\eta) - b(\eta_\lambda)\big)\big]/a(\phi) \\ &= \big\{\dot{b}(\eta)(\eta - \eta_\lambda) - \big(b(\eta) - b(\eta_\lambda)\big)\big\}/a(\phi), \end{aligned}$$

or its symmetrized version

$$\begin{aligned} \mathrm{SKL}(\eta, \eta_\lambda) &= \mathrm{KL}(\eta, \eta_\lambda) + \mathrm{KL}(\eta_\lambda, \eta) \\ &= \big(\dot{b}(\eta) - \dot{b}(\eta_\lambda)\big)(\eta - \eta_\lambda)/a(\phi) \\ &= (\mu - \mu_\lambda)(\eta - \eta_\lambda)/a(\phi), \end{aligned}$$

where $\mu = \dot{b}(\eta)$. To measure the performance of $\eta_\lambda(x)$ as an estimate of $\eta(x)$, a natural loss function is given by

$$L(\eta, \eta_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \big(\mu(x_i) - \mu_\lambda(x_i)\big)\big(\eta(x_i) - \eta_\lambda(x_i)\big), \tag{5.4}$$

which is proportional to the average symmetrized Kullback-Leibler distance over the sampling points; (5.4) reduces to (3.13) on page 65 for Gaussian data. The smoothing parameters that minimize $L(\eta, \eta_\lambda)$ represent the ideal

choices, given the data, and will be referred to as the optimal smoothing parameters. By the mean value theorem, one has

$$L(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n} w'(x_i)\big(\eta(x_i) - \eta_\lambda(x_i)\big)^2, \tag{5.5}$$

where $w'(x_i) = \ddot{b}\big(\eta'(x_i)\big)$ for $\eta'(x_i)$ a convex combination of $\eta(x_i)$ and $\eta_\lambda(x_i)$.

The performance-oriented iteration to be described below operates on (5.3), which has the same numerical structure as (3.9). In fact, (5.3) also has a stochastic structure similar to that of (3.9), as the following lemma asserts.

Lemma 5.1 *Suppose* $\ddot{b}\big(\eta(x_i)\big)$ *are bounded away from* 0 *and* $\ddot{b}\big(\eta'(x_i)\big) = \ddot{b}\big(\eta(x_i)\big)\big(1 + o(1)\big)$ *uniformly for* η' *any convex combination of* η *and* $\tilde{\eta}$. *One has*

$$\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i = \eta(x_i) - u_i^o/w_i^o + o_p(1),$$

where $u_i^o = -Y_i + \dot{b}\big(\eta(x_i)\big)$ *and* $w_i^o = \ddot{b}\big(\eta(x_i)\big)$.

Proof: We drop the subscripts and write $\tilde{\eta} = \tilde{\eta}(x)$ and $\eta = \eta(x)$. Write

$$\begin{aligned}\delta &= (\tilde{\eta} - \tilde{u}/\tilde{w}) - (\eta - u^o/w^o)\\ &= (\tilde{\eta} - \eta) - \big(\dot{b}(\tilde{\eta})/\ddot{b}(\tilde{\eta}) - \dot{b}(\eta)/\ddot{b}(\eta)\big) + Y\big(1/\ddot{b}(\tilde{\eta}) - 1/\ddot{b}(\eta)\big).\end{aligned}$$

It is easy to verify that

$$\begin{aligned}E[\delta] &= (\tilde{\eta} - \eta) - \big(\dot{b}(\tilde{\eta}) - \dot{b}(\eta)\big)/\ddot{b}(\tilde{\eta})\\ &= (\tilde{\eta} - \eta) - (\tilde{\eta} - \eta)\big(1 + o(1)\big) = o(\tilde{\eta} - \eta)\end{aligned}$$

and that

$$\mathrm{Var}[\delta] = \big\{\ddot{b}(\eta)a(\phi)/\ddot{b}^2(\eta)\big\}o(1) = o\big(a(\phi)/\ddot{b}(\eta)\big).$$

The lemma follows. □

Note that $E[u_i^o/w_i^o] = 0$ and $\mathrm{Var}[u_i^o/w_i^o] = a(\phi)/w_i^o$, so (5.3) is almost the same as (3.9), except that u_i^o/w_i^o is not normal and that the weights $\tilde{w}_i$ are not the same as w_i^o. Normality is not needed for Theorem 3.5 of §3.2.4 to hold, but one does need to take care of the "misspecified" weights in (5.3).

Theorem 5.2 *Consider the setting of Theorem 3.5. Suppose* $\sqrt{w_i}\epsilon_i$ *are independent with mean zero, variances* $v_i\sigma^2$, *and uniformly bounded fourth moments. Denote* $R_w(\lambda) = EL_w(\lambda)$ *and* $V = diag(v_i)$. *As* $n \to \infty$ *and* $\lambda \to 0$, *if* $nR_w(\lambda) \to \infty$, $\big\{n^{-1}trA_w(\lambda)\big\}^2/n^{-1}trA_w^2(\lambda) \to 0$, *and* $trA_w(\lambda)/\mathrm{tr}\big(VA_w(\lambda)\big) \to 1$, *then*

$$U_w(\lambda) - L_w(\lambda) - n^{-1}\boldsymbol{\epsilon}^T W\boldsymbol{\epsilon} = o_p\big(L_w(\lambda)\big),$$
$$V_w(\lambda) - L_w(\lambda) - n^{-1}\boldsymbol{\epsilon}^T W\boldsymbol{\epsilon} = o_p\big(L_w(\lambda)\big).$$

The proof of Theorem 5.2 follows straightforward modifications of the proofs of Theorems 3.1 and 3.3, and is left as an exercise (Problem 5.2).

Theorem 5.2 applies to (5.3) with $w_i = \tilde{w}_i$, $v_i = \tilde{w}_i/w_i^o$, and $\sigma^2 = a(\phi)$. Note that the condition $v_i = 1 + o(1)$ for Lemma 5.1 implies the condition $\mathrm{tr}A_w(\lambda)/\mathrm{tr}\big(VA_w(\lambda)\big) = 1 + o(1)$ for Theorem 5.2.

Denote by $\eta_{\lambda,\tilde{\eta}}$ the minimizer of (5.3) with varying smoothing parameters. By Theorem 5.2, the minimizer of $U_w(\lambda)$ or $V_w(\lambda)$ approximately minimizes $L_w(\lambda) = n^{-1}\sum_{i=1}^n \tilde{w}_i\big(\eta_{\lambda,\tilde{\eta}}(x_i) - \eta(x_i)\big)^2$, which is a proxy of $L(\eta, \eta_{\lambda,\tilde{\eta}})$; compare with (5.5). The set $\{\eta_{\lambda,\tilde{\eta}}\}$ may not necessarily intersect with the set $\{\eta_\lambda\}$, however.

For $\tilde{\eta} = \eta_{\lambda^o}$ with fixed $(\lambda^o, \theta_\beta^o)$, it is easy to see that $\eta_{\lambda^o,\eta_{\lambda^o}} = \eta_{\lambda^o}$, which is the fixed point of Newton iteration with the smoothing parameters in (5.1) fixed at $(\lambda^o, \theta_\beta^o)$. Unless $(\lambda^o, \theta_\beta^o)$ minimizes the corresponding $U_w(\lambda)$ or $V_w(\lambda)$ (which are η_{λ^o} dependent), one would not want to use $\eta_{\lambda^o,\eta_{\lambda^o}}$, because it is perceived to be inferior to the $\eta_{\lambda,\eta_{\lambda^o}}$ that minimizes the corresponding $U_w(\lambda)$ or $V_w(\lambda)$. Note that two sets of smoothing parameters come into play here: One set specifies $\tilde{\eta} = \eta_{\lambda^o}$, which, in turn, defines the scores $U_w(\lambda)$ and $V_w(\lambda)$, and the other set indexes $\eta_{\lambda,\tilde{\eta}}$ and is the argument in $U_w(\lambda)$ and $V_w(\lambda)$. The above discussion suggests that one should look for some $\eta_{\lambda^*,\eta_{\lambda^*}} = \eta_{\lambda^*}$ that minimizes the $U_w(\lambda)$ or $V_w(\lambda)$ scores defined by itself, provided such a "self-voting" η_{λ^*} exists. To locate such "self-voting" η_{λ^*}, a performance-oriented iteration procedure was proposed by Gu (1992a), which we discuss next.

In performance-oriented iteration, one iterates on (5.3) with the smoothing parameters updated according to $U_w(\lambda)$ or $V_w(\lambda)$. Instead of moving to a particular Newton update with fixed smoothing parameters, one chooses, from among a family of Newton updates, one that is perceived to be better performing according to $U_w(\lambda)$ or $V_w(\lambda)$. If the smoothing parameters stabilize at, say, $(\lambda^*, \theta_\beta^*)$ and the corresponding Newton iteration converges at η^*, then it is clear that $\eta^* = \eta_{\lambda^*}$ and one has found the solution. Note that the procedure never compares η_λ directly with each other but only tracks $L(\eta, \eta_{\lambda,\tilde{\eta}})$ through $U_w(\lambda)$ or $V_w(\lambda)$ in each iteration. In a neighborhood around η^*, where the corresponding (5.3) is a good approximation of (5.1) for smoothing parameters near $(\lambda^*, \theta_\beta^*)$, η_{λ,η^*}'s are hopefully close approximations of η_λ's, and through indirect comparison, η^*, in turn, is perceived to be better performing among the η_λ's in the neighborhood.

The existence of "self-voting" η_{λ^*} and the convergence of performance-oriented iteration remain open and do not appear to be tractable theoretically. Note that the numerical problem (5.3) as well as the scores

$U_w(\lambda)$ and $V_w(\lambda)$ change from iteration to iteration. With proper implementation, performance-oriented iteration is found to converge empirically in most situations, and when it converges, the fixed point of the iteration simply gives the desired "self-voting" η_{λ^*}.

The implementation suggested in Gu (1992a) starts at some $\tilde{\eta} = \eta_\lambda$ with λ large, and it limits the search range for smoothing parameters to a neighborhood of the previous ones during the minimization of $U_w(\lambda)$ or $V_w(\lambda)$ in each iteration. The idea is to start from the numerically more stable end of the trajectory $\{\eta_\lambda\}$ and to stay close to the trajectory, where the final solution will be located. Technical details are to be found in Gu (1992a).

Since $M(\lambda)$ also does a good job in tracking the mean square error loss in penalized least squares regression, as illustrated in simulations (see, e.g., §3.2.5), one may also use $M_w(\lambda)$ to drive the performance-oriented iteration by analogy. Such a procedure does not maximize any likelihood function with respect to the smoothing parameters, however.

To explore the mechanism that drives the performance-oriented iteration to convergence, a sample of binary data were generated on $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$ using a logit function

$$\eta(x) = 3\{10^5 x^{11}(1-x)^6 + 10^3 x^3 (1-x)^{10}\} - 2. \tag{5.6}$$

Set $\tilde{\eta} = \eta_{\tilde{\lambda}}$ in (5.3) for $\tilde{\lambda}$ on a grid $\log_{10} \tilde{\lambda} = -6(0.1)0$. The scores $U_w(\lambda)$ (with $a(\phi) = 1$), $V_w(\lambda)$, and $M_w(\lambda)$ were evaluated for λ on a grid $\log_{10} \lambda = -6(0.1)0$. Note that $\tilde{\lambda}$ here indexes $\tilde{\eta} = \eta_{\tilde{\lambda}}$ the minimizer of (5.1) and λ indexes $\eta_{\lambda,\tilde{\eta}}$ the minimizer of (5.3) given $\tilde{\eta}$. This gave 61×61 arrays of $U_w(\lambda)$, $V_w(\lambda)$, and $M_w(\lambda)$. These arrays are contoured in Fig. 5.1, where the horizontal axis is λ and the vertical axis is $\tilde{\lambda}$. An $\eta_{\tilde{\lambda}}$ that is not optimal can still be a good approximation of η for the purpose of Lemma 5.1, so for many of the horizontal slices in Fig. 5.1, one could expect the minima, marked as a circle or a star in the plots, to provide λ close to optimal for the weighted least squares problem (5.3). The stars in Fig. 5.1 indicate the respective "self-voting" λ^*, to which performance-oriented iteration converged. Note that although the iteration in general only visits the slice marked by the solid line on convergence, the scores associated with the intermediate iterates should have behavior similar to the horizontal slices in the plots.

5.2.2 Direct Cross-Validation

In order to compare η_λ directly, one needs some computable score that tracks $L(\eta, \eta_\lambda)$ of (5.4). One such score is the generalized approximate cross-validation (GACV) of Xiang and Wahba (1996), to be described below.

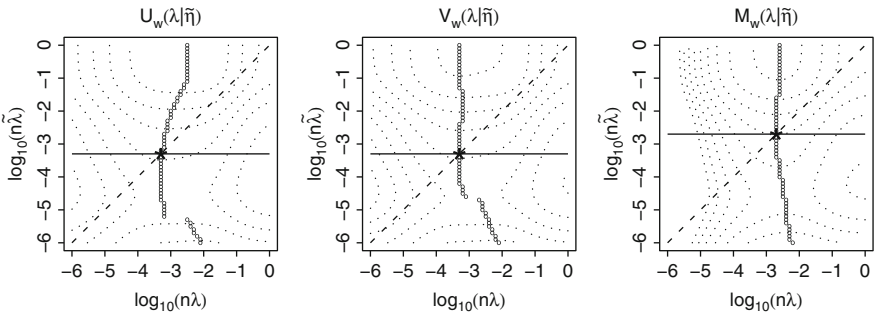


FIGURE 5.1. Contours of $U_w(\lambda|\eta_{\tilde{\lambda}})$, $V_w(\lambda|\eta_{\tilde{\lambda}})$, and $M_w(\lambda|\eta_{\tilde{\lambda}})$. The *circles* are minima of the horizontal slices with fixed $\tilde{\lambda}$. The *star* indicates the "self-voting" λ^*. Performance-oriented iteration visits the *solid slice* on convergence.

Without loss of generality, assume $a(\phi) = 1$. Consider the Kullback-Leibler distance

$$\mathrm{KL}(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{\mu(x_i)\big(\eta(x_i) - \eta_\lambda(x_i)\big) - \big(b(\eta(x_i)) - b(\eta_\lambda(x_i))\big)\big\}, \quad (5.7)$$

which is a proxy of $L(\eta, \eta_\lambda)$; roughly, $2\mathrm{KL}(\eta, \eta_\lambda) \approx L(\eta, \eta_\lambda)$. Dropping terms from (5.7) that do not involve η_λ, one gets the relative Kullback-Leibler distance

$$\mathrm{RKL}(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{ - \mu(x_i)\eta_\lambda(x_i) + b\big(\eta_\lambda(x_i)\big)\big\}. \quad (5.8)$$

Replacing $\mu(x_i)\eta_\lambda(x_i)$ by $Y_i\eta_\lambda^{[i]}(x_i)$, one obtains a cross-validation estimate of $\mathrm{RKL}(\eta, \eta_\lambda)$,

$$V_0(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{ - Y_i\eta_\lambda^{[i]}(x_i) + b\big(\eta_\lambda(x_i)\big)\big\}, \quad (5.9)$$

where $\eta_\lambda^{[k]}$ minimizes the "delete-one" version of (5.1),

$$-\frac{1}{n}\sum_{i\neq k}\big\{Y_i\eta(x_i) - b\big(\eta(x_i)\big)\big\} + \frac{\lambda}{2}J(\eta). \quad (5.10)$$

Note that $E[Y_i] = \mu(x_i)$ and that $\eta_\lambda^{[i]}$ is independent of Y_i. Write

$$V_0(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta_\lambda(x_i) - b\big(\eta_\lambda(x_i)\big)\big\} + \frac{1}{n}\sum_{i=1}^{n}Y_i\big(\eta_\lambda(x_i) - \eta_\lambda^{[i]}(x_i)\big), \quad (5.11)$$

where the first term is readily available, but the second term is impractical to compute. One needs computationally practical approximations of the second term to make use of $V_0(\lambda)$.

Through a series of first-order Taylor expansions, Xiang and Wahba (1996) propose to approximate the second term of (5.11) by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{h_{ii}Y_i\big(Y_i-\mu_\lambda(x_i)\big)}{1-h_{ii}\tilde{w}_i}, \tag{5.12}$$

where $\tilde{w}_i = \ddot{b}\big(\eta_\lambda(x_i)\big)$ and h_{ii} is the ith diagonal of a matrix H to be specified below. Recall matrices S and Q from §3.1 and let F_2 be an $n \times (n-m)$ orthogonal matrix satisfying $S^TF_2 = 0$. Write $W = \mathrm{diag}(\tilde{w}_i)$. The matrix H appearing in (5.12) is given by

$$H = \big(W + n\lambda F_2(F_2^TQF_2)^+F_2^T\big)^{-1},$$

where $(\cdot)^+$ denotes the Moore-Penrose inverse. Substituting the approximation into (5.11), one gets an approximate cross-validation (ACV) score

$$V_a(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta_\lambda(x_i) - b\big(\eta_\lambda(x_i)\big)\big\} + \frac{1}{n}\sum_{i=1}^{n}\frac{h_{ii}Y_i\big(Y_i-\mu_\lambda(x_i)\big)}{1-h_{ii}\tilde{w}_i}. \tag{5.13}$$

Replacing h_{ii} and $h_{ii}\tilde{w}_i$ in (5.13) by their respective averages $n^{-1}\mathrm{tr}H$ and $1-n^{-1}\mathrm{tr}(HW)$, one obtains the GACV score of Xiang and Wahba (1996),

$$\begin{aligned} V_g(\lambda) = &-\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta_\lambda(x_i) - b(\eta_\lambda(x_i))\big\} \\ &+ \frac{\mathrm{tr}H}{n-\mathrm{tr}(HW)}\frac{1}{n}\sum_{i=1}^{n}Y_i\big(Y_i-\mu_\lambda(x_i)\big). \end{aligned} \tag{5.14}$$

For n large, Q is often ill-conditioned and the computation of H can be numerically unstable.

As an alternative approach to the approximation of (5.11), Gu and Xiang (2001) substitute $\eta^{[i]}_{\lambda,\eta_\lambda}(x_i)$ for $\eta^{[i]}_\lambda(x_i)$, where $\eta^{[k]}_{\lambda,\eta_\lambda}$ minimizes the "delete-one" version of (5.3),

$$\frac{1}{n}\sum_{i\neq k}\tilde{w}_i\big(\tilde{Y}_i-\eta(x_i)\big)^2 + \lambda J(\eta), \tag{5.15}$$

for $\tilde{\eta} = \eta_\lambda$. Remember that $\eta_\lambda = \eta_{\lambda,\eta_\lambda}$. Trivial adaptation of Lemma 3.2 of §3.2.2 yields

$$\sqrt{\tilde{w}_i}\big(\eta_\lambda(x_i) - \eta^{[i]}_{\lambda,\eta_\lambda}(x_i)\big) = a_{i,i}\sqrt{\tilde{w}_i}\big(\tilde{Y}_i - \eta^{[i]}_{\lambda,\eta_\lambda}(x_i)\big),$$

where $a_{i,i}$ is the ith diagonal of the matrix $A_w(\lambda)$; see (3.11) and (3.12) on page 64. It follows that

$$\eta_\lambda(x_i) - \eta^{[i]}_{\lambda,\eta_\lambda}(x_i) = \frac{a_{i,i}}{1-a_{i,i}}\big(\tilde{Y}_i - \eta_\lambda(x_i)\big).$$

Recalling that $\tilde{Y}_i = \tilde{\eta}(x_i) - \tilde{u}_i/\tilde{w}_i$, one has

$$\eta_\lambda(x_i) - \eta^{[i]}_{\lambda,\eta_\lambda}(x_i) = \frac{a_{i,i}}{1-a_{i,i}} \frac{-\tilde{u}_i}{\tilde{w}_i}. \tag{5.16}$$

Substituting (5.16) into (5.11), one obtains an alternative ACV score

$$\begin{aligned} V^*_a(\lambda) = &-\frac{1}{n}\sum_{i=1}^n \big\{Y_i\eta_\lambda(x_i) - b\big(\eta_\lambda(x_i)\big)\big\} \\ &+ \frac{1}{n}\sum_{i=1}^n \frac{a_{i,i}}{1-a_{i,i}} \frac{Y_i(-\tilde{u}_i)}{\tilde{w}_i}. \end{aligned} \tag{5.17}$$

Parallel to (5.14), one may replace $a_{i,i}/\tilde{w}_i$ by $n^{-1}\sum_{i=1}^n a_{i,i}/\tilde{w}_i$ and $1-a_{i,i}$ by $1-n^{-1}\mathrm{tr}A_w$ to obtain an alternative GACV score:

$$\begin{aligned} V^*_g(\lambda) = &-\frac{1}{n}\sum_{i=1}^n \big\{Y_i\eta_\lambda(x_i) - b(\eta_\lambda(x_i))\big\} \\ &+ \frac{\mathrm{tr}(A_wW^{-1})}{n-\mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^n Y_i(-\tilde{u}_i). \end{aligned} \tag{5.18}$$

Remember that $\tilde{u}_i = -Y_i + \tilde{\mu}(x_i)$, and it can be shown (Problem 5.3) that when $F_2^TQF_2$ is nonsingular, $A_w(\lambda) = W^{1/2}HW^{1/2}$. Hence, $V_g(\lambda)$ and $V^*_g(\lambda)$ are virtually the same, and we shall remove the star in the notation from now on. The terms in (5.18) are numerically stable for all n.

For Gaussian data, $V_g(\lambda)$ of (5.18) reduces to

$$U^*(\lambda) = \frac{1}{n}\mathbf{Y}^T\big(I-A(\lambda)\big)^2\mathbf{Y} + \frac{2\mathrm{tr}A(\lambda)}{n}\frac{\mathbf{Y}^T\big(I-A(\lambda)\big)\mathbf{Y}}{\mathrm{tr}\big(I-A(\lambda)\big)}. \tag{5.19}$$

Under mild conditions, one can show that

$$U^*(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big).$$

See Problem 5.4.

With fixed smoothing parameters, the algorithms of §3.4 do not have any advantage over that of §3.5.3 even for $q = n$, so the weighted version of (3.63) will be used to calculate the minimizer of (5.3).

5.3 Inferential Tools

Based on (5.3) at the converged fit $\tilde{\eta} = \eta_\lambda$, one may calculate the posterior means and posterior variances as if it were weighted Gaussian regression, which can then be used to construct approximate Bayesian confidence intervals. For the "testing" of $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, one may calculate an estimate $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$ and compare it with its Kullback-Leibler projection in $\mathcal{H}_0$.

5.3.1 Approximate Bayesian Confidence Intervals

Consider $\eta = \eta_0 + \eta_1$, where η_0 and η_1 have independent mean zero Gaussian process priors with covariances $E\big[\eta_0(x)\eta_0(y)\big] = \tau^2 \sum_{\nu=1}^m \phi_\nu(x)\phi_\nu(y)$ and $E\big[\eta_1(x)\eta_1(y)\big] = bR_J(x,y)$. Write $\eta_0(x) = \sum_{\nu=1}^m \phi_\nu(x)\beta_\nu$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T \sim N(0, \tau^2 I)$. Write $\boldsymbol{\eta} = \big(\eta(x_1), \ldots, \eta(x_n)\big)^T$ and let $\tau^2 \to \infty$; the likelihood of $(\boldsymbol{\eta}, \boldsymbol{\beta})$ is proportional to

$$\exp\Big\{ -\frac{1}{2b}(\boldsymbol{\eta} - S\boldsymbol{\beta})^T Q^+ (\boldsymbol{\eta} - S\boldsymbol{\beta})\Big\}, \tag{5.20}$$

where S is $n \times m$ with the (i,ν)th entry $\phi_\nu(x_i)$ and Q^+ is the Moore-Penrose inverse of the $n \times n$ matrix Q with the (i,j)th entry $R_J(x_i, x_j)$; see Problem 5.5. Integrating out $\boldsymbol{\beta}$ from (5.20), the likelihood of $\boldsymbol{\eta}$ is seen to be

$$q(\boldsymbol{\eta}) \propto \exp\Big\{ -\frac{1}{2b}\boldsymbol{\eta}^T (Q^+ - Q^+ S(S^T Q^+ S)^{-1} S^T Q^+)\boldsymbol{\eta}\Big\}; \tag{5.21}$$

see Problem 5.6. The posterior likelihood of $\boldsymbol{\eta}$ given $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is proportional to the joint likelihood, which is of the form

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\eta})q(\boldsymbol{\eta}) \propto \exp\Big\{ &\frac{1}{a(\phi)} \sum_{i=1}^n \big(Y_i \eta(x_i) - b(\eta(x_i))\big) \\ &- \frac{1}{2b}\boldsymbol{\eta}^T (Q^+ - Q^+ S(S^T Q^+ S)^{-1} S^T Q^+)\boldsymbol{\eta}\Big\}. \end{aligned} \tag{5.22}$$

The following theorem extends the results of §2.5.

Theorem 5.3 *Suppose η_λ minimizes (5.1) with $n\lambda = a(\phi)/b$. For Q nonsingular, the fitted values $\boldsymbol{\eta}^* = \big(\eta_\lambda(x_1), \ldots, \eta_\lambda(x_n)\big)^T$ are the posterior mode of $\boldsymbol{\eta}$ given $\mathbf{Y}$.*

Proof: By (5.2), $\boldsymbol{\eta}^* = Q\mathbf{c} + S\mathbf{d}$, where $\mathbf{c} = (c_1, \ldots, c_n)^T$, $\mathbf{d} = (d_1, \ldots, d_m)^T$ minimize

$$-\frac{1}{n}\sum_{i=1}^n \big\{Y_i(\boldsymbol{\xi}_i^T \mathbf{c} + \boldsymbol{\phi}_i^T \mathbf{d}) - b(\boldsymbol{\xi}_i^T \mathbf{c} + \boldsymbol{\phi}_i^T \mathbf{d})\big\} + \frac{\lambda}{2}\mathbf{c}^T Q \mathbf{c}, \tag{5.23}$$

with $\boldsymbol{\xi}_i = \big(R_J(x_1, x_i), \ldots, R_J(x_n, x_i)\big)^T$ and $\boldsymbol{\phi}_i = \big(\phi_1(x_i), \ldots, \phi_m(x_i)\big)^T$. Taking derivatives of (5.23) with respect to $\mathbf{c}$ and $\mathbf{d}$ and setting them to zero, one has

$$\begin{aligned} Q\mathbf{u} + n\lambda Q\mathbf{c} &= 0, \\ S^T \mathbf{u} &= 0, \end{aligned} \tag{5.24}$$

where $\mathbf{u} = (u_1, \ldots, u_n)^T$ with $u_i = -Y_i + \dot{b}\big(\eta_\lambda(x_i)\big)$. For Q nonsingular, $Q^+ = Q^{-1}$. Taking derivatives of $-a(\phi)\log p(\mathbf{Y}|\boldsymbol{\eta})q(\boldsymbol{\eta})$ as given in (5.22)

with respect to $\boldsymbol{\eta}$, and plugging in $\boldsymbol{\eta}^* = Q\mathbf{c} + S\mathbf{d}$ with $\mathbf{c}$ and $\mathbf{d}$ satisfying (5.24), one has

$$\mathbf{u} + n\lambda(Q^{-1} - Q^{-1}S(S^TQ^{-1}S)^{-1}S^TQ^{-1})(Q\mathbf{c} + S\mathbf{d}) \\ = \mathbf{u} + n\lambda(\mathbf{c} - Q^{-1}S(S^TQ^{-1}S)^{-1}S^T\mathbf{c}) = 0.$$

The theorem follows. □

Replacing the exponent of $p(\mathbf{Y}|\boldsymbol{\eta})$ by its quadratic approximation at $\boldsymbol{\eta}^*$, one gets a Gaussian likelihood with observations $\tilde{Y}_i$ and variances $a(\phi)/\tilde{w}_i$, where $\tilde{Y}_i$ and $\tilde{w}_i$ are as specified in (5.3), all evaluated at $\tilde{\eta} = \eta_\lambda$. With such a Gaussian approximation of the sampling likelihood $p(\mathbf{Y}|\boldsymbol{\eta})$, the results of §3.3 yield approximate posterior means and variances for $\eta(x)$ and its components, which can be used to construct approximate Bayesian confidence intervals.

On the sampling points, for Q nonsingular, such an approximate posterior analysis of $\boldsymbol{\eta}$ is simply Laplace's method applied to the posterior distribution of $\boldsymbol{\eta}$, as ascertained by Theorem 5.3; see, e.g., Tierney and Kadane (1986) and Leonard et al. (1989) for discussions on Laplace's method. The statement, however, is generally not true even for a subset of $\boldsymbol{\eta}$, as the corresponding subset of $\boldsymbol{\eta}^*$ are, in general, not the exact mode of the respective likelihood. It appears that the exact Bayesian calculation can be sensitive to parameter specification. This also serves to explain why one would need Q to be nonsingular for Theorem 5.3 to hold.

With the Bayes model of §3.5.2 for efficient approximation, (5.20)–(5.22) hold after replacing Q^+ by RQ^+R^T, with R $n \times q$ having the (i,j)th entry $R_J(x_i, z_j)$ and Q $q \times q$ having the (j,k)th entry $R_J(z_j, z_k)$. Theorem 5.3 does not seem to hold in the setting, but approximate Bayesian confidence intervals can still be calculated based on the quadratic approximation of $p(\mathbf{Y}|\boldsymbol{\eta})$ at $\boldsymbol{\eta}^*$.

5.3.2 Kullback-Leibler Projection

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its Kullback-Leibler projection $\tilde{\eta} \in \mathcal{H}_0$ minimizes, over $\eta \in \mathcal{H}_0$, the Kullback-Leibler distance,

$$\mathrm{KL}(\hat{\eta}, \eta) = \frac{1}{n}\sum_{i=1}^{n}\big\{\hat{\mu}_i\big(\hat{\vartheta}_i - \vartheta(\eta(x_i))\big) - \big(b(\hat{\vartheta}_i) - b(\vartheta(\eta(x_i)))\big)\big\}, \qquad (5.25)$$

with $\hat{\mu}_i = \hat{\mu}(x_i)$ and $\hat{\vartheta}_i = \vartheta(\hat{\eta}(x_i))$. $\mathrm{KL}(\hat{\eta}, \eta)$ in (5.25) agrees with (5.7) for $\eta = \vartheta$ and is equivalent to (3.82) for Gaussian data with $\eta = \vartheta$, $b(\eta) = \eta^2/2$; the square error projection of §3.8 is thus a special case.

For $\eta_c \in \mathcal{H}_0$ a constant fit, one has (Problem 5.7)

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{\mu}_i - \hat{\mu}_i)\tilde{h}(x_i)\big(\tilde{\eta}(x_i) - \eta_c(x_i)\big) = 0, \qquad (5.26)$$

where $\tilde{\mu}_i = \tilde{\mu}(x_i)$ and $\tilde{h} = (d\vartheta/d\eta)\big|_{\tilde{\eta}}$. It is easy to verify that

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c) + \frac{1}{n}\sum_{i=1}^{n}(\tilde{\mu}_i - \hat{\mu}_i)\big(\tilde{\vartheta}(x_i) - \vartheta_c(x_i)\big),$$

where, by (5.26), the last term vanishes for $\eta = \vartheta$ the canonical link. The Kullback-Leibler decomposition $\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c)$ may still hold approximately for non-canonical links, depending on how accurate the first order approximation, $(\tilde{\mu} - \hat{\mu})(\tilde{\vartheta} - \vartheta_c) \approx (\tilde{\mu} - \hat{\mu})\tilde{h}(\tilde{\eta} - \eta_c)$, is.

The Kullbak-Leibler projection in an infinite-dimensional $\mathcal{H}_0$ is ill-posed, just like the special case of square error projection discussed in §3.8. To regulate the problem, one may use the efficient approximation of §3.5 with $q = o(n)$ and add a small but positive penalty term to (5.25); further details are as discussed in §3.8, except that one now iterates on weighted versions of (3.63).

5.4 Software, Customization, and Empirical Performance

The common structure of penalized likelihood regression warrants unified software implementation, yet distinctive characteristics of individual families require due customizations of the general methods. The empirical performances of the various methods provide insights concerning the method of choice in practice and guide the default software settings.

After a brief introduction of three suites of R functions for penalized likelihood regression, the specialization and customization of the general methods are spelled out for the binomial, Poisson, gamma, inverse Gaussian, and negative binomial families. The empirical performances of various cross-validation methods are presented for the individual families in their respective sections, along with simple software illustrations.

5.4.1 R Package `gss`: `gssanova`, `gssanova0`, and `gssanova1` Suites

Similar to the `ssanova` and `ssanova0` suites for Gaussian regression, the three suites for non-Gaussian regression largely share the same syntax but employ different numerical engines under the hood. The performance-oriented iteration of §5.2.1 is implemented in `gssanova0` and `gssanova1`, with the former using the algorithms of §3.4 to solve (5.3) with automatic smoothing parameters and the latter using the algorithms of §3.5.3; both suites allow the choices of `method="u"`, `"v"`, `"m"`, and `gssanova1` also takes `alpha`, with a default value 1.4, that modifies $U_w(\lambda)$, $V_w(\lambda)$ for `method="u"`, `"v"` by attaching a fudge factor $\alpha > 1$ in front of $\mathrm{tr}A_w(\lambda)$. The direct

cross-validation of §5.2.2 is implemented in `gssanova`. The Kullback-Leibler projection of §5.3.2 is implemented for `gssanova` and `gssanova1`, but not for `gssanova0`. The `gssanova0` suite is virtually the original `gssanova` suite referred to in the first edition of this book, delegating much of the numerical calculations to RKPACK routines.

For each of the families, only one link is used, one that is free of constraint. This is not much of a restriction, however, as splines are flexible.

5.4.2 Binomial Family

The binomial distribution Binomial(m, p) has a density

$$\binom{m}{y} p^y (1-p)^{m-y}$$

and a minus log likelihood

$$-y\eta + m\log(1+e^\eta) = l(\eta; y), \tag{5.27}$$

where the logit $\eta = \log\{p/(1-p)\}$ is the canonical parameter. The binary data of Example 5.2 is a special case with $m = 1$. To iterate on (5.3), it is easy to calculate $\tilde{u}_i = -Y_i + m_i\,\tilde{p}_i$ and $\tilde{w}_i = m_i\,\tilde{p}_i(1-\tilde{p}_i)$, where $\tilde{p}_i = \tilde{p}(x_i)$; see Problem 5.8.

Invariant Methods

The binomial responses Y_i are sums of binary responses, say $Y_i = \sum_{j=1}^{m_i} Y_{i,j}$, where $Y_{i,j} \in \{0, 1\}$. Using the same data, either in the individual form $(x_i, Y_{i,j})$ or in the grouped form (x_i, Y_i), one naturally expects the same end result. This calls for methods that are invariant to data grouping.

For the terms in (5.3), it is easy to verify that

$$\begin{aligned}\tilde{w}_i\big(\tilde{Y}_i - \eta(x_i)\big)^2 &= m_i\tilde{p}_i(1-\tilde{p}_i)\Big(\tilde{\eta}_i - \frac{m_i\tilde{p}_i - Y_i}{m_i\tilde{p}_i(1-\tilde{p}_i)} - \eta(x_i)\Big)^2 \\ &= \sum_{j=1}^{m_i} \tilde{p}_i(1-\tilde{p}_i)\Big(\tilde{\eta}_i - \frac{\tilde{p}_i - Y_{ij}}{\tilde{p}_i(1-\tilde{p}_i)} - \eta(x_i)\Big)^2 + C,\end{aligned}$$

where $\tilde{\eta}_i = \tilde{\eta}(x_i)$ and C does not involve $\eta(x_i)$. It is reassuring to see that (5.3) is invariant to data grouping.

The dispersion is known to be $a(\phi) = 1$, so intuitively, $U_w(\lambda)$ with $\sigma^2 = 1$ should be the preferred method to use in performance-oriented iteration. As seen in §3.2.4, $U(\lambda)$ for individual data $Y_{i,j}$ is equivalent to $U_w(\lambda)$ for grouped data Y_i/m_i with weights $w_i = m_i$; parallel calculations show that $U_w(\lambda)$ for individual data $Y_{i,j}$ with weights $w_{i,j} = p_i(1-p_i)$ is equivalent to $U_w(\lambda)$ for grouped data Y_i/m_i with weights $w_i = m_i p_i(1-p_i)$.

Hence, performance-oriented iteration driven by $U_w(\lambda)$ is invariant to data grouping. The same can not be said about $V_w(\lambda)$ or $M_w(\lambda)$, however.

For direct cross-validation, the verbatim application of (5.18) amounts to "delete-m" instead of "delete-one." One however could work under the equivalent binary setting, in which the matrices A_w and W are $N \times N$, where $N = \sum_{i=1}^{n} m_i$, and the entries associated with each x_i form homogeneous (by symmetry) blocks of sizes m_i; within each block the diagonals of A_w are $1/m_i$ of the binomial $a_{i,i}$, $\tilde{w}$ is $\tilde{p}_i(1-\tilde{p}_i)$, and $\tilde{u}$ is $\tilde{p}_i - Y_{ij}$. Applying (5.18) in the binary setting, simple algebra yields

$$V_g(\lambda) = -\frac{1}{N}\sum_{i=1}^{n}\left\{Y_i\eta_\lambda(x_i) - m_i\log(1+e^{\eta_\lambda(x_i)})\right\} + \alpha\frac{\mathrm{tr}(A_w M W^{-1})}{N - \mathrm{tr}A_w}\frac{1}{N}\sum_{i=1}^{n} Y_i(1-\tilde{p}_i) \tag{5.28}$$

for $\alpha = 1$, where $M = \mathrm{diag}(m_1, \ldots, m_n)$; a fudge factor $\alpha > 1$ might help if the unmodified cross-validation score delivers undersmoothing. Clearly, (5.28) is invariant to data grouping.

Empirical Performance

A simple simulation was performed to investigate the empirical performances of the methods discussed above. Binary samples were drawn on $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$ using the logit function given in (5.6) on page 181. For each replicate, five cubic spline fits were calculated with $q = n$, one minimizing the symmetrized Kullback-Leibler loss $L(\lambda) = L(\eta, \eta_\lambda)$ of (5.4), two minimizing $V_g(\lambda)$ of (5.28) with $\alpha = 1, 1.4$, and two resulting from performance-oriented iteration driven by $U_w(\lambda)$ with $\alpha = 1, 1.4$. The losses achieved by the five fits were recorded, which included the optimal $L(\lambda_o)$, two $L(\lambda_d)$'s from direct cross-validation, and two $L(\lambda_p)$'s from performance-oriented iteration.

The simulation was conducted on one hundred replicates of samples and the results are summarized in Fig. 5.2. In the left frame, the relative efficacy of the methods, $L(\lambda_o)/L(\lambda_d)$ or $L(\lambda_o)/L(\lambda_p)$, is shown in boxplots. In the center frame, methods modified by a fudge factor $\alpha = 1.4$ are compared with the respective standard ones. An $\alpha = 1.4$ in $V_g(\lambda)$ offers little benefit compared to $\alpha = 1$, warranting no further consideration. In the right frame, the performance-oriented iteration is compared against $V_g(\lambda)$. The direct cross-validation via $V_g(\lambda)$ emerges as the method of choice.

Software Illustration

The syntax of `gssanova` for the binomial family is similar to that of `glm`. The following sequence generates some synthetic data on a grid and calculates a cubic spline logistic fit:

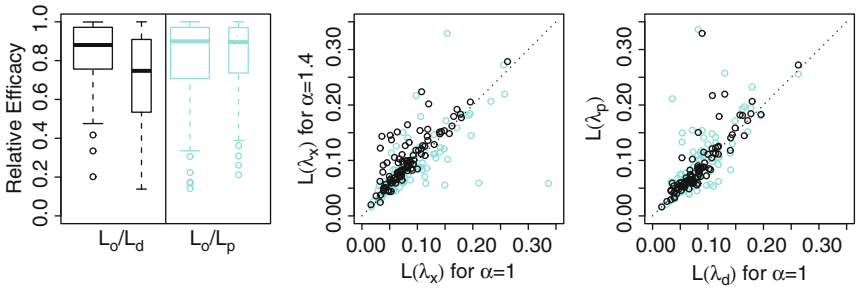


FIGURE 5.2. Effectiveness of $V_g(\lambda)$ and $U_w(\lambda)$ in logistic regression simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1$ versus $L(\lambda_p)$ with $\alpha = 1$ (*faded*) or $\alpha = 1.4$ (*solid*).

```
set.seed(5732)
test <- function(x)
        {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))-2}
x <- (0:100)/100
p <- 1-1/(1+exp(test(x)))
y <- rbinom(x,3,p)
fit.lgt <- gssanova(cbind(y,3-y)~x,family="binomial")
```

Equivalently, one may use a one-column response Y_i/m_i and enter $m_i = 3$ as weights:

```
fit.lgt <- gssanova(y/3~x,"binomial",weights=rep(3,101))
```

Due to the random selection of z_j, repeated calls to `gssanova` would return slightly different results unless `id.basis` is specified, as with `ssanova`. To evaluate the fit on the grid, use:

```
est <- predict(fit.lgt,data.frame(x=x),se=TRUE)
```

The fit is plotted in the left frame of Fig. 5.3, with the data and the test function superimposed:

```
plot(x,y/3,ylab="p",col=3); lines(x,p,lty=2)
lines(x,1-1/(1+exp(est$fit)))
lines(x,1-1/(1+exp(est$fit+1.96*est$se)),col=5)
lines(x,1-1/(1+exp(est$fit-1.96*est$se)),col=5)
```

Note that the prediction is on the logit scale. The working residuals and deviance residuals are also available:

```
resid(fit.lgt)
resid(fit.lgt,type="dev")
```

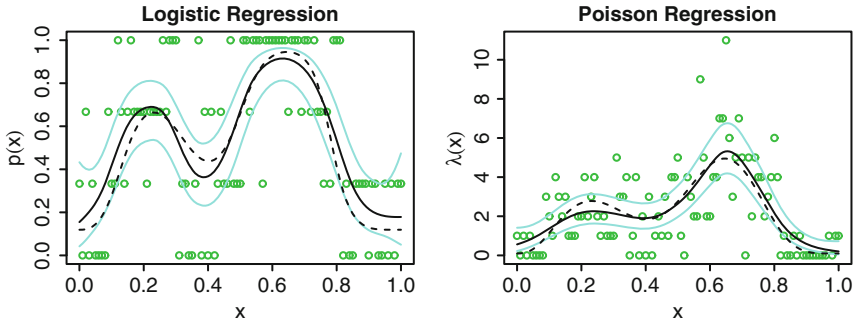


FIGURE 5.3. Cubic spline logistic and Poisson regression. The test functions are in *dashed lines*, the fits are in *solid lines*, and the 95 % Bayesian confidence intervals are in *faded lines*. The data are superimposed as *circles*.

The syntax of `gssanova0` and `gssanova1` is the same, unless one wants to override the default `method="u"` with `varht=1`, and for `gssanova1`, `alpha=1.4`.

5.4.3 Poisson Family

The Poisson distribution Poisson(λ) has a minus log likelihood

$$-y\log\lambda+\lambda=-y\eta+e^{\eta}=l(\eta;y), \tag{5.29}$$

where the log intensity $\eta=\log\lambda$ is the canonical parameter. To iterate on (5.3), one has $\tilde{u}_i=-Y_i+e^{\tilde{\eta}(x_i)}$ and $\tilde{w}_i=e^{\tilde{\eta}(x_i)}$; see Problem 5.9.

With a known dispersion $a(\phi)=1$, $U_w(\lambda)$ with $\sigma^2=1$ is still the preferred method to use in performance-oriented iteration. While there is no invariance to worry about here, the close relation between Poisson regression and density estimation suggests a direct cross-validation score that is more natural in the setting and works better than (5.18).

Poisson Regression as Density Estimation

Plugging the Poisson log likelihood (5.29) into (5.1) and adding and subtracting a term, one has

$$-\sum_{i=1}^{n}Y_i\left\{\eta(x_i)-\log\int e^{\eta}dx\right\}+\frac{n\lambda}{2}J(\eta)+\left\{\int e^{\eta}dx-N\log\int e^{\eta}dx\right\}, \tag{5.30}$$

where $\int e^{\eta}dx=\sum_{i=1}^{n}e^{\eta(x_i)}$ and $N=\sum_{i=1}^{n}Y_i$. The first two terms in (5.30) estimates the density $e^{\eta}/\int e^{\eta}dx$ on the discrete domain $\{x_1,\ldots,x_n\}$ using "binned data" with bin-size Y_i at x_i, and when $J(\eta)$ annihilates constant, the third term is separable from the other two, which fixes a constant in the log density η to make $\int e^{\eta}dx=N$; see §7.2.

The use of (5.18) in Poisson regression amounts to "delete-one-bin" in the density estimation context, and is far inferior to the "delete-one-count" cross-validation developed in §7.3; empirical comparisons can be found in Kim (2003). Applying (7.22) on page 245 in the current setting, one has

$$V(\lambda) = -\frac{1}{N}\sum_{i=1}^{n}\left\{Y_i\eta_\lambda(x_i) - e^{\eta_\lambda(x_i)}\right\} + \alpha\frac{\mathrm{tr}(P_y\tilde{R}H^{+}\tilde{R}^T P_y^T)}{N(N-1)}, \qquad (5.31)$$

where $\tilde{R} = (S, R)$, $P_y = (I - \mathbf{y}\mathbf{y}^T/N)\,\mathrm{diag}(\mathbf{y})$ for $\mathbf{y} = \left(\sqrt{Y_1}, \ldots, \sqrt{Y_n}\right)^T$, and $H = \left\{\tilde{R}^T(W - \mathbf{w}\mathbf{w}^T/N)\tilde{R} + n\lambda\begin{pmatrix} O & O \\ O & Q \end{pmatrix}\right\}/N$ for $W = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$ and $\mathbf{w} = (\tilde{w}_1, \ldots, \tilde{w}_n)^T$; S, R, and Q are as given in (3.63) on page 86, and remember that $\sum_{i=1}^n \tilde{w}_i = \sum_{i=1}^n e^{\tilde{\eta}(x_i)} = N$. Once again, a fudge factor $\alpha \geq 1$ is attached to the extra term beyond the minus log likelihood.

The score $V_g(\lambda)$ of (5.18) targets the relative Kullback-Leibler distance in the regression setting as given in (5.8),

$$\mathrm{RKL}(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{e^{\eta_\lambda(x_i)} - e^{\eta(x_i)}\eta_\lambda(x_i)\right\},$$

whereas $V(\lambda)$ of (5.31) is after the relative Kullback-Leibler distance in the density estimation setting as given in (7.14),

$$\log\int e^{\eta_\lambda}dx - \mu_\eta(\eta_\lambda) = \log\sum_{i=1}^{n} e^{\eta_\lambda(x_i)} - \frac{\sum_{i=1}^n e^{\eta(x_i)}\eta_\lambda(x_i)}{\sum_{i=1}^n e^{\eta(x_i)}}$$

Note that $\sum_{i=1}^n e^{\eta_\lambda(x_i)} = N$ for all λ and $\sum_{i=1}^n e^{\eta(x_i)}$ is independent of λ, so both are aiming to maximize $\sum_{i=1}^n e^{\eta(x_i)}\eta_\lambda(x_i)$.

Empirical Performance

Parallel to the logistic regression simulation, Poisson samples were generated on $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$ with log intensity

$$\eta(x) = 3\left\{10^5 x^{11}(1-x)^6 + 10^3 x^3(1-x)^{10}\right\} + 0.1.$$

Five cubic spline fits were calculated on each replicate and their performances in terms of $L(\lambda) = L(\eta, \eta_\lambda)$ of (5.4) were recorded. The results from one hundred replicates are summarized in Fig. 5.4. It is evident that the fudge factor $\alpha = 1.4$ helps here, for both the direct cross-validation via $V(\lambda)$ of (5.31) and the performance-oriented iteration driven by $U_w(\lambda)$. The two approaches gave nearly identical results, for $\alpha = 1.4$.

Software Illustration

For the Poisson family, both `gssanova` and `gssanova1` have a default `alpha=1.4`. The following sequence generates some Poisson responses on a grid and fits a cubic spline to the log intensity:

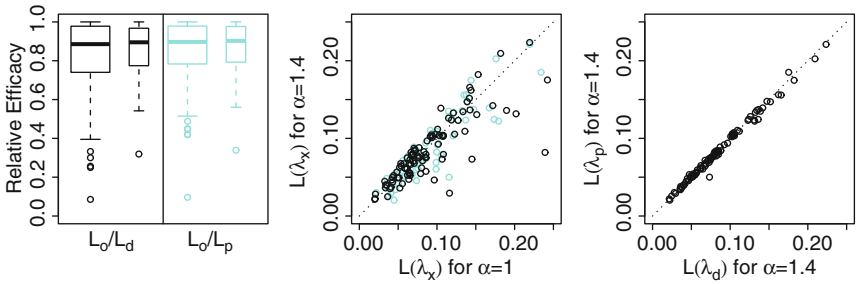


FIGURE 5.4. Effectiveness of $V(\lambda)$ and $U_w(\lambda)$ in Poisson regression simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1.4$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

```
set.seed(5732)
test <- function(x)
        {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))+.1}
x <- (0:100)/100
lam <- test(x)
y <- rpois(x,lam)
fit.pois <- gssanova1(y~x,family="poisson")
est <- predict(fit.pois,data.frame(x=x),se=TRUE)
```

The fit is shown in the right frame of Fig. 5.3, with the data and the test function superimposed:

```
plot(x,y,col=3); lines(x,lam,lty=2)
lines(x,exp(est$fit))
lines(x,exp(est$fit+1.96*est$se),col=5)
lines(x,exp(est$fit-1.96*est$se),col=5)
```

5.4.4 Gamma Family

The gamma distribution Gamma(α, β) has a density

$$\frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}, \quad \alpha, \beta, y > 0,$$

where α is the shape parameter and β is the scale parameter. When $\alpha = 1$, the gamma distribution reduces to the exponential distribution. Reparameterizing by (α, μ), where $\mu = \alpha\beta = E[Y]$, and dropping terms that do not involve μ, one has a minus log likelihood

$$\left\{\frac{y}{\mu} + \log \mu\right\}\alpha = \{ye^{-\eta} + \eta\}\alpha = l(\eta; y)/\sigma^2, \tag{5.32}$$

with $\sigma^2 = \alpha^{-1}$ being the dispersion parameter and $\eta = \log\mu$; see Problem 5.10. To avoid the constraint associated with the canonical parameter $-\mu^{-1}$, we choose to work with the log link $\eta = \log\mu$. It is easy to verify that $u = dl/d\eta = -y/\mu + 1$ and $w = d^2l/d\eta^2 = y/\mu$; $E[u] = 0$, $\mathrm{Var}[u] = \sigma^2$, and $E[w] = 1$. To iterate on (5.3), $\tilde{u}_i = -Y_i/\tilde{\mu}(x_i) + 1$ as usual, but we use $\tilde{w}_i = 1$, the expected value of w, as in Fisher's scoring.

To drive the performance-oriented iteration, one may use $V(\lambda)$ in general, or use $U(\lambda)$ when the dispersion is known such as with the exponential distribution.

Kullback-Leibler and Direct Cross-Validation

Using a non-canonical link, much of the general developments in §5.2 need due modifications. The Kullback-Leibler distance is given by

$$\mathrm{KL}(\eta, \tilde{\eta}) = -\mu(e^{-\eta} - e^{-\tilde{\eta}}) - (\eta - \tilde{\eta}) = (\mu/\tilde{\mu} - 1) - (\eta - \tilde{\eta}),$$

so (5.4) becomes

$$L(\lambda) = L(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\mu(x_i)}{\mu_\lambda(x_i)} + \frac{\mu_\lambda(x_i)}{\mu(x_i)} - 2\Big), \tag{5.33}$$

which will be used as the performance measure in the simulation below. The relative Kullback-Leibler distance of (5.8) is now

$$\mathrm{RKL}(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{\mu(x_i)}{\mu_\lambda(x_i)} + \eta_\lambda(x_i),$$

and (5.9) should look like

$$\begin{aligned} V_0(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{Y_i}{\mu_\lambda(x_i)} + \eta_\lambda(x_i)\Big\} + \frac{1}{n}\sum_{i=1}^{n} Y_i(e^{-\eta_\lambda^{[i]}(x_i)} - e^{-\eta_\lambda(x_i)}) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{Y_i}{\mu_\lambda(x_i)} + \eta_\lambda(x_i)\Big\} + \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\mu_\lambda(x_i)}\big(\eta_\lambda(x_i) - \eta_\lambda^{[i]}(x_i)\big). \end{aligned}$$

Replacing $\eta_\lambda^{[i]}(x_i)$ by $\eta_{\lambda,\eta_\lambda}^{[i]}(x_i) = \eta_\lambda(x_i) + a_{i,i}\tilde{u}_i/(1 - a_{i,i})$ [see (5.16) on page 184] and following the same procedures leading to (5.18), one obtains

$$\begin{aligned} V_g(\lambda) = &\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{Y_i}{\mu_\lambda(x_i)} + \eta_\lambda(x_i)\Big\} \\ &+ \alpha\frac{\mathrm{tr}A}{n - \mathrm{tr}A}\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\mu_\lambda(x_i)}\Big(\frac{Y_i}{\mu_\lambda(x_i)} - 1\Big), \end{aligned} \tag{5.34}$$

where a fudge factor $\alpha \geq 1$, not to be confused with the shape parameter σ^{-2}, is attached to the second term.

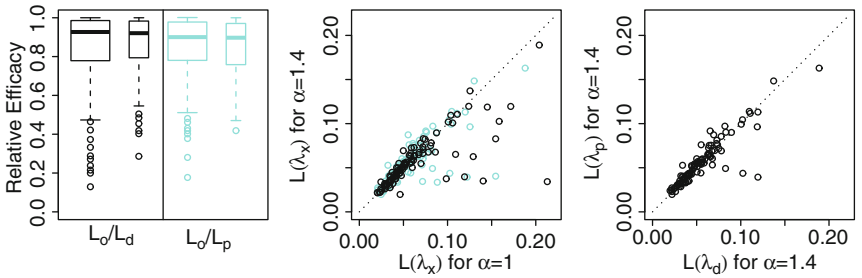


FIGURE 5.5. Effectiveness of $V_g(\lambda)$ and $V(\lambda)$ in gamma regression simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1.4$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

Empirical Performance

In a simulation study parallel to those for the binomial and Poisson families, gamma responses were generated on $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$ with a shape parameter $\sigma^{-2} = 2$ and a mean function

$$\mu(x) = 3\{10^5 x^{11}(1-x)^6 + 10^3 x^3 (1-x)^{10}\} + 0.1.$$

Five cubic spline fits were calculated on each replicate and their performances recorded in $L(\lambda)$ of (5.33). The performance-oriented iteration is now driven by $V(\lambda)$ of (3.23). The results from one hundred replicates are shown in Fig. 5.5. The fudge factor $\alpha = 1.4$ helps both methods, and with it, the performance-oriented iteration might have a tiny edge.

Software Illustration

For the gamma family, both `gssanova` and `gssanova1` have a default `alpha=1.4`. The following sequence generates some gamma responses on a grid and fits a cubic spline to the log mean:

```
set.seed(5732)
test <- function(x)
        {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))+.1}
x <- (0:100)/100
mu <- test(x)
y <- rgamma(x,shape=2,scale=mu/2)
fit.gamma <- gssanova1(y~x,family="Gamma")
est <- predict(fit.gamma,data.frame(x=x),se=TRUE)
```

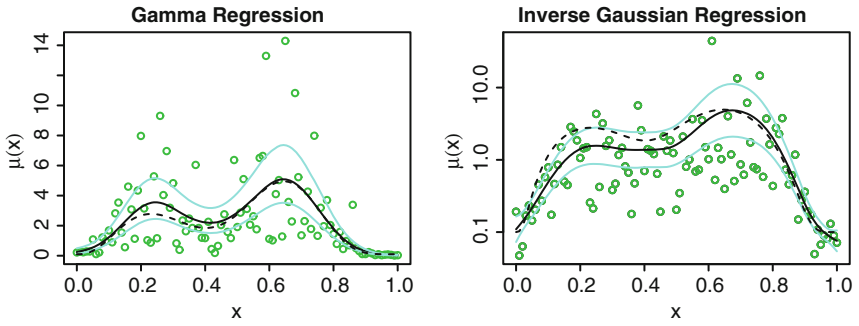


FIGURE 5.6. Cubic spline gamma and inverse Gaussian regression. The test functions are in *dashed lines*, the fits are in *solid lines*, and the 95 % Bayesian confidence intervals are in *faded lines*. The data are superimposed as *circles*.

The fit is shown in the left frame of Fig. 5.6, with the data and the test function superimposed:

```
plot(x,y,col=3); lines(x,mu,lty=2)
lines(x,exp(est$fit))
lines(x,exp(est$fit+1.96*est$se),col=5)
lines(x,exp(est$fit-1.96*est$se),col=5)
```

The dispersion σ^2 is needed for the calculation of standard errors, and is estimated using (3.26) on page 69 via (5.3) at $\tilde{\eta} = \eta_\lambda$.

5.4.5 Inverse Gaussian Family

The inverse Gaussian distribution $\mathrm{IG}(\mu, \sigma^2)$ has a density

$$\frac{1}{\sqrt{2\pi\sigma^2}} y^{-3/2} e^{-(y-\mu)^2/2\sigma^2\mu^2 y}, \quad \mu, \sigma^2, y > 0,$$

where $E[Y] = \mu$ and $\mathrm{Var}[Y] = \sigma^2\mu^3$. Dropping terms that do not involve μ, one has a minus log likelihood

$$\left\{\frac{y}{2\mu^2} - \frac{1}{\mu}\right\}\frac{1}{\sigma^2} = \{ye^{-2\eta}/2 - e^{-\eta}\}/\sigma^2 = l(\eta; y)/\sigma^2, \tag{5.35}$$

with σ^2 as the dispersion parameter and $\eta = \log\mu$; see Problem 5.11. Working with the log link $\eta = \log\mu$, one has $u = dl/d\eta = -y/\mu^2 + 1/\mu$ and $w = d^2l/d\eta^2 = 2y/\mu^2 - 1/\mu$; $E[u] = 0$, $\mathrm{Var}[u] = \sigma^2/\mu$, and $E[w] = 1/\mu$. To iterate on (5.3), we take $\tilde{u}_i = -Y_i/\tilde{\mu}^2(x_i) + 1/\tilde{\mu}(x_i)$ and $\tilde{w}_i = 1/\tilde{\mu}(x_i)$. To drive the performance-oriented iteration, one may use $V_w(\lambda)$.

Kullback-Leibler and Direct Cross-Validation

As with the gamma family, one needs to modify the calculations in §5.2. The Kullback-Leibler distance is given by

$$\mathrm{KL}(\eta,\tilde{\eta}) = -\frac{\mu}{2}(e^{-2\eta} - e^{-2\tilde{\eta}}) + (e^{-\eta} - e^{-\tilde{\eta}}) = \frac{\mu}{2\tilde{\mu}^2} - \frac{1}{\tilde{\mu}} + \frac{1}{2\mu}$$

and the equivalent of (5.4) looks like

$$L(\lambda) = L(\eta,\eta_\lambda) = \frac{1}{2n}\sum_{i=1}^{n}\Big(\frac{\mu(x_i)}{\mu_\lambda(x_i)} + \frac{\mu_\lambda(x_i)}{\mu(x_i)} - 2\Big)\Big(\frac{1}{\mu(x_i)} + \frac{1}{\mu_\lambda(x_i)}\Big), \quad (5.36)$$

The relative Kullback-Leibler distance of (5.8) is now

$$\mathrm{RKL}(\eta,\eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\mu(x_i)}{2\mu_\lambda^2(x_i)} - \frac{1}{\mu_\lambda(x_i)}\Big),$$

and (5.9) becomes

$$V_0(\lambda) \approx \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{Y_i}{2\mu_\lambda^2(x_i)} - \frac{1}{\eta_\lambda(x_i)}\Big\} + \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\mu_\lambda^2(x_i)}\big(\eta_\lambda(x_i) - \eta_\lambda^{[i]}(x_i)\big)$$

Replacing $\eta_\lambda^{[i]}(x_i)$ by $\eta_{\lambda,\eta_\lambda}^{[i]}(x_i) = \eta_\lambda(x_i) + (\tilde{u}_i/\tilde{w}_i)a_{i,i}/(1-a_{i,i})$ and following the procedures leading to (5.18), one has

$$\begin{aligned} V_g(\lambda) = {} & \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{Y_i}{2\mu_\lambda^2(x_i)} - \frac{1}{\mu_\lambda(x_i)}\Big\} \\ & + \alpha\frac{\mathrm{tr}(A_w W^{-1})}{n - \mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{\mu_\lambda^2(x_i)}\Big(\frac{Y_i}{\mu_\lambda^2(x_i)} - \frac{1}{\mu_\lambda(x_i)}\Big). \end{aligned} \quad (5.37)$$

Empirical Performance

Parallel to the simulations for previous families, inverse Gaussian responses were generated on $x_i = (i-0.5)/100, i = 1,\ldots,100$ with a dispersion $\sigma^2 = 1$ and a mean function

$$\mu(x) = 3\big\{10^5 x^{11}(1-x)^6 + 10^3 x^3(1-x)^{10}\big\} + 0.1.$$

Five cubic spline fits were calculated on each replicate and their performances recorded in $L(\lambda)$ of (5.36). Results from one hundred replicates are summarized in Fig. 5.7, The performance-oriented iteration driven by $V_w(\lambda)$, with $\alpha = 1.4$, emerges as the clear winner.

The inverse Gaussian family is numerically challenging; this might be related to the skewness of the distribution, which grows with μ. Initially,

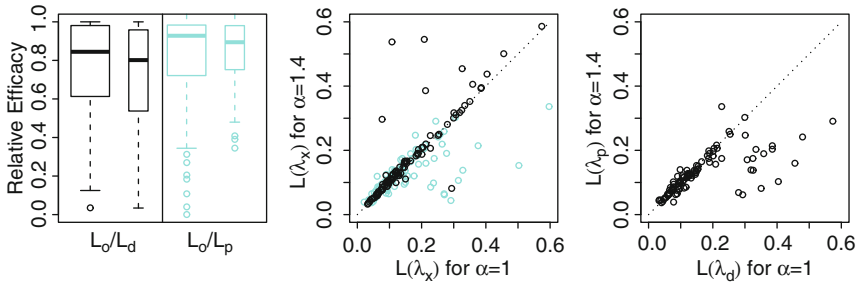


FIGURE 5.7. Effectiveness of $V_g(\lambda)$ and $V_w(\lambda)$ in inverse Gaussian regression simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$; four solid, two faded points are off the chart. *Right*: $L(\lambda_d)$ with $\alpha = 1$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

we had great difficulty trying to locate $L(\lambda_o)$ and $L(\lambda_d)$; optimization algorithms are easily trapped in plateaus at larger λ values, and we had to adjust internal settings in `gssanova` just to accommodate this family. Also, iterations on (5.3) with fixed-λ can experience more difficulties with none or slow convergence than performance-oriented iteration.

Software Illustration

The following sequence generates some inverse Gaussian responses on a grid and fits a cubic spline to the log mean; the function `rinvgauss` can be found in the R package `statmod` by Gordon Smyth:

```
set.seed(5732)
test <- function(x)
        {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))+.1}
x <- (0:100)/100
mu <- test(x)
y <- rinvgauss(x,mu)
fit.ig <- gssanova1(y~x,family="inverse.gaussian")
est <- predict(fit.ig,data.frame(x=x),se=TRUE)
```

The fit is shown in the right frame of Fig. 5.6, with the data and the test function superimposed:

```
plot(x,y,log="y",col=3); lines(x,mu,lty=2)
lines(x,exp(est$fit))
lines(x,exp(est$fit+1.96*est$se),col=5)
lines(x,exp(est$fit-1.96*est$se),col=5)
```

5.4.6 Negative Binomial Family

The negative binomial distribution has a density

$$\frac{\Gamma(\nu+y)}{y!\,\Gamma(\nu)}p^{\nu}(1-p)^{y}, \quad \nu>0,\ p\in(0,1),\ y=0,1,\ldots. \tag{5.38}$$

For ν an integer, the distribution describes the number of failures before the νth success in a sequence of Bernoulli trials with a success probability p. The distribution also describes the behavior of composite Poisson data with $Y\sim$ Poisson(λ) and $\lambda\sim$ Gamma$\big(\nu,(1-p)/p\big)$; see Problem 5.12. It can be shown that $E[Y]=\nu(1-p)/p$ and $\mathrm{Var}[Y]=\nu(1-p)/p^2$. Taking the logit link $\eta=\log\big\{p/(1-p)\big\}$ and dropping terms that do not involve η, one has a minus log likelihood

$$(\nu+y)\log(1+e^{\eta})-\nu\eta=l(\eta;y); \tag{5.39}$$

see Problem 5.13. It follows (Problem 5.14) that $u=dl/d\eta=(\nu+y)p-\nu$ and $w=d^2l/d\eta^2=(\nu+y)p(1-p)$; $E[u]=0$, $\mathrm{Var}[u]=\nu(1-p)$, and $E[w]=\nu(1-p)$. To iterate on (5.3), one may use $\tilde{u}_i=(\nu_i+Y_i)\tilde{p}(x_i)-\nu_i$ and $\tilde{w}_i=\nu_i\big(1-\tilde{p}(x_i)\big)$.

It is assumed that ν_i's are known. It is also possible to assume a common but unknown ν, under which one iterates between the estimations of ν and $\eta(x)$; given (Y_i,p_i), one may estimate ν via the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\big\{\log\Gamma(\nu)-\log\Gamma(\nu+Y_i)-\nu\log p_i\big\}. \tag{5.40}$$

Either way, the estimation of $\eta(x)$ is under known ν_i.

The dispersion is known to be $a(\phi)=1$, so performance-oriented iteration can be driven by $U_w(\lambda)$ with $\sigma^2=1$.

Kullback-Leibler and Direct Cross-Validation

The Kullback-Leibler distance is seen to be

$$\mathrm{KL}(\eta,\tilde{\eta})=\frac{\nu}{p}\log\frac{1-p}{1-\tilde{p}}+\nu(\eta-\tilde{\eta}),$$

and the counterpart of (5.4) is now

$$L(\lambda)=L(\eta,\eta_\lambda)=\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\nu_i}{p(x_i)}-\frac{\nu_i}{p_\lambda(x_i)}\Big)\log\frac{1-p(x_i)}{1-p_\lambda(x_i)}. \tag{5.41}$$

The relative Kullback-Leibler distance of (5.8) becomes

$$\mathrm{RKL}(\eta,\eta_\lambda)=-\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{\nu_i}{p(x_i)}\log\big(1-p_\lambda(x_i)\big)+\nu_i\,\eta_\lambda(x_i)\Big\},$$

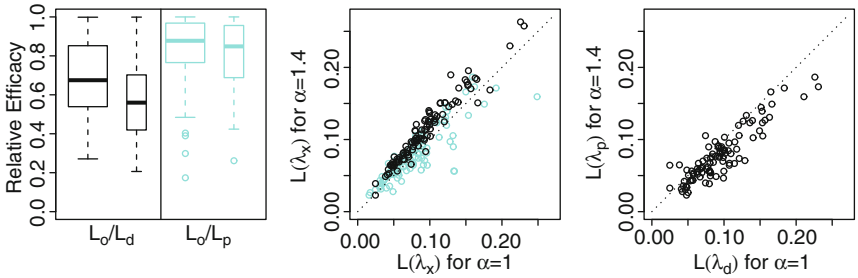


FIGURE 5.8. Effectiveness of $V_g(\lambda)$ and $V_w(\lambda)$ in negative binomial simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$; four solid, two faded points are off the chart. *Right*: $L(\lambda_d)$ with $\alpha = 1$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

and, noting that $E[\nu + y] = \nu/p$, (5.9) looks like

$$V_0(\lambda) \approx -\frac{1}{n}\sum_{i=1}^{n}\big\{(\nu_i + Y_i)\log\big(1 - p_\lambda(x_i)\big) + \nu_i\eta_\lambda(x_i)\big\} + \frac{1}{n}\sum_{i=1}^{n} Y_i\, p_\lambda(x_i)\big(\eta_\lambda^{[i]}(x_i) - \eta_\lambda(x_i)\big).$$

The same procedures leading to (5.18) yield

$$V_g(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\big\{(\nu_i + Y_i)\log\big(1 - p_\lambda(x_i)\big) + \nu_i\eta_\lambda(x_i)\big\} + \alpha\frac{\mathrm{tr}(A_w W^{-1})}{n - \mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^{n} Y_i\, p_\lambda(x_i)\big\{(\nu_i + Y_i)p_\lambda(x_i) - \nu_i\big\}. \quad (5.42)$$

Empirical Performance

Parallel to the simulations for the other families, negative binomial samples were drawn on $x_i = (i - 0.5)/100$, $i = 1, \ldots, 100$ with $\nu = 3$ and a mean function

$$\mu(x) = 3\big\{10^5 x^{11}(1-x)^6 + 10^3 x^3(1-x)^{10}\big\} + 0.1.$$

For each of the one hundred replicates generated, five cubic splines were fitted to the logit and their performances recorded in $L(\lambda)$ of (5.41). The results are summarized in Fig. 5.8. The fudge factor $\alpha = 1.4$ helps the performance-oriented iteration but not the direct cross-validation, and $U_w(\lambda)$ with $\alpha = 1.4$ emerges as the method of choice.

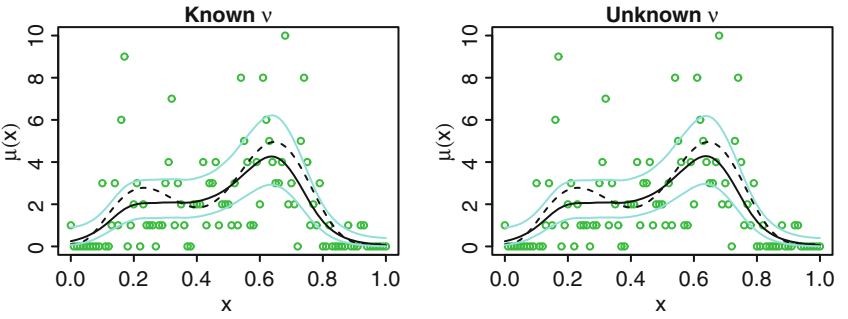


FIGURE 5.9. Cubic spline logistic fit to negative binomial data. The test functions are in *dashed lines*, the fits are in *solid lines*, and the 95% Bayesian confidence intervals are in *faded lines*. The data are superimposed as *circles*.

Software Illustration

Negative binomial responses can be entered in two ways, either in two columns of (Y_i, ν_i) or in a vector of Y_i; in the latter case, ν is unknown but assumed to be common to all observations. This is similar to the binomial family syntax-wise, but the two types of responses are not equivalent here.

The following sequence generates some negative binomial responses with $\nu = 3$ and fits a cubic spline to the logit:

```
set.seed(5732)
test <- function(x)
        {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))+.1}
x <- (0:100)/100
mu <- test(x); nu <- 3
p <- nu/(mu+nu)
y <- rnbinom(x,nu,p)
fit.nb <- gssanova1(cbind(y,nu)~x,family="nbinomial")
est <- predict(fit.nb,data.frame(x=x),se=TRUE)
```

The fit is shown in the left frame of Fig. 5.9, with the data and the test function superimposed:

```
plot(x,y,col=3); lines(x,mu,lty=2)
lines(x,nu/exp(est$fit))
lines(x,nu/exp(est$fit+1.96*est$se),col=5)
lines(x,nu/exp(est$fit-1.96*est$se),col=5)
```

One may also submit the responses as a vector:

```
fit.nb1 <- gssanova1(y~x,family="nbinomial",
                           id.basis=fit.nb$id.basis)
fit.nb1$nu
# 3.347354
```

with the resulting fit shown in the right frame of Fig. 5.9. The ν estimate in `fit.nb1` might be off, but together with the corresponding $\eta(x)$ estimate they produced virtually the same estimate for $\mu(x) = \nu e^{-\eta(x)}$.

With ν unknown, its updating is concurrent with $\eta(x)$. Every time a new set of $\tilde{\eta}(x_i)$ come from (5.3), ν is updated via the minimization of (5.40), and $\tilde{u}_i$ and $\tilde{w}_i$ are calculated using this updated ν to form the $\tilde{Y}_i$ for the next iteration; this is done for both the performance-oriented iteration and the fixed-λ iteration with direct cross-validation. The procedure is not guaranteed to converge, but we have yet to encounter any problems as of this writing. The selection of λ in the performance-oriented iteration is unaffected by this as comparisons are only made of estimates based on the same ν. The direct cross-validation compares estimates based on different ν's, however, so one needs to add back to (5.42) the terms,

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{\log\Gamma(\nu) - \log\Gamma(\nu + Y_i)\Big\},$$

which were earlier dropped from (5.39) as they do not involve η.

5.5 Case Studies

We now apply the techniques developed in this chapter to analyze a few real data sets. It will be seen that Poisson regression can be used to estimate a probability density and that gamma regression can be used to estimate the spectral density of a stationary time series.

5.5.1 Eruption Time of Old Faithful

Listed in Härdle (1991) are the duration and the waiting time to the next eruption gathered from 272 consecutive eruptions of the Old Faithful geyser in Yellowstone National Park. The data are available in R as a data frame `faithful` with elements `eruptions` and `waiting`, both in minutes. In this study, we use `eruptions` to estimate a continuous "mass spectrum" of the eruption duration.

The range of the eruption times is $[1.6, 5.1]$. Rounding the data to a histogram of 30 bins on $[1.5, 5.25]$, each of length 0.125, one has x_i as the middle points of the bins and Y_i as the frequencies of the bins:

```
data(faithful); erup <- faithful$eruptions
jk <- hist(erup,bre=seq(1.5,5.25,length=31),plot=FALSE)
x <- jk$mids
y <- jk$counts
```

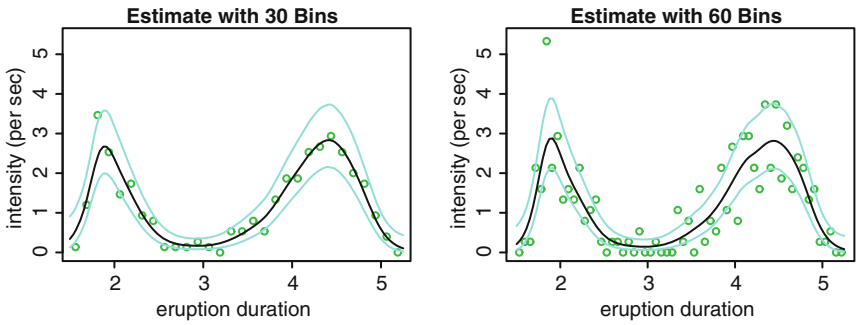


FIGURE 5.10. Mass spectrum of eruption duration of Old Faithful. The estimated Poisson intensity is in *solid lines* and the 95 % Bayesian confidence intervals in *faded lines*. The data are superimposed as *circles*.

The continuous "mass spectrum" can be estimated through a cubic spline Poisson regression, which is plotted in the left frame of Fig. 5.10:

```
fit.faith <- gssanova(y~x,family="poisson",
                      offset=rep(log(60/8),30))
xx <- seq(1.5,5.25,length=101)
est <- predict(fit.faith,data.frame(x=xx,offset=0),TRUE)
plot(x,y*8/60,col=3,ylim=c(0,6))
lines(xx,exp(est$fit))
lines(xx,exp(est$fit+1.96*est$se),col=5)
lines(xx,exp(est$fit-1.96*est$se),col=5)
```

The `offset` term scales the estimate to the unit of per-second intensity; note that Y_i are counts per 1/8 min. For the evenly binned data given here, the offset is not necessary, as one can always rescale the fit afterwards, but if the data are given in heterogeneous units, the offset provides a convenient device to align them to a common scale; see Problem 5.15.

Repeating the process with a histogram of 60 bins on $[1.5, 5.25]$, one gets the estimate in the right frame of Fig. 5.10.

Scaling the Poisson intensity to integrate to 1 on the domain $[1.5, 5.25]$, one gets a probability density of eruption duration; see §7.2. The Bayesian confidence intervals lose their meaning for a density, however. An analysis of the data using density estimation techniques will be shown in §7.5.2.

5.5.2 Spectrum of Yearly Sunspots

The yearly number of sunspots from 1700 to 1988 can be found in Tong (1990, page 471). Our task here is to estimate the frequency spectrum of the series.

For a stationary time series X_t, $t = 0, \pm 1, \pm 2, \ldots$ with covariance function $\gamma_k = \mathrm{Cov}(X_t, X_{t+k})$, the spectral density is defined by

$$f(\omega) = \frac{1}{\gamma_0} \sum_{k=-\infty}^{\infty} \gamma_k e^{-\mathbf{i}2\pi k\omega}, \quad \omega \in (-0.5, 0.5),$$

where $\mathbf{i} = \sqrt{-1}$, which satisfies

$$\gamma_k = \gamma_0 \int_{-0.5}^{0.5} f(\omega) e^{\mathbf{i}2\pi k\omega} d\omega.$$

See, e.g., Priestley (1981, §4.8.3) and Brockwell and Davis (1991, §4.3), where the frequency is parameterized by $\tilde{\omega} = 2\pi\omega \in (-\pi, \pi)$. The spectral density is an even function, so one only needs to estimate $f(\omega)$ on $(0, 0.5)$. Observing x_t, $t = 1, \ldots, T$, one may calculate the discrete Fourier transform (cf. §4.2.2)

$$\tilde{x}_\nu = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} x_t e^{-\mathbf{i}2\pi t\nu/T}, \quad \nu = 0, 1, \ldots, T-1, \tag{5.43}$$

which yields the periodogram $I(\omega_\nu) = |\tilde{x}_\nu|^2$ on the so-called Fourier frequencies $\omega_\nu = \nu/T$. Note that $I(\omega_\nu) = I(\omega_{T-\nu})$. For T large, it can be shown that $I(\omega_\nu)$, $\omega_\nu \in (0, 0.5)$, are asymptotically independent exponential random variables with means $E\big[I(\omega_\nu)\big] \propto f(\omega_\nu)$; see, e.g., Priestley (1981, page 425) and Brockwell and Davis (1991, Theorem 10.3.2). The estimation of the spectrum can thus be obtained from a gamma regression with $x_\nu = \omega_\nu$ and $Y_\nu = I(\omega_\nu)$.

The observed series are available in R as a `ts` object `sunspot.year`. The following sequence loads the data, calculates the periodogram, and sets up x_ν and Y_ν for gamma regression:

```
data(sunspot.year)
n <- length(sunspot.year)
ind <- 1:(ceiling(n/2)-1)
y <- (abs(fft(sunspot.year))^2/n)[-1][ind]
x <- ind/n
```

The R function `fft` calculates an unscaled discrete Fourier transform [i.e., the transform given in (5.43) but without $1/\sqrt{T}$]. A cubic spline can now be fitted to the log periodogram via gamma regression and plotted as in the right frame of Fig. 5.11:

```
set.seed(5732)
fit.sunspot <- gssanova(y~x,family="Gamma")
xx <- seq(0,.5,length=101)
est <- predict(fit.sunspot,data.frame(x=xx),se=TRUE)
```

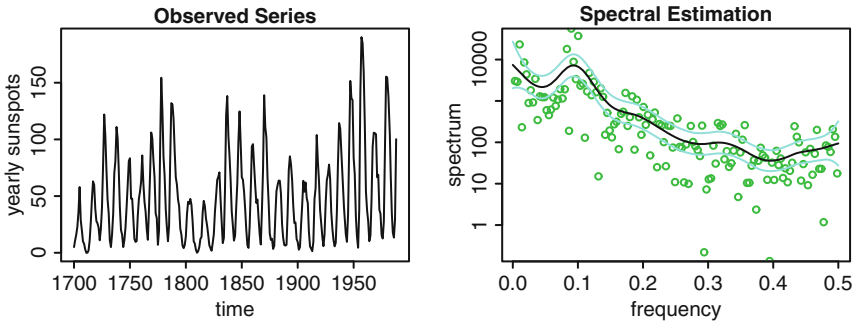


FIGURE 5.11. Spectrum of yearly sunspots. *Left*: Observed series. *Right*: Spectral estimate with 95 % Bayesian confidence intervals; the periodogram is superimposed as *circles*.

```
plot(x,y,log="y",col=3)
lines(xx,exp(est$fit))
lines(xx,exp(est$fit+1.96*est$se),col=5)
lines(xx,exp(est$fit-1.96*est$se),col=5)
```

Scaling the estimate to integrate to 0.5 on $(0, 0.5)$, one gets the spectral density. The Bayesian confidence intervals lose their meaning for a spectral density, however.

The performance-oriented iteration using $\tilde{w}_i = 1$ in (5.3) encountered numerical overflow within the first few steps, so we had to resort to `gssanova`. The performance-oriented iteration using $\tilde{w}_i = Y_i/\tilde{\mu}(x_i)$ did converge, however, as reported in the first edition of this book; the original `gssanova` used $\tilde{w}_i = Y_i/\tilde{\mu}(x_i)$ in performance-oriented iteration. As can be seen in the right frame of Fig. 5.11, Y_i here are extremely imbalanced in magnitude; $\tilde{u}_i/\tilde{w}_i = 1 - Y_i/\tilde{\mu}(x_i)$ for $\tilde{w}_i = 1$ inherit much of this imbalance whereas $\tilde{u}_i/\tilde{w}_i = \tilde{\mu}(x_i)/Y_i - 1$ for $\tilde{w}_i = Y_i/\tilde{\mu}(x_i)$ are moderated a bit. We nevertheless choose to use $\tilde{w}_i = 1$ in the implementation as they lead to much better performances by the performance-oriented iteration in simulations when things do converge. Direct cross-validation does not seem to be affected by the choice of $\tilde{w}_i$, however. The fit shown here is hardly distinguishable from the one presented in the first edition of this book.

5.5.3 Progression of Diabetic Retinopathy

The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) was an epidemiological study of a cohort of patients receiving their medical care in an 11-county area in southern Wisconsin, who were first examined in 1980–1982, then again in 1984–1986, 1990–1992, and 1994–1996. A subset derived from the WESDR data is distributed in GRKPACK (Wang 1997), to be found at

http://www.pstat.ucsb.edu/faculty/yuedong/software.html

which consists of the baseline measures of duration of diabetes in years, percent of glycosylated hemoglobin, body mass index, and a binary indicator of retinopathy progression at the first follow-up, of 669 patients. There were 278 positive cases among the 669 patients.

The data are included in `gss` as a data frame `wesdr` with elements `dur`, `gly`, `bmi`, and `ret`. A tensor product cubic spline can be fitted to the logit of retinopathy progression, with all interactions included, and the cosine diagnostics of §3.7 and the Kullback-Leibler projection of §5.3.2 inspected; the cosine diagnostics are based on the weighted least squares at the fit:

```
data(wesdr); set.seed(5732)
fit.wsd <- gssanova(ret~dur*bmi*gly,data=wesdr,
                    family="binomial")
sum.fit <- summary(fit.wsd,diag=TRUE)
round(sum.fit$kappa,2)
#   dur  bmi  gly dur:bmi dur:gly bmi:gly dur:bmi:gly
#  1.37 1.58 5.88    1.92    5.90    6.59        6.64
round(sum.fit$pi,2)
#   dur  bmi  gly dur:bmi dur:gly bmi:gly dur:bmi:gly
#  0.10 0.11 1.18    0.01   -0.05   -0.55        0.19
round(sum.fit$cos,2)
#        dur  bmi   gly dur:bmi dur:gly bmi:gly
# cos.y 0.11 0.06  0.32    0.02   -0.28   -0.29
# cos.e 0.03 0.01  0.00    0.01    0.00    0.00
# norm  4.86 7.45 14.75    6.66    0.65    7.50
# cos.y  dur:bmi:gly  yhat     y     e
# cos.e         0.26  0.40  1.00  0.93
# norm          0.00  0.02  0.93  1.00
#               2.97 10.54 27.99 25.67
project(fit.wsd,c("dur","bmi","gly"))$ratio
# 0.02744856
```

High concurvity and negative π_β's and $\cos(W^{1/2}\mathbf{Y}^*, W^{1/2}\mathbf{f}_\beta^*)$'s are associated with several interaction terms, so they might be just offsetting each other, and the Kullback-Leibler projection suggests the adequacy of an additive model. One can now fit a cubic spline additive model and evaluate the components on the sampling points:

```
fit.wsd.a <- gssanova(ret~dur+bmi+gly,"binomial",
                      data=wesdr,id.basis=fit.wsd$id)
sum.fit.a <- summary(fit.wsd.a,diag=TRUE)
round(sum.fit.a$kappa,2)
#  dur  bmi  gly
# 1.01 1.04 1.03
round(sum.fit.a$pi,2)
```

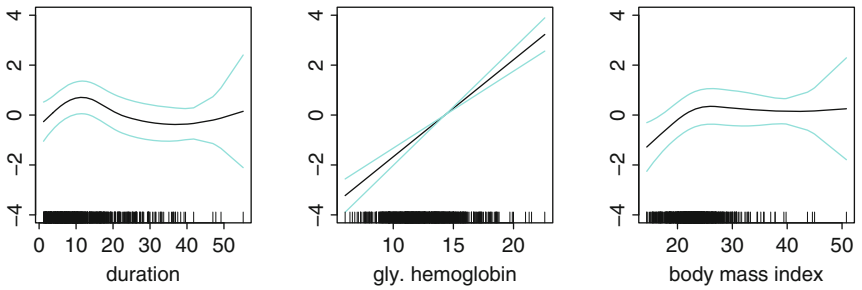


FIGURE 5.12. Factors affecting diabetic retinopathy progression. *Left*: Effect of duration of diabetes. *Center*: Effect of percent of glycosylated hemoglobin. *Right*: Effect of body mass index. The logit components are in *solid lines* and the 95 % Bayesian confidence intervals in *faded*. The rugs on the bottom mark the sampling points.

```
#  dur  bmi  gly
# 0.13 0.07 0.79
round(sum.fit.a$cos,2)
#       dur  bmi  gly  yhat     y     e
#cos.y 0.18 0.10 0.32  0.39  1.00  0.93
#cos.e 0.04 0.02 0.00  0.02  0.93  1.00
#norm  3.61 3.75 9.79 10.49 28.00 25.75
project(fit.wsd.a,c("dur","bmi"))$ratio
# 0.7306119
project(fit.wsd.a,c("dur","gly"))$ratio
# 0.08031866
project(fit.wsd.a,c("bmi","gly"))$ratio
# 0.08023298
est.dur <- predict(fit.wsd.a,wesdr,se=TRUE,inc="dur")
est.bmi <- predict(fit.wsd.a,wesdr,se=TRUE,inc="bmi")
est.gly <- predict(fit.wsd.a,wesdr,se=TRUE,inc="gly")
```

Binary data are intrinsically noisy and the weighted least squares is only a local approximation, so the consine diagnostics are not as easy to calibrate. The Kullback-Leibler projection however suggests that none of the remaining terms can be eliminated. The fitted logit components are plotted in Fig. 5.12 along with the respective Bayesian confidence intervals. The effect of glycosylated hemoglobin was linear and was dominant. The rugs on the bottom of the plots mark the marginal sampling points, and it is comforting to see that the standard errors are larger in sparse areas.

The deviance of the additive fit is 746.76 and that of the full interaction fit is 742.46. For comparison, a linear logistic regression yields a deviance of 780.98, with the duration effect insignificant (p-value 0.45).

5.5.4 Colorectal Cancer Mortality Rate

Information concerning cancer deaths in the United States is scattered in government registries. The county-wise death counts of colorectal cancer patients during years 2000–2004 in the state of Indiana were compiled by Tonglin Zhang and Ge Lin, along with selected demographic information from Census 2000 and geographic locations of the county governments. Part of the data are included in `gss` as a data frame `ColoCan`, with elements `pop` (population in 2000), `event` (death count of colorectal cancer), `sex`, `wrt` (proportion of whites), `brt` (proportion of blacks), `ort` (proportion of other minorities), `scrn` (screening rate for adults over 50), `lat` (latitude), `lon` (longitude), and `geog`; `geog` is a matrix with two columns of x-y coordinates as given in (4.32), where (ϕ_0, θ_0) is taken as Indianapolis, the state capital located in Marion county. The variables `pop` and `event` are given for males and females separately so there are a total of 184 rows for the 92 counties, with the top 92 rows containing the male data and the next 92 containing the female data; the racial proportions are however for both sexes together.

One may fit a standard Poisson regression model $Y \sim \text{Poisson}(e^{\eta(x)}\delta)$, where Y is the event count, δ is the population, and the log mortality rate $\eta(x)$ is a function of covariates:

```
data(ColoCan); set.seed(5732)
fit.cc.0 <- gssanova(event~sex*(geog+brt+ort+scrn),
                     "poisson",offset=log(pop),
                     data=ColoCan,nbasis=40)
```

where `nbasis=40` sets $q = 40$; `gssanova1` again ran into numerical problems with this data set whereas `gssanova` is reliable as usual. Note that only two of `wrt`, `brt`, and `ort` can be included as they add up to one. The terms `scrn`, `sex:brt`, `sex:ort`, and `sex:scrn` are negligible:

```
project(fit.cc.0,c("sex","geog",
        "sex:geog","brt","ort"))$ratio
# 0.01886593
```

Trying to remove one more term from the model, however, would result in $\text{KL}(\hat{\eta}, \tilde{\eta})/\text{KL}(\hat{\eta}, \eta_c) > 7.7\,\%$, so the remaining terms are indispensable.

The colorectal cancer screening rate for adults over 50 was found by Zhang and Lin (2009) to be a significant factor that impacted the mortality rate, but geography and racial proportions were not used as covariates there. In our fit here, the screening effect appears to have been accounted for by the other effects in the model.

We now fit the model with five terms in the log mortality rate:

```
fit.cc <- gssanova(event~sex*geog+brt+ort,"poisson",
                   offset=log(pop),data=ColoCan,
                   id.basis=1:184)
```

where `id.basis=1:184` sets $q = n$. The joint effect of `sex` and `geog` can be depicted in two mortality maps, one for each sex. To evaluate such a map on a grid, say that for male, one may try:

```
x.gd <- seq(-.024,.017,length=42)
y.gd <- seq(-.031,.033,length=65)
grid <- cbind(rep(x.gd,65),rep(y.gd,rep(42,65)))
est.g.m <- predict(fit.cc,data.frame(geog=I(grid),
                        sex=as.factor(rep("M",42*65))),
                    TRUE,c("sex","geog","sex:geog"))
```

where `sex` must be a factor. One can then plot the mortality map:

```
library(maps)
map("county","indiana",col=5)
m.lat <- ColoCan$lat[49]; m.lon <- ColoCan$lon[49];
lon.gd <- xy2ltln(cbind(x.gd,0),c(m.lat,m.lon))[,2]
lat.gd <- xy2ltln(cbind(0,y.gd),c(m.lat,m.lon))[,1]
contour(lon.gd,lat.gd,matrix(est.g.m$fit,42,65),
        lty=3,add=T)
contour(lon.gd,lat.gd,matrix(est.g.m$se,42,65),
        levels=.08,lty=5,add=T)
```

where the R function `xy2ltln` is from §4.3.4 on page 140 and `m.lat`, `m.lon` mark the geographic location of Indianapolis/Marion county. The effects of racial proportions can be similarly obtained:

```
est.brt <- predict(fit.cc,ColoCan,TRUE,"brt")
est.ort <- predict(fit.cc,ColoCan,TRUE,"ort")
```

Shown in Fig. 5.13 are the two mortality maps and the effects of racial proportions. The 0.08 contours of the standard errors of the mortality maps trace the state boundary closely. Indianapolis and its vicinity enjoy the lowest mortality rate whereas the highest is midway between Indianapolis and Chicago. To compare the mortality rates at Indianapolis/Marion county (49th) and at Purdue/Tippecanoe county (79th), one may try:

```
est1 <- predict(fit.cc,ColoCan[c(49,79,141,171),],
                inc=c("sex","geog","sex:geog"))
exp(est1[2]-est1[1])
# 1.741715
exp(est1[4]-est1[3])
# 1.637541
```

where the 141st and 171st data entries are for females in Marion and Tippecanoe counties, respectively. The effects of racial proportions turn out to be linear, with blacks suffering higher mortality rate and other minorities enjoying lower mortality rate; the equivalent fit using (`wrt`,`bwt`)

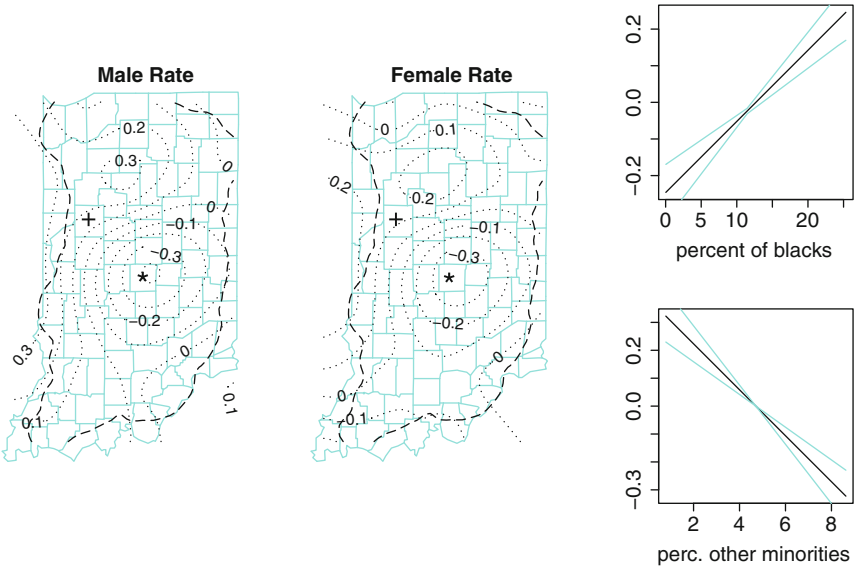


FIGURE 5.13. Components of log mortality rate of colorectal cancer. *Left* and *Center*: Geographic pattern for males (*left*) and females (*center*), with Marion county (*) and Tippecanoe county (+) marked; the 0.08 contours of standard errors are in *dashed lines*. *Right*: Effects of racial proportions.

has both slopes positive and that using (`wrt`,`owt`) has both slopes negative. It is well known that blacks have the highest colorectal cancer mortality rate, whites the second highest, with other races below them.

The deviance of `fit.cc.0` is 174.10 and that of `fit.cc` is 168.62.

5.6 Bibliographic Notes

Section 5.1

Penalized likelihood regression was formulated and studied by O'Sullivan, Yandell, and Raynor (1986); see also Silverman (1978) and Green and Yandell (1985). Fits with multiple penalty terms were found in Gu (1990) and Wahba, Wang, Gu, Klein, and Klein (1995), among others.

A standard reference on linear parametric regression with exponential family responses, better known as generalized linear models, is McCullagh and Nelder (1989), where extensive discussions can be found on the properties of exponential families and on the use of iterated weighted least squares in the fitting of generalized linear models.

Section 5.2

A suggestion in the early literature was to compute the minimizer η_λ of (5.1) for fixed λ, evaluate $V_w(\lambda|\tilde{\eta})$ at $\tilde{\eta} = \eta_\lambda$, and compare such $V_w(\lambda)$ values on a grid of λ. This amounts to comparing the scores on the dashed slice in Fig. 5.1. Since $V_w(\lambda|\tilde{\eta})$ with different $\tilde{\eta}$ are not comparable, this approach is ineffective, as was shown in Gu (1992a).

Performance-oriented iteration was used implicitly by Gu (1990), but the mechanism and the related issues were not understood until Gu (1992a). The direct cross-validation through $V_0(\lambda)$ of (5.9) was proposed by Cox and Chang (1990). Xiang and Wahba (1996) derived the more effective and computable GACV score $V_g(\lambda)$. Gu and Xiang (2001) derived the numerically stable, readily computable $V_g^*(\lambda)$ and proved the equivalence of $V_g(\lambda)$ and $V_g^*(\lambda)$.

Section 5.3

The adaptation of Bayesian confidence intervals for non-Gaussian regression was proposed and illustrated in Gu (1992c). Examples of componentwise intervals were shown in Wahba, Wang, Gu, Klein, and Klein (1995).

Hypothesis "testing" via Kullback-Leibler projection was developed in Gu (2004).

Section 5.4

The original `gssanova` suite was part of the `gss` package in its first public release dated back to 1999. GRKPACK, a collection of RATFOR routines implementing the performance-oriented iteration, was put together earlier by Wang (1997).

Extensive discussion of binomial, Poisson, and gamma distributions can be found in McCullagh and Nelder (1989). Facts concerning the inverse Gaussian distribution can be found in Chhikara and Folks (1989). Generalized linear model for the negative binomial family is discussed in Venables and Ripley (2002, §7.4).

The customizations of the direct cross-validation in the gamma, inverse Gaussian, and negative binomial families have not appeared in the literature.

Section 5.5

Various versions of the Old Faithful eruption data have been used in the literature to showcase regression and density estimation techniques; see,

e.g., Azzalini and Bowman (1990), Härdle (1991), and Scott (1992), among others. A nice discussion of density estimation through Poisson regression can be found in Lindsey (1997, Chap. 3).

The sunspot data or subsets thereof are among the most popular examples being used in textbooks and research articles on time series analysis. Spectral estimation through gamma regression was studied by Pawitan and O'Sullivan (1994); see also Cogburn and Davis (1974) and Wahba (1980).

Detailed descriptions of the WESDR data can be found in, e.g., Klein, Klein, Moss, Davis, and DeMets (1988, 1989), among others. The analysis presented here differs slightly from the one found in Wahba, Wang, Gu, Klein, and Klein (1995).

The Indiana colorectal cancer mortality data were compiled by Tonglin Zhang and Ge Lin and were analyzed in Zhang and Lin (2009).

5.7 Problems

Section 5.1

5.1 Consider the functional $L(f) = -\sum_{i=1}^{n} \{Y_i f(x_i) - b(f(x_i))\}$ in a reproducing kernel Hilbert space $\mathcal{H}$ with a square seminorm $J(f)$.

(a) Prove that $L(f)$ is continuous, convex, and Fréchet differentiable.

(b) Let $\{\phi_\nu, \nu = 1, \ldots, m\}$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$ and S be $n \times m$ with the (i, ν)th entry $\phi_\nu(x_i)$. Prove that if S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$.

(c) Prove that if S is of full column rank, then $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$.

Section 5.2

5.2 Prove Theorem 5.2.

5.3 Rewrite (3.12) on page 64 as

$$A_w(\lambda) = I - n\lambda W^{-1/2}(M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1})W^{-1/2},$$

where $M = Q + n\lambda W^{-1}$. Let $S = FR^* = (F_1, F_2)\binom{R}{O} = F_1 R$ be the QR-decomposition of S with F orthogonal and R upper-triangular.

(a) Show that

$$M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1} = F_2(F_2^T Q F_2 + n\lambda F_2^T W^{-1} F_2)^{-1} F_2^T.$$

(b) Let $H = (W + n\lambda F_2(F_2^T Q F_2)^+ F_2^T)^{-1}$. For $F_2^T Q F_2$ nonsingular, verify that $A_w(\lambda)(W^{1/2} H W^{1/2})^{-1} = I$.

5.4 Consider $U^*(\lambda)$ as given in (5.19) for penalized least squares regression.

(a) Show that $V_g(\lambda)$ of (5.18) reduces to $U^*(\lambda)$.

(b) Assume $n^{-1}\boldsymbol{\eta}^T\big(I - A(\lambda)\big)\boldsymbol{\eta} = o(1)$, where $\boldsymbol{\eta}^T = \big(\eta(x_1), \ldots, \eta(x_n)\big)$. If, in addition, Condition 3.2.1 of §3.2.1 also holds, that $nR(\lambda) \to \infty$, show that $U^*(\lambda) - L(\lambda) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L(\lambda)\big)$.

Section 5.3

5.5 Prove (5.20).

5.6 Prove (5.21).

5.7 Verify (5.26); set $\eta = \tilde{\eta} + \alpha(\tilde{\eta} - \eta_c)$ in (5.25) for α real and differentiate with respect to α.

Section 5.4

5.8 Show that $u = dl/d\eta = -y + mp$ and $w = d^2l/d\eta^2 = mp(1-p)$ for the binomial minus log likelihood $l(\eta; y)$ in (5.27) with $\eta = \log\big\{p/(1-p)\big\}$.

5.9 Show that $u = dl/d\eta = -y + \lambda$ and $w = d^2l/d\eta^2 = \lambda$ for the Poisson minus log likelihood $l(\eta; y)$ in (5.29) with $\eta = \log\lambda$.

5.10 Derive the minus log likelihood (5.32) for the gamma family.

5.11 Derive the minus log likelihood (5.35) for the inverse Gaussian family.

5.12 Derive the probability density (5.38) for composite Poisson data with $Y \sim \text{Poisson}(\lambda)$ and $\lambda \sim \text{Gamma}\big(\nu, (1-p)/p\big)$.

5.13 Derive the minus log likelihood (5.39) for the negative binomial family with $\eta = \log\big\{p/(1-p)\big\}$.

5.14 Show that for the negative binomial minus log likelihood $l(\eta; y)$ in (5.39) with $\eta = \log\big\{p/(1-p)\big\}$, $u = dl/d\eta = (\nu+y)p - \nu$ and $w = d^2l/d\eta^2 = (\nu + y)p(1-p)$.

Section 5.5

5.15 Round the `faithful` data into an uneven histogram using break points `seq(1.5,5.25,length=61)[-(1:20)*3]`. Estimate the per-second intensity using the uneven histogram and compare the estimate with the ones plotted in Fig. 5.10.

6
Regression with Correlated Responses

When responses are correlated in regression settings, (3.1) and (5.1) need to be modified to incorporate correlation. For the model components to be identifiable from each other, the correlation can not be arbitrary but structured around a limited number of parameters, say γ, and the correlation structure should not be dependent on the covariate x. Of primary interest is the selection of tuning parameters, which now consist of the smoothing parameters in $\lambda J(\eta)$ and the correlation parameters γ.

Commonly used correlation models include random effects and stationary time series. With random effects, the covariance matrix W^{-1} of the responses typically differ from $\sigma^2 I$ by some low-rank matrix update, and one may work with the joint likelihood of the fixed effect $\eta(x)$ and the random effects (§6.2); the variance components are effectively turned into "mean components," the tools developed for independent data are readily applicable, and the asymptotic optimality of cross-validation carries over. For correlation models with W^{-1} "far" from diagonal such as stationary time series models, optimal smoothing is possible via certain extensions of cross-validation (§6.3). Software tools are illustrated using simulated and real-data examples.

We are not aware of a mechanism in which one may characterize the "limiting behavior" of correlation structures, so no results are available concerning the asymptotic convergence of estimates based on correlated data.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_6,

6.1 Models for Correlated Data

Consider observations $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $(\epsilon_1, \ldots, \epsilon_n)^T = \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 W^{-1})$. When W is known, one may estimate $\eta = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}$ via the minimization of

$$(\mathbf{Y} - S\mathbf{d} - R\mathbf{c})^T W (\mathbf{Y} - S\mathbf{d} - R\mathbf{c}) + n\lambda\, \mathbf{c}^T Q \mathbf{c}, \tag{6.1}$$

where the notation is as in (3.61) and (3.62) of §3.5. Let $W^{-1} = C^T C$ be the Cholesky decomposition of W^{-1}, so $W = C^{-1} C^{-T}$. One may obtain a solution of (6.1) by replacing $(\mathbf{Y}, S, R)$ in (3.63) on page 86 with $(\mathbf{Y}_w, S_w, R_w) = C^{-T}(\mathbf{Y}, S, R)$. For $R = Q$ with $q = n$, one may alternatively solve (3.10) using the algorithms of §3.4, but with C^{-T} replacing $W^{1/2}$ in the definitions of the respective Q_w, $\mathbf{c}_w$, S_w and $\mathbf{Y}_w$ in (3.10). Relevant results in §3.2.4 also hold for W not diagonal.

When W involves unknown parameters, say γ, new tools are needed for the estimation of $\eta(x)$ with automatic tuning parameters (λ, γ). The tools are developed for some commonly used models of W, which include random effects for longitudinal or clustered data and time series models for data with serial correlations.

6.1.1 Random Effects

Let $\epsilon_i = a_i + \mathbf{z}_i^T \mathbf{b}$, where $a_i \sim N(0, \sigma^2)$ are independent of each other and of $\mathbf{b} \sim N(\mathbf{0}, \sigma^2 B)$. One has $W^{-1} = I + ZBZ^T$, where $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$.

The term $\mathbf{z}_i^T \mathbf{b}$ contains the random effects as opposed to the fixed effect $\eta(x_i)$. B is typically structured with unknown parameters. The terms of $\sigma^2 W^{-1} = \sigma^2(I + ZBZ^T)$ are also known as variance components.

Example 6.1 (Longitudinal data) Consider longitudinal data $Y_i = \eta(x_i) + b_{s_i} + a_i$, where Y_i is taken from subject $s_i \in \{1, \ldots, p\}$ with covariate x_i, where $b_s \sim N(0, \sigma_s^2)$ is the subject random effect, independent of the measurement error a_i and of each other. $B = \gamma I_p$ with $\gamma = \sigma_s^2/\sigma^2$ to be specified. □

Example 6.2 (Clustered data) Consider clustered data $Y_i = \eta(x_i) + b_{c_i} + a_i$, such as those from multi-center studies, where Y_i is taken from cluster $c_i \in \{1, \ldots, p\}$ with covariate x_i. The intra-cluster correlation within cluster c is seen to be $\sigma_c^2/(\sigma^2 + \sigma_c^2)$, $c = 1, \ldots, p$. $B = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ with p unknown parameters $\gamma_c = \sigma_c^2/\sigma^2$, as there is no reason to assume a common intra-cluster correlation. □

6.1.2 Stationary Time Series

The spectral density of a stationary time series can be approximated arbitrarily closely by that of an autoregressive-moving-average (ARMA)

process. Consider a stationary and invertible ARMA process of order (p, q) (ARMA(p,q)) for ϵ_i,

$$(1 - \varphi_1 B - \cdots - \varphi_p B^p)\epsilon_i = (1 - \theta_1 B - \cdots - \theta_q B^q)a_i, \tag{6.2}$$

where $p, q \geq 0$, $a_i \sim N(0, \sigma^2)$, $i = \ldots, -2, -1, 0, 1, 2, \ldots$ are independent, and B is the backward shift operator, $B\epsilon_i = \epsilon_{i-1}$ and $Ba_i = a_{i-1}$; the polynomials $\varphi(x) = 1 - \varphi_1 x - \cdots - \varphi_p x^p$ and $\theta(x) = 1 - \theta_1 x - \cdots - \theta_q x^q$ have all of their roots outside of the unit circle, and $\varphi(x)$ and $\theta(x)$ share no common root.

W^{-1} are generally not available in simple forms of φ_j's and θ_k's, but the (j, k)th entry of $\sigma^2 W^{-1}$ can be expressed as $\int_{-0.5}^{0.5} e^{\mathbf{i}2\pi(j-k)\omega} p(\omega) d\omega$, for $\mathbf{i} = \sqrt{-1}$, where $p(\omega) = \sigma^2 |\theta(e^{-\mathbf{i}2\pi\omega})|^2 / |\varphi(e^{-\mathbf{i}2\pi\omega})|^2$ is the power spectrum of the process.

Example 6.3 (AR(1) model) For $p = 1$ and $q = 0$, $\epsilon_i = \gamma\epsilon_{i-1} + a_i$, where $|\gamma| < 1$. One has

$$W^{-1} = \frac{1}{1 - \gamma^2} \begin{pmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{n-1} \\ \gamma & 1 & \gamma & \cdots & \gamma^{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma^{n-1} & \gamma^{n-2} & \gamma^{n-3} & \cdots & 1 \end{pmatrix}. \quad \square$$

Example 6.4 (MA(1) model) For $p = 0$ and $q = 1$, $\epsilon_i = a_i - \gamma a_{i-1}$, where $|\gamma| < 1$. One has

$$W^{-1} = \begin{pmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ -\gamma & 1 + \gamma^2 & -\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}. \quad \square$$

The W^{-1} matrices of AR(1) and MA(1) models are inverses of each other; see Problem 6.1.

6.2 Mixed-Effect Models and Penalized Joint Likelihood

Using the random-effect model of §6.1.1 for correlated data but with a slight change in notation, consider a mixed-effect model

$$Y_i = \eta(x_i) + \mathbf{z}_i^T \mathbf{b} + \epsilon_i, \tag{6.3}$$

$i = 1, \ldots, n$, where $\mathbf{b} \sim N(\mathbf{0}, \sigma^2 B)$, $\epsilon_i \sim N(0, \sigma^2)$, independent of $\mathbf{b}$ and of each other. We shall estimate the fixed effect $\eta(x)$ and the random effects $\mathbf{z}^T\mathbf{b}$ jointly via the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i) - \mathbf{z}_i^T\mathbf{b}\big)^2 + \frac{1}{n}\mathbf{b}^T\Sigma\mathbf{b} + \lambda J(\eta), \tag{6.4}$$

where $\Sigma = B^{-1} > 0$; note that we are working with the log joint likelihood of $(\eta, \mathbf{b})$ instead of the log marginal likelihood of η appearing in (6.1). When the random-effects $\mathbf{z}^T\mathbf{b}$ are not interpretable, such as in Example 6.2, the estimation via (6.4) turns variance components into "mean components."

The tuning parameters for (6.4), which include the smoothing parameters in $\lambda J(\eta)$ and the correlation parameters in Σ, can be selected using the methods of §3.2. The Bayes model is briefly noted, from which the Bayesian confidence intervals can be readily calculated. The selection of tuning parameters via generalized cross-validation is asymptotically optimal, and its empirical performance is assessed through simple simulations. The square error projection of §3.8 can be computed with the random effects $\mathbf{z}^T\mathbf{b}$ treated as an offset.

Mixed-effect models can also be used in non-Gaussian regression. Software tools are illustrated using simulated examples.

6.2.1 Smoothing Matrices

Plugging $\eta = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}$ into (6.4), one minimizes

$$(\mathbf{Y} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b})^T(\mathbf{Y} - S\mathbf{d} - R\mathbf{c} - Z\mathbf{b}) + \mathbf{b}^T\Sigma\mathbf{b} + n\lambda\,\mathbf{c}^T Q\mathbf{c} \tag{6.5}$$

with respect to $(\mathbf{d}, \mathbf{c}, \mathbf{b})$, where S, R, and Q are as in (3.62) on page 85 and $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T$ is $n \times p$. Write $\breve{R} = (S, R)$, $\breve{Q} = \mathrm{diag}(O, Q)$, and $\breve{\mathbf{c}}^T = (\mathbf{d}^T, \mathbf{c}^T)$. Differentiating (6.5) with respect to $\breve{\mathbf{c}}$ and $\mathbf{b}$ and setting the derivatives to 0, one has

$$\begin{pmatrix} \breve{R}^T\breve{R} + n\lambda\breve{Q} & \breve{R}^T Z \\ Z^T\breve{R} & Z^T Z + \Sigma \end{pmatrix}\begin{pmatrix} \breve{\mathbf{c}} \\ \mathbf{b} \end{pmatrix} = \begin{pmatrix} \breve{R}^T\mathbf{Y} \\ Z^T\mathbf{Y} \end{pmatrix}. \tag{6.6}$$

$\hat{\mathbf{Y}} = \breve{R}\breve{\mathbf{c}} + Z\mathbf{b} = A(\lambda, \gamma)\mathbf{Y}$ with the smoothing matrix

$$A(\lambda, \gamma) = (\breve{R}, Z)\begin{pmatrix} \breve{R}^T\breve{R} + n\lambda\breve{Q} & \breve{R}^T Z \\ Z^T\breve{R} & Z^T Z + \Sigma \end{pmatrix}^{+}\begin{pmatrix} \breve{R}^T \\ Z^T \end{pmatrix},$$

where γ denotes the correlation parameters in Σ and M^+ denotes the Moore-Penrose inverse of M. Using Problem 6.2(b), some algebra yields

$$A(\lambda, \gamma) = \tilde{A}(\lambda) + \big(I - \tilde{A}(\lambda)\big)Z\big(Z^T\big(I - \tilde{A}(\lambda)\big)Z + \Sigma\big)^{-1}Z^T\big(I - \tilde{A}(\lambda)\big), \tag{6.7}$$

where $\tilde{A}(\lambda) = \breve{R}(\breve{R}^T\breve{R} + n\lambda\breve{Q})^+\breve{R}^T$ is the smoothing matrix in the absence of random effects.

The scores $U(\lambda)$ of (3.14), $V(\lambda)$ of (3.23), and $M(\lambda)$ of (3.30) are in terms of the smoothing matrix $A(\lambda)$ given in (3.8) and (3.69). Substituting $A(\lambda, \gamma)$ in the place of $A(\lambda)$, one may use $U(\lambda, \gamma)$, $V(\lambda, \gamma)$, and $M(\lambda, \gamma)$ for the joint selection of (λ, γ).

6.2.2 Bayes Model

Under the Bayes model of §§2.5 and 3.5.2, which itself can be perceived as a mixed-effect model with the fixed effect diffusing in $\mathcal{N}_J$ and the random effects having proper priors, the random effects $\mathbf{z}^T\mathbf{b}$ simply augment terms with proper priors, just like the parametric terms in the partial spline models of §4.1 augment terms with diffuse priors.

Following §3.5.2, write $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2$ with independent components, where $\boldsymbol{\eta}_0 = S\mathbf{d}$ with $\mathbf{d}$ diffuse, $E[\boldsymbol{\eta}_1] = E[\boldsymbol{\eta}_2] = \mathbf{0}$, $E\big[\boldsymbol{\eta}_1\boldsymbol{\eta}_1^T\big] = (\sigma^2/n\lambda)RQ^+R^T$, and $E\big[\boldsymbol{\eta}_2\boldsymbol{\eta}_2^T\big] = \sigma^2 Z\,\Sigma^{-1}Z^T$; i.e., $E[\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2] = \mathbf{0}$ and

$$E\big[(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)^T\big] = \sigma^2\,(R, Z)\begin{pmatrix} Q^+/n\lambda & O \\ O & \Sigma^{-1}\end{pmatrix}\begin{pmatrix} R^T \\ Z^T\end{pmatrix}.$$

Comparing (3.63) with (6.6) but fully spelled out,

$$\begin{pmatrix} S^TS & S^TR & S^TZ \\ R^TS & R^TR + n\lambda Q & R^TZ \\ Z^TS & Z^TR & Z^TZ + \Sigma\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\\ \mathbf{b}\end{pmatrix} = \begin{pmatrix} S^T\mathbf{Y}\\ R^T\mathbf{Y}\\ Z^T\mathbf{Y}\end{pmatrix},$$

it is clear that everything in §3.5.2 remains intact with (R, Z) replacing R and $\mathrm{diag}(n\lambda Q, \Sigma)$ replacing $n\lambda Q$.

6.2.3 Optimality of Generalized Cross-Validation

We now present results parallel to Theorems 3.1 and 3.3 concerning the use of $U(\lambda, \gamma)$ and $V(\lambda, \gamma)$ for the selection of tuning parameters in (6.4). We shall motivate the ideas, discuss the conditions, and list the theorems. The proofs, to be found in Gu and Ma (2005b), are somewhat involved.

First consider the mean square error at the data points,

$$L_1(\lambda, \gamma) = \frac{1}{n}\sum_{i=1}^{n}\big(\hat{Y}_i - \eta(x_i) - \mathbf{z}_i^T\mathbf{b}\big)^2, \tag{6.8}$$

which is a natural loss when the random effects $\mathbf{z}^T\mathbf{b}$ are interpretable, or real. Parallel to (3.15) on page 65, one has

$$U(\lambda,\gamma) - L_1(\lambda,\gamma) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$$
$$= \frac{2}{n}(\boldsymbol{\eta} + Z\mathbf{b})^T\big(I - A(\lambda,\gamma)\big)\boldsymbol{\epsilon} - \frac{2}{n}\big(\boldsymbol{\epsilon}^T A(\lambda,\gamma)\boldsymbol{\epsilon} - \sigma^2 \mathrm{tr}A(\lambda,\gamma)\big); \quad (6.9)$$

see Problem 6.3. To bound $(Z\mathbf{b})^T\big(I - A(\lambda,\gamma)\big)\boldsymbol{\epsilon}$, one needs

Condition 6.2.1 $\Sigma\big(Z^T\big(I - \tilde{A}(\lambda)\big)Z + \Sigma\big)^{-1}\Sigma$ has eigenvalues bounded from above.

The condition holds for Σ with its largest eigenvalue bounded from above, or growing at a rate of up to $\sqrt{n}$ when that of $Z^T\big(I - \tilde{A}(\lambda)\big)Z$ grows at a rate of n. Write $R_1(\lambda,\gamma) = E\big[L_1(\lambda,\gamma)\big]$.

Condition 6.2.2 $nR_1(\lambda,\gamma) \to \infty$ as $n \to \infty$ and $\lambda \to 0$,

This is virtually Condition 3.2.1.

Theorem 6.1 *Under Conditions 6.2.1 and 6.2.2, as $n \to \infty$ and $\lambda \to 0$,*

$$U(\lambda,\gamma) - L_1(\lambda,\gamma) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L_1(\lambda,\gamma)\big).$$

Condition 6.2.3 $\big\{n^{-1}\mathrm{tr}A(\lambda,\gamma)\big\}^2/\big\{n^{-1}\mathrm{tr}A^2(\lambda,\gamma)\big\} \to 0$ as $n \to \infty$ and $\lambda \to 0$.

This is Condition 3.2.2. If $\mathrm{tr}\tilde{A}(\lambda) \asymp \lambda^{-1/r}$ and $\mathrm{tr}\tilde{A}^2(\lambda) \asymp \lambda^{-1/r}$ as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$ (see §4.2.3), then Condition 6.2.3 holds for p up to the order $O\big(\sqrt{n}\big)$; see Gu and Ma (2005b, Lemma 4.2).

Theorem 6.2 *Under Conditions 6.2.1–6.2.3, as $n \to \infty$ and $\lambda \to 0$,*

$$V(\lambda,\gamma) - L_1(\lambda,\gamma) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L_1(\lambda,\gamma)\big).$$

We now turn to the case where the random effects $\mathbf{z}^T\mathbf{b}$ are not interpretable, or latent, for which the loss $L_1(\lambda,\gamma)$ of (6.8) may not make much practical sense. Write $P_Z = Z(Z^TZ)^+Z^T$ and $P_Z^\perp = I - P_Z$. Consider the estimation of $P_Z^\perp\boldsymbol{\eta}$ by $P_Z^\perp\hat{\boldsymbol{\eta}}$, where $\hat{\boldsymbol{\eta}} = \breve{R}\breve{\mathbf{c}}$; the projection ensures the identifiability of the target function. Accounting for the error covariance $\sigma^2(I + Z\,\Sigma^{-1}Z^T)$, one may assess the estimation precision via the loss

$$L_2(\lambda,\gamma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T P_Z^\perp (I + Z\,\Sigma^{-1}Z^T)^{-1} P_Z^\perp (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}).$$

Now $(I + Z\,\Sigma^{-1}Z^T)^{-1} = I - Z(Z^TZ + \Sigma)^{-1}Z^T$ (Problem 6.4), so

$$L_2(\lambda,\gamma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})^T P_Z^\perp (\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}), \quad (6.10)$$

which is independent of Σ. Write $R_2(\lambda,\gamma) = E\big[L_2(\lambda,\gamma)\big]$.

Condition 6.2.4 $R_1(\lambda,\gamma) - R_2(\lambda,\gamma) = o\big(R_1(\lambda,\gamma)\big)$ as $n \to \infty$ and $\lambda \to 0$.

The condition is a mild one for p fixed; see Gu and Ma (2005b, Lemma 4.3). Conditions 6.2.2 and 6.2.4 together imply that $nR_2(\lambda,\gamma) \to \infty$.

Theorem 6.3 *Under Conditions 6.2.1, 6.2.2, and 6.2.4, as* $n \to \infty$ *and* $\lambda \to 0$,

$$U(\lambda,\gamma) - L_2(\lambda,\gamma) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L_2(\lambda,\gamma)\big).$$

If, in addition, Condition 6.2.3 also holds, then

$$V(\lambda,\gamma) - L_2(\lambda,\gamma) - n^{-1}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = o_p\big(L_2(\lambda,\gamma)\big).$$

With a mixture of real and latent random effects, say $Z\mathbf{b} = Z_1\mathbf{b}_1 + Z_2\mathbf{b}_2$, where $Z = (Z_1, Z_2)$ and $\mathbf{b}^T = (\mathbf{b}_1^T, \mathbf{b}_2^T)$ for $\mathbf{b}_1$ and $\mathbf{b}_2$ independent, one may use the loss

$$L_3(\lambda,\gamma) = \frac{1}{n}(\hat{\boldsymbol{\eta}} + Z_1\hat{\mathbf{b}}_1 - \boldsymbol{\eta} - Z_1\mathbf{b}_1)^T P_{Z_2}^{\perp}(\hat{\boldsymbol{\eta}} + Z_1\hat{\mathbf{b}}_1 - \boldsymbol{\eta} - Z_1\mathbf{b}_1). \quad (6.11)$$

Theorem 6.3 can be replicated for $L_3(\lambda,\gamma)$, but with $R_3(\lambda,\gamma) = E\big[L_3(\lambda,\gamma)\big]$ replacing $R_2(\lambda,\gamma)$ in Condition 6.2.4.

In summary, generalized cross-validation delivers asymptotically optimal smoothing for the estimation of mixed-effect models via (6.4), regardless of the nature of the random effects $\mathbf{z}^T\mathbf{b}$. The dimension of real random effects may grow at a rate of up to $\sqrt{n}$, but that of latent random effects should be fixed. Similar to Theorems 3.1 and 3.3, the available results only provide poor man's justification, as the theorems only concern deterministic tuning parameters.

6.2.4 Empirical Performance

Samples were drawn from $Y_i = \eta(x_i) + b_{s_i} + \epsilon_i$, $i = 1,\ldots,100$, where $\eta(x) = 1 + 3\sin(2\pi x - \pi)$, $x_i \sim U(0,1)$, $\epsilon_i \sim N(0, 0.5^2)$, $b_s \sim N(0, 0.5^2)$, and $s_i \in \{1,\ldots,5\}$, 20 each. With $B = \gamma I_5$, cubic spline estimates were calculated that minimized $L_1(\lambda,\gamma)$ of (6.8) at (λ_o,γ_o) and $V(\lambda,\gamma)$ at (λ_v,γ_v) with $\alpha = 1, 1.4$. The results from one hundred replicates are summarized in Fig. 6.1, with the relative efficacy $L_1(\lambda_o,\gamma_o)/L_1(\lambda_v,\gamma_v)$ in the boxplots in the left half of the left frame and $L_1(\lambda_v,\gamma_v)$ for $\alpha = 1$ versus that for $\alpha = 1.4$ in the center frame.

Perceiving the same data as clustered, estimates were also calculated with $B = \text{diag}(\gamma_1 \ldots \gamma_5)$ that minimized $L_2(\lambda,\gamma)$ of (6.10) and $V(\lambda,\gamma)$. Respective results are also summarized in Fig. 6.1, in the right half of the left frame and in the right frame.

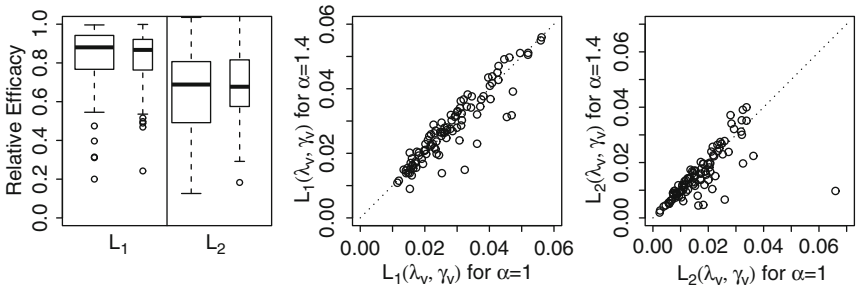


FIGURE 6.1. Effectiveness of $V(\lambda,\gamma)$ in mixed-effect simulations. *Left*: Relative efficacy $L(\lambda_o,\gamma_o)/L(\lambda_v,\gamma_v)$, for $\alpha=1$ (*wider boxes*) and $\alpha=1.4$ (*thinner boxes*). *Center*: $L_1(\lambda_v,\gamma_v)$ for $\alpha=1$ versus that for $\alpha=1.4$. *Right*: $L_2(\lambda_v,\gamma_v)$ for $\alpha=1$ versus that for $\alpha=1.4$.

6.2.5 Non-Gaussian Regression

The random effects $\mathbf{z}^T\mathbf{b}$ can also be used in non-Gaussian regression to model correlated data. Replacing $\eta(x)$ by $\eta(x)+\mathbf{z}^T\mathbf{b}$ in the families of Chap. 5, one may estimate $\eta(x)$ and $\mathbf{b}$ via the minimization of

$$\frac{1}{n}\sum_{i=1}^{n} l_i\big(\eta(x_i)+\mathbf{z}_i^T\mathbf{b};Y_i\big)+\frac{1}{2n}\mathbf{b}^T\Sigma\mathbf{b}+\frac{\lambda}{2}J(\eta), \tag{6.12}$$

where $l_i(\zeta;y)$ is the minus log likelihood associated with Y_i. For the minimization of (6.12), one may iterate on weighted versions of (6.4), and the tuning parameters can be selected through performance-oriented iteration driven by $U_w(\lambda,\gamma)/V_w(\lambda,\gamma)$ or via direct cross-validation. Approximate Bayesian confidence intervals can be calculated based on the quadratic approximation of (6.12) at the converged fit, and the Kullback-Leibler projection of §5.3.2 can be computed with $\mathbf{z}^T\mathbf{b}$ treated as an offset.

6.2.6 R Package `gss`: Optional Argument `random`

Mixed-effect models for Gaussian and non-Gaussian regression can be fitted using `ssanova`, `gssanova`, or `gssanova1` with an additional argument `random`, which can be a formula or a list.

The following sequence generates some synthetic longitudinal data and fits a model with $B=\gamma I_5$ as stipulated in Example 6.1:

```
id <- rep(1:5,rep(20,5))
b <- rnorm(5)/2
eps <- rnorm(100)/2+b[id[1:100]]
x <- runif(100)
y <- 1+3*sin(2*pi*x-pi)+eps
id <- as.factor(id)
fit.long <- ssanova(y~x,random=~1|id)
```

If the data are to be perceived as clustered, a model can be fitted with $B = \text{diag}(\gamma_1, \ldots, \gamma_5)$ as stipulated in Example 6.2:

```
fit.cluster <- ssanova(y~x,random=~id|id)
```

More generally, one may specify `random=~id1|id2`, with the levels of `id2` possibly "refining" those of `id1`, or the levels of `id1` possibly "collapsed" from those of `id2`, say

```
id1 <- rep(1:3,rep(4,3)); id2 <- rep(1:6,rep(2,6))
```

but not

```
id1 <- rep(1:3,rep(4,3)); id2 <- rep(1:6,2)
```

Each level of `id2` corresponds to a column in Z, and each level of `id1` is associated with a correlation parameter γ.

For general mixed-effect models, one may specify (Z, Σ) via

```
random=list(z=...,sigma=...,init=...)
```

where `z` contains the Z matrix, `sigma` gives Σ through

```
sigma$fun(gamma,sigma$env)
```

with γ in `gamma` and constants in `sigma$env`, and `init` provides initial values of γ; γ should be properly parameterized to be free of constraint.

6.3 Penalized Likelihood with Correlated Data

Working with (6.1) for W involving unknown parameters, one needs to select both the smoothing parameters and the correlation parameters. $M(\lambda)$ of §3.2.3 can be readily extended under the Bayes model, but effective counterparts of $U(\lambda)$ and $V(\lambda)$ take a few turns to derive.

We first discuss the Bayes model, then introduce extensions of $U(\lambda)$ and $V(\lambda)$ for tuning parameter selection. The asymptotic optimality of the selection methods are outlined, followed by the assessment of their empirical performances via simulations. Software tools are illustrated using simulated examples.

To cut down on clutter, the dependence of quantities on λ and γ are often omitted in the notation, except in the statements of conditions and theorems.

6.3.1 Bayes Model

Parallel to (3.63), the minimizer of (6.1) satisfies

$$\begin{pmatrix} S^T W S & S^T W R \\ R^T W S & R^T W R + n\lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T W \mathbf{Y} \\ R^T W \mathbf{Y} \end{pmatrix}. \tag{6.13}$$

Most of the calculations in §3.5.2 remain valid but with a redefined $M = RQ^+R^T + n\lambda W^{-1}$. Specifically, (3.65)–(3.68) hold verbatim for the modified M, with $\mathbf{d}$ and $\mathbf{c}$ given in (3.66) solving (6.13), and (3.69) becomes

$$A = I - n\lambda W^{-1}(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1}), \tag{6.14}$$

where $\hat{\mathbf{Y}} = S\mathbf{d} + R\mathbf{c} = A\mathbf{Y}$; note that WA is symmetric here, not A. For $W^{-1} = C^TC$, one has $\hat{\mathbf{Y}}_w = A_w\mathbf{Y}_w$, where $\mathbf{Y}_w = C^{-T}\mathbf{Y}$, $\hat{\mathbf{Y}}_w = C^{-T}\hat{\mathbf{Y}}$,

$$A_w = I - n\lambda C(M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})C^T; \tag{6.15}$$

see also Problem 6.5. $A_w = C^{-T}AC^T$ and $I - A_w = C^{-T}(I - A)C^T$.

With the modified M, (3.70) on page 87 still holds for REML, but now

$$(F_2^TMF_2)^{-1} = (n\lambda)^{-1}F_2^TW(I - A)F_2,$$
$$F_2(F_2^TMF_2)^{-1}F_2^T = (n\lambda)^{-1}W(I - A).$$

The numerator of (3.70) is thus

$$n^{-1}(n\lambda)^{-1}\mathbf{Y}^TW(I - A)\mathbf{Y} = n^{-1}(n\lambda)^{-1}\mathbf{Y}_w^T(I - A_w)\mathbf{Y}_w,$$

where $W(I - A) = C^{-1}(I - A_w)C^{-T}$. Using Problem 3.17,

$$\begin{aligned}\left|(n\lambda)^{-1}F_2^TMF_2\right| &= \left|(n\lambda)^{-1}F_2^TRQ^+R^TF_2 + F_2^TW^{-1}F_2\right| \\ &= \left|F_2^TW^{-1}F_2\right|\left|I + (n\lambda)^{-1}Q^+R^TF_2(F_2^TW^{-1}F_2)^{-1}F_2^TR\right| \\ &= \left|F_2^TW^{-1}F_2\right|\left|I + (n\lambda)^{-1}Q^+R_w^TF_w(F_w^TF_w)^{-1}F_w^TR_w\right|,\end{aligned}$$

where $R_w = C^{-T}R$ and $F_w = CF_2$. Let $S_w = (\tilde{F}_1, \tilde{F}_2)\binom{\tilde{R}}{O}$ be the QR-decomposition of $S_w = C^{-T}S$. Since $S_w^T\tilde{F}_2 = O = S^TF_2 = S_w^TF_w$, F_w and $\tilde{F}_2$ have the same column space, thus $F_w(F_w^TF_w)^{-1}F_w^T = \tilde{F}_2\tilde{F}_2^T$. The denominator of (3.70) is then seen to be

$$(n\lambda)^{-1}\left(\left|F_2^TW^{-1}F_2\right|\left|I - A_w\right|_+\right)^{1/(n-m)};$$

compare with (3.75) on page 89. Putting things together, one has

$$\tilde{M}(\lambda, \gamma) = \frac{n^{-1}\mathbf{Y}_w^T(I - A_w)\mathbf{Y}_w}{\left|I - A_w\right|_+^{1/(n-m)}}\frac{1}{\left|F_2^TW^{-1}F_2\right|^{1/(n-m)}}. \tag{6.16}$$

When W is known, $|F_2^TW^{-1}F_2|$ is a constant, in which case (6.16) is equivalent to (3.35) on page 72; this formally validates our earlier use of (3.35) for weighted data.

6.3.2 Extension of Cross-Validation

We now extend $U(\lambda)$ and $V(\lambda)$ of §3.2 for the joint selection of (λ, γ); playing a central role in the derivation is the minus log likelihood,

$$\frac{1}{2\sigma^2}(\mathbf{Y}-\boldsymbol{\eta})^T W(\mathbf{Y}-\boldsymbol{\eta}) - \frac{1}{2}\log|W| + \frac{n}{2}\log\sigma^2 + \frac{n}{2}\log 2\pi. \tag{6.17}$$

For $\sigma^2 W^{-1} = \sigma^2 C^T C$ known, one may select λ via the minimization of

$$U_w(\lambda) = \frac{1}{n}\mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w + 2\frac{\sigma^2}{n}\mathrm{tr}A_w, \tag{6.18}$$

where $\mathbf{Y}_w = C^{-T}\mathbf{Y}$ and A_w is as in (6.15); this is simply (3.33) but with W non-diagonal, and Theorem 3.5 still holds for $L_w(\lambda) = (\boldsymbol{\eta}_\lambda - \boldsymbol{\eta})^T W(\boldsymbol{\eta}_\lambda - \boldsymbol{\eta})$, where $\boldsymbol{\eta}_\lambda^T = \big(\eta_\lambda(x_i), \ldots, \eta_\lambda(x_n)\big)$. It is noted that $(n/2\sigma^2)U_w(\lambda)$ consists of the minus log likelihood plus a penalty term $\mathrm{tr}A_w$, but with terms in (6.17) that do not depend on λ dropped; note that

$$(\mathbf{Y}-\boldsymbol{\eta})^T W(\mathbf{Y}-\boldsymbol{\eta}) = \mathbf{Y}^T(I-A)^T W(I-A)\mathbf{Y} = \mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w.$$

With σ^2 known but $W = C^{-1}C^{-T}$ dependent on γ, one may add back the term $-(1/2)\log|W|$ in (6.17) and scale properly, yielding, for $\alpha = 1$,

$$\tilde{U}(\lambda,\gamma) = \frac{1}{n\sigma^2}\mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w - \frac{1}{n}\log|W| + \alpha\frac{2}{n}\mathrm{tr}A_w. \tag{6.19}$$

For σ^2 unknown, one may minimize (6.17) with respect to σ^2, plug into (6.17) the minimizer $\hat{\sigma}^2 = n^{-1}\big\{\mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w\big\}$ to obtain the profile likelihood, and then add the penalty term $\mathrm{tr}A_w$, properly scaled, to the profile likelihood. This yields

$$\tilde{V}(\lambda,\gamma) = \log\big\{n^{-1}\mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w\big\} - \frac{1}{n}\log|W| + \alpha\frac{2}{n}\mathrm{tr}A_w, \tag{6.20}$$

where terms free of (λ, γ) are dropped. For $W = I$ and $\mu = \mathrm{tr}A/n = o(1)$, (6.20) reduces to

$$\begin{aligned}&\log\big\{(n^{-1}\mathbf{Y}^T(I-A)^2\mathbf{Y})e^{2\mu}\big\}\\ &\quad = \log\left\{\frac{n^{-1}\mathbf{Y}^T(I-A)^2\mathbf{Y}}{(1-\mu)^2}\big(1+O(\mu^2)\big)\right\} = \log\big\{V(\lambda)\big(1+O(\mu^2)\big)\big\}.\end{aligned}$$

An obvious drawback of (6.20) is that the third term is bounded from above since $I - A_w \geq 0$, while the first term will go to $-\infty$ as A_w approaches I, favoring interpolation. To guard against this, one may use

$$\tilde{V}_*(\lambda,\gamma) = \log\big\{n^{-1}\mathbf{Y}_w^T(I-A_w)^2\mathbf{Y}_w\big\} - \frac{1}{n}\log|W| + \alpha\frac{2\,\mathrm{tr}A_w}{n-\mathrm{tr}A_w}; \tag{6.21}$$

when $\mu = \mathrm{tr}A_w/n = o(1)$, $\tilde{V}_*(\lambda,\gamma) - \tilde{V}(\lambda,\gamma) = 2\mu^2\big(1+o(1)\big)$.

6.3.3 Optimality of Cross-Validation

We now outline results parallel to Theorems 3.1 and 3.3 concerning the use of $\tilde{U}(\lambda,\gamma)$, $\tilde{V}(\lambda,\gamma)$, and $\tilde{V}_*(\lambda,\gamma)$ for the selection of (λ,γ) in (6.1). Technical details are to be found in Han and Gu (2008).

Let $f_{\boldsymbol{\eta},W}$ be the density of $N\big(\boldsymbol{\eta},\sigma^2W^{-1}\big)$. We use as loss the Kullback-Leibler distance of $f=f_{\boldsymbol{\eta},W}$ from the true density $f_0=f_{\boldsymbol{\eta}_0,W_0}$,

$$\frac{n}{2}L(\lambda,\gamma)=\mathrm{KL}(f_0,f)=\frac{1}{2\sigma^2}(\boldsymbol{\eta}-\boldsymbol{\eta}_0)^TW(\boldsymbol{\eta}-\boldsymbol{\eta}_0)+\frac{1}{2}\mathrm{tr}\big(WW_0^{-1}-I\big)-\frac{1}{2}\log\big|WW_0^{-1}\big|, \quad (6.22)$$

where the estimate f depends on λ and γ; see Problem 6.6. Parallel to (3.15), it is straightforward to verify that (Problem 6.7)

$$\begin{aligned}\tilde{U}(\lambda,\gamma)-L(\lambda,\gamma)-\frac{1}{n\sigma^2}\boldsymbol{\epsilon}^TW_0\boldsymbol{\epsilon}+\frac{1}{n}\log|W_0|&\\ =\frac{2}{n\sigma^2}\boldsymbol{\eta}_0^T(I-A)^TW\boldsymbol{\epsilon}-\frac{2}{n}\Big\{\frac{1}{\sigma^2}\boldsymbol{\epsilon}^TA^TW\boldsymbol{\epsilon}-\mathrm{tr}A_w\Big\}&\\ +\frac{1}{n}\Big\{\frac{1}{\sigma^2}\boldsymbol{\epsilon}^T(W-W_0)\boldsymbol{\epsilon}-\mathrm{tr}\big(WW_0^{-1}-I\big)\Big\}.&\quad (6.23)\end{aligned}$$

One can bound the terms on the right-hand side of (6.23) under regularity conditions. Write $R(\lambda,\gamma)=R\big[L(\lambda,\gamma)\big]$.

Condition 6.3.1 $R(\lambda,\gamma)\to 0$ and $nR(\lambda,\gamma)\to\infty$ as $\lambda\to 0$, $n\lambda^{1/r}\to\infty$, and $W_\gamma\to W_0$.

Condition 6.3.1 assures that the estimates are risk consistent, but concedes that the typical parametric convergence rates of $O(n^{-1})$ are not achievable. By $W_\gamma\to W_0$ we mean for W_γ in a shrinking neighborhood of $W_0=W_{\gamma_0}$, typically characterized by $\gamma\to\gamma_0$ at certain rates.

Condition 6.3.2 $\pm\big(W_\gamma^{1/2}W_0^{-1}W_\gamma^{1/2}-I\big)\le\rho_\gamma I$ for some positive $\rho_\gamma=O\big(R^{1/2}(\lambda,\gamma)\big)$ as $\lambda\to 0$, $n\lambda^{1/r}\to\infty$, and $W_\gamma\to W_0$.

Condition 6.3.2 requires W_γ to converge to W_0 at a certain rate so that the largest absolute eigenvalue of $W_\gamma W_0^{-1}-I$ is of the order $O\big(R^{1/2}(\lambda,\gamma)\big)$.

Condition 6.3.3 $\big\{n^{-1}\mathrm{tr}A_w(\lambda,\gamma)\big\}^2/\big\{n^{-1}\mathrm{tr}A_w^2(\lambda,\gamma)\big\}\to 0$ as $\lambda\to 0$ and $n\lambda^{1/r}\to\infty$.

Condition 6.3.3 holds in settings where $\mathrm{tr}A_w\asymp\mathrm{tr}A_w^2=o(n)$, and it implies that $\mu=n^{-1}\mathrm{tr}A_w\to 0$ as $n^{-1}\mathrm{tr}A_w^2\le n^{-1}\mathrm{tr}A_w\le 1$.

Theorem 6.4 *Under Conditions 6.3.1–6.3.3, as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$, *one has*

$$\tilde{U}(\lambda,\gamma) - L(\lambda,\gamma) - \frac{1}{n\sigma^2}\boldsymbol{\epsilon}^T W_0 \boldsymbol{\epsilon} + \frac{1}{n}\log|W_0| = o_p\big(L(\lambda,\gamma)\big).$$

To establish similar results for $\tilde{V}(\lambda,\gamma)$ and $\tilde{V}_*(\lambda,\gamma)$, one needs an additional condition.

Condition 6.3.4 $n^{-1}\mathrm{tr}\big(W_\gamma W_0^{-1} - I\big) = o\big(R^{1/2}(\lambda,\gamma)\big)$ as $\lambda \to 0$, $n\lambda^{1/r} \to \infty$, and $W_\gamma \to W_0$.

Note that Condition 6.3.2 only guarantees that $n^{-1}\mathrm{tr}\big(W_\gamma W_0^{-1} - I\big) = O\big(R^{1/2}(\lambda,\gamma)\big)$.

Theorem 6.5 *Under Conditions 6.3.1–6.3.4, as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$, *one has*

$$\tilde{V}(\lambda,\gamma) - L(\lambda,\gamma) - K = o_p\big(L(\lambda,\gamma)\big),$$
$$\tilde{V}_*(\lambda,\gamma) - L(\lambda,\gamma) - K = o_p\big(L(\lambda,\gamma)\big),$$

with $K = (n\sigma^2)^{-1}\boldsymbol{\epsilon}^T W_0 \boldsymbol{\epsilon} - n^{-1}\log|W_0| + \log\sigma^2 - 1$ *independent of* (λ,γ).

The proofs of the theorems under the conditions are straightforward albeit tedious, but the verifications of the conditions are much more involved; the limit process is delicate here. Some key lemmas used in the verifications of the conditions assume $c_1 I \le W^{-1} \le c_2 I$ for some $0 < c_1 < c_2 < \infty$, where a healthy lower bound seems to be essential for the stable empirical performance of $\tilde{V}_*(\lambda,\gamma)$; see §6.3.4 below.

For the ARMA(p,q) process of §6.1.2, Conditions 6.3.1–6.3.4 were verified in Han and Gu (2008) for γ over a compact set Γ; in Examples 6.3 and 6.4, $\Gamma = [-\bar{\gamma},\bar{\gamma}]$ for some $\bar{\gamma} < 1$.

For the longitudinal data of Example 6.1, Conditions 6.3.1–6.3.4 were verified when the number of observations from each subject is bounded from above; it is necessary that $p \asymp n$, invalidating the theory of §6.2.3 where p is allowed to grow but at a rate only up to $\sqrt{n}$.

The Kullback-Leibler loss of (6.22) involves W, so a γ that delivers a small $L(\lambda,\gamma)$ should be a good estimate of the true correlation parameter γ_0. In fact, the minimizers of $L(\lambda,\gamma)$, $\tilde{U}(\lambda,\gamma)$, $\tilde{V}(\lambda,\gamma)$, and $\tilde{V}_*(\lambda,\gamma)$ for fixed λ are $\sqrt{n}$-consistent under mild conditions, as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$.

Once again, the theory provides a poor man's justification for the practical use of $\tilde{U}(\lambda,\gamma)$, $\tilde{V}(\lambda,\gamma)$, and $\tilde{V}_*(\lambda,\gamma)$, as the theorems concern only deterministic (λ,γ).

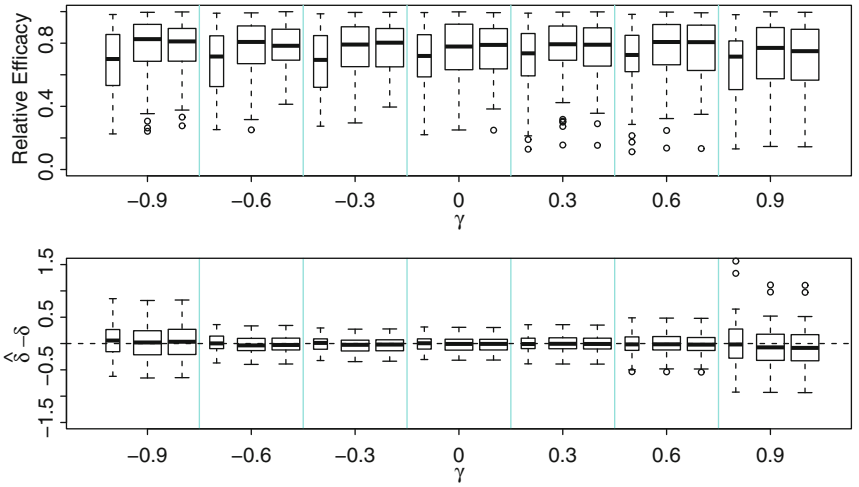


FIGURE 6.2. Effectiveness of $\tilde{V}_*(\lambda,\gamma)$ and $\tilde{M}(\lambda,\gamma)$ in AR(1) simulations. *Top*: Relative efficacy of $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order. *Bottom*: Estimation precision of $\delta = \log\{(1+\gamma)/(1-\gamma)\}$ using $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order.

6.3.4 Empirical Performance

We now assess the empirical performances of $\tilde{V}_*(\lambda,\gamma)$ and $\tilde{M}(\lambda,\gamma)$ via simulation studies. $U(\lambda,\gamma)$ and $V(\lambda,\gamma)$ are not as useful in practice, with the former assuming a known σ^2 and the latter having a global minimum at $\lambda = 0$.

Data were generated from $Y_i = \eta(x_i) + \epsilon_i$, $i = 1,\ldots,n$, where $\eta(x) = 1 + 3\sin(2\pi x - \pi)$, $x_i \sim U(0,1)$, and $\boldsymbol{\epsilon} \sim N\big(\mathbf{0}, 0.5^2 W^{-1}\big)$. Three sets of simulations were conducted, with the AR(1) model of Example 6.3, the MA(1) model of Example 6.4, and the longitudinal data of Example 6.1.

For the AR(1) and MA(1) simulations, samples of size $n = 200$ were drawn with $\gamma = 0, \pm 0.3, \pm 0.6, \pm 0.9$, one hundred replicates each. For each replicate, cubic spline estimates were calculated that minimized $L(\lambda,\gamma)$ of (6.22) at (λ_o,γ_o), $\tilde{M}(\lambda,\gamma)$ of (6.16) at (λ_m,γ_m), and $\tilde{V}_*(\lambda,\gamma)$ of (6.21) at (λ_v,γ_v) for $\alpha = 1, 1.4$. The relative efficacy $L(\lambda_o,\gamma_o)/L(\lambda_m,\gamma_m)$ and $L(\lambda_o,\gamma_o)/\,L(\lambda_v,\gamma_v)$ for the AR(1) simulations are summarized in the top frame of Fig. 6.2, and the estimation accuracy of γ, on the scale of $\delta = \log\{(1+\gamma)/\,(1-\gamma)\}$, is summarized in the bottom frame. Parallel results from the MA(1) simulations are shown in Fig. 6.3.

For the longitudinal simulations, $n = 200$ points were taken from 40 individuals, 5 each, with $W^{-1} = I + \gamma\,\mathrm{diag}(\mathbf{1}_5\mathbf{1}_5^T,\ldots,\mathbf{1}_5\mathbf{1}_5^T)$. Data were drawn with $\gamma = 0, 0.5, 1$, one hundred replicates each. The relative efficacy of $\tilde{M}(\lambda,\gamma)$ and $\tilde{V}_*(\lambda,\gamma)$ are summarized in Fig. 6.4, along with the estimation accuracy of γ on the scale of $\delta = \gamma/(1+\gamma)$.

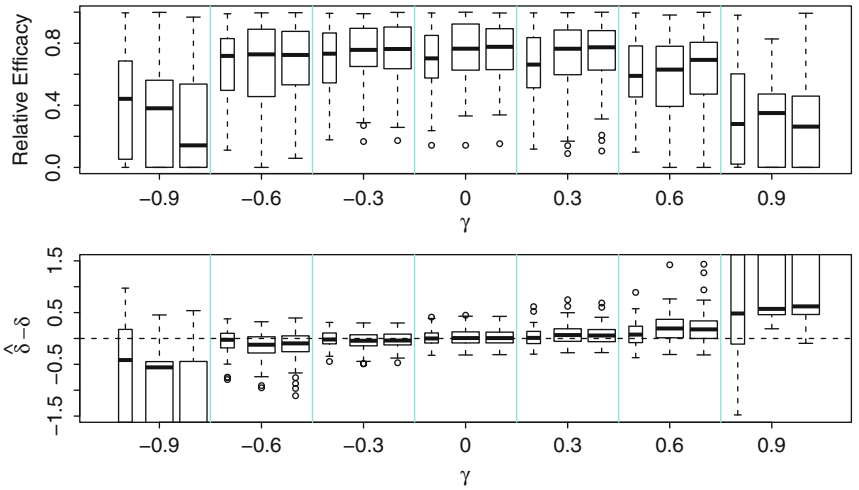


FIGURE 6.3. Effectiveness of $\tilde{V}_*(\lambda,\gamma)$ and $\tilde{M}(\lambda,\gamma)$ in MA(1) simulations. *Top*: Relative efficacy of $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order. *Bottom*: Estimation precision of $\delta = \log\{(1+\gamma)/(1-\gamma)\}$ using $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order.

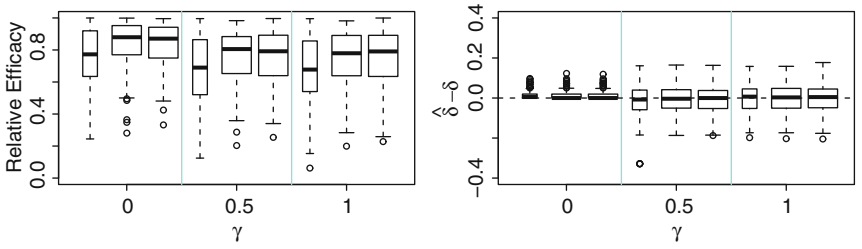


FIGURE 6.4. Effectiveness of $\tilde{V}_*(\lambda,\gamma)$ and $\tilde{M}(\lambda,\gamma)$ in longitudinal model simulations. *Left*: Relative efficacy of $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order. *Right*: Estimation precision of $\delta = \gamma/(1+\gamma)$ using $\tilde{M}(\lambda,\gamma)$ (*thinner boxes*) and $\tilde{V}_*(\lambda,\gamma)$ with $\alpha = 1, 1.4$, in order.

The results show that the methods do work in general, with $\tilde{V}_*(\lambda,\gamma)$ outperforming $\tilde{M}(\lambda,\gamma)$. The methods however demonstrated performance degradation in the MA(1) simulations with $\gamma = \pm 0.6$ and flatly broke down with $\gamma = \pm 0.9$.

In search for an explanation for the MA(1) results, we observe that $(1/4)I \le \big(1+|\gamma|\big)^{-2} I \le W^{-1}$ for the AR(1) model (Problem 6.1), $I \le W^{-1}$ for the longitudinal data, and $\big(1-|\gamma|\big)^2 I \le W^{-1}$ for the MA(1) model. For $|\gamma|$ close to 1, the W^{-1} matrix in the MA(1) model flirts with singularity.

For $W = I$, $\tilde{V}_*(\lambda,\gamma)$ of (6.21) reduces to

$$V_*(\lambda) = \log\left\{n^{-1}\mathbf{Y}^T(I-A)^2\mathbf{Y}\right\} + \alpha\frac{2\,\mathrm{tr}A}{n-\mathrm{tr}A}, \tag{6.24}$$

which is different from $V(\lambda)$ of (3.27). To compare the two cross-validation scores for smoothing parameter selection, samples were drawn from $Y_i = \eta(x_i)+\epsilon_i$, $i = 1, \ldots, 100$, where $\eta(x) = 1+3\sin(2\pi x-\pi)$, $x_i \sim U(0,1)$, and $\epsilon_i \sim N(0,1)$. For each of the one hundred replicates generated, estimates were calculated that minimized $V_*(\lambda)$ of (6.24) at λ_v^* and $V(\lambda)$ of (3.27) at λ_v, both with $\alpha = 1$. The 0, 25, 50, 75, and 100 % quantiles of the loss ratio $L(\lambda_v^*)/L(\lambda_v)$ are given by 0.50, 0.97, 1.00, 1.01, and 1.05, in order, where $L(\lambda)$ is as in (3.13). The respective quantiles of the loss ratio for $\alpha = 1.4$ are 0.95, 1.00 1.01, 1.01, and 1.03.

6.3.5 R Package gss: ssanova9 Suite

Penalized likelihood regression with correlated Gaussian data are implemented in the `ssanova9` suite. The syntax is pretty much the same as that of `ssanova`, except that the optional arguments `weights` and `random` are replaced by a mandatory argument `cov`. The following sequence generates data with independent noise but fits a model with AR(1) errors:

```
x <- runif(100)
y <- 1+3*sin(2*pi*x-pi)+rnorm(x)
fit.ar1 <- ssanova9(y~x,cov=list("arma",c(1,0)))
```

To obtain the estimated coefficient $\gamma = \varphi_1$, use

```
para.arma(fit.ar1)$a
```

The following sequence generates data with MA(1) noise, fits a model with MA(1) errors, and obtains the estimated $\gamma = \theta_1$:

```
eps <- rnorm(101); eps <- eps[-1]-.5*eps[-101]
x <- runif(100)
y <- 1+3*sin(2*pi*x-pi)+eps
fit.ma1 <- ssanova9(y~x,cov=list("arma",c(0,1)))
para.arma(fit.ma1)$b
```

For longitudinal data, one may enter `cov=list("long",id)`, where `id` is a factor of subject identification. One may also use `ssanova9` with a known W^{-1}, with `cov=list("known",w)`, where `w` contains the known W^{-1}.

More generally, one may pass W^{-1} onto `ssanova9` via

```
cov=list(fun=...,env=...,init=...)
```

where W^{-1} is to be calculated via `fun(gamma,env)`, `env` contains constants, and `init` contains initial values of γ; γ should be properly parameterized to be free of constraint.

To evaluate W^{-1} for an `ssanova9` `fit` at the estimated γ, one may use

```
fit$cov$fun(fit$zeta,fit$cov$env)
```

6.4 Case Studies

We now apply the techniques developed in this chapter to analyze a couple of real data sets.

6.4.1 Treatment of Bacteriuria

Patients with acute spinal cord injury and bacteriuria (bacteria in urine) were randomly assigned to two treatment groups. Patients in the first group were treated for all episodes of urinary tract infection, whereas those in the second group were treated only if two specific symptoms occurred. Weekly binary indicator of bacteriuria was recorded for every patient over 4–16 weeks. A total of 72 patients were represented in the data, with 36 each in the two treatment groups. The data are listed in Joe (1997, §11.4), where further details and references can be found. There are a total of 892 observations, but the week-1 bacteriuria indicator was positive for all patients. After removing the week-1 data, we have a sample size $n = 820$.

The data are included in `gss` as a data frame `bacteriuria` with elements `id` (patient id), `trt` (treatments), `time` (weeks after randomization), and `infect` (bacteriuria indicator); `trt` and `id` are factors. One may fit a logistic regression model to the data, with the infection probability p as a function of the treatment and the follow-up time.

```
data(bacteriuria)
id.basis <- (1:820)[bacteriuria$id%in%c(3,38)]
fit.bact0 <- gssanova(infect~trt*time,"binomial",
                      data=bacteriuria,random=~1|id,
                      id.basis=id.basis)
```

there are only 30 distinctive x_i's (15 time points by 2 treatment levels), and patients 3 and 38 had complete follow-up under the two treatments. The random patient effect appears as an additive term in the logit,

$$\log \frac{p}{1-p} = \eta(x) + b_s.$$

The interaction term is negligible, so an additive model is fitted.

```
project(fit.bact0,c("trt","time"))$ratio
# 0.01166707
fit.bact1 <- gssanova(infect~trt+time,"binomial",
                      data=bacteriuria,random=~1|id,
                      id.basis=id.basis)
```

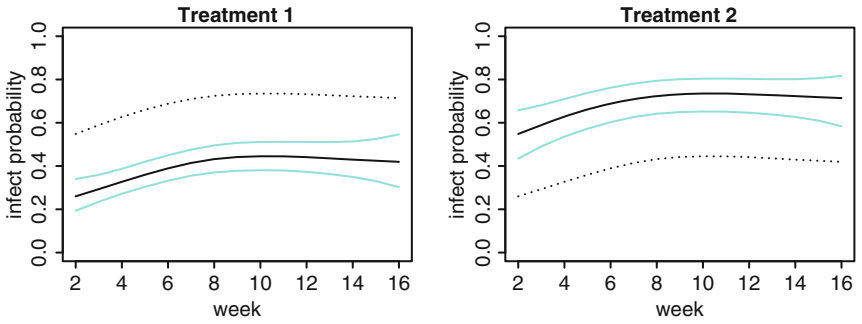


FIGURE 6.5. Bacteriuria infection probability. Estimated infection probability with 95 % Bayesian confidence intervals. The *dotted lines* mark the estimate under the other treatment.

Patients 1–36 were under treatment 1 and patients 37–72 were under treatment 2, and a quick check on the random patient effect reveals disparity between the two treatments.

```
var(fit.bact1$b[1:36])
# 0.05118155
var(fit.bact1$b[37:72])
# 0.2275906
```

Treatment 1 seems to allow less "individualism," so it appears appropriate to attach separate γ's to the two groups.

```
fit.bact2 <- gssanova(infect~trt+time,"binomial",
                      data=bacteriuria,random=~trt|id,
                      id.basis=id.basis)
var(fit.bact2$b[1:36])
# 1.582532e-15
```

The patient effect is in fact absent under treatment 1. The estimated infection probability as a function of time under the treatments can be evaluated and plotted as shown in Fig. 6.5.

```
new <- data.frame(trt=factor(rep(1,15)),time=2:16)
est.1 <- predict(fit.bact2,new,,se=TRUE)
plot(2:16,plogis(est.1$fit),type="l",ylim=c(0,1))
lines(2:16,plogis(est.1$fit-1.96*est.1$se),col=5)
lines(2:16,plogis(est.1$fit+1.96*est.1$se),col=5)
```

6.4.2 Ozone Concentration in Los Angeles Basin

We now revisit the ozone concentration data of §3.9.2. The fit shown in the bottom frames of Fig. 3.8 was estimated under $W = I$.

```
data(ozone, package="gss"); set.seed(5732)
fit.oz5 <- ssanova(log10(upo3)~sbtp+ibht+dgpg,data=ozone)
```

Inspections of the (partial) autocorrelation functions of the residuals suggest an AR(1) error structure, and one may refit the model using `ssanova9`.

```
acf(resid(fit.oz5)); pacf(resid(fit.oz5))
fit.oz6 <-ssanova9(log10(upo3)~sbtp+ibht+dgpg,data=ozone,
                   cov=list("arma",c(1,0)),id=fit.oz5$id)
para.arma(fit.oz6)$a
# 0.3095311
```

The cosine diagnostics are still available; premultiply (3.78) by C^{-T}, where $W^{-1} = C^T C$, and project the terms onto $\{C^{-T}\mathbf{1}\}^{\perp} = \{\mathbf{f} : \mathbf{f}^T W \mathbf{1} = 0\}$.

```
sum.oz6 <- summary(fit.oz6,TRUE)
round(sum.oz6$kappa,2)
# sbtp ibht dgpg
# 1.14 1.16 1.04
round(sum.oz6$pi,2)
# sbtp ibht dgpg
# 0.64 0.26 0.10
round(sum.oz6$cos,2)
#       sbtp ibht dgpg yhat    y    e
# cos.y 0.71 0.58 0.39 0.78 1.00 0.63
# cos.e 0.00 0.01 0.05 0.02 0.63 1.00
# norm  2.59 1.31 0.82 3.72 4.79 2.97
```

The terms of `fit.oz6` and `fit.oz5` are shown in Fig. 6.6, where the bottom frames are reproduced from Fig. 3.8. The `sbtp` effect in `fit.oz6` is virtually parametric, and the standard errors for the `ibht` effect of `fit.oz6` are slightly smaller than those of `fit.oz5`.

6.5 Bibliographic Notes

Section 6.1

Linear mixed-effect models, also known as variance component models, are extensively studied in the literature; see, e.g., Harville (1977) and Robinson (1991). The use of random effects in generalized linear models can be found in, e.g., Zeger and Karim (1991), Breslow and Clayton (1993), and McCulloch (1997).

Comprehensive treatments of stationary time series can be found in Box, Jenkins, and Reinsel (1994), Brockwell and Davis (1991), among others.

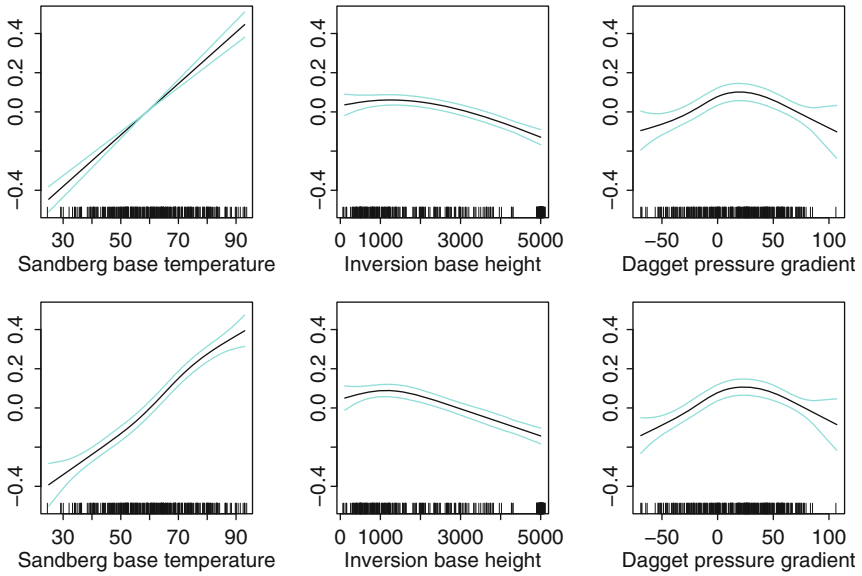


FIGURE 6.6. Terms in additive cubic spline fits to ozone data. The fits are in *solid lines* and the 95 % Bayesian confidence intervals in *faded*. *Top*: `fit.oz6` assuming AR(1) error. *Bottom*: `fit.oz5` assuming independent error. The rugs on the bottom mark the data points, slightly jittered.

Section 6.2

The materials of this section are mainly taken from Gu and Ma (2005b), and further results concerning non-Gaussian regression can be found in Gu and Ma (2005a). Penalized joint likelihood of $(\eta, \mathbf{b})$ allows one to use tools developed for independent data, resulting in structural simplicity and computational convenience. A thorough treatment of the strategy in parametric estimation can be found in Lee and Nelder (1996).

For treatments of nonparametric mixed-effect models via the marginal likelihood of η, see, e.g., Wang (1998a), Lin and Zhang (1999), and Karcher and Wang (2001).

Section 6.3

The materials of this section are mainly taken from Han and Gu (2008). Prior to that work, numerous *ad hoc* extensions of cross-validation had been proposed in the literature for use with correlated data, all demonstrating middling performances in the simulation studies of Wang (1998b), leaving the REML score $\tilde{M}(\lambda, \gamma)$ as the only viable solution at the time.

The parameterization of γ for the ARMA(p,q) model in `ssanova9` is taken from Jones (1980), which is free of constraint.

Section 6.4

The bacteriuria data were analyzed by Joe (1997) using Markov models with and without random effects. The analysis presented here is taken from Gu and Ma (2005a).

6.6 Problems

Section 6.1

6.1 Define

$$W = \begin{pmatrix} 1 & -\gamma & 0 & \cdots & 0 \\ -\gamma & 1+\gamma^2 & -\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

(a) Verify that

$$W^{-1} = \frac{1}{1-\gamma^2} \begin{pmatrix} 1 & \gamma & \gamma^2 & \cdots & \gamma^{n-1} \\ \gamma & 1 & \gamma & \cdots & \gamma^{n-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma^{n-1} & \gamma^{n-2} & \gamma^{n-3} & \cdots & 1 \end{pmatrix}.$$

(b) For $|\gamma| < 1$, show that $\big(1-|\gamma|\big)^2 I \le W \le \big(1+|\gamma|\big)^2 I$.

Section 6.2

6.2 The Moore-Penrose inverse M^+ of a non-negative definite matrix M satisfies $MM^+M = M$ and $M^+MM^+ = M^+$, with $MM^+ = M^+M$ being a projection matrix. Consider the matrix in (6.6),

$$M = \begin{pmatrix} \breve{R}^T\breve{R} + n\lambda\breve{Q} & \breve{R}^TZ \\ Z^T\breve{R} & Z^TZ + \Sigma \end{pmatrix} = \begin{pmatrix} E & \breve{R}^TZ \\ Z^T\breve{R} & D \end{pmatrix},$$

where $D > 0$.

(a) Show that $\tilde{D} = D - Z^T\breve{R}E^+\breve{R}^TZ > 0$.

(b) Show that

$$M^+ = \begin{pmatrix} E^+ + E^+\breve{R}^TZ\tilde{D}^{-1}Z^T\breve{R}E^+ & -E^+\breve{R}^TZ\tilde{D}^{-1} \\ -\tilde{D}^{-1}Z^T\breve{R}E^+ & \tilde{D}^{-1} \end{pmatrix}.$$

6.3 Verify (6.9) for $L_1(\lambda,\gamma)$ in (6.8) and

$$U(\lambda,\gamma)=\frac{1}{n}\mathbf{Y}^T\big(I-A(\lambda,\gamma)\big)^2\mathbf{Y}+2\frac{\sigma^2}{n}\mathrm{tr}A(\lambda,\gamma).$$

6.4 Verify that $(I+Z\,\Sigma^{-1}Z^T)^{-1}=I-Z(Z^TZ+\Sigma)^{-1}Z$.

Section 6.3

6.5 For $W^{-1}=C^TC$, $S_w=C^{-T}S$, $R_w=C^{-T}R$, and $M_w=R_wQ^+R_w^T+n\lambda I$, verify that A_w in (6.15) can be written as

$$A_w=I-n\lambda\big(M_w^{-1}-M_w^{-1}S_w(S_w^TM_w^{-1}S_w)^{-1}S_w^TM_w^{-1}\big).$$

6.6 Let $f_{\boldsymbol{\eta},W}$ be the density of $N\big(\boldsymbol{\eta},\sigma^2W^{-1}\big)$. Verify that the Kullback-Leibler distance of $f_1=f_{\boldsymbol{\eta}_1,W_1}$ from $f_0=f_{\boldsymbol{\eta}_0,W_0}$ is given by

$$E_{f_0}\big[\log(f_0/f_1)\big]=\frac{1}{2\sigma^2}(\boldsymbol{\eta}_1-\boldsymbol{\eta}_0)^TW_1(\boldsymbol{\eta}_1-\boldsymbol{\eta}_0)$$
$$+\frac{1}{2}\mathrm{tr}\big(W_1W_0^{-1}-I\big)-\frac{1}{2}\log\big|W_1W_0^{-1}\big|.$$

6.7 Verify (6.23).

7
Probability Density Estimation

For observational data, (1.5) of Example 1.2 defines penalized likelihood density estimation. Of interest are the selection of smoothing parameters, the computation of the estimates, and the asymptotic behavior of the estimates. Variants of (1.5) are also called for to accommodate samples subject to selection bias and samples from conditional distributions.

The precise formulation, the existence and uniqueness, and the computability of penalized likelihood density estimates are discussed in §7.1, and it is noted in §7.2 that the technique can be used to estimate inhomogeneous Poisson processes. The selection of smoothing parameters are discussed in §7.3, where a cross-validation score is derived and its empirical performance is assessed. Computational algorithms, inferential tools, and open-source software are discussed in §7.4, and the techniques are applied to analyze a few real data sets in §7.5. The estimation in the presence of sampling bias is treated in §§7.6 and 7.9. The estimation of the conditional density $f(y|x)$ is discussed in §7.7, with x and y on generic domains, which, for y discrete, leads to regression models with cross-classified responses (§7.8).

The computability of the estimates is through the notion of efficient approximation based on the asymptotic convergence rates, which will be discussed in Chap. 9.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_7,

7.1 Preliminaries

Let X_i, $i = 1, \ldots, n$, be independent and identically distributed (*i.i.d.*) random samples from a probability density $f(x)$ on a bounded domain $\mathcal{X}$. One is to estimate $f(x)$ from the observations X_i. When some parametric form of $f(x)$ is assumed, say $f \in P_\theta = \{f(x;\theta) : \theta \in \Theta\}$, where $f(x;\theta)$ is known up to a finite-dimensional parameter θ, density estimation reduces to parameter estimation, for which the maximum likelihood method is the standard technique possessing many favorable properties. When a parametric form is not available, however, a naive maximum likelihood density estimate without any nonintrinsic constraint (see the following paragraph for intrinsic constraints) is a sum of delta function spikes at the sample points, which, apparently, is not an appealing estimate when the domain $\mathcal{X}$ is continuous. In between the two extremes, one may use the penalized likelihood estimate.

Two intrinsic constraints that a probability density must satisfy are the positivity constraint that $f \geq 0$ and the unity constraint that $\int_{\mathcal{X}} f dx = 1$. Assuming $f > 0$ on $\mathcal{X}$, one can make a logistic density transform $f = e^\eta / \int_{\mathcal{X}} e^\eta dx$ and estimate η instead, which is free of the positivity and unity constraints. To make the transform one-to-one, one may impose a side condition on η, say $A\eta = 0$, where A is an averaging operator on $\mathcal{X}$; see §1.3.1 for a discussion of averaging operators. The estimate of η can then be obtained by minimizing the penalized likelihood functional,

$$-\frac{1}{n}\sum_{i=1}^{n} \eta(X_i) + \log \int_{\mathcal{X}} e^\eta dx + \frac{\lambda}{2} J(\eta), \tag{7.1}$$

in a reproducing kernel Hilbert space $\mathcal{H}$, in which the roughness penalty $J(\eta)$ is a square (semi) norm. The members of $\mathcal{H}$ have to comply with a side condition mentioned above to make the first term of (7.1) strictly convex. It is easy to construct such an $\mathcal{H}$ by dropping the constant term in a (one-way) ANOVA decomposition.

Let $L(f) = -n^{-1}\sum_{i=1}^{n} f(X_i) + \log \int_{\mathcal{X}} e^f dx$ be the minus log likelihood. When the maximum likelihood estimate exists in the null space $\mathcal{N}_J = \{f : Af = 0, J(f) = 0\}$, the following lemmas establish the existence and uniqueness of the minimizer of (7.1) via Theorem 2.9.

Lemma 7.1 *$L(f)$ is strictly convex for $f \in \mathcal{H} \subseteq \{f : Af = 0\}$.*

Proof: By Hölder's inequality, for $\alpha, \beta > 0$, $\alpha + \beta = 1$, and $f, g \in \mathcal{H}$,

$$\log \int_{\mathcal{X}} e^{\alpha f + \beta g} dx \leq \alpha \log \int_{\mathcal{X}} e^f dx + \beta \log \int_{\mathcal{X}} e^g dx,$$

where the equality holds if and only if $e^f \propto e^g$, which amounts to $f = g$ with $Af = Ag = 0$. □

Lemma 7.2 *If $e^{|f|}$ are Riemann integrable on $\mathcal{X}$ for all $f \in \mathcal{H}$, then $L(f)$ is continuous in $\mathcal{H}$. Furthermore, $L(f+\alpha g)$, $\forall f, g \in \mathcal{H}$, is infinitely differentiable as a function of α real.*

Proof: The claims follow from the Riemann sum approximations of related integrals and the continuity of evaluation. □

A simple example follows.

Example 7.1 (Cubic spline) Let $\mathcal{X} = [0,1]$ and $J(\eta) = \int_0^1 \ddot{\eta}^2 dx$. The null space of $J(\eta)$ without side condition is $\text{span}\{1, x\}$. One has the choice of at least two different formulations.

The first formulation employs the construction of §2.3.1. Take $Af = f(0)$. One has

$$\mathcal{H} = \left\{ f : f(0) = 0, \int_0^1 \ddot{f}^2 dx < \infty \right\} = \mathcal{N}_J \oplus \mathcal{H}_J,$$

where $\mathcal{N}_J = \text{span}\{x\}$ and

$$\mathcal{H}_J = \left\{ f : f(0) = \dot{f}(0) = 0, \int_0^1 \ddot{f}^2 dx < \infty \right\},$$

with $R_J(x,y) = \int_0^1 (x-u)_+(y-u)_+ du$.

The second formulation employs the construction of §2.3.3. Take $Af = \int_0^1 f dx$. One has

$$\mathcal{H} = \left\{ f : \int_0^1 f dx = 0, \int_0^1 \ddot{f}^2 dx < \infty \right\} = \mathcal{N}_J \oplus \mathcal{H}_J,$$

where $\mathcal{N}_J = \text{span}\{x - .5\}$ and

$$\mathcal{H}_J = \left\{ f : \int_0^1 f dx = \int_0^1 \dot{f} dx = 0, \int_0^1 \ddot{f}^2 dx < \infty \right\},$$

with $R_J(x,y) = k_2(x)k_2(y) - k_4(x-y)$; see (2.27) on page 39 for $k_2(x)$ and $k_4(x)$. □

With the same data and the same penalty, one would naturally expect that the two formulations of Example 7.1 would yield the same density estimate. It is indeed the case, as assured by the following proposition.

Proposition 7.3 *Let $\mathcal{H} \subseteq \left\{ f : J(f) < \infty \right\}$ and suppose that $J(f)$ annihilates constant. For any two different averaging operators A_1 and A_2, if η_1 minimizes (7.1) in $\mathcal{H}_1 = \mathcal{H} \cap \left\{ A_1 f = 0 \right\}$ and η_2 minimizes (7.1) in $\mathcal{H}_2 = \mathcal{H} \cap \left\{ A_2 f = 0 \right\}$, then $e^{\eta_1} / \int_{\mathcal{X}} e^{\eta_1} dx = e^{\eta_2} / \int_{\mathcal{X}} e^{\eta_2} dx$.*

Proof: For any $f \in \mathcal{H}_1$, it is easy to verify that $Pf = f - A_2 f \in \mathcal{H}_2$, $L(Pf) = L(f)$, and $J(Pf) = J(f)$. Similarly, for any $g \in \mathcal{H}_2$, $Qg = g - A_1 g \in \mathcal{H}_1$, $L(Qg) = L(g)$, and $J(Qg) = J(g)$. Now, for $f \in \mathcal{H}_1$, $Q(Pf) = Pf - A_1(Pf) = (f - A_2 f) - A_1(f - A_2 f) = f$, so there is an isomorphism between $\mathcal{H}_1$ and $\mathcal{H}_2$. Clearly, $e^f / \int_{\mathcal{X}} e^f dx = e^{Pf} / \int_{\mathcal{X}} e^{Pf} dx$. The proposition follows. □

Example 7.2 (Tensor product spline) Consider the domain $\mathcal{X}=[0,1]^3$. Multiple-term models can be constructed using the tensor product splines of §2.4, with an ANOVA decomposition

$$f = f_\emptyset + f_1 + f_2 + f_3 + f_{1,2} + f_{1,3} + f_{2,3} + f_{1,2,3},$$

where terms other than the constant $f_\emptyset$ satisfy certain side conditions. The constant shall be dropped for density estimation to maintain a one-to-one logistic density transform. The remaining seven components can all be included or excluded separately, resulting in 2^7 possible models of different complexities. The additive model implies the independence of the three coordinates, and it is easily seen to be equivalent to solutions of three separate problems on individual axes. Less trivial probability structures may also be built in via selective inclusion of the ANOVA terms. For example, the conditional independence of $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$ given $x_{\langle 3\rangle}$ may be incorporated by excluding $f_{1,2}$ and $f_{1,2,3}$ from the model.

The above discussion is simply a partial repeat of §1.3.3, where more discussions can be found. □

In addition to the evaluations $[x_i]\eta = \eta(x_i)$, the first term of (7.1) depends on η also through the integral $\int_{\mathcal{X}} e^\eta dx$. This breaks the argument of §2.3.2, so the solution expression (3.2) on page 62 no longer holds for the minimizer η_λ of (7.1) in the space $\mathcal{H} = \{f : Af = 0, J(f) < \infty\}$. Actually, η_λ is, in general, not computable. The notion of efficient approximation comes to rescue here, and one may calculate the minimizer η_λ^* of (7.1) in a (data-adaptive) finite-dimensional space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(Z_j, \cdot), j = 1, \ldots, q\}, \tag{7.2}$$

where $\{Z_j\}$ is a random subset of $\{X_i\}$. It is shown in §9.2.3 that η_λ^* and η_λ share the same asymptotic convergence rates with $q \asymp n^{2/(pr+1)+\epsilon}$ for some $r > 1$, $p \in [1, 2]$, and $\forall \epsilon > 0$, so there is no loss of efficiency in the substitution of $\mathcal{H}$ by $\mathcal{H}^*$. When the maximum likelihood estimate exists in the null space $\mathcal{N}_J$, the existence and uniqueness of η_λ^* follow from Lemmas 7.1 and 7.2.

Proposition 7.3 does not apply to η_λ^*; for the two different formulations in Example 7.1, $\mathcal{H}^*$ are different even for the same choice of $\{Z_j\}$. The asymptotic convergence results of §9.2 hold regardless which R_J is used, however, and the variability due to different choices of R_J is not much different from the variability due to different choices of $\{Z_j\}$.

In the rest of the chapter, we shall focus on η_λ^* but drop the star from the notation. Plugging the expression

$$\eta(x) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(x) + \sum_{j=1}^{q} c_j R_J(Z_j, x) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c} \tag{7.3}$$

into (7.1), the calculation of η_λ reduces to the minimization of

$$A_\lambda(\mathbf{c},\mathbf{d}) = -\frac{1}{n}\mathbf{1}^T(S\mathbf{d}+R\mathbf{c}) + \log\int_{\mathcal{X}} \exp\left(\boldsymbol{\phi}^T\mathbf{d}+\boldsymbol{\xi}^T\mathbf{c}\right)dx + \frac{\lambda}{2}\mathbf{c}^TQ\mathbf{c} \quad (7.4)$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where S is $n\times m$ with the (i,ν)th entry $\phi_\nu(X_i)$, R is $n\times q$ with the (i,j)th entry $\xi_j(X_i)=R_J(Z_j,X_i)$, and Q is $q\times q$ with the (j,k)th entry $R_J(Z_j,Z_k)$.

Write $\mu_f(g) = \int ge^f dx/\int e^f dx$, $V_f(g,h) = \mu_f(gh)-\mu_f(g)\mu_f(h)$, and $V_f(g)=V_f(g,g)$. Taking derivatives at $\tilde\eta = \boldsymbol{\phi}^T\tilde{\mathbf{d}}+\boldsymbol{\xi}^T\tilde{\mathbf{c}}\in\mathcal{H}^*$, one has

$$\begin{aligned}
\frac{\partial A_\lambda}{\partial \mathbf{d}} &= -S^T\mathbf{1}/n + \mu_{\tilde\eta}(\boldsymbol{\phi}) = -S^T\mathbf{1}/n+\mu_\phi,\\
\frac{\partial A_\lambda}{\partial \mathbf{c}} &= -R^T\mathbf{1}/n + \mu_{\tilde\eta}(\boldsymbol{\xi}) + \lambda Q\tilde{\mathbf{c}} = -R^T\mathbf{1}/n+\mu_\xi+\lambda Q\tilde{\mathbf{c}},\\
\frac{\partial^2 A_\lambda}{\partial\mathbf{d}\partial\mathbf{d}^T} &= V_{\tilde\eta}(\boldsymbol{\phi},\boldsymbol{\phi}^T) = V_{\phi,\phi},\\
\frac{\partial^2 A_\lambda}{\partial\mathbf{c}\partial\mathbf{c}^T} &= V_{\tilde\eta}(\boldsymbol{\xi},\boldsymbol{\xi}^T)+\lambda Q = V_{\xi,\xi}+\lambda Q,\\
\frac{\partial^2 A_\lambda}{\partial\mathbf{d}\partial\mathbf{c}^T} &= V_{\tilde\eta}(\boldsymbol{\phi},\boldsymbol{\xi}^T) = V_{\phi,\xi};
\end{aligned} \quad (7.5)$$

see Problem 7.1. The Newton updating equation is thus

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi}\\ V_{\xi,\phi} & V_{\xi,\xi}+\lambda Q\end{pmatrix}\begin{pmatrix}\mathbf{d}-\tilde{\mathbf{d}}\\ \mathbf{c}-\tilde{\mathbf{c}}\end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n-\mu_\phi\\ R^T\mathbf{1}/n-\mu_\xi-\lambda Q\tilde{\mathbf{c}}\end{pmatrix}. \quad (7.6)$$

After rearranging terms, (7.6) becomes

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi}\\ V_{\xi,\phi} & V_{\xi,\xi}+\lambda Q\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n-\mu_\phi+V_{\phi,\eta}\\ R^T\mathbf{1}/n-\mu_\xi+V_{\xi,\eta}\end{pmatrix}, \quad (7.7)$$

where $V_{\phi,\eta}=V_{\tilde\eta}(\boldsymbol{\phi},\tilde\eta)$ and $V_{\xi,\eta}=V_{\tilde\eta}(\boldsymbol{\xi},\tilde\eta)$; see Problem 7.2. Fixing the smoothing parameter λ, and θ_β hidden in R and Q for multiple-term models, one may iterate on (7.7) to calculate η_λ.

For prebinned data with replicate counts k_i at X_i, (7.4) becomes

$$-\frac{1}{N}\mathbf{k}^T(S\mathbf{d}+R\mathbf{c}) + \log\int_{\mathcal{X}}\exp\left(\boldsymbol{\phi}^T\mathbf{d}+\boldsymbol{\xi}^T\mathbf{c}\right)dx + \frac{\lambda}{2}\mathbf{c}^TQ\mathbf{c}, \quad (7.8)$$

where $\mathbf{k}=(k_1,\ldots,k_n)^T$ and $N=\sum_{i=1}^n k_i$, and (7.7) changes to

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi}\\ V_{\xi,\phi} & V_{\xi,\xi}+\lambda Q\end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} = \begin{pmatrix} S^T\mathbf{k}/N-\mu_\phi+V_{\phi,\eta}\\ R^T\mathbf{k}/N-\mu_\xi+V_{\xi,\eta}\end{pmatrix}. \quad (7.9)$$

On high-dimensional domains, the prohibitive cost of numerical integration renders (7.1) impractical. One however may use the penalized pseudo likelihood to be developed in §10.1, gaining numerical feasibility at the cost of degraded statistical performance.

7.2 Poisson Intensity

Consider a Poisson counting process on $\mathcal{X}$ with an intensity function $\lambda(x)$, where $\lambda(x)$ is not to be confused with the smoothing parameter λ. Observing N occurrences X_i, $i = 1, \ldots, N$, from the process, the joint likelihood of N and X_i can be shown to be

$$\left\{ \prod_{i=1}^{N} \lambda(X_i) \right\} \exp\left\{ -\int_{\mathcal{X}} \lambda(x)dx \right\} = \left\{ \prod_{i=1}^{N} \lambda_0(X_i) \right\} \left(\Lambda^N e^{-\Lambda}\right),$$

where $\Lambda = \int_{\mathcal{X}} \lambda(x)dx$ is the overall intensity of the process on $\mathcal{X}$ and $\lambda_0(x) = \lambda(x)/\Lambda$ is the occurrence density; see, e.g., Snyder (1975, §2.3). N is statistically sufficient for Λ and has a Poisson distribution with intensity Λ, and $X_i|N$ are conditionally independent with a probability density $\lambda_0(x)$. A penalized likelihood estimate of the Poisson intensity can be defined as the minimizer of

$$-\sum_{i=1}^{N} \log \lambda_0(X_i) - N \log \Lambda + \Lambda + J\big(\log \lambda_0(x) + \log \Lambda\big), \tag{7.10}$$

for $\log \lambda(x) \in \tilde{\mathcal{H}} \supset \{1\}$, where $\tilde{\mathcal{H}}$ is a general reproducing kernel Hilbert space and the smoothing parameter is absorbed into the roughness penalty $J(f)$ to avoid confusion with the intensity $\lambda(x)$. Decompose $\tilde{\mathcal{H}} = \{1\} \oplus \mathcal{H}$, where $\mathcal{H}$ satisfies a side condition, and write $\log \lambda(x) = C + \eta$, where C is a constant and $\eta \in \mathcal{H}$. Since $\log \lambda_0 = \eta - \log \int_{\mathcal{X}} e^{\eta} dx$ and $\log \Lambda = C + \log \int_{\mathcal{X}} e^{\eta} dx$, (7.10) can be written as

$$\left[-\sum_{i=1}^{N} \eta(X_i) + N \log \int_{\mathcal{X}} e^{\eta} dx + J(C + \eta) \right] + \left[-N\left(C + \log \int_{\mathcal{X}} e^{\eta} dx \right) + \exp\left(C + \log \int_{\mathcal{X}} e^{\eta} dx \right) \right]; \tag{7.11}$$

see Problem 7.3. Naturally, $J(f)$ should annihilate constant since smoothing should only apply to the occurrence density, so $J(C + \eta) = J(\eta)$. The minimization of (7.11) can then be achieved in two steps: first to minimize the sum in the first pair of square brackets in (7.11) with respect to $\eta \in \mathcal{H}$ to estimate the occurrence density $\lambda_0(x)$ and, second, to minimize the sum in the second pair of square brackets with respect to C to estimate the overall intensity Λ. The former is simply a penalized likelihood density estimation through (7.1) based on X_i, $i = 1, \ldots, N$, and the latter is a Poisson density estimation based on a single observation N.

When $J(f)$ annihilates constant, the two-step estimation in (7.11) may be manipulated to enforce an arbitrary positive value on Λ by modifying

the second part. Specifically, replacing $-N\log\Lambda+\Lambda$ by $-N\log\Lambda+N\Lambda$ in (7.11), one effectively enforces $\Lambda=1$. Dividing the functional thus modified by N, one has

$$-\frac{1}{N}\sum_{i=1}^{N}\tilde{\eta}(X_i)+\int_{\mathcal{X}}e^{\tilde{\eta}}dx+\tilde{J}(\tilde{\eta}), \tag{7.12}$$

where $\tilde{\eta}=\log\lambda(x)$ and $\tilde{J}(f)=J(f)/N$. Obviously, the minimizer $\tilde{\eta}^*$ of (7.12) satisfies $\int_{\mathcal{X}}e^{\tilde{\eta}^*}dx=1$; see Problem 7.4. This device was proposed by Silverman (1982) to enforce the unity constraint without imposing any side condition on the log density. Were a probability density defined to integrate to 2, one would use $\int_{\mathcal{X}}e^{\eta}dx/2$ in (7.12) instead of $\int_{\mathcal{X}}e^{\eta}dx$ to enforce the "unity" constraint $\int_{\mathcal{X}}e^{\tilde{\eta}^*}dx=2$.

7.3 Smoothing Parameter Selection

As with regression, smoothing parameter selection holds the key to any practical success of penalized likelihood density estimation. Similar to the situation with non-Gaussian regression in Chap. 5, the convex but non-quadratic functional (7.1) has to be minimized iteratively even for fixed smoothing parameters. Needed are effective methods to locate good estimates from among the η_λ's with varying smoothing parameters.

Similar to the developments in §5.2.2, a direct cross-validation score will be derived for density estimation. The Newton update for solving (7.1) no longer has its own statistical meaning as in (5.3), so there exists no alternative score to drive a possible performance-oriented iteration; the self-voting argument may still apply using the direct cross-validation score, but there is little numerical benefit to justify an indirect approach. The empirical performance of the cross-validation score and its modifications will be explored in simulation studies.

As in §§3.2 and 5.2, we only make the dependence of various entities on the smoothing parameter λ explicit, suppressing their dependence on θ_β in the notation.

7.3.1 Kullback-Leibler Loss

To measure the proximity of the estimate $f_\lambda=e^{\eta_\lambda}/\int_{\mathcal{X}}e^{\eta_\lambda}dx$ to the true density $f=e^{\eta}/\int_{\mathcal{X}}e^{\eta}dx$, consider the Kullback-Leibler distance

$$\mathrm{KL}(\eta,\eta_\lambda)=E_f\big[\log(f/f_\lambda)\big]=\mu_\eta(\eta-\eta_\lambda)-\log\int_{\mathcal{X}}e^{\eta}dx+\log\int_{\mathcal{X}}e^{\eta_\lambda}dx,$$

where $\mu_f(g)=\int ge^f dx/\int e^f dx$ as defined in §7.1, and the symmetrized version

$$L(\eta,\eta_\lambda)=\mathrm{SKL}(\eta,\eta_\lambda)=\mu_\eta(\eta-\eta_\lambda)+\mu_{\eta_\lambda}(\eta_\lambda-\eta). \tag{7.13}$$

Dropping terms in $\mathrm{KL}(\eta, \eta_\lambda)$ that do not involve η_λ, one has the relative Kullback-Leibler distance,

$$\mathrm{RKL}(\eta, \eta_\lambda) = \log \int_{\mathcal{X}} e^{\eta_\lambda} dx - \mu_\eta(\eta_\lambda). \tag{7.14}$$

The first term is readily computable, but the second term, $\mu_\eta(\eta_\lambda)$, involves the unknown density and will have to be estimated.

7.3.2 Cross-Validation

A naive estimate of $\mu_\eta(\eta_\lambda)$ is the sample mean $n^{-1} \sum_{i=1}^n \eta_\lambda(X_i)$, but the resulting estimate of the relative Kullback-Leibler distance would simply be the minus log likelihood, clearly favoring $\lambda = 0$. The naive sample mean is biased because the samples X_i contribute to the estimate η_λ. Standard cross-validation suggests an estimate $\tilde{\mu}_\eta(\eta_\lambda) = n^{-1} \sum_{i=1}^n \eta_\lambda^{[i]}(X_i)$, where $\eta_\lambda^{[i]}$ minimizes the delete-one version of (7.1),

$$-\frac{1}{n-1} \sum_{j \neq i} \eta(X_j) + \log \int_{\mathcal{X}} e^{\eta} dx + \frac{\lambda}{2} J(\eta). \tag{7.15}$$

Note that X_i does not contribute to $\eta_\lambda^{[i]}$, although $\eta_\lambda^{[i]}$ is not quite the same as η_λ. The delete-one estimates $\eta_\lambda^{[i]}$ are not analytically available, however; so it is impractical to compute $\tilde{\mu}_\eta(\eta_\lambda)$ directly.

For an analytically tractable approximation of $\eta_\lambda^{[i]}$, consider the quadratic approximation of (7.1) at η_λ. For $f, g \in \mathcal{H}$ and α real, define $L_{f,g}(\alpha) = \log \int_{\mathcal{X}} e^{f+\alpha g} dx$ as a function of α. It is easy to show that $\dot{L}_{f,g}(0) = \mu_f(g)$ (hence $L(f) = \log \int_{\mathcal{X}} e^f dx$ is Fréchet differentiable) and that $\ddot{L}_{f,g}(0) = V_f(g)$; see Problem 7.5. Setting $f = \tilde{\eta}$, $g = \eta - \tilde{\eta}$, and $\alpha = 1$, one has the Taylor expansion

$$\log \int_{\mathcal{X}} e^{\eta} dx = L_{\tilde{\eta}, \eta - \tilde{\eta}}(1) \approx L_{\tilde{\eta}, \eta - \tilde{\eta}}(0) + \mu_{\tilde{\eta}}(\eta - \tilde{\eta}) + \frac{1}{2} V_{\tilde{\eta}}(\eta - \tilde{\eta}). \tag{7.16}$$

Substituting the right-hand side of (7.16) for the term $\log \int_{\mathcal{X}} e^{\eta} dx$ in (7.1) and dropping terms that do not involve η, one obtains the quadratic approximation of (7.1) at $\tilde{\eta}$:

$$-\frac{1}{n} \sum_{i=1}^n \eta(X_i) + \mu_{\tilde{\eta}}(\eta) - V_{\tilde{\eta}}(\tilde{\eta}, \eta) + \frac{1}{2} V_{\tilde{\eta}}(\eta) + \frac{\lambda}{2} J(\eta). \tag{7.17}$$

Plugging (7.3) into (7.17) and solving for $\mathbf{c}$ and $\mathbf{d}$, one obtains (7.7); see Problem 7.6.

The delete-one version of (7.17),

$$-\frac{1}{n-1} \sum_{j \neq i} \eta(X_j) + \mu_{\tilde{\eta}}(\eta) - V_{\tilde{\eta}}(\tilde{\eta}, \eta) + \frac{1}{2} V_{\tilde{\eta}}(\eta) + \frac{\lambda}{2} J(\eta), \tag{7.18}$$

only involves changes in the first term. Set $\tilde{\eta} = \eta_\lambda$ and write $\breve{\boldsymbol{\xi}} = (\boldsymbol{\phi}^T, \boldsymbol{\xi}^T)^T$ and $\breve{\mathbf{c}} = (\mathbf{d}^T, \mathbf{c}^T)^T$. Rewrite (7.7) as

$$H\breve{\mathbf{c}} = \breve{R}^T\mathbf{1}/n + \mathbf{g},$$

where $H = V_{\tilde{\eta}}(\breve{\boldsymbol{\xi}}, \breve{\boldsymbol{\xi}}^T) + \operatorname{diag}(O, \lambda Q)$, $\breve{R}^T = \big(\breve{\boldsymbol{\xi}}(X_1), \ldots, \breve{\boldsymbol{\xi}}(X_n)\big) = (S, R)^T$, and $\mathbf{g} = V_{\tilde{\eta}}(\breve{\boldsymbol{\xi}}, \tilde{\eta}) - \mu_{\tilde{\eta}}(\breve{\boldsymbol{\xi}})$. The minimizer $\eta^{[i]}_{\lambda,\tilde{\eta}}$ of (7.18) has the coefficient

$$\breve{\mathbf{c}}^{[i]} = H^{-1}\left(\frac{\breve{R}^T\mathbf{1} - \breve{\boldsymbol{\xi}}(X_i)}{n-1} + \mathbf{g}\right) = \breve{\mathbf{c}} + \frac{H^{-1}\breve{R}^T\mathbf{1}}{n(n-1)} - \frac{H^{-1}\breve{\boldsymbol{\xi}}(X_i)}{n-1},$$

so

$$\eta^{[i]}_{\lambda,\tilde{\eta}}(X_i) = \breve{\boldsymbol{\xi}}(X_i)^T\breve{\mathbf{c}}^{[i]} = \breve{\boldsymbol{\xi}}(X_i)^T\breve{\mathbf{c}} - \frac{1}{n-1}\breve{\boldsymbol{\xi}}(X_i)^T H^{-1}\big(\breve{\boldsymbol{\xi}}(X_i) - \breve{R}^T\mathbf{1}/n\big). \tag{7.19}$$

Noting that $\breve{R}^T\mathbf{1}/n = n^{-1}\sum_{i=1}^n \breve{\boldsymbol{\xi}}(X_i)$, this leads to a cross-validation estimate of $\mu_\eta(\eta_\lambda)$,

$$\hat{\mu}_\eta(\eta_\lambda) = \frac{1}{n}\sum_{i=1}^n \eta^{[i]}_{\lambda,\tilde{\eta}}(X_i) = \frac{1}{n}\sum_{i=1}^n \eta_\lambda(X_i) - \frac{\operatorname{tr}\big(P_{\mathbf{1}}^{\perp}\breve{R}H^{-1}\breve{R}^T P_{\mathbf{1}}^{\perp}\big)}{n(n-1)}, \tag{7.20}$$

where $P_{\mathbf{1}}^{\perp} = I - \mathbf{1}\mathbf{1}^T/n$, and the corresponding estimate of the relative Kullback-Leibler distance,

$$V(\lambda) = -\frac{1}{n}\sum_{i=1}^n \eta_\lambda(X_i) + \log\int_{\mathcal{X}} e^{\eta_\lambda}dx + \alpha\frac{\operatorname{tr}\big(P_{\mathbf{1}}^{\perp}\breve{R}H^{-1}\breve{R}^T P_{\mathbf{1}}^{\perp}\big)}{n(n-1)}, \tag{7.21}$$

for $\alpha = 1$. Note that $\eta^{[i]}_{\lambda,\tilde{\eta}}$ is simply the one-step Newton update from η_λ for the minimization of (7.15).

For prebinned data, the delete-one operation should be done on the individual observations instead of the bins, yielding

$$\begin{aligned} V(\lambda) &= \log\int_{\mathcal{X}} e^{\eta_\lambda}dx - \frac{1}{N}\sum_{i=1}^n k_i\eta^{[i]}_{\lambda,\tilde{\eta}}(X_i) \\ &= -\frac{1}{N}\sum_{i=1}^n k_i\eta_\lambda(X_i) + \log\int_{\mathcal{X}} e^{\eta_\lambda}dx + \frac{\operatorname{tr}\big(P_{\tilde{\mathbf{k}}}^{\perp}\tilde{K}\breve{R}H^{-1}\breve{R}^T\tilde{K}P_{\tilde{\mathbf{k}}}^{\perp}\big)}{N(N-1)}, \end{aligned} \tag{7.22}$$

where $P_{\tilde{\mathbf{k}}}^{\perp} = I - \tilde{\mathbf{k}}\tilde{\mathbf{k}}^T/N$ with $\tilde{\mathbf{k}} = \big(\sqrt{k_1}, \ldots, \sqrt{k_n}\big)^T$, $\tilde{K} = \operatorname{diag}\big(\sqrt{k_i}\big)$, and $\eta^{[i]}_{\lambda,\tilde{\eta}}$ minimizes

$$-\frac{1}{N-1}\left\{\sum_{j=1}^n k_j\eta(X_j) - \eta(X_i)\right\} + \mu_{\tilde{\eta}}(\eta) - V_{\tilde{\eta}}(\tilde{\eta}, \eta) + \frac{1}{2}V_{\tilde{\eta}}(\eta) + \frac{\lambda}{2}J(\eta);$$

see Problem 7.7.

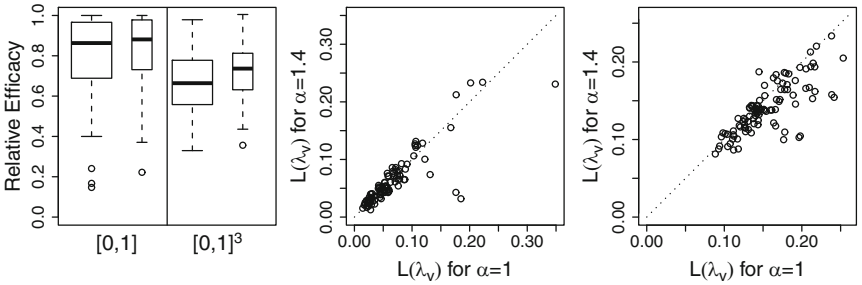


FIGURE 7.1. Effectiveness of cross-validation for density estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0, 1]$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0, 1]^3$.

7.3.3 Empirical Performance

Simple simulations were conducted to explore the empirical performance of cross-validation. On $\mathcal{X} = [0, 1]$, samples of size $n = 100$ were drawn from

$$f_1(x) \propto \tilde{f}_1(x) I_{x\in[0,1]} = \left\{ \frac{1}{3} e^{-50(x-0.3)^2} + \frac{2}{3} e^{-50(x-0.7)^2} \right\} I_{x\in[0,1]}, \quad (7.23)$$

which is a mixture of $N(0.3, 0.1^2)$ and $N(0.7, 0.1^2)$ truncated to $[0, 1]$. Using the second formulation of cubic spline as discussed in Example 7.1 and setting $q = n$ in (7.3), three estimates were calculated for each replicate, one minimizing $L(\lambda) = L(\eta, \eta_\lambda)$ of (7.13), another minimizing $V(\lambda)$ of (7.21) with $\alpha = 1$, and a third minimizing $V(\lambda)$ with $\alpha = 1.4$, yielding an optimal loss $L(\lambda_o)$ and two cross-validation losses $L(\lambda_v)$. The results from one hundred replicates are summarized in Fig. 7.1, with the relative efficacy $L(\lambda_o)/L(\lambda_v)$ shown in the left half of the left frame and the comparison of $\alpha = 1, 1.4$ in $V(\lambda)$ shown in the center frame.

On $\mathcal{X} = [0, 1]^3$, samples of size $n = 300$ were generated from

$$f_3(x) \propto e^{-12.5(x_{\langle 3\rangle}-0.5)^2} \tilde{f}_1(x_{\langle 1\rangle} - 0.3x_{\langle 3\rangle} + 0.1) \tilde{f}_1(x_{\langle 2\rangle} - 0.2x_{\langle 3\rangle} + 0.1) I_{x\in[0,1]^3}, \quad (7.24)$$

where $\tilde{f}_1(x)$ is as given in (7.23). Estimates with $q = 36$ were calculated using tensor product cubic splines of the form $\eta(x) = \eta_1+\eta_2+\eta_3+\eta_{1,3}+\eta_{2,3}$, where the conditional independence structure $(X_1 \perp X_2)|X_3$ is built in. The results from one hundred replicates are summarized in Fig. 7.1, with the relative efficacy in the right half of the left frame and the comparison of $\alpha = 1, 1.4$ in $V(\lambda)$ in the right frame.

On $\mathcal{X} = [0, 1]$, we set $q = n$ to take away the variability due to the choice of $\{Z_j\}$. On $\mathcal{X} = [0, 1]^3$, when q is large, we constantly ran into numerical

problems with the Newton iteration via (7.6) for cross-validated fits with $\alpha = 1$, so had to settle with the default $q = 10n^{2/9}$; simulations similar to those in §3.5.4 but in the density estimation setting can be found in Gu and Wang (2003), which suggested the default q value. We took care to use the same $\{Z_j\}$ for all the three estimates in each replicate, so the comparisons in Fig. 7.1 are adequate.

7.4 Computation, Inference, and Software

Fixing smoothing parameters, the computation involves the Newton iteration via (7.6) and the evaluation of the cross-validation score $V(\lambda)$ given in (7.21). To select smoothing parameters by cross-validation, quasi-Newton methods with numerical derivatives, such as those developed in Dennis and Schnabel (1996), can be employed to minimize $V(\lambda)$ with respect to the smoothing parameters.

Numerical integration is needed for the calculation of entities appearing in (7.6) and (7.21), which is nontrivial on a multidimensional $\mathcal{X}$.

For the "testing" of $H_0 : \eta \in \mathcal{H}_0$ versus $H_a : \eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, one again can make use of the Kullback-Leibler projection.

Software implementation of the techniques developed is embodied in the `ssden` suite in `gss`, whose usage is illustrated through simulated examples.

7.4.1 Newton Iteration

To perform the Newton iteration via (7.6), one calculates the Cholesky decomposition

$$H = \begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} = \begin{pmatrix} G_1^T & O \\ G_2^T & G_3^T \end{pmatrix} \begin{pmatrix} G_1 & G_2 \\ O & G_3 \end{pmatrix} = G^T G$$

for G upper-triangular, where $G_1^T G_1 = V_{\phi,\phi}$, $G_1^T G_2 = V_{\phi,\xi}$, and $G_3^T G_3 = (V_{\xi,\xi} - V_{\xi,\phi} V_{\phi,\phi}^{-1} V_{\phi,\xi}) + \lambda Q$, and then uses forward and back substitutions to calculate the update. Standard safeguard procedures such as step-halving might be called upon to ensure decreasing penalized likelihood scores in each step, and the iteration usually takes five to ten steps to converge given reasonable starting values. The Cholesky decomposition takes $O(q^3)$ flops and the substitutions take $O(q^2)$, usually dominated by the $O(dq^2)$ flops needed to form (7.6), where d is the quadrature size for numerical integration on $\mathcal{X}$.

On the convergence of the Newton iteration, the Cholesky decomposition $H = G^T G$ has already been computed. Back substitution yields $G^{-T} \breve{R}^T$ in $O(nq^2)$ flops, from which $\mathrm{tr}\big(P_{\mathbf{1}}^{\perp} \breve{R} H^{-1} \breve{R}^T P_{\mathbf{1}}^{\perp}\big)$ can be computed.

Care must be taken for numerically singular H, which may arise when $\xi_j(x) = R_J(Z_j, x)$ are linearly dependent. With a possible permutation of indices known as pivoting, G_3 in this case can be written as

$$G_3 = \begin{pmatrix} J_1 & J_2 \\ O & O \end{pmatrix} = \begin{pmatrix} J \\ O \end{pmatrix},$$

where J is of full row rank and $G_3^T G_3 = J^T J$. Define

$$\tilde{G}_3 = \begin{pmatrix} J_1 & J_2 \\ O & \delta I \end{pmatrix}, \quad \tilde{G} = \begin{pmatrix} G_1 & G_2 \\ O & \tilde{G}_3 \end{pmatrix},$$

for some $\delta > 0$, and partition $\tilde{G}_3^{-1} = (K, L)$. It follows that $JK = I$ and $JL = O$. This leads to $L^T G_3^T G_3 L = O$, and since $V_{\xi,\xi} - V_{\xi,\phi} V_{\phi,\phi}^{-1} V_{\phi,\xi}$ is non-negative definite, $L^T Q L = O$. Noting that $J(f)$ is a norm in the space $\text{span}\{\xi_1, \ldots \xi_q\}$ and $J(\boldsymbol{\xi}^T \mathbf{l}) = \mathbf{l}^T Q \mathbf{l}$, $L^T Q L = O$ implies $L^T \boldsymbol{\xi} = \mathbf{0}$, and, consequently, $L^T V_{\xi,\xi} = O$, $L^T V_{\xi,\phi} = O$, $L^T V_{\xi,\eta} = \mathbf{0}$, and $L^T \mu_\xi = \mathbf{0}$. Premultiply (7.7) by $\tilde{G}^{-T}$ and write $\tilde{\mathbf{c}} = \tilde{G} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix}$; straightforward algebra yields

$$\begin{pmatrix} I & O & O \\ O & I & O \\ O & O & O \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{c}}_1 \\ \tilde{\mathbf{c}}_2 \\ \tilde{\mathbf{c}}_3 \end{pmatrix} = \begin{pmatrix} * \\ * \\ \mathbf{0} \end{pmatrix}; \tag{7.25}$$

see Problem 7.8. This is the same exercise done in §3.5.3 leading up to (3.74) on page 89, and one may solve

$$\begin{pmatrix} G_1^T & O \\ G_2^T & \tilde{G}_3^T \end{pmatrix} \begin{pmatrix} G_1 & G_2 \\ O & \tilde{G}_3 \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \\ Q\mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \end{pmatrix},$$

which amounts to setting $\tilde{\mathbf{c}}_3 = \mathbf{0}$ in (7.25). In actual computation, one performs the Cholesky decomposition of H with pivoting, replaces the trailing O by δI with an appropriate δ, and proceeds as if H were nonsingular.

7.4.2 Numerical Integration

For the calculation of $\int_{\mathcal{X}} g(x)dx$, a quadrature/cubature is of the form $\sum_{i=1}^d w_i\, g(x_i)$, where x_i are the nodes and w_i are the associated weights; typically, one dimensional formulas are called quadratures and multidimensional ones are called cubatures. Within a family of formulas, the accuracy usually increases with the size d, along with the computational cost.

Certain methods are adaptive, attempting to achieve user-specified precision through sequential node addition guided by precision estimates. In our setting, $O(q^2)$ integrals involving the same $O(q)$ functions need to be calculated for each step of the Newton iteration, so formulas with fixed

nodes are actually more economical than the adaptive methods. Also, the H matrix is guaranteed to be non-negative definite with fixed nodes and positive weights.

In one dimension, a standard Gauss quadrature with d up to 200 is sufficient for our needs. The public domain FORTRAN subroutine `gaussq.f` archived at `http://www.netlib.org/go` can be used to generate the nodes and the weights.

On multidimensional cubes, product quadratures quickly become prohibitive. A system known as Smolyak algorithm has been developed in the literature for the derivation of efficient cubatures from univariate formulas. The efficiency of Smolyak cubatures is achieved by thinning out nodes from the product quadratures; some negative weights are introduced in the process. Some of the Smolyak cubatures can be found in Novak and Ritter (1996) and Petras (2001). A collection of public domain C routines are found in Knut Petras' SMOLPACK, which can be modified to return the nodes and the weights of Smolyak cubatures.

Smolyak cubatures are highly accurate with smooth integrands in general, but modifications are necessary for them to work in the current setting. Data for density estimation are typically away from the boundaries of the domain one specifies, but the placement of nodes in Smolyak cubatures is dense near the boundaries and sparse in the middle; gross errors result from such a misaligned resource allocation. To circumvent the problem, one may apply transformations on each coordinate of the cube to make the marginal data nearly uniformly distributed, then use the Smolyak formulas on the transformed domain.

To illustrate the strategy, consider integration on $\mathcal{X} = [0,1]^2$. One first estimates the marginal densities $f_1(x_{\langle 1\rangle})$ and $f_2(x_{\langle 2\rangle})$ with distribution functions F_1 and F_2; a bit oversmoothing does no harm for the purpose so one may use cross-validation with $\alpha = 2$. Transforming the domain by $\tilde{x}_{\langle 1\rangle} = F_1(x_{\langle 1\rangle})$ and $\tilde{x}_{\langle 2\rangle} = F_2(x_{\langle 2\rangle})$, the marginal observations are nearly uniformly distributed on the $\tilde{x}_{\langle 1\rangle}$ and $\tilde{x}_{\langle 2\rangle}$ scales. Let $(\tilde{x}_{i\langle 1\rangle}, \tilde{x}_{i\langle 2\rangle})$ be the Smolyak nodes and w_i be the associated weights, the integral

$$\int_{\mathcal{X}} g(x)dx = \int_0^1 \int_0^1 g\big(F_1^{-1}(\tilde{x}_{\langle 1\rangle}), F_2^{-1}(\tilde{x}_{\langle 1\rangle})\big) \frac{dx_{\langle 1\rangle}}{d\tilde{x}_{\langle 1\rangle}} \frac{dx_{\langle 2\rangle}}{d\tilde{x}_{\langle 2\rangle}} d\tilde{x}_{\langle 1\rangle} d\tilde{x}_{\langle 2\rangle}$$

can be approximated by

$$\sum_{i=1}^{d} \frac{w_i\, g\big(F_1^{-1}(\tilde{x}_{i\langle 1\rangle}), F_2^{-1}(\tilde{x}_{i\langle 2\rangle})\big)}{f_1\big(F_1^{-1}(\tilde{x}_{i\langle 1\rangle})\big) f_2\big(F_2^{-1}(\tilde{2}_{i\langle 2\rangle})\big)},$$

where $f_1\big(F_1^{-1}(\tilde{x}_{\langle 1\rangle})\big) = d\tilde{x}_{\langle 1\rangle}/dx_{\langle 1\rangle}$ and $f_2\big(F_2^{-1}(\tilde{x}_{\langle 2\rangle})\big) = d\tilde{x}_{\langle 2\rangle}/dx_{\langle 2\rangle}$.

An example of this is shown in Fig. 7.2, where the circles are 150 simulated observations and the filled dots are the nodes of the 449-point version of

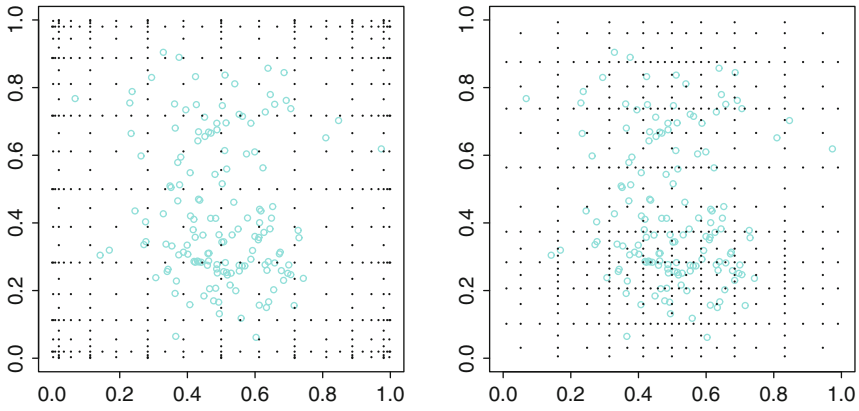


FIGURE 7.2. Smolyak cubature in two dimension. *Left*: Original scale. *Right*: Transformed scale. *Circles* are the data and *filled dots* are cubature nodes.

the so-called delayed Smolyak cubature in two dimension, on the original scale and on the transformed scale; the transformations are through the marginal density estimates based on the 150 observations.

7.4.3 Kullback-Leibler Projection

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its Kullback-Leibler projection $\tilde{\eta}$ in $\mathcal{H}_0$ minimizes

$$\mathrm{KL}(\hat{\eta}, \eta) = \mu_{\hat{\eta}}(\hat{\eta} - \eta) - \log \int_{\mathcal{X}} e^{\hat{\eta}} dx + \log \int_{\mathcal{X}} e^{\eta} dx$$

over $\eta \in \mathcal{H}_0$. Writing $A_{\tilde{\eta},g}(\alpha) = \mathrm{KL}(\hat{\eta}, \tilde{\eta} + \alpha g)$ for $g \in \mathcal{H}_0$, it is easy to verify that $0 = \dot{A}_{\tilde{\eta},g}(0) = \mu_{\tilde{\eta}}(g) - \mu_{\hat{\eta}}(g)$. It then follows, for $\eta_c \in \mathcal{H}_0$, that

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c).$$

One may take $\eta_c = 0$ as the uniform distribution on $\mathcal{X}$.

Unlike the projection in §5.3.2 for regression, this one is well-posed.

7.4.4 R Package `gss`: `ssden` Suite

Penalized likelihood density estimation is implemented in the `ssden` suite, whose usage shall be illustrated using a couple of synthetic examples. For density estimation in high dimensions, one should instead use the `ssden1` suite discussed in §10.1.5.

Example 7.3 ($\mathcal{X} = [0, 1]$) The following sequence generates a sample from (7.23) and fits a cubic spline to the log density, for λ minimizing $V(\lambda)$ of (7.21) with $\alpha = 1.4$:

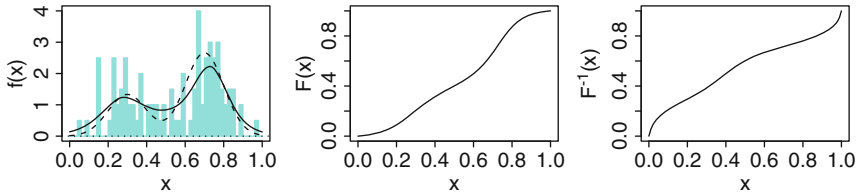


FIGURE 7.3. Density estimation on $\mathcal{X} = [0, 1]$. *Left*: Density estimate is in *solid line*, test density in *dashed line*, and data in finely binned histogram. *Center*: Cumulative distribution function $F(x)$. *Right*: Quantiles $F^{-1}(x)$.

```
rf1 <- function(n) {
  u <- runif(n); x0 <- rnorm(n)
  ifelse(u>2/3,x0/10+.3,x0/10+.7)
}
rtest1 <- function(n) {
  x <- rf1(n); ok <- (x>0)&(x<1)
  while(m<-sum(!ok)) {
    x[!ok] <- rf1(m); ok <- (x>0)&(x<1)
  }
  x
}
set.seed(5732); x <- rtest1(100)
fit <- ssden(~x,domain=data.frame(x=c(0,1)))
```

The domain $\mathcal{X}$ plays an active role in the estimation process as the density is normalized by $\int_{\mathcal{X}} e^{\eta} dx$, so it should be supplied by the user. A Gauss quadrature is used internally for the calculation of $\int_{\mathcal{X}} g(x)dx$. Shown in Fig. 7.3 are the estimated density along with the test density and the data, the cumulative distribution function, and the quantiles:

```
xx <- (0:100)/100
dtest1 <- function(x)
  (dnorm(x,.3,.1)/3+dnorm(x,.7,.1)*2/3)/.9986501
hist(x,breaks=(0:50)/50,border=5,col=5,prob=TRUE)
lines(xx,dssden(fit,xx))
lines(xx,dtest1(xx),lty=2)
plot(xx,pssden(fit,xx),type="l")
plot(xx,qssden(fit,xx),type="l")
```

`dssden` generally expects a data frame as input (like the `predict` function for `ssanova`) but does accept a vector in one-dimension, whereas `pssden` and `qssden` only work in one dimension and expect a vector. □

Example 7.4 ($\mathcal{X} = [0, 1]^3$) The following sequence generates a sample from (7.24) and fits a tensor product cubic spline to the log density:

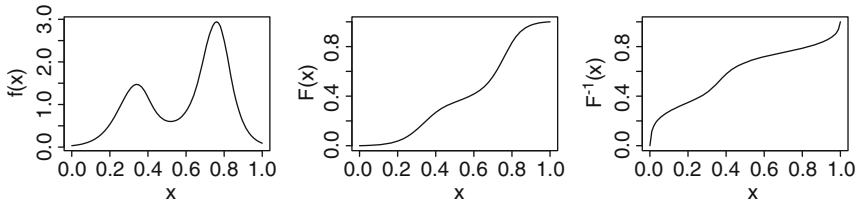


FIGURE 7.4. Density estimation on $\mathcal{X} = [0,1]^3$: Fitted conditional distribution $f(x_{\langle 1\rangle}|x_{\langle 2\rangle} = 0.5, x_{\langle 3\rangle} = 0.5)$. *Left*: Conditional density. *Center*: Conditional cumulative distribution function. *Right*: Quantiles of conditional distribution.

```
rtest3 <- function(n) {
  z <- .5+.2*rnorm(n)
  x <- rf1(n)-.1+.3*z; y <- rf1(n)-.1+.2*z
  ok <- (pmin(x,y,z)>0)&(pmax(x,y,z)<1)
  while(m<-sum(!ok)) {
    z[!ok] <- .5+.2*rnorm(m)
    x[!ok] <- rf1(m)-.1+.3*z[!ok]
    y[!ok] <- rf1(m)-.1+.2*z[!ok]
    ok <- (pmin(x,y,z)>0)&(pmax(x,y,z)<1)
  }
  cbind(x,y,z)
}
set.seed(5732); x <- rtest3(300)
x1 <- x[,1]; x2 <- x[,2]; x3 <- x[,3]; rg <- c(0,1)
my.domain <- data.frame(x1=rg,x2=rg,x3=rg)
fit <- ssden(~x1*x2*x3,domain=my.domain)
```

Three marginal densities are estimated internally to rescale the cube, and a Smolyak cubature is used on the rescaled cube for the calculation of $\int_{\mathcal{X}} g(x)dx$; see §7.4.2 for the strategy. A total of $3 + 3(3) + 7 = 19$ θ_β's are used in the fit, so the execution is a bit slow. The Kullback-Leibler projection suggests the elimination of the terms `x1:x2` and `x1:x2:x3`, and we refit without these terms:

```
project(fit,c("x1","x2","x3","x1:x3","x2:x3"))$ratio
# 0.01115107
fit <- ssden(~(x1+x2)*x3,domain=my.domain)
```

One may "slice out" the estimated density via conditional distributions, say $f(x_{\langle 1\rangle}|x_{\langle 2\rangle} = .5, x_{\langle 3\rangle} = .5)$, as shown in Fig. 7.4:

```
xx <- (0:100)/100; cond <- data.frame(x2=.5,x3=.5)
plot(xx,cdssden(fit,xx,cond=cond)$pdf,type="l")
plot(xx,cpssden(fit,xx,cond=cond),type="l")
plot(xx,cqssden(fit,xx,cond=cond),type="l")
```

where `cdssden` returns a list with elements `pdf` and `int`. □

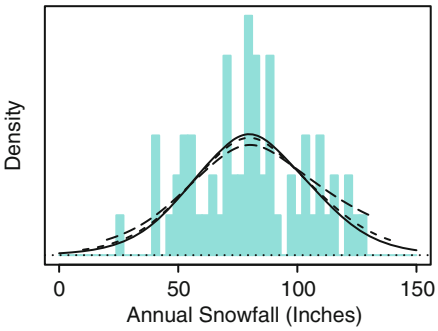


FIGURE 7.5. Distribution of Buffalo annual snowfall. The fits with $\mathcal{X} = [0, 150]$, $[10, 140]$, and $[20, 130]$ are in *solid*, *short-dashed*, and *long-dashed lines*, with the data superimposed as finely binned histogram.

7.5 Case Studies

We now apply the techniques developed so far to analyze a few real data sets. It will be seen that the specification of the domain $\mathcal{X}$ carries a rather heavy weight in the estimation process.

7.5.1 Buffalo Snowfall

The annual snowfall accumulations in Buffalo, New York from 1910 to 1973 are listed in Scott (1985), and are included in `gss` as a vector object `buffalo`. The data range from 25.0 to 126.4. To see how the domain $\mathcal{X}$ affects the estimate, three fits were calculated using $\mathcal{X} = [0, 150]$, $[10, 140]$, and $[20, 130]$, respectively:

```
data(buffalo)
fit.buf1 <- ssden(~buffalo,id.basis=1:63,
                  domain=data.frame(buffalo=c(0,150)))
fit.buf2 <- ssden(~buffalo,id.basis=1:63,
                  domain=data.frame(buffalo=c(10,140)))
fit.buf3 <- ssden(~buffalo,id.basis=1:63,
                  domain=data.frame(buffalo=c(20,130)))
```

where `id.basis=1:63` sets $q = n$ to take away the variability due to the selection of $\{Z_j\}$. The fits are shown in Fig. 7.5, along with the data as finely binned histogram:

```
hist(buffalo,breaks=(0:50)*3,border=5,col=5,prob=TRUE)
lines(0:150,dssden(fit.buf1,0:150),lty=1)
lines(10:140,dssden(fit.buf2,10:140),lty=2)
lines(20:130,dssden(fit.buf3,20:130),lty=5)
```

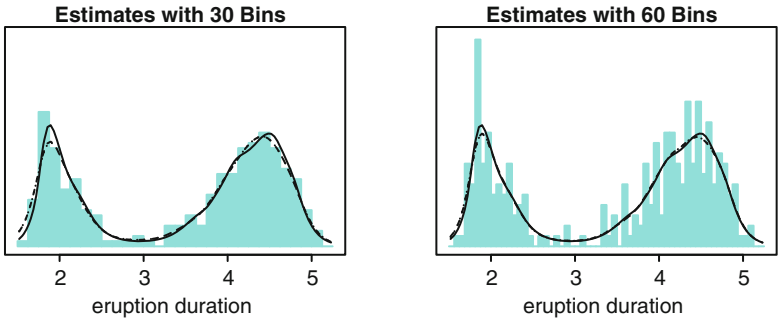


FIGURE 7.6. Density of eruption duration of Old Faithful. The fit based on the original data is in *solid lines*, those based on the histograms are in *dashed lines*, and those from Poisson regression are in *dotted lines*; the *dotted* and *dashed lines* coincide. The histograms are superimposed on the probability scale.

It is clear that as the domain $\mathcal{X}$ extends farther into the no-data area, the cross-validation tries harder to take away the mass assigned to the empty space by the smoothness of the estimates, resulting in less smoothing.

7.5.2 Eruption Time of Old Faithful

We now revisit the Old Faithful data discussed in §5.5.1:

```
data(faithful); erup <- faithful$eruptions
jk <- hist(erup,bre=seq(1.5,5.25,len=31),plot=FALSE)
x <- jk$mids; y <- jk$counts
```

Estimates using the original data and the binned data can be obtained, along with that using Poisson regression:

```
set.seed(5732)
fit.ori <- ssden(~erup,
                 domain=data.frame(erup=c(1.5,5.25)))
fit.bin <- ssden(~x,domain=data.frame(x=c(1.5,5.25)),
                 weights=y,subset=(y>0))
fit.poi <- gssanova(y~x,family="poisson")
```

The estimates can then be plotted along with the histogram, as shown in the left frame of Fig. 7.6:

```
xx <- ((1:100)-.5)/100*3.75+1.5
hist(erup,breaks=seq(1.5,5.25,length=31),
     prob=TRUE,border=5,col=5)
lines(xx,dssden(fit.ori,xx),lty=1)
lines(xx,dssden(fit.bin,xx),lty=2)
est <- predict(fit.poi,data.frame(x=xx))
lines(xx,exp(est)/sum(exp(est))*100/3.75,lty=3)
```

The estimate from Poisson regression is scaled into a probability density on $[1.5, 5.25]$, which coincides with the `ssden` fit using binned data. Parallel results using 60 bins are shown in the right frame.

7.5.3 AIDS Incubation

Details are in order concerning the AIDS incubation study discussed in §1.4.2. The data are included in `gss` as a data frame `aids` with elements `incu` (incubation time X), `infe` (time Y from infection to end of study), and `age`. Conditioning on the truncation mechanism, the density of (X, Y) is given by $f(x, y) = e^{\eta(x,y)} / \int_{\mathcal{T}} e^{\eta(x,y)} dxdy$, where $\mathcal{T} = \{x < y\}$.

The domain enters the estimation process only through the integrals $\int_{\mathcal{T}} g(x, y)dxdy$, so it is effectively specified via the quadrature. Lacking better alternatives, one may start with a crude rectangular grid on $[0, 100]^2$, eliminate points on $\{x > y\}$, and assign half weights along $\{x = y\}$:

```
qd.pt <- expand.grid(incu=2*(1:50)-1,infe=2*(1:50)-1)
qd.pt <- qd.pt[qd.pt$incu<=qd.pt$infe,]
qd.wt <- rep(1,nrow(qd.pt))
qd.wt[qd.pt$incu==qd.pt$infe] <- .5
qd.wt <- qd.wt/sum(qd.wt)*5e3
```

The following sequence loads the data, fits a tensor product cubic spline to log density, and checks for pretruncation independence:

```
data(aids); rg <- c(0,100); set.seed(5732)
fit.aids0 <- ssden(~incu*infe,data=aids,
                   domain=data.frame(incu=rg,infe=rg),
                   quad=list(pt=qd.pt,wt=qd.wt))
project(fit.aids0,c("incu","infe"))$ratio
# 0.01559929
```

One can then fit an additive model and plot, as shown in the bottom right frame of Fig. 7.7:

```
fit.aids <- ssden(~incu+infe,data=aids,
                  domain=data.frame(incu=rg,infe=rg),
                  quad=fit.aids0$quad,id=fit.aids0$id)
xx <- 2*(1:50)-1; grid <- expand.grid(incu=xx,infe=xx)
ff <- matrix(dssden(fit.aids,grid),50,50)
ff[outer(xx,xx,">")] <- NA
f.incu <- cdssden(fit.aids,xx,data.frame(infe=50))$pdf
f.infe <- cdssden(fit.aids,xx,data.frame(incu=50))$pdf
contour(xx,xx,log(ff)); lines(c(0,100),c(0,100),lty=2)
points(aids[,c("incu","infe")],col=3)
lines(xx,f.incu*1500); lines(100-f.infe*1500,xx,col=5)
```

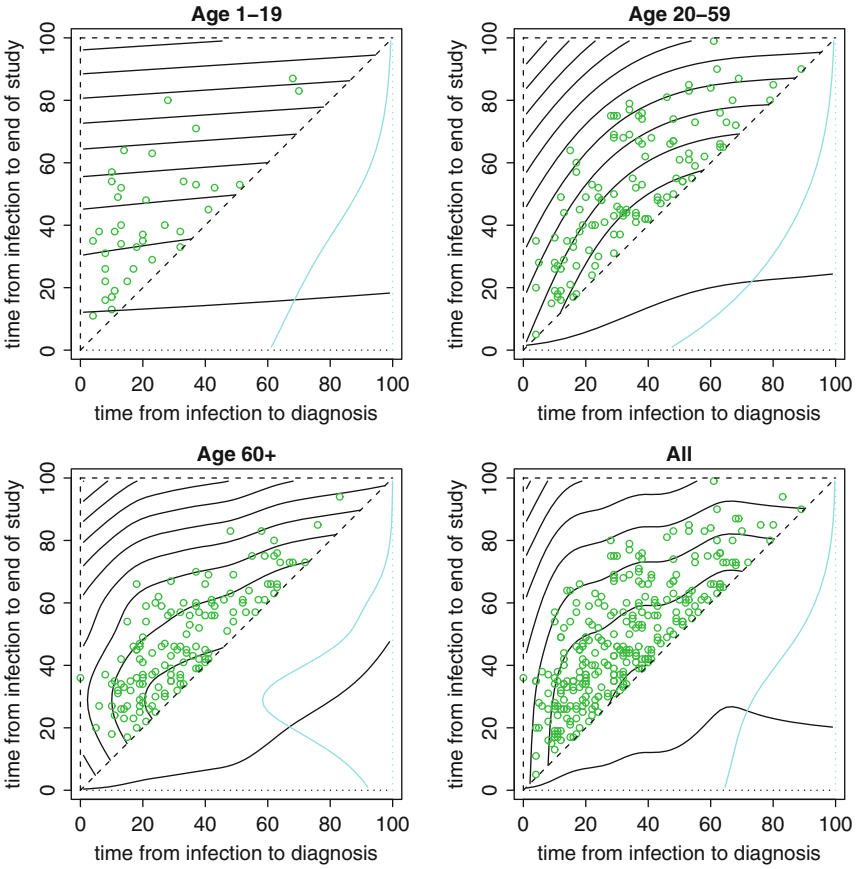


FIGURE 7.7. AIDS incubation and HIV infection: Fits with pretruncation independence. Contours are the fitted log density on the observable region surrounded by the *dashed lines*. *Circles* mark the observations. *Curves* over the *dotted lines* in the empty space are the fitted marginal densities.

Separate fits for the age groups as shown in the other frames of Fig. 7.7 can be obtained by adding a `subset` argument in the call to `ssden`, say `subset=(age>=60)` for the elderly.

Based on only 38 observations, the fit for the youth group is not to be taken too seriously. Due to the lack of information from the samples, $f(x)$ at the upper end and $f(y)$ at the lower end cannot be estimated accurately, and, indeed, the marginal estimates plotted near the lower-right corner demonstrate less consistency among different age groups. An interesting observation is the bump in $f(y)$ in the fit for the elderly, which appears to suggest that at the vicinity of January 1984 (30 months before July 1986), a batch of contaminated blood might have been distributed in the population from which the elderly data were collected.

7.6 Biased Sampling and Random Truncation

Independent and identically distributed samples may not always be available or may not be all that are available concerning the density $f(x)$. Biased sampling and random truncation are two sources from which non-*i.i.d.* samples may result.

A simple general formulation provides a unified framework for treating such data, and (7.1) can be easily modified to combine information from heterogeneous samples. The computation and smoothing parameter selection require only trivial modifications to the algorithms designed for (7.1). The empirical performance of cross-validation is explored via simple simulations, and the use of `ssden` under sampling bias is illustrated using simulated examples. The techniques can be used to estimate independent marginal densities of truncated data, allowing for an alternative analysis of the AIDS incubation data of §7.5.3.

7.6.1 Biased and Truncated Samples

Consider independent observations X_i on $\mathcal{X}$ sampled from densities proportional to $w_i(x)f(x)$, where $w_i(x) \geq 0$ are *known* biasing functions and $f(x)$ is to be estimated. Note that the data are actually the pairs (w_i, X_i). Let $\mathcal{T}$ be an index set and $w(t,x)$ a known function on $\mathcal{T} \times \mathcal{X}$ such that the set $\{w(t,\cdot), t \in \mathcal{T}\}$ includes all possible biasing functions and $w(t,\cdot) \neq w(t',\cdot)$ when $t \neq t'$. The "observed" biasing function w_i can then be written as $w(t_i,\cdot)$ for some $t_i \in \mathcal{T}$, and the data are now (t_i, X_i). Assume $0 < \int_{\mathcal{X}} w(t,x)f(x)dx < \infty$, $\forall t \in \mathcal{T}$, so that the densities $w(t,x)f(x)/\int_{\mathcal{X}} w(t,x)f(x)dx$ are well defined. Take t_i as observations from a probability density $m(t)$ on $\mathcal{T}$. The data (t_i, X_i) can then be treated as from a two-stage sampling.

Example 7.5 (Ordinary samples) Let $\mathcal{T} = \{1\}$ be a singleton and $w(1,x) = 1$. X_i are *i.i.d.* samples from $f(x)$. □

Example 7.6 (Length-biased samples) Let $\mathcal{T} = \{1\}$ be a singleton, $\mathcal{X} = [0,1]$, and $w(1,x) = x$. X_i are *i.i.d.* length-biased samples from the probability density $xf(x)/\int_0^1 xf(x)dx$. □

Example 7.7 (Ordinary and length-biased samples) Let $\mathcal{T}=\{1,2\}$, $\mathcal{X} = [0,1]$, $w(1,x) = 1$, and $w(2,x) = x$. $X_i|(t_i=1)$ are ordinary samples from $f(x)$ and $X_i|(t_i=2)$ are length-biased samples from $xf(x)/\int_0^1 xf(x)dx$. Examples 7.5 and 7.6 are special cases with $m(1) = 1$ and $m(1) = 0$, respectively. □

Example 7.8 (Finite-strata biased samples) Let $\mathcal{T} = \{1, \ldots, s\}$ and $\mathcal{X} = \bigcup_{t:m(t)>0} \{x : w(t,x) > 0\}$, where $w(t,x) \geq 0$ but otherwise arbitrary. $X_i|t_i$ are from the densities

$$\frac{w(t_i,x)f(x)}{\int_{\mathcal{X}} w(t_i,x)f(x)dx}.$$

Example 7.7 is a special case with $s = 2$. □

Example 7.9 (Truncated samples) Paired data (t, X) are generated from a joint density $g(t)f(x)$ on $\mathcal{T} \times \mathcal{X}$, but only those that fall on an observable region $A \subset \mathcal{T} \times \mathcal{X}$ are recorded and those that fall on A^c are lost. Of interest is the estimation of $f(x)$. It follows that $w(t,x) = I_{[(t,x)\in A]}$ and $m(t) \propto g(t) \int_{\mathcal{X}} I_{[(t,x)\in A]} f(x)dx$.

Note that t and X are interchangeable and that the truncation scheme is virtually arbitrary in this setting. The independence of t and X is necessary, for otherwise t would also carry information about $f(x)$.

For a specific case, consider $\mathcal{T} = \mathcal{X} = [0,1]$ and $A = \{t < x\}$. One has $w(t,x) = I_{[t<x]}$ and $m(t) \propto g(t) \int_t^1 f(x)dx$. □

7.6.2 Penalized Likelihood Estimation

Write $f(x) = e^{\eta(x)} / \int_{\mathcal{X}} e^{\eta(x)}dx$; the sampling likelihood of $X|t$ is seen to be

$$\frac{w(t,x)f(x)}{\int_{\mathcal{X}} w(t,x)f(x)dx} = \frac{w(t,x)e^{\eta(x)}}{\int_{\mathcal{X}} w(t,x)e^{\eta(x)}dx},$$

which leads to the penalized likelihood functional

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(X_i) - \log\int_{\mathcal{X}} w(t_i,x)e^{\eta(x)}dx\right\} + \frac{\lambda}{2}J(\eta). \tag{7.26}$$

For a singleton $\mathcal{T}$ such as the case with the length-biased samples of Example 7.6, (7.26) virtually reduces to (7.1) but with $\int_{\mathcal{X}} e^{\eta(x)}dx$ replaced by $\int_{\mathcal{X}} e^{\eta(x)}w(x)dx$, a substitution of the integration measure.

Removing dx from the notation and writing the integral as $\int_{\mathcal{X}} e^{\eta}$, (7.1) covers more ground than it first appears. Note that a probability density $f = e^{\eta}/\int_{\mathcal{X}} e^{\eta}$ is the Radon-Nikodym derivative of the probability measure with respect to a base measure, the integration measure that defines $\int_{\mathcal{X}} e^{\eta}$. By the chain rule of the Radon-Nikodym derivative, biased samples from $w(x)f(x)$ with respect to the uniform integration measure are simply ordinary samples from $f(x)$ with respect to the "biased" integration measure $\nu_w(A) = \int_A w(x)dx$. With such a change in notation, one no longer needs the domain $\mathcal{X}$ to be bounded, but only the integral $\int_{\mathcal{X}} 1$ over the domain to be finite so that the uniform distribution (with respect to the integration measure) is properly defined.

The minimizer of (7.26) in $\mathcal{H} = \{f : J(f) < \infty\}$ is generally not computable, but one again may calculate the efficient approximation in

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(Z_j, \cdot), j = 1, \ldots, q\};$$

see §9.2.5. Define

$$\mu_f(g|t) = \frac{\int_{\mathcal{X}} g(x)w(t,x)e^{f(x)}}{\int_{\mathcal{X}} w(t,x)e^{f(x)}}$$

and write $v_f(g, h|t) = \mu_f(gh|t) - \mu_f(g|t)\mu_f(h|t)$. Modify the definitions of $\mu_f(g)$ and $V_f(g, h)$ in §7.1 as

$$\mu_f(g) = \frac{1}{n}\sum_{i=1}^{n} \mu_f(g|t_i), \qquad V_f(g,h) = \frac{1}{n}\sum_{i=1}^{n} v_f(g,h|t_i). \tag{7.27}$$

The Newton updating formula (7.7) on page 241 holds verbatim for the minimization of (7.26) in $\mathcal{H}^*$, with the entries defined by the modified $\mu_f(g)$ and $V_f(g, h)$ (Problem 7.9).

Taking into account the sampling mechanism, the Kullback-Leibler distance of $e^{\eta_\lambda}/\int_{\mathcal{X}} e^{\eta_\lambda}$ from $e^{\eta}/\int_{\mathcal{X}} e^{\eta}$ should be modified as

$$\text{KL}(\eta, \eta_\lambda) = \int_{\mathcal{T}} m(t)\left\{\mu_\eta(\eta - \eta_\lambda|t) - \log\frac{\int_{\mathcal{X}} w(t,x)e^{\eta(x)}}{\int_{\mathcal{X}} w(t,x)e^{\eta_\lambda(x)}}\right\},$$

with the relative Kullback-Leibler distance

$$\text{RKL}(\eta, \eta_\lambda) = \int_{\mathcal{T}} m(t)\log\int_{\mathcal{X}} w(t,x)e^{\eta_\lambda(x)} - \int_{\mathcal{T}} m(t)\mu_\eta(\eta_\lambda|t). \tag{7.28}$$

The first term of (7.28) can be estimated by $n^{-1}\sum_{i=1}^{n}\log\int_{\mathcal{X}} w(t_i,x)e^{\eta_\lambda(x)}$. For the second term, $E\big[\eta_\lambda(X)\big]$, where X follows the marginal distribution under the sampling mechanism,

$$X \sim \int_{\mathcal{T}} m(t)\frac{w(t,x)e^{\eta(x)}}{\int_{\mathcal{X}} w(t,x)e^{\eta(x)}},$$

one may use the cross-validation estimate given by (7.20) on page 245, with the entries in the relevant matrices defined by the modified $\mu_f(g)$ and $V_f(g, h)$. The counterpart of (7.21) is easy to work out (Problem 7.10), and the computation following these lines can be accomplished via trivial modifications of the algorithms developed for (7.1).

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its Kullback-Leibler projection $\tilde{\eta}$ in $\mathcal{H}_0$ minimizes

$$\text{KL}(\hat{\eta}, \eta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mu_{\hat{\eta}}(\hat{\eta} - \eta|t_i) - \log\frac{\int_{\mathcal{X}} w(t_i,x)e^{\hat{\eta}(x)}}{\int_{\mathcal{X}} w(t_i,x)e^{\eta(x)}}\right\}.$$

For $\eta_c \in \mathcal{H}_0$, $\text{KL}(\hat{\eta}, \eta_c) = \text{KL}(\hat{\eta}, \tilde{\eta}) + \text{KL}(\tilde{\eta}, \eta_c)$.

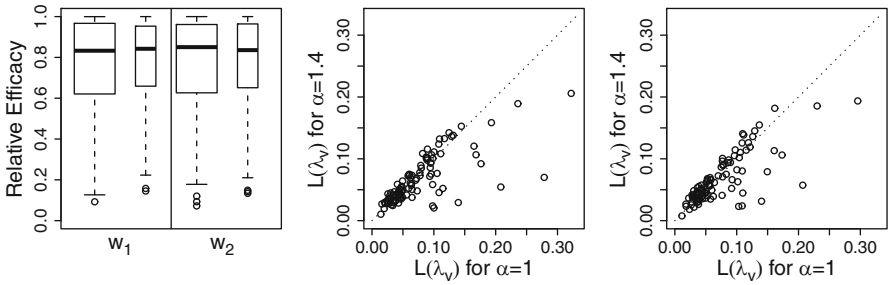


FIGURE 7.8. Effectiveness of cross-validation for density estimation under sampling bias. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $w_1(t,x) = x$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $w_2(t,x) = I_{[x>t]}$.

7.6.3 Empirical Performance

We now explore the empirical performance of the techniques outlined above through simple simulations. Samples (t_i, X_i) of size $n = 100$ were generated according to Example 7.9 with $A = \{t < x\}$, $g(t) = I_{[0<t<1]}$ uniform, and $f(x)$ as given in (7.23) on page 246. Note that the X_i thus generated are length-biased (Problem 7.11). Using the second formulation of cubic spline as discussed in Example 7.1 and setting $q = n$ in (7.3), estimates were calculated using two different biasing functions, $w_1(t,x) = x$ and $w_2(t,x) = I_{[x>t]}$; with w_1 one incorporates knowledge of $g(t)$ but discards t_i, whereas with w_2 one relies solely on the observed t_i.

For each replicate and each biasing function, three estimates were calculated, one minimizing the symmetrized Kullback-Leibler distance

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\mu_\eta(\eta - \eta_\lambda|t_i) + \mu_{\eta_\lambda}(\eta_\lambda - \eta|t_i)\right\},$$

another minimizing the duly modified $V(\lambda)$ (the counterpart of (7.21)) with $\alpha = 1$, and a third minimizing $V(\lambda)$ with $\alpha = 1.4$, yielding an optimal loss $L(\lambda_o)$ and two cross-validation losses $L(\lambda_v)$. The results from one hundred replicates are summarized in Fig. 7.8.

7.6.4 R Package `gss`: `ssden` Suite

Density estimation under sampling bias can be performed using `ssden` with an additional argument `bias`, which should be a list object with elements `t` ($\{t_k\} = \mathcal{T}$), `wt` ($m(t_k)$), and `fun` (biasing function $w(t,x)$); note that $\mathcal{T}$ is effectively discrete, t_i's do not need to be paired with X_i's, and only distinctive t_k's need to be listed.

The following function is modified from `rtest1` in Example 7.3, which generates truncated data (t_i, X_i) as in §7.6.3:

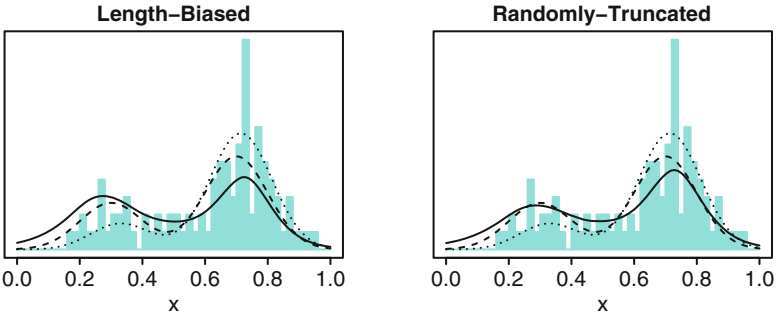


FIGURE 7.9. Density estimation under sampling bias. *Left*: Estimate using $w_1(t,x) = x$. *Right*: Estimate using $w_2(t,x) = I_{[x>t]}$. The estimates are in *solid lines*, the test density in *dashed lines*, the sampling density in *dotted lines*, and the data in finely binned histograms.

```
rtest.b <- function(n) {
  t <- runif(n); x <- rf1(n); ok <- (x>t)&(x<1)
  while(m<-sum(!ok)) {
    t[!ok] <- runif(m); x[!ok] <- rf1(m)
    ok <- (x>t)&(x<1)
  }
  cbind(x,t)
}
```

A sample of size $n = 100$ is generated, and $f(x)$ is estimated using biasing functions $w_1(t,x) = x$ and $w_2(t,x) = I_{[x>t]}$, respectively:

```
set.seed(5732); xt <- rtest.b(100)
x <- xt[,1]; t <- xt[,2]
bias1 <- list(t=1,wt=1,fun=function(t,x){x[,]})
fit1 <- ssden(~x,domain=list(x=c(0,1)),bias=bias1)
bias2 <- list(t=t,wt=rep(1/100,100),
              fun=function(t,x){x[,]>t})
fit2 <- ssden(~x,domain=list(x=c(0,1)),bias=bias2,
              id.basis=fit1$id.basis)
```

note that $\mathcal{T}$ is a singleton for w_1. The fit using w_1 can be plotted as in the left frame of Fig. 7.9, superimposed with the data, the test density $f(x)$ as given in (7.23), and the sampling density $\tilde{f}(x) \propto xf(x)$:

```
xx <- (0:100)/100
dtest <- function(x)
  (dnorm(x,.3,.1)/3+dnorm(x,.7,.1)*2/3)/.9986501
dtest.b <- function(x) dtest(x)*x/0.5665187
hist(x,breaks=(0:50)/50,border=5,col=5,prob=TRUE)
lines(xx,dssden(fit1,xx))
```

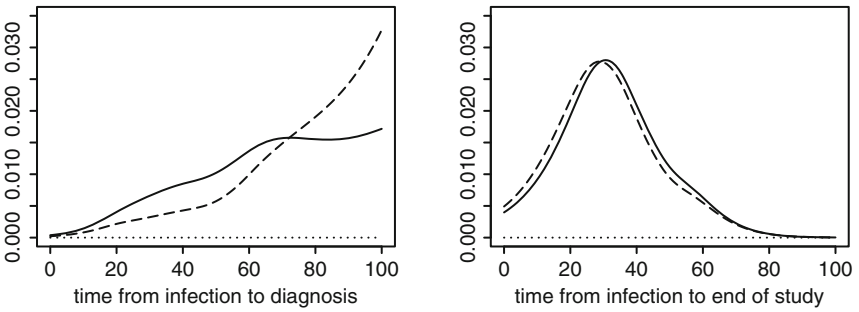


FIGURE 7.10. AIDS incubation and HIV infection for the elderly. *Left*: Incubation density $f(x)$ of X. *Right*: Infection density $f(y)$ of Y. The *solid lines* are the fits through (7.26). The *dashed lines* are taken from Fig. 7.7, *lower-left frame*.

```
lines(xx,dtest(xx),lty=2)
lines(xx,dtest.b(xx),lty=3)
```

Replacing `fit1` above by `fit2` yields the right frame.

7.6.5 Case Study: AIDS Incubation

We now apply the techniques developed in this section to the AIDS incubation data of §7.5.3. Assuming the independence of the incubation time X and the infection time Y, $f(x)$ can be estimated using $w(t,x) = I_{[x<t]}$ for $t = Y$ and $f(y)$ can be estimated using $w(t,y) = I_{[y>t]}$ for $t = X$.

The following sequence fits $f(x)$ and $f(y)$ for the elderly:

```
data(aids); n <- dim(aids)[1]; set.seed(5732)
bias.x <- list(t=aids$infe,wt=rep(1/n,n),
               fun=function(t,x){x[,]<t})
fit.x <- ssden(~incu,domain=data.frame(incu=c(0,100)),
               data=aids,subset=age>=60,bias=bias.x)
bias.y <- list(t=aids$incu,wt=rep(1/n,n),
               fun=function(t,y){y[,]>t})
fit.y <- ssden(~infe,domain=data.frame(infe=c(0,100)),
               data=aids,subset=age>=60,bias=bias.y,
               id.basis=fit.x$id.basis)
```

The estimated $f(x)$ is shown in the left frame of Fig. 7.10, with that from the joint estimation in §7.5.3 superimposed:

```
xx <- 0:100
plot(xx,dssden(fit.x,xx),type="l",ylim=c(0,.033))
f.incu <- cdssden(fit.aids,xx,data.frame(infe=50))$pdf
lines(xx,f.incu,lty=5)
```

where `fit.aids` is from §7.5.3 but with `subset=age>=60`. The right frame can be drawn in similar manner.

As can be seen in the lower-left frame of Fig. 7.7, information from data is scarce on the upper end of $f(x)$. The estimates appear to agree well, especially those of $f(y)$.

7.7 Conditional Densities

On a product domain $\mathcal{X} \times \mathcal{Y}$, the primary interest is often the estimation of the conditional density $f(y|x)$. Such a problem is typically known as regression, but unlike the formulations of Chaps. 3 and 5, no parametric assumption is made here on a generic $\mathcal{Y}$ axis, and the function to be estimated is "bivariate" in (x, y) instead of "univariate" only in x.

A logistic conditional density transform can be made one-to-one through side conditions on the $\mathcal{Y}$ axis, with which the penalized likelihood estimation is straightforward. The computation and smoothing parameter selection follow trivial modifications of the procedures developed for (7.1). For $\mathcal{Y}$ continuous, the empirical performance of cross-validation is assessed via simple simulation and software tools are illustrated using simulated and real-data examples.

When n is large or when $\mathcal{Y}$ involves multidimensional continuous domains, the high cost of numerical integration can cripple the computation, and one instead may have to use the penalized pseudo likelihood of §10.3 that trades statistical performance for computational efficiency.

For $\mathcal{Y}$ discrete, the approach leads to regression with cross-classified responses. Numerical integration is a non issue in such a setting, but a different set of modeling tools are needed, to be developed in §7.8.

7.7.1 Penalized Likelihood Estimation

Consider independent observations (X_i, Y_i) on a product domain $\mathcal{X} \times \mathcal{Y}$ from a density $f(x, y) = f(x)f(y|x)$. Of interest is the estimation of the conditional density $f(y|x) = f(x, y)/\int_{\mathcal{Y}} f(x, y)$ of Y given X. Since the marginal density $f(x)$ of X is only a nuisance parameter, the sampling of X_i can actually be arbitrary, random or deterministic, so long as $Y|X \sim f(y|x)$. For notational convenience, however, $f(x)$ will still be used to denote the "limiting distribution" of X_i's, even when they are deterministic.

The logistic conditional density transform, $f(y|x) = e^{\eta(x,y)}/\int_{\mathcal{Y}} e^{\eta(x,y)}$, can be employed to enforce the positivity and unity constraints. To make the transform one-to-one, $\eta(x, y)$ has to satisfy certain side conditions, say $A_y\eta(x, y) = 0$, $\forall x \in \mathcal{X}$, where the averaging operator A_y on the domain $\mathcal{Y}$ can, in principal, depend on x. A simple approach to achieving a one-to-one logistic conditional density transform is through term elimination in an ANOVA decomposition, as discussed in §1.3.2: For $\eta(x, y) = \eta_\emptyset + \eta_x +$

$\eta_y + \eta_{x,y}$ with averaging operators A_x and A_y,

$$f(y|x) = \frac{e^{\eta_\emptyset + \eta_x + \eta_y + \eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_\emptyset + \eta_x + \eta_y + \eta_{x,y}}} = \frac{e^{\eta_y + \eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{x,y}}}, \tag{7.29}$$

where $A_y(\eta_y + \eta_{x,y}) = 0$, $\forall x \in \mathcal{X}$; the side condition here is independent of x. Eliminating $\eta_\emptyset + \eta_x$ from $\eta(x,y)$, one may estimate $f(y|x) = e^{\eta(x,y)} / \int_{\mathcal{Y}} e^{\eta(x,y)}$ via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(X_i,Y_i) - \log\int_{\mathcal{Y}} e^{\eta(X_i,y)}\right\} + \frac{\lambda}{2}J(\eta) \tag{7.30}$$

in an appropriately assembled tensor product reproducing kernel Hilbert space.

Example 7.10 (Tensor product cubic spline) Consider $\mathcal{X}$=$[0,1]$ and $\mathcal{Y} = [0,1]$. Use the construction of Example 2.5 on page 44, with (x,y) replacing $(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ in the notation. Eliminating $\eta_\emptyset$ and η_x, one has the space

$$\mathcal{H} = \mathcal{H}_{00\langle x\rangle} \otimes (\mathcal{H}_{01\langle y\rangle} \oplus \mathcal{H}_{1\langle y\rangle}) \oplus (\mathcal{H}_{01\langle x\rangle} \oplus \mathcal{H}_{1\langle x\rangle}) \otimes (\mathcal{H}_{01\langle y\rangle} \oplus \mathcal{H}_{1\langle y\rangle}).$$

In the notation of Example 2.8, one may set

$$J(f,g) = \theta_{00,1}^{-1}(f,g)_{00,1} + \theta_{01,1}^{-1}(f,g)_{01,1} + \theta_{1,01}^{-1}(f,g)_{1,01} + \theta_{1,1}^{-1}(f,g)_{1,1},$$

which has the null space $\mathcal{N}_J = (\mathcal{H}_{00\langle x\rangle} \otimes \mathcal{H}_{01\langle y\rangle}) \oplus (\mathcal{H}_{01\langle x\rangle} \otimes \mathcal{H}_{01\langle y\rangle})$ spanned by $\phi_1 = y - 0.5$ and $\phi_2 = (x-0.5)(y-0.5)$, and the reproducing kernel

$$R_J = \theta_{00,1}R_{00,1} + \theta_{01,1}R_{01,1} + \theta_{1,01}R_{1,01} + \theta_{1,1}R_{1,1}.$$

Clearly, one has $\int_0^1 \eta(x,y)dy = 0$, $\forall x \in [0,1]$, for $\eta \in \mathcal{H}$. □

The minimizer of (7.30) in $\mathcal{H} = \{f : J(f) < \infty\}$ is generally not computable, but one may calculate an efficient approximation in

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\big\{R_J(V_j, \cdot), j = 1, \ldots, q\big\}$$

for $\{V_j\} \subseteq \big\{(X_i, Y_i)\big\}$ a random subset, which shares the same asymptotic convergence rates; see §9.2.6. Now, define $\mu_f(g|x) = \int_{\mathcal{Y}} ge^f / \int_{\mathcal{Y}} e^f$ and $v_f(g,h|x) = \mu_f(gh|x) - \mu_f(g|x)\mu_f(h|x)$. The Newton updating formula (7.7) on page 241 again holds verbatim for the minimization of (7.30) in $\mathcal{H}^*$, with $\mu_f(g)$ and $V_f(g,h)$ modified as follows,

$$\mu_f(g) = \frac{1}{n}\sum_{i=1}^{n} \mu_f(g|X_i), \qquad V_f(g,h) = \frac{1}{n}\sum_{i=1}^{n} v_f(g,h|X_i); \tag{7.31}$$

see Problem 7.12.

Weighted by the sampling proportion $f(x)$, the aggregated Kullback-Leibler distance of $f_\lambda(y|x) = e^{\eta_\lambda}/\int_{\mathcal{Y}} e^{\eta_\lambda}$ from $f(y|x) = e^{\eta}/\int_{\mathcal{Y}} e^{\eta}$ is now

$$\mathrm{KL}(\eta, \eta_\lambda) = \int_{\mathcal{X}} f(x)\Big\{\mu_\eta(\eta - \eta_\lambda|x) - \log\int_{\mathcal{Y}} e^{\eta} + \log\int_{\mathcal{Y}} e^{\eta_\lambda}\Big\}, \tag{7.32}$$

with the relative Kullback-Leibler distance

$$\mathrm{RKL}(\eta, \eta_\lambda) = \int_{\mathcal{X}} f(x)\log\int_{\mathcal{Y}} e^{\eta_\lambda} - \int_{\mathcal{X}} f(x)\mu_\eta(\eta_\lambda|x). \tag{7.33}$$

The first term of (7.33) can be estimated by $n^{-1}\sum_{i=1}^{n}\log\int_{\mathcal{Y}} e^{\eta_\lambda(X_i,y)}$; the second term $E\big[\eta_\lambda(X,Y)\big]$, where $(X,Y) \sim f(x)f(y|x) = f(x,y)$, can be estimated by $n^{-1}\sum_{i=1}^{n}\eta^{[i]}_{\lambda,\tilde{\eta}}(X_i,Y_i)$, which is given by (7.20) on page 245 with the entries in the relevant matrices defined by the modified $\mu_f(g)$ and $V_f(g,h)$. Parallel derivation yields the same cross-validation score $V(\lambda)$ of (7.21) but with the modified $\mu_f(g)$ and $V_f(g,h)$; see Problem 7.13.

While the formulas readily carry over from density estimation to conditional density estimation, the computation here can be prohibitive. The calculations of $\mu_f(g)$ and $V_f(g,h)$ as defined in §7.1 take $O(d)$ flops, where d is the quadrature size. The calculations of $\mu_f(g)$ and $V_f(g,h)$ as defined in (7.31) would in general take $O(nd)$ flops, however, unless X_i's are heavily duplicated. One nevertheless could trade statistical performance for computational efficiency/feasibility via an alternative treatment; see §10.3.

7.7.2 Empirical Performance of Cross-validation

Consider the test distribution on $\mathcal{X} = [0,1]$ and $\mathcal{Y} = [0,1]$,

$$f(y|x) \propto \phi\big((y-\mu_x)/\sigma_x\big)I_{[0<y<1]}, \tag{7.34}$$

where $\mu_x = x^3 - x^2 + x - 0.2$, $\sigma_x = 0.3$, and $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ is the standard normal density. Samples of size $n = 200$ were drawn with X_i on the grid $0.005(0.01)0.995$, two each. The tensor product cubic spline of Example 7.10 were used, and for each replicate, three fits were calculated, minimizing respectively the symmetrized Kullback-Leibler distance

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\big\{\mu_\eta(\eta - \eta_\lambda|X_i) + \mu_{\eta_\lambda}(\eta_\lambda - \eta|X_i)\big\}$$

and the cross-validation score $V(\lambda)$ with $\alpha = 1, 1.4$. The optimal $L(\lambda_o)$ and the two cross-validation $L(\lambda_v)$'s from one hundred replicates are summarized in Fig. 7.11, in the left half of the left frame and in the center frame.

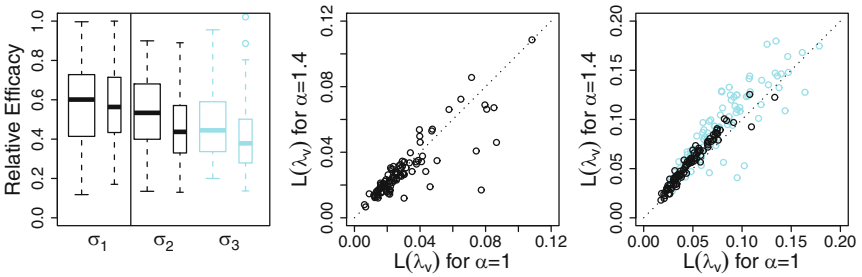


FIGURE 7.11. Effectiveness of cross-validation for conditional density estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*); $\sigma_1 = 0.3$, $\sigma_2 = 0.15(1 + x)$, $\sigma_3 = 0.15(2 - x)$. *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\sigma_x = 0.3$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\sigma_x = 0.15(1 + x)$ (*solid*) and $\sigma_x = 0.15(2 - x)$ (*faded*).

The experiments were repeated for two modified standard deviation functions, $\sigma_x = 0.15(1 + x)$ and $\sigma_x = 0.15(2 - x)$, respectively, with results from one hundred replicates also summarized in Fig. 7.11, in the right half of the left frame and in the right frame.

The relative efficacy is much worse than what we have seen so far in other settings, likely due to the more difficult task at hand; note that one only has two Y's per X in the simulated samples for the estimation of $f(y|x)$. The comparison of $\alpha = 1$ versus $\alpha = 1.4$ varies with the test distribution, but $\alpha = 1.4$ appears to be the safer choice.

7.7.3 Kullback-Leibler Projection

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its Kullback-Leibler projection $\tilde{\eta}$ in $\mathcal{H}_0$ minimizes

$$\mathrm{KL}(\hat{\eta}, \eta) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \mu_{\hat{\eta}}(\hat{\eta} - \eta | X_i) - \log \int_{\mathcal{Y}} e^{\hat{\eta}(X_i, y)} + \log \int_{\mathcal{Y}} e^{\eta(X_i, y)} \right\},$$

over $\eta \in \mathcal{H}_0$. Writing $A_{\tilde{\eta},g}(\alpha) = \mathrm{KL}(\hat{\eta}, \tilde{\eta} + \alpha g)$ for $g \in \mathcal{H}_0$, it is easy to verify that $0 = \dot{A}_{\tilde{\eta},g}(0) = \mu_{\tilde{\eta}}(g) - \mu_{\hat{\eta}}(g)$. It then follows, for $\eta_c \in \mathcal{H}_0$, that

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c).$$

One may take $\eta_c = \eta_y(y) = \eta_1(y_{\langle 1 \rangle}) + \cdots + \eta_\Gamma(y_{\langle \Gamma \rangle})$, where $\mathcal{Y} = \prod_{\gamma=1}^{\Gamma} \mathcal{Y}_\gamma$, with $Y_{\langle \gamma \rangle}$ independent of X and of each other.

7.7.4 R Package `gss`: `sscden` Suite

The `sscden` suite in `gss` implements the penalized likelihood conditional density estimation of (7.30) with (part of) $\mathcal{Y}$ continuous. For n large or with $\mathcal{Y}$ involving multidimensional continuous marginals, one should consider the `sscden1` suite (§10.3.4) instead, which runs much faster though

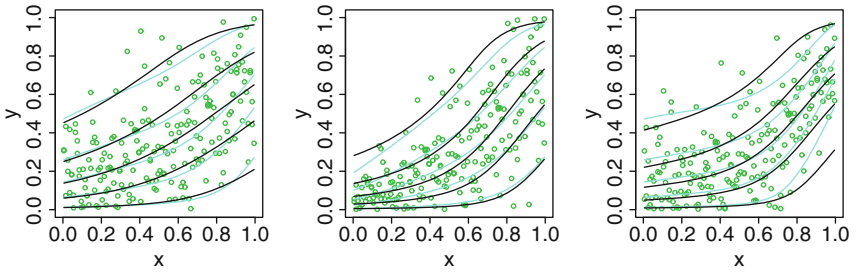


FIGURE 7.12. Conditional density estimation on $\mathcal{X} = [0,1]$ and $\mathcal{Y} = [0,1]$. The 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$ are in *solid lines*, those of the test distributions in *faded lines*, and the data in *circles*. From *left* to *right*: $\sigma_x = 0.3$, $0.15(1+x)$, $0.15(2-x)$.

generally returns worse-performing estimates. With $\mathcal{Y}$ discrete, one should use the `ssllrm` suite (§7.8.6) for regression with cross-classified responses.

The following sequence draws a sample from (7.34) with $\sigma_x = 0.15(1+x)$ and fits a tensor product cubic spline to the log conditional density, with smoothing parameters minimizing $V(\lambda)$ with $\alpha = 1.4$:

```
rfc2 <- function(x) {
  mu <- x^3-x^2+x-.2; sd=.15*(1+x)
  y <- (rnorm(length(x))*sd+mu)
  ok <- (y>0)&(y<1)
  while(m <- sum(!ok)) {
    y[!ok] <- (rnorm(m)*sd[!ok]+mu[!ok])
    ok <- (y>0)&(y<1)
  }
  y
}
xx <- ((1:100)-.5)/100; x <- rep(xx,2)
set.seed(5732); y <- rfc2(x)
fit <- sscden(~x*y,~y,ydomain=data.frame(y=c(0,1)))
```

where the first formula `~x*y` specifies model terms in the log conditional density and the second formula `~y` lists the y-variables; terms not involving y-variables are removed internally. The domain $\mathcal{Y}$ affects the normalization of the conditional density via $\int_{\mathcal{Y}} e^{\eta} dy$, which should be supplied through `ydomain`. A Gauss quadrature is used internally on an univariate $\mathcal{Y}$ for the calculation of $\int_{\mathcal{Y}} g(x,y)dy$.

Shown in the center frame of Fig. 7.12 are the 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$, with the data superimposed:

```
quan <- qsscden(fit,c(.05,.25,.5,.75,.95),
                data.frame(x=xx))
plot(x,y,col=3); for (i in 1:5) lines(xx,quan[i,])
```

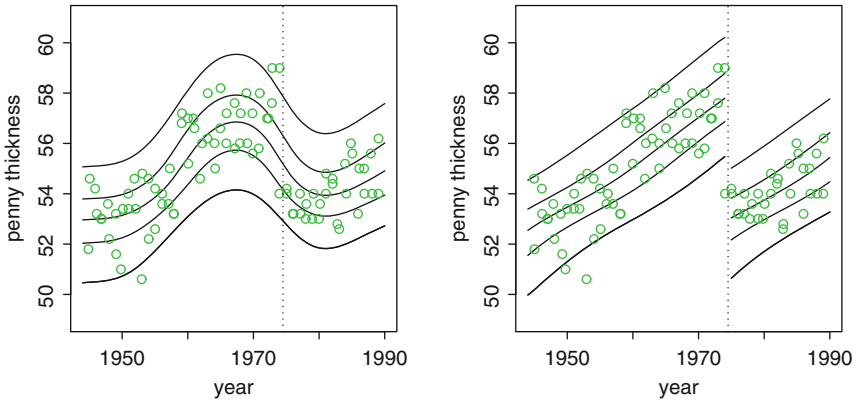


FIGURE 7.13. Thickness of U.S. Lincoln pennies. *Left*: Continuous fit. *Right*: Fit with built-in break. The *lines* are the 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$. The data, with the year jittered, are superimposed in *circles*. The *vertical dotted lines* mark the position of the break.

Also superimposed are the respective percentiles of the test distribution. Parallel results with $\sigma_x = 0.3$ and $\sigma_x = 0.15(2 - x)$ are shown in the left and the right frames, respectively.

7.7.5 Case Study: Penny Thickness

The thickness in mils of a sample of 90 U.S. Lincoln pennies is listed in Scott (1992, Appendix B.4). Two pennies from each year between 1945 and 1989 were measured. The data are included in `gss` as a data frame `penny` with elements `year` and `mil`, with the latter ranging between 50.6 and 59.

The following sequence loads the data, fits a tensor product cubic spline to the log conditional density, and plots the fit as shown in the left frame of Fig. 7.13, where the data, with the variable `year` slightly jittered to avoid overlap, are superimposed:

```
data(penny); set.seed(5732)
fit <- sscden(~year*mil,~mil,data=penny,
              ydomain=data.frame(mil=c(49,61)))
yy <- 1944+(0:92)/2
quan <- qsscden(fit,c(.05,.25,.5,.75,.95),
                data.frame(year=yy))
plot(penny$year+.1*rnorm(90),penny$mil,ylim=c(49,61))
for (i in 1:5) lines(yy,quan[i,])
```

The data show an abrupt downward shift of penny thickness from 1974 to 1975, perhaps due to equipment recalibration or the like at the time. To accommodate such discontinuity in the estimation, one may add to x a binary factor, with the result shown in the right frame of Fig. 7.13:

```
z <- factor(penny$year>1974.5); set.seed(5732)
fit1 <- sscden(~(year+z)*mil,~mil,data=penny,
               ydomain=data.frame(mil=c(49,61)))
yy <- 1944+(0:92)/2; zz <- factor(yy>1974.5)
quan1 <- qsscden(fit1,c(.05,.25,.5,.75,.95),
                 data.frame(year=yy,z=zz))
quan1[,yy==1974.5] <- NA
plot(penny$year+.1*rnorm(90),penny$mil,ylim=c(49,61))
for (i in 1:5) lines(yy,quan1[i,])
abline(v=1974.5,lty=3)
```

Apart from the downward shift from 1974 to 1975, the pennies were getting thicker steadily.

7.8 Regression with Cross-Classified Responses

For $\mathcal{Y} = \prod_{\gamma=1}^{\Gamma} \mathcal{Y}_\gamma$ with $\mathcal{Y}_\gamma$'s discrete, the conditional density estimation of §7.7 provides a means to regression with cross-classified responses. Beyond the standard developments of §§7.7.1 and 7.7.3, further modeling tools are available in the setting.

When $\mathcal{Y} = \{0, 1\}$, the method reduces to the logistic regression of Example 5.2, so it is an extension of logistic regression to general discrete responses. The association between y-variables can be characterized via odds ratios, for which some modeling options are briefly discussed. Bayesian confidence intervals can be developed for contrasts of $\log f(y|x)$ among "levels" of y and random effects can be included to accommodate correlated data. Empirical performances are explored through simple simulations and software tools are illustrated using simulated and real-data examples.

7.8.1 Logistic Regression

Set $\mathcal{Y} = \{0, 1\}$. A reproducing kernel Hilbert space $\mathcal{H}_{\langle y \rangle}$ on $\mathcal{Y}$ is the Euclidean space with a reproducing kernel $R_{\langle y \rangle}(y_1, y_2) = I_{[y_1 = y_2]}$, which can be decomposed, with ANOVA implications, as $R_{0\langle y \rangle} + R_{1\langle y \rangle} = I_{[y_1 = y_2 = 0]} + I_{[y_1 = y_2 = 1]}$ or $R_{0\langle y \rangle} + R_{1\langle y \rangle} = 1/2 + (I_{[y_1 = y_2]} - 1/2)$; both define an ANOVA decomposition $\eta = \eta_\emptyset + \eta_y$, with the former implying an averaging operator $A_y \eta(y) = \eta(0)$ and $\eta_y(0) = 0$, and the latter $A_y \eta(y) = \big(\eta(0) + \eta(1)\big)/2$ and $\eta_y(1) = -\eta_y(0)$. Taking tensor product with a reproducing kernel Hilbert space $\mathcal{H}_{\langle x \rangle}$ on $\mathcal{X}$ with an square (semi) norm $J_{\langle x \rangle}(\eta)$, the corresponding square (semi) norm in $\mathcal{H}_{\langle x \rangle} \otimes \mathcal{H}_{1\langle y \rangle}$ is given by $J(\eta) = J_{\langle x \rangle}\big(\eta(x, 1)\big)$ for either decompositions of $R_{\langle y \rangle}$. Write $\int_{\mathcal{Y}} g = g(0) + g(1)$.

For $R_{1\langle y\rangle} = I_{[y_1=y_2=1]}$ with $\eta(x,0) = 0$, (7.30) becomes

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{I_{[Y_i=1]}\tilde{\eta}(X_i) - \log\left(1+e^{\tilde{\eta}(X_i)}\right)\right\} + \frac{\lambda}{2}J_{\langle x\rangle}(\tilde{\eta}), \tag{7.35}$$

with $\tilde{\eta}(x) = \eta(x,1)$, which is the standard form of penalized likelihood logistic regression; see Problem 7.14.

For $R_{1\langle y\rangle} = I_{[y_1=y_2]} - 1/2$ with $\eta(x,1) = -\eta(x,0)$,

$$\eta(x,y) - \log\left(e^{\eta(x,1)} + e^{\eta(x,0)}\right) = 2\eta(x,1)I_{[y=1]} - \log\left(1+e^{2\eta(x,1)}\right),$$

so (7.30) becomes, for $\tilde{\eta}(x) = 2\eta(x,1)$,

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{I_{[Y_i=1]}\tilde{\eta}(X_i) - \log\left(1+e^{\tilde{\eta}(X_i)}\right)\right\} + \frac{\lambda}{8}J_{\langle x\rangle}(\tilde{\eta}),$$

which is the same as (7.35) since $\lambda > 0$ has yet to be selected.

Now let us look at cross-validation. With $\eta(x,0) = 0$,

$$\log\textstyle\int_{\mathcal{Y}} e^{\eta_\lambda(X_i,y)} = \log\left(1+e^{\eta_\lambda(X_i,1)}\right),$$
$$\eta_\lambda^{[i]}(X_i.Y_i) = I_{[Y_i=1]}\eta_\lambda^{[i]}(X_i,1),$$

so the relative Kullback-Leibler distance of (7.33) is estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(1+e^{\tilde{\eta}_\lambda(X_i)}\right) - \tilde{Y}_i\tilde{\eta}_\lambda(X_i)\right\} + \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i\left(\tilde{\eta}_\lambda(X_i) - \tilde{\eta}_\lambda^{[i]}(X_i)\right), \tag{7.36}$$

where $\tilde{\eta}(x) = \eta(x,1)$ and $\tilde{Y} = I_{[Y=1]}$; this is simply (5.11) on page 182. For $\eta(x,1) = -\eta(x,0)$,

$$\log\textstyle\int_{\mathcal{Y}} e^{\eta_\lambda(X_i,y)} = \log\left(e^{2\eta_\lambda(X_i,1)}+1\right) - \eta_\lambda(X_i,1),$$
$$\eta_\lambda^{[i]}(X_i,Y_i) = (2\tilde{Y}_i - 1)\eta_\lambda^{[i]}(X_i,1),$$

and (7.36) changes slightly to

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\log\left(1+e^{\tilde{\eta}_\lambda(X_i)}\right) - \tilde{Y}_i\tilde{\eta}_\lambda(X_i)\right\} + \frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_i - 0.5)\left(\tilde{\eta}_\lambda(X_i) - \tilde{\eta}_\lambda^{[i]}(X_i)\right),$$

where $\tilde{\eta}(x) = 2\eta(x,1)$; instead of the $\sum_i \mu(x_i)\eta_\lambda(x_i)$ appearing in (5.8) on page 182, one now estimates $\sum_i \big(\mu(x_i) - 0.5\big)\eta_\lambda(x_i)$.

Note that Proposition 7.3 does not apply here, as the marginal configurations of tensor product reproducing kernel Hilbert spaces affect more than just a constant. For symmetry, we shall use the averaging operator $A_\gamma\eta = \frac{1}{K_\gamma+1}\sum_{y_{\langle\gamma\rangle}=0}^{K_\gamma}\eta(y_{\langle\gamma\rangle})$ on $\mathcal{Y}_\gamma = \{0,\ldots,K_\gamma\}$ in the rest of the discussion; see Problem 7.15 for the construction of tensor product spaces with such a y-marginal. It then follows that $\int_{\mathcal{Y}}\eta(x,y) = 0$, where $\int_{\mathcal{Y}} f(y)$ is the summation of $f(y)$ over $y \in \mathcal{Y}$.

7.8.2 Log-Linear Regression Models

When x is absent, data on $\mathcal{Y} = \prod_{\gamma=1}^{\Gamma} \mathcal{Y}_\gamma$ are typically aggregated into contingency tables, for which the log-linear models are among standard analytical tools. The conditional density models add an x-axis to the log-linear models to disaggregate contingency tables, and will be referred to as log-linear regression models.

The log-linear model for an $(K_1+1)\times\cdots\times(K_\Gamma+1)$ table is a surrogate Poisson regression model on $\mathcal{Y} = \prod_{\gamma=1}^{\Gamma} \mathcal{Y}_\gamma$ for $\mathcal{Y}_\gamma = \{0,\ldots,K_\gamma\}$, which is equivalent to density estimation on $\mathcal{Y}$. Associations between the margins of contingency tables are typically characterized via the log odds ratios. Take a 2×2 table for example, with $\mathcal{Y} = \{0,1\}^2$ and $f(y) = e^{\eta_y}/\int_{\mathcal{Y}} e^{\eta_y}$ for $\eta_y(y) = \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{1,2}(y_{\langle 1\rangle}, y_{\langle 2\rangle})$. One has

$$\log \frac{f(1,1)f(0,0)}{f(1,0)f(0,1)} = \eta_{1,2}(1,1) - \eta_{1,2}(1,0) - \eta_{1,2}(0,1) + \eta_{1,2}(0,0),$$

thus the log odds ratio only depends on the interaction η_{12}.

Adding an x-axis, $f(y|x) = e^{\eta_y+\eta_{x,y}}/\int_{\mathcal{Y}} e^{\eta_y+\eta_{x,y}}$, where η_y is as above and $\eta_{x,y}(x,y) = \eta_{x,1}(x,y_{\langle 1\rangle}) + \eta_{x,2}(x,y_{\langle 2\rangle}) + \eta_{x,1,2}(x,y_{\langle 1\rangle},y_{\langle 2\rangle})$. The log odds ratio depends only on $\eta_{1,2}+\eta_{x,1,2}$. If $\eta_{x,1,2}=0$, the odds ratio is independent of x, with the model sitting in between the "saturated" model and the conditional independence model $(Y_{\langle 1\rangle} \perp Y_{\langle 2\rangle})|X$ with $\eta_{1,2}+\eta_{x,1,2}=0$.

7.8.3 Bayesian Confidence Intervals for y-Contrasts

As discussed in §5.3.1, one may derive approximate Bayesian confidence intervals for $\eta(x,y)$ based on the quadratic approximation of the log likelihood at η_λ, but such intervals are of little use here as $e^{\eta(x,y)}$ needs to be normalized to assume any meaning. Of interest are the y-contrasts of $\eta(x,y)$ over "levels" of y at fixed x values, for which the normalizing constant cancels out; the log odds ratios are y-contrasts of $\eta(x,y)$.

Write $\eta = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c} = \boldsymbol{\psi}^T\mathbf{a}$ as in (7.3) and refer η and $(\mathbf{d}^T,\mathbf{c}^T)^T = \mathbf{a}$ interchangeably. The quadratic approximation of (7.30) at $\tilde{\eta} = \eta_\lambda$ is seen to be

$$\frac{1}{2n}(\mathbf{a}-\tilde{\mathbf{a}})^T(nH)(\mathbf{a}-\tilde{\mathbf{a}}) + C,$$

where H is the matrix in the left-hand side of (7.7), $\tilde{\eta}(x,y) = \boldsymbol{\psi}^T(x,y)\tilde{\mathbf{a}}$, and C is a constant; (7.30) is the posterior likelihood of the data divided by n, so the posterior of $\mathbf{a}$ is approximately normal with mean $\tilde{\mathbf{a}}$ and covariance H^+/n, where H^+ is the Moore-Penrose inverse of H. The posterior of $\eta(x,y)$ is thus approximately normal with mean $\tilde{\eta}(x,y) = \boldsymbol{\psi}^T(x,y)\tilde{\mathbf{a}}$ and variance $\boldsymbol{\psi}^T(x,y)H^+\boldsymbol{\psi}(x,y)/n$. For any $x\in\mathcal{X}$, a y-contrast is of the form

$$\kappa(x) = \beta_1\eta(x,y_1) + \cdots + \beta_p\eta(x,y_p),$$

where $\beta_1 + \cdots + \beta_p = 0$; the log odds ratios of $f(y|x)$ are such y-contrasts. The posterior of $\kappa(x)$ is seen to have a mean $\tilde{\kappa}(x) = \tilde{\boldsymbol{\psi}}^T(x)\tilde{\mathbf{a}}$ and a variance $s^2(x) = \tilde{\boldsymbol{\psi}}^T(x)H^+\tilde{\boldsymbol{\psi}}(x)/n$, where $\tilde{\boldsymbol{\psi}}(x) = \beta_1\boldsymbol{\psi}(x, y_1) + \cdots + \beta_p\boldsymbol{\psi}(x, y_p)$. Bayesian confidence intervals of $\kappa(x)$ are given by $\tilde{\kappa}(x) \pm z_{1-\alpha/2}\, s(x)$.

7.8.4 Mixed-Effect Models for Correlated Data

The random effects of §6.1.1 can be extended to the current setting to model correlated data. Replace (7.29) by

$$f(y|x) = \frac{e^{\eta_y + \eta_{x,y} + \mathbf{z}^T\mathbf{b}_y}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{x,y} + \mathbf{z}^T\mathbf{b}_y}}, \tag{7.37}$$

where $\mathbf{b}_y \sim N(0, cB)$ varies with y, for c a constant. Parallel to $\int_{\mathcal{Y}} \eta(x, y) = 0$, we shall specify the correlations among $\mathbf{b}_y$ to ensure $\int_{\mathcal{Y}} \mathbf{z}^T\mathbf{b}_y = 0$.

For $\mathcal{Y} = \{0, \ldots, K\}$, write $\tilde{\mathbf{b}} = (\mathbf{b}_0^T, \ldots, \mathbf{b}_K^T)^T$. We shall specify

$$\tilde{\mathbf{b}} \sim N\big(0, c(I_{K+1} - \tfrac{1}{K+1}\mathbf{1}_{K+1}\mathbf{1}_{K+1}^T) \otimes B\big), \tag{7.38}$$

where $\otimes$ denotes the Kronecker product of matrices. For $\Gamma > 1$, we consider an additive model $\mathbf{b}_y = \mathbf{b}_{y\langle 1\rangle} + \cdots + \mathbf{b}_{y\langle\Gamma\rangle}$, with independent components $\mathbf{b}_{y\langle\gamma\rangle}$ specified as above; the structure of B should remain the same for all the components $\mathbf{b}_{y\langle\gamma\rangle}$, but the constant c may vary from margin to margin. For $\mathcal{Y} = \{0, 1\}$, this reduces to a mixed-effect logistic regression model seen in §6.4.1.

The formulation through (7.37) and (7.38) propagates random effects $\mathbf{z}^T\mathbf{b}$ for univariate responses to cross-classified responses. Note that by (7.38), $\mathbf{b}_0 + \cdots + \mathbf{b}_K = \mathbf{0}$, so one only needs K of the $K+1$ $\mathbf{b}_y$'s. Rewriting $\tilde{\mathbf{b}} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_K^T)^T$, the minus log likelihood of the random effects is seen to be proportional to $\tilde{\mathbf{b}}^T\Sigma\tilde{\mathbf{b}}$ for

$$\Sigma = c^{-1}(I_K + \mathbf{1}_K\mathbf{1}_K^T) \otimes B^{-1}, \tag{7.39}$$

where $I_K + \mathbf{1}_K\mathbf{1}_K = (I_K - \frac{1}{K+1}\mathbf{1}_K\mathbf{1}_K)^{-1}$. For $\Gamma > 1$, one may concatenate all the independent components of $\mathbf{b}_y$ in $\tilde{\mathbf{b}}$ with Σ block-diagonal with blocks of the form as in (7.39).

The model can then be estimated via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\Big\{\eta(X_i, Y_i) + \mathbf{z}_i^T\mathbf{b}_{Y_i} - \log\int_{\mathcal{Y}} e^{\eta(X_i,y) + \mathbf{z}_i^T\mathbf{b}_y}\Big\} + \frac{1}{2n}\tilde{\mathbf{b}}^T\Sigma\tilde{\mathbf{b}} + \frac{\lambda}{2}J(\eta). \tag{7.40}$$

The Newton iteration for the minimization of (7.40) follows straightforward modification of (7.6) (Problem 7.16) and the tuning parameters can be selected by cross-validation. The Kullback-Leibler projection of §7.7.3

can be computed with $\mathbf{z}^T\mathbf{b}_y$ treated as an offset, and Bayesian confidence intervals for y-contrasts follow the same calculus as in §7.8.3 but with $\mathbf{a} = (\mathbf{d}^T, \mathbf{c}^T, \mathbf{b}^T)^T$ and a modified H matrix to be derived in Problem 7.16.

7.8.5 Empirical Performance of Cross-Validation

The tuning parameters are to be selected by the cross-validation score $V(\lambda)$ of (7.21) but with $\mu_f(g)$ and $V_f(g,h)$ defined in (7.31). To assess the effectiveness of cross-validation, simulations were conducted on $\mathcal{Y} = \{0,1\} \times \{0,1\}$ and $\mathcal{X} = [0,1]$. Define

$$\begin{aligned}
\log \frac{p_1(x)}{1-p_1(x)} &= 400x^5(1-x)^3 - 1, \\
\log \frac{p_2(x)}{1-p_2(x)} &= 500x^7(1-x)^3 + 250x^2(1-x)^{10} - 1, \\
\log \frac{p_3(x)}{1-p_3(x)} &= 50x^2(1-x)^4.
\end{aligned}$$

A setting with $Y_{\langle 1\rangle} \perp Y_{\langle 2\rangle} | x$ would have

$$\big(f(0,0), f(0,1), f(1,0), f(1,1)\big) = (q_1q_2, q_1p_2, p_1q_2, p_1p_2),$$

where $q_k = 1 - p_k$, but we modify it by

$$\big(f(0,0), f(0,1), f(1,0), f(1,1)\big) \propto (q_1q_2p_3, q_1p_2q_3, p_1q_2q_3, p_1p_2p_3);$$

note that after the modification, $p_1(x)$ and $p_2(x)$ are no longer the marginal probabilities $P(y_{\langle 1\rangle}=1|x)$ and $P(y_{\langle 2\rangle}=1|x)$, but the log odds ratio is

$$\log \frac{f(0,0|x)f(1,1|x)}{f(1,0|x)f(0,1|x)} = 2\log \frac{p_3(x)}{1-p_3(x)} = 100x^2(1-x)^4.$$

Samples of size $n = 200$ were generated, for $x_i \sim U(0,1)$, with and without random effects. For samples with random effects, $\mathbf{z}^T\mathbf{b}_i = b_1(s_i, y_{\langle 1\rangle}) + b_2(s_i, y_{\langle 2\rangle})$, where $s_i \in \{1, \ldots, 10\}$, 20 each, $b_1(s,1) = -b_1(s,0) \sim N(0,1)$, and $b_2(s,1) = -b_2(s,0) \sim N(0,1)$. Models of the form

$$\begin{aligned}
\eta(x,y) = {} & \eta_1(y_{\langle 1\rangle}) + \eta_2(y_{\langle 2\rangle}) + \eta_{1,2}(y_{\langle 1\rangle}, y_{\langle 2\rangle}) \\
& + \eta_{x,1}(x, y_{\langle 1\rangle}) + \eta_{x,2}(x, y_{\langle 2\rangle}) + \eta_{x,1,2}(x, y_{\langle 1\rangle}, y_{\langle 2\rangle})
\end{aligned} \tag{7.41}$$

were fitted to the data.

To assess the performance of $\hat{f}(y|x)$ as an estimate of $f(y|x)$, one may use as loss the Kullback-Leibler distance

$$L(\lambda) = \mathrm{KL}(f, \hat{f}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{Y}} \log\left\{\frac{f(y|x_i)}{\hat{f}_\lambda(y|x_i)}\right\} f(y|x_i), \tag{7.42}$$

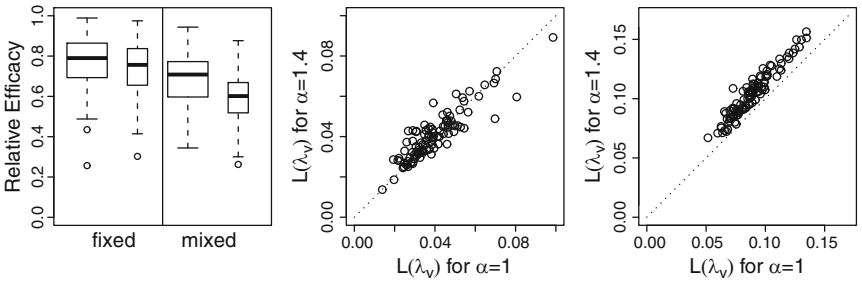


FIGURE 7.14. Effectiveness of cross-validation for log-linear regression models. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$, for $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ for $\alpha = 1$ versus that for $\alpha = 1.4$ in fixed-effect simulation. *Right*: $L(\lambda_v)$ for $\alpha = 1$ versus that for $\alpha = 1.4$ in mixed-effect simulation.

where the dependence of $\hat{f}_\lambda(y|x)$ on the tuning parameters is made explicit, with the subscript λ representing the λ in (7.30) or (7.40), the θ_β's hidden in $J(\eta)$, and also the Σ in (7.40) for mixed-effect models. The conditional density $f(y|x)$ is as in (7.37), which reduces to (7.29) when $\mathbf{z}^T\mathbf{b}_y$ is absent.

For both the fixed-effect (without random effects) and mixed-effect (with random effects) simulations, one hundred replicates were generated. Three estimates were calculated for each replicate, one minimizing $L(\lambda)$ of (7.42) at λ_o, and two minimizing $V(\lambda)$ of (7.21) at λ_v, for $\alpha = 1, 1.4$. The results are summarized in Fig. 7.14, with the relative efficacy $L(\lambda_o)/L(\lambda_v)$ in the left frame and $L(\lambda_v)$ for $\alpha = 1$ versus that for $\alpha = 1.4$ in the center and right frames. The choice of α appears a tossup in the fixed-effect simulation, but $\alpha = 1$ dominated $\alpha = 1.4$ in the mixed-effect simulation.

7.8.6 R Package `gss`: `ssllrm` Suite

Log-linear regression models can be fitted using the `ssllrm` suite. The following sequence generates a sample used in the fixed-effect simulation of §7.8.5 and fits a model of the form as in (7.41) with smoothing parameters selected by $V(\lambda)$ with $\alpha = 1$:

```
test <- function(x) {
  p1 <- plogis(400*x^5*(1-x)^3-1)
  p2 <- plogis(500*x^7*(1-x)^3+250*x^2*(1-x)^10-1)
  p3 <- plogis(50*x^2*(1-x)^4)
  q1 <- 1-p1; q2 <- 1-p2; q3 <- 1-p3
  p <- cbind(q1*q2*p3,q1*p2*q3,p1*q2*q3,p1*p2*p3)
  p/apply(p,1,sum)
}
set.seed(5732)
x <- runif(200); p <- test(x)
```

```
y1 <- y2 <- NULL
for (i in 1:200) {
  ywk <- rmultinom(1,1,p[i,])
  y1 <- c(y1,ywk[3]+ywk[4])
  y2 <- c(y2,ywk[2]+ywk[4])
}
y1 <- factor(y1); y2 <- factor(y2)
fit <- ssllrm(~y1*y2*x,~y1+y2)
```

The basic syntax of `ssllrm` is the same as that of `sscden`. To evaluate the fitted $f(y|x)$, say at $x = (0.3, 0.5)$, use

```
predict(fit,data.frame(x=c(.3,.5)))
```

which returns a 2×4 matrix with $f(y|0.3)$ and $f(y|0.5)$ in the rows; the ordering of the y values, $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$, can be obtained via `fit$qd.pt`. For a y-contrast on a grid, say $\log\{f(1,1|x)/f(1,0|x)\} = \eta(x,1,1) - \eta(x,1,0)$, use

```
xx <- seq(0,1,length=51)
est <- predict(fit,data.frame(x=xx),
               odds=c(0,0,-1,1),se=TRUE)
```

which can be plotted as in the top-left frame of Fig. 7.15, with the data and the test function superimposed:

```
plot(xx,exp(est$fit),type="l",log="y",ylim=c(0.1,10))
lines(xx,exp(est$fit+1.96*est$se),col=5)
lines(xx,exp(est$fit-1.96*est$se),col=5)
pp <- test(xx); lines(xx,pp[,4]/pp[,3],lty=3)
id3 <- (y1==1)&(y2==0); id4 <- (y1==1)&(y2==1)
points(x[id4],rep(10,sum(id4)),col=3)
points(x[id3],rep(0.1,sum(id3)),col=3)
```

Shown in the other three frames of Fig. 7.15 are $f(1,1|x)/f(0,1|x)$, $f(1,1|x)/f(0,0|x)$, and $f(1,1|x)f(0,0|x)/\{f(1,0|x)f(0,1|x)\}$.

7.8.7 Case Study: Eyetracking Experiments

In eyetracking experiments, participants in front of computer monitors listen to instructions such as "click on the purple bottle" and their eye fixation on the target (purple bottle), on some color competitor (e.g., purple pencil), on some object competitor (e.g., yellow bottle), or on something else is monitored on a fine time grid. The purpose of such studies is to explore how linguistic variables may affect the ease with which the listeners can select a visually available referred-to item.

As part of her dissertation research at The Ohio State University, eyetracking data were collected by Anouschka Foltz in 288 trials involving

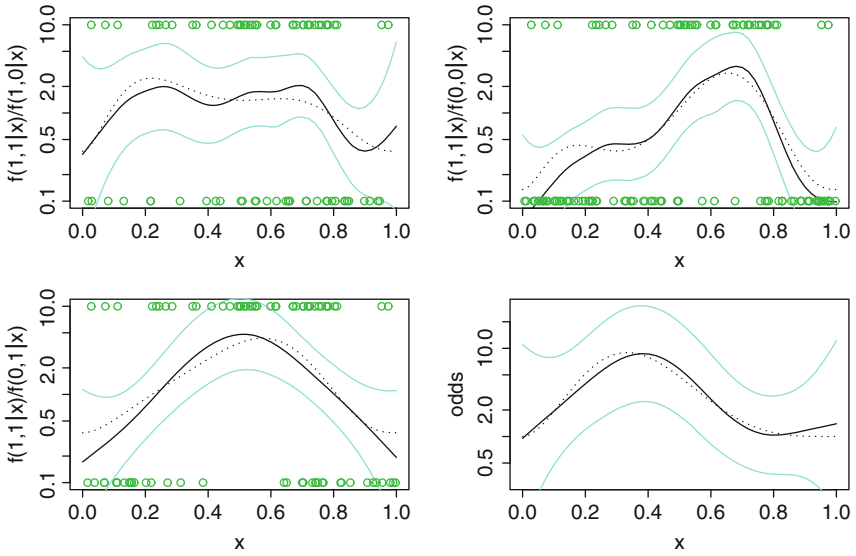


FIGURE 7.15. Log-linear regression model on $\mathcal{X} = [0,1]$ and $\mathcal{Y} = \{0,1\}^2$. Fitted $f(1,1|x)/f(1,0|x)$, $f(1,1|x)/f(0,1|x)$, and $f(1,1|x)/f(0,0|x)$ (*solid*) with 95 % Bayesian confidence intervals (*faded*); true functions (*dotted*) and data (*circles*) are superimposed. The odds ratio $f(1,1|x)f(0,0|x)/\{f(1,0|x)f(0,1|x)\}$ is in the *bottom-right frame*.

48 participants, 6 trials each. In each trial, the participant listened to three consecutive instructions, with the first being something like "click on the YELLOW pencil" and the second "click on the PURPLE bottle;" the common linguistic trait is the emphasized adjectives and the changes in both the adjectives and the nouns, but the particular word choices may vary from trial to trial. Data from the time segment associated with the second instructions are included in `gss` as a data frame `eyetrack`, with elements `time` (136 points at $(-867)(17)(1428)$ ms), `color` (binary, eye fixation on matching color), `object` (binary, eye fixation on matching object), `id` (participant's ID), and `cnt`; time 0 is at the onset of the noun, and the $136 \times 288 = 39168$ readings are merged into 13891 distinctive records with the multiplicity counts in `cnt`.

A model of the form as in (7.37) can be fitted to the data, with $\eta(x,y)$ as in (7.41) and $\mathbf{z}^T\mathbf{b}_y = b_1(s,y_{\langle 1\rangle}) + b_2(s,y_{\langle 2\rangle})$, where $b_1(s,1) = -b_1(s,0) \sim N(0,\sigma_1^2)$ and $b_2(s,1) = -b_2(s,0) \sim N(0,\sigma_2^2)$ are independent:

```
data(eyetrack)
fit.eye <- ssllrm(~time*color*object,~color+object,
                  data=eyetrack,weight=cnt,
                  id.basis=1:136,random=~1|id)
```

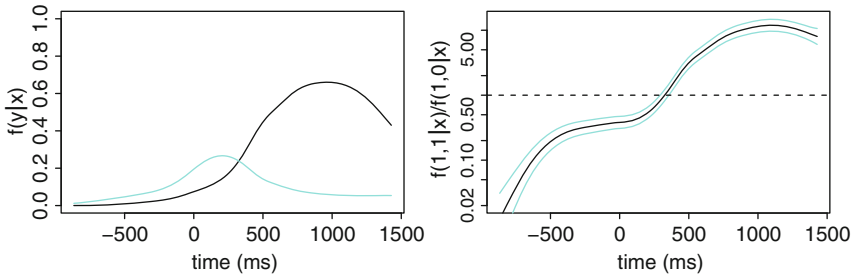


FIGURE 7.16. Fitted probabilities along time. *Left*: $f(1,1|x)$ (*solid*) versus $f(1,0|x)$ (*faded*). *Right*: $f(1,1|x)/f(1,0|x)$ along with 95 % Bayesian confidence intervals.

The `random` argument specifies random effects $\mathbf{z}^T\mathbf{b}$ for univariate responses as discussed in §6.2.6, which are then propagated into the $\mathbf{z}^T\mathbf{b}_y$ of §7.8.4. Due to the huge sample size, the fit may take a hour or two to execute on a modern desktop or laptop.

Upon hearing the emphasized adjective "PURPLE" but before the noun "bottle," one usually expects a noun repetition ("pencil") and starts to look for purple pencil on the monitor, and of interest is how long it takes for the participant to recover from the trap to focus on the target, the purple bottle. Setting $\mathbf{b} = \mathbf{0}$ in the fitted model, one is to compare $f(y|x) = e^{\eta(x,y)}/\int_{\mathcal{Y}} e^{\eta(x,y)}$ at $y = (1,1)$ (target) and $y = (1,0)$ (color competitor), as shown in Figure 7.16:

```
tt <- eyetrack$time[1:136]
p <- predict(fit.eye,data.frame(time=tt))
plot(tt,p[,4],type="l"); lines(tt,p[,3],col=5)
contr <- predict(fit.eye,data.frame(time=tt),
                 odds=c(0,0,-1,1),se=TRUE)
plot(tt,exp(contr$fit),log="y")
lines(tt,exp(contr$fit+1.96*contr$se),col=5)
lines(tt,exp(contr$fit-1.96*contr$se),col=5)
```

The Kullback-Leibler projection suggests that one may set $\eta_{x12} = 0$ but not $\eta_{12} + \eta_{x12} = 0$, so `color` and `object` are dependent but the odds ratio $f(1,1|x)f(0,0|x)/\{f(1,0|x)f(0,1|x)\}$ is largely independent of x:

```
project(fit.eye,c("color","object",
                  "time:color","time:object"))$ratio
# 0.3672533
project(fit.eye,c("color","object","color:object",
                  "time:color","time:object"))$ratio
# 0.03188852
```

The association between $Y_{\langle 1\rangle}$ and $Y_{\langle 2\rangle}$ however is not of primary interest in the current application.

7.9 Response-Based Sampling

In studies of rare events, data are often subject to a form of selection bias known as choice-based sampling in econometrics or case-control sampling in biostatistics. Samples largely from $f(x|y)$ have to be used to estimate $f(y|x)$ or part of it.

Because of the selection bias, $f(y|x)$ is estimable only through the joint density $f(x, y)$. The joint density is not always estimable, but when it is, the estimation through penalized likelihood method is straightforward. The odds ratio is available through either $f(x|y)$ or $f(y|x)$, so is always estimable.

7.9.1 Response-Based Samples

Consider a probability density $f(x, y)$ on a product domain $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$ is discrete. Let $\mathcal{Y}_j \subseteq \mathcal{Y}$, $j = 1, \ldots, s$, be s strata; $\bigcup_{j=1}^{s} \mathcal{Y}_j = \mathcal{Y}$. A stratum $\mathcal{Y}_j$ is selected with probability π_j, $\sum_{i=1}^{s} \pi_j = 1$, and given the stratum, observations are taken from $f(x, y)$ but restricted to the stratum $\mathcal{X} \times \mathcal{Y}_j$. Such data are known as choice-based samples in econometrics or case-control samples in biostatistics. Of interest is the estimation of the conditional density $f(y|x)$. Since the strata are defined by restricted y values, the sampling scheme is called response-based sampling.

Example 7.11 (Separate sampling) With $s = K$ and $\mathcal{Y}_j = \{y = j\}$, one gets a separate sample for case-control studies (Anderson 1972). □

Example 7.12 (Enriched choice-based sampling) With $s = K + 1$, $\mathcal{Y}_j = \{y = j\}$, $j = 1, \ldots, K$, and $\mathcal{Y}_{K+1} = \mathcal{Y}$, one obtains an enriched choice-based sample (Cosslett 1981). □

With response-based sampling, the data are largely from the "wrong" conditional distribution $f(x|y)$. Such sampling strategy is necessary when the categories of interest are rare in the population, in which case an informative random sample from $f(x, y)$ or $f(y|x)$ can be astronomical.

From $f(x, y) = e^{\eta_x + \eta_y + \eta_{x,y}} / \int_{\mathcal{X}} \int_{\mathcal{Y}} e^{\eta_x + \eta_y + \eta_{x,y}}$, one has

$$f(y|x) = \frac{e^{\eta_y + \eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_y + \eta_{x,y}}}, \qquad f(x|y) = \frac{e^{\eta_x + \eta_{x,y}}}{\int_{\mathcal{X}} e^{\eta_x + \eta_{x,y}}}.$$

Separate sampling does not warrant the estimation of $f(y|x)$ unless an independent estimate of the marginal density $f(y) \propto e^{\eta_y} \int_{\mathcal{X}} e^{\eta_x + \eta_{x,y}}$ is available, whereas an enriched sample does carry information about $f(x, y)$ and, hence, about $f(y|x)$. Note that the empirical π_j cannot be used to

estimate the marginal density $f(y)$ due to the very selection bias in the sampling scheme. It is easy to verify, however, that an odds ratio

$$\frac{f(y_1|x_1)/f(y_2|x_1)}{f(y_1|x_2)/f(y_2|x_2)} = \frac{f(y_1|x_1)f(y_2|x_2)}{f(y_1|x_2)f(y_2|x_1)}, \tag{7.43}$$

depends only on the interaction $\eta_{x,y}$, and, hence, is always estimable; see Problem 7.17.

In the case that none of the partitions $\{1, \ldots, s\} = A \cup A^c$ would satisfy $\big(\bigcup_{j\in A} \mathcal{Y}_j\big) \bigcap \big(\bigcup_{j\in A^c} \mathcal{Y}_j\big) = \emptyset$ (Cosslett 1981, Assumption 10), known as the connected case, $f(x, y)$ is identifiable from the sample, in the sense that the minus log likelihood

$$-\frac{1}{n}\sum_{i=1}^{n} \eta(X_i, Y_i) + \sum_{j=1}^{s} \frac{n_j}{n} \log \int_{\mathcal{X}} \int_{\mathcal{Y}_i} e^{\eta} \tag{7.44}$$

is strictly convex in $\eta = \eta_x + \eta_y + \eta_{x,y}$ that satisfies side conditions $A_x(\eta_x + \eta_{x,y}) = 0$, $\forall y$, and $A_y(\eta_y + \eta_{x,y}) = 0$, $\forall x$, where (X_i, Y_i) are the observed data and n_j are the sample sizes from the strata $\mathcal{Y}_j$, $\sum_{j=1}^{s} n_j = n$; see Problem 7.18. When there is a partition of $\{1, \cdots, s\} = A \cup A^c$ such that $\big(\bigcup_{j\in A} \mathcal{Y}_j\big) \bigcap \big(\bigcup_{j\in A^c} \mathcal{Y}_j\big) = \emptyset$, however, η_y is not identifiable although $\eta_x + \eta_{x,y}$ still is.

For the estimation of $f(x|y) = e^{\eta_x+\eta_{x,y}} / \int_{\mathcal{X}} e^{\eta_x+\eta_{x,y}}$, one can always cast the sampling scheme as separate sampling with $s = K$ and $\mathcal{Y}_j = \{y = j\}$, and the minus log conditional likelihood

$$-\frac{1}{n}\sum_{i=1}^{n} \eta(X_i, Y_i) + \sum_{j=1}^{K} \frac{n_j}{n} \log \int_{\mathcal{X}} e^{\eta(x,j)} \tag{7.45}$$

is strictly convex in $\eta = \eta_x + \eta_{x,y}$ that satisfies side conditions $A_x\eta = 0$, $\forall y$. Note that (7.45) is identical to (7.44) under separate sampling, with $\eta_y(j)$ in (7.44) canceling out.

7.9.2 Penalized Likelihood Estimation

The estimation of $f(x|y)$ has been treated in §7.7, so we only consider the connected case here. Write $\hat{\pi}_j = n_j/n$. The joint density $f(x, y) = e^{\eta} / \int_{\mathcal{X}} \int_{\mathcal{Y}} e^{\eta}$ can be estimated through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n} \eta(X_i, Y_i) + \sum_{j=1}^{s} \hat{\pi}_j \log \int_{\mathcal{X}} \int_{\mathcal{Y}_j} e^{\eta} + \frac{\lambda}{2} J(\eta), \tag{7.46}$$

for $\eta(x, y) = \eta_x + \eta_y + \eta_{x,y}$. The minimizer in $\mathcal{H} = \big\{f : J(f) < \infty\big\}$ is generally not computable, but that in $\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\big\{R_J((X_i, Y_i), \cdot)\big\}$ shares the same convergence rates; see §9.2. Define $\mu_f(g|j) = \int_{\mathcal{X}} \int_{\mathcal{Y}_j} ge^f / \int_{\mathcal{X}} \int_{\mathcal{Y}_j} e^f$

and $v_f(g,h|j) = \mu_f(gh|j) - \mu_f(g|j)\mu_f(h|x)$. The Newton updating formula (7.7) on page 241 again holds verbatim for the minimization of (7.46) in $\mathcal{H}^*$, with $\mu_f(g)$ and $V_f(g,h)$ modified as

$$\mu_f(g) = \sum_{j=1}^{K} \hat{\pi}_j \mu_f(g|j), \qquad V_f(g,h) = \sum_{j=1}^{K} \hat{\pi}_j v_f(g,h|j).$$

The Kullback-Leibler distance is now defined by

$$\mathrm{KL}(\eta,\eta_\lambda) = \sum_{j=1}^{K} \pi_j \left\{ \mu_\eta(\eta-\eta_\lambda|j) - \log \int_{\mathcal{X}} \int_{\mathcal{Y}_j} e^{\eta} + \log \int_{\mathcal{X}} \int_{\mathcal{Y}_j} e^{\eta_\lambda} \right\},$$

with the relative Kullback-Leibler distance given by

$$\mathrm{RKL}(\eta,\eta_\lambda) = \sum_{j=1}^{K} \pi_j \left\{ \log \int_{\mathcal{X}} \int_{\mathcal{Y}_j} e^{\eta_\lambda} - \mu_\eta(\eta_\lambda|j) \right\}.$$

The cross-validation and computation are again trivial to modify.

7.10 Bibliographic Notes

Sections 7.1 and 7.2

Penalized likelihood density estimation was pioneered by Good and Gaskins (1971), who used a square root transform for positivity and resorted to constrained optimization to enforce unity. The logistic density transform was introduced by Leonard (1978), and (7.12) was proposed by Silverman (1982) to ensure unity without numerically enforcing it. The early work was largely done in the univariate context, although the basic ideas are applicable in more general settings. Using B-spline basis with local support, O'Sullivan (1988a) developed a fast algorithm similar to that of §3.10.1 for the computation of Silverman's estimate.

The one-to-one logistic density transform through a side condition was introduced in Gu and Qiu (1993), where an asymptotic theory was developed that led to the computability of the estimate through $\mathcal{H}^*$ on generic domains. The estimation of the Poisson process and the link to Silverman's estimate was also noted by Gu and Qiu (1993).

Section 7.3

With a varying smoothing parameter λ in (7.7), a performance-oriented iteration similar to that in §5.2.1 was developed by Gu (1993b). This approach does not bode well with multiple smoothing parameters, however,

as analytical derivatives similar to those behind Algorithm 3.2 are lacking. The direct cross-validation presented here was developed in Gu and Wang (2003).

Section 7.4

The strategy for the handling of numerical singularity is similar to the one discussed in Gu (1993b, Appendix). The flop counts are largely taken from Golub and Van Loan (1989).

The rescaling of the domain for numerical integration on multidimensional cubes is discussed in Gu and Wang (2003).

The Kullback-Leibler projection was developed in Gu (2004).

Section 7.5

The Buffalo snowfall data have been analyzed by numerous authors using various density estimation methods such as density-quantile autoregressive estimation (Parzen 1979), average shifted histograms (Scott 1985), and regression spline extended linear models (Stone, Hansen, Kooperberg, and Truong 1997). The estimates presented here differ slight from the ones shown in Gu (1993b), where a performance-oriented iteration was used to select the smoothing parameter.

The analysis of the CDC blood transfusion data presented here differ slightly from the one in Gu (1998c), where a performance-oriented iteration was used to select the smoothing parameters.

Section 7.6

An early reference on length-biased sampling and its applications is Cox (1969). The empirical distributions for data in the settings of Examples 7.7 and 7.8 were derived and their asymptotic properties studied by Vardi (1982, 1985) and Gill, Vardi, and Wellner (1988). The empirical distribution for the truncated data of Example 7.9 was studied by Woodroofe (1985), Wang, Jewell, and Tsay (1986), Wang (1989), and Keiding and Gill (1990).

The smoothing of the empirical distribution for length-biased data of Example 7.6 through the kernel method was studied by Jones (1991). The general formulation of penalized likelihood density estimation for biased and truncated data as presented in this section is largely taken from an unpublished technical report (Gu 1992d).

Section 7.7

The materials of this section are largely taken from Gu (1995a). We estimate $f(y|x)$ as a "bivariate" function on generic domains $\mathcal{X}$ and $\mathcal{Y}$, where

$\mathcal{X}$ and $\mathcal{Y}$ can both be multivariate. A similar approach to conditional density estimation can be found in Stone, Hansen, Kooperberg, and Truong (1997).

Most nonparametric regression techniques, such as the local polynomial methods (Cleveland 1979; Fan and Gijbels 1996) with the kernel methods as special cases, primarily concern the conditional mean. Work has also been done for the estimation of conditional quantiles (Koenker, Ng, and Portnoy 1994). Cole (1988) and Cole and Green (1992) considered a three-parameter model on the y axis in the form of Box-Cox transformation and estimated the three parameters as functions of x through penalized likelihood; the conditional mean and conditional quantiles could be easily obtained from the three-parameter model.

Section 7.8

The materials of this section are mainly taken from Gu and Ma (2011) . It is a special case of conditional density estimation, yet it includes numerous models in the literature as special cases of its own.

Regression with multinomial responses has been studied by Kooperberg, Bose, and Stone (1997) and Lin (1998). For $\mathcal{Y} = \{0, 1\}^\Gamma$, a product of binary domains, Gao (1999) and Gao, Wahba, Klein, and Klein (2001) attempted a direct generalization of (5.1).

Section 7.9

The term response-based sampling was coined by Manski (1995). Parametric or partly parametric estimation of the odds ratio or the conditional density $f(y|x)$ under such a sampling scheme have been studied by Anderson (1972), Prentice and Pyke (1978), Cosslett (1981), and Scott and Wild (1986), among others. The empirical joint distribution based on enriched samples was derived by Morgenthaler and Vardi (1986) and was used as weights in their kernel estimate of $f(y|x)$. A version of penalized likelihood estimation adapted from Good and Gaskins (1971) was proposed by Anderson and Blair (1982) for the case of $K = 2$. The formulation of this section was largely taken from an unpublished technical report (Gu 1995b).

7.11 Problems

Section 7.1

7.1 Verify (7.5).

7.2 Verify (7.6) and (7.7).

Section 7.2

7.3 Verify (7.11).

7.4 Show that the minimizer $\tilde{\eta}^*$ of (7.12) satisfies the unity constraint $\int_{\mathcal{X}} e^{\tilde{\eta}^*} dx = 1$.

Section 7.3

7.5 For $L_{f,g}(\alpha) = \log \int_{\mathcal{X}} e^{f+\alpha g} dx$ as a function of α, verify that $\dot{L}_{f,g}(0) = \mu_f(g)$ and $\ddot{L}_{f,g}(0) = V_f(g)$.

7.6 Plugging (7.3) into (7.17), show that the minimizing coefficients satisfy (7.7).

7.7 Verify the cross-validation estimate given in (7.22) for prebinned data.

Section 7.4

7.8 Premultiply (7.7) by $\tilde{G}^{-T}$, and show that the linear system reduces to (7.25).

Section 7.6

7.9 Show that the Newton update for the minimization of (7.26) satisfies (7.7), with $\mu_f(g)$ and $V_f(g,h)$ modified as in (7.27).

7.10 Derive the counterpart of (7.21) for use with (7.26).

7.11 Show that with (t, X) generated according to the scheme of Example 7.9, with $A = \{t < x\}$, $f(x)$ supported on $(0, a)$, and $g(t)$ uniform on $(0, a)$, X is length-biased from a density proportional to $xf(x)$.

Section 7.7

7.12 Show that the Newton update for the minimization of (7.30) satisfies (7.7), with $\mu_f(g)$ and $V_f(g,h)$ defined as in (7.31).

7.13 Derive the counterpart of (7.21) for use with (7.30).

Section 7.8

7.14 Verify that (7.35) is the same as (5.1) applied to Bernoulli data.

7.15 Consider $\mathcal{H}_{\langle x\rangle}$ on $\mathcal{X}$, with a reproducing kernel $R_{\langle x\rangle}(x_1, x_2)$ and an inner product $J_{\langle x\rangle}(f, g)$, and $\mathcal{H}_{\langle y\rangle}$ on $\mathcal{Y} = \{0, \ldots, K\}$, with the reproducing kernel $R_{\langle y\rangle}(y_1, y_2) = I_{[y_1=y_2]} - \frac{1}{K+1}$ and the inner product

$$(f, g)_{\langle y\rangle} = f^T(I - \tfrac{1}{K+1}\mathbf{1}\mathbf{1}^T)g.$$

Verify that in the tensor product space $\mathcal{H}_{\langle x\rangle} \otimes \mathcal{H}_{\langle y\rangle}$, with a reproducing kernel $R_{\langle x\rangle}(x_1, x_2)R_{\langle y\rangle}(y_1, y_2)$, the inner product is given by

$$J(f, g) = \tfrac{1}{K+1} \sum_{y=0}^{K} J_{\langle x\rangle}\big((I - A_y)f, (I - A_y)g\big),$$

where $A_y f = \frac{1}{K+1} \sum_{y=0}^{K} f(y)$.

7.16 Plugging (7.3) into (7.40), derive the Newton updating equation for the minimization of (7.40) with respect to $(\mathbf{d}^T, \mathbf{c}^T, \tilde{\mathbf{b}}^T)^T$.

Section 7.9

7.17 Show that the odds ratio of (7.43) depends only on the interaction $\eta_{x,y}$.

7.18 Assuming the connected case and $n_j > 0$, $j = 1, \ldots, s$, show that the minus log likelihood of (7.44) is strictly convex in $\eta = \eta_x + \eta_y + \eta_{x,y}$.

8
Hazard Rate Estimation

For right-censored lifetime data with possible left-truncation, (1.6) of Example 1.3 defines penalized likelihood hazard estimation. Of interest are the selection of smoothing parameters, the computation of the estimates, and the asymptotic behavior of the estimates.

The existence and the computability of the penalized likelihood hazard estimates are discussed in §8.1, and it is shown that the numerical structure of hazard estimation parallels that of density estimation, as given in §7.1. In §8.2, a natural Kullback-Leibler loss is derived under the sampling mechanism, and a cross-validation scheme for smoothing parameter selection is developed to target the loss. It turns out that the algorithms for density estimation as developed in §§7.3 and 7.4 are readily applicable to hazard estimation after trivial modifications. Modeling tools such as Bayesian confidence intervals, Kullback-Leibler projection, and frailty models for correlated data are discussed in §8.3, along with open-source software. Real-data examples are given in §8.4. Also of interest are the estimation of relative risk in a proportional hazard model through penalized partial likelihood (§8.5), which is shown to be isomorphic to density estimation under biased sampling, and models that are parametric in time (§8.6), which can be fitted following the lines of non-Gaussian regression, as discussed in Chap. 5.

Similar to density estimation, the computability of the hazard estimates is through the notion of efficient approximation based on the asymptotic convergence rates, which will be discussed in Chap. 9.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_8,

8.1 Preliminaries

Let T be the lifetime of an item, Z be the left-truncation time at which the item enters study, and C be the right-censoring time beyond which the item is dropped from surveillance, independent of each other. Let U be a covariate and $T|U$ follow a lifetime distribution with survival function $S(t,u) = P(T > t \,|\, U = u)$. Observing independent samples $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, where $X = \min(T, C)$, $\delta = I_{[T \le C]}$, and $Z < X$, one is to estimate the hazard rate $\lambda(t,u) = -\partial \log S(t,u)/\partial t$.

When parametric models are assumed on the time axis, hazard estimation is not much different from non-Gaussian regression as treated in Chap. 5; see §8.6. Assuming a proportional hazard model $\lambda(t,u) = \lambda_0(t)\lambda_1(u)$, one may treat the base hazard $\lambda_0(t)$ as nuisance and estimate the "univariate" relative risk $\lambda_1(u)$ through penalized partial likelihood; see §8.5.

The main subject of this chapter is the estimation of the "bivariate" hazard function $\lambda(t,u) = e^{\eta(t,u)}$ through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)}dt\right\} + \frac{\lambda}{2}J(\eta) \tag{8.1}$$

in a reproducing kernel Hilbert space $\mathcal{H} = \{f : J(f) < \infty\}$ of functions defined on the domain $\mathcal{T} \times \mathcal{U}$. With $\mathcal{U}$ a singleton and $\lambda = 0$, the nonparametric maximum likelihood yields a delta sum estimate of $\lambda(t)$ corresponding to the Kaplan-Meier estimate of the survival function; see Kaplan and Meier (1958). With $\lambda = \infty$, one fits a parametric model in the null space $\mathcal{N}_J = \{f : J(f) = 0\}$ of the penalty. The time domain $\mathcal{T}$ is understood to be $[0, T^*]$ for some T^* finite, which is not much of a constraint, as all observations are finite in practice.

Let $L(f) = -n^{-1}\sum_{i=1}^{n}\{\delta_i f(X_i, U_i) - \int_{Z_i}^{X_i} e^{f(t,U_i)}dt\}$ be the minus log likelihood. When the maximum likelihood estimate uniquely exists in the null space $\mathcal{N}_J$, the following lemmas establish the existence of the minimizer of (8.1) through Theorem 2.9.

Lemma 8.1 *$L(f)$ is convex in f, and the convexity is strict if $f \in \mathcal{H}$ is uniquely determined by its restriction on $\bigcup_{i=1}^{n}\{(Z_i, X_i) \times \{U_i\}\}$.*

Proof: For $\alpha, \beta > 0$, $\alpha + \beta = 1$,

$$\begin{aligned}\int_Z^X e^{\alpha f(t,U)+\beta g(t,U)}dt &\le \left\{\int_Z^X e^{f(t,U)}dt\right\}^{\alpha}\left\{\int_Z^X e^{g(t,U)}dt\right\}^{\beta} \\ &= \exp\left\{\alpha \log\int_Z^X e^{f(t,U)}dt + \beta\log\int_Z^X e^{g(t,U)}dt\right\} \\ &\le \alpha\int_Z^X e^{f(t,U)}dt + \beta\int_Z^X e^{g(t,U)}dt,\end{aligned}$$

where the first inequality (Hölder's) is strict unless $e^{f(t,U)} \propto e^{g(t,U)}$ on (Z, X) and the second is strict unless $\int_Z^X e^{f(t,U)}dt = \int_Z^X e^{g(t,U)}dt$. The lemma follows. □

Lemma 8.2 *$L(f)$ is continuous in f if $f(t,u)$ is continuous in t, $\forall u \in \mathcal{U}$, $\forall f \in \mathcal{H}$.*

Proof: The lemma follows from the continuity of evaluation in $\mathcal{H}$ and the Riemann sum approximations of $\int_Z^X e^{f(t,U)}dt$. □

A few examples follow.

Example 8.1 (Cubic spline with no covariate) A singleton $\mathcal{U}$ characterizes the absence of covariate. Take $\mathcal{T} = [0,1]$ and $J(\eta) = \int_0^1 \ddot{\eta}^2 dt$. One has $\mathcal{N}_J = \text{span}\{1, t\}$. □

Example 8.2 (Cubic spline with binary covariate) Consider $\mathcal{U} = \{1, 2\}$. Take $\mathcal{T} = [0,1]$ and

$$\begin{aligned} J(\eta) &= \theta_m^{-1} \int_0^1 \big(\ddot{\eta}(t,1) + \ddot{\eta}(t,2)\big)^2 + \theta_c^{-1} \int_0^1 \big(\ddot{\eta}(t,1) - \ddot{\eta}(t,2)\big)^2 \\ &= \theta_m^{-1} J_m(\eta) + \theta_c^{-1} J_c(\eta), \end{aligned}$$

where $J_m(\eta)$ penalizes the mean log hazard and $J_c(\eta)$ penalizes the contrast. The null space is given by $\mathcal{N}_J = \text{span}\big\{I_{[u=1]}, I_{[u=2]}, tI_{[u=1]}, tI_{[u=2]}\big\}$. See Example 2.7 of §2.4.4.

Setting $\theta_c = 0$ and $\mathcal{N}_J = \text{span}\big\{I_{[u=1]}, I_{[u=2]}, t\big\}$, one obtains a proportional hazard model. The proportional hazard model can also be obtained from Example 8.1 using the partial spline technique of §4.1, by adding a term $\beta I_{[u=2]}$ to the log hazard, $\lambda(t,u) = e^{\eta(t)+\beta I_{[u=2]}}$. □

Example 8.3 (Tensor product cubic spline) Consider $\mathcal{U} = [0,1]$ and $\mathcal{T} = [0,1]$. The tensor product cubic spline of Example 2.5 in §2.4.3 can be used in (8.1) for the estimation of the log hazard; see also Example 2.8 in §2.4.5. An additive model characterizes a proportional hazard model. □

Similar to the situation for density estimation, a minimizer η_λ of (8.1) in $\mathcal{H} = \big\{f : J(f) < \infty\big\}$ is, in general, not computable, but one may calculate a minimizer η_λ^* in a data-adaptive finite-dimensional space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\big\{R_J\big((\tilde{T}_j, \tilde{U}_j), \cdot\big), j = 1, \ldots, q\big\}, \tag{8.2}$$

for $\big\{(\tilde{T}_j, \tilde{U}_j)\big\}_{j=1}^q \subseteq \big\{(X_i, U_i), \delta_i = 1\big\}$ a random subset of the failure cases, which shares the same asymptotic convergence rates as η_λ; see §9.3.4.

From now on, we shall focus on η_λ^* but drop the star from the notation. Plugging into (8.1) the expression

$$\eta(t,u) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t,u) + \sum_{j=1}^{q} c_j R_J\big((\tilde{T}_j, \tilde{U}_j), (t,u)\big) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \quad (8.3)$$

the calculation of η_λ reduces to the minimization of

$$A_\lambda(\mathbf{c},\mathbf{d}) = -\frac{1}{n}\boldsymbol{\delta}^T(\tilde{S}\mathbf{d} + \tilde{R}\mathbf{c}) + \frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i} \exp\big(\boldsymbol{\phi}_i^T\mathbf{d} + \boldsymbol{\xi}_i^T\mathbf{c}\big)dt + \frac{\lambda}{2}\mathbf{c}^T Q\mathbf{c} \quad (8.4)$$

with respect to $\mathbf{c}$ and $\mathbf{d}$, where $\tilde{S}$ is $n\times m$ with the (j,ν)th entry $\phi_\nu(X_i,U_i)$, $\tilde{R}$ is $n\times q$ with the (i,j)th entry $\xi_j(X_i,U_i) = R_J\big((\tilde{T}_j,\tilde{U}_j),(X_i,U_i)\big)$, Q is $q\times q$ with the (j,k)th entry $R_J\big((\tilde{T}_j,\tilde{U}_j),(\tilde{T}_k,\tilde{U}_k)\big)$, $\boldsymbol{\phi}_i$ is $m\times 1$ with the νth entry $\phi_\nu(t,U_i)$, and $\boldsymbol{\xi}_i$ is $q\times 1$ with the jth entry $\xi_j(t,U_i)$.

Write $\mu_f(g) = (1/n)\sum_{i=1}^n \int_{Z_i}^{X_i} g(t,U_i)e^{f(t,U_i)}dt$ and $V_f(g,h) = \mu_f(gh)$. Taking derivatives of A_λ in (8.4) at $\tilde{\eta} = \boldsymbol{\phi}^T\tilde{\mathbf{d}} + \boldsymbol{\xi}^T\tilde{\mathbf{c}} \in \mathcal{H}^*$, one has

$$\begin{aligned}
\frac{\partial A_\lambda}{\partial \mathbf{d}} &= -\tilde{S}^T\boldsymbol{\delta}/n + \mu_{\tilde{\eta}}(\boldsymbol{\phi}) = -S^T\mathbf{1}/n + \mu_\phi,\\
\frac{\partial A_\lambda}{\partial \mathbf{c}} &= -\tilde{R}^T\boldsymbol{\delta}/n + \mu_{\tilde{\eta}}(\boldsymbol{\xi}) + \lambda Q\tilde{\mathbf{c}} = -R^T\mathbf{1}/n + \mu_\xi + \lambda Q\tilde{\mathbf{c}},\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d}\partial \mathbf{d}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\phi}^T) = V_{\phi,\phi},\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{c}\partial \mathbf{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\xi},\boldsymbol{\xi}^T) + \lambda Q = V_{\xi,\xi} + \lambda Q,\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d}\partial \mathbf{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi},\boldsymbol{\xi}^T) = V_{\phi,\xi},
\end{aligned} \quad (8.5)$$

where S and R have $N = \sum_{i=1}^n \delta_i$ rows corresponding to observations with $\delta_i = 1$; this is virtually a carbon copy of (7.5) on page 241; see Problem 8.1. With the altered definitions of $\mu_f(g)$, $V_f(g,h)$, S, R, and Q, the Newton updating equations (7.6) and (7.7) also hold verbatim for the minimization of $A_\lambda(\mathbf{c},\mathbf{d})$ in (8.4).

Note that $\mu_f(g)$ as defined above generally involves $O(n)$ integrals unless U_i's are heavily duplicated, so one faces similar numerical burden with continuous covariates as with the conditional density estimation of §7.7. One again may trade statistical performance for numerical efficiency via penalized pseudo likelihood; see §10.4.

8.2 Smoothing Parameter Selection

Smoothing parameter selection for hazard estimation parallels that for density estimation. Performance-oriented iteration works fine when the covariate is absent, but it is numerically less efficient when the covariate is

present. The direct cross-validation is as effective as the indirect one and is simpler to implement.

A Kullback-Leibler distance is derived for hazard estimation under the sampling mechanism, and a cross-validation score is derived to track the Kullback-Leibler loss. The cross-validation procedure is nearly a carbon copy of the one derived for density estimation, so the computation follows trivially. The effectiveness of the cross-validation score is evaluated through simple simulation.

As in §§3.2, 5.2, and 7.3, the dependence of entities on θ_β is suppressed in the notation.

8.2.1 Kullback-Leibler Loss and Cross-Validation

Denote by $N(t) = I_{[t \le X, \delta=1]}$ the event process. Under independent censorship, the quantity $e^{\eta(t,u)}dt$ is the conditional probability that $N(t)$ makes a jump in $[t, t+dt)$ given that $t \le X$ and $U = u$; see, e.g., Fleming and Harrington (1991, p. 19). The Kullback-Leibler distance

$$e^{\eta}dt \log \frac{e^{\eta}dt}{e^{\eta_\lambda}dt} + (1 - e^{\eta}dt) \log \frac{1 - e^{\eta}dt}{1 - e^{\eta_\lambda}dt} = \big\{(\eta - \eta_\lambda)e^{\eta} - (e^{\eta} - e^{\eta_\lambda})\big\}dt + O\big((dt)^2\big)$$

measures the proximity of the estimate $e^{\eta_\lambda}dt$ to the true "success" probability $e^{\eta}dt$. Weighting by the at-risk probability

$$\tilde{S}(t,u) = P(Z < t \le X \,|\, U = u) = E\big[I_{[Z<t\le X]}\big|U = u\big]$$

and accumulating over $\mathcal{T} \times \mathcal{U}$, one has a Kullback-Leibler distance

$$\begin{aligned} \mathrm{KL}(\eta, \eta_\lambda) &= \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \big\{(\eta - \eta_\lambda)e^{\eta} - (e^{\eta} - e^{\eta_\lambda})\big\}\tilde{S}(t,u)dt \\ &= E\bigg[\int_{\mathcal{T}} \big\{\big(\eta(t,U) - \eta_\lambda(t,U)\big)e^{\eta(t,U)} - \big(e^{\eta(t,U)} - e^{\eta_\lambda(t,U)}\big)\big\}Y(t)dt\bigg], \end{aligned} \tag{8.6}$$

where $Y(t) = I_{[Z<t\le X]}$ is the at-risk process, $m(u)$ is the density of U, and the expectation is with respect to Z, X, and U. Dropping terms that do not involve η_λ, one obtains a relative Kullback-Leibler distance,

$$\mathrm{RKL}(\eta, \eta_\lambda) = E\bigg[\int_{\mathcal{T}} \big\{e^{\eta_\lambda(t,U)} - \eta_\lambda(t,U)e^{\eta(t,U)}\big\}Y(t)dt\bigg],$$

and its empirical version,

$$\frac{1}{n}\sum_{i=1}^{n} \int_{Z_i}^{X_i} e^{\eta_\lambda(t,U_i)}dt - \frac{1}{n}\sum_{i=1}^{n} \int_{Z_i}^{X_i} \eta_\lambda(t,U_i)e^{\eta(t,U_i)}dt. \tag{8.7}$$

The first term of (8.7) is readily computable, but the second term $\mu_\eta(\eta_\lambda)$ involves the unknown $\eta(t,u)$.

Write $A(t) = \int_0^t e^{\eta_0(s,U)}Y(s)ds$. Conditioning on Z and U, $M(t) = N(t) - A(t)$ is a martingale; see, e.g., Fleming and Harrington (1991, §1.3). For predictable function $h(t)$, the Stieltjes integral $\int_0^t h(s)dM(s)$ is also a martingale; see, e.g., Fleming and Harrington (1991, §2.4). A deterministic (meaning independent of $M(t)$) continuous function is predictable. For $h(t,u)$ continuous in t, $\forall u \in \mathcal{U}$, and independent of Z and X, $E\big[\int_{\mathcal{T}} h(t,U)dM(t)\big] = 0$, where $M(t)$ depends on Z, X, and U. "Estimating" 0 by the sample mean $n^{-1}\sum_{i=1}^n \int_{\mathcal{T}} h(t,U_i)dM_i(t)$, one has

$$\begin{aligned} 0 &\approx \frac{1}{n}\sum_{i=1}^n \left\{ \int_{\mathcal{T}} h(t,U_i)dN_i(t) - \int_{\mathcal{T}} h(t,U_i)I_{[Z_i<t\le X_i]}e^{\eta(t,U_i)}dt \right\} \\ &= \frac{1}{n}\sum_{i=1}^n \left\{ \delta_i h(X_i,U_i) - \int_{Z_i}^{X_i} h(t,U_i)e^{\eta(t,U_i)}dt \right\}, \end{aligned} \tag{8.8}$$

which, upon setting $h(t,U_i) = \eta^{[i]}_{\lambda,\tilde\eta}(t,U_i)$, yields

$$\tilde\mu_\eta(\eta_\lambda) = \frac{1}{n}\sum_{i=1}^n \int_{Z_i}^{X_i} \eta^{[i]}_{\lambda,\tilde\eta}(t,U_i)e^{\eta(t,U_i)}dt \approx \frac{1}{n}\sum_{i=1}^n \delta_i \eta^{[i]}_{\lambda,\tilde\eta}(X_i,U_i), \tag{8.9}$$

where $\eta^{[i]}_{\lambda,\tilde\eta}$ minimizes the delete-one version of the quadratic approximation of (8.1) at $\tilde\eta = \eta_\lambda$. The derivation of the quadratic approximation is left as an exercise (Problem 8.2).

Write $\breve{\boldsymbol{\xi}} = \big(\boldsymbol{\phi}^T, \boldsymbol{\xi}^T\big)^T$, $\breve R = (S,R)$, and $H = V_{\tilde\eta}\big(\breve{\boldsymbol{\xi}}, \breve{\boldsymbol{\xi}}^T\big) + \mathrm{diag}(O, \lambda Q)$. Similar to (7.19) on page 245,

$$\eta^{[i]}_{\lambda,\tilde\eta}(X_i,U_i) = \eta_\lambda(X_i,U_i) - \frac{1}{n-1}\breve{\boldsymbol{\xi}}(X_i,U_i)^T H^{-1}\big(\delta_i\breve{\boldsymbol{\xi}}(X_i,U_i) - \breve R^T\mathbf{1}/n\big), \tag{8.10}$$

where $\sum_{i=1}^n \delta_i\breve{\boldsymbol{\xi}}(X_i,U_i) = \breve R^T\mathbf{1}$; see Problem 8.3. It follows that

$$\tilde\mu_\eta(\eta_\lambda) = \frac{1}{n}\sum_{i=1}^n \delta_i\eta_\lambda(X_i,U_i) - \frac{\mathrm{tr}\big(\breve R H^{-1}\breve R^T\big)}{n(n-1)} + \frac{\mathrm{tr}\big(\mathbf{1}^T\breve R H^{-1}\breve R^T\mathbf{1}\big)}{n^2(n-1)}. \tag{8.11}$$

Substituting (8.11) for the second term in (8.7), one gets a cross-validation estimate of the relative Kullback-Leibler distance,

$$\begin{aligned} V(\lambda) = -\frac{1}{n}\sum_{i=1}^n &\left\{ \delta_i\eta_\lambda(X_i,U_i) - \int_{Z_i}^{X_i} e^{\eta_\lambda(t,U_i)}dt \right\} \\ &+ \alpha\left\{ \frac{\mathrm{tr}\big(\breve R H^{-1}\breve R^T\big)}{n(n-1)} - \frac{\mathrm{tr}\big(\mathbf{1}^T\breve R H^{-1}\breve R^T\mathbf{1}\big)}{n^2(n-1)} \right\}, \end{aligned} \tag{8.12}$$

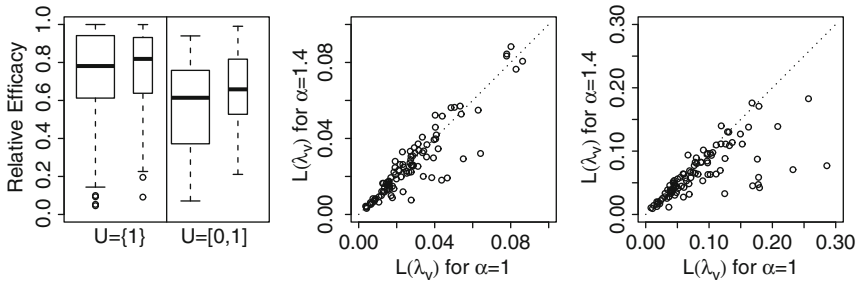


FIGURE 8.1. Effectiveness of cross-validation for hazard estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ for a singleton $\mathcal{U}$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ for $\mathcal{U} = [0, 1]$.

for $\alpha = 1$, where the first term is the minus log likelihood of η_λ. The computation of cross-validated hazard estimates requires little change to the algorithm developed for density estimation.

8.2.2 Empirical Performance

Simple simulations were conducted to explore the empirical performance of cross-validation. Take a singleton $\mathcal{U}$ and a test hazard

$$\lambda_0(t) = e^{\eta(t)} = 24(t - 0.35)^2 + 2.$$

Samples of size $n = 150$ were generated with T_i from $\lambda_0(t)$, C_i from a truncated exponential distribution with $P(C > c) = I_{[c \le 1]} e^{-4c/3}$, and Z_i from an exponential distribution with $P(Z > z) = e^{-5z}$. Using the cubic spline of Example 8.1 and setting $q = N$ in (8.3), three estimates were calculated for each replicate, one minimizing the symmetrized Kullback-Leibler distance

$$L(\lambda) = L(\eta, \eta_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \big(\eta(t, U_i) - \eta_\lambda(t, U_i)\big)\big(e^{\eta(t,U_i)} - e^{\eta_\lambda(t,U_i)}\big) dt \tag{8.13}$$

and the other two minimizing $V(\lambda)$ of (8.12) with $\alpha = 1, 1.4$, yielding an optimal loss $L(\lambda_o)$ and two cross-validation losses $L(\lambda_v)$. The results from one hundred replicates are summarized in Fig. 8.1, with the relative efficacy $L(\lambda_o)/L(\lambda_v)$ shown in the left half of the left frame and the comparison of $\alpha = 1, 1.4$ in $V(\lambda)$ shown in the center frame; two cases are off the chart in the center frame, $(0.406, 0.111)$ and $(0.236, 0.056)$, both in favor of $\alpha = 1.4$. The observed number of failures $N = \sum_{i=1}^{150} \delta_i$ ranged from 90 to 117 over the one hundred replicates, and the overall empirical censoring rate was $4{,}743/15{,}000 = 31.6\,\%$.

The experiment was repeated with $\mathcal{U} = [0, 1]$ and a test hazard

$$\lambda_2(t, u) = e^{\eta(t,u)} = \big(24(t - 0.35)^2 + 2\big)\big(3(u - 0.5)^2 + 0.5\big). \tag{8.14}$$

Samples of size $n = 150$ were generated with $U_i \sim U(0,1)$, $T_i|U_i$ from $\lambda_2(t, U_i)$, and C_i and Z_i as above, and estimates were calculated using the tensor product cubic spline of Example 8.3; the interaction term $\eta_{t,u}$ was included in estimation even though $\eta_2(t,u)$ had an additive structure. Instead of $q = N$, we used $q = 31 \approx 10n^{2/9}$ ξ_j's in (8.3) and the same set was used in the three estimates for each sample. The results from one hundred replicates are also summarized in Fig. 8.1, in the right half of the left frame and in the right frame; the relative efficacy is similar to that in conditional density estimation seen in Fig. 7.11. The observed number of failures N ranged from 81 to 104 over the one hundred replicates, and the overall empirical censoring rate was 5,805/15,000=38.7%.

8.3 Inference and Software

Numerically, hazard estimation has much in common with density estimation, and the Kullback-Leibler projection is well-posed. Without the complication of a normalizing constant, hazards are also like regression functions, on which one may apply tools such as Bayesian confidence intervals and mixed-effect models for correlated data.

Software implementation of the tools is embodied in the `sshzd` suite in `gss`, whose usage is illustrated via simulated examples. For large data sets with continuous covariates, one may have to sacrifice some performance, using instead the `sshzd1` suite of §10.4.

8.3.1 Bayesian Confidence Intervals

Following the calculus of §7.8.3, write $\eta = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c} = \boldsymbol{\psi}^T\mathbf{a}$ as in (8.3) and refer η and $(\mathbf{d}^T, \mathbf{c}^T)^T = \mathbf{a}$ interchangeably. The quadratic approximation of (8.1) at $\tilde{\eta} = \eta_\lambda$ can be written as

$$\frac{1}{2n}(\mathbf{a} - \tilde{\mathbf{a}})^T(nH)(\mathbf{a} - \tilde{\mathbf{a}}) + C,$$

where H is as in (8.10), $\tilde{\eta} = \boldsymbol{\psi}^T\tilde{\mathbf{a}}$, and C is a constant; (8.1) is the posterior likelihood of the data divided by n, so the posterior of $\mathbf{a}$ is approximately normal with mean $\tilde{\mathbf{a}}$ and covariance H^+/n, where H^+ is the Moore-Penrose inverse of H. The posterior of $\eta(t,u)$ is thus approximately normal with mean $\tilde{\eta}(t,u) = \boldsymbol{\psi}^T(t,u)\tilde{\mathbf{a}}$ and variance $s^2(t,u) = \boldsymbol{\psi}^T(t,u)H^+\boldsymbol{\psi}(t,u)/n$. Bayesian confidence intervals of $\eta(t,u)$ are given by $\tilde{\eta}(t,u) \pm z_{1-\alpha/2}\, s(t,u)$.

8.3.2 Kullback-Leibler Projection

Given $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, its Kullback-Leibler projection $\tilde{\eta}$ in $\mathcal{H}_0$ minimizes

$$\mathrm{KL}(\hat{\eta}, \eta) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \{ e^{\hat{\eta}(t,U_i)} \big(\hat{\eta}(t,U_i) - \eta(t,U_i)\big) - \big(e^{\hat{\eta}(t,U_i)} - e^{\eta(t,U_i)}\big) \} dt$$

over $\eta \in \mathcal{H}_0$. Writing $A_{\tilde{\eta},g}(\alpha) = \mathrm{KL}(\hat{\eta}, \tilde{\eta} + \alpha g)$ for $g \in \mathcal{H}_0$, one has

$$0 = \dot{A}_{\tilde{\eta},g}(0) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \big(e^{\hat{\eta}(t,U_i)} - e^{\tilde{\eta}(t,U_i)}\big) g(t,U_i) dt.$$

It then follows, for $\eta_c \in \mathcal{H}_0$, that

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c).$$

One may take $e^{\eta_c} = \sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} (X_i - Z_i)$, the maximum likelihood estimate of a constant hazard model; see Problem 8.4.

8.3.3 Frailty Models for Correlated Data

Adding random effects $\mathbf{z}^T \mathbf{b}$ to the log hazard $\eta(t,u)$, where $\mathbf{b} \sim N(\mathbf{0}, B)$, one obtains frailty models for correlated survival data. The estimation is via the minimization of

$$-\frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i \big(\eta(X_i,U_i) + \mathbf{z}_i^T \mathbf{b}\big) - \int_{Z_i}^{X_i} e^{\eta(t,U_i) + \mathbf{z}_i^T \mathbf{b}} dt \right\} + \frac{1}{2n} \mathbf{b}^T \Sigma \mathbf{b} + \frac{\lambda}{2} J(\eta), \tag{8.15}$$

where $\Sigma = B^{-1}$, often structured, contains correlation parameters, say γ. The Newton updating equation is straightforward to derive (Problem 8.5), and the tuning parameters (λ, γ) can be jointly selected via the cross-validation of §8.2. Bayesian confidence intervals follow the same calculus as in §8.3.1 but with $\mathbf{a} = (\mathbf{d}^T, \mathbf{c}^T, \mathbf{b}^T)^T$ and a modified H matrix to be derived in Problem 8.5. The Kullback-Leibler projection can be computed with $\mathbf{z}^T \mathbf{b}$ treated as an offset.

To fit a frailty model for correlated data, one may use the `random` argument discussed in §6.2.6 in a `sshzd` call.

8.3.4 R Package `gss`: `sshzd` Suite

Penalized likelihood hazard estimation is implemented in the `sshzd` suite, whose usage shall be illustrated using a synthetic example. The following

sequence generates a sample of size $n = 150$ with $T|U$ from $\lambda_2(t,u)$ of (8.14) and fits a tensor product cubic spline to the log hazard:

```
rhzd2 <- function(n) {
  u <- runif(n); wk0 <- 3*(u-.5)^2+.5
  wk1 <- (-log(runif(n))/wk0-.343)/8
  wk1 <- sign(wk1)*abs(wk1)^(1/3)+.35
  wk2 <- -log(runif(n))/2/wk0
  cbind(pmin(wk1,wk2),u)
}
rtest2 <- function(n) {
  wk <- rhzd2(n); tt <- wk[,1]; u <- wk[,2]
  cens <- pmin(-log(runif(n))*3/4,1)
  z <- -log(runif(n))/5
  x <- pmin(tt,cens)
  delta <- tt<=cens
  ok <- x>z
  while(m <- sum(!ok)) {
    wk[!ok] <- rhzd2(m)
    tt[!ok] <- wk[!ok,1]; u[!ok] <- wk[!ok,2]
    cens[!ok] <- pmin(-log(runif(m))*3/4,1)
    z[!ok] <- -log(runif(m))/5
    x[!ok] <- pmin(tt[!ok],cens[!ok])
    delta[!ok] <- tt[!ok]<=cens[!ok]
    ok <- x>z
  }
  cbind(x,delta,z,u)
}
set.seed(2375)
xdzu <- rtest2(150)
x <- xdzu[,1]; delta <- xdzu[,2]
z <- xdzu[,3]; u <- xdzu[,4]
fit <- sshzd(Surv(x,delta,z)~x*u)
```

where the follow-up time `x` must appear in the right-hand side of the model formula. Projecting the fit into the space of additive models, one has

```
project(fit,inc=c("x","u"))$ratio
# 0.1589023
```

In this case, the Kullback-Leibler projection failed to detect the additive structure of the true log hazard.

To evaluate the fitted hazard, say at $(t,u) = (0.5, 0.5)$, one may use

```
hzdrate.sshzd(fit,data.frame(x=.5,u=.5))
# 1.360889
```

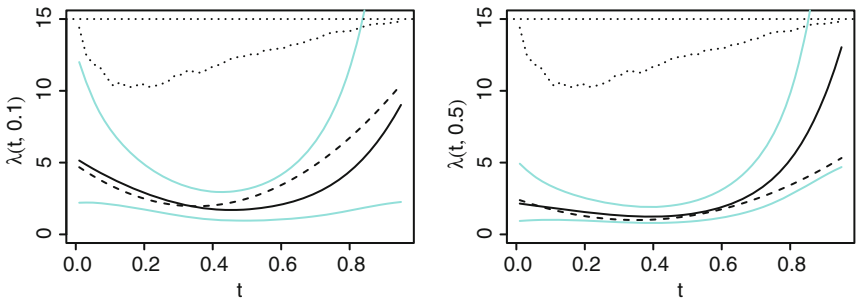


FIGURE 8.2. Hazard estimation on $\mathcal{T} = [0, 1]$ and $\mathcal{U} = [0, 1]$. The estimated $e^{\eta(t,u)}$ are in *solid lines*, the 95 % Bayesian confidence intervals in *faded lines*, and the test hazard $\lambda_2(t, u) = \{24(t - 0.35)^2 + 2\}\{3(u - 0.5)^2 + 0.5\}$ in *dashed lines*. *Left*: $u = 0.1$. *Right*: $u = 0.5$. The *dotted lines* from above are proportional to the size of the risk set, $\sum_{i=1}^{n} I_{[Z_i < t \le X_i]}$.

To evaluate $e^{\eta(t,u)}$ on a grid of t at selected u values, try something like

```
tt <- seq(.01,.95,length=48)
est <- hzdcurve.sshzd(fit,tt,data.frame(u=c(.1,.5)),
                      se=TRUE)
```

which can then be plotted along with Bayesian confidence intervals, the test hazard, and the size of the risk set $\sum_{i=1}^{n} I_{[Z_i < t \le X_i]}$ as in Fig. 8.2:

```
plot(tt,est$fit[,1],type="l",ylim=c(0,15))
lines(tt,est$fit[,1]*exp(1.96*est$se[,1]),col=5)
lines(tt,est$fit[,1]/exp(1.96*est$se[,1]),col=5)
hzd2 <- function(t,u) (24*(t-.35)^2+2)*(3*(u-.5)^2+.5)
lines(tt,hzd2(tt,.1),lty=2)
risk <- apply(outer(tt,z,">")&outer(tt,x,"<"),1,sum)
lines(tt,15-risk/15,lty=3)
```

Note that `est$fit` is the estimated hazard $e^{\eta(t,u)}$ but `est$se` is the standard error of the log hazard $\eta(t, u)$. It is reassuring to see that the Bayesian confidence intervals are tighter at $u = 0.5$ than at $u = 0.1$. The peak size of the risk set was 71.

8.4 Case Studies

We now apply the techniques developed so far to analyze a few real data sets.

8.4.1 Treatments of Gastric Cancer

The survival times of 90 gastric cancer patients are listed in Moreau et al. (1985). Half of the patients were treated by chemotherapy, the other half by

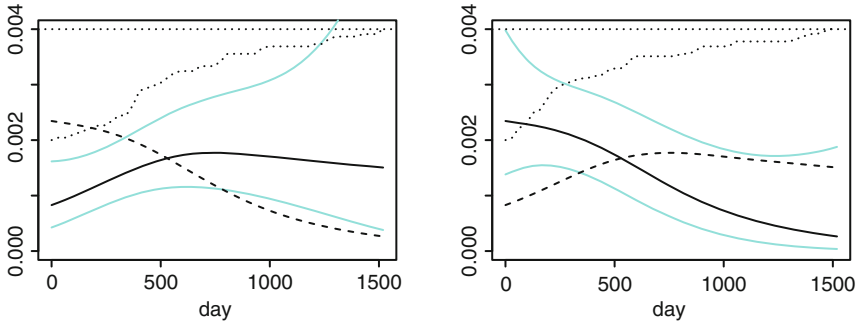


FIGURE 8.3. Treatments of gastric cancer. *Left*: Chemotherapy. *Right*: Combined therapy. The estimated $e^{\eta(t,u)}$ are in *solid lines*, the 95 % Bayesian confidence intervals in *faded lines*, and the hazard $e^{\eta(t,u)}$ under the other treatment in *dashed lines*. The *dotted lines* from the above are proportional to the size of the risk set.

chemotherapy combined with radiotherapy. There were 37 recorded deaths and 8 censorings in each of the treatment groups. The follow-up times ranged from 1 to 1,519 days. The data are included in `gss` as a data frame `gastric` with elements `futime`, `status`, and `trt`.

The following sequence loads the data and fits the model specified in Example 8.2; $\mathcal{T}$ is mapped onto $[0, 1]$ internally:

```
data(gastric)
fit.gastric <- sshzd(Surv(futime,status)~futime*trt,
                     data=gastric,nbasis=90)
```

The option `nbasis=90` allows q up to $n = 90$ but the maximum it can take is $q = N = \sum_{i=1}^{n} \delta_i$. The fit can then be plotted as in Fig. 8.3:

```
tt <- seq(0,1519,length=50)
est <- hzdcurve.sshzd(fit.gastric,tt,
                data.frame(trt=as.factor(1:2)),TRUE)
plot(tt,est$fit[,1],type="l",ylim=c(0,.004))
lines(tt,est$fit[,1]*exp(1.96*est$se[,1]),col=5)
lines(tt,est$fit[,1]/exp(1.96*est$se[,1]),col=5)
lines(tt,est$fit[,2],lty=2)
r1 <- apply(outer(tt,gastric$futime,"<")[,1:45],1,sum)
lines(tt,.004-r1/45*.002,lty=3); abline(h=.004,lty=3)
```

The combined therapy appeared to take a heavier toll than chemotherapy alone in the early going, but for those who survived beyond about 500 days, the comparison was reversed. This, however, does not necessarily mean that radiation would eventually benefit. The stronger patients would probably survive a long time anyway, regardless of the therapy, but for the rest of the patients, radiation seemed to kill many of them before long.

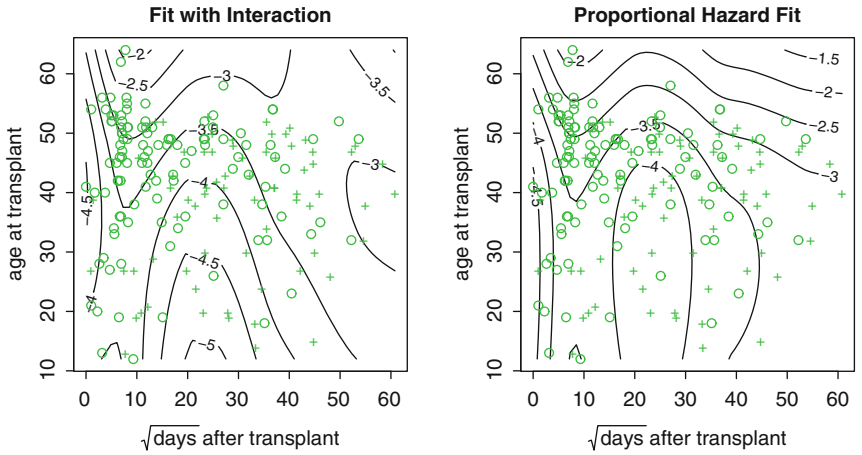


FIGURE 8.4. Hazard after heart transplant. The contours are the estimated $\log \tilde{\lambda}(t^*, u)$, with deceased (*circles*) and censored (*pluses*) patients superimposed.

8.4.2 Survival After Heart Transplant

We shall now fill in more details concerning the analysis of the Stanford heart transplant data previewed in §1.4.3. The data are included in `gss` as a data frame `stan` with elements `time`, `status`, `age`, and `futime`, where `futime` is the square root of `time`. The follow-up times after transplant were between 0 and 3,695 days, and the ages of patients at transplant were between 12 and 64. As mentioned in §1.4.3, a square root transform $t^* = \sqrt{t}$ was applied on the time axis to spread the data more evenly.

The following sequence loads the data and fits a tensor product cubic spline to the log hazard $\log \tilde{\lambda}(t^*, u) = \tilde{\eta}(t^*, u)$:

```
data(stan)
fit.stan <- sshzd(Surv(futime,status)~futime*age,
                  data=stan,nbasis=200)
```

Projecting into the space of additive models, one has

```
project(fit.stan,inc=c("futime","age"))$ratio
# 0.09302142
```

The strength of the interaction term is moderate, and one may also fit a proportional hazard model:

```
fit1.stan <- sshzd(Surv(futime,status)~futime+age,
                   data=stan,nbasis=200)
```

The fits can then be plotted as contours as shown in Fig. 8.4:

```
t.gd <- seq(0,max(stan$futime),length=51)
u.gd <- seq(min(stan$age),max(stan$age),length=51)
```

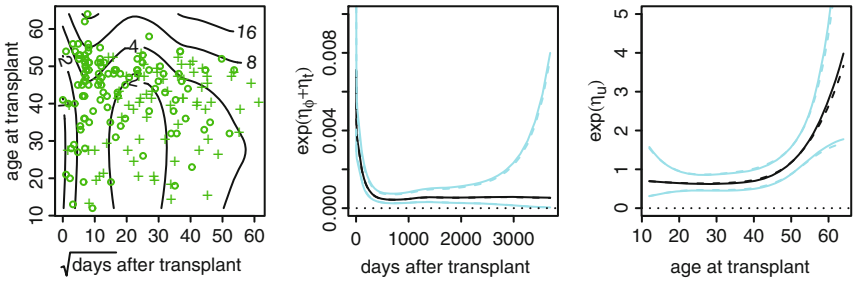


FIGURE 8.5. Hazard after heart transplant: Proportional hazard fit. *Left*: Contours of $100\tilde{\lambda}(t^*, u)$, with deceased (*circles*) and censored (*pluses*) patients superimposed. *Center*: Base hazard $e^{\eta_\emptyset + \eta_t}$ with 95 % Bayesian confidence intervals, on the original time scale. *Right*: Age effect e^{η_u} with 95 % Bayesian confidence intervals. Estimates via the penalized partial likelihood of §8.5 are superimposed in *dashed lines* in the *center* and *right* frames.

```
grid <- expand.grid(futime=t.gd,age=u.gd)
est <- hzdrate.sshzd(fit.stan,grid)
dead <- stan$status==1
contour(t.gd,u.gd,matrix(log(est),51,51))
points(stan$futime[!dead],stan$age[!dead],pch="+",col=3)
points(stan$futime[dead],stan$age[dead],col=3)
```

The two fits are visually close to each other, especially in data-dense areas.

Figure 1.4 for the proportional hazard fit is reproduced in Fig. 8.5, with the contours of $100\tilde{\lambda}(t^*, u)$:

```
est1 <- hzdrate.sshzd(fit1.stan,grid)
contour(t.gd,u.gd,matrix(100*est1,51,51))
points(stan$futime[!dead],stan$age[!dead],pch="+",col=3)
points(stan$futime[dead],stan$age[dead],col=3)
```

the base hazard $e^{\eta_\emptyset + \eta_t} = e^{\tilde{\eta}_\emptyset + \tilde{\eta}_t}/(2\sqrt{t})$ on the original time scale:

```
est.b <- hzdrate.sshzd(fit1.stan,data.frame(futime=t.gd),
                       se=TRUE,inc=c("1","futime"))
plot(t.gd^2,est.b$fit/2/t.gd,type="l",ylim=c(0,.01))
lines(t.gd^2,est.b$fit/2/t.gd*exp(1.96*est.b$se),col=5)
lines(t.gd^2,est.b$fit/2/t.gd/exp(1.96*est.b$se),col=5)
abline(h=0,lty=3)
```

and the age effect e^{η_u}:

```
est.a <- hzdrate.sshzd(fit1.stan,data.frame(age=u.gd),
                       se=TRUE,inc=c("age"))
plot(u.gd,est.a$fit,type="l",ylim=c(0,5))
lines(u.gd,est.a$fit*exp(1.96*est.a$se),col=5)
```

```
lines(u.gd,est.a$fit/exp(1.96*est.a$se),col=5)
abline(h=0,lty=3)
```

It is seen that once a patient survived the initial shock, the hazard rate would remain stable over extended time period. The relative risk was flat for younger patients up to about 40 years of age, then quickly took off for older patients.

8.5 Penalized Partial Likelihood

Assume a proportional hazard model $\lambda(t,u) = \lambda_0(t)\lambda_1(u)$. Treating the base hazard $\lambda_0(t)$ as a nuisance parameter, one may estimate the relative risk $\lambda_1(u)$ using penalized partial likelihood.

The estimation of relative risk through penalized partial likelihood is isomorphic to density estimation under biased sampling, as treated in §7.6, so no new estimation techniques are needed here. Models for the relative risk have much in common with regression models, for which one may add parametric (partial) terms as in §4.1, add random effects as in §8.3.3, and calculate Bayesian confidence intervals as in §§7.8.3 and 8.3.1. Software tools are illustrated using simulated and real data examples.

8.5.1 Partial Likelihood and Biased Sampling

Let $Y_i(t) = I_{[Z_i<t\le X_i]}$ be the at-risk process of the ith observation. For the estimation of the relative risk $\lambda_1(u) = e^{\eta(u)}$, Cox (1972) proposed to work with the partial likelihood,

$$\prod_{i=1}^{n}\left(\frac{e^{\eta(U_i)}}{\sum_{k=1}^{n} Y_k(X_i)e^{\eta(U_k)}}\right)^{\delta_i} = \prod_{j=1}^{N}\left(\frac{e^{\eta(U_j^*)}}{\sum_{k=1}^{n} Y_k(T_j)e^{\eta(U_k)}}\right), \tag{8.16}$$

where (T_j, U_j^*) are the observed lifetimes and the corresponding covariates. Note that the relative risk is defined only up to a multiplicative constant, so a side condition $A\eta = 0$ on the log relative risk would be needed to pin down the function to be estimated; see related discussion on logistic density transform in §7.1.

Writing $\int f = \sum_{k=1}^{n} f(U_k)$, $e^\eta / \int e^\eta$ defines a probability density on the discrete domain $\{U_k, k = 1,\ldots,n\}$. One may write

$$\frac{e^{\eta(U_j^*)}}{\sum_{k=1}^{n} Y_k(T_j)e^{\eta(U_k)}} = \frac{w_j(U_j^*)e^{\eta(U_j^*)}}{\int w_j(u)e^{\eta(u)}},$$

where $w_j(u)$ is defined by $w_j(U_k) = Y_k(T_j)$. Hence, the partial likelihood of (8.16) can be cast as a likelihood for density estimation under biased

sampling; see §7.6. The estimation of relative risk can be conducted via the minimization of the penalized partial likelihood functional

$$\frac{1}{N}\sum_{j=1}^{N}\left\{\eta(U_j^*)-\log\sum_{i=1}^{n}Y_i(T_j)e^{\eta(U_i)}\right\}+\frac{\lambda}{2}J(\eta), \tag{8.17}$$

which is in fact a special case of (7.26); computation and smoothing parameter selection follow the procedures outlined in §7.6.2. Further details are left as exercises (Problems 8.6 and 8.7).

8.5.2 Inference

Following §§7.8.3 and 8.3.1, one may write $\eta = \boldsymbol{\phi}^T\mathbf{d}+\boldsymbol{\xi}^T\mathbf{c} = \boldsymbol{\psi}^T\mathbf{a}$, plug it into (8.17), and derive Bayesian confidence intervals for η based on the quadratic approximation of (8.17) at its minimizer η_λ.

Given $\hat\eta \in \mathcal{H}_0 \oplus \mathcal{H}_1$, one may calculate its Kullback-Leibler projection $\tilde\eta$ in $\mathcal{H}_0$ via the minimization of

$$\mathrm{KL}(\hat\eta,\eta)=\frac{1}{N}\sum_{j=1}^{N}\left\{\frac{\sum_{i=1}^{n}(\hat\eta-\eta)(U_i)Y_i(T_j)e^{\hat\eta(U_i)}}{\sum_{i=1}^{n}Y_i(T_j)e^{\hat\eta(U_i)}}-\log\frac{\sum_{i=1}^{n}Y_i(T_j)e^{\hat\eta(U_i)}}{\sum_{i=1}^{n}Y_i(T_j)e^{\eta(U_i)}}\right\}$$

over $\eta \in \mathcal{H}_0$. It is easy to verify that $\mathrm{KL}(\hat\eta,\eta_c) = \mathrm{KL}(\hat\eta,\tilde\eta)+\mathrm{KL}(\tilde\eta,\eta_c)$, where $\eta_c \in \mathcal{H}_0$ is a constant. As is the case in regression settings, the minimization of $\mathrm{KL}(\hat\eta,\eta)$ can be ill-posed.

As in §8.3.3, mixed-effect (frailty) models can be used to accommodate correlated data; the fitting function `sscox` to be discussed below also has the optional argument `random` described in §6.2.6. The computation, cross-validation, and Bayesian confidence intervals follow straightforward modifications, and the Kullback-Leibler projection can be computed with the random effects treated as an offset.

8.5.3 R Package `gss`: `sscox` Suite

Tools for penalized partial likelihood are implemented in the `sscox` suite, whose usage shall be illustrated using synthetic example. We recycle the simulated data used in §8.3.4, with $T|U$ from $\lambda_2(t,u)$ of (8.14):

```
set.seed(2375); xdzu <- rtest2(150)
x <- xdzu[,1]; delta <- xdzu[,2]
z <- xdzu[,3]; u <- xdzu[,4]
```

where `rtest2` is listed in §8.3.4. To estimate the relative risk, one may use

```
fit.cox <- sscox(Surv(x,delta,z)~u,
                 type=list(u=list("cubic",c(0,1))))
```

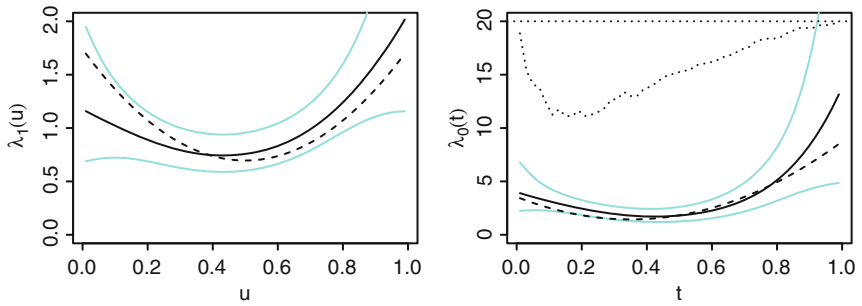


FIGURE 8.6. Estimation of relative risk and base hazard. *Left*: $\lambda_1(u) = e^{\eta(u)}$ with 95 % Bayesian confidence intervals. *Right*: $\lambda_0(t) = e^{\zeta(t)}$ with 95 % Bayesian confidence intervals; the *dotted line* from above is proportional to the size of the risk set, $\sum_{i=1}^{n} I_{[Z_i<t\le X_i]}$. The test functions are superimposed in *dashed lines*.

which can be plotted on a grid as in the left frame of Fig. 8.6:

```
gd <- ((1:50)-.5)/50
est.u <- predict(fit.cox,data.frame(u=gd),se=TRUE)
plot(gd,est.u$fit,type="l",ylim=c(0,2))
lines(gd,est.u$fit*exp(1.96*est.u$se),col=5)
lines(gd,est.u$fit/exp(1.96*est.u$se),col=5)
lam1 <- (3*(gd-.5)^2+.5); cc <- mean(log(lam1))
lines(gd,lam1/exp(cc),lty=2)
```

`predict` returns the relative risk $e^{\eta(u)}$ but the standard error is for $\eta(u)$. We took care to specify the domain $\mathcal{U} = [0,1]$ in `fit.cox` so that $\int_0^1 \eta(u)du = 0$, allowing a definitive factorization of $\lambda_2(t,u) = \lambda_0(t)\lambda_1(u)$ as plotted in Fig. 8.6 in dashed lines.

Treating the estimated relative risk $\lambda_1(u) = e^{\eta(u)}$ as known, the base hazard $\lambda_0(t) = e^{\zeta(t)}$ can be estimated via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\zeta(X_i) - \int_{Z_i}^{X_i} e^{\zeta(t)+o_i}dt\right\} + \frac{\lambda}{2}J(\zeta),$$

where $o_i = \eta(U_i) = \log\lambda_1(U_i)$. This can be achieved using `sshzd` with an `offset` term:

```
risk <- predict(fit.cox,data.frame(u=u))
fit.base <- sshzd(Surv(x,delta,z)~x,offset=log(risk))
```

The base hazard can then be plotted on a grid as in the right frame of Fig. 8.6:

```
est.t <- hzdcurve.sshzd(fit.base,gd,se=TRUE)
plot(gd,est.t$fit,type="l",ylim=c(0,20))
lines(gd,est.t$fit*exp(1.96*est.t$se),col=5)
```

```
lines(gd,est.t$fit/exp(1.96*est.t$se),col=5)
lines(gd,(24*(gd-.35)^2+2)*exp(cc),lty=2)
r.set <- apply(outer(gd,z,">")&outer(gd,x,"<="),1,sum)
lines(gd,20-r.set/8,lty=3); abline(h=20,lty=3)
```

We now add two more terms to $\eta(u)$ that should not be there, one parametric and one nonparametric:

```
set.seed(5732); u2 <- runif(150); u3 <- runif(150)
fit1.cox <- sscox(Surv(x,delta,z)~u+u2,partial=~u3)
```

The Kullback-Leibler projection can be calculated to assess these terms:

```
project(fit1.cox,inc=c("u"))$ratio
# 0.2348213
project(fit1.cox,inc=c("u2","u3"))$ratio
# 0.8118954
```

The estimate seems to contain structures that are not in the test hazard.

8.5.4 Case Study: Survival After Heart Transplant

For the Stanford heart transplant data of §§1.4.3 and 8.4.2, the following sequence estimates, evaluates, and plots the relative risk $e^{\eta(u)}$ as shown in the right frame of Fig. 8.5 in dashed lines:

```
fit2.stan <- sscox(Surv(futime,status)~age,
                   data=stan,nbasis=200)
u.gd <- seq(min(stan$age),max(stan$age),length=51)
est2.a <- predict(fit2.stan,data.frame(age=u.gd),se=TRUE)
plot(u.gd,est2.a$fit,type="l",ylim=c(0,5))
lines(u.gd,est2.a$fit*exp(1.96*est2.a$se),col=5)
lines(u.gd,est2.a$fit/exp(1.96*est2.a$se),col=5)
```

Note that one may simply use the untransformed `time` in the place of `futime` to obtain the same fit. Pretending the estimated relative risk as known, one may estimate, evaluate, and plot the base hazard as shown in the center frame of Fig. 8.5 in dashed lines:

```
risk <- predict(fit2.stan,stan)
fit2.b <- sshzd(Surv(futime,status)~futime,data=stan,
                offset=log(risk),nbasis=200)
t.gd <- seq(0,max(stan$futime),length=51)
est2.b <- hzdcurve.sshzd(fit2.b,t.gd,se=TRUE)
plot(t.gd^2,est2.b$fit/2/t.gd,type="l",ylim=c(0,.01))
lines(t.gd^2,est2.b$fit/2/t.gd*exp(1.96*est2.b$se),col=5)
lines(t.gd^2,est2.b$fit/2/t.gd/exp(1.96*est2.b$se),col=5)
abline(h=0,lty=3)
```

Visually, the estimates through penalized partial likelihood are nearly indistinguishable from those resulting from the joint estimation via (8.1).

8.6 Models Parametric in Time

When parametric models are assumed on the time axis, one usually needs to estimate a parameter of the lifetime distribution as a function of the covariate. The problem is similar to non-Gaussian regression as treated in Chap. 5, although the response likelihood may not belong to an exponential family.

We discuss the accelerated life models through location-scale families for the log lifetime. Details are then spelled out, in parallel to §§5.4.2–5.4.6, concerning the Weibull, log normal, and log logistic families; software tools are in the `gssanova`, `gssanova0`, and `gssanova1` suites.

8.6.1 Location-Scale Families and Accelerated Life Models

Let $F(z)$ be a cumulative distribution function on $(-\infty, \infty)$ and $f(z)$ be its density. A location-scale family is given by $P(X \le x|\mu, \sigma) = F\big((x-\mu)/\sigma\big)$, where μ is the location parameter and $\sigma > 0$ is the scale parameter.

Assume a location-scale family for $\log T$. The survival function and the hazard function are easily seen to be

$$S(t) = 1 - F(z), \qquad \lambda(t) = \frac{1}{\sigma t}\,\frac{f(z)}{1-F(z)}, \tag{8.18}$$

where $z = (\log t - \mu)/\sigma$. We shall write $\eta = \mu$ for the rest of the section.

Let σ be a constant and η be a function of a covariate u with $\eta(u_0) = 0$ at a "control" point u_0. It follows that

$$S(t|u) = 1 - F\big((\log t - \eta(u))/\sigma\big) = 1 - F\big(\log(te^{-\eta(u)})/\sigma\big) = S(te^{-\eta(u)}|u_0),$$

so the covariate is effectively rescaling the time axis. Such models are known as accelerated life models.

Example 8.4 (Extreme value and Weibull distributions) Setting $F(z) = 1 - e^{-w}$ with $f(z) = we^{-w}$, where $w = e^z$, one has the extreme value distribution. When $\log T$ follows an extreme value distribution, T follows a Weibull distribution with survival function and hazard function

$$\begin{aligned} S(t) &= \exp\big\{-e^{(\log t-\eta)/\sigma}\big\} = \exp\big\{-(t/e^{\eta})^{1/\sigma}\big\} = \exp\big\{-(t/\beta)^{\nu}\big\}, \\ \lambda(t) &= \frac{1}{\sigma t}e^{(\log t-\eta)/\sigma} = \frac{1}{\sigma t}\left(\frac{t}{e^{\eta}}\right)^{1/\sigma} = \frac{\nu}{t}\left(\frac{t}{\beta}\right)^{\nu}, \end{aligned} \tag{8.19}$$

where $\nu = 1/\sigma$ is called the shape parameter and $\beta = e^{\eta}$ is called the scale parameter. When $\nu = 1$, the Weibull distribution reduces to the exponential distribution. □

Example 8.5 (Normal and log normal distributions) Setting $F(z) = \Phi(z)$, the cumulative distribution function of the standard normal with $f(z) = \phi(z) = e^{-z^2/2}/\sqrt{2\pi}$, one has the normal distribution. When $\log T$ follows a normal distribution, T is log normal with survival function and hazard function

$$S(t) = 1 - \Phi(z), \qquad \lambda(t) = \frac{1}{\sigma t}\,\frac{\phi(z)}{1-\Phi(z)}, \tag{8.20}$$

where $z = (\log t - \eta)/\sigma$. □

Example 8.6 (Logistic and log logistic distributions) Setting $F(z) = w/(1+w)$ with $f(z) = w/(1+w)^2$, where $w = e^z$, one has the logistic distribution. When $\log T$ follows a logistic distribution, T follows a log logistic distribution with survival function and hazard function

$$S(t) = \frac{1}{1+e^z}, \qquad \lambda(t) = \frac{1}{\sigma t}\,\frac{e^z}{1+e^z}, \tag{8.21}$$

where $z = (\log t - \eta)/\sigma$. □

The minus log likelihood of (Z, X, δ) is seen to be

$$-\left\{\delta \log \lambda(X;\eta,\sigma) - \textstyle\int_Z^X \lambda(t;\eta,\sigma)dt\right\} = l(\eta,\sigma), \tag{8.22}$$

where $\lambda(t;\eta,\sigma)$ spells out the dependence of $\lambda(t)$ on the parameters η and σ; see Problem 1.2. Observing $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, one may estimate η via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \log \lambda(X_i;\eta_i,\sigma) - \int_{Z_i}^{X_i} \lambda(t;\eta_i,\sigma)\,dt\right\} + \frac{\lambda}{2}J(\eta), \tag{8.23}$$

where $\eta_i = \eta(U_i)$; the smoothing parameter λ is not to be confused with the hazard rate $\lambda(t,u) = \lambda\big(t;\eta(u),\sigma\big)$. To calculate the minimizer η_λ of (8.22), one may iterate on (5.3) (p. 177). Fix σ and define

$$h_1(t;\eta) = -\frac{\partial \log \lambda(t;\eta,\sigma)}{\partial \eta}, \qquad h_2(t;\eta) = \frac{\partial h_1(t;\eta)}{\partial \eta}.$$

One has

$$\begin{aligned} u &= \frac{dl}{d\eta} = \delta h_1(X;\eta) - \int_Z^X h_1(t;\eta)\lambda(t;\eta,\sigma)dt = \int h_1(t;\eta)dM(t), \\ w &= \frac{d^2 l}{d\eta^2} = \delta h_2(X;\eta) - \int_Z^X h_2(t;\eta)\lambda(t;\eta,\sigma)dt + \int_Z^X h_1^2(t;\eta)\lambda(t;\eta,\sigma)dt \\ &= \int h_2(t;\eta)dM(t) + \int h_1^2(t;\eta)dA(t), \end{aligned}$$

where $M(t) = N(t) - A(t)$ is a martingale, $N(t) = I_{[t \le X, \delta = 1]}$ is the event process, and $A(t) = \int_0^t I_{[Z < s \le X]} \lambda(s; \eta, \sigma) ds$; see §8.2.1. By martingale properties, one has $E[u] = 0$ and $E[u^2] = E[w]$; see, e.g., Fleming and Harrington (1991, §2.7). See also §9.3.1. Since $\int h_2(t; \eta) dM(t)$ can be negative, one may set it to its mean value zero and use only the second term of w, $\int h_1^2(t; \eta) dA(t)$, which is always positive.

8.6.2 Kullback-Leibler and Cross-Validation

Following the lines of §8.2.1, one has the Kullback-Leibler distance

$$\mathrm{KL}(\eta, \eta_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \left\{ \lambda(t, \eta_i) \log \frac{\lambda(t, \eta_i)}{\lambda(t, \eta_{\lambda,i})} - \lambda(t, \eta_i) + \lambda(t, \eta_{\lambda,i}) \right\} dt \tag{8.24}$$

and the relative Kullback-Leibler distance

$$\mathrm{RKL}(\eta, \eta_\lambda) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \left\{ \lambda(t, \eta_{\lambda,i}) - \lambda(t, \eta_i) \log \lambda(t, \eta_{\lambda,i}) \right\} dt, \tag{8.25}$$

where $\eta_{\lambda,i} = \eta_\lambda(U_i)$. Following (8.9), (8.25) is to be estimated by the cross-validation score

$$\frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \lambda(t, \eta_{\lambda,i}) dt - \frac{1}{n} \sum_{i=1}^{n} \delta_i \log \lambda\big(X_i, \eta_{\lambda,i}^{[i]}\big), \tag{8.26}$$

where $\eta_{\lambda,i}^{[i]} = \eta_\lambda^{[i]}(U_i)$ for $\eta_\lambda^{[i]}$ the delete-one estimate of η. The performance of η_λ can be assessed through the symmetrized Kullback-Leibler distance

$$L(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \big(\lambda(t, \eta_i) - \lambda(t, \eta_{\lambda,i})\big) \log \frac{\lambda(t, \eta_i)}{\lambda(t, \eta_{\lambda,i})} dt. \tag{8.27}$$

8.6.3 Weibull Family

For the Weibull family of Example 8.4, one has, for $\nu = 1/\sigma$,

$$l(\eta, \nu) = -\delta\big\{\nu(\log X - \eta) + \log \nu\big\} + (X^\nu - Z^\nu) e^{-\nu\eta}; \tag{8.28}$$

see Problem 8.8. Note that $\log \lambda(t, u) = \nu\big(\log t - \eta(u)\big) + \log(\nu/t)$, so the Weibull model is also a proportional hazard model, with the relative risk proportional to $e^{-\nu\eta(u)}$. It is easily seen that $h_1(t; \eta) = \nu$ and $h_2 = 0$.

Fixing ν, one may iterate on (5.3) using

$$\tilde{u}_i = \nu\big(\delta_i - (X_i^\nu - Z_i^\nu) e^{-\nu\tilde{\eta}_i}\big),$$
$$\tilde{w}_i = \nu^2 (X_i^\nu - Z_i^\nu) e^{-\nu\tilde{\eta}_i},$$

where $\tilde{\eta}_i = \tilde{\eta}(U_i)$. Fixing $\eta_i = \eta(U_i)$, one may estimate ν by minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\big(\nu(\log X_i - \eta_i) + \log\nu\big) - (X_i^\nu - Z_i^\nu)e^{-\nu\eta_i}\right\}$$

The situation is the same as in §5.4.6 for regression with negative binomial responses, and for ν unknown, one may alternate the updating of η and ν. To drive performance-oriented iteration, one may use $U_w(\lambda)$ with $\sigma^2 = 1$.

Kullback-Leibler and Direct Cross-Validation

With $\log\lambda(t,\eta) = \nu(\log t - \eta) + \log(\nu/t)$, (8.26) looks like

$$\frac{1}{n}\sum_{i=1}^{n}\left\{(X_i^\nu - Z_i^\nu)e^{-\nu\eta_{\lambda,i}} - \delta_i\big(\nu(\log X_i - \eta_{\lambda,i}) + \log\nu\big)\right\} + \frac{\nu}{n}\sum_{i=1}^{n}\delta_i\big(\eta_{\lambda,i}^{[i]} - \eta_{\lambda,i}\big),$$

where $\delta_i\big(\eta_{\lambda,i}^{[i]} - \eta_{\lambda,i}\big) = \delta_i\big\{\eta_\lambda^{[i]}(U_i) - \eta_\lambda(U_i)\big\}$ are non-negative. Replacing $\delta_i\big(\eta_{\lambda,i}^{[i]} - \eta_{\lambda,i}\big)$ by $\delta_i\big|\eta_{\lambda,\eta_\lambda}^{[i]}(U_i) - \eta_\lambda(U_i)\big|$, the lines leading to (5.18) yields

$$V_g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{(X_i^\nu - Z_i^\nu)e^{-\nu\eta_{\lambda,i}} - \delta_i\big(\nu(\log X_i - \eta_{\lambda,i}) + \log\nu\big)\right\} + \alpha\frac{\mathrm{tr}(A_w W^{-1})}{n - \mathrm{tr}A_w}\frac{\nu}{n}\sum_{i=1}^{n}\delta_i|\tilde{u}_i| \quad (8.29)$$

for $\alpha = 1$, where terms not involving η can be dropped for ν known but are necessary for ν unknown. Fixing ν, (8.27) reads

$$L(\lambda) = \frac{\nu}{n}\sum_{i=1}^{n}(X_i^\nu - Z_i^\nu)(e^{-\nu\eta_i} - e^{-\nu\eta_{\lambda,i}})(\eta_{\lambda,i} - \eta_i), \quad (8.30)$$

and the Kullback-Leibler projection of $\hat{\eta}$ minimizes

$$\mathrm{KL}(\hat{\eta},\eta) = \frac{1}{n}\sum_{i=1}^{n}(X_i^\nu - Z_i^\nu)\big\{\nu e^{-\nu\hat{\eta}_i}(\eta_i - \hat{\eta}_i) + e^{-\nu\eta_i} - e^{-\nu\hat{\eta}_i}\big\},$$

where $\hat{\eta}_i = \hat{\eta}(U_i)$.

Empirical Performance

Parallel to the simulations for the families of §5.4, Weibull failure times $T_i|u_i$ were drawn on $u_i = (i - 0.5)/100$, $i = 1,\ldots,100$ with $\nu = 2$ and

$$\beta(u) = e^{\eta(u)} = 3\big\{10^5 u^{11}(1-u)^6 + 10^3 u^3(1-u)^{10}\big\} + 1, \quad (8.31)$$

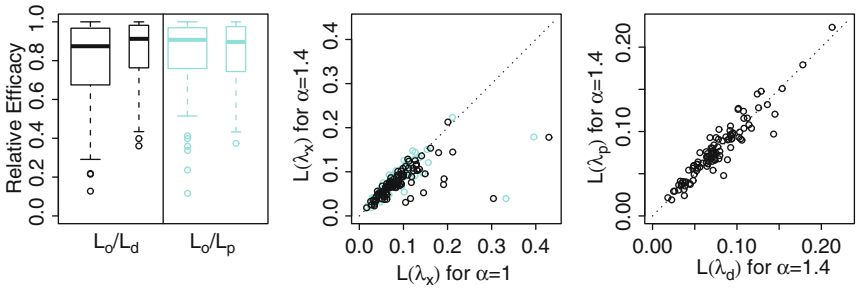


FIGURE 8.7. Effectiveness of $V_g(\lambda)$ and $U_w(\lambda)$ in Weibull simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1.4$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

along with exponential censoring times satisfying $P(C_i \geq c) = e^{-c/2\beta(u_i)}$ and truncation times satisfying $P(Z_i \geq z) = e^{-2z/\beta(u_i)}$.

For each of the one hundred replicates generated, five cubic splines were fitted to the log scale function $\eta(u)$, one minimizing $L(\lambda)$ of (8.30) at $L(\lambda_o)$, two from performance-oriented iteration driven by $U_w(\lambda)$ for $\alpha = 1, 1.4$ with performances $L(\lambda_p)$, and two minimizing $V_g(\lambda)$ of (8.29) for $\alpha = 1, 1.4$ with performances $L(\lambda_d)$. The results are summarized in Fig. 8.7. The fudge factor $\alpha = 1.4$ helps both methods, but the choice between direct and indirect cross-validation seems to be a toss up.

Software Illustration

The following sequence generates a sample of $(X_i, \delta_i, Z_i)|u_i$ used in the simulation above and fits a cubic spline to the log scale function using performance-oriented iteration, with $\nu = 2$ known:

```
test <- function(x)
          {.3*(1e6*(x^11*(1-x)^6)+1e4*(x^3*(1-x)^10))+1}
rtest.wei <- function(u) {
  mu <- test(u)
  tt <- rweibull(u,2,mu)
  cens <- rweibull(u,1,2*mu)
  z <- rweibull(u,1,mu/2)
  x <- pmin(tt,cens)
  delta <- tt<=cens
  ok <- x>z
  while(m <- sum(!ok)) {
    tt[!ok] <- rweibull(m,2,mu[!ok])
    cens[!ok] <- rweibull(m,1,2*mu[!ok])
    z[!ok] <- rweibull(m,1,mu[!ok]/2)
    x[!ok] <- pmin(tt[!ok],cens[!ok])
```

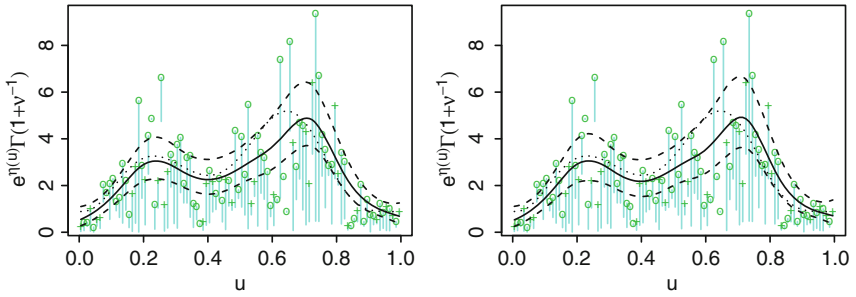


FIGURE 8.8. Cubic spline Weibull regression with censored and truncated data. The estimated $E[T|u] = e^{\eta(u)}\Gamma(1+\nu^{-1})$ are in *solid lines*, the 95 % Bayesian confidence intervals in *dashed lines*, and the test function in *dotted lines*. The data are superimposed as *circles* (failures) or *pluses* (censorings) along with at-risk processes $I_{[Z_i<t\le X_i]}$ in *faded vertical lines*. *Left*: Estimate via indirect cross-validation with a known $\nu = 2$. *Right*: Estimate via direct cross-validation with an estimated $\nu = 1.88$.

```
    delta[!ok] <- tt[!ok]<=cens[!ok]
    ok <- x>z
  }
  cbind(x,delta,z)
}
u <- ((1:100)-.5)/100
set.seed(2375); y <- rtest.wei(u)
fit1.wei <- gssanova1(y~u,"weibull",nu=2)
```

where `y` should have at least two columns containing (X_i, δ_i). The fit can then be plotted as in the left frame of Fig. 8.8:

```
est1 <- predict(fit1.wei,data.frame(u=u),se=TRUE)
plot(u,y[,1],type="n")
for (i in 1:100)
  lines(c(u[i],u[i]),c(y[i,1],y[i,3]),col=5)
points(u,y[,1],pch=c("+","o")[y[,2]+1])
gg <- gamma(1+1/fit1.wei$nu)
lines(u,gg*exp(est1$fit))
lines(u,gg*exp(est1$fit+1.96*est1$se),lty=5)
lines(u,gg*exp(est1$fit-1.96*est1$se),lty=5)
lines(u,gg*test(u),lty=3)
```

A fit through direct cross-validation can be similarly obtained, as plotted in the right frame of Fig. 8.8, with an estimated $\nu = 1.88$:

```
fit.wei <- gssanova(y~u,"weibull",id.basis=fit1.wei$id)
fit.wei$nu
# 1.881233
```

8.6.4 Log Normal Family

For the log normal family of Example 8.5, one has, for $\nu = 1/\sigma$,

$$l(\eta, \nu) = -\delta\big(\log\phi(\breve{z}) - \log\big(1 - \Phi(\breve{z})\big) + \log\nu\big) + \log\frac{1 - \Phi(\tilde{z})}{1 - \Phi(\breve{z})}, \quad (8.32)$$

where $\breve{z} = \nu(\log X - \eta)$ and $\tilde{z} = \nu(\log Z - \eta)$; see Problem 8.9. It is easy to verify that $h_1(t;\eta) = \nu\big\{\phi(z)/\big(1 - \Phi(z)\big) - z\big\}$, where $z = \nu(\log t - \eta)$. Fixing ν, one may iterate on (5.3) using

$$\tilde{u}_i = \nu\,\delta_i\left(\frac{\phi(\breve{z}_i)}{1 - \Phi(\breve{z}_i)} - \breve{z}_i\right) - \nu\left(\frac{\phi(\breve{z}_i)}{1 - \Phi(\breve{z}_i)} - \frac{\phi(\tilde{z}_i)}{1 - \Phi(\tilde{z}_i)}\right),$$

$$\tilde{w}_i = \int_{Z_i}^{X_i} h_1^2(t;\eta_i)\lambda(t;\eta_i,\nu)dt = \int_{Z_i}^{X_i} \nu^2\left(\frac{\phi(z)}{1 - \Phi(z)} - z\right)^2 \frac{\phi(z)}{1 - \Phi(z)}\frac{\nu dt}{t},$$

where $\breve{z}_i = \nu(\log X_i - \eta_i)$ and $\tilde{z}_i = \nu(\log Z_i - \eta_i)$. It can be shown that

$$\tilde{w}_i = \nu^2\Bigg\{\left(\frac{1}{2}\left(\frac{\phi(\breve{z}_i)}{1 - \Phi(\breve{z}_i)}\right)^2 - \frac{\breve{z}_i\phi(\breve{z}_i)}{1 - \Phi(\breve{z}_i)} - \log\big(1 - \Phi(\breve{z}_i)\big)\right)$$
$$- \left(\frac{1}{2}\left(\frac{\phi(\tilde{z}_i)}{1 - \Phi(\tilde{z}_i)}\right)^2 - \frac{\tilde{z}_i\phi(\tilde{z}_i)}{1 - \Phi(\tilde{z}_i)} - \log\big(1 - \Phi(\tilde{z}_i)\big)\right)\Bigg\}; \quad (8.33)$$

see Problem 8.10. Fixing $\eta_i = \eta(U_i)$, one may estimate ν via minimizing

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\big(\log\phi(\breve{z}_i) - \log\big(1 - \Phi(\breve{z}_i)\big) + \log\nu\big) - \log\frac{1 - \Phi(\tilde{z}_i)}{1 - \Phi(\breve{z}_i)}\right\}.$$

To drive performance-oriented iteration, one may use $U_w(\lambda)$ with $\sigma^2 = 1$.

Kullback-Leibler and Direct Cross-Validation

With $\log\lambda(t,\eta) = \log\phi(z) - \log\big(1 - \Phi(z)\big) + \log(\nu/t)$, (8.26) looks like

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{1 - \Phi(\tilde{z}_i)}{1 - \Phi(\breve{z}_i)} - \delta_i\big(\log\phi(\breve{z}_i) - \log\big(1 - \Phi(\breve{z}_i)\big) + \log\nu\big)\right\}$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\delta_i\big(\log\lambda(X_i,\eta_{\lambda,i}) - \log\lambda(X_i,\eta_{\lambda,i}^{[i]})\big).$$

Replacing the non-negative $\delta_i\big(\log\lambda(X_i,\eta_{\lambda,i}) - \log\lambda(X_i,\eta_{\lambda,i}^{[i]})\big)$ by a linear approximation $\delta_i\big|h_1(X_i,\eta_{\lambda,i})\big(\eta_{\lambda,\eta_\lambda}^{[i]}(U_i) - \eta_\lambda(U_i)\big)\big|$, one is led to

$$V_g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{1 - \Phi(\tilde{z}_i)}{1 - \Phi(\breve{z}_i)} - \delta_i\big(\log\phi(\breve{z}_i) - \log\big(1 - \Phi(\breve{z}_i)\big) + \log\nu\big)\right\}$$
$$+ \alpha\frac{\mathrm{tr}(A_wW^{-1})}{n - \mathrm{tr}A_w}\frac{1}{n}\sum_{i=1}^{n}\delta_i\big|h_1(X_i,\eta_{\lambda,i})\tilde{u}_i\big| \quad (8.34)$$

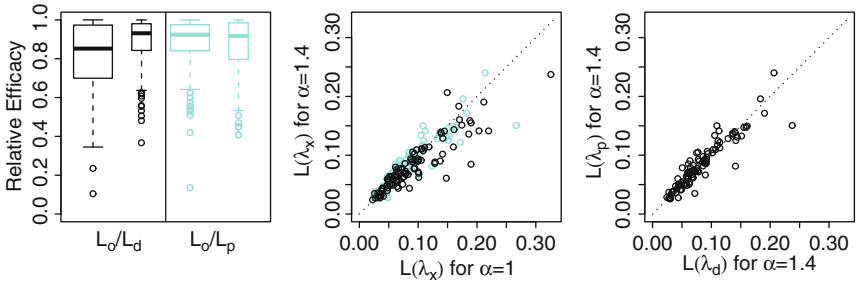


FIGURE 8.9. Effectiveness of $V_g(\lambda)$ and $U_w(\lambda)$ in log normal simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1.4$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

for $\alpha = 1$. $L(\lambda)$ of (8.27) does not simplify further and the Kullback-Leibler projection of $\hat{\eta}$ minimizes $\mathrm{KL}(\hat{\eta}, \eta)$ as defined in (8.24).

Empirical Performance

Log normal failure times $T_i|u_i$ were drawn, with $\nu = 2$ and $\beta(u) = e^{\eta(u)}$ as in (8.31), on $u_i = (i - 0.5)/100$, $i = 1, \ldots, 100$, along with exponential censoring times satisfying $P(C_i \geq c) = e^{-c/2\beta(u_i)}$ and truncation times satisfying $P(Z_i \geq z) = e^{-2z/\beta(u_i)}$. Results from one hundred replicates are shown in Fig. 8.9; one replicate is off the chart in the center frame, with $L(\lambda_d)$ at $(0.579, 0.061)$ and $L(\lambda_p)$ at $(0.445, 0.061)$.

Software Illustration

The following sequence generates a sample of $(X_i, \delta_i, Z_i)|u_i$ used in the simulation above and fits a cubic spline to $\eta(u)$ using performance-oriented iteration, with $\nu = 2$ known; `rtest.lognorm` is nearly a duplicate of `rtest.wei` in §8.6.3 so only a few lines are listed here:

```
rtest.lognorm <- function(u) {
  mu <- test(u)
  tt <- exp(rnorm(u)/2+log(mu))
  ...
  while(m <- sum(!ok)) {
    tt[!ok] <- exp(rnorm(m)/2+log(mu[!ok]))
    ...
  }
  cbind(x,delta,z)
}
u <- ((1:100)-.5)/100
set.seed(2375); y <- rtest.lognorm(u)
fit1.lognorm <- gssanova1(y~u,"lognorm",nu=2)
```

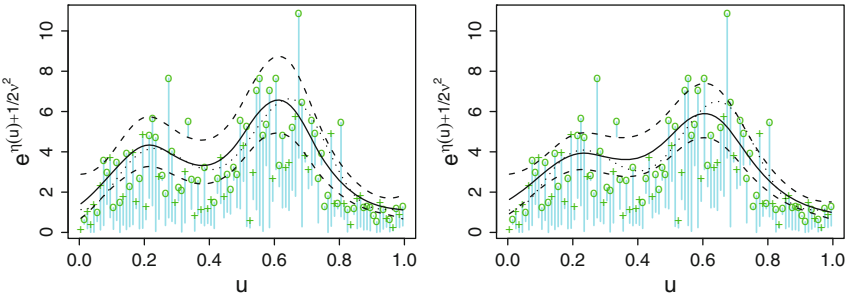


FIGURE 8.10. Cubic spline log normal regression with censored and truncated data. The estimated $E[T|u] = e^{\eta(u)+1/2\nu^2}$ are in *solid lines*, the 95 % Bayesian confidence intervals in *dashed lines*, and the test function in *dotted lines*. The data are superimposed as *circles* (failures) or *pluses* (censorings) along with at-risk processes $I_{[Z_i<t\le X_i]}$ in *faded vertical lines*. *Left*: Estimate via indirect cross-validation with a known $\nu = 2$. *Right*: Estimate via direct cross-validation with an estimated $\nu = 2.07$.

The fit can then be plotted as in the left frame of Fig. 8.10:

```
est1 <- predict(fit1.lognorm,data.frame(u=u),se=TRUE)
plot(u,y[,1],type="n")
for (i in 1:100)
  lines(c(u[i],u[i]),c(y[i,1],y[i,3]),col=5)
points(u,y[,1],pch=c("+","o")[y[,2]+1])
gg <- exp(1/2/fit1.lognorm$nu^2)
lines(u,gg*exp(est1$fit))
lines(u,gg*exp(est1$fit+1.96*est1$se),lty=5)
lines(u,gg*exp(est1$fit-1.96*est1$se),lty=5)
lines(u,gg*test(u),lty=3)
```

A fit through direct cross-validation can be similarly obtained, as plotted in the right frame of Fig. 8.10, with an estimated $\nu = 2.07$:

```
fit.lognorm <- gssanova(y~u,"lognorm",
                        id.basis=fit1.lognorm$id.basis)
fit.lognorm$nu
# 2.067286
```

8.6.5 Log Logistic Family

For the log logistic family of Example 8.6, one has, for $\nu = 1/\sigma$,

$$l(\eta,\nu) = -\delta\big(\check{z} - \log(1+e^{\check{z}}) + \log\nu\big) + \log\frac{1+e^{\check{z}}}{1+e^{\bar{z}}}, \tag{8.35}$$

where $\check{z} = \nu(\log X - \eta)$ and $\tilde{z} = \nu(\log Z - \eta)$; see Problem 8.11. Since $h_1(t;\eta) = \nu/(1+e^z)$, where $z = \nu(\log t - \eta)$, one may iterate on (5.3) using

$$\tilde{u}_i = \frac{\nu\delta_i}{1+e^{\check{z}_i}} - \nu\left(\frac{1}{1+e^{\check{z}_i}} - \frac{1}{1+e^{\tilde{z}_i}}\right),$$
$$\tilde{w}_i = \int_{Z_i}^{X_i} h_1^2(t;\eta_i)\lambda(t;\eta_i,\nu)dt = \int_{Z_i}^{X_i} \frac{\nu^2}{(1+e^z)^2}\frac{e^z}{1+e^z}\frac{\nu dt}{t}$$
$$= \frac{\nu^2}{2}\left(\frac{1}{(1+e^{\tilde{z}_i})^2} - \frac{1}{(1+e^{\check{z}_i})^2}\right).$$

Fixing $\eta_i = \eta(U_i)$, one may estimate ν through the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\big(\check{z}_i - \log(1+e^{\check{z}_i}) + \log\nu\big) - \log\frac{1+e^{\check{z}_i}}{1+e^{\tilde{z}_i}}\right\},$$

To drive performance-oriented iteration, one may use $U_w(\lambda)$ with $\sigma^2 = 1$.

Kullback-Leibler and Direct Cross-Validation

With $\log\lambda(t,\eta) = z - \log(1+e^z) + \log(\nu/t)$, (8.26) looks like

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{1+e^{\check{z}_i}}{1+e^{\tilde{z}_i}} - \delta_i\big(\check{z}_i - \log(1+e^{\check{z}_i}) + \log\nu\big)\right\}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\delta_i\big(\log\lambda(X_i,\eta_{\lambda,i}) - \log\lambda(X_i,\eta_{\lambda,i}^{[i]})\big),$$

and (8.34) becomes

$$V_g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\log\frac{1+e^{\check{z}_i}}{1+e^{\tilde{z}_i}} - \delta_i\big(\check{z}_i - \log(1+e^{\check{z}_i}) + \log\nu\big)\right\}$$
$$+\alpha\frac{\mathrm{tr}(A_wW^{-1})}{n - \mathrm{tr}A_w}\frac{\nu}{n}\sum_{i=1}^{n}\frac{\delta_i|\tilde{u}_i|}{1+e^{\check{z}_i}}. \quad (8.36)$$

$L(\lambda)$ of (8.27) does not simplify further and the Kullback-Leibler projection of $\hat{\eta}$ minimizes $\mathrm{KL}(\hat{\eta},\eta)$ as defined in (8.24).

Empirical Performance

Log logistic failure times $T_i|u_i$ were drawn, with $\nu = 2$ and $\beta(u) = e^{\eta(u)}$ as in (8.31), on $u_i = (i-0.5)/100$, $i = 1,\ldots,100$, along with exponential censoring times satisfying $P(C_i \geq c) = e^{-c/2\beta(u_i)}$ and truncation times satisfying $P(Z_i \geq z) = e^{-2z/\beta(u_i)}$. Results from one hundred replicates are shown in Fig. 8.11; two faded points are off the chart in the center frame, with $L(\lambda_p)$ at $(0.635, 0.018)$ and $(0.865, 0.186)$.

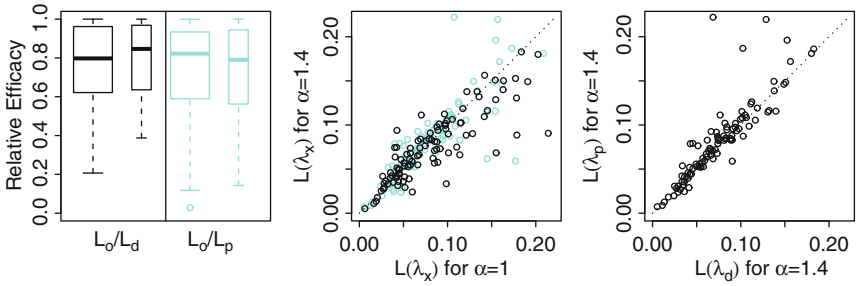


FIGURE 8.11. Effectiveness of $V_g(\lambda)$ and $U_w(\lambda)$ in log logistic simulation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_d)$ (*solid*) and $L(\lambda_o)/L(\lambda_p)$ (*faded*), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_d)$ (*solid*) or $L(\lambda_p)$ (*faded*) with $\alpha = 1$ versus those with $\alpha = 1.4$. *Right*: $L(\lambda_d)$ with $\alpha = 1.4$ versus $L(\lambda_p)$ with $\alpha = 1.4$.

Software Illustration

The following sequence generates a sample of $(X_i, \delta_i, Z_i)|u_i$ used in the simulation above and fits a cubic spline to $\eta(u)$ using performance-oriented iteration, with $\nu = 2$ known; `rtest.loglogis` is nearly a duplicate of `rtest.wei` in §8.6.3 so only a few lines are listed here:

```
rtest.loglogis <- function(u) {
  mu <- test(u)
  tt <- exp(rlogis(u)/2+log(mu))
  ...
  while(m <- sum(!ok)) {
    tt[!ok] <- exp(rlogis(m)/2+log(mu[!ok]))
    ...
  }
  cbind(x,delta,z)
}
u <- ((1:100)-.5)/100
set.seed(2375); y <- rtest.loglogis(u)
fit1.loglogis <- gssanova1(y~u,"loglogis",nu=2)
```

The fit can then be plotted as in the left frame of Fig. 8.12:

```
est1 <- predict(fit1.loglogis,data.frame(u=u),se=TRUE)
plot(u,y[,1],type="n")
for (i in 1:100)
  lines(c(u[i],u[i]),c(y[i,1],y[i,3]),col=5)
points(u,y[,1],pch=c("+","o")[y[,2]+1])
gg <- gamma(1+1/fit1.loglogis$nu)*
        gamma(1-1/fit1.loglogis$nu)
lines(u,gg*exp(est1$fit))
lines(u,gg*exp(est1$fit+1.96*est1$se),lty=5)
```

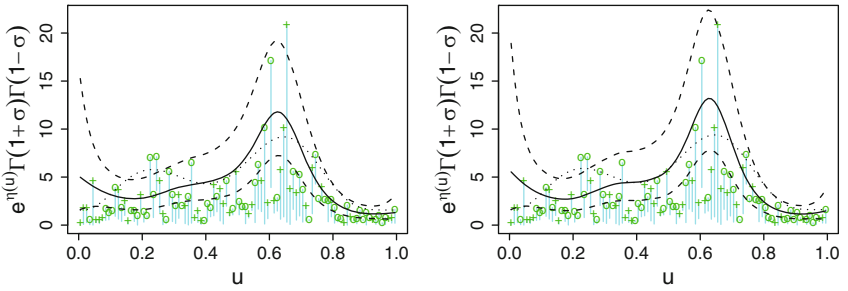


FIGURE 8.12. Cubic spline log logistic regression with censored and truncated data. The estimated $E[T|u] = e^{\eta(u)}\Gamma(1+\sigma)\Gamma(1-\sigma)$ are in *solid lines*, the 95 % Bayesian confidence intervals in *dashed lines*, and the test function in *dotted lines*. The data are superimposed as *circles* (failures) or *pluses* (censorings) along with at-risk processes $I_{[Z_i<t\le X_i]}$ in *faded vertical lines*. *Left*: Estimate via indirect cross-validation with a known $\nu = 2$. *Right*: Estimate via direct cross-validation with an estimated $\nu = 1.96$.

```
lines(u,gg*exp(est1$fit-1.96*est1$se),lty=5)
lines(u,gg*test(u),lty=3)
```

A fit through direct cross-validation can be similarly obtained, as plotted in the right frame of Fig. 8.12, with an estimated $\nu = 1.96$:

```
fit.loglogis <- gssanova(y~u,"loglogis",
                         id.basis=fit1.loglogis$id.basis)
fit.loglogis$nu
# 1.960357
```

8.6.6 Case Study: Survival After Heart Transplant

The following sequence loads the Stanford heart transplant data of §§1.4.3 and 8.4.2 and fits a Weibull model to the data:

```
data(stan)
fit3.stan <- gssanova(cbind(time+.01,status)~age,
                      data=stan,family="weibull",
                      nbasis=200)
```

The follow-up times in the records were rounded to whole days and there was a recorded death at 0, and we choose to add 0.01 to the follow-up times instead of deleting the 0. With an ANOVA decomposition $\eta(u) = \eta_\emptyset + \eta_u(u)$, the relative risk is given by $\lambda_1(u) = e^{-\nu\eta_u(u)}$, which can be plotted as shown in the left frame of Fig. 8.13, where the estimate via penalized partial likelihood seen in the right frame of Fig. 8.5 is superimposed:

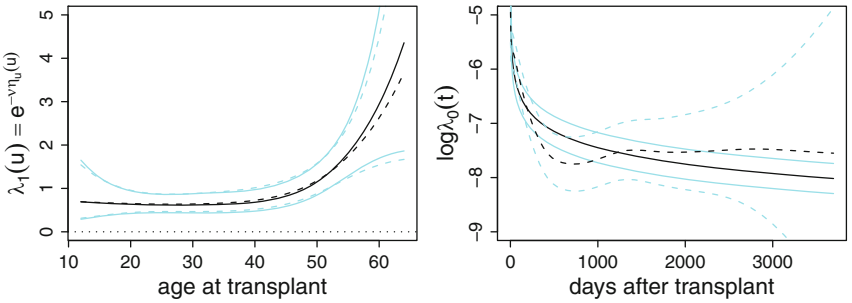


FIGURE 8.13. Hazard after heart transplant: Weibull fit. *Left*: The fitted relative risk $\lambda_1(u) = e^{-\nu\eta_u(u)}$ (*solid*) along with 95 % Bayesian confidence intervals (*faded*). *Right*: The fitted log base hazard $\log\lambda_0(t) = \nu(\log y - \eta_\emptyset) + \log(\nu/t)$ (*solid*) along with 95 % Bayesian confidence intervals (*faded*). The estimates via penalized partial likelihood seen in the *right* and *center* frames of Fig. 8.5 are superimposed in *dashed lines*.

```
nu <- fit3.stan$nu
u.gd <- seq(min(stan$age),max(stan$age),length=51)
est3 <- predict(fit3.stan,data.frame(age=u.gd),
                se=TRUE,inc="age")
plot(u.gd,exp(-nu*est3$fit),type="l",ylim=c(0,5))
lines(u.gd,exp(-nu*(est3$fit-1.96*est3$se)),col=5)
lines(u.gd,exp(-nu*(est3$fit+1.96*est3$se)),col=5)
lines(u.gd,est2.a$fit,lty=2)
lines(u.gd,est2.a$fit*exp(1.96*est2.a$se),lty=2,col=5)
lines(u.gd,est2.a$fit/exp(1.96*est2.a$se),lty=2,col=5)
```

where `est2.a` is from §8.5.4. The log base hazard is seen to be $\log\lambda_0(t) = \nu(\log t - \eta_\emptyset) + \log(\nu/t)$, which can be plotted as shown in the right frame of Fig. 8.13, where the estimate through penalized partial likelihood seen in the center frame of Fig. 8.5 is superimposed:

```
est3.b <- predict(fit3.stan,data.frame(age=35),
                  se=TRUE,inc="1")
t.gd <- seq(0,max(stan$futime),length=51)
lhzd <- nu*(2*log(t.gd)-est3.b$fit)+log(nu/t.gd^2)
plot(t.gd^2,lhzd,type="l",ylim=c(-9,-5))
lines(t.gd^2,lhzd-nu*1.96*est3.b$se,col=5)
lines(t.gd^2,lhzd+nu*1.96*est3.b$se,col=5)
lines(t.gd^2,log(est2.b$fit/2/t.gd),lty=2)
lines(t.gd^2,log(est2.b$fit/2/t.gd)+1.96*est2.b$se,
      lty=2,col=5)
lines(t.gd^2,log(est2.b$fit/2/t.gd)-1.96*est2.b$se,
      lty=2,col=5)
```

where `est2.b` is from §8.5.4. The estimates of relative risk are close to each other, while the estimates of the base hazard differ quite a bit, as can be expected.

8.7 Bibliographic Notes

Section 8.1

Absent of covariate, penalized likelihood hazard estimation was studied by Anderson and Senthilselvan (1980), Bartoszyński, Brown, McBride, and Thompson (1981), O'Sullivan (1988a), Antoniadis (1989), and Gu (1994). With covariate, the estimation of the "bivariate" hazard function through penalized full likelihood was formulated and studied by Gu (1996, 1998c).

Section 8.2

A performance-oriented iteration similar to that in §5.2.1 and Gu (1993b) was proposed and illustrated by Gu (1994) for $\mathcal{U}$ a singleton, where a martingale moment estimate similar to (8.9) was used to derive an indirect cross-validation score. The direct cross-validation score presented here is adapted from §7.3.

A comprehensive treatment of the counting process approach to survival analysis and the related martingale structure can be found in Fleming and Harrington (1991). A technically less demanding exposition can be found in Gill (1984).

Section 8.3

Bayesian confidence intervals for log hazard were derived and illustrated in Du and Gu (2006).

The Kullback-Leibler projection was developed in Gu (2004).

The frailty models for correlated data were studied in Du and Ma (2010).

Section 8.4

The gastric cancer data was used as an example by Moreau, O'Quigley, and Mesbah (1985) to illustrate their goodness-of-fit test for the proportional hazard model; the p-value of the test calculated on the data was between 0.01 and 0.02, indicating the inadequacy of the proportional hazard model.

The analysis of the Stanford heart transplant data presented here differs slightly from the one in Gu (1998c), where a performance-oriented iteration was used to select the smoothing parameters.

Section 8.5

Partial likelihood was proposed by Cox (1972) based on a conditioning argument, and maximum partial likelihood has become the golden standard for the parametric estimation of relative risk. Penalized partial likelihood was studied by O'Sullivan (1988b); see also Hastie and Tibshirani (1986) and Gray (1992). The isomorphism between partial likelihood and likelihood under biased sampling has its root in Cox's conditioning argument.

Zucker and Karr (1990) considered a generalization of the proportional hazard model of the form $\lambda(t,u) = \lambda_0(t)\lambda_1\big(\beta(t),u\big)$, where $\lambda_1\big(\beta(t),u\big)$ was parametric in u with a time-varying parameter $\beta(t)$, and $\beta(t)$ was estimated via penalized partial likelihood.

Section 8.6

Accelerated life models are among classical tools in reliability and survival analysis; see, e.g., Kalbfleisch and Prentice (1980, §2.3). Basic properties of the Weibull, the log normal, and the log logistic distributions can be found in Kalbfleisch and Prentice (1980, §2.2) along with properties of other lifetime distributions. Parametric linear models for $\mu(u)$ have been implemented by Terry Therneau in his `survival` package, ported to R from the Splus original by Thomas Lumley.

The direct cross-validation scores have not appeared in the literature.

8.8 Problems

Section 8.1

8.1 Verify (8.5).

Section 8.2

8.2 Using the calculus leading to (7.16) on page 244, one can obtain the quadratic approximation of (8.1).

(a) Define $L_{f,g}(\alpha) = (1/n)\sum_{i=1}^{n}\int_{Z_i}^{X_i} e^{(f+\alpha g)(t,U_i)}dt$, where f and g are functions and α is real. Calculate $\dot{L}_{f,g}(0)$ and $\ddot{L}_{f,g}(0)$.

(b) Obtain the quadratic approximation of (8.1) at $\tilde{\eta}$.

8.3 Verify (8.10).

Section 8.3

8.4 Consider a constant hazard model $\lambda(t,u)=\lambda$ for (X_i,δ_i,Z_i,U_i). Show that the maximum likelihood estimate is given by

$$\hat{\lambda}=\sum_{i=1}^n \delta_i/\sum_{i=1}^n (X_i-Z_i).$$

8.5 Plugging (8.3) into (8.15), derive the Newton updating equation for minimizing (8.15) with respect $(\mathbf{d}^T,\mathbf{c}^T,\mathbf{b}^T)^T$.

Section 8.5

8.6 Discuss basic properties of (8.17), such as the existence and uniqueness of the minimizer.

8.7 Applying the techniques developed in §7.6 to the estimation of relative risk via the minimization of (8.17), characterize the Kullback-Leibler loss that is targeted by cross-validation.

Section 8.6

8.8 Verify the minus log likelihood (8.28) for the Weibull family.

8.9 Verify the minus log likelihood (8.32) for the log normal family.

8.10 Verify (8.33) for $\int_{Z_i}^{X_i} h_1^2(t;\eta_i)\lambda(t;\eta_i,\nu)dt$.

(a) Verify that

$$\frac{d}{dz}\left(\frac{\phi(z)}{1-\Phi(z)}\right)=\left(\frac{\phi(z)}{1-\Phi(z)}-z\right)\frac{\phi(z)}{1-\Phi(z)}.$$

(b) Calculate (8.33) via integration by parts.

8.11 Verify the minus log likelihood (8.35) for the log logistic family.

9
Asymptotic Convergence

In this chapter, we develop an asymptotic theory concerning the rates of convergence of penalized likelihood estimates to the target functions as the sample size goes to infinity. The rates are calculated in terms of problem-specific loss functions derived from the respective stochastic settings.

The primary tool used in the development is the eigenvalue analysis in a Hilbert space, of which a brief introduction is given in §9.1. Convergence rates are established in §9.2 for the density estimates of Chap. 7, in §9.3 for the hazard estimates of §§8.1–8.4, and in §9.4 for the regression estimates of Chaps. 3, 5 and §8.6. For density estimation and hazard estimation, the notion of efficient approximation allows the practical computation of the estimates. For regression, the theory is developed in a setting more general than that of §5.1.

When an estimate is sought in a space $\mathcal{H}$ for the target function $\eta_0 \notin \mathcal{H}$, the estimate converges to a Kullback-Leibler projection η_0^* of η_0 in $\mathcal{H}$, at the same rates as established for the convergence to $\eta_0 \in \mathcal{H}$.

9.1 Preliminaries

Let $V(f)$ be a quadratic functional that defines a statistically interpretable metric so that a small $V(\hat{\eta} - \eta)$ indicates a good estimate $\hat{\eta}$ of η. The asymptotic convergence rates of penalized likelihood estimates can be characterized through an eigenvalue analysis of $J(f)$ with respect to

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_9,

$V(f)$, to be discussed below. Following the convention of §2.1.1, abstract concepts are set in boldface at the point of definition and are followed by simple examples set in italic.

A quadratic functional B is said to be **completely continuous** with respect to another quadratic functional A, if for any $\epsilon > 0$, there exist a finite number of linear functionals $L_1, \ldots, L_k$ such that $L_j f = 0$, $j = 1, \ldots, k$, implies that $B(f) \le \epsilon A(f)$; see Weinberger (1974, §3.3).

Consider the space $\mathcal{P}[0,1]$ of periodic functions permitting the Fourier series expansion (4.2) on page 127. Define $B(f) = 2\int_0^1 f^2 dx$, $A(f) = 2\int_0^1 \left(f^{(m)}\right)^2 dx$, and

$$L_{2\mu} f = \int_0^1 f(x) \sin 2\pi\mu x \, dx,$$
$$L_{2\mu+1} f = \int_0^1 f(x) \cos 2\pi\mu x \, dx, \qquad \mu = 0, 1, \ldots.$$

A function f satisfying $L_j f = 0$, $j = 1, \ldots, 2k-1$, has an expression

$$f(x) = \sum_{\mu=k}^{\infty} (a_\mu \cos 2\pi\mu x + b_\mu \sin 2\pi\mu x)$$

and, consequently,

$$B(f) = \sum_{\mu=k}^{\infty} (a_\mu^2 + b_\mu^2) \le \frac{1}{(2\pi k)^{2m}} \sum_{\mu=k}^{\infty} (a_\mu^2 + b_\mu^2)(2\pi\mu)^{2m} = \frac{1}{(2\pi k)^{2m}} A(f).$$

Hence, B is completely continuous with respect to A.

When B is completely continuous with respect to A and, hence, to $A + B$, there exist **eigenvalues** λ_ν and the associated **eigenfunctions** ψ_ν such that

$$B(\psi_\nu, \psi_\mu) = \lambda_\nu \delta_{\nu,\mu}, \qquad (A + B)(\psi_\nu, \psi_\mu) = \delta_{\nu,\mu},$$

where $\delta_{\nu,\mu}$ is the Kronecker delta and $1 \ge \lambda_\nu \downarrow 0$; see Theorem 3.1 of Weinberger (1974, p. 52). Write $\phi_\nu = \lambda_\nu^{-1/2} \psi_\nu$. It follows that

$$B(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}, \qquad A(\phi_\nu, \phi_\mu) = \rho_\nu \delta_{\nu,\mu},$$

where $0 \le \rho_\nu = \lambda_\nu^{-1} - 1 \uparrow \infty$. We refer to ρ_ν as the eigenvalues of A with respect to B and to ϕ_ν as the associated eigenfunctions. Functions satisfying $A(f) < \infty$ can be expressed as a **Fourier series expansion** $f = \sum_\nu f_\nu \phi_\nu$, where $f_\nu = B(f, \phi_\nu)$ are the **Fourier coefficients**.

Take $\phi_{2\mu} = \sin 2\pi\mu x$, $\phi_{2\mu+1} = \cos 2\pi\mu x$, $\mu = 0, 1, \ldots$, in the periodic function example given above. It is easy to see that

$$B(\phi_\nu, \phi_\mu) = \delta_{\nu,\mu}, \quad A(\phi_\nu, \phi_\mu) = (2\pi\lfloor \nu/2 \rfloor)^{2m} \delta_{\nu,\mu}, \qquad \nu, \mu = 1, 2, \ldots,$$

where $\lfloor \nu/2 \rfloor$ *is the integer part of* $\nu/2$. *The eigenvalues* $\rho_\nu = (2\pi\lfloor \nu/2 \rfloor)^{2m}$ *grow at a rate* ν^{2m}. *The Fourier coefficients are given by* $f_{2\mu} = b_\mu$, $f_{2\mu+1} = a_\mu$, $\mu = 0, 1, \ldots$.

To possibly achieve noise reduction in estimation, the effective dimension of the model space has to be kept finite, and to make the procedure non-restrictive, the dimension has to be expandable when more data become available. When V is completely continuous with respect to J, this can be achieved through constraints of the form $J(f) \le \rho$ with $\rho \to \infty$ as $n \to \infty$ or, equivalently, by Theorem 2.12, through penalized likelihood with $\lambda \to 0$ as $n \to \infty$. The growth rate of the eigenvalues ρ_ν of J with respect to V, which typically is at ν^r for some $r > 1$, dictates how fast λ should approach 0, as will be seen in the sections to follow.

A few examples are given in the rest of the section.

Example 9.1 (Polynomial splines) Consider $J(f) = \int_0^1 \left(f^{(m)}\right)^2 dx$ and $V(f) = \int_0^1 f^2 w(x) dx$ on $\mathcal{X} = [0, 1]$, where $w(x)$ satisfies $0 < c_1 < w(x) < c_2 < \infty$ for some c_1, c_2. V is known to be completely continuous with respect to J, and it can be shown that $\rho_\nu \asymp \nu^{2m}$. See, e.g., Utreras (1981).

For $J(f) = \int_0^1 (Lf)^2 dx$ with L given in (4.75) on page 157, the same results hold as $\int_0^1 (Lf)^2 dx$ is equivalent to $\int_0^1 \left(f^{(m)}\right)^2 dx$. □

Let $\{\varphi_\nu\}$ be a sequence of functions on $[0, 1]$ satisfying $\int_0^1 \varphi_\nu \varphi_\mu dx = \delta_{\nu,\mu}$ and $\int_0^1 \ddot{\varphi}_\nu \ddot{\varphi}_\mu dx = \sigma_\nu \delta_{\nu,\mu}$, where $\nu^4 \asymp \sigma_\nu \uparrow \infty$. The first two entries are $\varphi_1 = 1$ and $\varphi_2 = \sqrt{12}(\cdot - 0.5)$, with $\sigma_1 = \sigma_2 = 0$.

Example 9.2 (Tensor product cubic spline) Consider $\mathcal{X} = [0, 1]^2$. Write $\tilde{V}(f) = \int_0^1 \int_0^1 f^2 dx_{\langle 1 \rangle} dx_{\langle 2 \rangle}$ and

$$\tilde{J}(f) = J_{1,00}(f) + J_{00,1}(f) + J_{1,01}(f) + J_{01,1}(f) + J_{1,1}(f),$$

where

$$\begin{aligned}
J_{1,00}(f) &= \int_0^1 \left\{ \int_0^1 \ddot{f}_{11} dx_{\langle 2 \rangle} \right\}^2 dx_{\langle 1 \rangle}, \\
J_{00,1}(f) &= \int_0^1 \left\{ \int_0^1 \ddot{f}_{22} dx_{\langle 1 \rangle} \right\}^2 dx_{\langle 2 \rangle}, \\
J_{1,01}(f) &= \int_0^1 \left\{ \int_0^1 f^{(3)}_{112} dx_{\langle 2 \rangle} \right\}^2 dx_{\langle 1 \rangle}, \\
J_{01,1}(f) &= \int_0^1 \left\{ \int_0^1 f^{(3)}_{122} dx_{\langle 1 \rangle} \right\}^2 dx_{\langle 2 \rangle}, \\
J_{1,1}(f) &= \int_0^1 \int_0^1 (f^{(4)}_{1122})^2 dx_{\langle 1 \rangle} dx_{\langle 2 \rangle}.
\end{aligned}$$

The sequence $\{\varphi_\nu(x_{\langle 1\rangle})\varphi_\mu(x_{\langle 2\rangle})\}$ are orthonormal with respect to $\tilde{V}(f,g)$ and are orthogonal with respect to $\tilde{J}(f,g)$. More precisely, $J_\beta(f)$'s define square norms in $\mathcal{H}_\beta$, where

$$\begin{aligned}
\mathcal{H}_{1,00} &= \{\varphi_\nu(x_{\langle 1\rangle})\varphi_1(x_{\langle 2\rangle})\}_{\nu\geq 3},\\
\mathcal{H}_{00,1} &= \{\varphi_1(x_{\langle 1\rangle})\varphi_\nu(x_{\langle 2\rangle})\}_{\nu\geq 3},\\
\mathcal{H}_{1,01} &= \{\varphi_\nu(x_{\langle 1\rangle})\varphi_2(x_{\langle 2\rangle})\}_{\nu\geq 3},\\
\mathcal{H}_{01,1} &= \{\varphi_2(x_{\langle 1\rangle})\varphi_\nu(x_{\langle 2\rangle})\}_{\nu\geq 3},\\
\mathcal{H}_{1,1} &= \{\varphi_\nu(x_{\langle 1\rangle})\varphi_\mu(x_{\langle 2\rangle})\}_{\nu,\mu\geq 3}.
\end{aligned}$$

The null space of $\tilde{J}(f)$ is given by $\mathcal{N}_{\tilde{J}} = \{\varphi_\nu(x_{\langle 1\rangle})\varphi_\mu(x_{\langle 2\rangle})\}_{\nu,\mu=1,2}$. Putting $\{\sigma_\nu\sigma_\mu\}_{\nu,\mu\geq 3}$ in an increasing order as $\{\tilde{\sigma}_\nu\}$, it can be shown that $\tilde{\sigma}_\nu$ grow at a rate faster than $(\nu/\log\nu)^4$ but slower than ν^4; see, e.g., Wahba (1990, §12.1).

When $w(x)$ is bounded away from 0 and ∞, $V(f) = \int_0^1\int_0^1 wf^2 dx_{\langle 1\rangle}dx_{\langle 2\rangle}$ is equivalent to $\tilde{V}(f)$. For $\theta_\beta > 0$, $\beta = \{1,00\},\{00,1\},\{1,01\},\{01,1\}$, and $\{1,1\}$, $J(f) = \sum_\beta \theta_\beta J_\beta(f)$ is equivalent to $\tilde{J}(f)$. V is thus completely continuous with respect to J, and the eigenvalues ρ_ν of J with respect to V satisfy $\beta_1\nu^{4-\epsilon} < \rho_\nu < \beta_2\nu^4$ for some $0 < \beta_1 < \beta_2 < \infty$ and ν sufficiently large, $\forall\epsilon > 0$. If $\mathcal{H}_{1,1}$ is eliminated with $\theta_{1,1} = 0$, ϵ can be set to 0. □

Example 9.3 (Thin-plate splines) For the thin-plate splines of §4.3, $J_m^d(f)$ in (4.17) on page 134 is defined on the unbounded domain $(-\infty,\infty)^d$, on which the usual L_2 norm is not defined.

Consider a bounded domain Ω satisfying certain boundary conditions. Let $J(f)$ be the integral of (4.17) restricted to Ω and $V(f) = \int_\Omega f^2 dx$. It can be shown that V is completely continuous with respect to J and $\rho_\nu \asymp \nu^{2m/d}$; see Cox (1984) and Utreras (1988). This does not address the thin-plate splines directly, but appears to be as close as one can get. □

Example 9.4 (Spherical splines) For the spherical splines of §4.4, $V(f) = \int_{\mathcal{S}} f^2(x)dx$ is completely continuous with respect to $\tilde{J}_m(f)$ of (4.46) on page 146, and $\rho_\nu \asymp \nu^m$. □

9.2 Rates for Density Estimates

Denote by $e^{\eta_0}/\int_{\mathcal{X}} e^{\eta_0}$ the density to be estimated and by $e^{\hat{\eta}}/\int_{\mathcal{X}} e^{\hat{\eta}}$ the estimate through the minimization of (7.1). We shall establish the asymptotic convergence rates in terms of the symmetrized Kullback-Leibler distance

$$\mathrm{SKL}(\eta_0,\hat{\eta}) = \mu_{\eta_0}(\eta_0 - \hat{\eta}) + \mu_{\hat{\eta}}(\hat{\eta} - \eta_0),$$

where $\mu_\eta(f) = \int_{\mathcal{X}} f e^\eta / \int_{\mathcal{X}} e^\eta$, and in terms of $V(\hat{\eta} - \eta_0) = V_{\eta_0}(\hat{\eta} - \eta_0)$, where $V_\eta(f) = \mu_\eta(f^2) - \mu_\eta^2(f)$.

The rates are first established for the minimizer $\tilde{\eta}$ of the quadratic approximation of (7.1) at η_0, then extended to $\hat{\eta}$ by bounding the magnitude of $\hat{\eta} - \tilde{\eta}$. The rates are further extended to the minimizer $\hat{\eta}^*$ of (7.1) in $\mathcal{H}^*$ of (7.2), by bounding the magnitudes of $\hat{\eta} - \eta^*$ and $\eta^* - \hat{\eta}^*$, where η^* is the projection of $\hat{\eta}$ in $\mathcal{H}^*$. The geometry in the spaces and the Fourier series expansion provide convenient tools throughout the analysis.

When $\eta_0 \notin \mathcal{H}$, the estimates are seen to converge to a Kullback-Leibler projection of η_0 in $\mathcal{H}$ at the same rates. The theory can also be easily adapted for the analysis of conditional density estimates and of estimates based on samples that are subject to selection bias.

9.2.1 Linear Approximation

Take $V(f) = V_{\eta_0}(f)$. The following conditions are needed in our analysis.

Condition 9.2.1 V is completely continuous with respect to J.

Condition 9.2.2 For ν sufficiently large and some $\beta > 0$, the eigenvalues ρ_ν of J with respect to V satisfy $\rho_\nu > \beta\nu^r$, where $r > 1$.

Consider the quadratic approximation of (7.1) at η_0, which is given by

$$-\frac{1}{n}\sum_{i=1}^{n}\eta(X_i) + \mu_{\eta_0}(\eta) + \frac{1}{2}V(\eta - \eta_0) + \frac{\lambda}{2}J(\eta); \tag{9.1}$$

see (7.16) on page 244. Plugging the Fourier series expansions $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ into (9.1), one has

$$\sum_\nu \left\{ -\eta_\nu \left(\frac{1}{n}\sum_{i=1}^{n}\phi_\nu(X_i) - \mu_{\eta_0}(\phi_\nu) \right) + \frac{1}{2}(\eta_\nu - \eta_{\nu,0})^2 + \frac{\lambda}{2}\rho_\nu\eta_\nu^2 \right\}. \tag{9.2}$$

Write $\beta_\nu = n^{-1}\sum_{i=1}^{n}\phi_\nu(X_i) - \mu_{\eta_0}(\phi_\nu)$. The Fourier coefficients that minimize (9.2) are given by

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu).$$

The minimizer $\tilde{\eta} = \sum_\nu \tilde{\eta}_\nu \phi_\nu$ of (9.1) is called a linear approximation of $\hat{\eta}$ since it is linear in $\phi_\nu(X_i)$. Straightforward calculation yields

$$V(\tilde{\eta} - \eta_0) = \sum_\nu (\tilde{\eta}_\nu - \eta_{\nu,0})^2 = \sum_\nu \frac{\beta_\nu^2 - 2\beta_\nu\lambda\rho_\nu\eta_{\nu,0} + \lambda^2\rho_\nu^2\eta_{\nu,0}^2}{(1 + \lambda\rho_\nu)^2},$$

$$\lambda J(\tilde{\eta} - \eta_0) = \sum_\nu \lambda\rho_\nu(\tilde{\eta}_\nu - \eta_{\nu,0})^2 = \sum_\nu \lambda\rho_\nu \frac{\beta_\nu^2 - 2\beta_\nu\lambda\rho_\nu\eta_{\nu,0} + \lambda^2\rho_\nu^2\eta_{\nu,0}^2}{(1 + \lambda\rho_\nu)^2}.$$

Note that $E[\beta_\nu] = 0$ and $E\big[\beta_\nu^2\big] = n^{-1}$. It follows that

$$\begin{aligned} E\big[V(\tilde{\eta} - \eta_0)\big] &= \frac{1}{n}\sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2} + \lambda \sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2}\rho_\nu \eta_{\nu,0}^2, \\ E\big[\lambda J(\tilde{\eta} - \eta_0)\big] &= \frac{1}{n}\sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} + \lambda \sum_\nu \frac{(\lambda\rho_\nu)^2}{(1+\lambda\rho_\nu)^2}\rho_\nu \eta_{\nu,0}^2. \end{aligned} \tag{9.3}$$

These quantities can be bounded with the help of the following lemma.

Lemma 9.1 *Under Condition 9.2.2, as $\lambda \to 0$, one has*

$$\begin{aligned} \sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} &= O\big(\lambda^{-1/r}\big), \\ \sum_\nu \frac{1}{(1+\lambda\rho_\nu)^2} &= O\big(\lambda^{-1/r}\big), \\ \sum_\nu \frac{1}{1+\lambda\rho_\nu} &= O\big(\lambda^{-1/r}\big). \end{aligned}$$

Proof: We prove the first equation.

$$\begin{aligned} \sum_\nu \frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} &= \Bigg(\sum_{\nu<\lambda^{-1/r}} + \sum_{\nu\ge\lambda^{-1/r}}\Bigg)\frac{\lambda\rho_\nu}{(1+\lambda\rho_\nu)^2} \\ &= O\big(\lambda^{-1/r}\big) + O\bigg(\int_{\lambda^{-1/r}}^\infty \frac{\lambda x^r}{(1+\lambda x^r)^2}dx\bigg) \\ &= O\big(\lambda^{-1/r}\big) + \lambda^{-1/r} O\bigg(\int_1^\infty \frac{x^r}{(1+x^r)^2}dx\bigg) \\ &= O\big(\lambda^{-1/r}\big). \end{aligned}$$

The other two follow similar arguments. □

Theorem 9.2 *Assume $J(\eta_0) < \infty$. Under Conditions 9.2.1 and 9.2.2, as $n \to \infty$ and $\lambda \to 0$,*

$$(V + \lambda J)(\tilde{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda\big).$$

Proof: Note that $\sum_\nu \rho_\nu \eta_{\nu,0}^2 = J(\eta_0) < \infty$. The theorem follows from (9.3) and Lemma 9.1. □

When η_0 is "supersmooth," in the sense that $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ for some $p > 1$, the rates can be improved to $O\big(n^{-1}\lambda^{-1/r} + \lambda^p\big)$, for p up to 2; see Problem 9.1.

9.2.2 Approximation Error and Main Results

We now turn to the approximation error $\hat{\eta} - \tilde{\eta}$. Define

$$A_{f,g}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}(f+\alpha g)(X_i) + \log\int_{\mathcal{X}} e^{f+\alpha g} + \frac{\lambda}{2}J(f+\alpha g),$$

$$B_{f,g}(\alpha) = -\frac{1}{n}\sum_{i=1}^{n}(f+\alpha g)(X_i) + \mu_{\eta_0}(f+\alpha g) + \frac{1}{2}V(f+\alpha g-\eta_0) + \frac{\lambda}{2}J(f+\alpha g).$$

It is easy to verify that (Problem 9.2)

$$\dot{A}_{f,g}(0) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \mu_f(g) + \lambda J(f,g), \tag{9.4}$$

$$\dot{B}_{f,g}(0) = -\frac{1}{n}\sum_{i=1}^{n} g(X_i) + \mu_{\eta_0}(g) + V(f-\eta_0, g) + \lambda J(f,g). \tag{9.5}$$

Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (9.4), one has

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\tilde{\eta})(X_i) + \mu_{\hat{\eta}}(\hat{\eta}-\tilde{\eta}) + \lambda J(\hat{\eta}, \hat{\eta}-\tilde{\eta}) = 0, \tag{9.6}$$

and setting $f = \tilde{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (9.5) yields

$$-\frac{1}{n}\sum_{i=1}^{n}(\hat{\eta}-\tilde{\eta})(X_i) + \mu_{\eta_0}(\hat{\eta}-\tilde{\eta}) + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \lambda J(\tilde{\eta}, \hat{\eta}-\tilde{\eta}) = 0. \tag{9.7}$$

Combining (9.6) and (9.7), it follows that

$$\mu_{\hat{\eta}}(\hat{\eta}-\tilde{\eta}) - \mu_{\tilde{\eta}}(\hat{\eta}-\tilde{\eta}) + \lambda J(\hat{\eta}-\tilde{\eta}) = V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \mu_{\eta_0}(\hat{\eta}-\tilde{\eta}) - \mu_{\tilde{\eta}}(\hat{\eta}-\tilde{\eta}). \tag{9.8}$$

Now, define

$$C(\alpha) = \mu_{\eta_0+\alpha(\tilde{\eta}-\eta_0)/\sigma}(\hat{\eta}-\tilde{\eta}) - \mu_{\eta_0}(\hat{\eta}-\tilde{\eta}),$$

where $\sigma = \{V(\tilde{\eta}-\eta_0)\}^{1/2} = o_p(1)$. A Taylor expansion gives $C(\alpha) = \alpha\big(1+o(1)\big)V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta})/\sigma$, where $o(1)$ is with respect to $\alpha \to 0$. This leads to

$$\mu_{\tilde{\eta}}(\hat{\eta}-\tilde{\eta}) - \mu_{\eta_0}(\hat{\eta}-\tilde{\eta}) = C(\sigma) = V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta})\big(1+o_p(1)\big), \tag{9.9}$$

as $\lambda \to 0$ and $n\lambda^{1/r} \to \infty$. Now, define $D(\alpha) = \mu_{\tilde{\eta}+\alpha(\hat{\eta}-\tilde{\eta})}(\hat{\eta}-\tilde{\eta})$. It can be shown that $\dot{D}(\alpha) = V_{\tilde{\eta}+\alpha(\hat{\eta}-\tilde{\eta})}(\hat{\eta}-\tilde{\eta})$. By the mean value theorem,

$$\mu_{\hat{\eta}}(\hat{\eta}-\tilde{\eta}) - \mu_{\tilde{\eta}}(\hat{\eta}-\tilde{\eta}) = D(1) - D(0) = \dot{D}(\alpha) = V_{\tilde{\eta}+\alpha(\hat{\eta}-\tilde{\eta})}(\hat{\eta}-\tilde{\eta}), \tag{9.10}$$

for some $\alpha \in [0,1]$. The following condition is needed to proceed.

Condition 9.2.3 For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$, $c_1 V(f) \le V_\eta(f)$ holds uniformly for some $c_1 > 0$.

Condition 9.2.3 is satisfied when the members of B_0 have uniform upper and lower bounds on domain $\mathcal{X}$.

Theorem 9.3 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.2.1–9.2.3, as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \tilde{\eta}) = o_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Consequently,

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: From (9.8)–(9.10), and Condition 9.2.3,

$$\begin{aligned} c_1 V(\hat{\eta} - \tilde{\eta}) + \lambda J(\hat{\eta} - \tilde{\eta}) &\le o_p\big(V(\tilde{\eta} - \eta_0, \hat{\eta} - \tilde{\eta})\big) \\ &= o_p\big(\{V(\hat{\eta} - \tilde{\eta}) V(\tilde{\eta} - \eta_0)\}^{1/2}\big). \end{aligned}$$

The theorem follows from Theorem 9.2 after trivial manipulation. □

Theorem 9.4 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.2.1–9.2.3, as* $\lambda \to 0$ *and* $n\lambda^{1/r} \to \infty$,

$$\mathrm{SKL}(\eta_0, \hat{\eta}) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \eta_0$ in (9.4), one has

$$\begin{aligned} &\mu_{\eta_0}(\eta_0 - \hat{\eta}) + \mu_{\hat{\eta}}(\hat{\eta} - \eta_0) \\ &\qquad = \left\{ \frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta_0)(X_i) - \mu_{\eta_0}(\hat{\eta} - \eta_0) \right\} - \lambda J(\hat{\eta}, \hat{\eta} - \eta_0) \end{aligned} \tag{9.11}$$

For the first term on the right-hand side of (9.11), write

$$\frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta_0)(X_i) - \mu_{\eta_0}(\hat{\eta} - \eta_0) = \sum_\nu (\hat{\eta}_\nu - \eta_{\nu,0}) \beta_\nu,$$

where $\hat{\eta}_\nu$ are the Fourier coefficients of $\hat{\eta}$ and $\beta_\nu = n^{-1} \sum_{i=1}^n \phi_\nu(X_i) - \mu_{\eta_0}(\phi_\nu)$. By the Cauchy-Schwartz inequality,

$$\sum_\nu \big|(\hat{\eta}_\nu - \eta_{\nu,0})\beta_\nu\big| \le \big\{ \textstyle\sum_\nu \alpha_\nu^2 (\hat{\eta}_\nu - \eta_{\nu,0})^2 \big\}^{1/2} \big\{ \textstyle\sum_\nu \alpha_\nu^{-2} \beta_\nu^2 \big\}^{1/2},$$

for some sequence α_ν. Setting $\alpha_\nu^2 = 1 + \lambda\rho_\nu$, one has

$$\left| \frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta_0)(X_i) - \mu_{\eta_0}(\hat{\eta} - \eta_0) \right| \le \big\{ (V + \lambda J)(\hat{\eta} - \eta_0) \big\}^{1/2} O_p\big(n^{-1/2} \lambda^{-1/2r}\big), \tag{9.12}$$

where

$$\sum_\nu (1+\lambda\rho_\nu)(\hat\eta_\nu-\eta_{\nu,0})^2 = (V+\lambda J)(\hat\eta-\eta_0) = O_p\big(n^{-1}\lambda^{-1/r}+\lambda^p\big)$$

by Theorem 9.3, and $E\big[\sum_\nu (1+\lambda\rho_\nu)^{-1}\beta_\nu^2\big] = O\big(n^{-1}\lambda^{-1/r}\big)$ by Lemma 9.1 and the fact that $E\big[\beta_\nu^2\big] = n^{-1}$. Hence,

$$\left|\frac{1}{n}\sum_{i=1}^n (\hat\eta-\eta_0)(X_i) - \mu_{\eta_0}(\hat\eta-\eta_0)\right| = O_p\big(n^{-1}\lambda^{-1/r}+n^{-1/2}\lambda^{-1/2r+p/2}\big). \tag{9.13}$$

Similarly, $\lambda J(\hat\eta,\hat\eta-\eta_0) = \lambda J(\hat\eta-\eta_0)+\lambda J(\eta_0,\hat\eta-\eta_0)$, where

$$\begin{aligned}
\lambda J(\eta_0,\hat\eta-\eta_0) &= \sum_\nu \lambda\rho_\nu\eta_{\nu,0}(\hat\eta_\nu-\eta_{\nu,0}) \\
&\le \left\{\sum_\nu (1+\lambda\rho_\nu)(\hat\eta_\nu-\eta_{\nu,0})^2\right\}^{1/2} \\
&\quad\times\left\{\lambda^p\sum_\nu \frac{(\lambda\rho_\nu)^{2-p}}{1+\lambda\rho_\nu}\rho_\nu^p\eta_{\nu,0}^2\right\}^{1/2} \\
&= \big\{(V+\lambda J)(\hat\eta-\eta_0)\big\}^{1/2} O\big(\lambda^{p/2}\big).
\end{aligned}$$

By Theorem 9.3, $\big|\lambda J(\hat\eta,\hat\eta-\eta_0)\big| = O_p\big(n^{-1}\lambda^{-1/r}+\lambda^p\big)$. Combining this with (9.13), the theorem follows. □

9.2.3 Efficient Approximation

As was noted in §7.1, the minimizer $\hat\eta$ of (7.1) in $\mathcal{H}$ is, in general, not computable. The minimizer $\hat\eta^*$ in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\big\{R_J(Z_j,\cdot), j=1,\ldots,q\big\}$$

was computed instead, where $\{Z_j\}$ is a random subset of $\{X_i\}$ and hence also an *i.i.d.* sample from $e^{\eta_0(x)}/\int_{\mathcal{X}} e^{\eta_0(x)}$. We shall now establish the same convergence rates for $\hat\eta^*$ under an extra condition.

Condition 9.2.4 $V(\phi_\nu\phi_\mu) \le c_2$ holds uniformly for some $c_2 > 0$, $\forall\nu,\mu$.

Condition 9.2.4 virtually calls for uniformly bounded fourth moments of $\phi_\nu(X)$. The condition appears mild, as ϕ_ν typically grow in roughness but not necessarily in magnitude, but since ϕ_ν are generally not available in explicit forms, the condition is extremely difficult to verify from more primitive conditions, if at all possible.

Lemma 9.5 *Under Conditions 9.2.1, 9.2.2, and 9.2.4, as* $\lambda\to 0$ *and* $q\lambda^{2/r}\to\infty$, $V(h) = o_p\big(\lambda J(h)\big)$, $\forall h\in\mathcal{H}\ominus\mathcal{H}^*$.

Note that $q \le n$, so when $\lambda \to 0$ and $q\lambda^{2/r} \to \infty$, $n\lambda^{1/r} \to \infty$. The computational cost of $\hat{\eta}^*$ is of the order $O(nq^2)$, thus a smaller q is preferred. The optimal convergence rate $O_p\big(n^{-pr/(pr+1)}\big)$ is achieved at $\lambda \asymp n^{-r/(pr+1)}$, hence it is sufficient to have $q \asymp n^{2/(pr+1)+\epsilon}$, $\forall \epsilon > 0$.

Proof of Lemma 9.5: For $h \in \mathcal{H} \ominus \mathcal{H}^*$, since $h(Z_j) = J\big(R_J(Z_j, \cdot), h\big) = 0$, $\sum_{j=1}^q h^2(Z_j) = 0$. Write $h = \sum_\nu h_\nu \phi_\nu$. It follows that

$$\begin{aligned} V(h) \le \mu_{\eta_0}(h^2) &= \sum_\nu \sum_\mu h_\nu h_\mu \mu_{\eta_0}(\phi_\nu \phi_\mu) \\ &= \sum_\nu \sum_\mu h_\nu h_\mu \left\{ \mu_{\eta_0}(\phi_\nu \phi_\mu) - \frac{1}{q} \sum_{j=1}^q \phi_\nu(Z_j)\phi_\mu(Z_j) \right\} \\ &\le \left\{ \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \right. \\ &\qquad \left. \times \left\{ \frac{1}{q} \sum_{j=1}^q \phi_\nu(Z_j)\phi_\mu(Z_j) - \mu_{\eta_0}(\phi_\nu\phi_\mu) \right\}^2 \right\}^{1/2} \\ &\qquad \times \left\{ \sum_\nu \sum_\mu (1+\lambda\rho_\nu)(1+\lambda\rho_\mu) h_\nu^2 h_\mu^2 \right\}^{1/2} \\ &= O_p\big(q^{-1/2}\lambda^{-1/r}\big)(V + \lambda J)(h), \end{aligned}$$

where Lemma 9.1 and the fact that

$$E\left[\frac{1}{q} \sum_{j=1}^q \phi_\nu(Z_j)\phi_\mu(Z_j) - \mu_{\eta_0}(\phi_\nu\phi_\mu) \right]^2 \le \frac{c_2}{q}$$

are used. The lemma follows. □

Let η^* be the projection of $\hat{\eta}$ in $\mathcal{H}^*$. Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \eta^*$ in (9.4), one has

$$-\frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta^*)(X_i) + \mu_{\hat{\eta}}(\hat{\eta} - \eta^*) + \lambda J(\hat{\eta}, \hat{\eta} - \eta^*) = 0. \tag{9.14}$$

Adding and subtracting $\mu_{\eta_0}(\hat{\eta} - \eta^*)$, and noting that $J(\eta^*, \hat{\eta} - \eta^*) = 0$,

$$\left\{ \frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta^*)(X_i) - \mu_{\eta_0}(\hat{\eta} - \eta^*) \right\} - \big(\mu_{\hat{\eta}}(\hat{\eta} - \eta^*) - \mu_{\eta_0}(\hat{\eta} - \eta^*)\big) = \lambda J(\hat{\eta} - \eta^*). \tag{9.15}$$

Similar to (9.12), one has

$$\left| \frac{1}{n} \sum_{i=1}^n (\hat{\eta} - \eta^*)(X_i) - \mu_{\eta_0}(\hat{\eta} - \eta^*) \right| = O_p\big(n^{-1/2}\lambda^{-1/2r}\big) \big\{ (V + \lambda J)(\hat{\eta} - \eta^*) \big\}^{1/2}, \tag{9.16}$$

and similar to (9.9), it can be shown that

$$\mu_{\hat\eta}(\hat\eta - \eta^*) - \mu_{\eta_0}(\hat\eta - \eta^*) = V(\hat\eta - \eta_0, \hat\eta - \eta^*)\big(1 + o_p(1)\big). \tag{9.17}$$

Theorem 9.6 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.2.1–9.2.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$\lambda J(\hat\eta - \eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big),$$
$$V(\hat\eta - \eta^*) = o_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Combining (9.15)–(9.17) and applying Theorem 9.3,

$$\lambda J(\hat\eta - \eta^*) = O_p\big(n^{-1/2}\lambda^{-1/2r} + \lambda^{p/2}\big)\big\{(V + \lambda J)(\hat\eta - \eta^*)\big\}^{1/2}.$$

The theorem follows from Lemma 9.5. □

We can now obtain the rates for $(V + \lambda J)(\hat\eta^* - \eta^*)$ and, in turn, for $(V + \lambda J)(\hat\eta^* - \hat\eta)$. Condition 9.2.3 needs to be modified to include $\hat\eta^*$ and η^* in the convex set B_0.

Theorem 9.7 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.2.1–9.2.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat\eta^* - \eta^*) = o_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big),$$
$$(V + \lambda J)(\hat\eta - \hat\eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Setting $f = \hat\eta^*$ and $g = \hat\eta^* - \eta^* \in \mathcal{H}^*$ in (9.4), one has

$$-\frac{1}{n}\sum_{i=1}^n (\hat\eta^* - \eta^*)(X_i) + \mu_{\hat\eta^*}(\hat\eta^* - \eta^*) + \lambda J(\hat\eta^*, \hat\eta^* - \eta^*) = 0. \tag{9.18}$$

Setting $f = \hat\eta$ and $g = \hat\eta - \hat\eta^*$ in (9.4), one gets

$$-\frac{1}{n}\sum_{i=1}^n (\hat\eta - \hat\eta^*)(X_i) + \mu_{\hat\eta}(\hat\eta - \hat\eta^*) + \lambda J(\hat\eta, \hat\eta - \hat\eta^*) = 0. \tag{9.19}$$

Adding (9.18), (9.19) and subtracting (9.14), some algebra yields

$$\mu_{\hat\eta^*}(\hat\eta^* - \eta^*) - \mu_{\eta^*}(\hat\eta^* - \eta^*) + \lambda J(\hat\eta^* - \eta^*) = \mu_{\hat\eta}(\hat\eta^* - \eta^*) - \mu_{\eta^*}(\hat\eta^* - \eta^*);$$

remember that $J(\hat\eta - \eta^*, \eta^*) = J(\hat\eta - \eta^*, \hat\eta^*) = 0$. In view of (9.9), (9.10), and Condition 9.2.3,

$$c_1 V(\hat\eta^* - \eta^*) + \lambda J(\hat\eta^* - \eta^*) \le \big|V(\hat\eta - \eta^*, \hat\eta^* - \eta^*)\big|\big(1 + o_p(1)\big).$$

The theorem follows after applying the Cauchy-Schwartz inequality and Theorem 9.6. □

Theorem 9.8 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.2.1–9.2.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat\eta^* - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big),$$
$$\mathrm{SKL}(\eta_0, \hat\eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: The first part of the theorem follows from Theorems 9.3, 9.6, and 9.7. For the second part, set $f = \hat\eta$ and $h = \hat\eta^* - \eta_0$ in (9.4). This yields

$$-\frac{1}{n}\sum_{i=1}^n (\hat\eta^* - \eta_0)(X_i) + \mu_{\hat\eta}(\hat\eta^* - \eta_0) + \lambda J(\hat\eta, \hat\eta^* - \eta_0) = 0.$$

Hence,

$$\begin{aligned}
&\mu_{\eta_0}(\eta_0 - \hat\eta^*) + \mu_{\hat\eta^*}(\hat\eta^* - \eta_0)\\
&\quad = \mu_{\eta_0}(\eta_0 - \hat\eta^*) + \mu_{\hat\eta^*}(\hat\eta^* - \eta_0)\\
&\qquad + \frac{1}{n}\sum_{i=1}^n (\hat\eta^* - \eta_0)(X_i) - \mu_{\hat\eta}(\hat\eta^* - \eta_0) + \lambda J(\hat\eta, \eta_0 - \hat\eta^*)\\
&\quad = \lambda J(\hat\eta, \eta_0 - \hat\eta^*) + \left\{\frac{1}{n}\sum_{i=1}^n (\hat\eta^* - \eta_0)(X_i) - \mu_{\eta_0}(\hat\eta^* - \eta_0)\right\}\\
&\qquad + \big\{\mu_{\hat\eta^*}(\hat\eta^* - \eta_0) - \mu_{\hat\eta}(\hat\eta^* - \eta_0)\big\}.
\end{aligned}$$

The first term on the right-hand side is of the order $O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big)$ by arguments similar to ones used in the proof of Theorem 9.4. The second and the third terms are of the same order in view of (9.16) and (9.17), Theorem 9.7, and the first part of this theorem. □

9.2.4 Convergence Under Incorrect Model

It has been implicitly assumed thus far that $\eta_0 \in \mathcal{H}$. In the case $\eta_0 \notin \mathcal{H}$, say an additive model is fitted while the interaction is present in η_0, modifications are needed in the problem formulation. The convergence rates remain valid under the modified formulation, however.

Suppose the minimizer of $\mathrm{RKL}(\eta_0, \eta) = \log \int_{\mathcal{X}} e^\eta - \mu_{\eta_0}(\eta)$ exists in $\mathcal{H}$, then it is the Kullback-Leibler projection of η_0 in $\mathcal{H}$, to be denoted by η_0^*, which is probably the best proxy of η_0 one can hope to estimate in the context. It is known that $\mu_{\eta_0^*}(h) = \mu_{\eta_0}(h)$, $\forall h \in \mathcal{H}$. Substituting η_0^* for η_0 everywhere in §§9.2.1–9.2.3, all results and arguments remain valid if

$$E\left[\frac{1}{n}\sum_{i=1}^n \phi_\nu(X_i) - \mu_{\eta_0^*}(\phi_\nu)\right]^2 = \frac{1}{n},$$
$$E\left[\frac{1}{n}\sum_{i=1}^n \phi_\nu(X_i)\phi_\mu(X_i) - \mu_{\eta_0^*}(\phi_\nu\phi_\mu)\right]^2 \le \frac{c_2}{n}.$$

These equations hold under an extra condition.

Condition 9.2.0 $gh - C_{gh} \in \mathcal{H}$ for some constant C_{gh}, $\forall g, h \in \mathcal{H}$.

Note that if $\mu_{\eta_0}(gh - C_{gh}) = \mu_{\eta_0^*}(gh - C_{gh})$, then $\mu_{\eta_0}(gh) = \mu_{\eta_0^*}(gh)$. The key requirement here is that $J(gh) < \infty$ whenever $J(g) < \infty$, $J(h) < \infty$; the constant C_{gh} takes care of the side condition on log density. Condition 9.2.0 is satisfied by all the spaces appearing in the examples in Chap. 7.

9.2.5 Estimation Under Biased Sampling

Now, consider the setting of §7.6. Observations (t_i, X_i) are taken from $\mathcal{T} \times \mathcal{X}$ with $X|t \sim w(t,x)e^{\eta_0(x)}/\int_{\mathcal{X}} w(t,x)e^{\eta_0(x)}$, and the density estimate $e^{\hat{\eta}}/\int_{\mathcal{X}} e^{\hat{\eta}}$ is obtained via the minimization of (7.26). The theory developed in the proceeding sections remain valid with due modifications, although some of the intermediate o_p rates might have to be replaced by the respective O_p rates.

Let $m(t)$ be the limiting density of t_i on $\mathcal{T}$. Write

$$\mu_\eta(f|t) = \frac{\int_{\mathcal{X}} f(x)w(t,x)e^{\eta(x)}}{\int_{\mathcal{X}} w(t,x)e^{\eta(x)}}, \quad v_\eta(f|t) = \mu_\eta(f^2|t) - \mu_\eta^2(f|t),$$

and define

$$\mu_\eta(f) = \int_{\mathcal{T}} m(t)\mu_\eta(f|t), \quad V_\eta(f) = \int_{\mathcal{T}} m(t)v_\eta(f|t).$$

The convergence rates are given in terms of

$$\mathrm{SKL}(\eta_0, \hat{\eta}) = \int_{\mathcal{T}} m(t)\big\{\mu_{\eta_0}(\eta_0 - \hat{\eta}|t) + \mu_{\hat{\eta}}(\hat{\eta} - \eta_0|t)\big\}$$

and $V(\hat{\eta} - \eta_0)$, where $V(f) = V_{\eta_0}(f)$.

For the theory of §§9.2.1–9.2.3 to hold in this setting, Conditions 9.2.1 and 9.2.2 need little change except for the definition of V. Conditions 9.2.3 and 9.2.4 shall be modified as follows.

Condition 9.2.3b For η in a convex set B_0 around η_0 containing $\tilde{\eta}$, $\hat{\eta}$, η^*, and $\hat{\eta}^*$, $c_1 v_{\eta_0}(f|t) \le v_\eta(f|t) \le c_2 v_{\eta_0}(f|t)$ holds uniformly for some $0 < c_1 < c_2 < \infty$, $\forall f \in \mathcal{H}$, $\forall t \in \mathcal{T}$.

Condition 9.2.4b $\int_{\mathcal{T}} m(t)\big\{v_{\eta_0}(\phi_\nu, \phi_\mu|t)\big\}^2 \le c_3$ holds uniformly for some $c_3 < \infty$, $\forall \nu, \mu$.

To apply the arguments of §9.2.4, the relative Kullback-Leibler distance shall be modified as $\mathrm{RKL}(\eta_0, \eta) = \int_{\mathcal{T}} m(t)\big\{\log \int_{\mathcal{X}} w(t,x)e^{\eta(x)} - \mu_{\eta_0}(\eta|t)\big\}$. Details are straightforward to work out and are left as an exercise (Problem 9.3).

9.2.6 Estimation of Conditional Density

For the estimation of the conditional density $f(y|x) = e^{\eta_0(x,y)}/\int_{\mathcal{Y}} e^{\eta_0(x,y)}$ via the minimization of (7.30), the theory is also easy to modify.

Let $f(x)$ be the marginal density of X on $\mathcal{X}$. Write

$$\mu_\eta(g|x) = \frac{\int_{\mathcal{Y}} g(x,y)e^{\eta(x,y)}}{\int_{\mathcal{Y}} e^{\eta(x,y)}}, \quad v_\eta(g|x) = \mu_\eta(g^2|x) - \mu_\eta^2(g|x)$$

and define

$$\mu_\eta(g) = \int_{\mathcal{X}} f(x)\mu_\eta(g|x), \quad V_\eta(g) = \int_{\mathcal{X}} f(x)v_\eta(g|x).$$

The convergence rates are given in terms of

$$\mathrm{SKL}(\eta_0, \hat{\eta}) = \int_{\mathcal{X}} f(x)\big\{\mu_{\eta_0}(\eta_0 - \hat{\eta}|x) + \mu_{\hat{\eta}}(\hat{\eta} - \eta_0|x)\big\}$$

and $V(\hat{\eta} - \eta_0)$, where $V(g) = V_{\eta_0}(g)$.

For the theory of §§9.2.1–9.2.3 to hold for conditional density estimates, Conditions 9.2.1 and 9.2.2 need little change except for the definition of V. Conditions 9.2.3 and 9.2.4 shall be modified as follows.

Condition 9.2.3c For η in a convex set B_0 around η_0 containing $\tilde{\eta}$, $\hat{\eta}$, η^*, and $\hat{\eta}^*$, $c_1 v_{\eta_0}(g|x) \le v_\eta(g|x) \le c_2 v_{\eta_0}(g|x)$ holds uniformly for some $0 < c_1 < c_2 < \infty$, $\forall g \in \mathcal{H}$, $\forall x \in \mathcal{X}$.

Condition 9.2.4c There exist $c_3, c_4, c_5 < \infty$, such that

$$\begin{aligned}
&\textstyle\int_{\mathcal{X}} f(x)\big\{v_{\eta_0}(\phi_\nu, \phi_\mu|x)\big\}^2 \le c_3,\\
&\textstyle\int_{\mathcal{X}} f(x)v_{\eta_0}(\phi_\nu\phi_\mu, \phi_\nu\phi_\mu|x) \le c_4,\\
&\textstyle\int_{\mathcal{X}} f(x)\big\{\mu_{\eta_0}(\phi_\nu\phi_\mu|x) - \mu_{\eta_0}(\phi_\nu\phi_\mu)\big\}^2 \le c_5,
\end{aligned}$$

hold uniformly, $\forall \nu, \mu$,

To apply the arguments of §9.2.4, the relative Kullback-Leibler distance shall be modified as $\mathrm{RKL}(\eta_0, \eta) = \int_{\mathcal{X}} f(x)\big\{\log \int_{\mathcal{Y}} e^{\eta} - \mu_{\eta_0}(\eta|x)\big\}$, and the constant C_{gh} in Condition 9.2.0 may be a function of x. Details are left as an exercise (Problem 9.4).

9.2.7 Estimation Under Response-Based Sampling

Consider the connected case in the setting of §7.9, where the strata $\mathcal{Y}_j$ are sampled with probability π_j, and the samples $(X, Y)|\mathcal{Y}_j$ are taken from $e^{\eta_0(x,y)}/\int_{\mathcal{X}\times\mathcal{Y}_j} e^{\eta_0(x,y)}$. Write

$$\mu_\eta(f|j) = \frac{\int_{\mathcal{X}\times\mathcal{Y}_j} fe^{\eta}}{\int_{\mathcal{X}\times\mathcal{Y}_j} e^{\eta}}, \quad v_\eta(f|j) = \mu_\eta(f^2|j) - \mu_\eta^2(f|j)$$

and define

$$\mu_\eta(f) = \sum_{j=1}^{s} \pi_j \mu_\eta(f|j), \quad V_\eta(f) = \sum_{j=1}^{s} \pi_j v_\eta(f|j).$$

The rates for the minimizers of (7.46) can be derived in terms of

$$\mathrm{SKL}(\eta_0, \hat{\eta}) = \sum_{j=1}^{s} \pi_j \big\{ \mu_{\eta_0}(\eta_0 - \hat{\eta}|j) + \mu_{\hat{\eta}}(\hat{\eta} - \eta_0|j) \big\}$$

and $V(\hat{\eta} - \eta_0)$, where $V(f) = V_{\eta_0}(f)$. The conditions needed are similar to those for conditional density estimates. The relative Kullback-Leibler distance is defined by $\mathrm{RKL}(\eta_0, \eta) = \sum_{j=1}^{s} \pi_j \big\{ \log \int_{\mathcal{X}\times\mathcal{Y}_j} e^{\eta} - \mu_{\eta_0}(\eta|j) \big\}$. Further details are left as an exercise (Problem 9.5).

9.3 Rates for Hazard Estimates

The convergence rates for the minimizers of (8.1) are to be established in this section. The martingale structure of censored lifetime data, which was mentioned in §§8.2.1 and 8.6.1, serves as the primary tool for the stochastic calculations involved.

Some basic facts concerning the martingale structure are summarized, and a quadratic functional V is derived under the sampling structure. The rates are given in terms of $V(\hat{\eta} - \eta_0)$ and in terms of the symmetrized version of $\mathrm{KL}(\eta_0, \hat{\eta})$ as defined in (8.6). The analysis parallels that in §9.2.

9.3.1 Martingale Structure

Write $N(t) = I_{[X \le t, \delta = 1]}$, $Y(t) = I_{[Z < t \le X]}$, and $A(t) = \int_0^t e^{\eta_0(s,U)} Y(s) ds$, as in §8.2.1. Under independent censorship, $M(t) = N(t) - A(t)$ is a martingale conditional on U and Z. We shall now summarize some martingale properties needed in the asymptotic analysis. The results are quoted from Fleming and Harrington (1991, §2.7) and Gill (1984).

First of all, one has $E\big[M(t)\big|U, Z\big] = 0$ and

$$E\big[M^2(t)\big|U, Z\big] = E\big[A(t)\big|U, Z\big] = \int_0^t e^{\eta_0(s,U)} E\big[Y(s)\big|U, Z\big] ds.$$

For any deterministic function $h(t, u)$ continuous in t, $\forall u$ (so it is locally bounded predictable), the Stieltjes integral $\int_0^t h(s, U) dM(s)$ is a martingale as long as $\int_{\mathcal{T}} h^2(t, U) e^{\eta_0(t,U)} E\big[Y(t)\big|U, Z\big] dt < \infty$. It follows that

$$E\big[\textstyle\int_0^t h(s, U) dM(s)\big|U, Z\big] = 0,$$

$$E\big[\big\{\textstyle\int_0^t h(s, U) dM(s)\big\}^2\big|U, Z\big] = \int_0^t h^2(s, U) e^{\eta_0(s,U)} E\big[Y(s)\big|U, Z\big] ds.$$

This yields

$$E\left[\int_0^t h\,dN(s)\right] - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds = E\left[\int_0^t h\,dM(s)\right] = 0, \quad (9.20)$$

$$E\left[\left\{\int_0^t h\,dM(s)\right\}^2\right] = E\left[\int_0^t h^2 dA(s)\right] = \int_{\mathcal{U}} m(u)\int_0^t h^2 e^{\eta_0}\tilde{S}ds, \quad (9.21)$$

where $\tilde{S}(t,u) = E\big[Y(t)\big|U=u\big] = P\big(Z < t \le X \,\big|\, U=u\big)$. Furthermore,

$$\begin{aligned} &E\left[\left\{\int_0^t h\,dN(s) - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds\right\}^2\right] \\ &= E\left[\left\{\int_0^t h\,dM(s) + \int_0^t h\,e^{\eta_0}Y(s)ds - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds\right\}^2\right] \\ &= E\left[\left\{\int_0^t h\,dM(s)\right\}^2\right] \\ &\quad + E\left[\left\{\int_0^t h\,e^{\eta_0}Y(s)ds - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds\right\}^2\right], \end{aligned} \quad (9.22)$$

where $E\big[\int_0^t h\,dM(s)\big\{\int_0^t h\,e^{\eta_0}Y(s)ds - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds\big\}\big|U,Z\big] = 0$ because $\int_0^t h\,e^{\eta_0}Y(s)ds - \int_{\mathcal{U}} m(u)\int_0^t h\,e^{\eta_0}\tilde{S}ds$ is predictable.

Note that $\delta\,\eta(X,U) = \int_{\mathcal{T}} \eta(t,U)dN(t)$, $\int_Z^X e^{\eta(t,U)}dt = \int_{\mathcal{T}} e^{\eta(t,U)}Y(t)dt$. The penalized likelihood functional (8.1) on page 286 shall be written as

$$-\frac{1}{n}\sum_{i=1}^n\left\{\int_{\mathcal{T}}\eta_i dN_i(t) - \int_{\mathcal{T}} e^{\eta_i}Y_i dt\right\} + \frac{\lambda}{2}J(\eta), \quad (9.23)$$

where $\eta_i(t) = \eta(t,U_i)$. Define

$$V(f) = \int_{\mathcal{U}} m(u)\int_{\mathcal{T}} f^2(t,u)e^{\eta_0(t,u)}\tilde{S}(t,u)dt. \quad (9.24)$$

Convergence rates for the minimizer $\hat{\eta}$ of (9.23) shall be established in terms of $V(\hat{\eta}-\eta_0)$ and

$$\mathrm{SKL}(\eta_0,\hat{\eta}) = \int_{\mathcal{U}} m(u)\int_{\mathcal{T}}\big(e^{\hat{\eta}(t,u)} - e^{\eta_0(t,u)}\big)\big(\hat{\eta}(t,u)-\eta_0(t,u)\big)\tilde{S}(t,u)dt,$$

which is the symmetrized version of $\mathrm{KL}(\eta_0,\hat{\eta})$ defined in (8.6) on page 289.

9.3.2 Linear Approximation

The following conditions are needed in our analysis, which are carbon copies of Conditions 9.2.1 and 9.2.2 but with V as defined in (9.24).

Condition 9.3.1 V is completely continuous with respect to J.

Condition 9.3.2 For ν sufficiently large and some $\beta > 0$, the eigenvalues ρ_ν of J with respect to V satisfy $\rho_\nu > \beta\nu^r$, where $r > 1$.

Consider the quadratic functional

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}\eta_i dN_i(t) - \int_{\mathcal{T}}\eta_i e^{\eta_{0,i}}Y_i dt\right\} + \frac{1}{2}V(\eta-\eta_0) + \frac{\lambda}{2}J(\eta), \quad (9.25)$$

where $\eta_{0,i}(t) = \eta_0(t, U_i)$. Plugging the Fourier expansions $\eta = \sum_\nu \eta_\nu\phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ into (9.25), the minimizer $\tilde{\eta}$ of (9.25) has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1+\lambda\rho_\nu),$$

where $\beta_\nu = n^{-1}\sum_{i=1}^{n}\int_{\mathcal{T}}\phi_{\nu,i}dM_i(t)$ with $\phi_{\nu,i}(t) = \phi_\nu(t, U_i)$. From (9.20), (9.21), and the fact that $\int_{\mathcal{U}} m(u)\int_{\mathcal{T}}\phi_\nu^2 e^{\eta_0}\tilde{S}dt = V(\phi_\nu) = 1$, it is easy to see that $E[\beta_\nu] = 0$ and $E\big[\beta_\nu^2\big] = n^{-1}$. See Problem 9.6.

Theorem 9.9 *Assume* $\sum_\nu \rho_\nu^p\eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.3.1 and 9.3.2, as* $n\to\infty$ *and* $\lambda\to 0$,

$$(V+\lambda J)(\tilde{\eta}-\eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: See the proof of Theorem 9.2. □

9.3.3 Approximation Error and Main Results

We now turn to the approximation error $\hat{\eta} - \tilde{\eta}$. Define

$$\begin{aligned}
A_{f,g}(\alpha) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(f+\alpha g)_i dN_i(t) - \int_{\mathcal{T}} e^{(f+\alpha g)_i}Y_i dt\right\} \\
&\quad + \frac{\lambda}{2}J(f+\alpha g), \\
B_{f,g}(\alpha) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(f+\alpha g)_i dN_i(t) - \int_{\mathcal{T}}(f+\alpha g)_i e^{\eta_{0,i}}Y_i dt\right\} \\
&\quad + \frac{1}{2}V(f+\alpha g-\eta_0) + \frac{\lambda}{2}J(f+\alpha g).
\end{aligned}$$

It can be shown that

$$\dot{A}_{f,g}(0) = -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}} g_i dN_i(t) - \int_{\mathcal{T}} g_i e^{f_i}Y_i dt\right\} + \lambda J(f,g), \quad (9.26)$$

$$\begin{aligned}
\dot{B}_{f,g}(0) &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}} g_i dN_i(t) - \int_{\mathcal{T}} g_i e^{\eta_{0,i}}Y_i dt\right\} \\
&\quad + V(f-\eta_0, g) + \lambda J(f,g). \quad (9.27)
\end{aligned}$$

Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (9.26), one has

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i dN_i(t) - \int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i e^{\hat{\eta}_i} Y_i dt\right\} + \lambda J(\hat{\eta}, \hat{\eta}-\tilde{\eta}) = 0, \quad (9.28)$$

and setting $f = \tilde{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (9.27), one gets

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i dN_i(t) - \int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i e^{\eta_{0,i}} Y_i dt\right\} + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \lambda J(\tilde{\eta}, \hat{\eta}-\tilde{\eta}) = 0. \quad (9.29)$$

Subtracting (9.29) from (9.28), some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i(e^{\hat{\eta}} - e^{\tilde{\eta}})_i Y_i dt + \lambda J(\hat{\eta}-\tilde{\eta}) = V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i(e^{\tilde{\eta}} - e^{\eta_0})_i Y_i dt. \quad (9.30)$$

One needs the following conditions in addition to Conditions 9.3.1 and 9.3.2 to proceed.

Condition 9.3.3 For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$, $c_1 \le e^{\eta(t,u)-\eta_0(t,u)} \le c_2$ holds uniformly for some $0 < c_1 < c_2 < \infty$.

Condition 9.3.4 $\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu^2 \phi_\mu^2 e^{k\eta_0} \tilde{S} dt \le c_3$, $\forall \nu, \mu$, for some $c_3 < \infty$, $k = 1, 2$.

By the mean value theorem, Condition 9.3.3 implies the equivalence of $V(\eta - \eta_0)$ and $\mathrm{SKL}(\eta_0, \eta)$ for η in B_0. When η_0 is bounded, Condition 9.3.4 essentially asks for a uniform bound on the fourth moments of ϕ_ν.

Lemma 9.10 *Under Conditions 9.3.1, 9.3.2, and 9.3.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}} f_i^2 e^{\eta_{0,i}} Y_i dt = V(f) + o_p\big((V + \lambda J)(f)\big),$$

where $f_i = f(t, U_i)$. *Similarly,*

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}} f_i g_i e^{\eta_{0,i}} Y_i dt = V(f, g) + o_p\big(\{(V + \lambda J)(f)(V + \lambda J)(g)\}^{1/2}\big).$$

Proof: We only prove the first statement. The same arguments apply to the second. Write $\tau(f) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} f e^{\eta_0} \tilde{S} dt$. Using the Fourier series expansion $f = \sum_\nu f_\nu \phi_\nu$, one has

$$\begin{aligned}
&\left| \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} f_i^2 e^{\eta_{0,i}} Y_i dt - V(f) \right| \\
&\quad = \left| \sum_\nu \sum_\mu f_\nu f_\mu \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} e^{\eta_{0,i}} Y_i dt - \tau(\phi_\nu \phi_\mu) \right\} \right| \\
&\quad \le \left\{ \sum_\nu \sum_\mu \frac{1}{1+\lambda\rho_\nu} \frac{1}{1+\lambda\rho_\mu} \right. \\
&\qquad \left. \times \left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} e^{\eta_{0,i}} Y_i dt - \tau(\phi_\nu \phi_\mu) \right\}^2 \right\}^{1/2} \\
&\qquad \times \left\{ \sum_\nu \sum_\mu (1+\lambda\rho_\nu)(1+\lambda\rho_\mu) f_\nu^2 f_\mu^2 \right\}^{1/2} \\
&\quad = O_p\big(n^{-1/2} \lambda^{-1/r}\big)(V + \lambda J)(f),
\end{aligned}$$

where the Cauchy-Schwartz inequality, Lemma 9.1, and the fact that

$$E\left[\left\{ \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} e^{\eta_{0,i}} Y_i dt - \tau(\phi_\nu \phi_\mu) \right\}^2 \right] = O(n^{-1}) \tag{9.31}$$

are used. To see (9.31), note that

$$\begin{aligned}
&E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} Y dt - \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} \tilde{S} dt \right\}^2 \right] \\
&\quad = E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} (Y - \tilde{S}) dt \right\}^2 \right] \\
&\qquad + E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} \tilde{S} dt - \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} \tilde{S} dt \right\}^2 \right] \\
&\quad \le E\left[\left(\int_{\mathcal{T}} |\phi_\nu \phi_\mu| e^{\eta_0} \tilde{S}^{1/2} dt \right) \left(\int_{\mathcal{T}} |\phi_\nu \phi_\mu| e^{\eta_0} \tilde{S}^{-1/2} E\big[(Y - \tilde{S})^2 \big| U\big] dt \right) \right] \\
&\qquad + E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} \tilde{S} dt \right\}^2 \right] \\
&\quad \le E\left[\left\{ \int_{\mathcal{T}} |\phi_\nu \phi_\mu| e^{\eta_0} \tilde{S}^{1/2} dt \right\}^2 \right] + \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu^2 \phi_\mu^2 e^{2\eta_0} \tilde{S}^2 dt \\
&\quad \le 2c_3.
\end{aligned}$$

This completes the proof. □

Theorem 9.11 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.3.1–9.3.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \tilde{\eta}) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Consequently,

$$\begin{aligned}(V + \lambda J)(\hat{\eta} - \eta_0) &= O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big),\\ \mathrm{SKL}(\eta_0, \hat{\eta}) &= O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).\end{aligned}$$

Proof: By the mean value theorem, Condition 9.3.3, and Lemma 9.10, (9.30) leads to

$$\begin{aligned}&(c_1 V + \lambda J)(\hat{\eta} - \tilde{\eta})\big(1 + o_p(1)\big)\\ &\quad \le \big\{\big(|1-c|V + \lambda J\big)(\hat{\eta} - \tilde{\eta})\big\}^{1/2} O_p\big(\big\{\big(|1-c|V + \lambda J\big)(\tilde{\eta} - \eta_0)\big\}^{1/2}\big)\end{aligned}$$

for some $c \in [c_1, c_2]$. The theorem follows Theorem 9.9. □

9.3.4 Efficient Approximation

As was noted in §8.1, the minimizer $\hat{\eta}$ of (8.1) in $\mathcal{H}$ is, in general, not computable. The minimizer $\hat{\eta}^*$ in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \mathrm{span}\big\{R_J\big((\tilde{X}_j, \tilde{U}_j), \cdot\big), \tilde{\delta}_j = 1\big\}$$

was computed instead, where $\big\{(\tilde{X}_j, \tilde{U}_j, \tilde{\delta}_j)\big\}_{j=1}^q \subseteq \big\{(X_i, U_i, \delta_i)\big\}_{i=1}^n$ is a random subset. We now establish the convergence rates for $\hat{\eta}^*$.

For $h \in \mathcal{H} \ominus \mathcal{H}^*$, one has $\tilde{\delta}_j h(\tilde{X}_j, \tilde{U}_j) = \tilde{\delta}_j J\big(R_J\big((\tilde{X}_j, \tilde{U}_j), \cdot\big), h\big) = 0$, so $\sum_{j=1}^q \int_{\mathcal{T}} h_j^2 d\tilde{N}_j(t) = \sum_{j=1}^q \tilde{\delta}_j h^2(\tilde{X}_j, \tilde{U}_j) = 0$, where $\tilde{N}_j(t) = I_{[\tilde{X}_j \le t, \tilde{\delta}_j = 1]}$ and $h_j(t) = h(t, \tilde{U}_j)$.

Lemma 9.12 *Under Conditions 9.3.1, 9.3.2, and 9.3.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$, $V(h) = o_p\big(\lambda J(h)\big)$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$.

Proof: Define $\tau(f) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} f e^{\eta_0} \tilde{S} dt$. From (9.20)–(9.22), Condition 9.3.4, and the proof of (9.31), one has

$$\begin{aligned}&E\left[\left\{\int_{\mathcal{T}} \phi_\nu \phi_\mu dN(t) - \tau(\phi_\nu \phi_\mu)\right\}^2\right]\\ &\quad = E\left[\left\{\int_{\mathcal{T}} \phi_\nu \phi_\mu dM(t)\right\}^2\right] + E\left[\left\{\int_{\mathcal{T}} \phi_\nu \phi_\mu e^{\eta_0} Y dt - \tau(\phi_\nu \phi_\mu)\right\}^2\right]\\ &\quad \le \tau(\phi_\nu^2 \phi_\mu^2) + 2c_3 \le 3c_3.\end{aligned}$$

By the same arguments used in the proof of Lemma 9.10,

$$V(h) = \left| \frac{1}{q} \sum_{j=1}^{q} \int_{\mathcal{T}} h_j^2 d\tilde{N}_j(t) - V(h) \right| = O_p\big(q^{-1/2}\lambda^{-1/r}\big)(V + \lambda J)(h).$$

The lemma follows. □

Theorem 9.13 *Let η^* be the projection of $\hat{\eta}$ in $\mathcal{H}^*$. Assume $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ for some $p \in [1,2]$. Under Conditions 9.3.1–9.3.4, as $\lambda \to 0$ and $q\lambda^{2/r} \to \infty$,*

$$\lambda J(\hat{\eta} - \eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big),$$
$$V(\hat{\eta} - \eta^*) = o_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \eta^*$ in (9.26), one has

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{ \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i dN_i(t) - \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i e^{\hat{\eta}_i} Y_i dt \right\} + \lambda J(\hat{\eta}, \hat{\eta} - \eta^*) = 0. \tag{9.32}$$

Some algebra yields

$$\begin{aligned} \lambda J(\hat{\eta} - \eta^*) = {} & \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i dM_i(t) \\ & - \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i (e^{\hat{\eta}} - e^{\eta_0})_i Y_i dt; \end{aligned} \tag{9.33}$$

remember that $J(\eta^*, \hat{\eta} - \eta^*) = 0$. Now, with $\beta_\nu = n^{-1}\sum_{i=1}^{n} \int_{\mathcal{T}} \phi_{\nu,i} dM_i(t)$,

$$\begin{aligned} \left| \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i dM_i(t) \right| &= \left| \sum_\nu (\hat{\eta}_\nu - \eta_\nu^*)\beta_\nu \right| \\ &= \left\{ \sum_\nu (1 + \lambda\rho_\nu)(\hat{\eta}_\nu - \eta_\nu^*)^2 \right\}^{1/2} \left\{ \sum_\nu (1 + \lambda\rho_\nu)^{-1} \beta_\nu^2 \right\}^{1/2} \\ &= \big\{(V + \lambda J)(\hat{\eta} - \eta^*)\big\}^{1/2} O_p\big(n^{-1/2}\lambda^{-1/2r}\big). \end{aligned} \tag{9.34}$$

By the mean value theorem, Condition 9.3.3, and Lemmas 9.10 and 9.12,

$$\begin{aligned} &\left| \frac{1}{n}\sum_{i=1}^{n} \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i (e^{\hat{\eta}} - e^{\eta_0})_i Y_i dt \right| \\ &\qquad = o_p\big(\big\{\lambda J(\hat{\eta} - \eta^*)(V + \lambda J)(\hat{\eta} - \eta_0)\big\}^{1/2}\big); \end{aligned} \tag{9.35}$$

see Problem 9.7. Plugging (9.34) and (9.35) into (9.33) and applying Theorem 9.11 and Lemma 9.12, one has

$$\lambda J(\hat{\eta}-\eta^*) = \{\lambda J(\hat{\eta}-\eta^*)\}^{1/2}\{O_p(n^{-1/2}\lambda^{-1/2r}) + o_p(\lambda^{p/2})\}.$$

The theorem follows. □

We shall now calculate $(V+\lambda J)(\hat{\eta}^*-\eta^*)$. Setting $f=\hat{\eta}^*$ and $g=\hat{\eta}^*-\eta^*$ in (9.26), one has

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i dN_i(t) - \int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i e^{\hat{\eta}_i^*}Y_i dt\right\} + \lambda J(\hat{\eta}^*, \hat{\eta}^*-\eta^*) = 0. \quad (9.36)$$

Setting $f=\hat{\eta}$ and $g=\hat{\eta}-\hat{\eta}^*$ in (9.26), one gets

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}}(\hat{\eta}-\hat{\eta}^*)_i dN_i(t) - \int_{\mathcal{T}}(\hat{\eta}-\hat{\eta}^*)_i e^{\hat{\eta}_i}Y_i dt\right\} + \lambda J(\hat{\eta}, \hat{\eta}-\hat{\eta}^*) = 0. \quad (9.37)$$

Adding (9.36), (9.37) and subtracting (9.32), and noting that $J(\hat{\eta}-\eta^*, \eta^*) = J(\hat{\eta}-\eta^*, \hat{\eta}^*) = 0$, some algebra yields

$$\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i(e^{\hat{\eta}^*}-e^{\eta^*})_i Y_i dt + \lambda J(\hat{\eta}^*-\eta^*) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i(e^{\hat{\eta}}-e^{\eta^*})_i Y_i dt \quad (9.38)$$

see Problem 9.8. Condition 9.3.3 has to be modified to include η^* and $\hat{\eta}^*$ in the convex set B_0.

Theorem 9.14 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.3.1–9.3.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V+\lambda J)(\hat{\eta}^*-\eta^*) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

Consequently,

$$(V+\lambda J)(\hat{\eta}^*-\eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p),$$
$$\mathrm{SKL}(\eta_0, \hat{\eta}^*) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

Proof: By the mean value theorem, Condition 9.3.3, Lemma 9.10, and Theorem 9.13, (9.38) leads to

$$(c_1 V + \lambda J)(\hat{\eta}^*-\eta^*) \le \{(V+\lambda J)(\hat{\eta}^*-\eta^*)\}^{1/2} O_p(n^{-1/2}\lambda^{-1/2r} + \lambda^{p/2}).$$

The first part of the theorem follows. The rest is straightforward. □

9.3.5 Convergence Under Incorrect Model

For $\eta_0 \notin \mathcal{H}$, one defines the relative Kullback-Leibler distance as

$$\mathrm{RKL}(\eta_0, \eta) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \{e^{\eta(t,u)} - \eta(t,u) e^{\eta_0(t,u)}\} \tilde{S}(t) dt.$$

The minimizer η_0^* of $\mathrm{RKL}(\eta_0, \eta)$ in $\mathcal{H}$, when it exists, satisfies

$$\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} f(t,u) \{e^{\eta_0^*(t,u)} - e^{\eta_0(t,u)}\} \tilde{S}(t) dt = 0, \quad \forall f \in \mathcal{H}.$$

Substituting η_0^* for η_0 in §§9.3.1–9.3.4, the analysis remain valid under a couple of extra conditions.

Condition 9.3.0 $fg \in \mathcal{H}$, $\forall f, g \in \mathcal{H}$.

Condition 9.3.5 $\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu^2 \big(e^{\eta_0} - e^{\eta_0^*}\big)^2 \tilde{S} dt < c_4$ holds uniformly for some $c_4 < \infty$, $\forall \nu$.

Further details are left as an exercise (Problem 9.9).

9.4 Rates for Regression Estimates

We now establish convergence rates for regression estimates, which include those discussed in Chaps. 3, 5 and §8.6.

A formulation more general than (5.1) is presented, and a quadratic functional V is defined in the general setting. Rates are established in terms of $V(\hat{\eta} - \eta_0)$. The first step is, once again, the analysis of a linear approximation $\tilde{\eta}$.

9.4.1 General Formulation

Denote by $l(\eta; y)$ a minus log likelihood of η with observation y. We shall consider the penalized likelihood functional

$$\frac{1}{n} \sum_{i=1}^{n} l\big(\eta(x_i); Y_i\big) + \frac{\lambda}{2} J(\eta). \tag{9.39}$$

When η is the canonical parameter of an exponential family distribution, (9.39) reduces to (5.1) on page 176. The general formulation of (9.39) covers the noncanonical links used in the gamma family, the inverse Gaussian family, and the negative binomial family of §5.4. It also covers the log likelihoods of §8.6, where $\eta(x)$ was written as $\mu(u)$ and y consisted of several

components. The dispersion parameter of an exponential family distribution can be absorbed into λ, known or unknown, but the ν parameter in the negative binomial family of §5.4.6 and in the accelerated life models of §8.6 is assumed to be known.

Write $u(\eta; y) = dl/d\eta$ and $w(\eta; y) = d^2l/d\eta^2$; it is assumed that

$$E\big[u\big(\eta_0(x); Y\big)\big] = 0, \quad E\big[u^2\big(\eta_0(x); Y\big)\big] = \sigma^2 E\big[w\big(\eta_0(x); Y\big)\big], \tag{9.40}$$

which hold for all the log likelihoods appearing in §§5.4 and 8.6, where σ^2 is a constant. Let $f(x)$ be the limiting density of x_i. Write $v_\eta(x) = E\big[w\big(\eta(x); Y\big)\big]$ and define

$$V(g) = \int_{\mathcal{X}} g^2(x) v_{\eta_0}(x) f(x) dx. \tag{9.41}$$

The specific forms of V for the families of §§5.4 and 8.6 are easy to work out; see Problem 9.10. Convergence rates for the minimizer $\hat{\eta}$ of (9.39) shall be established in terms of $V(\hat{\eta} - \eta_0)$.

9.4.2 Linear Approximation

The following conditions are needed in our analysis, which are carbon copies of Conditions 9.2.1 and 9.2.2 but with V as defined in (9.41) in the regression setting.

Condition 9.4.1 V is completely continuous with respect to J.

Condition 9.4.2 For ν sufficiently large and some $\beta > 0$, the eigenvalues ρ_ν of J with respect to V satisfy $\rho_\nu > \beta\nu^r$, where $r > 1$.

Consider the quadratic functional

$$\frac{1}{n}\sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big)\eta(x_i) + \frac{1}{2}V(\eta - \eta_0) + \frac{\lambda}{2}J(\eta). \tag{9.42}$$

Plugging the Fourier series expansions $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ into (9.42), it is easy to show that the minimizer $\tilde{\eta}$ of (9.42) has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu),$$

which are linear in $\beta_\nu = -n^{-1}\sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big)\phi_\nu(x_i)$; see Problem 9.11. Note that $E[\beta_\nu] = 0$ and $E\big[\beta_\nu^2\big] = \sigma^2/n$. The following theorem can be easily proved parallel to Theorem 9.2.

Theorem 9.15 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 9.4.1 and 9.4.2, as* $n \to \infty$ *and* $\lambda \to 0$,

$$(V + \lambda J)(\tilde{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

9.4.3 Approximation Error and Main Result

We now turn to the approximation error $\hat{\eta} - \tilde{\eta}$. Define

$$A_{g,h}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} l\big((g+\alpha h)(x_i); Y_i\big) + \frac{\lambda}{2} J(g+\alpha h),$$

$$B_{g,h}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big)(g+\alpha h)(x_i) + \frac{1}{2} V(g+\alpha h-\eta_0) + \frac{\lambda}{2} J(g+\alpha h).$$

It can be easily shown that

$$\dot{A}_{g,h}(0) = \frac{1}{n}\sum_{i=1}^{n} u\big(g(x_i); Y_i\big)h(x_i) + \lambda J(g,h), \tag{9.43}$$

$$\dot{B}_{g,h}(0) = \frac{1}{n}\sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big)h(x_i) + V(g-\eta_0, h) + \lambda J(g,h). \tag{9.44}$$

Setting $g = \hat{\eta}$ and $h = \hat{\eta} - \tilde{\eta}$ in (9.43), one has

$$\frac{1}{n}\sum_{i=1}^{n} u\big(\hat{\eta}(x_i); Y_i\big)(\hat{\eta}-\tilde{\eta})(x_i) + \lambda J(\hat{\eta}, \hat{\eta}-\tilde{\eta}) = 0, \tag{9.45}$$

and setting $g = \tilde{\eta}$ and $h = \hat{\eta} - \tilde{\eta}$ in (9.44), one gets

$$\frac{1}{n}\sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big)(\hat{\eta}-\tilde{\eta})(x_i) + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \lambda J(\tilde{\eta}, \hat{\eta}-\tilde{\eta}) = 0. \tag{9.46}$$

Subtracting (9.46) from (9.45), some algebra yields

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} \big\{u\big(\hat{\eta}(x_i); Y_i\big) - u\big(\tilde{\eta}(x_i); Y_i\big)\big\}(\hat{\eta}-\tilde{\eta})(x_i) + \lambda J(\hat{\eta}-\tilde{\eta}) \\ &= V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) - \frac{1}{n}\sum_{i=1}^{n} \big\{u\big(\tilde{\eta}(x_i); Y_i\big) - u\big(\eta_0(x_i); Y_i\big)\big\}(\hat{\eta}-\tilde{\eta})(x_i).\end{aligned} \tag{9.47}$$

By the mean value theorem,

$$\begin{aligned}u\big(\hat{\eta}(x_i); Y_i\big) - u\big(\tilde{\eta}(x_i); Y_i\big) &= w\big(\eta_1(x_i); Y_i\big)(\hat{\eta}-\tilde{\eta})(x_i), \\ u\big(\tilde{\eta}(x_i); Y_i\big) - u\big(\eta_0(x_i); Y_i\big) &= w\big(\eta_2(x_i); Y_i\big)(\tilde{\eta}-\eta_0)(x_i),\end{aligned} \tag{9.48}$$

where η_1 is a convex combination of $\hat{\eta}$ and $\tilde{\eta}$, and η_2 is that of $\tilde{\eta}$ and η_0.

To proceed, one needs the following conditions in addition to Conditions 9.4.1 and 9.4.2.

Condition 9.4.3 For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$, $c_1 w\big(\eta_0(x);Y\big) \le w\big(\eta(x);Y\big) \le c_2 w\big(\eta_0(x);Y\big)$ holds uniformly for some $0 < c_1 < c_2 < \infty$, $\forall x \in \mathcal{X}$, $\forall Y$.

Condition 9.4.4 $\mathrm{Var}\big[\phi_\nu(X)\phi_\mu(X)w\big(\eta_0(X),Y\big)\big] \le c_3$ for some $c_3 < \infty$, $\forall \nu, \mu$.

To understand the practical meanings of these conditions, one needs to work out their specific forms for the families of §§5.4 and 8.6 (Problem 9.12). Roughly speaking, Condition 9.4.3 concerns the equivalence of the information in B_0 and Condition 9.4.4 asks for a uniform bound for the fourth moments of $\phi_\nu(X)$.

Lemma 9.16 *Under Conditions 9.4.1, 9.4.2, and 9.4.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i)h(x_i)w\big(\eta_0(x_i);Y_i\big) = V(g,h) + o_p\big(\big\{(V+\lambda J)(g)(V+\lambda J)(h)\big\}^{1/2}\big).$$

Proof: Write $\tau(g) = \int_{\mathcal{X}} g(x)v_{\eta_0}(x)f(x)dx$. Under Condition 9.4.4,

$$\frac{1}{n}\sum_{i=1}^{n} \phi_\nu(x_i)\phi_\mu(x_i)w\big(\eta_0(x_i);Y_i\big) - \tau(\phi_\nu\phi_\mu) \le \frac{c_3}{n}.$$

Write $g = \sum_\nu g_\nu\phi_\nu$ and $h = \sum_\nu h_\nu\phi_\nu$. Similar to the proofs of Lemmas 9.5 and 9.10, as $n\lambda^{2/r} \to \infty$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} g(x_i)h(x_i)w\big(\eta_0(x_i);Y_i\big) - V(g,h)\right| \\
&\quad= \left|\sum_\nu\sum_\mu g_\nu h_\mu\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_\nu(x_i)\phi_\mu(x_i)w\big(\eta_0(x_i);Y_i\big) - \tau(\phi_\nu\phi_\mu)\right\}\right| \\
&\quad\le \left\{\sum_\nu\sum_\mu \frac{1}{1+\lambda\rho_\nu}\frac{1}{1+\lambda\rho_\mu}\right. \\
&\qquad\left.\times\left\{\frac{1}{n}\sum_{i=1}^{n}\phi_\nu(x_i)\phi_\mu(x_i)w\big(\eta_0(x_i);Y_i\big) - \tau(\phi_\nu\phi_\mu)\right\}^2\right\}^{1/2} \\
&\qquad\times\left\{\sum_\nu\sum_\mu (1+\lambda\rho_\nu)(1+\lambda\rho_\mu)g_\nu^2 h_\mu^2\right\}^{1/2} \\
&\quad= \big\{(V+\lambda J)(g)(V+\lambda J)(h)\big\}^{1/2} O_p\big(n^{-1/2}\lambda^{-1/r}\big) \\
&\quad= o_p\big(\big\{(V+\lambda J)(g)(V+\lambda J)(h)\big\}^{1/2}\big),
\end{aligned}$$

where the Cauchy-Schwartz inequality and Lemma 9.1 are used. □

Theorem 9.17 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 9.4.1–9.4.4, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \tilde{\eta}) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Consequently,

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Substituting (9.48) into (9.47), and applying Condition 9.4.3 and Lemma 9.16, one has

$$\begin{aligned}&(c_1 V + \lambda J)(\hat{\eta} - \tilde{\eta})\big(1 + o_p(1)\big)\\ &\quad \le \big\{\big(|1 - c|V + \lambda J\big)(\hat{\eta} - \tilde{\eta})\big\}^{1/2} O_p\big(\big\{\big(|1 - c|V + \lambda J\big)(\tilde{\eta} - \eta_0)\big\}^{1/2}\big),\end{aligned}$$

for some $c \in [c_1, c_2]$. The theorem follows Theorem 9.15. □

9.4.4 Efficient Approximation

While the minimizer $\hat{\eta}$ of (9.39) in $\mathcal{H}$ is always computable, the computation is in general of order $O(n^3)$. For more scalable computation, one may consider the minimizer $\hat{\eta}^*$ in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\big\{R_J(z_j, \cdot), j = 1, \ldots, q\big\},$$

where $\{z_j\}$ is a random subset of $\{x_i\}$. We now establish the convergence rates for $\hat{\eta}^*$.

Lemma 9.18 *Under Conditions 9.4.1, 9.4.2, and 9.4.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$, $V(h) = o_p\big(\lambda J(h)\big)$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$.

Proof: For $h \in \mathcal{H} \ominus \mathcal{H}^*$, $h(z_j) = J\big(R_J(z_j, \cdot), h\big) = 0$. Denote by $\tilde{Y}_j$ the response associated with z_j. Similar to the proof of Lemma 9.16,

$$V(h) = \left|V(h) - \frac{1}{q}\sum_{j=1}^{q} h^2(z_j) w\big(\eta_0(z_j); \tilde{Y}_j\big)\right| = O_p\big(q^{-1/2}\lambda^{-1/r}\big)(V + \lambda J)(h).$$

The lemma follows. □

Let η^* be the projection of $\hat{\eta}$ in $\mathcal{H}^*$; one also needs to include η^* and $\hat{\eta}^*$ in the convex set B_0 in Condition 9.4.3. Note that $\hat{\eta} - \eta^* \in \mathcal{H} \ominus \mathcal{H}^*$, so $J(\eta^*, \hat{\eta} - \eta^*) = 0$. Setting $g = \hat{\eta}$ and $h = \hat{\eta} - \eta^*$ in (9.43), one has

$$\frac{1}{n}\sum_{i=1}^{n} u\big(\hat{\eta}(x_i); Y_i\big)(\hat{\eta} - \eta^*)(x_i) + \lambda J(\hat{\eta}, \hat{\eta} - \eta^*) = 0. \tag{9.49}$$

By the mean value theorem and Condition 9.4.3, (9.49) leads to

$$\begin{aligned}\lambda J(\hat\eta-\eta^*) &= -\frac{1}{n}\sum_{i=1}^n \big\{u\big(\hat\eta(x_i);Y_i\big)-u\big(\eta_0(x_i);Y_i\big)\big\}(\hat\eta-\eta^*)(x_i)\\ &\quad -\frac{1}{n}\sum_{i=1}^n u\big(\eta_0(x_i);Y_i\big)(\hat\eta-\eta^*)(x_i)\\ &= -\frac{c}{n}\sum_{i=1}^n w\big(\eta_0(x_i);Y_i\big)(\hat\eta-\eta_0)(x_i)(\hat\eta-\eta^*)(x_i)\\ &\quad -\frac{1}{n}\sum_{i=1}^n u\big(\eta_0(x_i);Y_i\big)(\hat\eta-\eta^*)(x_i) \qquad (9.50)\end{aligned}$$

for some constant c; remember that $J(\eta^*,\hat\eta-\eta^*)=0$. By Lemma 9.16, the first term in (9.50) is of the order

$$\begin{aligned}&\left|\frac{c}{n}\sum_{i=1}^n w\big(\eta_0(x_i);Y_i\big)(\hat\eta-\eta_0)(x_i)(\hat\eta-\eta^*)(x_i)\right|\\ &\qquad = \big\{(V+\lambda J)(\hat\eta-\eta_0)(V+\lambda J)(\hat\eta-\eta^*)\big\}^{1/2}O_p(1). \qquad (9.51)\end{aligned}$$

For the second term, one has

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n u\big(\eta_0(x_i);Y_i\big)(\hat\eta-\eta^*)(x_i)\right|\\ &\qquad = \left|\sum_\nu(\hat\eta-\eta^*)_\nu\Big\{\frac{1}{n}\sum_{i=1}^n u(\eta_0(x_i);Y_i)\phi_\nu(x_i)\Big\}\right|\\ &\qquad \le \left\{\sum_\nu(\hat\eta-\eta^*)_\nu^2(1+\lambda\rho_\nu)\right\}^{1/2}\left\{\sum_\nu\frac{\beta_\nu^2}{1+\lambda\rho_\nu}\right\}^{1/2}\\ &\qquad = \big\{(V+\lambda J)(\hat\eta-\eta^*)\big\}^{1/2}O_p\big(n^{-1/2}\lambda^{-1/2r}\big), \qquad (9.52)\end{aligned}$$

where $\hat\eta-\eta^*=\sum_\nu(\hat\eta-\eta^*)_\nu\phi_\nu$ and $E\big[\beta_\nu^2\big]=\sigma^2/n$. Combining (9.50)–(9.52) and applying Theorem 9.17 and Lemma 9.18, trivial manipulation yields the following theorem.

Theorem 9.19 *Assume $\sum_\nu \rho_\nu^p\eta_{\nu,0}^2<\infty$ for some $p\in[1,2]$. Under Conditions 9.4.1–9.4.4, as $\lambda\to0$ and $q\lambda^{2/r}\to\infty$,*

$$(V+\lambda J)(\hat\eta-\eta^*)=O_p\big(n^{-1}\lambda^{-1/r}+\lambda^p\big).$$

We now turn to $\hat\eta^*-\eta^*$. Setting $g=\hat\eta$ and $h=\hat\eta-\hat\eta^*$ in (9.43), one has

$$\frac{1}{n}\sum_{i=1}^n u\big(\hat\eta(x_i);Y_i\big)(\hat\eta-\hat\eta^*)(x_i)+\lambda J(\hat\eta,\hat\eta-\hat\eta^*)=0. \qquad (9.53)$$

Setting $g = \hat{\eta}^*$ and $h = \hat{\eta}^* - \eta^*$ in (9.43) leads to

$$\frac{1}{n}\sum_{i=1}^{n} u\big(\hat{\eta}^*(x_i); Y_i\big)(\hat{\eta}^* - \eta^*)(x_i) + \lambda J(\hat{\eta}^*, \hat{\eta}^* - \eta^*) = 0. \tag{9.54}$$

Adding (9.53), (9.54) and subtracting (9.49), some algebra yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \big\{u\big(\hat{\eta}^*(x_i); Y_i\big) - u\big(\eta^*(x_i); Y_i\big)\big\}(\hat{\eta}^* - \eta^*)(x_i) + \lambda J(\hat{\eta}^* - \eta^*) \\ = \frac{1}{n}\sum_{i=1}^{n} \big\{u\big(\hat{\eta}(x_i); Y_i\big) - u\big(\eta^*(x_i); Y_i\big)\big\}(\hat{\eta}^* - \eta^*)(x_i);\end{aligned} \tag{9.55}$$

note that $J(\hat{\eta} - \eta^*, \eta^* - \hat{\eta}^*) = 0$. By the mean value theorem, Condition 9.4.3, and Lemma 9.16, the first term on the left-hand side of (9.55) is equal to

$$cV(\hat{\eta}^* - \eta^*) + o_p\big(\lambda J(\hat{\eta}^* - \eta^*)\big)$$

for some constant $c \geq c_1 > 0$. Similarly, the right-hand side is of the order

$$\big\{(V + \lambda J)(\hat{\eta}^* - \eta^*)(V + \lambda J)(\hat{\eta} - \eta^*)\big\}^{1/2}\big(1 + o_p(1)\big).$$

Combining these with Theorems 9.17 and 9.19, one obtains the following theorem.

Theorem 9.20 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 9.4.1–9.4.4, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta}^* - \eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Consequently,

$$\begin{aligned}(V + \lambda J)(\hat{\eta}^* - \hat{\eta}) &= O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big), \\ (V + \lambda J)(\hat{\eta}^* - \eta_0) &= O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).\end{aligned}$$

9.4.5 Convergence Under Incorrect Model

For $\eta_0 \notin \mathcal{H}$, one may define the relative Kullback-Leibler distance as

$$\mathrm{RKL}(\eta_0, \eta) = \int_{\mathcal{X}} E\big[l\big(\eta(x); Y\big)\big] f(x)dx,$$

where the expectation is taken under η_0. The minimizer η_0^* of $\mathrm{RKL}(\eta_0, \eta)$ in $\mathcal{H}$, when it exists, satisfies

$$\int_{\mathcal{X}} g(x) E\big[u\big(\eta_0^*(x); Y\big)\big] f(x)dx = 0, \quad \forall g \in \mathcal{H}.$$

Substituting η_0^* for η_0 in (9.42) but not in the definition of V, Theorem 9.15 may be proved under some extra condition that assures uniformly bounded $E[\beta_\nu^2]$. For Theorem 9.17 to hold, further conditions are also needed to ensure the uniform boundedness of

$$E\left[\left\{\phi_\nu(X)\phi_\mu(X)w\big(\eta_0^*(X);Y\big) - \tau(\phi_\nu\phi_\mu)\right\}^2\right];$$

details are tedious. It would be easier to work with specific families than with the general setting; see Problem 9.13.

9.5 Bibliographic Notes

Section 9.1

An general theory of eigenvalue analysis can be found in Weinberger (1974). Results on eigenvalues related to smoothing splines can be found in, e.g., Cox (1984, 1988) and Utreras (1981, 1983, 1988), among others. Example 9.2 is taken from Gu (1996).

Section 9.2

An asymptotic theory was developed by Silverman (1982) for the minimizer of (7.12), which laid the groundwork for later analysis. Cox and O'Sullivan (1990) developed a general asymptotic theory for penalized likelihood estimates, of which the estimate of Silverman (1982) was listed as an example. The materials of §§9.2.1–9.2.3 represent a refinement of the analysis found in Gu and Qiu (1993, §§5 and 6), where the efficient approximation was first proposed and studied. The analysis of §9.2.4 was noted by Gu (1998b). The adaptations of §§9.2.5–9.2.7 are found in Gu (1992d, 1995a, 1995b).

Section 9.3

The materials of this section are a refined version of the analysis found in Gu (1996). For $\mathcal{U}$ a singleton, the analyses of Antoniadis (1989) and Cox and O'Sullivan (1990) apply, but not to the efficient approximation.

Section 9.4

The analysis in the general setting as presented is adapted from that of Gu and Qiu (1994), where η was taken as the canonical parameter of an exponential family distribution, as in §5.1. The analysis of Cox and O'Sullivan (1990) also applies in the setting of §5.1. The efficient approximation is taken from Gu and Kim (2002).

Convergence rates for penalized least squares estimates have been studied extensively in the literature. For results on multidimensional domains, see Cox (1984), Utreras (1988), Chen (1991), and Lin (2000).

9.6 Problems

Section 9.2

9.1 Assume $\sum_\nu \nu^{pr}\eta_{\nu,0}^2 < \infty$ for some $p > 1$. Show that the rates in Theorem 9.2 can be improved to $O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big)$, with p up to 2.

9.2 Verify (9.4) and (9.5).

9.3 In the setting of §9.2.5, state and prove the counterparts of all the lemmas and theorems appearing in §§9.2.1–9.2.3.

9.4 In the setting of §9.2.6, state and prove the counterparts of all the lemmas and theorems appearing in §§9.2.1–9.2.3.

9.5 In the setting of §9.2.7, state and prove the counterparts of all the lemmas and theorems appearing in §§9.2.1–9.2.3.

Section 9.3

9.6 Show that the minimizer $\tilde{\eta}$ of (9.25) has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu),$$

where

$$\beta_\nu = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}} \phi_{\nu,i}\, dM_i(t)$$

for $\phi_{\nu,i}(t) = \phi_\nu(t, U_i)$ satisfy $E[\beta_\nu] = 0$ and $E\big[\beta_\nu^2\big] = n^{-1}$.

9.7 Prove (9.35).

9.8 Prove (9.38).

9.9 When $\eta_0 \notin \mathcal{H}$, substituting η_0^* of §9.3.5 for η_0 in §§9.3.1–9.3.4, show that the convergence rates remain valid under Conditions 9.3.0–9.3.5.

Section 9.4

9.10 Specify the definition $V(g) = \int_{\mathcal{X}} g^2(x) v_{\eta_0}(x) f(x) dx$ for the families of §§5.4 and 8.6.

9.11 Show that the minimizer $\tilde{\eta}$ of (9.42) has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu),$$

where

$$\beta_\nu = -\frac{1}{n} \sum_{i=1}^{n} u\big(\eta_0(x_i); Y_i\big) \phi_\nu(x_i)$$

satisfy $E[\beta_\nu] = 0$ and $E\big[\beta_\nu^2\big] = \sigma^2/n$.

9.12 Specify Conditions 9.4.3 and 9.4.4 for the families of §§5.4 and 8.6.

9.13 For the families of §§5.4 and 8.6, specify the extra conditions needed to extend Theorems 9.15 and 9.17 to the case $\eta_0 \notin \mathcal{H}$.

10
Penalized Pseudo Likelihood

The density estimation of (7.1) is infeasible on high-dimensional $\mathcal{X}$ due to the prohibitive cost of $\int_{\mathcal{X}} e^{\eta(x)}$ via multivariate numerical integration. As an alternative, Jeon and Lin (2006) proposed a certain penalized pseudo likelihood, replacing $\int_{\mathcal{X}} e^{\eta(x)}$ by an integral of the form $\int_{\mathcal{X}} \eta(x)\rho(x)$ for some $\rho(x)$, which is computable as sums of products of univariate integrals.

The conditional density estimation of (7.30) with a continuous $\mathcal{Y}$ can be crippled computationally by repeated numerical integrations, so can the hazard estimation of (8.1) with continuous covariates U_i. Pseudo likelihoods can also be devised in these settings to avoid repeated numerical integrations, gaining numerical efficiency at the cost of performance degradation.

Parallel developments are presented in the settings of density estimation (§10.1), conditional density estimation (§10.3), and hazard estimation (§10.4). For the approach to be practically viable, one needs smoothing parameter selection methods that are also void of the offending numerical ingredients. Likewise, the Kullback-Leibler projection is to be replaced by certain square error projections in the respective settings. Open-source software is illustrated using simulated and real-data examples, and empirical comparisons are made against the respective penalized likelihood methods in terms of numerical efficiency and statistical performance.

Under the technical framework developed in Chap. 9, one can also calculate the asymptotic convergence rates for the estimates obtained via penalized pseudo likelihood (§§10.2, 10.3.6, and 10.5).

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_10,

10.1 Density Estimation on Product Domains

For the computation of (7.1), integrals of the form $\int_{\mathcal{X}} h(x)e^{\eta(x)}$ have to be performed over and over while $\eta(x)$ is being updated iteratively, as seen in (7.5)–(7.7). Numerical integration is costly on high dimensional domains, which limits the practical applicability of the method. In an effort to relieve the numerical burden associated with multidimensional integrations, Jeon and Lin (2006) proposed to calculate the minimizer η_λ of

$$\frac{1}{n}\sum_{i=1}^{n} e^{-\eta(X_i)} + \int_{\mathcal{X}} \eta(x)\rho(x) + \frac{\lambda}{2}J(\eta) \tag{10.1}$$

for some known density $\rho(x)$ on $\mathcal{X}$, and the resulting density estimate is of the form $f(x) \propto e^{\eta_\lambda(x)}\rho(x)$.

An informal analysis reveals how (10.1) works, and the existence and the computation of the minimizer of (10.1) are similar to that of (7.1). A cross-validation scheme is devised for smoothing parameter selection, and a certain square error projection provides an alternative to Kullback-Leibler projection in the setting; the key here is to avoid integrals of the form $\int_{\mathcal{X}} h(x)e^{\eta(x)}$. Empirical performances are assessed and software tools are illustrated using simulated and real-data examples.

The asymptotic convergence rates of the minimizers of (10.1) are to be found in §10.2.

10.1.1 Pseudo and Genuine Likelihoods

To see how (10.1) works, let $n \to \infty$ and $\lambda \to 0$ in (10.1) and consider the limiting convex functional $P(\eta) = \int_{\mathcal{X}} \{e^{-\eta(x)}f(x) + \eta(x)\rho(x)\}$. Write $A_{\tilde\eta,h}(\alpha) = P(\tilde\eta + \alpha h)$, where $\tilde\eta$ minimizes $P(\eta)$ and α is a scalar. One has

$$\dot A_{\tilde\eta,h}(0) = \int_{\mathcal{X}} \{\rho(x) - e^{-\tilde\eta(x)}f(x)\}h(x) = 0, \quad \forall h,$$

so $f(x) = e^{\tilde\eta(x)}\rho(x)$. The quadratic approximation of $P(\eta)$ at $\tilde\eta$ is thus

$$\begin{aligned} P(\eta) = A_{\tilde\eta,\eta-\tilde\eta}(1) &\approx A_{\tilde\eta,\eta-\tilde\eta}(0) + \dot A_{\tilde\eta,\eta-\tilde\eta}(0) + \tfrac{1}{2}\ddot A_{\tilde\eta,\eta-\tilde\eta}(0) \\ &= P(\tilde\eta) + \tfrac{1}{2}\int_{\mathcal{X}} \big(\eta(x) - \tilde\eta(x)\big)^2\rho(x) \end{aligned}$$

Parallel analysis can be performed on the limiting functional of (7.1), $G(\eta) = -\int_{\mathcal{X}} \eta(x)f(x) + \log\int_{\mathcal{X}} e^{\eta(x)}$, with $B_{\tilde\eta,h}(\alpha) = G(\tilde\eta + \alpha h)$,

$$\dot B_{\tilde\eta,h}(0) = \int_{\mathcal{X}} \big\{\big(\int_{\mathcal{X}} e^{\tilde\eta(x)}\big)^{-1} e^{\tilde\eta(x)} - f(x)\big\}h(x) = 0, \quad \forall h,$$

and

$$G(\eta) \approx G(\tilde\eta) + \tfrac{1}{2}\Big[\int_{\mathcal{X}} \big(\eta(x) - \tilde\eta(x)\big)^2 f(x) - \big\{\int_{\mathcal{X}} \big(\eta(x) - \tilde\eta(x)\big)f(x)\big\}^2\Big];$$

see Problem 10.1. The contrast between the pseudo likelihood and the genuine likelihood can be perceived via $P(\eta)$ and $G(\eta)$.

10.1.2 Preliminaries

Write $L(f) = n^{-1}\sum_{i=1}^n e^{-f(X_i)} + \int_{\mathcal{X}} f(x)\rho(x)$. It is easy to verify that $L(f)$ is continuous, convex, and Fréchet differentiable. Let $\{\phi_\nu\}_{\nu=1}^m$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$ and S be an $n \times m$ matrix with the (i,ν)th entry $\phi_\nu(X_i)$. If S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$, and $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$. See Problem 10.2. By Theorem 2.9, the minimizer of (10.1) uniquely exists when S is of full column rank, which we will assume.

Suppose $J(f)$ annihilates constant and consider a tensor sum decomposition $\mathcal{H} = \{1\} \oplus \mathcal{G}$. Writing $\eta = d + g$ with $g \in \mathcal{G}$, (10.1) becomes

$$\frac{1}{n}\sum_{i=1}^n e^{-g(X_i)-d} + \int_{\mathcal{X}} \{g(x) + d\}\rho(x) + \frac{\lambda}{2}J(g). \tag{10.2}$$

Fixing $g(x)$, noting that $\int_{\mathcal{X}} \rho(x) = 1$, the d that minimizes (10.2) is given by $e^d = n^{-1}\sum_{i=1}^n e^{-g(X_i)}$; the minimizer of (10.1) is seen to be "normalized" to satisfy $n^{-1}\sum_{i=1}^n e^{-\eta(X_i)} = 1$. Plugging this back into (10.2) and dropping terms not involving $g(x)$, one has a "profile" functional

$$\log\left\{\frac{1}{n}\sum_{i=1}^n e^{-g(X_i)}\right\} + \int_{\mathcal{X}} g(x)\rho(x)dx + \frac{\lambda}{2}J(g). \tag{10.3}$$

Without loss of inferential efficiency, one may minimize (10.1) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\{R_J(Z_j, \cdot), j = 1, \ldots, q\}, \tag{10.4}$$

where $\{Z_j\}$ is a random subset of $\{X_i\}$; see §10.2.3. One has an expression,

$$g(x) = \sum_\nu d_\nu\phi_\nu(x) + \sum_j c_j R_J(Z_j, x) = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{10.5}$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J \ominus \{1\}$ and $\xi_j(x) = R_J(Z_j, x)$. Plugging (10.5) into (10.3), one has

$$A_\lambda(\mathbf{c}, \mathbf{d}) = \log\left\{\frac{1}{n}\sum_{i=1}^n e^{-\boldsymbol{\phi}_i^T\mathbf{d} - \boldsymbol{\xi}_i^T\mathbf{c}}\right\} + \mathbf{b}_\phi^T\mathbf{d} + \mathbf{b}_\xi^T\mathbf{c} + \frac{\lambda}{2}\mathbf{c}^T Q\mathbf{c}, \tag{10.6}$$

where $\boldsymbol{\phi}_i = \boldsymbol{\phi}(X_i)$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i)$, $\mathbf{b}_\phi = \int_{\mathcal{X}} \boldsymbol{\phi}(x)\rho(x)$, $\mathbf{b}_\xi = \int_{\mathcal{X}} \boldsymbol{\xi}(x)\rho(x)$, and Q is $q \times q$ with the (j,k)th entry $J(\xi_j, \xi_k) = R_J(Z_j, Z_k)$.

For $\mathcal{X} = \prod_\gamma \mathcal{X}_\gamma$ a product domain and $R_J(x,y) = \sum_\beta \theta_\beta R_\beta(x,y)$, $\phi_\nu(x)$ and $R_\beta(Z_j, x)$ are products of functions on the marginal domains, thus one may set $\rho(x) = \prod_\gamma \rho_\gamma(x_{\langle\gamma\rangle})$ and compute the integrals $\mathbf{b}_\phi$ and $\mathbf{b}_\xi$ as (sums of) products of univariate integrals; with such a $\rho(x)$, conditional independence implications of the ANOVA structures in $\eta(x)$ also remain intact. Among good choices of $\rho_\gamma(x_{\langle\gamma\rangle})$ are density estimates on the marginal domains, parametric or nonparametric.

Taking derivatives of $A_\lambda(\mathbf{c},\mathbf{d})$ in (10.6) at $\tilde g = \boldsymbol\phi^T\tilde{\mathbf{d}} + \boldsymbol\xi^T\tilde{\mathbf{c}} \in \mathcal{H}^*$, one has

$$\begin{aligned}
\frac{\partial A_\lambda}{\partial \mathbf{d}} &= -\mu_{\tilde g}(\boldsymbol\phi) + \mathbf{b}_\phi = -\mu_\phi + \mathbf{b}_\phi,\\
\frac{\partial A_\lambda}{\partial \mathbf{c}} &= -\mu_{\tilde g}(\boldsymbol\xi) + \mathbf{b}_\xi + \lambda Q\tilde{\mathbf{c}} = -\mu_\xi + \mathbf{b}_\xi + \lambda Q\tilde{\mathbf{c}},\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d}\partial \mathbf{d}^T} &= V_{\tilde g}(\boldsymbol\phi,\boldsymbol\phi^T) = V_{\phi,\phi},\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{c}\partial \mathbf{c}^T} &= V_{\tilde g}(\boldsymbol\xi,\boldsymbol\xi^T) + \lambda Q = V_{\xi,\xi} + \lambda Q,\\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d}\partial \mathbf{c}^T} &= V_{\tilde g}(\boldsymbol\phi,\boldsymbol\xi^T) = V_{\phi,\xi},
\end{aligned} \tag{10.7}$$

where $\mu_g(f) = \sum_{i=1}^n e^{-g(X_i)}f(X_i)/\sum_{i=1}^n e^{-g(X_i)}$ and $V_g(f,h) = \mu_g(fh) - \mu_g(g)\mu_g(h)$. The Newton updating equation is thus

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi}+\lambda Q \end{pmatrix}\begin{pmatrix}\mathbf{d}\\ \mathbf{c}\end{pmatrix} = \begin{pmatrix}\mu_\phi - \mathbf{b}_\phi + V_{\phi,g}\\ \mu_\xi - \mathbf{b}_\xi + V_{\xi,g}\end{pmatrix}, \tag{10.8}$$

where $V_{\phi,g} = V_{\tilde g}(\boldsymbol\phi,\tilde g)$ and $V_{\xi,g} = V_{\tilde g}(\boldsymbol\xi,\tilde g)$; see Problem 10.3.

10.1.3 Smoothing Parameter Selection

To make (10.1) work in practice, one needs an accompanying method for smoothing parameter selection. Integrals of the form $\int_{\mathcal{X}} h(x)e^{\eta(x)}$ are to be avoided, so the cross-validation of §7.3 does not work here. As an alternative to the Kullback-Leibler distance, consider a loss function

$$\tilde L(\eta,\eta_\lambda) = \int_{\mathcal{X}} \big\{e^{(\eta-\eta_\lambda)(x)} - (\eta-\eta_\lambda)(x) - 1\big\}\rho(x), \tag{10.9}$$

where $e^{\eta(x)}\rho(x) = f(x)$ is the true density and $\eta_\lambda(x)$ is the minimizer of (10.1); note that $e^x - x - 1$ has a unique minimum at $x = 0$. Dropping terms not involving η_λ, one has the relative loss

$$\int_{\mathcal{X}} e^{-\eta_\lambda(x)}f(x) + \int_{\mathcal{X}} \eta_\lambda(x)\rho(x), \tag{10.10}$$

where the first term may be estimated by a cross-validated sample mean, $n^{-1}\sum_{i=1}^n e^{-\eta_\lambda^{[i]}(X_i)}$, for $\eta_\lambda^{[i]}$ minimizing some delete-one version of (10.1).

Write $\eta = d + g = d + \boldsymbol\xi^T\mathbf{c}$ in (10.1) and denote its minimizer by $\eta_\lambda = \tilde\eta = \tilde d + \tilde g = \tilde d + \boldsymbol\xi^T\tilde{\mathbf{c}}$, where in an abuse of notation we merge $(\boldsymbol\phi,\boldsymbol\xi)$, $(\mathbf{d},\mathbf{c})$ and rewrite $\boldsymbol\phi^T\mathbf{d} + \boldsymbol\xi^T\mathbf{c}$ in (10.5) as $\boldsymbol\xi^T\mathbf{c}$. Fixing $\tilde d$, consider the quadratic approximation of (10.1) at $\tilde\eta$ as a function of $\mathbf{c}$,

$$\frac{1}{n}\sum_{i=1}^n w_i\Big\{1 - \boldsymbol\xi_i^T(\mathbf{c}-\tilde{\mathbf{c}}) + \frac{1}{2}(\mathbf{c}-\tilde{\mathbf{c}})^T\boldsymbol\xi_i\boldsymbol\xi_i^T(\mathbf{c}-\tilde{\mathbf{c}})\Big\} + \tilde d + \mathbf{b}^T\mathbf{c} + \frac{\lambda}{2}\mathbf{c}^T\tilde Q\mathbf{c}, \tag{10.11}$$

where $\mathbf{b}^T = (\mathbf{b}_\phi^T, \mathbf{b}_\xi^T)$, $\tilde{Q} = \text{diag}(O, Q)$, and

$$w_i = e^{-\tilde{\eta}(X_i)} = e^{-\tilde{d}-\tilde{g}(X_i)} = ne^{-\tilde{g}(X_i)} / \sum_{l=1}^n e^{-\tilde{g}(X_l)}.$$

The solution of (10.11) is $\tilde{\mathbf{c}}$, with an expression $\tilde{\mathbf{c}} = H^{-1}\mathbf{d}$, where $H = n^{-1}\sum_{i=1}^n w_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T + \lambda\tilde{Q}$ and $\mathbf{d} = n^{-1}\sum_{i=1}^n w_i(1+\tilde{g}_i)\boldsymbol{\xi}_i - \mathbf{b}$ for $\tilde{g}_i = \boldsymbol{\xi}_i^T\tilde{\mathbf{c}} = \tilde{g}(X_i)$. Solving a delete-one version of (10.11),

$$\frac{1}{n}\sum_{j\neq i} w_j \Big\{ 1 - \boldsymbol{\xi}_j^T(\mathbf{c}-\tilde{\mathbf{c}}) + \frac{1}{2}(\mathbf{c}-\tilde{\mathbf{c}})^T\boldsymbol{\xi}_j\boldsymbol{\xi}_j^T(\mathbf{c}-\tilde{\mathbf{c}}) \Big\} + \tilde{d} + \mathbf{b}^T\mathbf{c} + \frac{\lambda}{2}\mathbf{c}^T Q\mathbf{c},$$

one has $\tilde{\mathbf{c}}^{[i]} = \big(H - n^{-1}w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\big)^{-1}\big(\mathbf{d} - n^{-1}w_i(1+\tilde{g}_i)\boldsymbol{\xi}_i\big)$. One may use $\tilde{d} + \tilde{g}_i^{[i]} = \tilde{d} + \boldsymbol{\xi}_i^T\tilde{\mathbf{c}}^{[i]}$ as $\eta_\lambda^{[i]}(X_i)$. Since

$$\big(H - n^{-1}w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\big)^{-1} = H^{-1} + \frac{n^{-1}w_i H^{-1}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T H^{-1}}{1 - n^{-1}w_i\boldsymbol{\xi}_i^T H^{-1}\boldsymbol{\xi}_i},$$

some algebra yields $\tilde{g}_i^{[i]} = \tilde{g}_i - a_i/(1-a_i)$, where $a_i = n^{-1}w_i\boldsymbol{\xi}_i^T H^{-1}\boldsymbol{\xi}_i$; see Problem 10.4. A cross-validation estimate of (10.10) is thus

$$V(\lambda) = \frac{1}{n}\sum_{i=1}^n e^{-\eta_\lambda(X_i)} + \int_{\mathcal{X}} \eta_\lambda(x)\rho(x) + \alpha\frac{1}{n}\sum_{i=1}^n e^{-\eta_\lambda(X_i)}\big(e^{a_i/(1-a_i)} - 1\big) \tag{10.12}$$

for $\alpha = 1$, which is the pseudo likelihood plus an extra term.

As outlined in §3.5.3, the minimization of cross-validation scores typically involves quasi-Newton iteration using starting values from Algorithm 3.3 on page 84. For $V(\lambda)$ in (10.12) to deliver adequate performances, however, one *must* stop at the starting values and forgo the quasi-Newton iteration. As a univariate function of λ for fixed θ_β's, $V(\lambda)$ in (10.12) follows (10.10) reasonably well, but as a multivariate function of θ_β's, it often loses track of its target, yielding poor performances or even outright disasters.

Empirical Performance

Simulations were conducted to explore the empirical performance of cross-validation. On $[0,1]^3$, samples of size $n = 300$ were taken from the test density $f_3(x)$ given in (7.24) on page 246. Note that $(X_1 \perp X_2)|X_3$ here, so the correct model has log density of the form

$$\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_{1,3} + \eta_{2,3}.$$

Using tensor product cubic splines under the correct model specification and setting $q = 100$ in (10.4), three estimates were calculate for each replicate, two with the smoothing parameters λ_v "minimizing" the cross-validation score (10.12) with $\alpha = 1, 1.4$, respectively, and the other with λ_o minimizing the symmetrized Kullback-Leibler distance

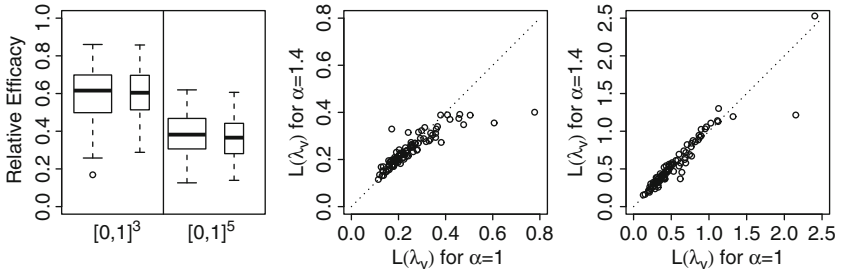


FIGURE 10.1. Effectiveness of cross-validation for density estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0,1]^3$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0,1]^5$.

$$L(\lambda) = \int_{\mathcal{X}} (\eta - \eta_\lambda)(x) f(x) dx + \int_{\mathcal{X}} (\eta_\lambda - \eta)(x) f_\lambda(x) dx,$$

where $f_\lambda \propto e^{\eta_\lambda(x)}\rho(x)$ is the estimated density; despite the use of $\tilde{L}(\eta, \eta_\lambda)$ in (10.9) for the derivation of $V(\lambda)$, we still use the standard symmetrized Kullback-Leibler distance to assess the performance. As noted above, only two passes of fixed-θ minimization were performed to locate λ_v through Algorithm 3.3, but λ_o did minimize $L(\lambda)$ as a multivariate function. The results from one hundred replicates are summarized in Fig. 10.1, with the relative efficacy $L(\lambda_o)/L(\lambda_v)$ shown in the left half of the left frame and the comparison of $\alpha = 1, 1.4$ in $V(\lambda)$ shown in the center frame; $\alpha = 1.4$ is prefered to $\alpha = 1$.

On $[0,1]^5$, consider $(X_2, X_3, X_4)^T \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0.5)\mathbf{1}$ and $\Sigma^{-1} = \begin{pmatrix} 62 & -15 & 0 \\ -15 & 62 & -30 \\ 0 & -30 & 62 \end{pmatrix}$, $X_1 = Y_1 - 0.4X_2 - 0.1$, and $X_5 = Y_2 + 0.3X_4 - 0.1$, then truncate to $\mathcal{X} = [0,1]^5$, where $Y_1, Y_2 \sim f_1(y)$ the normal mixture given in (7.23), independent of $(X_2, X_3, X_4)^T$ and of each other. Note that $(X_i \perp X_j)|(\text{the rest})$ for $(i,j) = (1,3), (1,4), (1,5), (2,4), (2,5), (3,5)$, and the correct model has log density of the form

$$\eta = \eta_\emptyset + \eta_1 + \eta_2 + \eta_3 + \eta_4 + \eta_5 + \eta_{1,2} + \eta_{2,3} + \eta_{3,4} + \eta_{4,5}.$$

Sample of size $n = 600$ were generated and estimates were calculated with $q = 100$. The results from one hundred replicates are shown in Fig. 10.1, with the relative efficacy in the right half of the left frame and the comparison of $\alpha = 1, 1.4$ in the right frame.

Comparison Against Penalized Likelihood

For each of the replicates used in Fig. 10.1, the two cross-validated estimates via (10.1) were recalculated using $q = 10n^{2/9}$ in (10.4), along with the estimate through (7.1) using the same $\xi_j(x) = R_J(Z_j, x)$ and with the default $\alpha = 1.4$ in $V(\lambda)$ of (7.21) on page 245; the quasi-Newton step was

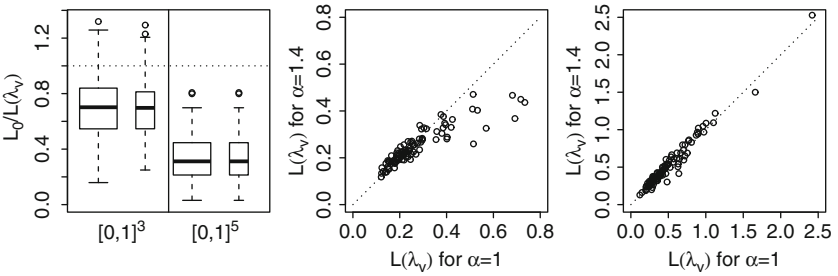


FIGURE 10.2. Comparison against penalized likelihood. *Left*: L_0 via (7.1) over $L(\lambda_v)$ via (10.1), with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0,1]^3$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$ on $[0,1]^5$.

also skipped for the estimates via (7.1) to put things on equal footing. The results are shown in Fig. 10.2, where L_0 are the performances achieved by the estimates via (7.1), and $L(\lambda_v)$ are the performances achieved by the estimates via (10.1), the same as in Fig. 10.1.

For the one hundred replicates on $[0,1]^3$ with $n = 300$ and $q = 36$, estimates via (10.1) with $\alpha = 1.4$ took a total of 62.5 CPU seconds on a linux server, the estimates via (7.1) using a 2,527-point quadrature took 296.4 CPU seconds, and (7.1) with a 3,679-point quadrature took 405.9 CPU seconds. For the one hundred replicates on $[0,1]^5$ with $n = 600$ and $q = 42$, the estimates via (10.1) with $\alpha = 1.4$ took 180.3 CPU seconds, the estimates via (7.1) using a 10,063-point quadrature took 1839.7 CPU seconds, and (7.1) with a 17,103-point quadrature took 3,232.8 CPU seconds.

The computation of (10.1) are $O(nq^2)$, whereas that of (7.1) largely depends on the quadrature size. As the dimension goes up, adequate quadrature sizes become astronomical, rendering (7.1) numerically infeasible.

10.1.4 Square Error Projection

To compute the Kullback-Leibler projection of §7.4.3, one needs integrals of the form $\int_{\mathcal{X}} h(x)e^{\eta(x)}$, which is to be avoided here. As an alternative, one may consider $\tilde{V}(\hat{\eta}-\eta) = \int_{\mathcal{X}}(\hat{\eta}-\eta)^2(x)\rho(x)dx - \big\{\int_{\mathcal{X}}(\hat{\eta}-\eta)(x)\rho(x)dx\big\}^2$ for $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, and calculate the square error projection of $\hat{\eta}$ in $\mathcal{H}_0$ by minimizing $\tilde{V}(\hat{\eta}-\eta)$ over $\eta \in \mathcal{H}_0$; $\tilde{V}(\hat{\eta}-\eta)$ is a proxy of the symmetrized Kullback-Leibler distance (see §10.2), and it is invariant to the normalizing constant.

Let $\tilde{\eta}$ be the square error projection of $\hat{\eta}$ in $\mathcal{H}_0$ and consider $A_{\tilde{\eta},h}(\alpha) = \tilde{V}\big(\hat{\eta} - (\tilde{\eta} + \alpha h)\big)$ for $h \in \mathcal{H}_0$. Since $\dot{A}_{\tilde{\eta},h} = 0$, $\tilde{V}(\hat{\eta} - \tilde{\eta}, h) = 0$, $\forall h \in \mathcal{H}_0$.

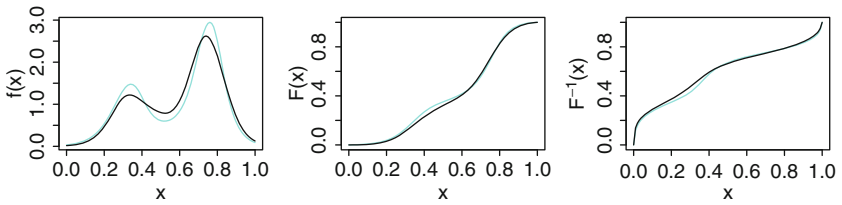


FIGURE 10.3. Density estimation on $\mathcal{X} = [0,1]^3$: Fitted conditional distribution $f(x_{\langle 1\rangle}|x_{\langle 2\rangle} = 0.5, x_{\langle 3\rangle} = 0.5)$. *Left*: Conditional density. *Center*: Conditional cumulative distribution function. *Right*: Quantiles of conditional distribution. Fit via (10.1) is in *solid* and fit via (7.1) in *faded*.

The uniform distribution corresponds to $\eta_u = -\log \rho(x)$, and when $\eta_u \in \mathcal{H}_0$, $\tilde{V}(\hat{\eta} - \tilde{\eta}, \tilde{\eta} - \eta_u) = 0$, so $\tilde{V}(\hat{\eta} - \eta_u) = \tilde{V}(\hat{\eta} - \tilde{\eta}) + \tilde{V}(\tilde{\eta} - \eta_u)$. When the ratio $\tilde{V}(\hat{\eta} - \tilde{\eta})/\tilde{V}(\hat{\eta} - \eta_u)$ is small, one may safely cut out $\mathcal{H}_1$.

With $\rho(x) = \prod_\gamma \rho_\gamma(x_{\langle\gamma\rangle})$, the calculations involved are sums of products of univariate integrals, and $\eta_u \in \mathcal{H}_0$ when $\mathcal{H}_0$ includes all the main effects.

10.1.5 R Package `gss`: `ssden1` Suite

Density estimation via penalized pseudo likelihood is implemented in the `ssden1` suite. The following sequence generates a sample from $f_3(x)$ given in (7.24) on page 246 and fits a tensor product cubic spline to the log density, where `rtest3` is listed in Example 7.4 on page 251:

```
set.seed(5732); x <- rtest3(300)
x1 <- x[,1]; x2 <- x[,2]; x3 <- x[,3]; rg <- c(0,1)
domain <- data.frame(x1=rg,x2=rg,x3=rg)
fit <- ssden1(~x1*x2*x3,domain=domain)
```

Three marginal densities $\rho_\gamma(x_{\langle\gamma\rangle})$, $\gamma = 1, 2, 3$ are estimated internally via (7.1) to form $\rho(x) = \prod_{\gamma=1}^3 \rho_\gamma(x_{\langle\gamma\rangle})$. The square error projection suggests the elimination of the terms `x1:x2` and `x1:x2:x3`, and one may refit without these terms; only interactions need to be listed in `project`, as all main effects are automatically included internally:

```
project(fit,c("x1:x3","x2:x3"))$ratio
# 0.02169527
fit <- ssden1(~(x1+x2)*x3,domain=domain)
```

The utility functions `dssden`, `cdssden`, `cpssden`, and `cqssden` are shared by `ssden` and `ssden1`, though the results from `dssden` are unnormalized for `ssden1` fits. The conditional distribution $f(x_{\langle 1\rangle}|x_{\langle 2\rangle} = .5, x_{\langle 3\rangle} = .5)$ based on the `ssden1` fit is shown in Fig. 10.3 in solid lines, with that based on the `ssden` fit seen in Fig. 7.4 superimposed in faded lines.

With all interactions included, a `ssden` fit took 22.29 CPU seconds on a linux laptop and a `ssden1` fit took 0.63 CPU seconds. With only `x1:x3` and `x2:x3` included, a `ssden` fit took 8.67 CPU seconds and a `ssden1` fit took 0.5 CPU seconds.

10.1.6 Case Study: Transcription Factor Association

Gene expression is largely regulated by transcription mechanisms, in which transcription factors bind to DNA segments in the promoter regions of the target genes to turn on or shut off gene expression. Some transcription factor association strength scores, normalized to be between 0 and 5.132242, were compiled by Ouyang, Zhou, and Wong (2009) for 12 transcription factors and 18,936 genes, with a higher score indicating the proximity of the gene to the binding sites of the transcription factor along the genome. The data are available at

```
http://www.pnas.org/content/suppl/2009/12/04
                    /0904863106.DCSupplemental/SD2.txt
```

and one may read the data into R as a data frame:

```
SD2<-read.table("SD2.txt",header=TRUE); SD2<-SD2[,-(1:2)]
```

with elements `E2f1`, `Mycn`, `Zfx`, `Myc`, `Klf4`, `Tcfcp2l1`, `Esrrb`, `Nanog`, `Oct4`, `Sox2`, `Stat3`, and `Smad1`, which are the 12 transcription factors.

A log density involving all main effects and two-way interactions was fitted to `SD2`:

```
mn <- apply(SD2,2,min); mx <- apply(SD2,2,max)
domain <- data.frame(rbind(mn,mx)); set.seed(5732)
fit.sd2<-ssden1(~(E2f1+Mycn+Zfx+Myc+Klf4+Tcfcp2l1
                  +Esrrb+Nanog+Oct4+Sox2+Stat3+Smad1)^2,
                domain=domain,data=SD2)
```

where `domain` specifies the domain $\mathcal{X} = [0, 5.132242]^{12}$ used in (10.1). To check how irreplaceable each interaction is, one may try:

```
lab.sd2 <- fit.sd2$terms$labels[-(1:12)]
r.sd2 <- project(fit.sd2,lab.sd2,drop1=TRUE)$ratio
```

where `lab.sd2` collects the $\binom{12}{2} = 66$ interaction terms and `drop1=TRUE` in the call to `project` orders 66 "drop-one-term" projections, with `r.sd2` containing 66 "drop-one-term" $\tilde{V}(\hat{\eta} - \tilde{\eta})/\tilde{V}(\hat{\eta} - \eta_u)$ ratios labeled by the dropped terms; these may be perceived as the "strengths" of the terms. Projecting into the space with all the main effects plus the top six interactions, one has $\tilde{V}(\hat{\eta} - \tilde{\eta})/\tilde{V}(\hat{\eta} - \eta_u) = 2.92\,\%$:

```
project(fit.sd2,lab.sd2[rev(order(r.sd2))[1:6]])
# 0.0292398
```

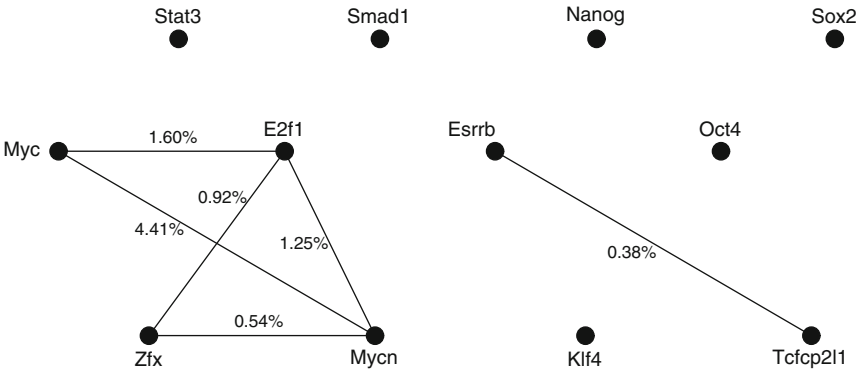


FIGURE 10.4. Stronger interactions in the `SD2` fit. The labels on the edges indicate the "strengths" of the interactions.

A graph illustrating the six terms in `lab.sd2[rev(order(r.sd2))[1:6]]` is shown in Fig. 10.4. Apart from the first six terms, the rest of the terms all have "strengths" no better than 0.21 %, thus are individually dispensable. The overall "weakness" of the interactions suggests weak correlations among the variables.

The transcription factors `E2f1`, `Mycn`, `Zfx`, and `Myc` seem to work in concert, so do `Tcfcp2l1` and `Esrrb` but to a lesser extent; the rest of the field appear to act independently.

Due to the huge sample size and the large number of terms, one needs ample RAM to run the analysis. On a linux server with 32 Gb RAM (though 16 should be sufficient), `fit.sd2` took around 25 min to obtain, and `project(...,drop1=TRUE)` took about 20 min to execute.

10.2 Density Estimation: Asymptotic Convergence

The analysis of §9.2 can be adapted to study the asymptotic convergence of the density estimates via (10.1). Let $f_0(x) = e^{\eta_0(x)}\rho(x)$ be the density to be estimated satisfying $\int_{\mathcal{X}} e^{\eta_0(x)}\rho(x) = 1$ and $\hat{\eta}(x)$ be the minimizer of (10.1); in general, $\int_{\mathcal{X}} e^{\hat{\eta}(x)}\rho(x) \neq 1$. Define $V(f) = \int_{\mathcal{X}} f^2(x)\rho(x)$. Convergence rates in this section are in terms of $V(\eta - \eta_0)$.

Consider $\tilde{V}(f) = \int_{\mathcal{X}}\{f(x) - \int_{\mathcal{X}} f(x)\rho(x)\}^2\rho(x) < V(f)$; $\tilde{V}(\eta - \eta_0)$ is invariant to the normalizing constant, and rates in $V(\eta - \eta_0)$ imply rates in $\tilde{V}(\eta - \eta_0)$. For a density $f(x) \propto e^{\eta(x)}\rho(x)$, the symmetrized Kullback-Leibler distance between f_0 and f is seen to be

$$\mathrm{SKL}(\eta_0, \eta) = \int_{\mathcal{X}} \{(\eta - \eta_0)(x) - \int_{\mathcal{X}} (\eta - \eta_0)(x)\tilde{f}(x)\}^2 \tilde{f}(x), \qquad (10.13)$$

where $\tilde{f}(x) \propto e^{\tilde{\eta}(x)}\rho(x)$ for $\tilde{\eta}$ a convex combination of η and η_0; see Problem 10.5. $\tilde{V}(\eta - \eta_0)$ can thus be viewed as a proxy of $\mathrm{SKL}(\eta_0, \eta)$.

For comparison, convergence rates of the estimates via (7.1) are in terms of $\int_{\mathcal{X}} \left\{(\eta - \eta_0)(x) - \int_{\mathcal{Y}} (\eta - \eta_0)(x) f_0(x)\right\}^2 f_0(x)$, as seen in §9.2.

10.2.1 Linear Approximation

As in §9.2, the following conditions are needed for the analysis.

Condition 10.2.1 V is completely continuous with respect to J.

Condition 10.2.2 For ν sufficiently large and some $\beta > 0$, the eigenvalues ρ_ν of J with respect to V satisfy $\rho_\nu > \beta\nu^r$, where $r > 1$.

Consider the quadratic functional

$$\frac{1}{n}\sum_{i=1}^{n} -e^{-\eta_0(X_i)}\eta(X_i) + \int_{\mathcal{X}} \eta(x)\rho(x) + \frac{1}{2}V(\eta - \eta_0) + \frac{\lambda}{2}J(\eta). \tag{10.14}$$

Plugging the Fourier series expansions $\eta = \sum_\nu \eta_\nu \phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ into (10.14), its minimizer $\tilde{\eta}$ has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu),$$

where $\beta_\nu = n^{-1}\sum_{i=1}^n \left\{e^{-\eta_0(X_i)}\phi_\nu(X_i) - \int_{\mathcal{X}} \phi_\nu(x)\rho(x)\right\}$. It is easy to verify that $E[\beta_\nu] = 0$ and $E[\beta_\nu^2] \le n^{-1}\int_{\mathcal{X}} \phi_\nu^2(x)e^{-\eta_0(x)}\rho(x)$.

Condition 10.2.3 For some $c_3 < \infty$, $e^{-\eta_0(x)} < c_3$.

Under Condition 10.2.3, $E[\beta_\nu^2] \le c_3/n$, noting that $V(\phi_\nu) = 1$.

Theorem 10.1 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 10.2.1–10.2.3, as* $n \to \infty$ *and* $\lambda \to 0$,

$$(V + \lambda J)(\tilde{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: See the proof of Theorem 9.2. □

10.2.2 Approximation Error and Main Results

We now turn to the approximation error $\hat{\eta} - \tilde{\eta}$. Define

$$A_{f,g}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} e^{-(f+\alpha g)(X_i)} + \int_{\mathcal{X}} (f + \alpha g)(x)\rho(x) + \frac{\lambda}{2}J(f + \alpha g),$$

$$B_{f,g}(\alpha) = \frac{1}{n}\sum_{i=1}^{n} -e^{-\eta_0(X_i)}(f + \alpha g)(X_i) + \int_{\mathcal{X}} (f + \alpha g)(x)\rho(x)$$
$$+ \frac{1}{2}V(f + \alpha g - \eta_0) + \frac{\lambda}{2}J(f + \alpha g).$$

It is easy to verify that

$$\dot{A}_{f,g}(0) = \frac{1}{n}\sum_{i=1}^{n} -e^{-f(X_i)}g(X_i) + \int_{\mathcal{X}} g(x)\rho(x) + \lambda J(f,g), \tag{10.15}$$

$$\dot{B}_{f,g}(0) = \frac{1}{n}\sum_{i=1}^{n} -e^{-\eta_0(X_i)}g(X_i) + \int_{\mathcal{X}} g(x)\rho(x) + V(f-\eta_0, g) + \lambda J(f,g). \tag{10.16}$$

Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (10.15), one has

$$\frac{1}{n}\sum_{i=1}^{n} -e^{-\hat{\eta}(X_i)}(\hat{\eta}-\tilde{\eta})(X_i) + \int_{\mathcal{X}} (\hat{\eta}-\tilde{\eta})(x)\rho(x) + \lambda J(\hat{\eta}, \hat{\eta}-\tilde{\eta}) = 0, \tag{10.17}$$

and setting $f = \tilde{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (10.16) yields

$$\frac{1}{n}\sum_{i=1}^{n} -e^{-\eta_0(X_i)}(\hat{\eta}-\tilde{\eta})(X_i) + \int_{\mathcal{X}} (\hat{\eta}-\tilde{\eta})(x)\rho(x) + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \lambda J(\tilde{\eta}, \hat{\eta}-\tilde{\eta}) = 0. \tag{10.18}$$

Subtracting (10.18) from (10.17), one has

$$\begin{aligned}\lambda J(\hat{\eta}-\tilde{\eta}) &- \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\hat{\eta}(X_i)} - e^{-\tilde{\eta}(X_i)}\right\}(\hat{\eta}-\tilde{\eta})(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\tilde{\eta}(X_i)} - e^{-\eta_0(X_i)}\right\}(\hat{\eta}-\tilde{\eta})(U_i) + V(\hat{\eta}-\tilde{\eta}, \tilde{\eta}-\eta_0). \end{aligned} \tag{10.19}$$

Condition 10.2.4 For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$, $c_1 < e^{\eta_0(x)-\eta(x)} < c_2$ holds uniformly for some $0 < c_1 < c_2 < \infty$.

Condition 10.2.5 $\int_{\mathcal{X}} \phi_\nu^2(x)\phi_\mu^2(x)e^{-\eta_0(x)}\rho(x) < c_4$ for some $c_4 < \infty$, $\forall \nu, \mu$.

Under Condition 10.2.4, by the mean value theorem, one has

$$\frac{c_1}{n}\sum_{i=1}^{n} e^{-\eta_0(X_i)}(\hat{\eta}-\tilde{\eta})^2(X_i) \le -\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\hat{\eta}(X_i)} - e^{-\tilde{\eta}(X_i)}\right\}(\hat{\eta}-\tilde{\eta})(X_i), \tag{10.20}$$

and for some $c \in (c_1, c_2)$,

$$\begin{aligned} -\frac{c}{n}\sum_{i=1}^{n} e^{-\eta_0(X_i)}&(\hat{\eta}-\tilde{\eta})(X_i)(\tilde{\eta}-\eta_0)(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\tilde{\eta}(X_i)} - e^{-\eta_0(X_i)}\right\}(\hat{\eta}-\tilde{\eta})(X_i). \end{aligned} \tag{10.21}$$

Under Condition 10.2.5, parallel to Lemma 9.16 on page 344, one has

$$\Big|\frac{1}{n}\sum_{i=1}^{n} e^{-\eta_0(X_i)} g(X_i)h(X_i) - V(g,h)\Big| = O_p\big(n^{-1/2}\lambda^{-1/r}\big)\big\{(V+\lambda J)(g)(V+\lambda J)(h)\big\}^{1/2}; \quad (10.22)$$

see Problem 10.6. Substituting (10.20)–(10.22) into (10.19), some manipulations yield, as $\lambda \to 0$ and $n\lambda^{2/r} \to \infty$,

$$(c_1 V + \lambda J)(\hat{\eta} - \tilde{\eta}) \le \big(|1-c| + o_p(1)\big)\big\{(V+\lambda J)(\hat{\eta}-\tilde{\eta})(V+\lambda J)(\tilde{\eta}-\eta_0)\big\}^{1/2},$$

which, in combination with Theorem 10.1, leads to the following theorem.

Theorem 10.2 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 10.2.1–10.2.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$(V+\lambda J)(\hat{\eta} - \eta_0) = O_p(n^{-1}\lambda^{-1/r} + \lambda^p).$$

10.2.3 Efficient Approximation

Now consider the minimizer $\hat{\eta}^*$ of (10.1) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(Z_j, \cdot), j = 1, \ldots, q\},$$

where $\{Z_j\}$ is a random subset of $\{X_i\}$.

Lemma 10.3 *Under Conditions 10.2.1–10.2.3 and 10.2.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$, $V(h) = o_p\big(\lambda J(h)\big)$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$.

Proof: For $h \in \mathcal{H} \ominus \mathcal{H}^*$, $h(Z_j) = J\big(R_J(Z_j, \cdot), h\big) = 0$. Similar to (10.22),

$$V(h) = \Big|V(h) - \frac{1}{q}\sum_{j=1}^{q} e^{-\eta_0(Z_j)} h^2(Z_j)\Big| = O_p\big(q^{-1/2}\lambda^{-1/r}\big)(V+\lambda J)(h).$$

The lemma follows. □

Let η^* be the projection of $\hat{\eta}$ in $\mathcal{H}^*$; $J(\eta^*, \hat{\eta} - \eta^*) = 0$. The convex set B_0 in Condition 10.2.4 should also contain $\hat{\eta}^*$ and η^*. Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \eta^*$ in (10.15), one has

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{\eta}(X_i)}(\hat{\eta} - \eta^*)(X_i) + \int_{\mathcal{X}} (\hat{\eta} - \eta^*)(x)\rho(x) + \lambda J(\hat{\eta}, \hat{\eta} - \eta^*) = 0,$$

which can be rearranged as

$$\begin{aligned}\lambda J(\hat{\eta} - \eta^*) =& \frac{1}{n}\sum_i \big\{e^{-\hat{\eta}(X_i)} - e^{-\eta_0(X_i)}\big\}(\hat{\eta} - \eta^*)(X_i) \\ &+ \frac{1}{n}\sum_i e^{-\eta_0(X_i)}(\hat{\eta} - \eta^*)(X_i) - \int_{\mathcal{X}} (\hat{\eta} - \eta^*)(x)\rho(x). \quad (10.23)\end{aligned}$$

The first term on the right-hand side of (10.23) is $\big(c+o_p(1)\big)V(\eta_0-\hat\eta,\hat\eta-\eta^*)$ for some c by (10.21) and (10.22); parallel to (9.16) on page 328, the second term is of the order $O_p(n^{-1/2}\lambda^{-1/2r})\big\{(V+\lambda J)(\hat\eta-\eta^*)\big\}^{1/2}$. Combining these with Lemma 10.3 and Theorem 10.2, one has the following theorem.

Theorem 10.4 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 10.2.1–10.2.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V+\lambda J)(\hat\eta-\eta^*) = O_p(n^{-1}\lambda^{-1/r}+\lambda^p).$$

Setting $f=\hat\eta$ and $g=\hat\eta-\hat\eta^*$ in (10.15), one has

$$-\frac{1}{n}\sum_{i=1}^n e^{-\hat\eta(X_i)}(\hat\eta-\hat\eta^*)(X_i)+\int_{\mathcal{X}}(\hat\eta-\hat\eta^*)(x)\rho(x)+\lambda J(\hat\eta-\hat\eta^*,\hat\eta)=0, \quad (10.24)$$

and setting $f=\hat\eta^*$ and $g=\hat\eta^*-\eta^*$ in (10.15) leads to

$$-\frac{1}{n}\sum_{i=1}^n e^{-\hat\eta^*(X_i)}(\hat\eta^*-\eta^*)(X_i)+\int_{\mathcal{X}}(\hat\eta^*-\eta^*)(x)\rho(x) + \lambda J(\hat\eta^*-\eta^*,\hat\eta^*)=0. \quad (10.25)$$

Adding (10.24), (10.25) and subtracting (10.23), some algebra yields

$$\lambda J(\hat\eta^*-\eta^*)-\frac{1}{n}\sum_{i=1}^n\big\{e^{-\hat\eta^*(X_i)}-e^{-\eta^*(X_i)}\big\}(\hat\eta^*-\eta^*)(X_i) = -\frac{1}{n}\sum_{i=1}^n\big\{e^{-\hat\eta(X_i)}-e^{-\eta^*(X_i)}\big\}(\hat\eta^*-\eta^*)(X_i); \quad (10.26)$$

noting that $J(\hat\eta^*-\eta^*,\hat\eta-\eta^*)=0$. By Condition 10.2.4 and (10.22), the left-hand side of (10.26) is no less than $\big(c_1+o_p(1)\big)V(\hat\eta^*-\eta^*)+\lambda J(\hat\eta^*-\eta^*)$, and the right-hand side is $\big(c+o_p(1)\big)V(\hat\eta-\eta^*,\hat\eta^*-\eta^*)$. These, in combination with Theorems 10.2 and 10.4, lead to the following theorem.

Theorem 10.5 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 10.2.1–10.2.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V+\lambda J)(\hat\eta^*-\eta_0) = O_p(n^{-1}\lambda^{-1/r}+\lambda^p).$$

10.3 Conditional Density Estimation

As an alternative to (7.30) of §7.7, one may estimate the conditional density $f(y|x) \propto e^{\eta(x,y)}\rho(x,y)$ via the minimization of

$$\frac{1}{n}\sum_{i=1}^n\left\{e^{-\eta(X_i,Y_i)}+\int_{\mathcal{Y}}\eta(X_i,y)\rho(X_i,y)\right\}+\frac{\lambda}{2}J(\eta), \quad (10.27)$$

where $\rho(x, y)$ is some known conditional density on the domain $\mathcal{X} \times \mathcal{Y}$ satisfying $\int_{\mathcal{Y}} \rho(x, y) = 1$, $\forall x \in \mathcal{X}$. This works similar to (10.1), and the integral $\int_{\mathcal{Y}} \eta(X_i, y)\rho(X_i, y)$ can largely be pre-computed.

With an ANOVA decomposition $\eta = \eta_\emptyset + \eta_x + \eta_y + \eta_{x,y}$, $\eta_\emptyset + \eta_x$ does not cancel out in (10.27), contrasting (7.30), though it does in the estimate $f(y|x) = e^{\eta(x,y)}\rho(x, y)/\int_{\mathcal{Y}} e^{\eta(x,y)}\rho(x, y)$. It is necessary to include $\eta_\emptyset + \eta_x$ in η for (10.27) to work; see §10.3.6.

A good portion of the developments here nearly duplicate those in §10.1, for which the discussions will be brief; these include the existence and the computation of the minimizer of (10.27), the cross-validation score for smoothing parameter selection, and the square error projection. Software tools are illustrated via simulated and real-data examples, and comparisons are made against the penalized likelihood estimates through (7.30).

The asymptotic analysis of §10.2 applies to the minimizer of (10.27) with trivial modifications (§10.3.6).

10.3.1 Preliminaries

One shall minimize (10.27) in a reproducing kernel Hilbert space $\mathcal{H}$ on $\mathcal{X} \times \mathcal{Y}$ with a square (semi) norm $J(f)$. Write

$$L(f) = n^{-1} \sum_{i=1}^{n} \left\{ e^{-f(X_i,Y_i)} + \int_{\mathcal{Y}} f(X_i, y)\rho(X_i, y) \right\}.$$

It is easy to verify that $L(f)$ is continuous, convex, and Fréchet differentiable. Let $\{\phi_\nu\}_{\nu=1}^m$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$ and S be an $n \times m$ matrix with the (i, ν)th entry $\phi_\nu(X_i, Y_i)$. The minimizer of (10.27) uniquely exists when S is of full column rank, which we assume.

When $J(f)$ annihilates constant, the minimizer of (10.27) satisfies

$$\frac{1}{n} \sum_{i=1}^{n} e^{-\eta(X_i,Y_i)} = \frac{1}{n} \sum_{i=1}^{n} e^{-d-g(X_i,Y_i)} = 1,$$

where $g \in \mathcal{G} = \mathcal{H} \ominus \{1\}$ minimizes a "profile" functional parallel to (10.3),

$$\log \left\{ \frac{1}{n} \sum_{i=1}^{n} e^{-g(X_i,Y_i)} \right\} + \frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{Y}} g(X_i, y)\rho(X_i, y) + \frac{\lambda}{2} J(g). \tag{10.28}$$

Without loss of inferential efficiency, one may minimize (10.27) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(V_j, \cdot), j = 1, \ldots, q\}, \tag{10.29}$$

where $\{V_j\}$ is a random subset of $\{(X_i, Y_i)\}$. One has, for $u = (x, y)$,

$$g(u) = \sum_{\nu} d_\nu \phi_\nu(u) + \sum_{j} c_j R_J(V_j, u) = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{10.30}$$

where $\{\phi_\nu\}$ is a basis of $\mathcal{N}_J \ominus \{1\}$ and $\xi_j(u) = R_J(V_j, u)$. Plugging (10.30) into (10.28), one has

$$A_\lambda(\mathbf{c}, \mathbf{d}) = \log\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-\boldsymbol{\phi}_i^T\mathbf{d} - \boldsymbol{\xi}_i^T\mathbf{c}}\right\} + \mathbf{b}_\phi^T\mathbf{d} + \mathbf{b}_\xi^T\mathbf{c} + \frac{\lambda}{2}\mathbf{c}^T Q\mathbf{c}, \tag{10.31}$$

where $\boldsymbol{\phi}_i = \boldsymbol{\phi}(X_i, Y_i)$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i, Y_i)$, $\mathbf{b}_\phi = n^{-1}\sum_{i=1}^n \int_{\mathcal{Y}} \boldsymbol{\phi}(X_i, y)\rho(X_i, y)$, $\mathbf{b}_\xi = n^{-1}\sum_{i=1}^n \int_{\mathcal{Y}} \boldsymbol{\xi}(X_i, y)\rho(X_i, y)$, and Q is $q \times q$ with the (j,k)th entry $J(\xi_j, \xi_k) = R_J(V_j, V_k)$; (10.31) appears as a carbon copy of (10.6), and (10.7), (10.8) hold verbatim but with a modified definition of

$$\mu_g(f) = \textstyle\sum_{i=1}^n e^{-g(X_i,Y_i)} f(X_i, Y_i) / \sum_{i=1}^n e^{-g(X_i,Y_i)}.$$

Note that the integrals $\mathbf{b}_\phi$, $\mathbf{b}_\xi$ can be computed once for all, which is the key to the numerical efficiency of (10.27).

The $\rho(x, y)$ function is an important part of (10.27). A simple choice is to set $\rho(x, y) = e^{\eta(y)} / \int_{\mathcal{Y}} e^{\eta(y)}$, an estimate of the marginal density on $\mathcal{Y}$. Alternatively, one may pretend $Y \sim N\big(\mu(x), \sigma^2\big)$, for Y on $[a, b]$, estimate $\mu(x)$ and σ^2 using the techniques of Chap. 3, then sets

$$\rho(x, y) = \frac{\phi\big((y - \mu(x))/\sigma\big)}{\Phi\big((b - \mu(x))/\sigma\big) - \Phi\big((a - \mu(x))/\sigma\big)}, \tag{10.32}$$

where $\phi(x)$ is the standard normal density and $\Phi(x)$ is the distribution function. For $\mathcal{Y} = \prod_\gamma [a_\gamma, b_\gamma]$, one may use (10.32) on marginal domains and take their product as $\rho(x, y)$.

10.3.2 Smoothing Parameter Selection

To devise a cross-validation scheme for smoothing parameter selection with the minimizer η_λ of (10.27), consider a loss function

$$\tilde{L}(\eta, \eta_\lambda) = \int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} \big\{e^{(\eta - \eta_\lambda)(x,y)} - (\eta - \eta_\lambda)(x, y) - 1\big\}\rho(x, y), \tag{10.33}$$

where $f(y|x) = e^{\eta(x,y)}\rho(x, y)$ and $f(x)$ is the limiting density of X_i. Dropping terms not involving η_λ, one has the relative loss

$$\int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} e^{-\eta_\lambda(x,y)} f(y|x) + \int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} \eta_\lambda(x, y)\rho(x, y). \tag{10.34}$$

The second term in (10.34) can be substituted by its empirical version $n^{-1}\sum_{i=1}^n \int_{\mathcal{Y}} \eta_\lambda(X_i, y)\rho(X_i, y)$, and the first term, $E\big[e^{-\eta_\lambda(X,Y)}\big]$, may be estimated by a cross-validated sample mean, $n^{-1}\sum_{i=1}^n e^{\eta_\lambda^{[i]}(X_i,Y_i)}$, where

$\eta_\lambda^{[i]}$ minimizes some delete-one version of (10.27). The derivation leading to (10.12) then yields a cross-validation estimate of (10.34),

$$V(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\eta_\lambda(X_i,Y_i)} + \int_{\mathcal{Y}} \eta_\lambda(X_i,y)\rho(X_i,y)\right\} + \alpha\frac{1}{n}\sum_{i=1}^{n} e^{-\eta_\lambda(X_i,Y_i)}\big(e^{a_i/(1-a_i)} - 1\big), \quad (10.35)$$

for $\alpha = 1$, where $a_i = n^{-1}w_i\boldsymbol{\xi}_i^T H^{-1}\boldsymbol{\xi}_i$ as in (10.12) but with the modified $w_i = ne^{-\tilde{g}(X_i,Y_i)}/\sum_{l=1}^{n} e^{-\tilde{g}(X_l,Y_l)}$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i,Y_i)$, and in turn $H = n^{-1}\sum_{i=1}^{n} w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \lambda\tilde{Q}$. Unlike $V(\lambda)$ in (10.12), however, one does not have to stop at the starting values when minimizing $V(\lambda)$ in (10.35).

Empirical Performance

Recall the simulations of §7.7.2 on $\mathcal{X} = [0,1]$ and $\mathcal{Y} = [0,1]$; the test distribution is as given in (7.34), $f(y|x) \propto \phi\big((y-\mu_x)/\sigma_x\big)I_{[0<y<1]}$, with $\mu_x = x^3 - x^2 + x - 0.2$ and three versions of σ_x: $\sigma_1 = 0.3$, $\sigma_2 = 0.15(1+x)$, and $\sigma_3 = 0.15(2-x)$. Samples of size $n = 200$ were drawn with X_i on the grid 0.005(0.01)0.995, two each.

Tensor product cubic splines were calculated as minimizers of (10.27) with two versions of $\rho(x,y)$, $\rho_1(x,y)$ a penalized likelihood estimate of the marginal density $f(y)$ and $\rho_2(x,y)$ as specified in (10.32). Estimates were obtained with the smoothing parameters minimizing the symmetrized Kullback-Leibler distance

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{Y}}\left\{\log\frac{f(y|X_i)}{f_\lambda(y|X_i)}f(y|X_i) + \log\frac{f_\lambda(y|X_i)}{f(y|X_i)}f_\lambda(y|X_i)\right\}$$

at λ_o, and minimizing $V(\lambda)$ of (10.35) at λ_v with $\alpha = 1, 1.4$. The same set $\{V_j\}$ of size $q = 33 \approx 10(200)^{2/9}$ were used in (10.29) for all the estimates based on the same sample.

One hundred replicates were drawn with each of the three σ_x and the simulation results are summarized in Fig. 10.5, parallel to Fig. 7.11 on page 266. Despite the use of $\tilde{L}(\eta,\eta_\lambda)$ in (10.33) for the derivation of (10.35), performance is measured by the same $L(\lambda)$ used in §7.7.2. It appears that $\rho_2(x,y)$ works much better in the simulation settings and $\alpha = 1.4$ is generally preferred to $\alpha = 1$.

Comparison Against Penalized Likelihood

Penalized likelihood estimates via (7.30) were also calculated for the replicates, for smoothing parameters minimizing the cross-validation score in (7.21), duly modified for use with (7.30), with the default $\alpha = 1.4$.

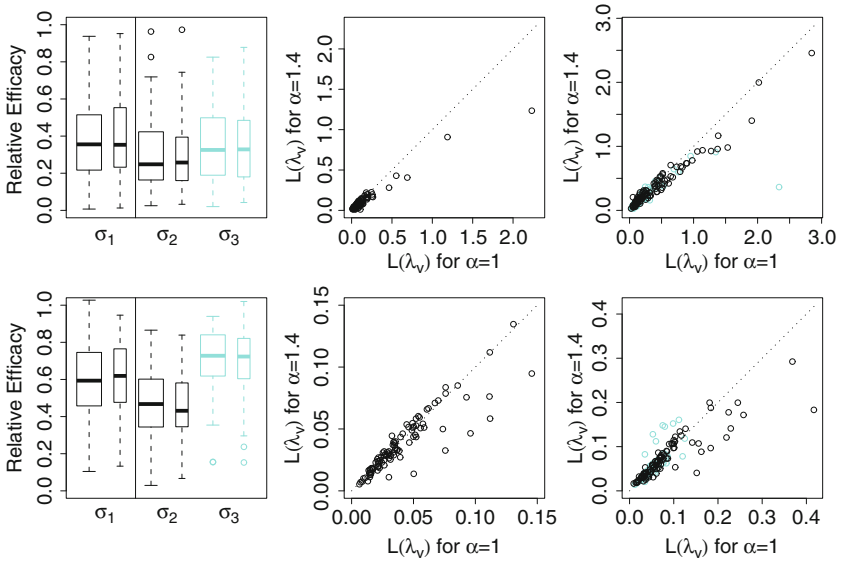


FIGURE 10.5. Effectiveness of cross-validation for conditional density estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*); $\sigma_1 = 0.3$, $\sigma_2 = 0.15(1 + x)$, $\sigma_3 = 0.15(2 - x)$. *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\sigma_x = 0.3$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\sigma_x = 0.15(1 + x)$ (*solid*) and $\sigma_x = 0.15(2 - x)$ (*faded*). The *top row* corresponds to $\rho_1(x, y)$ and the *bottom row* to $\rho_2(x, y)$.

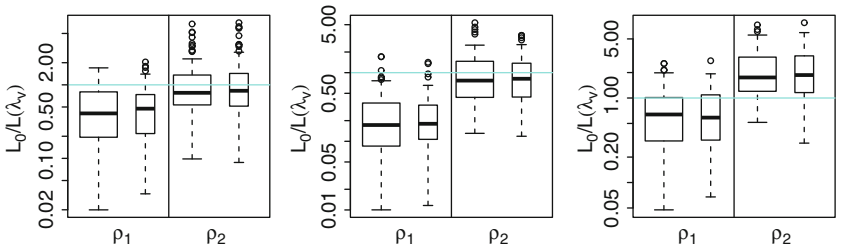


FIGURE 10.6. Performance comparisons of penalized likelihood of (7.30) versus penalized pseudo likelihood of (10.27). L_0 achieved by (7.30) with $\alpha = 1.4$ in (7.21) over $L(\lambda_v)$ achieved by (10.27). From *left to right*: $\sigma_x = 0.3, 0.15(1 + x), 0.15(2 - x)$. *Wider boxes* correspond to $\alpha = 1$ in (10.35) and *thinner boxes* to $\alpha = 1.4$. The *faded lines* mark equal performance.

Performance comparisons of the penalized likelihood of (7.30) versus the penalized pseudo likelihood of (10.27) are shown in Fig. 10.6. It is a bit surprising that for the test distribution with $\sigma_x = 0.15(2 - x)$, (10.27) with $\rho_2(x, y)$ actually outperforms (7.30).

The one hundred estimates via (7.30) in the center frame of Fig. 10.6 took about 5,400 CPU seconds to compute on a linux server, and those through (10.27) with $\rho_2(x, y)$ and $\alpha = 1.4$ took about 280 CPU seconds.

10.3.3 Square Error Projection

The Kullback-Leibler projection of §7.7.3 involves integrals of the form $\int_{\mathcal{Y}} h(X_i, y) e^{\eta(X_i, y)}$ as with (7.30), which we strive to avoid here. As an alternative, one may consider a square error

$$\tilde{V}(\hat{\eta} - \eta) = \int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} \left\{ (\hat{\eta} - \eta)(x, y) - \int_{\mathcal{Y}} (\hat{\eta} - \eta)(x, y) \rho(x, y) \right\}^2 \rho(x, y)$$

for $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, and calculate the square error projection of $\hat{\eta}$ in $\mathcal{H}_0$ by minimizing $\tilde{V}(\hat{\eta} - \eta)$ over $\eta \in \mathcal{H}_0$; $\tilde{V}(\hat{\eta} - \eta)$ is a proxy of the symmetrized Kullback-Leibler distance, and is invariant to the normalizing constants.

Let $\tilde{\eta}$ be the square error projection of $\hat{\eta}$ in $\mathcal{H}_0$. One has $\tilde{V}(\hat{\eta} - \tilde{\eta}, h) = 0$, $\forall h \in \mathcal{H}_0$. The uniform conditional density corresponds to $\eta_u = -\log \rho(x, y)$, and when $\eta_u \in \mathcal{H}_0$, $\tilde{V}(\hat{\eta} - \eta_u) = \tilde{V}(\hat{\eta} - \tilde{\eta}) + \tilde{V}(\tilde{\eta} - \eta_u)$.

While the projection tool is easy to derive, model selection is more involveds here. The conditional density is of the form $f(y|x) \propto e^{\eta(x,y)} \rho(x, y)$, and ANOVA structures in $\eta(x, y)$ may not have conditional independence implications when $\rho(x, y)$ gets in the way. A $\rho(x, y)$ constant along $\mathcal{X}$ is less intruding, but it could perform poorly as seen in the simulations of §10.3.2.

10.3.4 R Package `gss`: `sscden1` Suite

The `sscden1` suite in `gss` implements conditional density estimation via the minimization of (10.27). The following sequence draws a sample from (7.34) on page 265 with $\sigma_x = 0.15(2 - x)$ and calculates a cross-validated estimate with $\rho(x, y)$ given by (10.32):

```
rfc3 <- function(x) {
  mu <- x^3-x^2+x-.2; sd=.15*(2-x)
  y <- (rnorm(length(x))*sd+mu)
  ok <- (y>0)&(y<1)
  while(m <- sum(!ok)) {
    y[!ok] <- (rnorm(m)*sd[!ok]+mu[!ok])
    ok <- (y>0)&(y<1)
  }
  y
}
xx <- ((1:100)-.5)/100; x <- rep(xx,2)
set.seed(5732); y <- rfc3(x)
fit <- sscden1(~x*y,~y,ydomain=data.frame(y=c(0,1)))
```

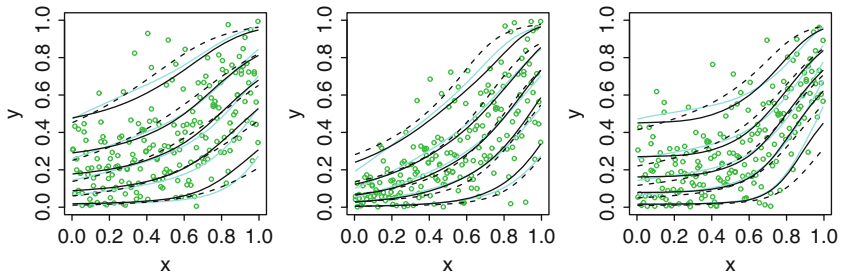


FIGURE 10.7. Conditional density estimation via (10.27) on $\mathcal{X} = [0,1]$ and $\mathcal{Y} = [0,1]$. The 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$ are in *solid lines*, those of the test distributions in *faded lines*, and the data in *circles*. Estimates via (7.30) are also shown in *dashed lines*. Form *left to right*: $\sigma_x = 0.3$, $0.15(1+x)$, $0.15(2-x)$.

Shown in the right frame of Fig. 10.7 are the 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$, with the data superimposed:

```
quan <- qsscden(fit,c(.05,.25,.5,.75,.95),
                data.frame(x=xx))
plot(x,y,col=3); for (i in 1:5) lines(xx,quan[i,])
```

Also superimposed are the respective percentiles of the test distribution (faded) and those of an estimate via (7.30) (dashed); the latter was obtained using the `sscden` suite discussed in §7.7.4 and was shown in Fig. 7.12 on page 267. Parallel results with $\sigma_x = 0.3$ and $\sigma_x = 0.15(1+x)$ are shown in the left and the center frames of Fig. 10.7, respectively.

The syntax of `sscden1` is largely identical to that of `sscden`, except that one needs to specify $\rho(x,y)$. In the call above, we used the default `rho=list("xy")`, which generates $\rho(x,y)$ internally using (10.32); with `rho=list("y")`, an estimate of the marginal density on $\mathcal{Y}$ will be generated internally to use as $\rho(x,y)$.

One may also generate $\rho(x,y)$ externally and pass it into `sscden1` via the argument `rho`, to be evaluated through

```
rho$fun(x,y,rho$env,outer.prod)
```

where `rho$env` contains constants and the logical flag `outer.prod` indicates whether to return a vector of $\rho(x_i,y_i)$ or the matrix $\rho(\mathbf{x},\mathbf{y}^T)$; `rho$env` must be a list object at least containing a quadrature on $\mathcal{Y}$ in the elements `rho$env$qd.pt` and `rho$env$qd.wt`.

From left to right in Fig. 10.7, the three solid fits using `sscden1` took 1.51, 1.40, and 1.63 CPU seconds on a linux laptop, in order. The respective dashed fits using `sscden` took 21.4, 38.2, and 43.9 CPU seconds.

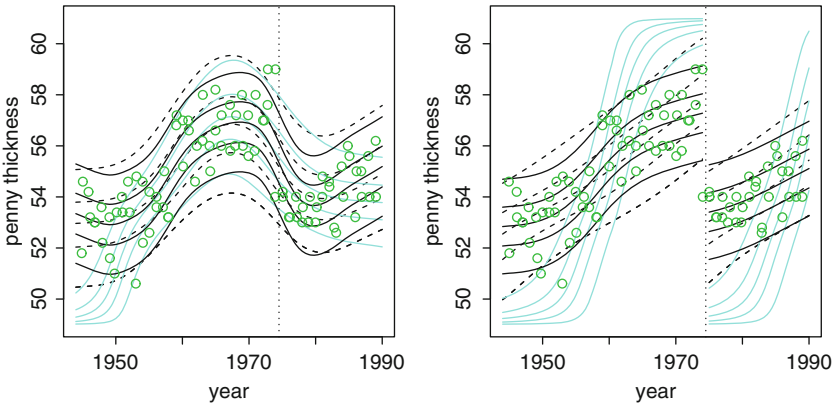


FIGURE 10.8. Thickness of U.S. Lincoln Pennies. *Left*: Continuous fits. *Right*: Fits with built-in break. The *lines* are the 5th, 25th, 50th, 75th, and 95th percentiles of the fitted $f(y|x)$, with the default `sscden1` fits in *solid*, the `sscden1` fits with `rho=list("y")` in *faded*, and the `sscden` fits in *dashed*. The data, with the year jittered, are superimposed in *circles*. The *vertical dotted lines* mark the position of the break.

10.3.5 Case Study: Penny Thickness

We now look at a `sscden1` fit to the penny thickness data of §7.7.5, shown in the left frame of Fig. 10.8 in solid lines with the data superimposed:

```
data(penny); set.seed(5732)
fit <- sscden1(~year*mil,~mil,data=penny,
               ydomain=data.frame(mil=c(49,61)))
yy <- 1944+(0:92)/2
quan <- qsscden(fit,c(.05,.25,.5,.75,.95),
                data.frame(year=yy))
plot(penny$year+.1*rnorm(90),penny$mil,ylim=c(49,61))
for (i in 1:5) lines(yy,quan[i,])
```

Also superimposed are a `sscden` fit in dashed lines and a `sscden1` fit with `rho=list("y")` in faded lines. Parallel fits with a built-in break at `year=1974.5` are shown in the right frame of Fig. 10.8.

The support of $f(y|x)$ seems to vary greatly with x, and `rho=list("y")` barely holds up in the left frame but completely breaks down in the right. The default `sscden1` fits with `rho=list("xy")` appear to be closer to the data than the `sscden` fits.

The solid fits using `sscden1` in the left and the right frames of Fig. 10.8 took around 1 CPU second each on a linux laptop. The respective dashed fits using `sscden` took about 20 and 11 CPU seconds.

10.3.6 Asymptotic Convergence

The theory of §10.2 can be readily modified for the conditional density estimation via (10.27). Denote by $e^{\eta_0(x,y)}\rho(x,y)$ the conditional density to be estimated satisfying $\int_{\mathcal{Y}} e^{\eta_0(x,y)}\rho(x,y) = 1$, $\forall x \in \mathcal{X}$, and by $\hat{\eta}(x,y)$ the minimizer of (10.27). It is clear that the space $\mathcal{H}$ must contain the ANOVA components $\eta_\emptyset$ and η_x in order for $\eta_0 \in \mathcal{H}$.

Define $V(g) = \int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} g^2(x,y)\rho(x,y)$, where $f(x)$ is the limiting density of X_i. Apart from the modified definition of V, little change is needed in Conditions 10.2.1 and 10.2.2 and in the statements of the theorems. Conditions 10.2.3–10.2.5 shall be trivially modified as follows.

Condition 10.3.3 For some $c_3 < \infty$, $e^{-\eta_0(x,y)} < c_3$.

Condition 10.3.4 For η in a convex set B_0 around η_0 containing $\hat{\eta}$, $\tilde{\eta}$, $\hat{\eta}^*$, and η^*, $c_1 < e^{\eta_0(x,y)-\eta(x,y)} < c_2$ holds uniformly for some $0 < c_1 < c_2 < \infty$.

Condition 10.3.5 $\int_{\mathcal{X}} f(x) \int_{\mathcal{Y}} \phi_\nu^2(x,y)\phi_\mu^2(x,y)e^{-\eta_0(x,y)}\rho(x,y) < c_4$, $\forall \nu, \mu$, for some $c_4 < \infty$.

The efficient approximation $\hat{\eta}^*$ minimizes (10.27) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J(V_j, \cdot), j = 1, \ldots, q\},$$

where $\{V_j\}$ is a random subset of $\{(X_i, Y_i)\}$.

10.4 Hazard Estimation

As an alternative to (8.1), one may estimate a covariate-dependent hazard $e^{\eta(t,u)}$ via the minimization of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i e^{-\eta(X_i,U_i)}\rho(X_i,U_i) + \int_{Z_i}^{X_i} \eta(t,U_i)\rho(t,U_i)dt\right\} + \frac{\lambda}{2}J(\eta), \quad (10.36)$$

where $\rho(t,u)$ is a known positive function. This works similar to (10.1), and the integral $n^{-1}\sum_{i=1}^{n}\int_{Z_i}^{X_i}\eta(t,U_i)\rho(t,U_i)dt$ can largely be pre-computed.

The existence and the computation of the minimizer of (10.36) is similar to that of (10.1) and (10.27), and a cross-validation score similar to (10.35) can be used for smoothing parameter selection. The Bayesian confidence intervals can be adapted, and a square error projection replaces the Kullback-Leibler projection. Software tools are illustrated via simulated and real-data examples, and comparisons are made against the penalized likelihood estimates through (8.1).

Parallel to the analysis of §9.3, asymptotic convergence rates can be calculated for the minimizer of (10.36), which is the subject of §10.5.

10.4.1 Preliminaries

One minimizes (10.36) in $\mathcal{H}$ on $\mathcal{T} \times \mathcal{U}$ with a square (semi) norm $J(f)$. Write $L(f) = n^{-1} \sum_{i=1}^{n} \{\delta_i e^{-f(X_i,U_i)} \rho(X_i, U_i) + \int_{Z_i}^{X_i} f(t, U_i)\rho(t, U_i)dt\}$. It is easy to verify that $L(f)$ is continuous, convex, and Fréchet differentiable. Let $\{\phi_\nu\}_{\nu=1}^{m}$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$, $(T_j, \tilde{U}_j)$ be the $N = \sum_{i=1}^{n} \delta_i$ observed lifetimes, and S be $N \times m$ with the (j, ν)th entry $\phi_\nu(T_j, \tilde{U}_j)$. If S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$, and $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$. See Problem 10.7. By Theorem 2.9, the minimizer of (10.36) uniquely exists when S is of full column rank, which we will assume.

Without loss of inferential efficiency, one may minimize (10.36) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \text{span}\{R_J((\tilde{T}_j, \tilde{U}_j), \cdot), j = 1, \ldots, q\}, \tag{10.37}$$

where $\{(\tilde{T}_j, \tilde{U}_j)\}_{j=1}^{q} \subseteq \{(X_i, U_i), \delta_i = 1\}$ is a random subset of the failure cases; see §10.5.3. One has an expression

$$\eta(t, u) = \sum_{\nu} d_\nu \phi_\nu(t, u) + \sum_{j} c_j R_J(\tilde{T}_j, \tilde{U}_j; t, u) = \boldsymbol{\phi}^{\boldsymbol{T}} \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}. \tag{10.38}$$

Plugging (10.38) into (10.36), one has

$$A_\lambda(\mathbf{c}, \mathbf{d}) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \rho_i e^{-\boldsymbol{\phi}_i^T \mathbf{d} - \boldsymbol{\xi}_i^T \mathbf{c}} + \mathbf{b}_\phi^T \mathbf{d} + \mathbf{b}_\xi^T \mathbf{c} + \frac{\lambda}{2} \mathbf{c}^T Q \mathbf{c}, \tag{10.39}$$

where $\rho_i = \rho(X_i, U_i)$, $\boldsymbol{\phi}_i = \boldsymbol{\phi}(X_i, U_i)$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(X_i, U_i)$,

$$\mathbf{b}_\phi = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \boldsymbol{\phi}(t, U_i)\rho(t, U_i)dt, \quad \mathbf{b}_\xi = \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \boldsymbol{\xi}(t, U_i)\rho(t, U_i)dt,$$

and Q is $q \times q$ with the (j, k)th entry $J(\xi_j, \xi_k) = R_J((\tilde{T}_j, \tilde{U}_j), (\tilde{T}_k, \tilde{U}_k))$; note that $\mathbf{b}_\phi$ and $\mathbf{b}_\xi$ can be computed once for all.

Taking derivatives of $A_\lambda(\mathbf{c}, \mathbf{d})$ at $\tilde{\eta} = \boldsymbol{\phi}^T \tilde{\mathbf{d}} + \boldsymbol{\xi}^T \tilde{\mathbf{c}} \in \mathcal{H}^*$, one has

$$\begin{aligned}
\frac{\partial A_\lambda}{\partial \mathbf{d}} &= -\mu_{\tilde{\eta}}(\boldsymbol{\phi}) + \mathbf{b}_\phi = -\mu_\phi + \mathbf{b}_\phi, \\
\frac{\partial A_\lambda}{\partial \mathbf{c}} &= -\mu_{\tilde{\eta}}(\boldsymbol{\xi}) + \mathbf{b}_\xi + \lambda Q \tilde{\mathbf{c}} = -\mu_\xi + \mathbf{b}_\xi + \lambda Q \tilde{\mathbf{c}}, \\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d} \partial \mathbf{d}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T) = V_{\phi,\phi}, \\
\frac{\partial^2 A_\lambda}{\partial \mathbf{c} \partial \mathbf{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T) + \lambda Q = V_{\xi,\xi} + \lambda Q, \\
\frac{\partial^2 A_\lambda}{\partial \mathbf{d} \partial \mathbf{c}^T} &= V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T) = V_{\phi,\xi},
\end{aligned} \tag{10.40}$$

where $\mu_f(g) = n^{-1}\sum_{i=1}^n \delta_i\rho_i e^{-f(X_i,U_i)}g(X_i,U_i)$ and $V_f(g,h) = \mu_f(gh)$. This simply duplicates (10.7) but with modified definitions of entities. The Newton updating equation is virtually the same as (10.8); see Problem 10.8.

The $\rho(t,u)$ function acts to replace $e^{\eta(t,u)}$ as the weight $w(t,u)$ in a weighted mean square error $V(\hat\eta-\eta) = \int_{\mathcal{U}} m(u)\int_{\mathcal{T}}(\hat\eta-\eta)^2 w\tilde S dt$, where $m(u)$ is the density of U and $\tilde S(t,u) = P(Z < t \le X|U=u)$ is the at-risk probability; see §10.5. One may set $\rho(t,u) = e^{\eta(t)}$ as a hazard estimate via (8.1) absent of covariate, or set $\rho(t,u)$ as an estimate parametric in t using the techniques of §8.6.

10.4.2 Smoothing Parameter Selection

For smoothing parameter selection with the minimizer η_λ of (10.36), consider a loss function similar to (10.33),

$$\tilde L(\eta,\eta_\lambda) = E\big[\int_{\mathcal{T}}\big\{e^{(\eta-\eta_\lambda)(t,U)} - (\eta-\eta_\lambda)(t,U) - 1\big\}\rho(t,U)Y(t)dt\big].$$

Dropping terms not involving η_λ, one has the relative loss,

$$E\big[\int_{\mathcal{T}} e^{-\eta_\lambda(t,U)}\rho(t,U)e^{\eta(t,U)}Y(t)dt\big] + E\big[\int_{\mathcal{T}}\eta_\lambda(t,U)\rho(t,U)Y(t)dt\big]. \tag{10.41}$$

A cross-validation estimate of the first term in (10.41) is available by setting $h(t,U_i) = e^{-\eta_\lambda^{[i]}(t,U_i)}\rho(t,U_i)$ in (8.8) on page 290, where $\eta_\lambda^{[i]}$ minimizes some delete-one version of (10.36), and the second term may be substituted by its empirical version, yielding

$$\frac{1}{n}\sum_{i=1}^n \delta_i\rho_i e^{-\eta_\lambda^{[i]}(X_i,U_i)} + \frac{1}{n}\sum_{i=1}^n\int_{Z_i}^{X_i}\eta_\lambda(t,U_i)\rho(t,U_i)dt. \tag{10.42}$$

With the same abuse of notation as in (10.11), write $\eta = \boldsymbol{\xi}^T\mathbf{c}$ in (10.36) and denote its minimizer by $\eta_\lambda = \tilde\eta = \boldsymbol{\xi}^T\tilde{\mathbf{c}}$. The quadratic approximation of (10.36) at $\tilde\eta$ virtually duplicates (10.11), but with $w_i = \delta_i\rho_i e^{-\tilde\eta_i}$ for $\tilde\eta_i = \tilde\eta(X_i,U_i)$. Solving the delete-one version, one again has

$$\eta_\lambda^{[i]}(X_i,U_i) = \boldsymbol{\xi}_i^T\mathbf{c}^{[i]} = \boldsymbol{\xi}_i^T\mathbf{c} - \frac{a_i}{1-a_i} = \eta_\lambda(X_i,U_i) - \frac{a_i}{1-a_i}$$

where $a_i = n^{-1}w_i\boldsymbol{\xi}_i^T H^{-1}\boldsymbol{\xi}_i$ as in (10.12) but with the modified w_i and in turn $H = n^{-1}\sum_{i=1}^n w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T + \lambda\tilde Q$. Plugging this into (10.42), one gets

$$\begin{aligned} V(\lambda) =& \frac{1}{n}\sum_{i=1}^n\left\{\delta_i\rho_i e^{-\eta_\lambda(X_i,U_i)} + \int_{Z_i}^{X_i}\eta_\lambda(t,U_i)\rho(t,U_i)dt\right\} \\ &+ \alpha\frac{1}{n}\sum_{i=1}^n \delta_i\rho_i e^{-\eta_\lambda(X_i,U_i)}\big(e^{a_i/(1-a_i)} - 1\big), \end{aligned} \tag{10.43}$$

for $\alpha = 1$. Unlike (10.12) but similar to (10.35), one does not need to stop early for (10.43) to work.

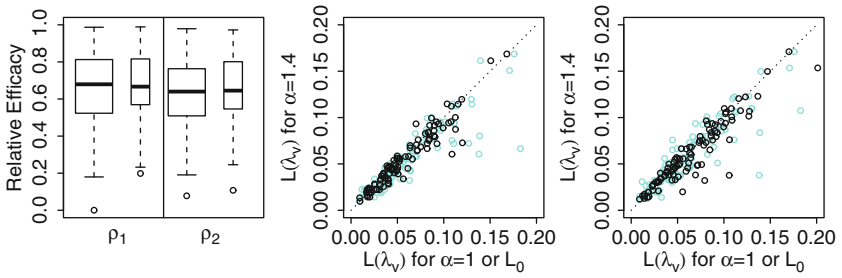


FIGURE 10.9. Effectiveness of cross-validation for hazard estimation. *Left*: Relative efficacy $L(\lambda_o)/L(\lambda_v)$ with $\alpha = 1$ (*wider boxes*) and $\alpha = 1.4$ (*thinner boxes*). *Center*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\rho_1(t, u)$. *Right*: $L(\lambda_v)$ with $\alpha = 1$ versus $L(\lambda_v)$ with $\alpha = 1.4$, for $\rho_2(t, u)$. *Faded points* are $L(\lambda_v)$ with $\alpha = 1.4$ versus L_0 via (8.1).

Empirical Performance

Recall the simulation of §8.2.2 on $\mathcal{U} = [0, 1]$ with the test hazard $\lambda_2(t, u)$ as given in (8.14) on page 291. Estimates were also calculated via (10.36) using the same samples and the same $\{(\tilde{T}_j, \tilde{U}_j)\}$ of size $q = 31$ in (10.37), with two versions of $\rho(t, u)$, $\rho_1(t, u) = e^{\eta(t)}$ a hazard estimate via (8.1) absent of covariate, and $\rho_2(t, u) = (\nu/t)e^{\nu\{\log t - \eta(u)\}}$ obtained via the Weibull regression of §8.6.3.

For each replicate and each version of $\rho(t, u)$, three estimates were calculated, one minimizing the symmetrized Kullback-Leibler distance $L(\lambda)$ of (8.13) at λ_o, and two minimizing $V(\lambda)$ of (10.42) with $\alpha = 1, 1.4$ at λ_v. The results are summarized in Fig. 10.9, where comparisons against the estimates via (8.1) are also shown in the center and right frames in faded points; a solid point in the center frame is off the chart. The two versions of $\rho(t, u)$ delivered similar performances in the setting, $\alpha = 1.4$ was slightly preferred to $\alpha = 1$, and the estimates via (10.36) actually did slightly better than those via (8.1).

The one hundred replicates with $\rho_1(t, u)$ and $\alpha = 1.4$ took 233.2 CPU seconds on a linux server, those with $\rho_2(t, u)$ and $\alpha = 1.4$ took 232.5 CPU seconds, and those via (8.1) took 5451.5 CPU seconds.

10.4.3 Inference

The inferential and modeling tools of §8.3 are readily adapted.

Bayesian Confidence Intervals

With the same abuse of notation as in (10.11), write $\eta = \boldsymbol{\xi}^T \mathbf{c}$ in (10.36) and refer η and $\mathbf{c}$ interchangeably. The quadratic approximation of (10.36)

at $\tilde{\eta} = \eta_\lambda$ can be written as

$$\frac{1}{2n}(\mathbf{c} - \tilde{\mathbf{c}})^T (nH)(\mathbf{c} - \tilde{\mathbf{c}}) + C,$$

where $H = n^{-1} \sum_{i=1}^n w_i \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T + \lambda \tilde{Q}$, for $w_i = \delta_i \rho_i e^{-\tilde{\eta}_i}$ and $\tilde{Q} = \text{diag}(O, Q)$. This may be perceived as an approximate posterior likelihood of $\mathbf{c}$, with mean $\tilde{\mathbf{c}}$ and covariance H^+/n, where H^+ is the Moore-Penrose inverse of H. The posterior of $\eta(t,u)$ is thus approximately normal with mean $\tilde{\eta}(t,u) = \boldsymbol{\xi}^T(t,u)\tilde{\mathbf{c}}$ and variance $s^2(t,u) = \boldsymbol{\xi}^T(t,u) H^+ \boldsymbol{\xi}(t,u)/n$. Bayesian confidence intervals of $\eta(t,u)$ are given by $\tilde{\eta}(t,u) \pm z_{1-\alpha/2}\, s(t,u)$.

Square Error Projection

Consider the empirical version of $V(\hat{\eta} - \eta) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} (\hat{\eta} - \eta)^2 \rho \tilde{S} dt$,

$$\tilde{V}(\hat{\eta} - \eta) = \frac{1}{n} \sum_{i=1}^n \int_{Z_i}^{X_i} \big\{(\hat{\eta} - \eta)(t, U_i)\big\}^2 \rho(t, U_i) dt.$$

For $\hat{\eta} \in \mathcal{H}_0 \oplus \mathcal{H}_1$, one may calculate its square error projection in $\mathcal{H}_0$ by minimizing $\tilde{V}(\hat{\eta} - \eta)$ over $\eta \in \mathcal{H}_0$. Let $\tilde{\eta}$ be the square error projection of $\hat{\eta}$ in $\mathcal{H}_0$ and consider $A_{\tilde{\eta},h}(\alpha) = \tilde{V}\big(\hat{\eta} - (\tilde{\eta} + \alpha h)\big)$ for $h \in \mathcal{H}_0$. One has $\dot{A}_{\tilde{\eta},h}(0) = \tilde{V}(\hat{\eta} - \tilde{\eta}, h) = 0$, $\forall h \in \mathcal{H}_0$.

For $\eta_c \in \mathcal{H}_0$, $\tilde{V}(\hat{\eta} - \tilde{\eta}, \tilde{\eta} - \eta_c) = 0$, so $\tilde{V}(\hat{\eta} - \eta_c) = \tilde{V}(\hat{\eta} - \tilde{\eta}) + \tilde{V}(\tilde{\eta} - \eta_c)$. When the ratio $\tilde{V}(\hat{\eta} - \tilde{\eta})/\tilde{V}(\hat{\eta} - \eta_c)$ is small, one may safely cut out $\mathcal{H}_1$. One may take $e^{\eta_c} = \sum_{i=1}^n \delta_i \rho_i / \sum_{i=1}^n \int_{Z_i}^{X_i} \rho(t, U_i) dt$, which is the constant hazard minimizing $\sum_{i=1}^n \big\{\delta_i \rho_i e^{-\eta} + \int_{Z_i}^{X_i} \eta \rho(t, U_i) dt\big\}$.

Frailty Models for Correlated Data

The frailty model of §8.3.3 can be estimated via the minimization of

$$\frac{1}{n} \sum_{i=1}^n \left\{ \delta_i e^{-(\eta(X_i, U_i) + \mathbf{z}_i^T \mathbf{b})} \rho(X_i, U_i) + \int_{Z_i}^{X_i} \big(\eta(t, U_i) + \mathbf{z}_i^T \mathbf{b}\big) \rho(t, U_i) dt \right\}$$
$$+ \frac{1}{2n} \mathbf{b}^T \Sigma \mathbf{b} + \frac{\lambda}{2} J(\eta). \quad (10.44)$$

The Newton updating equation is straightforward to derive, and the tuning parameters can be jointly selected via (10.43). Bayesian confidence intervals are straightforward to adapt and the square error projection can be computed with $\mathbf{z}^T \mathbf{b}$ treated as an offset.

10.4.4 R Package `gss`: `sshzd1` Suite

Hazard estimation via (10.36) is implemented in the `sshzd1` suite. The following sequence generates a sample of size $n = 150$ with $T|U$ from $\lambda_2(t,u)$ of (8.14) and fits a tensor product cubic spline to the log hazard:

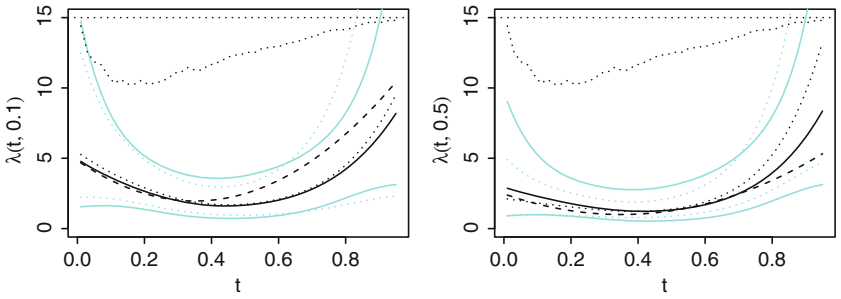


FIGURE 10.10. Hazard estimation on $\mathcal{T} = [0, 1]$ and $\mathcal{U} = [0, 1]$. The estimated $e^{\eta(t,u)}$ are in *solid lines*, the 95 % Bayesian confidence intervals in *faded lines*, and the test hazard $\lambda_2(t, u) = \{24(t - 0.35)^2 + 2\}\{3(u - 0.5)^2 + 0.5\}$ in *dashed lines*. *Left*: $u = 0.1$. *Right*: $u = 0.5$. The estimate via (8.1) is superimposed in *dotted lines*. The *dotted lines* from above are proportional to the size of the risk set, $\sum_{i=1}^{n} I_{[Z_i < t \le X_i]}$.

```
set.seed(2375)
xdzu <- rtest2(150)
x <- xdzu[,1]; delta <- xdzu[,2]
z <- xdzu[,3]; u <- xdzu[,4]
fit <- sshzd1(Surv(x,delta,z)~x*u)
```

where `rtest2` was listed in §8.3.4. Projecting the fit into the space of additive models, one has

```
project(fit,inc=c("x","u"))$ratio
# 0.02643945
```

To evaluate the fitted hazard, say at $(t, u) = (0.5, 0.5)$, one may use

```
hzdrate.sshzd(fit,data.frame(x=.5,u=.5))
# 1.320611
```

The estimated $\lambda(t, u) = e^{\eta(t,u)}$ is shown in Fig. 10.10, superimposed with the estimate via (8.1) seen in Fig. 8.2.

The syntax of `sshzd1` is largely identical to that of `sshzd`, except for the specification of $\rho(t, u)$; the default `rho=list("marginal")` specifies a covariate-free $\rho(t, u) = e^{\eta(t)}$ via (8.1), and `rho=list("weibull")` uses $\rho(t, u) = (\nu/t)e^{\nu\{\log t - \eta(u)\}}$ with $\eta(u)$ and ν from the Weibull regression of §8.6.3, both calculated internally. One may also create $\rho(t, u)$ externally and pass it into `sshzd1` via `rho`, to be evaluated through

```
rho$fun(t,u,rho$env,outer.prod)
```

This is similar to `sscden1` of §10.3.4, but one does not need to supply in `rho$env` a quadrature on $\mathcal{T}$ as it is generated internally.

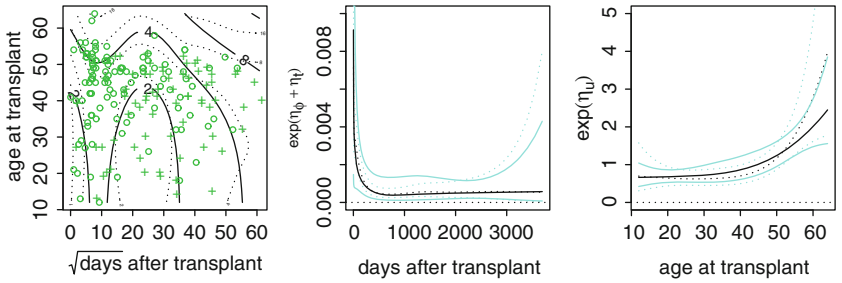


FIGURE 10.11. Hazard after heart transplant: Proportional hazard fit. *Left*: Contours of $100\tilde{\lambda}(t^*, u)$, with deceased (*circles*) and censored (*pluses*) patients superimposed. *Center*: Base hazard $e^{\eta_0+\eta_t}$ with 95 % Bayesian confidence intervals, on the original time scale. *Right*: Age effect e^{η_u} with 95 % Bayesian confidence intervals. The Estimate via (8.1) is superimposed in *dotted lines*.

10.4.5 Case Study: Survival After Heart Transplant

For a quick analysis of the Stanford heart transplant data of §1.4.3 and §8.4.2, one may try:

```
data(stan)
fit.stan <- sshzd1(Surv(futime,status)~futime*age,
                   data=stan,nbasis=200)
project(fit.stan,inc=c("futime","age"))$ratio
# 0.03536646
fit1.stan <- sshzd1(Surv(futime,status)~futime+age,
                    data=stan,nbasis=200)
```

The proportional hazard fit is shown in Fig. 10.11, superimposed with the estimate via (8.1) seen in Fig. 8.5.

The solid fit in Fig. 10.11 using `sshzd1` took 8.6 CPU seconds on a linux laptop; the dotted fit using `sshzd` took 51.8 CPU seconds.

10.5 Hazard Estimation: Asymptotic Convergence

Denote by $e^{\eta_0(t,u)}$ the hazard to be estimated and by $\hat{\eta}(t, u)$ the minimizer of (10.36). Define

$$V(f) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} f^2(t, u)\rho(t, u)\tilde{S}(t, u)dt, \tag{10.45}$$

where $\rho(t, u)$ replaces $e^{\eta_0(t,u)}$ in (9.24) on page 334 for the definition of $V(f)$. Convergence rates here are in terms of $V(\hat{\eta}-\eta_0)$ as defined in (10.45).

The analysis is adapted from that of §9.3, from which much of the notation is inherited. It is convenient to write (10.36) as

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}} e^{-\eta_i}\rho_i\, dN_i(t) + \int_{\mathcal{T}} \eta_i\rho_i Y_i dt\right\} + \frac{\lambda}{2}J(\eta), \tag{10.46}$$

where $\rho_i = \rho(t, U_i)$ and the rest of the terms are as in (9.23).

10.5.1 Linear Approximation

Conditions 9.3.1 and 9.3.2 are recycled, but with V as defined in (10.45).

Condition 10.5.1 V is completely continuous with respect to J.

Condition 10.5.2 For ν sufficiently large and some $\beta > 0$, the eigenvalues ρ_ν of J with respect to V satisfy $\rho_\nu > \beta\nu^r$, where $r > 1$.

Consider the quadratic functional

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}} \eta_i e^{-\eta_{0,i}}\rho_i dN_i(t) + \int_{\mathcal{T}} \eta_i\rho_i Y_i dt\right\} + \frac{1}{2}V(\eta-\eta_0) + \frac{\lambda}{2}J(\eta), \tag{10.47}$$

where $\eta_{0,i}(t) = \eta_0(t, U_i)$. Plugging the Fourier expansions $\eta = \sum_\nu \eta_\nu\phi_\nu$ and $\eta_0 = \sum_\nu \eta_{\nu,0}\phi_\nu$ into (10.47), the minimizer $\tilde{\eta}$ of (10.47) has Fourier coefficients

$$\tilde{\eta}_\nu = (\beta_\nu + \eta_{\nu,0})/(1 + \lambda\rho_\nu),$$

where $\beta_\nu = n^{-1}\sum_{i=1}^{n}\int_{\mathcal{T}} \phi_{\nu,i}e^{-\eta_{0,i}}\rho_i dM_i(t)$ with $\phi_{\nu,i}(t) = \phi_\nu(t, U_i)$. From (9.20) and (9.21), one has $E[\beta_\nu] = 0$, $E\left[\beta_\nu^2\right] = n^{-1}\int_{\mathcal{U}} m(u)\int_{\mathcal{T}} \phi_\nu^2 e^{-\eta_0}\rho^2\tilde{S}dt$.

Condition 10.5.3 For some $c_3 < \infty$, $e^{-\eta_0(t,u)}\rho(t,u) < c_3$.

Under Condition 10.5.3, $E\left[\beta_\nu^2\right] \le c_3/n$, noting that $\int_{\mathcal{U}} m(u)\int_{\mathcal{T}} \phi_\nu^2\rho\tilde{S}dt = V(\phi_\nu) = 1$.

Theorem 10.6 *Assume* $\sum_\nu \rho_\nu^p\eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 10.5.1–10.5.3, as* $n \to \infty$ *and* $\lambda \to 0$,

$$(V + \lambda J)(\tilde{\eta} - \eta_0) = O_p\left(n^{-1}\lambda^{-1/r} + \lambda^p\right).$$

Proof: See the proof of Theorem 9.2. □

10.5.2 Approximation Error and Main Results

We now turn to the approximation error $\hat{\eta} - \tilde{\eta}$. Define

$$A_{f,g}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{\int_{\mathcal{T}} e^{-(f+\alpha g)_i}\rho_i dN_i(t) + \int_{\mathcal{T}}(f+\alpha g)_i\rho_i Y_i dt\right\} + \frac{\lambda}{2}J(f+\alpha g),$$

$$B_{f,g}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}}(f+\alpha g)_i e^{-\eta_{0,i}}\rho_i dN_i(t) + \int_{\mathcal{T}}(f+\alpha g)_i\rho_i Y_i dt\right\} + \frac{1}{2}V(f+\alpha g-\eta_0) + \frac{\lambda}{2}J(f+\alpha g).$$

It can be shown that

$$\dot{A}_{f,g}(0) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}} g_i e^{-f_i}\rho_i dN_i(t) + \int_{\mathcal{T}} g_i\rho_i Y_i dt\right\} + \lambda J(f,g), \tag{10.48}$$

$$\dot{B}_{f,g}(0) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}} g_i e^{-\eta_{0,i}}\rho_i dN_i(t) + \int_{\mathcal{T}} g_i\rho_i Y_i dt\right\} + V(f-\eta_0, g) + \lambda J(f,g). \tag{10.49}$$

Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (10.48), one has

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i e^{-\hat{\eta}_i}\rho_i dN_i(t) + \int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i\rho_i Y_i dt\right\} + \lambda J(\hat{\eta}, \hat{\eta}-\tilde{\eta}) = 0, \tag{10.50}$$

and setting $f = \tilde{\eta}$ and $g = \hat{\eta} - \tilde{\eta}$ in (10.49), one gets

$$\frac{1}{n}\sum_{i=1}^{n}\left\{-\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i e^{-\eta_{0,i}}\rho_i dN_i(t) + \int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i\rho_i Y_i dt\right\} + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}) + \lambda J(\tilde{\eta}, \hat{\eta}-\tilde{\eta}) = 0. \tag{10.51}$$

Subtracting (10.51) from (10.50), some algebra yields

$$\lambda J(\hat{\eta}-\tilde{\eta}) - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i(e^{-\hat{\eta}} - e^{-\tilde{\eta}})_i\rho_i dN_i(t) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}-\tilde{\eta})_i(e^{-\tilde{\eta}} - e^{-\eta_0})_i\rho_i dN_i(t) + V(\tilde{\eta}-\eta_0, \hat{\eta}-\tilde{\eta}). \tag{10.52}$$

Condition 10.5.4 For η in a convex set B_0 around η_0 containing $\hat{\eta}$ and $\tilde{\eta}$, $c_1 \le e^{\eta_0(t,u)-\eta(t,u)} \le c_2$ holds uniformly for some $0 < c_1 < c_2 < \infty$.

Condition 10.5.5 $\int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu^2 \phi_\mu^2 \rho^k \tilde{S} dt \le c_4$, $\forall \nu, \mu$, for some $c_4 < \infty$ and $k = 1, 2$.

Parallel to Lemma 9.10, one has the following lemma.

Lemma 10.7 *Under Conditions 10.5.1–10.5.3 and 10.5.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} f_i g_i e^{-\eta_{0,i}} \rho_i dN_i(t) = V(f,g) + o_p\big(\{(V+\lambda J)(f)(V+\lambda J)(g)\}^{1/2}\big).$$

Proof: The proof parallels that of Lemma 9.10. One needs to bound

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{n} \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} e^{-\eta_{0,i}} \rho_i dN_i(t) - \tau(\phi_\nu \phi_\mu) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} e^{-\eta_{0,i}} \rho_i dM_i(t) + \int_{\mathcal{T}} \phi_{\nu,i} \phi_{\mu,i} \rho_i Y_i dt - \tau(\phi_\nu \phi_\mu) \right\}, \end{aligned}$$

where $\tau(f) = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} f \rho \tilde{S} dt$. Under Conditions 10.5.3 and 10.5.5,

$$E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu e^{-\eta_0} \rho \, dM(t) \right\}^2 \right] = \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu^2 \phi_\mu^2 e^{-\eta_0} \rho^2 \tilde{S} dt \le c_3 c_4.$$

By the arguments behind (9.31), $E\big[\{\int_{\mathcal{T}} \phi_\nu \phi_\mu \rho Y dt - \tau(\phi_\nu \phi_\mu)\}^2\big] \le 2c_4$; see Problem 10.9. The lemma follows. □

Theorem 10.8 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1,2]$. *Under Conditions 10.5.1–10.5.5, as* $\lambda \to 0$ *and* $n\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: By the mean value theorem, Condition 10.5.4, and Lemma 10.7, (10.52) leads to

$$(c_1 V + \lambda J)(\hat{\eta} - \tilde{\eta}) \le \big(|1-c| + o_p(1)\big)\big\{(V+\lambda J)(\hat{\eta}-\tilde{\eta})(V+\lambda J)(\tilde{\eta}-\eta_0)\big\}^{1/2}$$

for some $c \in [c_1, c_2]$. The theorem follows Theorem 10.6. □

10.5.3 Efficient Approximation

Now consider the minimizer $\hat{\eta}^*$ of (10.46) in a space

$$\mathcal{H}^* = \mathcal{N}_J \oplus \operatorname{span}\big\{R_J\big((\tilde{X}_j, \tilde{U}_j), \cdot\big), \tilde{\delta}_j = 1\big\},$$

where $\big\{(\tilde{X}_j, \tilde{U}_j, \tilde{\delta}_j)\big\}_{j=1}^q \subseteq \big\{(X_i, U_i, \delta_i)\big\}_{i=1}^n$ is a random subset. The following lemma replicates Lemma 9.12.

Lemma 10.9 *Under Conditions 10.5.1–10.5.3 and 10.5.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$, $V(h) = o_p\big(\lambda J(h)\big)$, $\forall h \in \mathcal{H} \ominus \mathcal{H}^*$.

Proof: For $h \in \mathcal{H} \ominus \mathcal{H}^*$, $\tilde{\delta}_j h(\tilde{X}_j, \tilde{U}_j) = \tilde{\delta}_j J\big(R_J\big((\tilde{X}_j, \tilde{U}_j), \cdot\big), h\big) = 0$, so $\sum_{j=1}^q \int_{\mathcal{T}} h_j^2 e^{-\eta_{0,j}} \rho_j d\tilde{N}_j(t) = \sum_{j=1}^q \tilde{\delta}_j h^2(\tilde{X}_j, \tilde{U}_j) e^{-\eta_0(\tilde{X}_j, \tilde{U}_j)} \rho(\tilde{X}_j, \tilde{U}_j) = 0$, where $h_j(t) = h(t, \tilde{U}_j)$, $\eta_{0,j}(t) = \eta_0(t, \tilde{U}_j)$, $\rho_j(t) = \rho(t, \tilde{U}_j)$, and $\tilde{N}_j(t) = I_{[\tilde{X}_j \le t, \tilde{\delta}_j = 1]}$. By the arguments in the proofs of Lemmas 9.10 and 10.7,

$$V(h) = \left| V(h) - \frac{1}{q} \sum_{j=1}^q \int_{\mathcal{T}} h_j^2 e^{-\eta_{0,j}} \rho_j d\tilde{N}_j(t) \right| = O_p\big(q^{-1/2} \lambda^{-1/r}\big)(V + \lambda J)(h).$$

The lemma follows. □

Let η^* be the projection of $\hat{\eta}$ in $\mathcal{H}^*$; $J(\eta^*, \hat{\eta} - \eta^*) = 0$. The convex set B_0 in Condition 10.5.4 should also contain $\hat{\eta}^*$ and η^*.

Theorem 10.10 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 10.5.1–10.5.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta} - \eta^*) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

Proof: Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \eta^*$ in (10.48), one has

$$\frac{1}{n} \sum_{i=1}^n \left\{ - \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i e^{-\hat{\eta}_i} \rho_i dN_i(t) + \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i \rho_i Y_i dt \right\} + \lambda J(\hat{\eta}, \hat{\eta} - \eta^*) = 0, \tag{10.53}$$

which can be rearranged as

$$\begin{aligned} \lambda J(\hat{\eta} - \eta^*) = {} & \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i (e^{-\hat{\eta}} - e^{-\eta_0})_i dN_i(t) \\ & + \frac{1}{n} \sum_{i=1}^n \int_{\mathcal{T}} (\hat{\eta} - \eta^*)_i e^{-\eta_{0,i}} \rho_i dM_i(t). \end{aligned} \tag{10.54}$$

By the mean value theorem, Condition 10.5.4, and Lemma 10.7, the first term on the right-hand side of (10.54) is $\big(c + o_p(1)\big) V(\eta_0 - \hat{\eta}, \hat{\eta} - \eta^*)$ for some $c \in (c_1, c_2)$; parallel to (9.16), the second term is of the order $O_p(n^{-1/2}\lambda^{-1/2r}) \big\{(V + \lambda J)(\hat{\eta} - \eta^*)\big\}^{1/2}$. Combining these with Lemme 10.9 and Theorem 10.8, the theorem follows. □

Setting $f = \hat{\eta}^*$ and $g = \hat{\eta}^* - \eta^*$ in (10.48), one has

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \left\{ - \int_{\mathcal{T}} (\hat{\eta}^* - \eta^*)_i e^{-\hat{\eta}_i^*} \rho_i dN_i(t) + \int_{\mathcal{T}} (\hat{\eta}^* - \eta^*)_i \rho_i Y_i dt \right\} & \\ + \lambda J(\hat{\eta}^*, \hat{\eta}^* - \eta^*) = 0, & \end{aligned} \tag{10.55}$$

Setting $f = \hat{\eta}$ and $g = \hat{\eta} - \hat{\eta}^*$ in (10.48), one gets

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{-\int_{\mathcal{T}}(\hat{\eta}-\hat{\eta}^*)_i e^{-\hat{\eta}_i}\rho_i dN_i(t) + \int_{\mathcal{T}}(\hat{\eta}-\hat{\eta}^*)_i\rho_i Y_i dt\Big\} + \lambda J(\hat{\eta}, \hat{\eta}-\hat{\eta}^*) = 0, \tag{10.56}$$

Adding (10.55), (10.56) and subtracting (10.53), some algebra yields

$$\lambda J(\hat{\eta}^* - \eta^*) - \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i(e^{\hat{\eta}^*} - e^{\eta^*})_i\rho_i dN_i(t)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}(\hat{\eta}^*-\eta^*)_i(e^{\hat{\eta}} - e^{\eta^*})_i\rho_i dN_i(t). \tag{10.57}$$

By the mean value theorem, Condition 10.5.4, and Lemma 10.7, the left-hand side of (10.57) is no less than $\big(c_1+o_p(1)\big)V(\hat{\eta}^*-\eta^*)+\lambda J(\hat{\eta}^*-\eta^*)$, and the right-hand side is $\big(c+o_p(1)\big)V(\hat{\eta}-\eta^*, \hat{\eta}^*-\eta^*)$. These, in combination with Theorems 10.8 and 10.10, lead to the following theorem.

Theorem 10.11 *Assume* $\sum_\nu \rho_\nu^p \eta_{\nu,0}^2 < \infty$ *for some* $p \in [1, 2]$. *Under Conditions 10.5.1–10.5.5, as* $\lambda \to 0$ *and* $q\lambda^{2/r} \to \infty$,

$$(V + \lambda J)(\hat{\eta}^* - \eta_0) = O_p\big(n^{-1}\lambda^{-1/r} + \lambda^p\big).$$

10.6 Bibliographic Notes

Sections 10.1 and 10.2

Density estimation through the minimization of (10.1) was proposed by Jeon and Lin (2006). The cross-validation of (10.12), the square error projection of §10.1.4, and the asymptotic analysis of §10.2 are found in Gu, Jeon, and Lin (2013).

Section 10.3

The materials of this section are largely taken from Gu (2011).

Sections 10.4 and 10.5

Hazard estimation via (10.36) was studied in Du and Gu (2009). The asymptotic analysis of §10.5 is adapted from §9.3; a brief outline is found in Du and Gu (2009) in an appendix.

10.7 Problems

Section 10.1

10.1 Let $G(\eta) = -\int_{\mathcal{X}} \eta(x)f(x) + \log \int_{\mathcal{X}} e^{\eta(x)}$ and $B_{\tilde{\eta},h}(\alpha) = G(\tilde{\eta} + \alpha h)$. Calculate $\dot{B}_{\tilde{\eta},h}(0)$ and $\ddot{B}_{\tilde{\eta},h}(0)$, where $\tilde{\eta}$ minimizes $G(\eta)$.

10.2 Let $\{\phi_\nu, \nu = 1, \ldots, m\}$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$ and S be $n \times m$ with the (i, ν)th entry $\phi_\nu(X_i)$. Consider

$$L(f) = \frac{1}{n}\sum_{i=1}^{n} e^{-f(X_i)} + \int_{\mathcal{X}} f(x)\rho(x).$$

(a) Prove that $L(f)$ is continuous, convex, and Fréchet differentiable.

(b) Prove that if S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$.

(c) Prove that if S is of full column rank, then $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$.

10.3 Verify the Newton updating equation (10.8).

10.4 For $\tilde{\mathbf{c}} = H^{-1}\mathbf{d}$ and $\tilde{\mathbf{c}}^{[i]} = \big(H - n^{-1}w_i\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\big)^{-1}\big(\mathbf{d} - n^{-1}w_i(1+\tilde{g}_i)\boldsymbol{\xi}_i\big)$, verify that $\boldsymbol{\xi}_i^T\tilde{\mathbf{c}}^{[i]} = \boldsymbol{\xi}_i^T\tilde{\mathbf{c}} - a_i/(1-a_i)$, where $a_i = n^{-1}w_i\boldsymbol{\xi}_i^T H^{-1}\boldsymbol{\xi}_i$.

Section 10.2

10.5 Consider densities $f_0(x) \propto e^{\eta_0(x)}\rho(x)$ and $f(x) \propto e^{\eta(x)}\rho(x)$. Write $\mathrm{SKL}(\eta_0, \eta) = E_f \log\big(f(X)/f_0(X)\big) + E_{f_0} \log\big(f_0(X)/f(X)\big)$.

(a) Verify that

$$\mathrm{SKL}(\eta_0, \eta) = \frac{\int_{\mathcal{X}}(\eta - \eta_0)(x)e^{\eta(x)}\rho(x)}{\int_{\mathcal{X}} e^{\eta(x)}\rho(x)} - \frac{\int_{\mathcal{X}}(\eta - \eta_0)(x)e^{\eta_0(x)}\rho(x)}{\int_{\mathcal{X}} e^{\eta_0(x)}\rho(x)}.$$

(b) Define $A(\alpha) = \mathrm{SKL}\big(\eta_0, \eta_0 + \alpha(\eta - \eta_0)\big)$. Verify (10.13) using the mean value theorem.

10.6 Under Conditions 10.2.1, 10.2.2 and 10.2.5, prove (10.22) using arguments similar to those in the proof of Lemma 9.16.

Section 10.4

10.7 Let $\{\phi_\nu, \nu = 1, \ldots, m\}$ be a basis of $\mathcal{N}_J = \{f : J(f) = 0\}$, $(T_j, \tilde{U}_j)$ be the $N = \sum_{i=1}^n \delta_i$ observed lifetimes, and S be $N \times m$ with the (j, ν)th entry $\phi_\nu(T_j, \tilde{U}_j)$. Consider

$$L(f) = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i e^{-f(X_i, U_i)} \rho(X_i, U_i) + \int_{Z_i}^{X_i} f(t, U_i) \rho(t, U_i) dt \right\}.$$

(a) Prove that $L(f)$ is continuous, convex, and Fréchet differentiable.

(b) Prove that if S is of full column rank, then $L(f)$ is strictly convex in $\mathcal{N}_J$.

(c) Prove that if S is of full column rank, then $L(f) + \lambda J(f)$ is strictly convex in $\mathcal{H}$.

10.8 State the Newton updating equation for the minimization of (10.39).

Section 10.5

10.9 Under Condition 10.5.5, verify that

$$E\left[\left\{ \int_{\mathcal{T}} \phi_\nu \phi_\mu \, \rho \, Y dt - \int_{\mathcal{U}} m(u) \int_{\mathcal{T}} \phi_\nu \phi_\mu \, \rho \, \tilde{S} dt \right\}^2 \right] \leq 2c_4.$$

Appendix A
R Package `gss`

In this appendix, we outline the overall design of the R package `gss`. The code is assembled from three primary components, (i) utilities for the creation of the null space basis ϕ_ν and the reproducing kernels R_β, (ii) utilities implementing various modeling and data analytical tools, and (iii) the numerical engines that perform the bulk of the computation.

A.1 Model Construction

The utilities for the creation of ϕ_ν and R_β consists of numerous `mkphi` and `mkrk` functions and the assembler `mkterm` that puts things together using inputs from the model formula and the `type` argument.

For an example, consider the model formula in an `ssanova` call

```
ssanova(y~x1*x2)
```

with `x1` and `x2` both numerical vectors for which the default type is the cubic spline; this is Example 2.5 on page 44. The model formula yields four model terms in an ANOVA decomposition, `1`, `x1`, `x2`, and `x1:x2`, with `1` containing one $\phi_\nu = 1$ and no R_β, `x1` and `x2` each containing one ϕ_ν and one R_β, and `x1:x2` containing one ϕ_ν and three R_β's.

The $\phi_\nu(x)$ are to be evaluated via

```
phi$fun(x,nu,phi$env)
```

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7,

where x is the argument and phi$env contains constants. Similarly, $R_\beta(x, y)$ are to be evaluated through

```
rk$fun(x,y,nu,rk$env,outer.prod)
```

where x and y are the arguments, rk$env contains constants, and one may calculate $R_\beta(\mathbf{x}, \mathbf{y}^T)$ with outer.prod=TRUE.

In the rest of the section, we spell out how the marginal spaces are configured, how tensor product spaces are constructed, and how one may enter marginal configurations that are not "canned" in the package.

A.1.1 Marginal Configurations

For the construction of tensor product reproducing kernel Hilbert spaces discussed in §2.4, one simply takes the products of marginal kernels. The marginal spaces are individually configured, independent of each other.

The marginal configurations are directly used for the main effects in an ANOVA decomposition.

Numerical Vectors

For x a numerical vector, the default type is the cubic spline with a parametric contrast in $\text{span}\{\phi(x)\}$ with reproducing kernel $R_p(x, y) = \phi(x)\phi(y)$ and a nonparametric contrast in the space generated by the reproducing kernel $R_n(x, y)$, where $\phi(x) = k_1(x)$ and $R_n(x, y) = k_2(x)k_2(y) - k_4(x - y)$ after the domain $[a, b]$ is mapped onto $[0, 1]$; this is the formulation of §2.3.3 with $m = 2$. The default domain is the data range extended by 5 % on both ends, and to override the default, one may specify it via something like

```
type=list(x=list("cubic",c(a,b)))
```

Replacing "cubic" by "linear", with or without direct domain specification, one configures a linear spline with no "parametric contrast" and $R_n = k_1(x)k_1(y) + k_2(x - y)$.

To configure the periodic splines of §4.2.1, one may use

```
type=list(x=list("per",c(a,b)))
```

where "per" is the short version of "cubic.per" and the domain $[a, b]$ must be specified; there is no parametric contrast and $R_n(x, y) = -k_4(x-y)$ after mapping $[a, b]$ onto $[0, 1]$. Replacing "per" by "linear.per", one has the linear periodic spline with $R_n(x, y) = k_2(x - y)$.

To configure the trigonometric spline of (4.63) on page 152, one may use

```
type=list(x=list("trig",c(a,b)))
```

where, after mapping $[a, b]$ onto $[0, 1]$, one has the parametric contrast in $\text{span}\{\phi_1(x), \phi_2(x)\}$, for $\phi_1(x) = \sqrt{2}\cos 2\pi x$ and $\phi_2(x) = \sqrt{2}\sin 2\pi x$, with $R_p(x, y) = \phi_1(x)\phi_1(y) + \phi_2(x)\phi_2(y)$ and the nonparametric contrast generated by $R_n(x, y) = -k_4(x - y) - 2\cos 2\pi(x - y)/(2\pi)^4$.

Numerical Matrices

For `x` a numerical matrix, the default type is the thin-plate splines of §4.3. The default order is $m = 2$, which may or may not satisfy $2m - d > 0$. The default "normalizing mesh" $\{u_i\}$ in (4.23), $(f, g)_0 = \sum_i p_i f(u_i) g(u_i)$, are taken as the sampling points $\{x_i\}$ with $p_i \propto 1$. The parametric contrast is in span$\{\phi_\nu(x)\}$ of dimension $\binom{d+m-1}{d} - 1$ with $R_p(x, y) = \sum_\nu \phi_\nu(x)\phi_\nu(y)$, where $\phi_\nu(x)$ satisfying $(\phi_\nu, 1)_0 = 0$ are obtained numerically. The nonparametric contrast are generated by the reproducing kernel $R_n(x, y) = (I - P_{(x)})(I - P_{(y)})E(|x - y|)$ as given in (4.26). To override the default m, $\{u_i\}$, or p_i, use something like

```
type=list(x=list("tp",list(order=m,mesh=u,weight=p)))
```

To configure the spherical splines of §4.4 for `x` two-dimensional, one uses something like

```
type=list(x=list("sphere",2))
```

where the order $m = 2$ is the default so can be omitted in the `type` specification; other orders available are $m = 3, 4$. There is no parametric contrast and $R_n(x, y) = \frac{q_{2m-2}(x \cdot y) - 1/(2m-1)}{2\pi(2m-2)!}$ as given in (4.45). It is assumed that `x[,1]` is the latitude in degrees in the range of $[-90, 90]$ and `x[,2]` is the longitude in degrees in the range of $[-180, 180]$.

Factors

For `x` a factor, we use the constructions of §2.2. The contrast is finite-dimensional so technically is always parametric, but we decide to penalize it when the number of levels $K \geq 3$. Hence, for $\mathcal{X} = \{1, 2\}$, one has a parametric contrast in span$\{I_{[x=1]} - 1/2\}$ with $R_p(x, y) = I_{[x=y]} - 1/2$. For $\mathcal{X} = \{1, \ldots, K\}$, $K \geq 3$, one has a nonparametric contrast generated by $R_n(x, y) = I_{[x=y]} - 1/K$

For `x` an ordered factor with $K \geq 3$, one has $R_n(x, y) = B(x, y)$, where $B = (C^T C)^+$ for a $(K-1) \times K$ matrix C given by

$$C = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \ldots & 0 & \ldots & -1 & 1 \end{pmatrix}.$$

A.1.2 Construction of Interaction Terms

For interaction terms in an ANOVA decomposition, one takes products of the R_p's and R_n's of the marginals involved. A product containing at least one R_n is penalized, adding an R_β to the scene. A product containing only R_p's is unpenalized, contributing ϕ_ν's.

For an example, consider a two-way interaction `x1:x2`, where `x1` is configured as a cubic spline with a one-dimensional $R_{p\langle 1\rangle}$ and `x2` is configured as a thin-plate spline on $(-\infty,\infty)^2$ of order $m=2$ with $R_{p\langle 2\rangle}$ of dimension $\binom{2+2-1}{2}-1=2$. The term contains two ϕ_ν's, namely $\phi_{\langle 1\rangle}\phi_{1\langle 2\rangle}$ and $\phi_{\langle 1\rangle}\phi_{2\langle 2\rangle}$, and three R_β's, namely $R_{n\langle 1\rangle}R_{p\langle 2\rangle}$, $R_{p\langle 1\rangle}R_{n\langle 2\rangle}$, and $R_{n\langle 1\rangle}R_{n\langle 2\rangle}$.

A.1.3 Custom Types

The built-in support for marginal configurations as listed in §A.1.1 should satisfy most practical needs. In case some applications call for configurations not on the list, all is not lost, as the user can enter his own configurations via

```
type=list(x=list("custom",par))
```

where `par` is a list object with elements `nphi`, `mkphi`, `mkrk`, and `env`.

As an example, consider an reimplementation of the trigonometric spline of (4.63) with

```
par <- list(nphi=2,mkphi=mkphi.trig,
            mkrk=mkrk.trig,env=c(a,b))
```

where `nphi=2` specifies the dimension of $\text{span}\{\phi_\nu(x)\}$, `env=c(a,b)` specifies the domain $[a,b]$, `mkphi` takes `env` as input to create $\phi_\nu(x)$,

```
mkphi.trig <- function(env) {
  ## save constants
  env <- list(min=min(env),max=max(env))
  ## create phi
  fun <- function(x,nu,env) {
    x <- (x-env$min)/(env$max-env$min)
    switch(nu,cos(2*pi*x),sin(2*pi*x))
  }
  ## return phi and constants
  list(fun=fun,env=env)
}
```

and `mkrk` takes `env` as input to create $R_n(x,y)$,

```
mkrk.trig <- function(env) {
  ## save constants
  env <- list(min=min(env),max=max(env))
  ## create rk
  fun <- function(x,y,env,outer.prod=FALSE) {
    x <- (x-env$min)/(env$max-env$min)
    y <- (y-env$min)/(env$max-env$min)
    rk <- function(x,y) {
```

```
      k4 <- function(x) ((x-.5)^4-(x-.5)^2/2+7/240)/24
      -k4(abs(x-y))-2*cos(2*pi*(x-y))/(2*pi)^4
    }
    if (outer.prod) outer(x,y,rk)
    else rk(x,y)
  }
  ## return rk and constants
  list(fun=fun,env=env)
}
```

The precise scaling of ϕ_ν is not of much practical importance, but when `nphi` is 2 or more, the relative scaling of ϕ_ν has real implications if the variable is to be involved in interactions, as $R_p(x, y) = \sum_\nu \phi_\nu(x)\phi_\nu(y)$. When `nphi=0`, there is no parametric contrast and `mkphi` is not used.

A.2 Modeling and Data Analytical Tools

Besides the model formula and type specifications that dictate the model construction through ϕ_ν and R_β, other model components can be entered via optional arguments such as `weights`, `offset`, `partial`, and `random`. Primary data analytical tools include the Kullback-Leibler projection and the Bayesian confidence intervals.

All fitting functions but `ssanova9` accept an optional argument `weights`. For the penalized least squares regression of `ssanova` and `ssanova0`, the argument provides the w_i in (3.9) on page 64. For everything else, the argument provides the multiplicity counts of replicated observations. Weights for `ssanova9` are entered via the mandatory argument `cov`.

The optional argument `offset` is a familiar component in standard modeling suites such as `lm` and `glm`. The regression suites accept `offset`, so does `sshzd`. For (conditional) density estimation that requires normalization, offset does not make much practical sense. For the estimation of log hazard, information is rarely available to justify an offset, except for the estimation of the base hazard following the estimation of relative risk via `sscox`, as shown in §8.5.

Parametric terms can be entered through an optional argument `partial` as discussed in §4.1; it is assumed to be a formula of numerical vectors. The regression suites and hazard estimation suites accept `partial`, while the density estimation suites do not due to normalization. For a binary variable, one may either enter it through `partial` as a numerical vector or in the model formula as a factor, but a partial term can not take part in tensor products.

Parametric random effects can be entered via the optional argument `random`, which is accepted by `ssanova`, `gssanova`, `gssanova1`, `ssllrm`, and the hazard estimation suites. The algorithms of §3.4 are incompatible

TABLE A.1. Modeling and data analytical tools implemented for gss suites.

	weights	offset	partial	random	project	CI
ssanova	○	×	×	×	×	×
ssanova9		×	×		×	×
ssanova0	○	×	×			×
gssanova	×	×	×	×	×	×
gssanova1	×	×	×	×	×	×
gssanova0	×	×	×			×
ssden	×				×	
ssden1	×				×	
sscden	×				×	
sscden1	×				×	
ssllrm	×			×	×	○
sshzd	×	×	×	×	×	×
sshzd1	×		×	×	×	×
sscox	×		×	×	×	×

with (6.4) on page 218, so ssanova0 and gssanova0 can not accommodate random. The approach implemented in ssanova9 is an alternative, not in addition, to the mixed-effect models of §6.2, and weights and random are replaced in ssanova9 by the mandatory argument cov. Random effects do not make much practical sense in density estimation due to normalization, except that in ssllrm they can be propagated into versions for multivariate responses as shown in §7.8.4.

The Kullback-Leibler/square-error projection is implemented for all but ssanova0 and gssanova0 fits. The random effects, if present, are treated as an offset.

Bayesian confidence intervals can be calculated for η using the fitted values and the associated standard errors, for regression estimates and hazard estimates. For density estimation, normalization invalidates the notion of interval estimate. For hazard estimates, the fitted values returned from hzdrate.sshzd and predict.sscox are e^{η} but the standard errors are for η. For ssllrm fits, Bayesian confidence intervals only make sense for the y-contrasts as discussed in §7.8.3.

The discussions above are summarized in Table A.1, where the ×'s mark the "usual" meaning/implementation and the ○'s mark "unusual" meaning or restricted implementation. Some setting-specific entries are also worth noting, which include the argument domain for ssden and ssden1, ydomain for sscden and sscden1, and the cosine diagnostics of §3.7 for Gaussian and non-Gaussian regression fits.

A.3 Numerical Engines

The algorithms of §3.4, while highly efficient, rely on a special structure not available in general, and the legacy RKPACK routines are only used to power the `ssanova0` and `gssanova0` suites. For the other suites, computational strategies are as outlined in §3.5.3, with the likes of cross-validation scores minimized via quasi-Newton iterations using numerical derivatives.

For the density estimation and hazard estimation suites plus `gssanova`, the computation consists of two nested iteration loops, with the inner loop calculating penalized likelihood estimates with fixed tuning parameters, and the outer loop minimizing the likes of cross-validation scores for tuning parameter selection. With a single tuning parameter, the outer loop is performed through an R function `nlm0` for univariate minimization that operates on three-point quadratic interpolation with golden-section safe-guard (Gill et al. 1981, §§4.1.2.3–4.1.2.4) . With multiple tuning parameters, the outer loop is carried out via the R function `nlm` that implements the quasi-Newton algorithm of Dennis and Schnabel (1996). The inner loop Newton iteration, with safe-guards such as step-halving, is executed in FORTRAN routines, that in turn call BLAS and LINPACK routines for numerical linear algebra operations.

For the `ssanova` and `ssanova9` suites, the inner loop is unnecessary, as the penalized least squares estimates are directly available from numerical linear algebra operations. For the `gssanova1` suite, the performance-oriented iteration executes the algorithms for `ssanova` in each step.

With multiple smoothing parameters, we use Algorithm 3.3 on page 84 to obtain starting values of θ_β for quasi-Newton iteration:

1. Set $\tilde\theta_\beta^{-1} \propto \mathrm{tr}(Q_\beta)$ so that $\mathrm{tr}(\tilde\theta_\beta Q_\beta)$ contribute equally to $\mathrm{tr}(Q)$ for $Q = \sum_\beta \tilde\theta_\beta Q_\beta$, then calculate $\hat\eta = \sum_\nu \phi_\nu + \sum_\beta \eta_\beta$ with a single smoothing parameter λ, where $\eta_\beta = \tilde\theta_\beta \sum_j c_j R_\beta(z_j, \cdot)$.

2. Set $\theta_{\beta,0} \propto (\eta,\eta)_\beta = (\eta_\beta,\eta_\beta)_\beta = \tilde\theta_\beta^2\, \mathbf{c}^T Q_\beta\, \mathbf{c}$, then minimize the selection criterion with a single smoothing parameter λ at λ_0.

One then fix $\lambda = \lambda_0$ and iterate on θ_β using $\theta_{\beta,0}$ as starting values. Such a starting value algorithm is invariant of the relative scaling of R_β.

The starting value algorithm proves to be highly effective, and multivariate quasi-Newton optimization with numerical derivatives is computationally costly, so the θ iteration from $\theta_{\beta,0}$ could be chasing the "last 20 %" performance at a cost many times over the initial one. For all the fitting functions except `ssanova0`, `gssanova0`, and `ssden1`, one may choose to skip the θ iteration by setting `skip.iter=TRUE`; the skipping of the θ iteration is enforced in `ssden1` as noted in §10.1.3. In the presence of correlation parameters, however, as in the mixed-effect models or in `ssanova9`, the computational savings via `skip.iter=TRUE` could be less significant.

Appendix B
Conceptual Critiques

In this appendix, we discuss a few conceptual issues concerning nonparametric statistical models. The arguments are presented in the context of penalty smoothing, but the implications likely reach beyond. Empirical evidences in supprt of the arguments are obtained through simple simulations in the setting of penalized least squares regression.

The central issue in our discussion concerns the proper indexing of nonparametric models, and it will be argued that the usual, easy-to-work-with model indices do not properly "register" estimates based on different samples from the same source. Consequently, some widely accepted notions and perceptions are on wrong footings, and some popular practices seem misguided.

B.1 Model Indexing

Consider $Y_i = \eta(x_i) + \epsilon_i$, $x_i = (i - 0.5)/n$, $i = 1, \ldots, n$, where $n = 100$,

$$\eta(x) = 1 + 3\sin(2\pi x - \pi),$$

and $\epsilon_i \sim N(0, 1)$. One hundred replicates were generated from the setting, and for each replicate, cubic spline estimates minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1 \big(\ddot{\eta}(x)\big)^2 dx$$

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7,

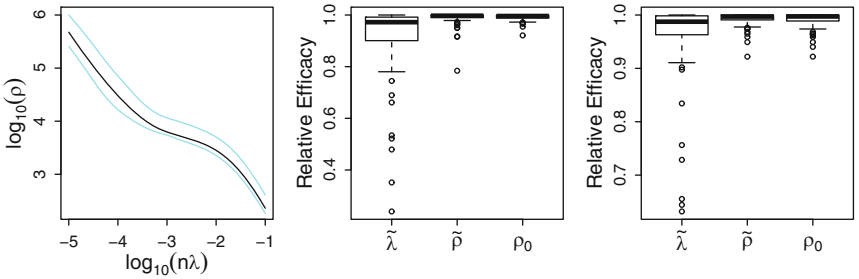


FIGURE B.1. Model indices λ and ρ. *Left*: A $\lambda \leftrightarrow \rho$ mapping (*solid*) in cubic spline simulation and the envelop (*faded*) containing one hundred such mappings. *Center*: Relative efficacy of $\tilde{\lambda}$, $\tilde{\rho}$, and ρ_0 in cubic spline simulation. *Right*: Relative efficacy of $\tilde{\lambda}$, $\tilde{\rho}$, and ρ_0 in linear spline simulation.

were calculated for λ on a grid $\log_{10}(n\lambda) = (-5)(.05)(-1)$. Recorded for each of the estimates η_λ are the mean square error

$$L(\eta, \eta_\lambda) = n^{-1} \sum_{i=1}^{n} \big(\eta_\lambda(x_i) - \eta(x_i)\big)^2 \tag{B.1}$$

and a roughness index

$$\rho = \int_0^1 \big(\ddot{\eta}_\lambda(x)\big)^2 dx.$$

Associated with the optimal η_λ on the grid that minimizes $L(\eta, \eta_\lambda)$ for each replicate, one has the optimal λ_o and the optimal ρ_o; the one hundred $\log_{10}(n\lambda_o)$ range between -3.45 and -2.25 with the median at -2.85, and the one hundred $\log_{10}\rho_o$ range between 3.776 and 3.923 with the median at 3.858. The smoothing parameter λ has no place in the data generation setting, whereas the test function $\eta(x)$ has a roughness index $\log_{10}\rho_0 = \log_{10}\big((12\pi^2)^2/2\big) = 3.846$.

Remember the equivalence between (1.1) and (1.2); see Theorem 2.12. The mapping $\lambda \leftrightarrow \rho$ is one-to-one, but the mapping varies from sample to sample. Plotted in the left frame of Fig. B.1 are one of the $\lambda \leftrightarrow \rho$ mappings from the simulation (solid) and the envelop containing all one hundred such mappings (faded). The envelop is not too wide so rates of λ and ρ should be comparable across-replicates, but with exact quantification, at most one of λ and ρ can be used to "register" estimates based on different replicates.

The much tighter range of ρ_o as compared to the range of λ_o is not quite enough to put ρ over λ, as one could argue that the scales of λ and ρ may not be comparable. Instead, we set the median $\log_{10}(n\tilde{\lambda}) = -2.85$ as a "typical" optimal λ value and the median $\log_{10}\tilde{\rho} = 3.858$ as a "typical" optimal ρ value, and assess the relative efficacy of these choices. The relative efficacy of $\tilde{\lambda}$ is simply $L(\eta, \eta_{\lambda_o})/L(\eta, \eta_{\tilde{\lambda}})$, where λ_o varies from replicate to replicate. For $\tilde{\rho}$, we have to settle with approximations, using for each replicate the

estimate on the λ grid that has the smallest $|\log_{10}\rho - 3.858|$. The relative efficacy of $\tilde{\lambda}$ and $\tilde{\rho}$ are summarized in the center frame of Fig. B.1 along with that of ρ_0.

For the same one hundred replicates of simulated data, we also calculated linear spline estimates minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda\int_0^1 \big(\dot{\eta}(x)\big)^2 dx$$

for λ on a grid $\log_{10}(n\lambda) = (-2.5)(.05)(1.5)$. The roughness index is now

$$\rho = \textstyle\int_0^1 \big(\dot{\eta}_\lambda(x)\big)^2 dx,$$

with $\log_{10}\rho_0 = \log_{10}\big((6\pi)^2/2\big) = 2.250$. The corresponding $\log_{10}(n\lambda_o)$ have a range of $[-1.2, -0.5]$ with the median at $\log_{10}(n\tilde{\lambda}) = -0.9$, and the corresponding $\log_{10}\rho_o$ have a range of $[2.181, 2.356]$ with the median at $\log_{10}\tilde{\rho} = 2.255$. The relative efficacy of such $\tilde{\lambda}$, $\tilde{\rho}$, and ρ_0 are shown in the right frame of Fig. B.1.

Statistical estimation is a compromise between the data and the model, where the model is best characterized by a set of constraints. Model constraints are clearly spelled out in standard parametric models, but are vague or implicit at best with nonparametric estimation. The equivalence between penalized and constrained optimizations provides a means for one to study the subtle issue of model indexing in the context of penalty smoothing, and the empirical results shown in the center and right frames of Fig. B.1 confirm the fact that, across-replicates, estimates with the same ρ have more in common than estimates with the same λ.

While ρ is the conceptually "correct" model index, it is impossible to work with in practice, both in numerical computation and in theoretical analysis. Throughout this book, we have worked exclusively with λ, and the results remain valid, for they either concern only rates but not exact quantifications, or they are replicate-specific so the mapping $\lambda \leftrightarrow \rho$ is one-to-one in the context, or both. The ρ index appears useless operationwise, but it can help to explain a few "mysterious" phenomena that led to misguided perceptions and practices in the literature.

B.2 Optimal and Cross-Validation Indices

Despite the asymptotic optimality established by Li (1986) and the largely excellent empirical performances in simulations and applications, cross-validation had over the years received its share of criticisms in the literature. Some of the concerns are valid, such as the occasional wild failures, which can be tamed by the use of a fudge factor. Other concerns mainly involve

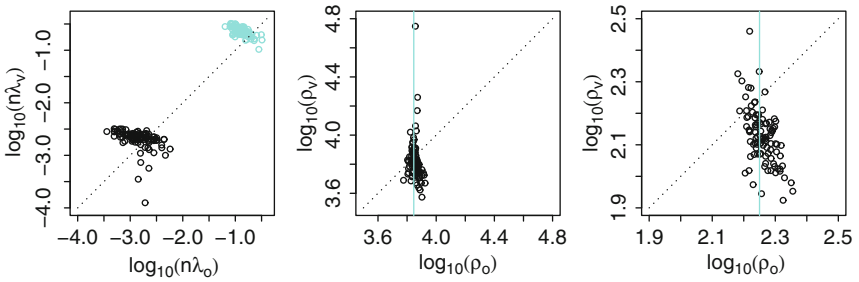


FIGURE B.2. Optimal and cross-validation λ and ρ. *Left*: λ_o versus λ_v in cubic spline simulation (*solid*) and in linear spline simulation (*faded*). *Center*: ρ_o versus ρ_v in cubic spline simulation. *Right*: ρ_o versus ρ_v in linear spline simulation. The *vertical faded lines* in the center and right frames mark the respective ρ_0.

the "unfavorable" behaviors of the minimizers of the cross-validation scores, which can be misperceived.

For each of the estimates in the simulations of §B.1, also recorded are the cross-validation score with a fudge factor $\alpha = 1.4$,

$$V(\lambda) = \frac{n^{-1}\mathbf{Y}^T\big(I - A(\lambda)\big)^2\mathbf{Y}}{\big\{n^{-1}\mathrm{tr}\big(I - \alpha A(\lambda)\big)\big\}^2}.$$

Associated with the η_λ that minimizes $V(\lambda)$ on the grid for each replicate, one has the cross-validation indices λ_v and ρ_v. Plotted in the left frame of Fig. B.2 are λ_o versus λ_v for the one hundred replicates in the cubic spline simulation (solid) and in the linear spline simulation (faded), where the negative correlation between λ_o and λ_v is evident. Such negative correlation was well publicized in the literature concerning a few versions of cross-validation scores in various settings, and in light of this, cross-validation was charged as acting "counter-intuitively," prompting the developments of alternative approaches to smoothing parameter selection; see, e.g., Scott and Terrell (1987) and Hall and Johnstone (1992).

Were the λ index comparable across-replicates, such negative correlation would indeed signal trouble. Given the discussion of §B.1, however, the negative correlation in λ is inconsequential. Plotted in the center and right frames of Fig. B.2 are the respective ρ_o versus ρ_v in the cubic and linear spline simulations, where negative correlation is nowhere to be found.

Further discussions on this and related issues can be found in Gu (1998a).

B.3 Loss, Risk, and Smoothing Parameter Selection

The mean square error $L(\lambda) = L(\eta, \eta_\lambda)$ of (B.1) is a replicate-specific loss function, and the optimal indices λ_o and ρ_o vary from replicate to replicate. If one must take expectation of the loss, lining up estimates with

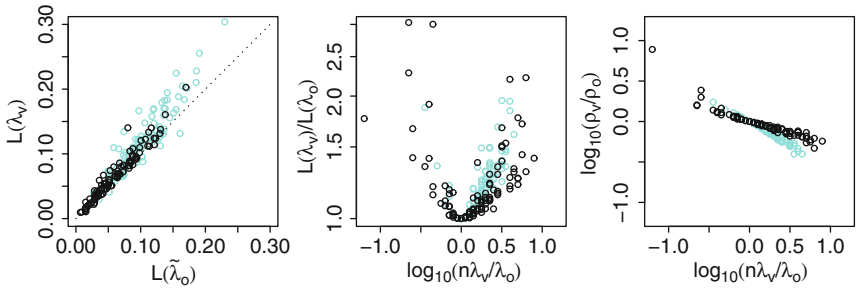


FIGURE B.3. Optimal and cross-validation λ and ρ. *Left*: $L(\tilde{\lambda}_o)$ versus $L(\lambda_v)$. *Center*: λ_v/λ_o versus $L(\lambda_v)/L(\lambda_o)$. *Right*: λ_v/λ_o versus ρ_v/ρ_o. Results from cubic spline simulation are in *solid* and those from linear spline simulation in *faded*.

common ρ appears conceptually "correct" though practically impossible to perform, whereas expectation with fixed-λ is effectively mixing oranges with tangerines and grapefruits.

If risk could be calculated with fixed-ρ, then from the center and right frames of Fig. B.1, one could infer that the risk-optimal ρ would largely match the performance of the loss-optimal ρ_o. The λ-indexed risk function $R(\lambda) = E\big[L(\lambda)\big]$ has its conceptual flaws, and we now evaluate it empirically. Averaging over the replicates in the simulations of §B.1, we obtained empirical versions of $R(\lambda)$, whose minimizers on the grids gave the "risk-optimal" $\log_{10}(n\tilde{\lambda}_o) = -2.8 \approx -2.85 = \log_{10}(n\tilde{\lambda})$ for cubic spline estimates and $\log_{10}(n\tilde{\lambda}_o) = -0.9 = \log_{10}(n\tilde{\lambda})$ for linear spline estimates. Plotted in the left frame of Fig. B.3 are $L(\tilde{\lambda}_o)$ versus $L(\lambda_v)$ in the cubic spline simulation (solid) and in the linear spline simulation (faded). The "risk-optimal" $\tilde{\lambda}_o$ did do better, but was helped by extra knowledge unknown to cross-validation. The very existence of points below the dotted line, 19 solid and 18 faded, speaks to the fact that $\tilde{\lambda}_o$ is *not* optimal. It is one thing to calculate the rate of $L(\lambda)$ via $R(\lambda)$, as was done in the asymptotics of §3.2, but it is a different matter to define the notion of optimality through the exact minimization of $R(\lambda)$.

Merit-wise, the loss $L(\lambda)$ is no doubt more appealing than the risk $R(\lambda)$ as the performance measure, but questions were raised in the literature concerning the practical feasibility of pursuing $L(\lambda)$, with the main argument being the slow convergence rates of the likes of $\hat{\lambda}_o - \lambda_o$; see, e.g., Hall and Marron (1991). We however shall argue below that the slow convergence of $\hat{\lambda}_o - \lambda_o$ could be as inconsequential as the negative correlation between λ_o and λ_v as seen in the left frame of Fig. B.2.

Aiming to minimize $L(\lambda)$ via a selection method such as cross-validation, the success/failure of the method is naturally assessed through the likes of relative efficacy $L(\lambda_o)/L(\lambda_v)$. The loss curve could be steep or flat near λ_o,

and could have different slopes on different sides of λ_o, thus the difference $\lambda_v - \lambda_o$ could be a poor proxy of $L(\lambda_o)/L(\lambda_v)$. Furthermore, the λ index has no place in the data generation setting, and the optimal λ_o assumes its meaning only via the loss function $L(\lambda)$, so its "estimation" accuracy should also be assessed through $L(\lambda)$. Shown in the center frame of Fig. B.3 are $L(\lambda_v)/L(\lambda_o)$ versus λ_v/λ_o in the simulations, where the distance between λ_v and λ_o is measured on the more natural log scale; $L(\lambda_v)/L(\lambda_o)$ does generally increase as λ_v moves away from λ_o, as expected, but the exact quantification is far too scattered for λ_v/λ_o to be a reliable proxy of $L(\lambda_v)/L(\lambda_o)$. Plotted in the right frame of Fig. B.3 are λ_v/λ_o versus ρ_v/ρ_o in the simulations, showing that the raw distance between "λ_o" and "$\hat{\lambda}_o$" may also depend on the particular "λ" (model index) in use.

In summary, a "risk-optimal" λ based on $R(\lambda)$ has its flaws conceptually and empirically, and the slow convergence rates of the likes of $\hat{\lambda}_o - \lambda_o$ may not have any bearing on the practical feasibility of targeting $L(\lambda)$ in smoothing parameter selection. In fact, the asymptotic optimality of cross-validation in terms of losses, as discussed in §§3.2, 6.2.3, and 6.3.3, provide direct, positive solutions to loss-minimizing smoothing parameter selection.

B.4 Degrees of Freedom

Model constraints in nonparametric estimation are intrinsically adaptive and typically also implicit, whereas those in parametric models are prespecified explicitly. Despite the fundamental difference, numerous attempts have been made in the literature to extend familiar notions and practices in parametric statistics to nonparametric estimation. One popular notion of such is the so-called "degrees of freedom" as a model complexity index in nonparametric regression, which we shall scrutinize below.

Recall the smoothing matrix $A(\lambda)$ introduced in Chap. 3 satisfying $\hat{\mathbf{Y}} = A(\lambda)\mathbf{Y}$, which resembles the hat matrix $H = X(X^TX)^{-1}X^T$ for a linear regression model $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The trace of the smoothing matrix, $\mathrm{tr}A(\lambda)$, was deemed by many as the "effective number of parameters," or the "degrees of freedom," and suggestion was made to possibly select the smoothing parameters by specifying the "degrees of freedom;" see, e.g., Hastie and Tibshirani (1990).

Write $\nu = \mathrm{tr}A(\lambda)$. Given the sampling points x_i, the mapping $\lambda \leftrightarrow \nu$ is one-to-one, independent of Y_i, so ν is simply a reparameterization of λ. The trace of a matrix is much more intuitive than a λ in front of a roughness penalty, however, and the smoothing matrix can be defined for all nonparametric regression methods, so ν appears to provide an intuitive, universal index for model complexity. Unfortunately, the very appeal of the ν index is where it falters.

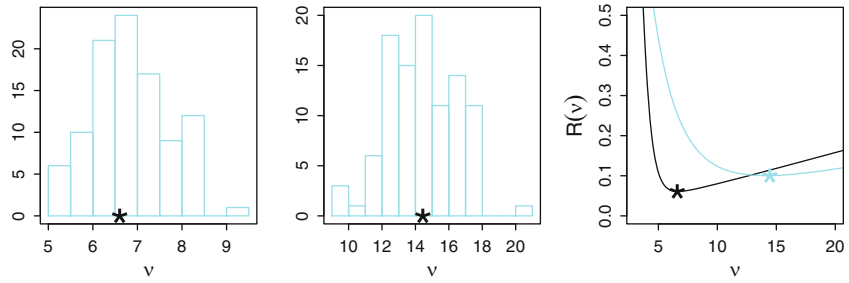


FIGURE B.4. Loss-optimal ν_o and empirical risk $R(\nu)$. *Left*: Histogram of ν_o in cubic spline simulation. *Center*: Histogram of ν_o in linear spline simulation. *Right*: $R(\nu)$ in cubic (*solid*) and linear (*faded*) spline simulations. The *stars* in the left and center frames mark the respective "risk-optimal" $\tilde{\nu}_o$; the stars in the right frame mark the minima.

Recall the simulations of §B.1, where for cubic splines $\log_{10}(n\lambda_o)$ range between -3.45 and -2.25 with the median $\log_{10}(n\tilde{\lambda}) = -2.85$, which translate into a ν_o range of $[5.08, 9.14]$ and the median $\tilde{\nu} = 6.77$; for linear splines, the $\log_{10}(n\lambda_o)$ range $[-3.45, -2.25]$ corresponds to a ν_o range of $[9.35, 20.00]$ and the median $\log_{10}(n\tilde{\lambda}) = -0.9$ to $\tilde{\nu} = 14.44$. Histograms of ν_o are shown in the left and center frames of Fig. B.4. Depicted in the right frame of Fig. B.4 are the empirical risk functions $R(\nu)$ indexed by ν in the cubic and linear spline simulations. Within the respective families of estimates, namely the cubic splines and the linear splines, the ν index is equivalent to the λ index, sharing its conceptual flaws but offering nothing new. Across different families of estimates, it is hard to reconcile a "cubic-spline-optimal" $\nu \approx 7$ with a "linear-spline-optimal" $\nu \approx 14$; the "risk-optimal" $\tilde{\nu}_o$ are 6.60 and 14.44 in the cubic and linear spline simulations, respectively, corresponding to $\log_{10}(n\tilde{\lambda}_o)$ values of -2.8 and -0.9. When the "optimal" values are territory-dependent, an index perceived to be "universal" only serves to mislead.

In parametric statistics, the degrees of freedom code the dimensions of the prospective model spaces. The notion is not defined through the trace of any matrix, and in many settings there is no matrix to talk about yet degrees of freedom are indispensable in inference. The fact that the trace of the hat matrix in linear regression models matches the dimension of the model space is conceptually a coincidence. In the context of nonparametric regression, model complexity depends on a variety of factors including the structure of the smoothing matrix, but loading everything on a matrix trace oversimplifies the matter.

References

Abramowitz M, Stegun IA (1964) Handbook of mathematical functions with formulas, graphs, and mathematical tables. National Bureau of Standards, Washington, DC.

Akhiezer NI, Glazman IM (1961) Theory of linear operators in Hilbert space. Ulgar, New York.

Anderson JA (1972) Separate sampling logistic regression. Biometrika 59:19–35.

Anderson JA, Blair V (1982) Penalized maximum likelihood estimation in logistic regression and discrimination. Biometrika 69:123–136.

Anderson JA, Senthilselvan A (1980) Smooth estimates for the hazard function. J R Stat Soc Ser B 42:322–327.

Ansley CF, Kohn R, Wong C-M (1993) Nonparametric spline regression with prior information. Biometrika 80:75–88.

Antoniadis A (1989) A penalty method for nonparametric estimation of the intensity function of a counting process. Ann Inst Stat Math 41:781–807.

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68:337–404.

Azzalini A, Bowman AW (1990) A look at some data on the Old Faithful geyser. Appl Stat 39:357–365.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7,

Barry D (1986) Nonparametric Bayesian regression. Ann Stat 14:934–953.

Bartoszyński R, Brown BW, McBride CM, Thompson JR (1981) Some nonparametric techniques for estimating the intensity function of a cancer related nonstationary Poisson process. Ann Stat 9:1050–1060.

Bates DM, Lindstrom M, Wahba G, Yandell B (1987) GCVPACK – routines for generalized cross validation. Commun Stat Simul Comput 16:263–297.

Becker RA, Chambers JM, Wilks AR (1988) The new S language. Wadsworth & Brooks/Cole, Pacific Grove.

Berger JO (1985) Statistical decision theory and Bayesian analysis, 2nd edn. Springer, New York.

Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis, 3rd edn. Prentice Hall, Englewood Cliffs.

Breiman L (1991) The $\prod$ method for estimating multivariate functions from noisy data. Technometrics 33:125–160 (with discussions).

Breiman L, Friedman JH (1985) Estimating optimal transformations for multiple regression and correlation. J Am Stat Assoc 80:580–598 (with discussions).

Breslow NE, Clayton DG (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25.

Brinkman ND (1981) Ethanol fuel – a single-cylinder engine study of efficiency and exhaust emissions. SAE Trans 90:1410–1424.

Brockwell PJ, Davis RA (1991) Time series: theory and methods, 2nd edn. Springer, New York.

Buja A, Hastie TJ, Tibshirani RJ (1989) Linear smoothers and additive models. Ann Stat 17:453–555 (with discussions).

Byerly WE (1959) An elementary treatise on Fourier's series and spherical, cylindrical, and ellipsoidal harmonics. Dover, New York. Reprint of the 1893 original.

Chambers JM, Hastie TJ (eds) (1992) Statistical models in S. Chapman & Hall, New York.

Chen Z (1991) Interaction spline models and their convergence rates. Ann Stat 19:1855–1868.

Chhikara RS, Folks JL (1989) The inverse Gaussian distribution: theory, methodology, and applications. Marcel Dekker, New York.

Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 83:829–836.

Cleveland WS, Devlin SJ (1988) Locally weighted regression: an approach to regression analysis by local fitting. J Am Stat Assoc 83:596–610.

Cogburn R, Davis HT (1974) Periodic splines and spectral estimation. Ann Stat 2:1108–1126.

Cole TJ (1988) Fitting smoothed centile curves to reference data. J R Stat Soc Ser A 151:385–418 (with discussions).

Cole TJ, Green PJ (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. Stat Med 11:1305–1319.

Cosslett SR (1981) Maximum likelihood estimator for choice-based samples. Econometrika 49:1289–1316.

Cox DR (1969) Some sampling problems in technology. In: Johnson NL, Smith Jr H (eds) New developments in survey sampling. Wiley, New York, pp 506–527.

Cox DR (1972) Regression models and life tables. J R Stat Soc Ser B 34:187–220 (with discussions).

Cox DD (1984) Multivariate smoothing spline functions. SIAM J Numer Anal 21:789–813.

Cox DD (1988) Approximation of method of regularization estimators. Ann Stat 16:694–712.

Cox DD, Chang Y (1990) Iterated state space algorithms and cross validation for generalized smoothing splines. Technical Report 49, Department of Statistics, University of Illinois, Champion.

Cox DD, O'Sullivan F (1990) Asymptotic analysis of penalized likelihood and related estimators. Ann Stat 18:124–145.

Craven P, Wahba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403.

Crowley J, Hu M (1977) Covariance analysis of heart transplant survival data. J Am Stat Assoc 72:27–36.

Dalzell CJ, Ramsay JO (1993) Computing reproducing kernels with arbitrary boundary constraints. SIAM J Sci Comput 14:511–518.

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

de Boor C, Lynch R (1966) On splines and their minimum properties. J Math Mach 15:953–969.

Dennis JE, Schnabel RB (1996) Numerical methods for unconstrained optimization and nonlinear equations. SIAM, Philadelphia. Corrected reprint of the 1983 original.

Dongarra JJ, Moler CB, Bunch JR, Stewart GW (1979) LINPACK user's guide. SIAM, Philadelphia.

Douglas A, Delampady M (1990) Eastern lake survey – phase I: documentation for the data base and the derived data sets. Technical Report

160 (SIMS), Department of Statistics, University of British Columbia, Vancouver.

Du P, Gu C (2006) Penalized likelihood hazard estimation: efficient approximation and bayesian confidence intervals. Stat Probab Lett 76:244–254.

Du P, Gu C (2009) Penalized pseudo-likelihood hazard estimation: a fast alternative to penalized likelihood. J Stat Plan Inference 139:891–899.

Du P, Ma S (2010) Frailty model with spline estimated nonparametric hazard function. Stat Sin 20:561–580.

Duchon J (1977) Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: Schemp W, Zeller K (eds) Constructive theory of functions of several variables. Springer, Berlin, pp 85–100.

Elden L (1984) A note on the computation of the generalized cross validation function for ill-conditioned least square problems. BIT 24:467–472.

Fan J, Gijbels I (1996) Local polynomial modelling and its applications. Chapman & Hall, London.

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, New York.

Gao F (1999) Penalized multivariate logistic regression with a large data set. Ph.D. thesis, University of Wisconsin, Madison.

Gao F, Wahba G, Klein R, Klein BEK (2001) Smoothing spline ANOVA for multivariate Bernoulli observations, with application to ophthalmology data. J Am Stat Assoc 96:127–160 (with discussions).

Gill RD (1984) Understanding Cox's regression model: a martingale approach. J Am Stat Assoc 79:441–447.

Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic, New York.

Gill RD, Vardi Y, Wellner JA (1988) Large sample theory of empirical distributions in biased sampling models. Ann Stat 16:1069–1112.

Girard DA (1989) A fast "Monte-Carlo cross validation" procedure for large least squares problems with noisy data. Numer Math 56:1–23.

Girard DA (1991) Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. Ann Stat 19:1950–1963.

Golub G, Van Loan C (1989) Matrix computations, 2nd edn. The Johns Hopkins University Press, Baltimore.

Good IJ, Gaskins RA (1971) Nonparametric roughness penalties for probability densities. Biometrika 58:255–277.

Gray RJ (1992) Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. J Am Stat Assoc 87:942–951.

Green PJ, Yandell B (1985) Semi-parametric generalized linear models. In: Gilchrist R, Francis B, Whittaker J (eds) Proceedings of the GLIM85 conference. Springer, Berlin, pp 44–55.

Green PJ, Silverman BW (1994) Nonparametric regression and generalized linear models. Chapman & Hall, London.

Gu C (1989) RKPACK and its applications: fitting smoothing spline models. In: ASA proceedings of statistical computing section. ASA, Alexandria pp 42–51.

Gu C (1990) Adaptive spline smoothing in non Gaussian regression models. J Am Stat Assoc 85:801–807.

Gu C (1992a) Cross-validating non-Gaussian data. J Comput Graph Stat 1:169–179.

Gu C (1992b) Diagnostics for nonparametric regression models with additive term. J Am Stat Assoc 87:1051–1058.

Gu C (1992c) Penalized likelihood regression: a Bayesian analysis. Stat Sin 2:255–264.

Gu C (1992d) Smoothing spline density estimation: biased sampling and random truncation. Technical Report 92–03, Department of Statistics, Purdue University, West Lafayette.

Gu C (1993a) Interaction splines with regular data: automatically smoothing digital images. SIAM J Sci Comput 14:218–230.

Gu C (1993b) Smoothing spline density estimation: a dimensionless automatic algorithm. J Am Stat Assoc 88:495–504.

Gu C (1994) Penalized likelihood hazard estimation: algorithm and examples. In: Berger JO, Gupta SS (eds) Statistical decision theory and related topics, V. Springer, New York, pp 61–72.

Gu C (1995a) Smoothing spline density estimation: conditional distribution. Stat Sin 5:709–726.

Gu C (1995b) Smoothing spline density estimation: response-based sampling. Technical Report 267, Department of Statistics, University of Michigan, Ann Arbor.

Gu C (1996) Penalized likelihood hazard estimation: a general procedure. Stat Sin 6:861–876.

Gu C (1998a) Model indexing and smoothing parameter selection in nonparametric function estimation. Stat Sin 8:607–646 (with discussions).

Gu C (1998b) Penalized likelihood estimation: convergence under incorrect model. Stat Probab Lett 36:359–364.

Gu C (1998c) Structural multivariate function estimation: some automatic density and hazard estimates. Stat Sin 8:317–335.

Gu C (2004) Model diagnostics for smoothing spline ANOVA models. Can J Stat 32:347–358.

Gu C (2011) Practical nonparametric conditional density estimation. Preprint.

Gu C, Kim Y-J (2002) Penalized likelihood regression: general formulation and efficient approximation. Can J Stat 30:619–628.

Gu C, Ma P (2005a) Generalized nonparametric mixed-effect models: computation and smoothing parameter selection. J Comput Graph Stat 14:485–504.

Gu C, Ma P (2005b) Optimal smoothing in nonparametric mixed-effect models. Ann Stat 33:1357–1379.

Gu C, Ma P (2011) Nonparametric regression with cross-classified responses. Can J Stat 39:591–609.

Gu C, Qiu C (1993) Smoothing spline density estimation: theory. Ann Stat 21:217–234.

Gu C, Qiu C (1994) Penalized likelihood regression: a simple asymptotic analysis. Stat Sin 4:297–304.

Gu C, Wahba G (1991a) Discussion of "multivariate adaptive regression splines" by J. Friedman. Ann Stat 19:115–123.

Gu C, Wahba G (1991b) Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. SIAM J Sci Stat Comput 12:383–398.

Gu C, Wahba G (1993a) Semiparametric analysis of variance with tensor product thin plate splines. J R Stat Soc Ser B 55:353–368.

Gu C, Wahba G (1993b) Smoothing spline ANOVA with component-wise Bayesian "confidence intervals". J Comput Graph Stat 2:97–117.

Gu C, Wang J (2003) Penalized likelihood density estimation: direct cross-validation and scalable approximation. Stat Sin 13:811–826.

Gu C, Xiang D (2001) Cross-validating non-Gaussian data: generalized approximate cross-validation revisited. J Comput Graph Stat 10:581–591.

Gu C, Bates DM, Chen Z, Wahba G (1989) The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. SIAM J Matrix Anal Appl 10:457–480.

Gu C, Heckman N, Wahba G (1992) A note on generalized cross-validation with replicates. Stat Probab Lett 14:283–287.

Gu C, Jeon Y, Lin Y (2013) Nonparametric density estimation in high dimensions. Stat Sin 23:000-000.

Hall P, Johnstone I (1992) Empirical functionals and efficient smoothing parameter selection. J R Stat Soc Ser B 54:475–530 (with discussions).

Hall P, Marron JS (1991) Lower bounds for bandwidth selection in density estimation. Probab Theory Relat Fields 90:149–173.

Han C, Gu C (2008) Optimal smoothing with correlated data. Sankhya 70-A:38–72.

Härdle W (1991) Smoothing techniques with implementation in S. Springer, New York.

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–340 (with discussions).

Hastie TJ, Tibshirani RJ (1986) Generalized additive models. Stat Sci 1:297–318 (with discussions).

Hastie TJ, Tibshirani RJ (1990) Generalized additive models. Chapman & Hall, London.

Heckman NE, Ramsay JO (2000) Penalized regression with model-based penalties. Can J Stat 28:241–258.

Hornik K (2010) The R FAQ. ISBN 3-900051-08-9.

Hutchinson MF (1989) A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. Commun Stat Simul Comput 18:1059–1076.

Hutchinson MF, de Hoog FR (1985) Smoothing noisy data with spline functions. Numer Math 47:99–106.

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314.

Jeon Y, Lin Y (2006) An effective method for high dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. Stat Sin 16:353–374.

Joe H (1997) Multivariate models and dependence concepts. Chapman & Hall, London.

Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis, 3rd edn. Prentice Hall, Englewood Cliffs.

Jones RH (1980) Maximum likelihood fitting of ARMA models to time series with missing observations. Technometrics 20:389–395.

Jones MC (1991) Kernel density estimation for length biased data. Biometrika 78:511–520.

Kalbfleisch JD, Lawless JF (1989) Inference based on retrospective ascertainment: an analysis of the data on transfusion-related AIDS. J Am Stat Assoc 84:360–372.

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley, New York.

Kaplan EL, Meier P (1958) Nonparametric estimator from incomplete observations. J Am Stat Assoc 53:457–481.

Karcher P, Wang Y (2001) Generalized nonparametric mixed effects models. J Comput Graph Stat 10:641–655.

Karlin S, Studden WJ (1966) Tchebycheff systems: with applications in analysis and statistics. Wiley, New York/London/Sydney.

Keiding N, Gill RD (1990) Random truncation models and markov processes. Ann Stat 18:582–602.

Kernighan BW (1975) Ratfor – a preprocessor for a rational Fortran. In: UNIX programmer's manual. Bell Laboratories, Murray Hill.

Kim Y-J (2003) Smoothing spline regression: scalable computation and cross-validation. Ph.D. thesis, Purdue University, West Lafayette.

Kim Y-J, Gu C (2004) Smoothing spline Gaussian regression: more scalable computation via efficient approximation. J R Stat Soc Ser B 66:337–356.

Kimeldorf G, Wahba G (1970a) A correspondence between Bayesian estimation of stochastic processes and smoothing by splines. Ann Math Stat 41:495–502.

Kimeldorf G, Wahba G (1970b) Spline functions and stochastic processes. Sankhya Ser A 32:173–180.

Kimeldorf G, Wahba G (1971) Some results on Tchebycheffian spline functions. J Math Anal Appl 33:82–85.

Klein R, Klein BEK, Moss SE, Davis MD, DeMets DL (1988) Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. J Am Med Assoc 260:2864–2871.

Klein R, Klein BEK, Moss SE, Davis MD, DeMets DL (1989) The Wisconsin epidemiologic study of diabetic retinopathy. X. Four incidence and progression of diabetic retinopathy when age at diagnosis is 30 or more years. Arch Ophthalmol 107:244–249.

Koenker R, Ng P, Portnoy S (1994) Quantile smoothing splines. Biometrika 81:673–680.

Kooperberg C, Bose S, Stone CJ (1997) Polychotomous regression. J Am Stat Assoc 92:117–127.

Lauritzen SL (1996) Graphical models. Oxford University Press, New York.

Lee Y, Nelder JA (1996) Hierarchical generalized linear models. J R Stat Soc Ser B 58:619–678 (with discussions).

Lehmann EL, Casella G (1998) Theory of point estimation, 2nd edn. Springer, New York.

Leonard T (1978) Density estimation, stochastic processes and prior information. J R Stat Soc Ser B 40:113–146 (with discussions).

Leonard T, Hsu JSJ, Tsui K-W (1989) Bayesian marginal inference. J Am Stat Assoc 84:1051–1058.

Li K-C (1986) Asymptotic optimality of C_L and generalized cross-validation in the ridge regression with application to spline smoothing. Ann Stat 14:1101–1112.

Lin X (1998) Smoothing spline analysis of variance for polychotomous response data. Ph.D. thesis, University of Wisconsin, Madison.

Lin Y (2000) Tensor product space ANOVA models. Ann Stat 28:734–755.

Lin X, Zhang D (1999) Inference in generalized additive mixed models by using smoothing splines. J R Stat Soc Ser B 61:381–400.

Lindsey JK (1997) Applying generalized linear models. Springer, New York.

Mallows CL (1973) Some comments on C_P. Technometrics 15:661–675.

Manski CF (1995) Identification problems in the social sciences. Harvard University Press, Cambridge.

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, London.

McCulloch CE (1997) Maximum likelihood algorithms for generalized linear mixed models. J Am Stat Assoc 92:162–170.

Meinguet J (1979) Multivariate interpolation at arbitrary points made simple. J Appl Math Phys (ZAMP) 30:292–304.

Miller RG (1976) Least squares regression with censored data. Biometrika 63:449–464.

Miller RG, Halpern J (1982) Regression with censored data. Biometrika 69:521–531.

Moreau T, O'Quigley J, Mesbah M (1985) A global goodness-of-fit statistic for the proportional hazards model. Appl Stat 34:212–218.

Morgenthaler S, Vardi Y (1986) Choice-based samples: a non-parametric approach. J. Econom 32:109–125.

Novak E, Ritter K (1996) High dimensional integration of smooth functions over cubes. Numer Math 75:79–97.

Nychka D (1988) Bayesian confidence intervals for smoothing splines. J Am Stat Assoc 83:1134–1143.

O'Sullivan F (1985) Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. Silverman J R Stat Soc Ser B 47:39–40.

O'Sullivan F (1988a) Fast computation of fully automated log-density and log-hazard estimators. SIAM J Sci Stat Comput 9:363–379.

O'Sullivan F (1988b) Nonparametric estimation of relative risk using splines and cross-validation. SIAM J Sci Stat Comput 9:531–542.

O'Sullivan F, Yandell B, Raynor W (1986) Automatic smoothing of regression functions in generalized linear models. J Am Stat Assoc 81:96–103.

Ouyang Z, Zhou Q, Wong WH (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci USA 106:21521–21526. doi:10.1073/pnas.0904863106.

Parzen E (1979) Nonparametric statistical data modeling. J Am Stat Assoc 74:105–131 (with discussions).

Pawitan Y, O'Sullivan F (1994) Nonparametric spectral density estimation using penalized Whittle likelihood. J Am Stat Assoc 89:600–610.

Petras K (2001) Asymptotically minimal smolyak cubature. Preprint.

Prentice RL, Pyke R (1978) Logistic disease incidence models and case-control studies. Biometrika 66:403–411.

Priestley MB (1981) Spectral analysis and times series. Academic, London.

Ramsay JO, Dalzell CJ (1991) Some tools for functional data analysis. J R Stat Soc Ser B 53:539–572 (with discussions).

Rao CR (1973) Linear statistical inference and its applications. Wiley, New York.

Robinson GK (1991) That BLUP is a good thing: the estimation of the random effects. Stat Sci 6:15–51 (with discussions).

Scheffe H (1959) The analysis of variance. Wiley, New York.

Schoenberg IJ (1964) Spline functions and the problem of graduation. Proc Natl Acad Sci USA 52:947–950.

Schumaker L (1981) Spline functions: basic theory. Wiley, New York.

Scott DW (1985) Averaged shifted histograms: effective nonparametric density estimators in several dimensions. Ann Stat 13:1024–1040.

Scott DW (1992) Multivariate density estimation: theory, practice and visualization. Wiley, New York.

Scott DW, Terrell GR (1987) Biased and unbiased cross-validation in density estimation. J Am Stat Assoc 82:1131–1146.

Scott AJ, Wild CJ (1986) Fitting logistic models under case-control and choice-based sampling. J R Stat Soc Ser B 48:170–182.

Seber GAF (1977) Linear regression analysis. Wiley, New York.

Silverman BW (1978) Density ratios, empirical likelihood and cot death. Appl Stat 27:26–33.

Silverman BW (1982) On the estimation of a probability density function by the maximum penalized likelihood method. Ann Stat 10:795–810.

Snyder DL (1975) Random point processes. Wiley, New York.

Stein ML (1993) Spline smoothing with an estimated order parameter. Ann Stat 21:1522–1544.

Stewart GW (1987) Collinearity and least square regression. Stat Sci 2:68–100 (with discussions).

Stone CJ (1985) Additive regression and other nonparametric models. Ann Stat 13:689–705.

Stone CJ, Hansen MH, Kooperberg C, Truong YK (1997) Polynomial splines and their tensor products in extended linear modeling. Ann Stat 25:1371–1470 (with discussions).

Tapia RA, Thompson JR (1978) Nonparametric probability density estimation. Johns Hopkins University Press, Baltimore.

Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81:82–86.

Tong H (1990) Nonlinear time series. Oxford University Press, New York.

Turnbull BW, Brown BW, Hu M (1974) Survivorship analysis of heart transplant data. J Am Stat Assoc 69:74–80.

Utreras F (1981) Optimal smoothing of noisy data using spline functions. SIAM J Sci Stat Comput 2:349–362.

Utreras F (1983) Natural spline functions: their associated eigenvalue problem. Numer Math 42:107–117.

Utreras F (1988) Convergence rates for multivariate smoothing spline functions. J Approx Theory 52:1–27.

Vardi Y (1982) Nonparametric estimation in the presence of length bias. Ann Stat 10:616–620.

Vardi Y (1985) Empirical distributions in selection bias models. Ann Stat 13:178–203.

Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York.

Wahba G (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. J R Stat Soc Ser B 40:364–372.

Wahba G (1980) Automatic smoothing of the log periodogram. J Am Stat Assoc 75:122–132.

Wahba G (1981) Spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 2:5–16.

Wahba G (1982) Erratum: spline interpolation and smoothing on the sphere. SIAM J Sci Stat Comput 3:385–386.

Wahba G (1983) Bayesian "confidence intervals" for the cross-validated smoothing spline. J R Stat Soc Ser B 45:133–150.

Wahba G (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann Stat 13:1378–1402.

Wahba G (1986) Partial and interaction spline models for the semiparametric estimation of functions of several variables. In: Computer science and statistics: proceedings of the 18th symposium on the interface. ASA, Washington, DC, pp 75–80.

Wahba G (1990) Spline models for observational data. CBMS-NSF regional conference series in applied mathematics, vol 59. SIAM, Philadelphia.

Wahba G, Wendelberger J (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. Mon Weather Rev 108:1122–1145.

Wahba G, Wang Y, Gu C, Klein R, Klein BEK (1995) Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. Ann Stat 23:1865–1895.

Wang M-C (1989) A semiparametric model for randomly truncated data. J Am Stat Assoc 84:742–748.

Wang Y (1997) GRKPACK: fitting smoothing spline ANOVA models for exponential families. Commun Stat Simul Comput 26:765–782.

Wang Y (1998a) Mixed-effects smoothing spline ANOVA. J R Stat Soc Ser B 60:159–174.

Wang Y (1998b) Smoothing spline models with correlated random errors. J Am Stat Assoc 93:341–348.

Wang Y, Brown MB (1996) A flexible model for human circadian rhythms. Biometrics 52:588–596.

Wang M-C, Jewell NP, Tsay W-Y (1986) Asymptotic properties of the product limit estimate under random truncation. Ann Stat 14:1597–1605.

Weinberger HF (1974) Variational methods for eigenvalue approximation. CBMS-NSF regional conference series in applied mathematics, vol 15. SIAM, Philadelphia.

Whittaker ET (1923) On a new method of graduation. Proc Edinb Math Soc 41:63–75.

Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, Chichester.

Williamson B (1899) An elementary treatise on the differential calculus. Longman, Green, London.

Woodroofe M (1985) Estimating a distribution function with truncated data. Ann Stat 13:163–177.

Xiang D, Wahba G (1996) A generalized approximate cross validation for smoothing splines with non-Gaussian data. Stat Sin 6:675–692.

Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc 86:79–86.

Zhang T, Lin G (2009) Cluster detection based on spatial associations and iterated residuals in generalized linear mixed models. Biometrics 65:353–360.

Zucker DM, Karr AF (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. Ann Stat 18:329–353.

Author Index

Abramowitz, M., 37, 55, 167
Akhiezer, N. I., 55
Anderson, J. A., 19, 278, 282, 316
Ansley, C. F., 167
Antoniadis, A., 316, 348
Aronszajn, N., 55
Azzalini, A., 212

Barry, D., 56
Bartoszyński, R., 316
Bates, D. M., 115, 116
Becker, R. A., 94
Berger, J. O., 56, 116
Blair, V., 282
Bose, S., 282
Bowman, A. W., 212
Box, G. E. P., 233
Breiman, L., 106, 107, 117
Breslow, N. E., 233
Brinkman, N. D., 106
Brockwell, P. J., 204, 233
Brown, B. W., 316
Brown, M. B., 167
Buja, A., 103, 107, 117
Bunch, J. R., 116
Byerly, W. E., 143

Casella, G., 55
Chambers, J. M., 94
Chang, Y., 211
Chen, Z., 56, 115, 116, 349
Chhikara, R. S., 211
Clayton, D. G., 233
Cleveland, W. S., 106, 107, 117, 282
Cogburn, R., 212
Cole, T. J., 282
Cosslett, S. R., 278, 279, 282
Cox, D. D., 211, 322, 348, 349
Cox, D. R., 20, 281, 299, 317
Craven, P., 19, 55, 64, 68, 116, 167
Crowley, J., 20

Dalzell, C. J., 167
Davis, H. T., 212
Davis, M. D., 212
Davis, R. A., 204, 233

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7,

Delampady, M., 12
DeMets, D. L., 212
Dennis, J. E., 88, 162, 247, 393
Devlin, S. J., 106, 107, 117
de Boor, C., 3, 19, 112, 113, 118
de Hoog, F. R., 118
Dongarra, J. J., 81, 93, 116
Douglas, A., 12
Du, P., 316, 383
Duchon, J., 134, 135, 167

Elden, L., 82, 116

Fan, J., 282
Fleming, T. R., 20, 289, 290, 305, 316, 333
Folks, J. L., 211
Foltz, A., 275
Friedman, J. H., 107

Gao, F., 282
Gaskins, R. A., vii, 19, 280, 282
Gentleman, R., 94
Gijbels, I., 282
Gill, P. E., 56, 84, 393
Gill, R. D., 281, 316, 333
Girard, D. A., 114, 115, 118
Glazman, I. M., 55
Golub, G., 63, 79–81, 88, 112–114, 116, 281
Good, I. J., vii, 19, 280, 282
Gray, R. J., 317
Green, P. J., vii, 166, 210, 282
Gu, C., 19, 20, 56, 70, 79, 81, 83, 89, 93, 115–117, 167, 180, 181, 183, 210–212, 219–221, 226, 227, 234, 235, 247, 280–282, 316, 348, 383, 398

Härdle, W., 202, 212
Hall, P., 398, 399
Halpern, J., 15
Han, C., 226, 227, 234
Hansen, M. H., 281, 282
Harrington, D. P., 20, 289, 290, 305, 316, 333
Harville, D. A., 116, 233
Hastie, T. J., 19, 94, 117, 118, 317, 400
Heckman, N. E., 116, 165, 167
Hornik, K., 94
Hu, M., 20
Hutchinson, M. F., 118

Ihaka, R., 94

Jenkins, G. M., 233
Jeon, Y., 351, 352, 383
Jewell, N. P., 281
Joe, H., 231, 235
Johnson, R. A., 49, 76
Johnstone, I., 398
Jones, M. C., 281
Jones, R. H., 234

Kadane, J., 186
Kalbfleisch, J. D., 20, 317
Kaplan, E. L., 286
Karcher, P., 234
Karim, M. R., 233
Karlin, S., 153
Karr, A. F., 19, 317
Ke, C., 147
Keiding, N., 281
Kernighan, B. W., 93
Kim, Y.-J., 70, 89, 116, 117, 167, 192, 348
Kimeldorf, G., vii, 19, 55, 56, 167
Klein, B. E. K., 56, 210–212, 282
Klein, R., 56, 210–212, 282
Koenker, R., 282
Kohn, R., 167
Kooperberg, C., 281, 282

Lauritzen, S. L., 19
Lawless, J. F., 20
Lee, Y., 234
Lehmann, E. L., 55
Leonard, T., 19, 186, 280

Li, K.-C., 66, 68, 69, 115, 116, 397
Lin, G., 208, 212
Lin, X., 234, 282
Lin, Y., 349, 351, 352, 383
Lindsey, J. K., 212
Lindstrom, M., 116
Lumley, T., 317
Lynch, R., 19

Ma, P., 56, 219–221, 234, 235, 282
Ma, S., 316
Mallows, C. L., 116
Manski, C. F., 282
Marron, J. S., 399
McBride, C. M., 316
McCullagh, P., 177, 210, 211
McCulloch, C. E., 233
Meier, P., 286
Meinguet, J., 134, 135, 138, 167
Mesbah, M., 316
Miller, R. G., 15, 20
Moler, C. B., 116
Moreau, T., 295, 316
Morgenthaler, S., 282
Moss, S. E., 212
Murray, W., 56

Nelder, J. A., 177, 210, 211, 234
Ng, P., 282
Novak, E., 249
Nychka, D., 78, 116

O'Quigley, J., 316
O'Sullivan, F., 19, 113, 118, 210, 212, 280, 316, 317, 348
Ouyang, Z., 359

Parzen, E., 281
Pawitan, Y., 212
Petras, K., 249
Portnoy, S., 282
Prentice, R. L., 20, 282, 317
Priestley, M. P., 166, 204
Pyke, R., 282

Qiu, C., 19, 56, 280, 348

Ramsay, J. O., 165, 167
Rao, C. R., 54
Raynor, W., 19, 210
Reinsel, G. C., 233
Ripley, B. D., 94, 161, 211
Ritter, K., 249
Robinson, G. K., 116, 233

Scheffe, H., 19
Schnabel, R. B., 88, 162, 247, 393
Schoenberg, I. J., 3, 19, 56
Schumaker, L., 3, 19, 55, 112, 113, 118, 153, 155–158, 167
Schur, I., 55
Scott, A. J., 282
Scott, D. W., 212, 253, 268, 281, 398
Seber, G. A. F., 19
Senthilselvan, A., 19, 316
Silverman, B. W., vii, 19, 56, 166, 210, 243, 280, 348
Smyth, G., 198
Snyder, D. L., 242
Stegun, I. A., 37, 55, 167
Stein, M. L., 167
Stewart, G. W., 99, 116, 117
Stone, C. J., 19, 281, 282
Studden, W. J., 153

Tapia, R. A., 52, 56
Terrell, G. R., 398
Therneau, T. M., 317
Thompson, J. R., 52, 56, 316
Tibshirani, R. J., 19, 117, 118, 317, 400
Tierney, L., 186
Tong, H., 203
Truong, Y. K., 281, 282
Tsay, W.-Y., 281
Turnbull, B. W., 20

Utreras, F., 321, 322, 348, 349

Van Loan, C., 63, 79–81, 88, 112–114, 116, 281

Vardi, Y., 281, 282
Venables, W. N., 94, 161, 211

Wahba, G., vii, 19, 20, 55, 56, 61, 64, 68, 71, 74, 75, 78, 79, 83, 115, 116, 134, 135, 138, 143, 146, 166, 167, 178, 181, 183, 210–212, 282, 322
Wang, J., 116, 117, 247, 281
Wang, M.-C., 14, 20, 281
Wang, Y., 56, 147, 167, 205, 210–212, 234
Weinberger, H. F., 320, 348
Wellner, J. A., 281
Wendelberger, J., 134, 135, 138, 167
Whittaker, E. T., 19
Whittaker, J., 12, 19
Wichern, D. W., 49, 76
Wild, C. J., 282
Williamson, B., 170
Woltring, H. J., 113
Wong, C.-M., 167
Wong, W.-H., 359
Woodroofe, M., 281
Wright, M. H., 56

Xiang, D., 178, 181, 183, 211

Yan, L., 55
Yandell, B., 19, 116, 210

Zeger, S. L., 233
Zhang, D., 234
Zhang, T., 208, 212
Zhou, Q., 359
Zucker, D. M., 19, 317

Subject Index

Accelerated life models, 303–317
Additive models, 6, 9–10, 12, 19, 47, 49, 59, 104, 110, 122, 141, 206, 231, 240, 255, 272, 287, 294, 297, 330, 377
AIDS incubation, 14, 255, 262
Algorithm
  convergence of, *see* Convergence
  for density estimation, 247–250
  for hazard estimation, 291
  for L-splines, 165–166
  for penalized least squares, 79–85, 111–115
  for penalized likelihood regression, 177–181
  starting value, 84, 393
ANOVA decomposition, 6–12, 21, 33, 36, 39–41, 44, 93, 98, 104, 128, 138, 139, 153, 238, 240, 264, 269, 314, 365, 372, 387–389
ANOVA models, 6, 7, 9, 10, 19, 33, 48, 49
Approximation
  efficient, 85–93, 117, 176, 240, 259, 264, 327–330, 338–340, 345–348, 363–364, 372, 381–383
  linear, 309, 323, 334, 342, 361, 379
  Monte Carlo, 114, 118
  quadratic, 177, 186, 222, 244, 271, 290, 292, 300, 317, 323, 352, 354, 374, 376
At-risk
  probability, 289, 374
  process, 289, 299
Averaging operator, 6–9, 11, 21, 33, 36, 39, 41, 42, 44, 58, 128, 138, 238, 239, 264, 269, 270

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7,

B-splines, 19, 113, 118, 280
  basis, *see* Basis
Bacteriuria data, *see* Data sets
Base hazard, *see* Hazard
Basis
  B-spline, 113, 280
  local-support, 112, 113, 280
  of null space, 36, 39, 47, 62, 77, 104, 118, 135, 137, 139, 154, 155, 157, 158, 169, 176, 212, 353, 365, 366, 373, 384, 385, 387
  orthonormal, 37, 57, 137, 139, 144, 154, 155, 157, 158, 169, 322
Bayes model, 48–51, 56, 70–72, 75–79, 86–88, 185–186, 219, 223–224
  empirical, 48, 56
Bayesian confidence intervals, 20, 75–79, 84, 86–87, 95, 106, 116, 128, 136, 163, 164, 185–186, 207, 211, 222, 271–273, 292–293, 295, 300, 316, 375, 376, 391, 392
Bernoulli polynomials, 37, 38, 55
Biased sampling, *see* Sampling
Biasing function, 257, 260, 261
Bilinear form, 25, 27, 28, 37, 62, 176
Binomial
  distribution, *see* Distribution
  family, *see* Family
BLAS, 93, 393
Blood transfusion data, *see* Data sets
Buffalo snowfall data, *see* Data sets

Cancer mortality data, *see* Data sets
Cauchy sequence, 26, 31, 57
CDC blood transfusion data, *see* Data sets
Censored data, 5, 15, 20, 286, 296, 297, 333
Censoring, 15, 289, 296, 333
  rate, 291, 292
  time, 5, 15, 286, 307, 310, 312
Chebyshev
  space, 158, 165, 166
  spline, 55, 153–157, 167
  system, 153, 158
Cholesky decomposition, 80, 88, 89, 216, 247, 248
  band, 81, 112
  modified, 84
Climate data, *see* Data sets
Closed
  set, 26
  space, 26, 28, 29, 32, 36, 56, 57
$\mathcal{C}^{(m)}[0,1]$ space, 34–40, 55, 153, 157
Collinearity, 98, 101, 103, 117
  indices, 99, 117
Colorectal cancer mortality data, *see* Data sets
Complete continuity, 320–323, 335, 342, 361, 379
Complete space, 26, 27, 31
Concurvity, 103, 105, 109–111, 117, 206
Conditional
  density, 263–278, 282, 332, 364–372
  distribution, 11, 252, 278, 358
  independence, 10–12, 240, 242, 246, 271, 353, 369
  non-negative definiteness, 136
Conditional density estimation, *see* Density estimation
Consistency, 66, 68–70, 72, 226, 227

Constrained
least squares, 3
optimization, 51, 53–54, 56, 136, 280
Continuity, 24–26, 29, 32, 37, 51, 53, 57, 62, 112, 115, 118, 123, 134, 177, 212, 239, 287, 353, 365, 373, 384, 385
complete, *see* Complete continuity
Contrast, 7, 34, 36, 39, 40, 44, 48, 71, 72, 128, 153, 287, 388, 389, 391, 392
y-, 271–273, 275
Convergence, 25, 26, 31, 53, 58
of algorithms, 79, 83, 84, 88, 115, 180, 181, 198, 202, 205, 247
rates, 66, 85, 90, 226, 240, 258, 264, 280, 287, 322–349, 360–364, 378–383
Convex
combination, 179, 343, 360
functional, *see* Functional
set, 325, 329, 331, 332, 336, 340, 344, 345, 362, 363, 372, 380, 382
Convexity, 51–54, 60, 62, 118, 177, 212, 238, 279, 284, 286, 330, 353, 365, 373, 384, 385
Correlation parameter, 218, 223, 227, 393
Covariance
function, 48–51, 71, 76, 77, 86, 87, 185, 203
matrix, 49, 120, 220, 271, 292, 376
posterior, 77
Cross-validation, 67–68, 116, 234, 260, 265, 270, 272, 280, 293, 318, 383, 393, 397–400
for density estimation, 244–247, 265–266, 273–274, 354–357, 366–369
direct, 181–184, 188, 189, 191–192, 194, 197, 199–200, 202, 205, 211, 222, 281, 306, 309–310, 312, 316, 317
generalized, 64, 68–71, 116, 162, 167, 219–221
generalized approximate, 178, 181–184, 211
for hazard estimation, 289–292, 374–375
indirect, 178–181, 316
Monte Carlo, 114–115
for non-Gaussian regression, 177–184, 188–189, 191–192, 194–200, 305–307, 309–310, 312
for regression with correlated data, 225–230
Cubic spline, 2–3, 36, 39, 42, 74, 86, 91, 92, 95–97, 99, 101, 102, 106, 108, 122, 128, 136, 138, 140, 156, 159, 161, 162, 189, 192, 195, 197, 198, 200, 201, 203, 204, 206, 221, 228, 239, 246, 251, 255, 260, 264, 265, 267, 268, 287, 291, 292, 294, 297, 307, 310, 313, 321, 355, 358, 367, 376, 387–389, 395
with jump, 126
CV, *see* Cross-validation

Data sets
bacteriuria, 231–232, 235
Buffalo snowfall, 253–254, 281

CDC blood transfusion, 14, 20, 255–256, 262–263, 281
climate, 147–149
colorectal cancer mortality, 208–210, 212
EPA lake acidity, 12–14, 20, 140–142
eyetracking, 275–277
gastric cancer, 295–296, 316
Los Angeles ozone, 107–111, 118, 232–233
nitrogen oxides (NO_x), 106–107, 117
Old Faithful eruption, 202–203, 212, 254–255
Stanford heart transplant, 15–17, 20, 297–299, 302, 314–316, 378
transcription factor association, 359–360
US penny thickness, 268, 269, 371, 371
weight loss, 161–165
WESDR, 205–207, 212
yearly sunspots, 203–205, 212
Decomposition
ANOVA, *see* ANOVA decomposition
Cholesky, *see* Cholesky decomposition
eigenvalue, *see* Eigenvalue
of kernel, *see* Reproducing kernel
QR, *see* QR-Decomposition
of reproducing kernel, *see* Reproducing kernel
spectral, *see* Spectral
tensor sum, *see* Tensor sum decomposition
Density estimation, 4, 14, 19, 191–192, 212, 238–256, 280, 281, 352–360, 383, 391–393
conditional, 263–278, 282, 332, 364–371, 391
convergence rates for, 322–333, 360–364, 372
Poisson, 242
under sampling bias, 257–263, 278–281, 300
Dispersion, 176, 188, 191, 193, 196, 197, 199, 212, 308, 342
Distance, 25, 26
Euclidean, *see* Euclidean
Kullback-Leibler, *see* Kullback-Leibler distance
Distribution
binomial, 188, 211
conditional, 11, 252, 278, 358
empirical, 20, 281
exponential, 193, 291
exponential family, 176, 178, 341, 348
extreme value, 303
gamma, 193, 211
inverse Gaussian, 196, 211
lifetime, 286, 303, 317
log logistic, 304, 317
log normal, 304, 317
logistic, 304
negative binomial, 199
normal, 49, 76, 304
Poisson, 191, 211, 242
posterior, 76–78, 186
uniform, 10, 250, 258, 358
Weibull, 303, 317

Eigenfunction, 145, 320
Eigenvalue, 33, 63, 66, 70, 71, 90, 118, 130, 145, 146, 149, 220, 226, 320–323, 335, 342, 348, 361, 379
decomposition, 33, 66, 90
Eigenvector, 33

EPA lake acidity data, *see* Data sets
Estimate
  Bayes, 48–51, 85
  cross-validation, 182, 245, 259, 283, 290, 355, 367, 374
  existence of, *see* Existence of estimates
  interval, 75–79, 185–186, 222, 271–272, 292, 300, 375, 376, 392
  martingale moment, 316
  maximum likelihood, 4, 52, 238, 240, 286, 293, 318
  penalized least squares, 348, 393
  penalized likelihood, 52, 56, 238, 242, 319, 348, 367, 393
  shrinkage, 34, 48–49, 55
  unbiased, 65, 72
  variance, 69–71, 75, 78, 116
Estimation
  density, *see* Density estimation
  hazard, *see* Hazard estimation
  of relative risk, *see* Relative risk
  spectral, *see* Spectral
Euclidean
  distance, 25, 135
  inner product, 28
  norm, 25, 99
  space, 25, 26, 28–30, 33, 57, 269
Evaluation functional, 29, 30, 32, 134, 176, 239, 240, 287
  representer of, 29, 33, 35, 38, 55
Existence of estimates, 52–53, 56, 62, 126, 135, 177, 238, 240, 286, 318, 353, 365, 373
Exponential distribution, *see* Distribution
Exponential family distribution, *see* Distribution
Extreme value distribution, *see* Distribution
Eyetracking data, *see* data sets

Family
  binomial, 188–191
  exponential, *see* Exponential family distribution
  gamma, 193–196, 211, 213, 341
  inverse Gaussian, 196–198, 211, 213, 341
  location-scale, 303
  log logistic, 311–314, 318
  log normal, 309–311, 318
  negative binomial, 199–202, 211, 213, 341, 342
  Poisson, 191–193, 213
  Weibull, 305–308, 318
Fixed effects, 48, 49, 51, 77, 120, 216, 218, 219
Fourier
  coefficient, 145, 320, 321, 323, 326, 335, 342, 349, 350, 361, 379
  expansion, 127, 166, 320, 323, 335, 337, 342, 361, 379
  frequency, 204
  matrix, 129, 168
  transform (discrete), 129, 166, 204
Fréchet differentiability, 53, 54, 118, 212, 244, 353, 365, 373, 384, 385
Function
  biasing, *see* Biasing function
  covariance, *see* Covariance
  on discrete domains, 6, 28, 32, 41
  generating, 144

Green's, *see* Green's function
hazard, *see* Hazard
Legendre, *see* Legendre
non-negative definite, 30, 40
periodic, 37, 127, 150, 152, 320
predictable, 290, 333, 334
survival, *see* Survival
Functional, 25, 26, 53
continuous, 25, 29, 52, 53
convex, 52, 53, 60, 243, 352
evaluation, *see* Evaluation functional
least squares, 56, 62, 64, 118
linear, 25, 29, 32, 57, 320
log likelihood, 62, 65, 177, 212
penalized least squares, 51, 62, 64, 72, 73, 145, 177, 218
penalized likelihood, 29, 51, 176, 222, 238, 243, 258, 264, 272, 279, 286, 304, 334, 341
penalized partial likelihood, 300
penalized pseudo likelihood, 352, 364, 372, 376
pseudo likelihood, 353, 365, 373, 384, 385
quadratic, 4, 25, 52, 60, 134, 168, 319, 320, 333, 335, 341, 342, 361, 379

GACV (generalized approximate cross-validation), *see* Cross-validation
Gamma
distribution, *see* Distribution
family, *see* Family
regression, *see* Regression
Gastric cancer data, *see* Data sets
Gaussian
inverse, *see* Inverse Gaussian
prior, *see* Prior
process, 50, 51, 71, 76, 77, 86, 185
regression, *see* Regression
GCV (generalized cross-validation), *see* Cross-validation
GML (generalized maximum likelihood), 71, 72
Graphical models, 10–12, 19
Green's function, 154, 157, 158
GRKPACK, 205, 211

Hazard
base, 12, 16, 286, 298, 299, 301, 302, 315, 316
function, 5, 286, 303, 304
model, *see* Proportional hazard model
Hazard estimation, 5, 15–17, 19, 20, 286–299, 316, 372–378, 383, 391–393
convergence rates for, 333–341, 378–383
Heart transplant data, *see* Data sets
Hilbert space, 24–29, 52–54, 56, 57, 60
finite-dimensional, 29, 57
reproducing kernel, *see* Reproducing kernel Hilbert space

Independence, 5, 11, 14, 48, 49, 51, 76, 77, 120, 123, 182, 185, 216, 218, 219, 221, 240, 255, 258, 262, 266, 272, 276, 286, 289, 290, 333, 356, 360
conditional, *see* Conditional
linear, 24, 25, 166

Inequality
Cauchy-Schwarz, 25, 31, 56, 119, 132, 133, 326, 329, 337, 345
Hölder's, 238, 287
triangle, 25, 26, 56, 57
Inner product, 25–28, 31–35, 37, 42–47, 51, 54, 55, 57, 58, 62, 128, 138, 139, 145, 150–155, 158, 160, 161, 169, 173, 176, 284
semi, 27, 28, 37, 45, 54, 138
Interaction, 6, 8–13, 42, 44, 45, 59, 97, 100, 101, 104, 108, 109, 117, 122, 139, 141, 206, 231, 271, 279, 284, 292, 297, 330, 358–360, 389, 391
Invariance, 12, 68, 84, 134, 139, 144, 168, 170, 188–189, 357, 360, 369, 393
Inverse, 33, 58
Moore-Penrose, 33, 42, 86, 183, 185, 235, 271, 292, 376
Inverse Gaussian
distribution, *see* Distribution
family, *see* Family
regression, *see* Regression
Isomorphism, 29, 57, 127, 239, 299, 317

Kernel
reproducing, *see* Reproducing kernel
semi-, 135–136, 139
Knot, 3, 112, 123
Kronecker
delta, 37, 57, 132, 137, 144, 153, 320
product, 42, 272
Kullback-Leibler distance, 178, 182, 186, 187, 192, 194, 197, 199, 226, 236, 243, 250, 259, 265, 266, 273, 280, 289, 293, 300, 305, 306
relative, 182, 192, 194, 197, 199, 244, 245, 259, 265, 270, 280, 289, 305, 330–333, 341, 347
symmetrized, 178, 189, 243, 260, 265, 291, 305, 322, 326, 329, 331–334, 336, 338, 340, 355, 357, 360, 367, 369, 375, 384
Kullback-Leibler projection, *see* Projection

L-splines, 73, 149–167
$\mathcal{L}_2[0,1]$ space, 27–29, 34, 38, 42, 44, 127, 158
Lagrange
method, 2, 3
multiplier, 2–4, 53
Lake acidity data, *see* Data sets
Least squares
constrained, 3
functional, *see* Functional
parametric, 135, 162
penalized, *see* Penalized least squares
weighted, 64, 73, 79, 177, 206, 210
Legendre
function, 143, 167
polynomial, 144
Likelihood, 4, 5, 20, 24, 29, 52, 54, 62, 64, 71, 119, 185, 186, 188, 191, 193, 196, 199, 213, 218, 225, 238, 242, 244, 258, 272, 279, 284, 286, 291, 299, 304, 305, 309, 311, 317, 318, 341, 352
maximum, *see* Maximum likelihood
partial, 299, 317

penalized, *see* Penalized likelihood
pseudo, 352, 353, 355, 365, 373
Limit point, 25, 26, 31
Linear
approximation, *see* Approximation
dependence, 248
functional, *see* Functional
independence, 24, 25, 166
regression, *see* Regression
spline, *see* Spline
system, 63, 79, 86, 89, 112, 114, 115, 118, 130, 136, 248, 283
Linear models, 2, 19, 25, 94, 98, 103, 105, 233, 317, 400, 401
generalized, 210, 211, 233
log, 19, 271
retrospective, 98, 99, 121
Linear space, 24–32, 36, 52, 53, 56, 57, 112, 127, 157, 240
dimension of, 24, 25
Linearity, 2, 3, 53, 54, 68, 207, 210, 323, 342
non-, 163
LINPACK, 81, 93, 393
Log logistic
distribution, *see* Distribution
family, *see* Family
regression, *see* Regression
Log normal
distribution, *see* Distribution
family, *see* Family
regression, *see* Regression
Logistic
distribution, *see* Distribution
regression, *see* Regression
spline, 161
Logistic density transform, 10, 238, 240, 280, 299
conditional, 11, 263, 264
Los Angeles ozone data, *see* Data sets
Loss, 65, 72, 74, 91, 92, 178, 189, 219–221, 226, 243, 246, 260, 273, 289, 306, 318, 354, 366, 374, 398–400
relative, 65, 72, 354, 366, 374

Main effect, 6, 8–11, 42, 44, 45, 59, 104, 108, 109, 122, 358, 359, 388
multivariate, 139
Matrix
banded, 112, 165
column space of, 32, 33, 49, 87, 90
conditionally non-negative definite, 136
covariance, *see* Covariance
Fourier, 129, 168
Gram, 137
hat, 400, 401
projection, 33, 87, 90, 121, 122, 235
smoothing, *see* Smoothing matrix
sparse, 114
Wronskian, 153, 157, 159–161
Maximum likelihood, 2, 116, 238, 286, 317
estimate, *see* Estimate
generalized, 71, 72
restricted, *see* REML
Mean
(as opposed to contrast), 36, 39, 40, 48, 128, 287
posterior, 48–50, 75, 77, 87, 92, 95, 96, 186, 271, 292, 376
sample, 244, 290, 354, 367

Mean square error, 74, 92, 93, 219, 374, 396, 398
Mean value theorem, 179, 325, 336, 338–340, 343, 362, 381–384
Mixed-effect models, 48, 49, 51, 77, 116, 217–223, 233, 234, 272–274, 293, 300, 376, 391, 393
Model
  accelerated life, 303–317
  additive, *see* Additive models
  ANOVA, *see* ANOVA models
  Bayes, *see* Bayes model
  frailty, 293, 300, 376
  graphical, 10–12, 19
  index, 397, 400
  linear, *see* Linear models
  mixed-effect, *see* Mixed-effect models
  nonlinear, 162
  parametric, 2, 79, 103, 149, 162, 286, 303, 317, 397
  proportional hazard, *see* Proportional hazard model
  semiparametric, 126

Negative binomial
  distribution, *see* Distribution
  family, *see* Family
Nitrogen oxides (NO_x) data, *see* Data sets
Non-negative definiteness, 25, 27, 30, 32, 35, 40, 118, 136, 138, 235, 248, 249
  conditional, 136
Norm, 25, 27, 31, 36, 38, 48, 52, 53, 87, 89, 138, 238, 248, 269, 322, 365, 373
  Euclidean, *see* Euclidean
  semi, 27, 48, 52, 53, 118, 134, 138, 212, 238, 269, 365, 373, 384, 385
Normal
  density, 265, 366
  distribution, *see* Distribution
Normality, 6, 49, 64, 66, 114, 179, 271, 292, 376
Null space, 2, 4, 27–29, 34, 36, 37, 45, 47, 48, 52, 62, 134, 135, 137, 139, 149–153, 155–158, 162, 176, 238–240, 264, 286, 287, 322
  basis, *see* Basis

Old Faithful eruption data, *see* Data sets
Operator
  averaging, *see* Averaging operator
  backward shift, 217
  differential, 144, 149–154, 156, 157, 159, 162, 172
  identity, 7, 137
  projection, 137
Optimality, 65, 69, 70, 74, 90, 131, 134, 179, 181, 189, 219, 221, 226, 246, 260, 265, 291, 328, 396–401
Optimization, 88, 181, 355, 393
  constrained, 51, 53–54, 56, 280
  penalized, 3, 51, 53–54, 56
Orthogonal complement, 26, 28, 35, 48, 57
Orthogonality, 31, 32, 36, 49, 63, 66, 73, 80, 82, 83, 99, 104, 120, 127, 132, 168, 169, 183, 212, 322
Orthonormal basis, *see* Basis
Ozone data, *see* Data sets

Penalized least squares, 2–4, 19, 62, 115, 176, 213, 391, 395
  estimate, *see* Estimate
  functional, *see* Functional
Penalized likelihood, 2, 4–5, 14, 15, 19, 20, 176, 177, 210, 230, 238, 242, 243, 250, 258, 264, 266, 270, 272, 279–282, 286, 304, 316, 321, 367
  estimate, *see* Estimate
  functional, *see* Functional
  joint, 218, 222
  partial, 300, 317
  pseudo, 352, 358, 364, 367, 372, 376
Penalized optimization, 51, 53–54, 56
Penny thickness data, *see* Data sets
Periodicity, 38, 58
Periodogram, 204
Poisson
  composite, 199, 213
  distribution, *see* Distribution
  family, *see* Family
  process, 242, 280
  regression, *see* Regression
Polynomial spline, 34–40, 44, 49–51, 55, 56, 112, 116, 149, 155, 156, 321
  periodic, 128
Polynomials, 2, 35, 37, 112, 134, 137, 139, 149, 168, 169, 217
  Bernoulli, 37, 38, 55
  Legendre, 144
  piecewise, 3, 10, 86, 112, 123
Positive definiteness, 25, 27, 83, 84
Prior, 48, 49, 87
  diffuse, 48, 49, 51, 76, 77, 86, 219
  Gaussian, 48–51, 76, 77, 86, 185
  proper, 120, 219
Process
  ARMA, 217, 227
  at-risk, 289, 299
  counting, 242, 316
  event, 289, 305
  Gaussian, *see* Gaussian
  Poisson, 242, 280
Projection, 26, 32, 41, 45, 47, 51, 53, 56, 98, 99, 137, 220, 323, 328, 339, 345, 363, 382
  Kullback-Leibler, 117, 186–188, 206–208, 211, 222, 250, 252, 259, 266, 272, 277, 281, 293, 294, 300, 302, 306, 310, 312, 316, 330, 357, 369, 391, 392
  matrix, *see* Matrix
  operator, 137
  square error, 104–105, 110, 117, 218, 357, 358, 369, 376, 383
Proportional hazard model, 12, 15, 16, 19, 286, 287, 297–299, 305, 316, 378

QR-Decomposition, 63, 80, 83, 139, 169, 212, 224
Quadratic
  approximation, *see* Approximation
  form, 25
  functional, *see* Functional

R, 94, 106, 140, 147, 149, 162, 202, 204, 206, 209, 317
  `assist` package, 147
  `gss` package, 94–99, 106, 107, 117
  `gss` package, 126, 128, 140, 147, 152, 187–193,

195–196, 198, 201–202, 206, 208, 211, 222–223, 230–231, 250–253, 255, 260–262, 266–268, 274–276, 293–297, 300–302, 307–308, 310–311, 313–314, 358, 369–370, 376–377
`maps` package, 142, 147
`MASS` package, 161
`statmod` package, 198
`survival` package, 317
Random effects, 48, 49, 51, 77, 120, 216, 218–222, 233, 235, 272–274, 276, 293, 300, 391, 392
Regression, 2, 4, 19, 33, 115, 210, 212, 213, 263, 391, 392, 395
convergence rates for, 341–348
with correlated data, 216–233, 272–273
with cross-classified responses, 269–277
gamma, 193–196, 203–205, 212
Gaussian, 4, 62, 94, 106–111, 117, 176, 216, 230, 232–233
inverse Gaussian, 196–198
linear, 2, 207, 210, 400, 401
log logistic, 311–314
log normal, 309–311
log-linear, 271, 274–277
logistic, 20, 176, 177, 188–191, 199–202, 205–207, 231–232, 269–270, 272
multinomial, 282
non-Gaussian, 187, 211, 222, 234, 286, 303
nonlinear, 161
nonparametric, 19, 282, 400, 401
parametric, 2, 103, 210
Poisson, 20, 191–193, 202–203, 208–210, 212, 254, 255, 271
ridge, 114, 116
Weibull, 305–308, 314–316, 375, 377
Relative risk, 12, 16, 286, 299, 301, 302, 305, 314, 316
estimation of, 299–302, 317, 318, 391
REML (restricted maximum likelihood), 70–72, 87, 88, 116, 224, 234
Representer, 29
of evaluation functional, *see* Evaluation functional
Reproducing kernel, 29–48, 51, 55, 57, 58, 62, 128, 135–139, 145, 146, 150–155, 158, 160, 161, 166, 176, 264, 269, 284, 387–389
decomposition of, 33, 36, 39, 138, 152
marginal, 40, 41, 55
product, 42
Reproducing kernel Hilbert space, 29–32, 34, 40–48, 51, 55, 57, 58, 62, 118, 138, 212, 238, 242, 269, 284, 286, 365, 388
Riesz representation theorem, 29, 55
RKPACK, 81, 93, 94, 98, 117, 188, 393
G-, 205, 211

Sampling
biased, 257–263, 281, 299, 300, 317, 331
case-control, 278
choice-based, 278
length-biased, 257, 260, 281

response-based, 278–280, 282, 332
separate, 278, 279
Sampling points, 3, 36, 65, 67, 68, 75, 78, 87, 92, 98, 103, 111, 112, 140, 178, 186, 206, 207, 389, 400
Side condition, 4, 6–9, 11, 14, 34, 41, 238–240, 242, 243, 263, 264, 279, 280, 299, 316, 331
Singularity, 63, 73, 229
non-, 50, 76, 88, 137, 158, 184–186, 213, 248
numerical, 81, 248, 281
Smoothing, 19, 34, 44, 106, 112, 114, 221, 242, 254, 281, 395, 397
over-, 249
under-, 70, 75, 98, 140, 189
Smoothing matrix, 63, 73, 78, 114, 115, 118, 120, 183, 218, 219, 400, 401
Smoothing parameter, 2, 4, 5, 62, 64, 65, 78–80, 82, 88, 91–93, 95–97, 112, 114, 130, 177–181, 184, 188, 218, 241–243, 247, 267, 281, 304, 367, 393, 396, 400
selection of, 5, 88, 104, 114, 116, 117, 127, 130, 140, 162, 177–181, 219–221, 225–230, 243–247, 257, 274, 281, 296, 298, 300, 316, 354–357, 366–369, 374–375, 398, 400
Smoothing spline, 3, 19, 34, 47, 48, 51, 55, 56, 78, 94, 112, 348
on discrete domain, 32–34, 49
interpolation, 19
Space
Chebyshev, *see* Chebyshev
closed, *see* Closed
$\mathcal{C}^{(m)}[0,1]$, *see* $\mathcal{C}^{(m)}[0,1]$ space
complete, *see* Complete space
Euclidean, *see* Euclidean
finite-dimensional, 4, 26, 27, 29, 32, 36, 42, 51–53, 60, 62, 112, 134, 158, 240, 287, 323
Hilbert, *see* Hilbert space
$\mathcal{L}_2[0,1]$, *see* $\mathcal{L}_2[0,1]$ space
linear, *see* Linear space
marginal, 42, 44, 388
null, *see* Null space
tensor product, *see* Tensor product
vector, 24, 25, 27, 28, 32, 42, 54, 55
Spectral
analysis, 167
decomposition, 127, 129, 130, 133, 167
density, 204, 205, 217
estimation, 203–205, 212
Spectrum, 203, 204
mass, 202, 203
power, 217
Spherical
coordinates, 143, 144
harmonics, 143–145, 167
spline, 144–149, 322, 389
Spline
B, *see* B-splines
Chebyshev, 55, 153–157, 167
cubic, *see* Cubic spline
exponential, 155, 156, 159
hyperbolic, 156
L, 73, 149–167
linear, 35, 39, 42, 388, 397
logistic, 161

natural, 3, 56, 112, 113, 118, 123
partial, 126–127, 150, 166, 287, 391
periodic, 127–128, 130, 153, 388
polynomial, *see* Polynomial spline
regression, 281
smoothing, *see* Smoothing spline
spherical, 144–149, 322, 389
tensor product, *see* Tensor product
thin-plate, *see* Thin-plate spline
trigonometric, 150–153, 160, 388, 390
Splus, 94, 317
Stanford heart transplant data, *see* Data sets
Survival, 15, 296, 297, 299, 302, 314, 378
analysis, 19, 20, 316, 317
function, 5, 286, 303, 304
probability, 20
time, 15, 296
Symmetry, 25, 29, 39, 62, 113

Taylor expansion, 34, 39, 55, 182, 244, 325
generalized, 39, 155, 156, 159
Tensor product, 269, 391
space, 40–48, 55, 59, 264, 269, 270, 284, 388
spline, 12, 20, 40–48, 55, 56, 59, 91, 92, 96, 97, 99, 101, 102, 106, 108, 122, 128, 135, 137, 139, 140, 167, 206, 240, 246, 251, 255, 264, 265, 267, 268, 287, 292, 294, 297, 321, 355, 358, 367, 376
Tensor sum decomposition, 26, 27, 32, 33, 35–37, 39, 41, 42, 44, 45, 47, 353
Thin-plate spline, 12, 20, 134–142, 167, 322, 389
Transcription factor association data, *see* Data sets
Transform
cubic root, 107
discrete Fourier, *see* Fourier
log, 13, 106, 108
logistic density, *see* Logistic density transform
monotone, 71
orthogonal, 68, 169
square root, 16, 280, 297
Truncation, 14, 20, 255, 257, 258, 281, 291
left-, 5, 286

US penny thickness data, *see* Data sets

Variance
components, 116, 216, 218, 233
estimate, *see* Estimate
inflation factor, 99
posterior, 75–77, 84, 87, 112, 136, 165, 186, 271, 292, 376

Weibull
distribution, *see* Distribution
family, *see* Family
regression, *see* Regression
Weight loss data, *see* Data sets
WESDR data (Wisconsin Epidemiological Study of Diabetic Retinopathy), *see* Data sets

Yearly sunspots data, *see* Data sets