Theoretical and Practical Aspects of Penalized Spline Smoothing

Dissertation zur Erlangung des Grades
eines Doktors der Wirtschaftswissenschaften (Dr. rer. pol.)
der Fakultät für Wirtschaftswissenschaften
der Universität Bielefeld

Vorgelegt von
MSc. Fin. Math. Tatyana Krivobokova

Bielefeld, im August 2006

Dekan:	Prof. Dr. Alfred Greiner
Gutachter:	Prof. Dr. Göran Kauermann Prof. Dr. Ludwig Fahrmeir

Abstract

Parametric regression models that describe the dependence of the mean of some response variable on a set of covariates play a fundamental role in statistics. Allowing for simple interpretation and estimation these models, however, are often not flexible enough for describing the data at hand. In the last 15 to 20 years with the development of computer technology and statistical software, another approach - nonparametric regression - has received more attention and recognition. The mean of a response is thereby modelled as a smooth, but otherwise unspecified function of covariates.
The large domain of nonparametric regression models includes local techniques like kernel or locally-weighted smoothers and spline methods. The main focus of this thesis is on *penalized splines* (P-splines), which have become a very powerful and applicable smoothing technique over the last decade. This nonparametric method can be viewed as a generalization of smoothing splines with a more flexible choice of bases and penalties. The main attraction of P-spline smoothing is its ties with ridge regression, mixed models and Bayesian statistics. This allows the adoption of different techniques, like Markov chain Monte Carlo or likelihood ratio tests for penalized spline methodology. Smoothing, in particular, can be performed with any mixed model or Bayesian software.
This thesis addresses several problems of nonparametric techniques that can be successively handled with penalized spline smoothing, due to its link to mixed models. First, smoothing in the presence of correlated errors is shown to be more robust if performed in the mixed models framework. This property is used to estimate the term structure of interest rates. Next, the problem of smoothing of locally heterogeneous functions is treated by representing the adaptive penalized splines as a hierarchical mixed model. Application of Laplace approximation for parameter estimation of this model results in the fast and efficient method for adaptive smoothing, which is implemented in the R package *AdaptFit*. Investigation of the asymptotic rate at which the spline basis dimension is supposed to grow to minimize mean squared error concludes the thesis.

Acknowledgements

I am deeply grateful and indebted to my principal advisor Prof. Dr. Göran Kauermann for his permanent support and encouragement, for sharing his great ideas, knowledge and enthusiasm for statistics and for giving me the opportunity to work in a most inspiring and friendly environment. I would like also to thank Theophanis Archontakis, Ciprian M. Crainiceanu, PhD, and Prof. Dr. Ludwig Fahrmeir for the productive and stimulating collaboration.

Table of Contents

1 Introduction **1**

1.1 Overview of nonparametric techniques 1

1.2 Objectives of Thesis and Outline 2

2 Penalized Spline Smoothing **5**

2.1 Idea of Penalized Smoothing 5

2.1.1 Regression and Penalized Splines 5

2.1.2 Spline Bases and Penalties 8

2.1.3 Basic Definitions 13

2.1.4 Smoothing Parameter Selection 14

2.2 Mixed Model Representation 16

2.2.1 Mixed Models 17

2.2.2 Penalized Splines as Mixed Models 21

2.3 Confidence Intervals 23

2.4 Extensions 24

2.4.1 Additive Models 24

2.4.2 Bivariate Smoothing 26

2.4.3 Smoothing with Generalized Response 29

2.5 Computational Issues 33

2.5.1 Choice of Knots and Basis 33

2.5.2 Fast and Stable Penalized Smoothing 35

2.5.3 R Packages *mgcv*, *nlme* and *SemiPar* 35

2.6 Bayesian Model for Smoothing 38

3 Smoothing with Correlated Errors **41**

3.1 Motivation 41

3.2 Smoothing Parameter Selection 42

3.2.1 Akaike and REML 42

3.2.2 Smoothing with misspecified correlation . . . 44
3.2.3 Simulation . . . 47
3.3 Examples and Applications . . . 50
3.4 Extensions . . . 54
3.4.1 Additive Models . . . 54
3.4.2 Non-normal Response . . . 55
3.5 Smoothing Parameter Estimation . . . 56
3.5.1 REML estimate . . . 56
3.5.2 AIC estimate . . . 58
3.5.3 Relation of AIC and REML based smoothing parameters . . . 59
3.6 Computational Issues . . . 60
3.7 Discussion . . . 63

4 Estimation of the Term Structure of Interest Rates 65
4.1 Motivation . . . 65
4.2 Data . . . 67
4.3 Spline Models for the Term Structure . . . 69
4.3.1 Bivariate spline smoothing . . . 69
4.3.2 Spline Smoothing with correlated errors . . . 70
4.3.3 Empirical Results . . . 72
4.4 Discussion . . . 74

5 Fast Adaptive Penalized Smoothing 75
5.1 Motivation . . . 75
5.2 Smoothly varying local penalties for P-spline regression . . . 76
5.2.1 Hierarchical penalty model . . . 76
5.2.2 Restricted maximum likelihood . . . 80
5.2.3 Variance estimation . . . 80
5.2.4 Numerical implementation . . . 81
5.2.5 Simulations and comparisons with other univariate smoothers . . . 82
5.3 Spatial smoothing . . . 84
5.3.1 Hierarchical modelling . . . 84
5.3.2 Simulations and comparisons with other surface fitting methods . . . 85
5.4 Non-normal response model . . . 87
5.4.1 Hierarchical modelling . . . 87

5.4.2 Simulations for the logistic regression example 88
5.5 Example 89
5.6 R Package *AdaptFit* 91
5.7 Discussion 92

6 Some Asymptotics on Penalized Splines **93**
6.1 Introduction 93
6.2 Generalized P-Spline Smoothing 95
6.3 P-Spline Smoothing and Mixed Models 99
6.3.1 Laplace Approximation 99
6.3.2 Posterior Cumulants 103
6.3.3 Maximum Likelihood Estimation 105
6.4 Asymptotic Behavior 106
6.5 Discussion 112

7 Summary **113**

List of Figures

2.1 Function $m(x)$ estimates (bold) based on 5 (left) and 20 (right) knots . . 6

2.2 Function $m(x)$ estimates (bold) based on 20 knots and $\lambda = 0.2$ (left) and $\lambda = 0.002$ (right) 7

2.3 B-spline of degree 2 (left) and degree 3 (right) 9

2.4 Truncated lines (left) and B-spline basis functions (right) of degree 1 . . 10

2.5 Radial basis functions of degree 1 (left) and 3 (right) 12

2.6 Criteria $\exp(AIC)$ (bold), GCV (dashed), Mallow's C_p (dotted) and estimated curve (bold) with confidence bands (dashed) 17

2.7 Tensor product of B-splines of degree 2 at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$. 28

2.8 Low-rank radial splines at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$. 30

3.1 Estimated curves with AIC (dashed) and REML (bold) based smoothing parameter choice (left) and partial autocorrelation function corresponding to the true function (right). 47

3.2 Boxplots for log-transformed smoothing parameter choice in 100 simulations. Upper row shows $\log \hat{\lambda}_{AIC}$ and $\log \hat{\lambda}_{REML}$ without accounting for correlation. The two bottom plots correspond to AIC and REML smoothing parameters when the true correlation structure is explicitly taken into account. 48

3.3 Boxplots for average squared error in 100 simulations. Upper row shows $\mathrm{ASE}(\hat{\lambda}_{AIC})$ and $\mathrm{ASE}(\hat{\lambda}_{REML})$ without accounting for correlation. The two bottom plots correspond to ASE of fits with AIC and REML smoothing parameters correspondingly, when the true correlation structure is explicitly taken into account. 49

3.4 Estimated curves (bold) with AIC and REML based smoothing parameter choice (top row) with boxplots for log-transformed smoothing parameter choice in 100 simulations (middle row) and boxplots for average squared error in 100 simulations (bottom row). 51
3.5 Estimated curves with (bold) and without (dashed) accounting for correlation using AIC (left) and REML based (right) choice of smoothing parameter (top row) with partial autocorrelation functions for the AIC (left) and REML based (right) fits without accounting for correlation (bottom row). Dotted lines show confidence bands for the REML fit which accounts for correlation. 52
3.6 Estimated curves with an AR(2) (bold) and AR(1) (dashed) correlation structure using AIC (left) and REML based (right) choice of smoothing parameter (top row) with partial autocorrelation functions for the AIC (left) and REML based (right) fits accounting for an AR(1) correlation structure (bottom row). 53
3.7 Right: Estimated curves with AIC (dashed) and REML (bold) based smoothing parameter choice. Left: Partial autocorrelation function corresponding to the REML estimate. 55

4.1 Design of independent covariates time t and maturity m 68
4.2 Yield development of a 6 years bond, smoothed with (solid line) and without (dashed line) accounting for correlation, with the corresponding partial autocorrelation function of residuals. 71
4.3 Bivariate fit of the term structure 72
4.4 Estimated term structure with prediction intervals 73
4.5 Estimated yield development with prediction intervals 74

5.1 Estimated regression functions $m_1(x)$ (left) and $m_2(x)$ (right) with confidence intervals (dashed) and true function. 82
5.2 Pointwise MSE with a smoother of the points (left) and smoothed pointwise coverage probabilities of 95% confidence intervals (right) for 500 simulated datasets with function $m_1(x)$ 83
5.3 Pointwise MSE with a smoother of the points (left) and smoothed pointwise coverage probabilities of 95% confidence intervals (right) for 500 simulated datasets with function $m_2(x)$. 84
5.4 True regression function $m_3(x_1, x_2)$, adaptive and non-adaptive estimates 86

5.5 Smoothed coverage probability of 95% confidence intervals for 500 simulated datasets with function $m_3(x_1, x_2)$. 86
5.6 Estimated regression function $\pi = \text{logit}^{-1}[m_2(x)]$ with adaptive penalty (bold), with global smoothing parameter (dashed) and true function for 1000 grouped binomial data ($n_i = 5$) (left) and 5000 binary data (right). 89
5.7 Estimated regression function $P(d = t|d \geq t, c) = \text{logit}^{-1}(m(t, c))$ with global (left) and local (right) smoothing parameter. 91

1 Introduction

Modelling the dependence of the mean of some response variable y on a set of covariates $x_1, ..., x_d$ is one of the main objectives of regression analysis. The intention is to specify a function $m(.)$ such that

$$E(y|x_1, ..., x_d) = m(x_1, ..., x_d)$$

or in case of non-normal response (e.g. count or binary data)

$$E(y|x_1, ..., x_d) = h[m(x_1, ..., x_d)],$$

with $h(.)$ as the inverse of a link function. Modelling regression function $m(.)$ as a linear combination of some known functions of covariates, e.g. $m(x_1, ..., x_d) = \beta_0 + \sum_i \beta_i x_i$, leads to a parametric (generalized) linear model. These models possess a well-developed theory, they are easy to estimate and to interpret. However, the underlying assumptions are often too restrictive and not supported by the data at hand. A need for more flexible approaches has led to the development of a number of nonparametric methods, where function $m(.)$ is modelled as some unspecified smooth function of covariates.

1.1 Overview of nonparametric techniques

Nonparametric models are conceptually different from linear regression. The functional dependence between response and covariates is exposed without imposing any particular parametric assumption about this dependence. We give here a short overview of nonparametric techniques for estimating the model with one metric covariate $E(y|x) = m(x)$, extension to the multiple covariate and generalized response is then straightforward.
The simple *running mean* or *moving average* method estimates the regression function as $m(x_i) = \sum_{j \in N_i^k} y_j/k$, with N_i^k as a neighborhood of x_i containing k observations. The approach is popular in time series analysis, but produces an estimate which is hardly "smooth". A convenient generalization is the *locally-weighted running-line smoother*,

also known as *loess*. Instead of calculating a mean value one computes a weighted least-squares line in each neighbourhood. The smoothness of the estimate is controlled by the "size" of the neighbourhood, expressed as a percentage or span of the data points. Wider spans result in smoother fits. Another enhancement of local smoothing is achieved by using local averaging with the so-called kernels weights. The regression function is then estimated as $m(x_i) = \sum_j K[(x_i - x_j)/\lambda]y_j / \sum_j K[(x_i - x_j)/\lambda]$, with some fixed constant λ and $K(.)$ as a kernel function, e.g. the standard Gaussian density. Here, the tuning parameter is bandwidth λ, with larger values leading to a smoother estimate.
Another approach to the nonparametric regression is to find $m(.)$ as a solution to the optimization problem

$$\sum_i [y_i - m(x_i)]^2 + \lambda \int [m''(t)]^2 dt, \tag{1.1}$$

with λ as a fixed constant. The first term in (1.1) ensures the closeness of the estimate to the data, while the second penalizes the curvature of the function. It has been shown that the *natural cubic spline* with knots at x_i is the unique solution of (1.1). Parameter λ plays the same role as the bandwidth in kernel smoothing or span in loess. Small λ values imply an interpolating estimate, while large smoothing parameter forces $m''(x) \to 0$, yielding the least squares line fit. Modelling $m(.)$ with spline functions (natural cubic or B-splines) without penalization, but with appropriate choice of number and location of knots, defines *regression spline* smoothing.
Penalized spline smoothing is a nonparametric technique which has become very popular over the last decade. It can be seen as a generalization of spline smoothing with a more flexible choice of bases, penalties and knots. Namely, one chooses a spline basis based on some sufficiently large set of knots and penalizes unnecessary structure. One of the main strengths of this approach is its link to mixed and Bayesian models. This allows application of techniques such as likelihood ratio tests or Markov chain Monte Carlo to the penalized spline methodology. In particular, smoothing can be performed with mixed models or Bayesian software.

1.2 Objectives of Thesis and Outline

This thesis aims to investigate some aspects of penalized spline smoothing. After presenting the theoretical background we concentrate on three main issues.
First, smoothing in presence of correlated errors in considered. We show that the mixed model representation of penalized splines results in smoothing parameter estimation that is more robust to misspecification of the correlation structure, when compared to stan-

dard methods using mean squared error based criteria for smoothing parameter choice. We demonstrate with a number of real data examples how this property can help to discover an underlying variance structure. In particular, estimation of a long term trend in macroeconomic or financial time series can be approached with this method. As illustration a two dimensional smoothing of the term structure of interest rates is performed. This latter modelling exercise has two main challenges: untypical correlation structure of the data and very large sample size (more than 126 000 points). Penalized spline smoothing allows handling both problems successfully, resulting in a fast and robust smoothing.

Secondly, we face the problem of the smoothing of a function of locally varying complexity, that is, if the regression function is changing rapidly in some regions while in other regions it is very smooth. Estimation of such functions with a global smoothing parameter is not efficient and a number of solutions have been suggested. We approach the problem in that we model the smoothing parameter as a smooth function, which has to be estimated as well. This assumption leads to a hierarchical mixed model and its likelihood function results in an intractable integral. We avoid, however, numerically extensive MCMC techniques and employ simple Laplace approximation for the parameter estimation. This results in a fast and efficient adaptive smoothing method, which can be readily extended to bivariate smoothing and models with generalized response. As illustration, we apply our approach to a dataset on absenteeism of workers of a medium-sized German industrial company. During the observation period (1981 - 1998) the company went through a major downsizing process (1992 - 1993) which changed the absenteeism behaviour of employees, increasing the probability of returning to work after a sick leave noticeably. We show that adaptive smoothing captures such untypical data structure more appropriate than non-adaptive. We also provide the R package *AdaptFit* to make application of the technique convenient and accessible.

Finally, we investigate asymptotics issues of penalized smoothing. In particular, we are interested how fast should the spline basis dimension grow, so that the mean squared error is minimized. In the mixed model framework this question relates to the order of the error in Laplace approximation.

This thesis is based on the following papers

- Krivobokova, T. and Kauermann, G. (2006). A Note on Penalized Spline Smoothing with Correlated Errors (submitted to *Journal of American Statistical Association*).

- Krivobokova, T., Kauermann, G. and Archontakis, T. (2006). Estimating the

term structure of interest rates using penalized splines. *Statistical Papers*, 47(3): 443-459.

- Krivobokova, T., Crainiceanu, C.M., Kauermann, G. (2006). Fast Adaptive Penalized Splines (submitted to *Journal of Computational and Graphical Statistics*).

- Kauermann, G., Krivobokova, T. and Fahrmeir, L. (2006). Some Asymptotics on Generalized P-Spline Smoothing (*working paper*).

2 Penalized Spline Smoothing

This chapter introduces penalized splines as smoothing technique. Beginning with the main idea of P-spline smoothing as a type of ridge regression, we extend it to the different basis functions and discuss the link to the mixed models. Computational issues, as well as extensions to additive and generalized models, are considered.

2.1 Idea of Penalized Smoothing

Penalized spline smoothing is a very flexible concept. Different basis functions, form of the penalties, amount and location of knots all provide a wide spectrum of smoothers. Some of them are discussed here.

2.1.1 Regression and Penalized Splines

We introduce the idea of penalized spline smoothing with the following model

$$y_i \sim N(m(x_i), \sigma_\epsilon^2), \quad i = 1, ..., n, \tag{2.1}$$

where $m(x)$ is a smooth, but otherwise unspecified, function of some univariate covariate x, that needs to be estimated from y_i, x_i. To be able to capture a complex non-linear structure of $m(x)$, we define K knots $k_1, ..., k_K$ and extend a parametric polynomial model with the truncated polynomial basis functions; i.e. we model

$$m(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{j=1}^{K} u_j (x - k_j)_+^p,$$

where $(x - k_j)_+ = \max\{0, (x - k_j)\}$. Denoting with $X = [1, x_i, .., x_i^p]_{1 \le i \le n}$, $Z = [(x_i - k_1)_+^p, ..., (x_i - k_K)_+^p]_{1 \le i \le n}$, $\beta = (\beta_0, ..., \beta_p)^T$ and $u = (u_1, ..., u_K)^T$ we can rewrite (2.1) as

$$y \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \tag{2.2}$$

with $y = (y_1, ..., y_n)^T$. The model (2.2) is purely parametric and can be easily estimated with ordinary least squares

$$\hat{y} = C(C^T C)C^T y,$$

with $C = [X, Z]$. This approach is referred to as regression spline smoothing. More general versions are defined for other basis functions, e.g. B-splines (see de Boor, 1978). Inherent with the advantages of parametric modelling, regression splines possess, however, a serious drawback - a proper strategy for selecting the number and location of knots is needed. As illustration we estimate with regression splines the function $m(x) = 4 + \sin(2\pi x)$ for 300 x equally spaced on $[0, 1]$ and independent $\epsilon_i \sim N(0, 0.3^2)$, $i = 1, ..., 300$. In Figure 2.1 the estimated curves (bold) are shown together with the used cubic truncated polynomial basis functions. While the left hand side fit is based on 5 equidistant knots, the plot on the right makes use of 20 equidistant knots. It appears

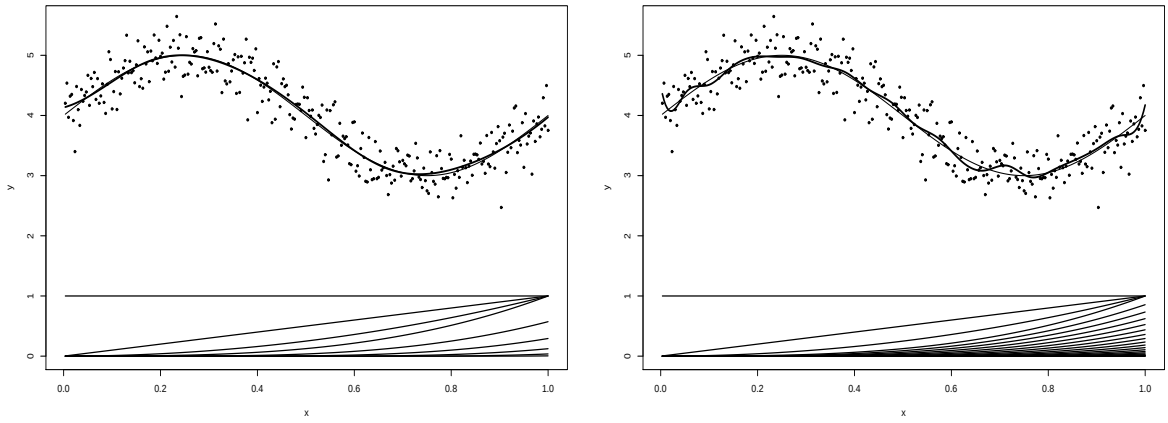

Figure 2.1: Function $m(x)$ estimates (bold) based on 5 (left) and 20 (right) knots

that the right hand side fit is "too" flexible, i.e. the data are "overfitted". It seems natural to employ some selection strategies for choosing an optimal model, similarly to the multiple linear regression, like the Akaike or (generalized) cross validation criteria. However, flexibility of regression splines also implies a huge amount of "candidate" models. For example for some fixed K knots there are $\sum_{i=0}^{K} \binom{K}{i} = 2^K$ possible models. Since the location of knots has additionally a marked effect on the fit, the usual selection procedures become unfeasible. Although a number of approaches for choosing the amount and position of knots have been suggested (see e.g. Fried & Silverman, 1989, Stone, Hansen, Kooperberg & Truong, 1997 or Smith & Kohn, 1996), all of them are rather

complicated and computationally intensive.
An alternative approach to optimize the fit is achieved by imposing a penalty on spline coefficients. Specifically, one chooses a large amount of knots (e.g. $\min\{n/4, 40\}$ as suggested in Ruppert, 2002) and prevents overfitting by putting a constraint on spline coefficients, i.e. one finds

$$\min_{\beta,u} \|y - X\beta - Zu\|^2, \text{ subject to } \|u\|^2 \leq c,$$

for some nonnegative constant c. Using the Lagrange multiplier, this minimization problem can be written as

$$\min_{\beta,u} \left\{ \|y - X\beta - Zu\|^2 + \lambda u^T u \right\} = \min_{\theta} \left\{ \|y - C\theta\|^2 + \lambda \theta^T D \theta \right\},$$

with $\theta = (\beta^T, u^T)^T$, $D = \text{blockdiag}(0_{(p+1)\times(p+1)}, I_K)$ and some $\lambda \geq 0$. The resulting estimate is given by

$$\hat{y} = C(C^T C + \lambda D)^{-1} C^T y. \tag{2.3}$$

Note that (2.3) is a type of ridge regression, which is used in parametric regression to reduce variability of estimates (see e.g. Draper & Smith, 1998). The smoothness of

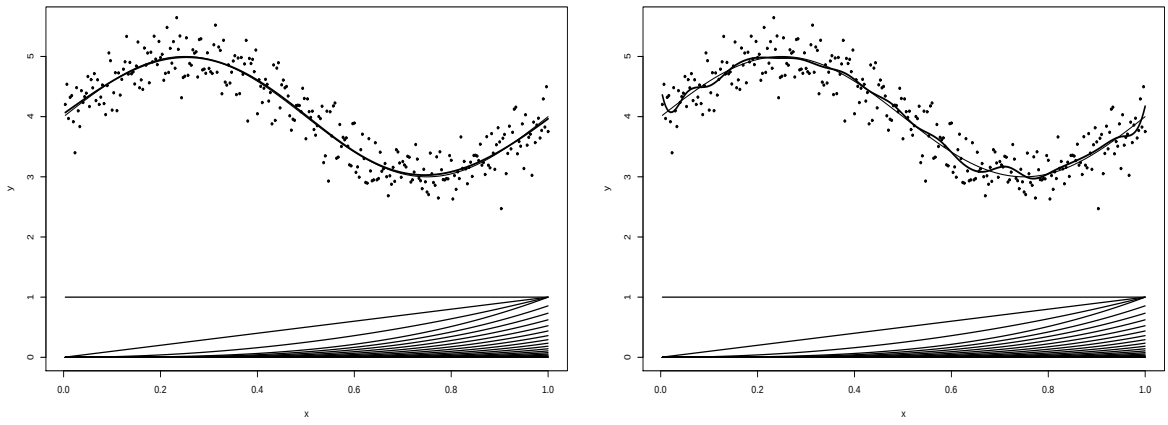

Figure 2.2: Function $m(x)$ estimates (bold) based on 20 knots and $\lambda = 0.2$ (left) and $\lambda = 0.002$ (right)

the estimate varies now continuously as a function of the single smoothing parameter λ. The larger the values of the smoothing parameter λ, the more the fit shrinks towards a polynomial fit, while smaller values of λ result in a wiggly "overfitted" estimate. This is

visible in Figure 2.2, where the above data are estimated with $K = 20$ and smoothing parameters $\lambda = 0.2$ (left hand side fit) and $\lambda = 0.002$ (right hand side fit). Such smoothing technique is known as penalized spline smoothing.

2.1.2 Spline Bases and Penalties

The truncated polynomial basis is simple, but not always numerically stable. When the number of knots is large and the smoothing parameter λ close to zero the inversion of $(C^TC + \lambda D)$ can lead to numerical problems. In this case the computation has to be organized carefully, involving QR or Demmler-Reinsch decomposition (see Section 2.5.2 or Ruppert, Wand & Carroll, 2003). However, numerically superior alternatives are available, like B-splines and radial basis functions, which will be presented subsequently.

B-splines

The idea of penalized spline smoothing traces back to O'Sullivan (1986), but it was Eilers & Marx (1996) who introduced the combination of B-splines and difference penalties which they called P-splines. B-splines based on a set of knots $k_1, ..., k_K$ are defined in de Boor (1978) (see also Dierckx, 1993) with a recursive formula

$$\begin{aligned} B_j^0(x) &= I_{[k_j, k_{j+1}]}(x), \\ B_j^p(x) &= \frac{x - k_j}{k_{j+p} - k_j} B_j^{p-1}(x) + \frac{k_{j+1} - x}{k_{j+p+1} - k_{j+1}} B_{j+1}^{p-1}, \end{aligned}$$

where $B_j^p(x)$ denotes jth B-spline of degree p. Note that additional $2p + 2$ knots are necessary for constructing the full B-spline basis of degree p. The general properties of a B-spline of degree p as given in Eilers & Marx (1996) are the following

- it consists of $p + 1$ polynomial pieces of degree p;
- the polynomial pieces join at p inner knots;
- at the joining points the derivatives up to order $p - 1$ are continuous;
- the B-spline is positive on a domain spanned by $p + 2$ knots, otherwise it is zero;
- except at the boundaries, it overlaps with $2p$ polynomial pieces of its neighbours;
- at a given x, $p + 1$ B-splines are non-zero.

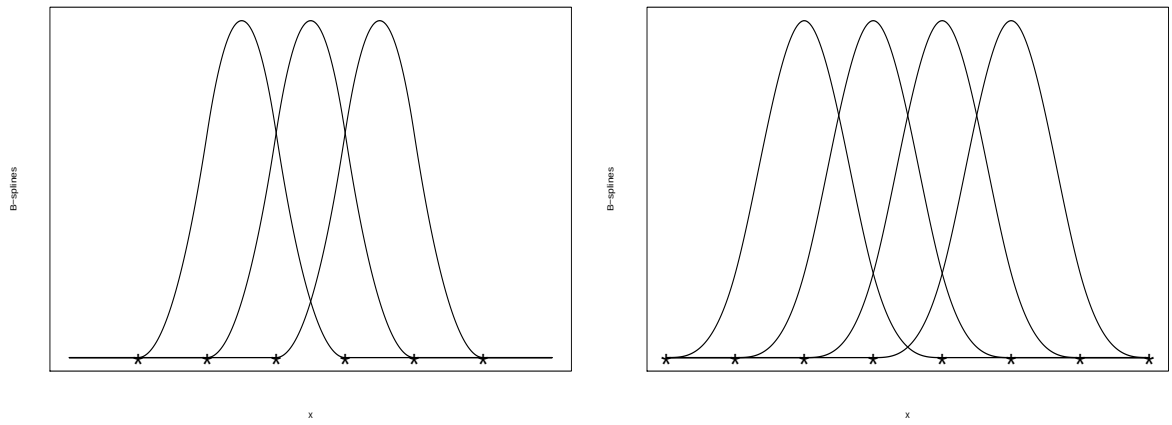

Figure 2.3: B-spline of degree 2 (left) and degree 3 (right)

Figure 2.3 shows an example of B-spline of degree 2 and 3, the position of knots is indicated by the stars.

B-splines can also be computed by the differencing of corresponding truncated polynomials, as shown in Eilers & Marx (2004). For example, the B-spline of degree $p = 1$ based on equidistant knots can be computed with $B_j^1(x) = (x - k_j)_+ - 2(x - k_{j-1})_+ + (x - k_{j-2})_+ =: \Delta^2 Z_j^1(x)$. The general formula is given by

$$B_j^p(x = (-1)^{p+1} \Delta^{p+1} Z_j^p(x)/(h^p p!),$$

with $h = k_{j-1} - k_j$, $Z_j^p(x) = (x - k_j)_+^p$ and the difference operator Δ^{p+1} defined through

$$\begin{aligned} \Delta^1 a_j &= a_j - a_{j-1}, \\ \Delta^2 a_j &= \Delta^1(\Delta^1 a_j) = a_j - 2a_{j-1} + a_{j-2}, \\ \Delta^q a_j &= \Delta^1(\Delta^{q-1} a_j). \end{aligned}$$

Somewhat more complicated results can be also obtained for arbitrary chosen knots, as shown in de Boor (2001). Note that one needs extra $2p + 2$ truncated polynomial basis functions to generate a complete B-spline basis of degree p.

A complete B-spline basis matrix of degree p for n observations based on K knots has dimension $n \times (K + 1 + p)$. It is not difficult to see that truncated polynomials and B-spline basis matrices of the same degree and based on the same knots are equivalent,

i.e. there exists a square invertible matrix L, such that

$$B = CL. \tag{2.4}$$

For example, for linear truncated polynomials based on $K = 5$ equidistant knots placed over $x_i \in [0, 1]$ matrix L has the form

$$L = 6 \begin{pmatrix} 1/6 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}, \tag{2.5}$$

with corresponding basis functions shown in Figure 2.4. Substituting (2.4) into (2.3)

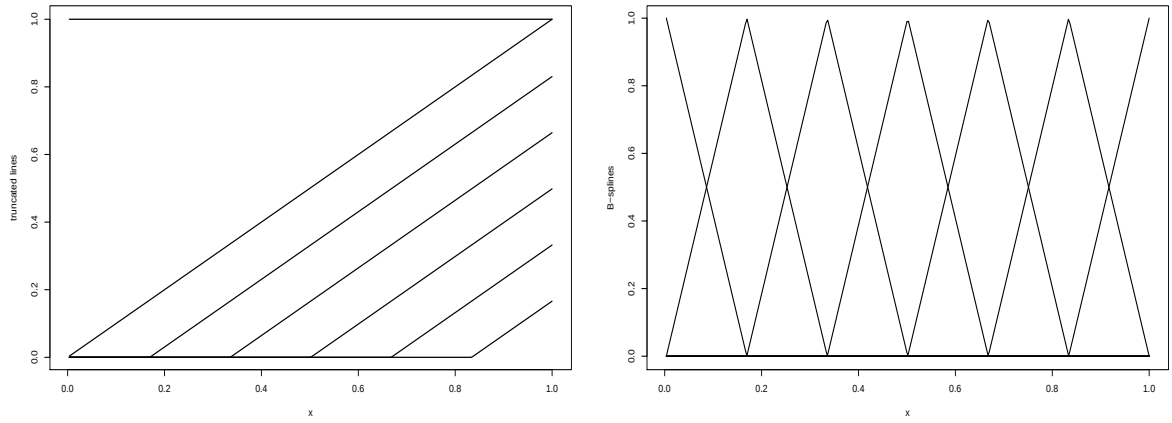

Figure 2.4: Truncated lines (left) and B-spline basis functions (right) of degree 1

results in

$$\hat{y} = B(B^T B + \lambda \tilde{D})^{-1} B^T y, \tag{2.6}$$

where $\tilde{D} = L^T D L$. For more details see Eilers & Marx (2004) or Ruppert, Wand & Carroll (2003).

From (2.6) we obtain also the form of the penalty used with B-spline basis, which in fact equals the difference penalty suggested by Eilers & Marx (1996), i.e. $\tilde{D} = \Delta_q^T \Delta_q$, where

Δ_q is a $(K+1+p-q) \times (K+1+p)$ matrix representation of the difference operator of order $q = p+1$. For example, the difference matrix of the second order Δ_2 has the form

$$\Delta_2 = \begin{pmatrix} 1 & -2 & 1 & & \\ & 1 & -2 & 1 & \\ & & \ddots & \ddots & \ddots \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

As argued by Eilers & Marx (1996) this penalty is a "good discrete approximation to the integrated square of the qth derivative". It follows immediately from the formula for the qth derivative of the B-spline of degree p, given in de Boor (1978)

$$h^q \sum_j \theta_j (B_j^p)^{(q)}(x) = \sum_j \Delta_q \theta_j B_j^{p-q}(x),$$

with θ as spline coefficients. The details can be found in Eilers & Marx (1996). The smoothness penalty as integrated square of the second derivative has become standard, following the work on smoothing splines by Reinsch (1967). Eilers & Marx (1996) also suggested using higher order differences, choosing thereby the difference order q and spline degree p independently. B-splines of degree three and second order difference penalty have become a common choice. Note that the difference order of the penalty determines the limiting fit - large smoothing parameter λ with $(q+1)$ order penalty shrink the fit toward a qth degree polynomial. For a truncated polynomial basis the limiting fit is defined by the polynomial degree p.

Radial basis functions

Radial basis functions are defined in Ruppert, Wand & Carroll (2003) as follows

$$1, x, ..., x^{r-1}, |x-k_1|^{2r-1}, ..., |x-k_K|^{2r-1}, \tag{2.7}$$

for $r = 1, 2, ...$ Figure 2.5 shows the radial basis functions for $r = 1$ and $r = 2$ based on $K = 5$ equidistant knots. The fitting criterion with the radial basis is given as

$$\min_{\beta,u} \left\{ \|y - X\beta - Z_R u\|^2 + \lambda u^T \Omega u \right\},$$

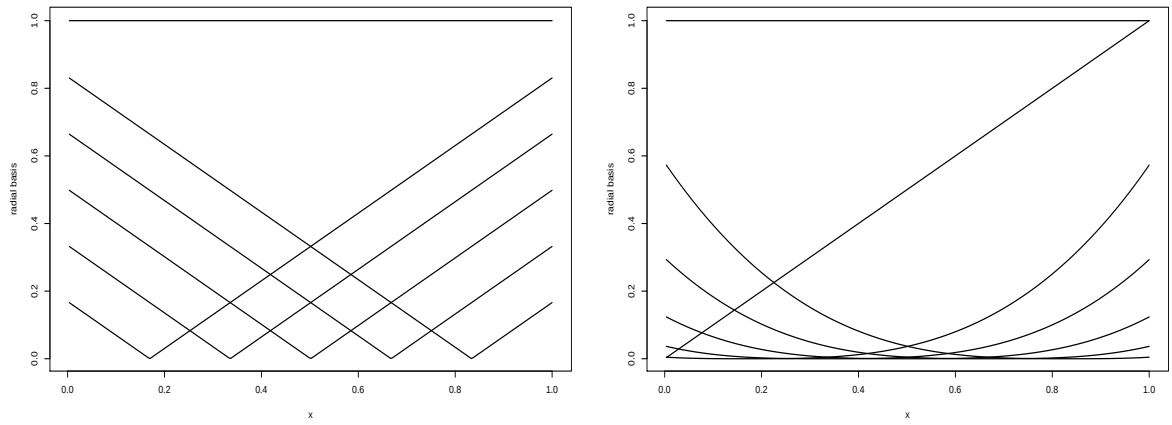

Figure 2.5: Radial basis functions of degree 1 (left) and 3 (right)

with model matrices $X = [1, x_i, .., x_i^{r-1}]_{1\leq i\leq n}$, $Z_R = [|x_i - k_1|^{2r-1}, ..., |x_i - k_K|^{2r-1}]_{1\leq i\leq n}$ and penalty matrix $\Omega = [|k_j - k_1|^{2r-1}, ..., |k_j - k_K|^{2r-1}]_{1\leq j\leq K}$. The motivation of this smoother comes from the equivalence of natural cubic spline and

$$m(x) = \beta_0 + \beta_1 x + \sum_{i=1}^{n} u_i |x - x_i|^3, \qquad \text{subject to } \sum_{i=1}^{n} u_i = \sum_{i=1}^{n} u_i x_i = 0, \tag{2.8}$$

as shown in Green & Silverman (1994). The constraints in (2.8) provide essentially the identifiability of basis coefficients β_0 and β_1. The integrated squared second derivative of function $m(x)$, used as a penalty with natural cubic splines, takes for (2.8) the form $u^T \tilde{Z}_R u$ with $\tilde{Z}_R = [|x_i - x_1|^3, ..., |x_i - x_n|^3]_{1\leq i\leq n}$. To fit n data with this approach one needs to estimate n model parameters and a smoothing parameter, which is computationally intensive, especially in case of more than one covariate ($O(n^3)$ operations) and a large n. Approximation to natural cubic splines can be obtained by specifying knots $k_1, ..., k_K$ and using basis functions $C_R = [X, Z_R]$ with the penalty Ω with $r = 2$, which reduces significantly numerical effort (see Ruppert, Wand & Carroll, 2003 or Wood, 2003). Extension to arbitrary odd degree results in (2.7) (see also Nychka, 2000).
Note that the radial basis functions depend on the distance between observations and knots only. This allows extension to higher dimensional predictors, e.g. by defining for $\mathbf{x}, \mathbf{k}_i \in R^d$ basis functions $r(\|\mathbf{x} - \mathbf{k}_i\|)$, with $\|.\|$ as Euclidean norm and some appropriate positive function $r(.)$. We postpone the details until Section 2.4.2.

2.1.3 Basic Definitions

A *penalized smoother* is defined through $C\hat{\theta}$, with $\hat{\theta}$ as a minimizer of

$$\|y - C\theta\|^2 + \lambda\theta^T D\theta,$$

for some positive definite matrix D and $\lambda > 0$. The model matrix C can contain any of the basis functions defined above with the penalty matrix D appropriately chosen. In fact, one has to make the following choice when applying penalized splines:

- amount and location of knots;
- spline basis functions;
- degree of the spline and the penalty matrix.

Penalized spline estimates and least squares fits share the key feature that they are both linear functions of the response variable. For a linear regression model $y \sim N(X\beta, \sigma_\epsilon^2 I_n)$ the ordinary least squares fit is obtained as $\hat{y} = X(X^TX)^{-1}X^Ty = Hy$, where H is the hat matrix. The penalized spline fit results in $\hat{y} = C(C^TC + \lambda D)^{-1}C^Ty$ and the matrix

$$S_\lambda = C(C^TC + \lambda D)^{-1}C^T$$

is called a *smoothing matrix*. Generalizing also the definition of degrees of freedom from linear models as the trace of the hat matrix, we can define with

$$df = \mathrm{tr}(S_\lambda)$$

the *degree of the smoother*, corresponding to the smoothing parameter λ. It can be interpreted as the equivalent number of fitted parameters.
The *residuals degrees of freedom* can be obtained analogously to the linear model from

$$E[RSS] = E[y^T(I - S_\lambda)^T(I - S_\lambda)y] = \sigma_\epsilon^2 \mathrm{tr}(I - 2S_\lambda + S_\lambda S_\lambda) + \|m(x)(I - S_\lambda)\|^2 .$$

Assuming that the bias $\|m(x)(I - S_\lambda)\|^2$ is negligible, we get an unbiased estimate for σ_ϵ^2 as RSS/df_{res} with

$$df_{res} = n - \mathrm{tr}(2S_\lambda - S_\lambda S_\lambda).$$

An alternative definition of residual degrees of freedom $n - \mathrm{tr}(S_\lambda)$, also common for linear smoothers, is used as well.

A common measure of error of the smoother for a fixed point x is the *mean squared error* (MSE), which is defined through

$$MSE[\hat{m}(x)] = \mathrm{Var}[\hat{m}(x)] + \{E[\hat{m}(x)] - m(x)\}^2 .$$

Since an entire fit is usually of interest, rather then individual points, one considers the *mean average squared error*

$$MASE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} MSE[\hat{m}(x_i)] = \frac{1}{n}\left\{\sigma_\epsilon^2 \mathrm{tr}(S_\lambda S_\lambda) + \|m(x)(I - S_\lambda)\|^2\right\}. \qquad (2.9)$$

The first term in (2.9) represents the average variance, the second is the average squared bias contribution. In general (2.9) reflects the so-called *bias-variance trade-off*; larger values of the smoothing parameter λ lead to a smaller variance but increase the bias, smaller λ values attain the opposite result. Thus, the optimal amount of smoothing has to be chosen by compromising goodness of fit with complexity of the estimated function. This issue is the subject of the next section.

2.1.4 Smoothing Parameter Selection

In this section some approaches to data-driven smoothing parameter choices are discussed.

(Generalized) Cross Validation

The residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is known as a common measure of goodness-of-fit in regression. However, estimation of an optimal smoothing parameter by minimizing of RSS will result in the fit that is closest to interpolation. Instead, one minimizes the cross validation expression

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{-i})^2,$$

where $\hat{y}^{-i}$ denotes the fit computed by leaving out the ith data point. For a linear smoother one can show that $\hat{y}_i^{-i}$ equals

$$\hat{y}_i^{-i} = \sum_{j \neq i} \frac{S_\lambda^{ij}}{1 - S_\lambda^{ii}} y_i,$$

with S_λ^{ij} denoting an ijth element of the smoothing matrix. With this it can easily be shown (see e.g. Hastie & Tibshirani, 1990) that

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - S_\lambda^{ii}}\right\}^2. \tag{2.10}$$

A modification of this criterion was suggested by Craven & Wahba (1979), which replaced S_λ^{ii} with $\mathrm{tr}(S_\lambda)/n$, resulting in the generalized cross validation criterion

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - \mathrm{tr}(S_\lambda)/n}\right\}^2 = \frac{RSS/n}{(1 - \mathrm{tr}(S_\lambda)/n)^2}.$$

Note that both GCV and CV in expectation approximate the mean average squared error. Indeed,

$$E(RSS/n) = MASE(\lambda) + \sigma_\epsilon^2 - 2\sigma_\epsilon^2 \mathrm{tr}(S_\lambda)/n. \tag{2.11}$$

Approximating now $(1-\mathrm{tr}(S_\lambda)/n)^{-2} = 1+2\mathrm{tr}(S_\lambda)/n+o(n^{-1})$ and assuming $\mathrm{tr}(S_\lambda)/n \approx \mathrm{tr}(S_\lambda S_\lambda)/n$ we find

$$\begin{aligned} E\left[GCV(\lambda)\right] &\approx \frac{1}{n}\left\{\sigma_\epsilon^2 \mathrm{tr}(S_\lambda S_\lambda)\left[1 - 2\frac{\mathrm{tr}(S_\lambda)}{n}\right] + \|m(x)(I - S_\lambda)\|^2\left[1 + 2\frac{\mathrm{tr}(S_\lambda)}{n}\right]\right\} + \sigma_\epsilon^2 \\ &= MASE(\lambda) + \sigma_\epsilon^2 + o(n^{-1}) \end{aligned}$$

A similar relationship holds for cross validation.

Mallow's C_p

Expression (2.11) motivates also another criterion, known as Mallow's C_p. If the variance σ_ϵ^2 where known, we could correct (2.11) by adding $2\sigma_\epsilon^2 \mathrm{tr}(S_\lambda)/n$. Substituting σ_ϵ^2 with its estimate results in the so-called Mallow's C_p statistic

$$C_p(\lambda) = RSS(\lambda)/n + 2\hat{\sigma}_\epsilon^2 \mathrm{tr}(S_\lambda)/n.$$

The disadvantage of this criterion is that it requires a proper prior estimate for σ_ϵ^2. It is suggested to choose $\hat{\sigma}_\epsilon^2 = RSS(\lambda)/df_{res}(\lambda)$ with a rather small λ (even $\lambda = 0$ would be appropriate for penalized smoothing) in order to minimize the bias (see Hastie & Tibshirani, 1990 or Ruppert, Wand & Carroll, 2003).

Akaike Information Criterion

Motivation of the Akaike criterion is different from the criteria above (Akaike, 1969). It is based on the Kullback-Leibler distance between the unknown true density $g(y)$ of the distribution generating the data y and the approximate model $f(y)$ used for fitting the data

$$I(f,g) = \int_{-\infty}^{\infty} \{\log[g(y)] - \log[f(y)]\}\, g(y)dy,$$

which has to be minimized with respect to $f(.)$. This is equivalent to minimizing of $-E_g[log(f(y)]$, which is for a normal density approximated by

$$AIC(\lambda) = \log[RSS(\lambda)] + 2\text{tr}(S_\lambda)/n. \tag{2.12}$$

Moreover, for a normal distribution it is easy to see that $E\{\exp[AIC(\lambda)]\}$ is also approximately equal to MASE. Indeed, analogously to GCV we approximate $\exp[2\text{tr}(S_\lambda)/n] = 1 + 2\text{tr}(S_\lambda)/n + o(n^{-1})$ and find

$$E\left\{\exp[AIC(\lambda)]\right\} = MASE(\lambda) + \sigma_\epsilon^2 + o(n^{-1}).$$

There were some modified versions of AIC criterion suggested, e.g. Hurvich, Simonoff & Tsai (9958) proposed for nonparametric regression using

$$AIC(\lambda) = \log[RSS(\lambda)] + \frac{2\text{tr}(S_\lambda) + 2}{n - \text{tr}(S_\lambda) - 2}.$$

Identification of an optimal smoothing parameter λ with the above criteria is carried out with a grid search. An efficient algorithm employing Demmler-Reinsch orthogonalisation is available and will be discussed in Section 2.5.2. Figure 2.6 shows all three described criteria and the resulting fit for the simulated data described in Section 2.1.1. For fitting a squared truncated polynomial basis based on 20 equidistant knots was used.

2.2 Mixed Model Representation

Mixed models are regression models that incorporate random effects. Having a wide application spectrum - from longitudinal studies to survival analysis - mixed models are also closely related to smoothing. Penalized spline smoothing corresponds exactly to optimal prediction in a mixed model framework. This makes it possible to use mixed

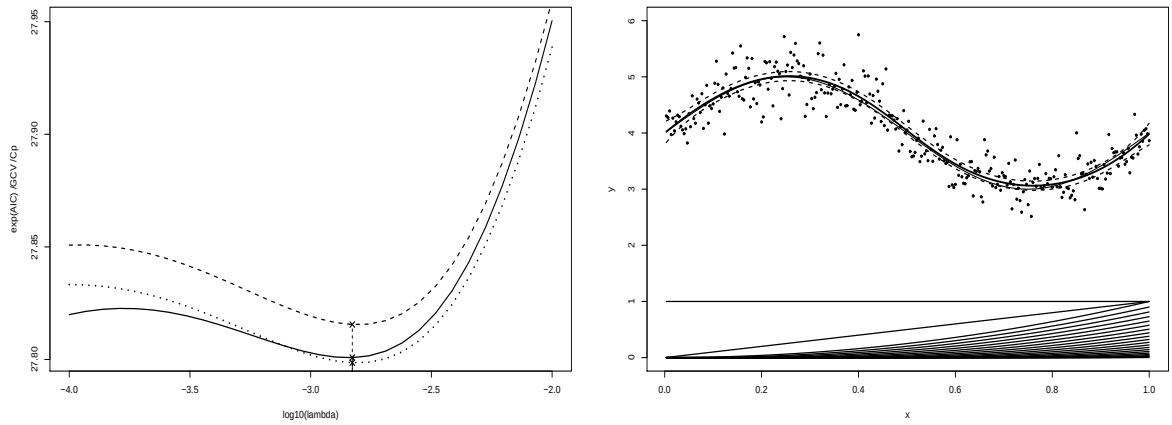

Figure 2.6: Criteria exp(AIC) (bold), GCV (dashed), Mallow's C_p (dotted) and estimated curve (bold) with confidence bands (dashed)

model methodology and software for penalized spline regression.

2.2.1 Mixed Models

We begin with a short review of linear mixed models, which can be defined as

$$y = X\beta + Zu + \epsilon.$$

Thereby y is a vector of n observable random variables, β is the $p+1$ dimensional vector of fixed effects, also known as "marginal" or "population-averaged" effects. The model matrices X and Z can be quite general, depending on application, we do not specify any form for these matrices at the moment. K dimensional vector u of random or "subject-specific" effects and n dimensional error term ϵ are unobservable random variables, such that

$$E\begin{bmatrix} u \\ \epsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \text{Cov}\begin{bmatrix} u \\ \epsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix},$$

where G and R are positive definite covariance matrices. Usually it is assumed that the random effects and the error term are normally distributed.

Fixed effects estimation

Estimation of the fixed effects β can be carried out from the linear model

$$y = X\beta + \epsilon^*,$$

where $\epsilon^* = Zu + \epsilon$ with $\mathrm{Cov}(\epsilon^*) = ZGZ^T + R =: V$. For a given covariance matrix V the fixed effects estimate results in

$$\tilde{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \tag{2.13}$$

Estimate $\tilde{\beta}$ is referred to as generalized least squares (GLS) and is the best linear unbiased estimator (BLUE) for β. For multivariate normal y estimate (2.13) can be derived as a maximum likelihood estimate.

Prediction

The best predictor of a random vector w based on an observable random vector v, such that $v = w + \epsilon$ with independent w and ϵ, is defined as a solution of $E[\|w - v\|^2]$. Commonly one assumes $\tilde{w} = a + Bv$, with some vector a and matrix B, restricting the family of predictors to be linear. Minimizing $E\{\|w - (a + Bv)\|^2\}$ with respect to a and B one easily finds that the best linear predictor for w is given by $\tilde{w} = E(w) + \mathrm{Cov}(w, v)\mathrm{Cov}^{-1}(v)[v - E(v)]$ (for more details see e.g. Searle, Casella, & McCulloch, 1992 or McCulloch & Searle, 2001). Note, from the standard results on multivariate normal distribution the best predictor for bivariate normal w and v is linear. Random effects u can now be predicted from $y - X\tilde{\beta} = Zu + \epsilon$ resulting in the best linear predictor

$$\tilde{u} = GZ^T V^{-1}(y - X\tilde{\beta}). \tag{2.14}$$

For known G and V predictor $\tilde{u}$ is shown in Robinson (1991) to be the best linear unbiased predictor (BLUP). Note that the difference between "estimator" and "predictor" is only that the target is deterministic for the former and random for the latter. However, some authors argue that the distinction is not necessary.
Another way to derive the BLUP solutions under normality assumption is given in Henderson (1950), who suggested to maximize the joint density of u and y

$$(2\pi)^{-\frac{n+K}{2}} \left(\det \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}\right)^{-1/2} \exp\left\{\frac{1}{2} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}^T \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}^{-1} \begin{pmatrix} u \\ y - X\beta - Zu \end{pmatrix}\right\},$$

leading to the criterion

$$(y - X\beta - Zu)^T R^{-1}(y - X\beta - Zu) + u^T G^{-1} u, \tag{2.15}$$

which is minimized by (2.13) and (2.14). From (2.15) is easy to see that $\tilde{\theta} = (\tilde{\beta}^T, \tilde{u}^T)^T$ can be also written as $\tilde{\theta} = (C^T R^{-1} C + B)^{-1} C^T R^{-1} y$, with $B = \text{blockdiag}(0, G^{-1})$, yielding

$$\tilde{y} = C\tilde{\theta} = C(C^T R^{-1} C + B)^{-1} C^T R^{-1} y \tag{2.16}$$

as best linear predictor for y.

Covariance matrix

It remains to estimate the covariance matrix V. With the normality assumption this can be done from the profile log-likelihood

$$-2l^p(V) = (y - X\tilde{\beta})^T V^{-1}(y - X\tilde{\beta}) + \log|V|. \tag{2.17}$$

Parameterizing V with some vector φ, one can find its estimate $\hat{\varphi}$, e.g. with Fisher scoring. General references are Harville (1977) and Lindstrom & Bates (1988) as well as Searle, Casella, & McCulloch (1992).
Replacing now covariance matrix by its estimate in (2.13) and (2.14) we get the estimated best linear estimate (EBLUE) and the estimated best linear predictor (EBLUP), respectively, as

$$\begin{aligned} \hat{\beta} &= (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y, \\ \hat{u} &= \hat{G} Z^T \hat{V}^{-1}(y - X\hat{\beta}). \end{aligned} \tag{2.18}$$

Estimated BLUE and BLUP have, therefore, an additional source of variability, due to estimation of V.

Restricted maximum likelihood

It is known that maximum likelihood estimates of variance are biased, since they do not take into account the degrees of freedom used for fixed effects estimation. Modifying the standard likelihood function using generalized least squares residuals, as suggested in Patterson & Thompson (1971) and later in Harville (1974), results in the method known

as *restricted (or residual) maximum likelihood* (REML). More precisely, in REML estimation one maximizes the log-likelihood for the residual vector $\tilde{e} = y - X\tilde{\beta}$, which due to the independence of $\tilde{e}$ and $\tilde{\beta}$ equals

$$\begin{aligned} l(\tilde{e}) = l(\beta, V) - l(\tilde{\beta}) &= -\frac{1}{2}[\log|V| + (y - X\beta)^T V^{-1}(y - X\beta) \\ &+ \log|X^T V^{-1} X| - (\tilde{\beta} - \beta)^T (X^T V^{-1} X)(\tilde{\beta} - \beta)] \\ &= l^p(V) - \frac{1}{2}\log|X^T V^{-1} X|. \end{aligned}$$

An alternative way to justify REML estimate is using Bayesian approach with a non-informative prior distribution for β, e.g. $\beta \sim N(0, \sigma_\beta^2)$ with a large σ_β^2. Integrating β out to obtain the corresponding marginal likelihood leads to the above result (see Laird & Ware, 1982). This suggests using the profile restricted likelihood for the estimation of V

$$l_R^p(V) = l^p(V) - \frac{1}{2}\log|X^T V^{-1} X|$$

instead of (2.17) (see also Harville, 1977).

Variance estimates

For the inference about estimated parameters we need to derive the corresponding variances. From (2.16) we immediately find

$$\mathrm{Cov}(\tilde{\theta}|u) = (C^T R^{-1} C + B)^{-1} C^T R^{-1} C (C^T R^{-1} C + B)^{-1}$$

and consequently $\mathrm{Cov}(C\tilde{\theta}|u) = C\mathrm{Cov}(\tilde{\theta}|u)C^T$. However, often the inference about precision of BLUPS is more relevant, i.e. one looks for

$$\mathrm{Cov}(\tilde{\theta} - \theta) = \mathrm{Cov}\begin{bmatrix} \tilde{\beta} - \beta \\ \tilde{u} - u \end{bmatrix} = \mathrm{Cov}\begin{bmatrix} \tilde{\beta} \\ \tilde{u} - u \end{bmatrix}.$$

Again, using (2.16) we get

$$\mathrm{Cov}\begin{bmatrix} \tilde{\beta} \\ \tilde{u} - u \end{bmatrix} = (C^T R^{-1} C + B)^{-1} C^T V C (C^T R^{-1} C + B)^{-1} - \begin{bmatrix} 0 & 0 \\ 0 & G \end{bmatrix} = (C^T R^{-1} C + B)^{-1},$$

yielding

$$\mathrm{Cov}[C\tilde{\theta} - C\theta] = C\mathrm{Cov}(\tilde{\theta} - \theta)C^T = C(C^T R^{-1} C + B)^{-1} C^T \tag{2.19}$$

Replacing V with its estimate provides the variance estimates for inference. These estimates, however, ignore the additional variability due to estimation of V. While this is acceptable for large datasets, it can make a difference for small sample sizes. This variability can be taken into account with a full Bayesian approach, presented in Section 2.6.

2.2.2 Penalized Splines as Mixed Models

Comparing (2.3) with (2.16) we can see that the penalized fit (2.3) can be obtained by assuming coefficient u to be random. We consider the following model

$$y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \qquad u \sim N(0, \sigma_u^2 I_K), \tag{2.20}$$

with matrices X containing polynomial and Z truncated polynomial basis functions as defined in Section 2.1. With the results of the previous section we get for $R = \sigma_\epsilon^2 I_n$ and $G = \sigma_u^2 I_K$

$$\tilde{y} = C(C^T C + \frac{\sigma_\epsilon^2}{\sigma_u^2} D)^{-1} C^T y, \tag{2.21}$$

with $D = \text{blockdiag}(0_{(p+1)\times(p+1)}, I_K^{-1})$. Thus, the ratio of variances $\sigma_\epsilon^2/\sigma_u^2$ in the mixed model framework plays the role of the smoothing parameter λ. With this in mind penalized spline smoothing is equivalent to the parameter estimation in a linear mixed model, which can be carried out with any standard mixed model software.
Note that the inverse of the penalty matrix imposed on spline coefficients has to be a proper covariance matrix - symmetric and positive definite. While this is unproblematic for truncated polynomials as shown above (covariance matrix is just identity), other basis functions with corresponding penalties need to be adjusted in order to be represented by a linear mixed model.

B-splines for mixed models

Let us consider penalized spline smoothing with B-spline basis of degree p with difference penalty of order q, based on K knots, that is

$$\|y - B\theta\|^2 + \lambda \theta^T \Delta_q^T \Delta_q \theta.$$

The difference matrix Δ_q has the dimensions $(K + 1 + p) \times (K + 1 + p - q)$ with the penalty matrix $\Delta_q^T \Delta_q$ being singular and having the rank $K + 1 + p - q$. A singular value

decomposition leads to $\Delta_q^T\Delta_q = U\text{diag}(b)U^T$ with U as eigenvectors and eigenvalues in vector b arranged in descending order, so that $K+1+p-q$ eigenvalues are strictly positive and the remaining q are zeros. Thus, we can represent $U = [U_+, U_0]$ and $b = (b_+^T, 0_q^T)^T$ with U_+ of dimension $(K+1+p) \times (K+1+p-q)$, corresponding to non-zero elements of vector b and rewrite

$$\begin{aligned} B\theta &= BUU^T\theta = B[U_0U_0^T\theta + U_+\text{diag}(b_+^{-1/2})\text{diag}(b_+^{1/2})U_+^T\theta] \\ &=: B[U_0\beta + U_+\text{diag}(b_+^{-1/2})u] =: X\beta + Z_Bu \end{aligned} \tag{2.22}$$

Moreover,

$$\theta^T\Delta_q^T\Delta_q\theta = \theta^TU\text{diag}(b)U^T\theta = \theta^TU_0\text{diag}(0_q)U_0^T\theta + \theta^TU_+\text{diag}(b_+)U_+^T\theta = u^Tu, \tag{2.23}$$

implying that only coefficients u are penalized with the penalty matrix $I_{K+1+p-q}$. Hence, the mixed model representation is available and results in

$$y|u \sim N(X\beta + Z_Bu, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_{K+1+p-q}).$$

However, the representation (2.22) is not unique due to the singularity of $\Delta_q^T\Delta_q$. In fact, any one-to-one transformation of spline coefficients $B\theta = B[W_\beta\beta + W_uu]$ with matrices W_β and W_u of dimensions $(K+1+p) \times q$ and $(K+1+p) \times (K+1+p-q)$, respectively, such that

- $[W_\beta, W_u]$ is of full rank;
- $W_\beta^TW_u = W_u^TW_\beta = 0$;
- $W_\beta^T\Delta_q^T\Delta_qW_\beta = 0$;
- $W_u^T\Delta_q^T\Delta_qW_u = I_{K+1+p-q}$

can be applied (for more details see Green, 1987 or Fahrmeir, Kneib & Lang, 2004). While the first condition ensures the uniqueness of transformation, the remaining ones provide that only coefficients u are penalized with identity penalty matrix. A common choice has become to define $W_\beta = [1, w, ..., w^{q-1}]$ with $w = (1, 2, ..., K+p+1)^T$ and $W_u = \Delta_q^T(\Delta_q\Delta_q^T)^{-1}$ (see e.g. Durban & Currie, 2003) yielding the transformation

$$B\theta = B[W_\beta\beta + \Delta_q^T(\Delta_q\Delta_q^T)^{-1}u] =: X\beta + Z_Bu.$$

Note, that X results here in a polynomial of degree q.

Radial basis for mixed models

Radial basis smoother

$$\|y - X\beta - Z_R u\|^2 + \lambda u^T \Omega u,$$

can be represented as a mixed model (2.20) with the simple adjustment of the spline basis matrix

$$y|u \sim N(X\beta + Z_R \Omega^{-1/2} u, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_K),$$

since the inverse of the penalty matrix $\Omega = [|k_j - k_1|^{2r-1}, ..., |k_j - k_K|^{2r-1}]_{1 \le j \le K}$ is a proper covariance matrix.

2.3 Confidence Intervals

Let us first consider the conditional model $y \sim N(m(x), \sigma_\epsilon^2 I_n)$, with $m(x) = X\beta + Zu$, where the parameters β and u are assumed to be fixed and estimated from penalized least squares, yielding $\hat{m}(x) = S_\lambda y$. With this, $\text{Cov}[\hat{m}(x)] = \sigma_\epsilon^2 S_\lambda S_\lambda^T$. The confidence intervals can now be constructed from

$$\hat{m}(x) \sim N(E[\hat{m}(x)], \sigma_\epsilon^2 S_\lambda S_\lambda^T). \tag{2.24}$$

However, the confidence interval obtained from (2.24) covers $E[\hat{m}(x)]$ rather than $m(x)$, since in the conditional model $\hat{m}(x)$ is not an unbiased estimate for $m(x)$.
The confidence bands resulting due to mixed model representation of penalized splines, i.e. from the marginal model $y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n)$, $u \sim N(0, \sigma_u^2 I_K)$, do not have such a drawback. In the mixed model framework the function $m(x) = X\beta + Zu$ is random due to the randomness of u and $\tilde{m}(x) = X\tilde{\beta} + Z\tilde{u}$ is unbiased for $m(x)$. Thus, we can construct the confidence interval for $m(x)$ from

$$\tilde{m}(x) - m(x) \sim N(0, \sigma_\epsilon^2 S_\lambda),$$

where the $\text{Cov}[\tilde{m}(x) - m(x)] = \sigma_\epsilon^2 S_\lambda$ according to (2.19). Replacing σ_ϵ^2 with the estimate results in the 95 % confidence bands for $m(x_i)$ as

$$\hat{m}(x_i) \pm 2\hat{\sigma}_\epsilon \sqrt{S_\lambda^{ii}}.$$

Thus, at some fixed x the confidence bands in conditional model use diagonal elements of $S_\lambda S_\lambda$, while in the mixed model framework variability bands are bias adjusted and use diagonal entries of S_λ.

2.4 Extensions

2.4.1 Additive Models

Additive models introduced by Hastie & Tibshirani (1990) extend the usual multiple linear models by incorporating nonparametric techniques. More precisely, these models allow for nonlinear covariate effects which remain additive; i.e.

$$y_i \sim N(\beta_0 + m_1(x_{i1}) + ... + m_d(x_{id}), \sigma_\epsilon^2), \quad i = 1, ..., n$$

with $m_l(.)$ as smooth, but otherwise unspecified, functions of n dimensional covariates x_l, $l = 1, ..., d$. A widely used approach for fitting additive models is the backfitting algorithm introduced in Hastie & Tibshirani (1990). Thereby, functions $m_l(.)$ are estimated iteratively by smoothing partial residuals, which arise from the model without $m_l(.)$. However, the penalized spline approach allows for the fitting of additive models simultaneously for all variables. For example, modelling $m_l(.)$ with a linear truncated polynomial basis, we represent each function as $m_l(x_l) = x_l\beta_l + Z_l u_l$, with K_l dimensional vectors u_l, $l = 1, ..., d$. Hence, we can estimate parameters either from

$$\|y - X\beta - Zu\|^2 + \lambda_1 u_1^T u_1 + ... + \lambda_d u_d^T u_d \tag{2.25}$$

or from the linear mixed model

$$y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \text{blockdiag}[\sigma_{u_1}^2 I_{K_1}, ..., \sigma_{u_d}^2 I_{K_d}]),$$

with $\beta = (\beta_0, \beta_1, ..., \beta_d)$, $u = (u_1^T, ..., u_d^T)^T$, $X = [1, x_{i1}, ..., x_{id}]_{1 \le i \le n}$ and $Z = [Z_1, ..., Z_d]$, where Z_l is the linear truncated polynomial basis matrix of dimension $n \times K_l$. Both models result in the estimate $\hat{y} = C(C^T C + \bar{D})^{-1} C^T y$ with $C = [X, Z]$ and $\bar{D} = \text{blockdiag}[0_{(d+1)\times(d+1)}, \lambda_1 I_{K_1}, ..., \lambda_d I_{K_d}]$. As in the univariate case, in the mixed model framework smoothing parameter $\lambda_l = \sigma_\epsilon^2 / \sigma_{u_l}^2$ is a ratio of variances.

Additive models with B-spline basis

Fitting additive models with B-splines is described in detail in Durban & Currie (2003). To achieve identifiability we center the B-spline basis matrices B_l of degree p_l based on K_l knots leading to $\tilde{B}_l = (I - 1_n 1_n^T/n)B_l$, $l = 1, ..., d$. This gives the model

$$y \sim N(B\theta, \sigma_\epsilon^2 I_n),$$

with $B\theta = \beta_0 1_n + \tilde{B}_1\theta_1 + ... + \tilde{B}_d\theta_d$, which is estimated in the penalized form

$$\|y - B\theta\|^2 + \lambda_1 \theta_1^T \Delta_{q_1}^T \Delta_{q_1} \theta_1 + ... + \lambda_d \theta_d^T \Delta_{q_d}^T \Delta_{q_d} \theta_d.$$

To use the mixed model representation the coefficient θ have to be modified so that $B\theta = B(W_\beta \beta + W_u u)$ with matrices W_β and W_u having those properties described in Section 2.2.2 and $u \sim N(0, \text{blockdiag}[\sigma_{u_1}^2 I_{K_1}, ..., \sigma_{u_d}^2 I_{K_d}])$. Durban & Currie (2003) suggested to choose matrix $W_u = \text{blockdiag}(W_{u_1}, ..., W_{u_d})$ with $W_{u_l} = \Delta_{q_l}^T (\Delta_{q_l}^T \Delta_{q_l})^{-1}$ and matrix $W_\beta = \text{blockdiag}(1, W_{\beta_1}, ..., W_{\beta_d})$ with $W_{\beta_l} = (w_l, w_l^2, ..., w_l^{q_l - 1})$ and $w_l = (1, 2, ..., K_l + p_l + 1)$, $l = 1, ..., d$.

Additive models with radial basis

Modelling $m_l(x_l) = \beta_l x_l + Z_{R_l} u_l$, $l = 1, ..., d$ we can estimate the model either from penalized least squares

$$\|y - X\beta - Z_R u\|^2 + \lambda_1 u_1^T \Omega_1 u_1 + ... + \lambda_d u_d^T \Omega_d u_d$$

or from the linear mixed model

$$y|u \sim N(X\beta + Z_R \Omega^{-1/2} u, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \text{blockdiag}[\sigma_{u_1}^2 I_{K_1}, ..., \sigma_{u_d}^2 I_{K_d}]),$$

with $X = [1, x_{i1}, ..., x_{id}]_{1 \le i \le n}$, $Z_R = [Z_{R_1}, ..., Z_{R_d}]$ and $\Omega = \text{blockdiag}[\Omega_1, ..., \Omega_d]$, where Z_{R_l} and Ω_l are respectively radial basis function and corresponding penalty matrix based on K_l knots.

Smoothing parameters estimation

In the mixed model framework smoothing parameters selection brings no additional challenges - variance parameters σ_ϵ^2 and $\sigma_{u_l}^2$ are estimated from the (restricted) likelihood along with the other parameters. However, estimation of smoothing parameters

in the conditional model, e.g. (2.25), requires solving a d-dimensional optimization problem. The BRUTO algorithm introduced in Hastie & Tibshirani (1990) combines backfitting and smoothing parameter selection, avoiding a d-dimensional minimization. The method, applicable to general linear smoothers, is based on iterative univariate minimization of the GCV criterion for additive models. For smoothing splines Gu & Wahba (1991) suggested a modified Newton procedure, which is extended in Wood (2000) and Wood (2004) to a wider range of smoothers, including penalized splines. This approach is implemented by S.N. Wood in his R package *mgcv*, which is discussed in Section 2.5.3. In general, it can be difficult to find a global minimum when d is large.

2.4.2 Bivariate Smoothing

The general bivariate smoothing model has the form

$$y_i \sim N(m(x_{i1}, x_{i2}), \sigma_\epsilon^2), \quad i = 1, ..., n.$$

Clearly bivariate smoothing requires bivariate basis functions and there are essentially two possible approaches: tensor product of one dimensional truncated polynomials (see Ruppert, Wand & Carroll, 2003) or B-spline basis functions and radial basis functions defined in two dimensions, known also as low rank thin plate splines. We consider these methods consecutively.

Tensor product of truncated polynomials bases

Since it is common in linear regression to model interaction by adding a product term $y \sim N(\beta_0+\beta_1^1 x_1+\beta_1^2 x_2+\beta^{12} x_1 x_2, \sigma_\epsilon^2 I_n)$, the natural extension for truncated polynomials would be to represent the regression function by

$$\begin{aligned} m(x_1, x_2) &= \beta_0 + \beta_1^1 x_1 + \sum_{i=1}^{K_1} u_i^1 (x_1 - k_i^1)_+ + \beta_1^2 x_2 + \sum_{i=1}^{K_2} u_i^2 (x_2 - k_i^2)_+ \\ &+ \beta^{12} x_1 x_2 + x_2 \sum_{i=1}^{K_1} u_i^1 (x_1 - k_i^1)_+ + x_1 \sum_{i=1}^{K_2} u_i^2 (x_2 - k_i^2)_+ \\ &+ \sum_{i=1}^{K_{12}} u_i^{12} (x_1 - k_i^1)_+ (x_2 - k_i^2)_+. \end{aligned}$$

Thus, the basis functions are formed as a tensor product of two one dimensional bases

$$C_1 = [1, x_{i1}, (x_{i1} - k_1^1)_+, ..., (x_{i1} - k_{K^1}^1)_+]_{1 \leq i \leq n},$$

$$C_2 = [1, x_{i2}, (x_{i2} - k_1^2)_+, ..., (x_{i2} - k_{K^2}^2)_+]_{1 \leq i \leq n},$$

which for any polynomial degree can be written as

$$C_{12} = C_1 \otimes C_2 = [X_1, Z_1] \otimes [X_2, Z_2],$$

with C_1 and C_2 denoting basis matrices of truncated polynomials of degree p for covariates x_1 and x_2 respectively. Rearranging the basis matrix according to the parameter vector $\theta = (\beta^T, u_1^T, u_2^T, u_{12}^T)^T$ as

$$C_{12} = [X_1 \otimes X_2, X_2 \otimes Z_1, X_1 \otimes Z_2, Z_1 \otimes Z_2]$$

and estimating the parameters from the penalized likelihood

$$\|y - C_{12}\theta\|^2 + \lambda_1 u_1^T u_1 + \lambda_2 u_2^T u_2 + \lambda_{12} u_{12}^T u_{12}$$

or from the linear mixed model

$$y|u \sim N(C_{12}\theta, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \text{blockdiag}[\sigma_{u_1}^2 I_{K_1}, \sigma_{u_2}^2 I_{K_2}, \sigma_{u_{12}}^2 I_{K_{12}}]),$$

we get the following estimate $\hat{y} = C_{12}({C_{12}}^T C_{12} + \bar{D})^{-1} {C_{12}}^T y$ with the penalty matrix $\bar{D} = \text{blockdiag}[0_{(p+1)^2 \times (p+1)^2}, \lambda_1 I_{K_1}, \lambda_2 I_{K_2}, \lambda_{12} I_{K_{12}}]$. Thus, one can consider bivariate smoothing as a type of additive model with three smoothing parameters to be chosen: one for each covariate and one for an interaction effect. Consequently, smoothing parameter estimation methods described in the previous section can be directly applied to the bivariate smoothing.

Tensor product of B-splines

A truncated polynomial basis is not necessarily a perfect choice for bivariate smoothing when it comes to numerical stability. One may prefer to perform bivariate smoothing with a superior alternative, e.g. with the tensor product of B-splines $B_{12}(x_1, x_2) =$

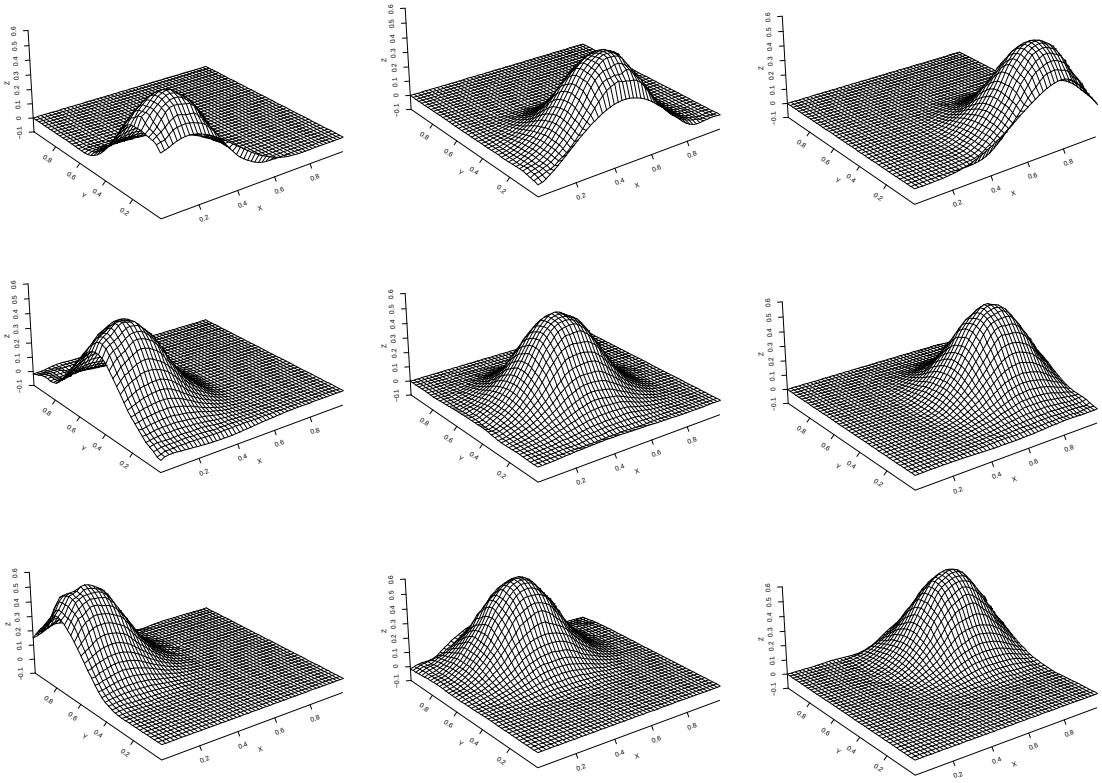

Figure 2.7: Tensor product of B-splines of degree 2 at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$.

$B^1(x_1) \otimes B^2(x_2)$ (see Dierckx, 1993 or de Boor, 2001), representing

$$m(x_1, x_2) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} B_i^1(x_1) B_j^2(x_2) \theta_{ij}.$$

Figure (2.7) shows a part of B-splines tensor product basis at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$. Parameters can now be estimated from the penalized least squares

$$\|y - B_{12}\theta\|^2 + \lambda_1 \theta^T (\Delta_{q_1}^T \Delta_{q_1} \otimes I_{K_2})\theta + \lambda_2 \theta^T (I_{K_1} \otimes \Delta_{q_2}^T \Delta_{q_2})\theta,$$

where smoothing parameters λ_1 and λ_2 steer the amount of smoothing in directions x_1 and x_2 respectively. Eilers & Marx (2003) also suggest adding an overall penalty $\lambda_{12}\theta^T\theta$. Mixed model representation can be derived similar to additive models.

Low-rank radial smoother

Extension from a one-dimensional radial basis to a bivariate case is straightforward, since basis functions depend only on distance between observations and knots $|x - k_j|$. Having $\mathbf{x}_i, \mathbf{k}_j \in R^2$, $i = 1, ..., n$, $j = 1, ..., K$, we just replace the absolute value with the Euclidean norm and define

$$Z_R = \left[\|\mathbf{x}_i - \mathbf{k}_j\|^{2r-2} \log \|\mathbf{x}_i - \mathbf{k}_j\|\right]_{1\leq i\leq n,\, 1\leq j\leq K}, \Omega = \left[\|\mathbf{k}_s - \mathbf{k}_t\|^{2r-2} \log \|\mathbf{k}_s - \mathbf{k}_t\|\right]_{1\leq s,t\leq K},$$

with $r \geq 2$. Addition of the log(.) factor arises from the multivariate extension of integrated squared derivative penalty of natural cubic splines (see Section 2.1.2). Figure 2.8 represents radial basis functions at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$. The parameter estimates can be obtained either from penalized least squares

$$\|y - X\beta - Z_R u\|^2 + \lambda u^T \Omega u$$

or from the linear mixed model

$$y|u \sim N(X\beta + Z_R \Omega^{-1/2} u, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_K).$$

This smoothing method (also referred to as low-rank thin plate smoother) is closely related to kriging and is rotationally invariant, which is important for geographical smoothing. For more details and discussion see Ruppert, Wand & Carroll (2003).

2.4.3 Smoothing with Generalized Response

Let us consider the following generalized response model

$$E(y|x) = \mu(x) = h\left[m(x)\right], \ \mathrm{Var}(y|x) = \phi v(\mu),$$

with function $h(.)$ as the inverse of a link function, $v(.)$ as some specified variance function and ϕ as a dispersion parameter. We assume that the observations are drawn from the exponential family

$$y|x \sim \exp\left\{(y\vartheta(x) - b\left[\vartheta(x)\right])/\phi + c(y, \phi)\right\}.$$

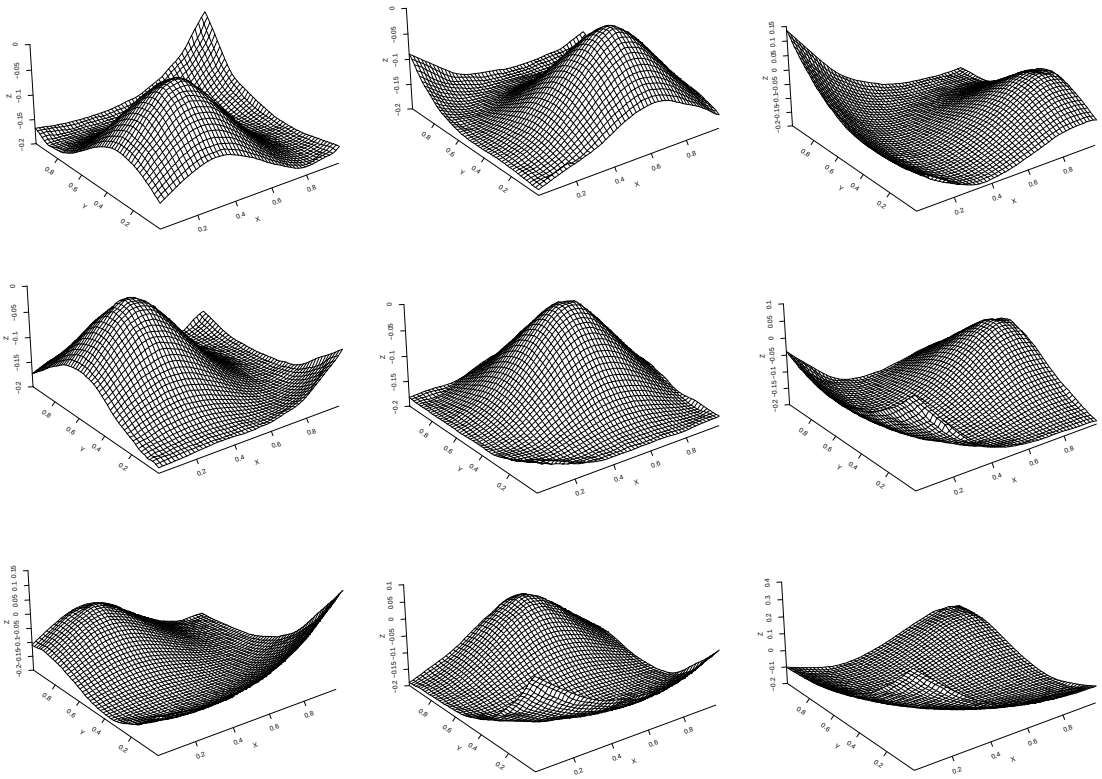

Figure 2.8: Low-rank radial splines at knots $k_1^{x_1} = k_1^{x_2} = 0.3$, $k_2^{x_1} = k_2^{x_2} = 0.6$ and $k_3^{x_1} = k_3^{x_2} = 0.9$.

Functions $\vartheta(x)$ and $\mu(x)$ stand in the unique relationship $b'(\vartheta) = \mu$, so that $\vartheta = b'^{-1}\{h[m(x)]\}$. Choosing $h(.) = b'(.)$ results in a generalized model with the natural link. Modelling $m(x)$ with penalized splines using for example truncated polynomial basis we can estimate the parameter either from penalized log-likelihood

$$\frac{1}{\phi}\left\{y^T\vartheta(X\beta + Zu) - 1_n^T b\left[\vartheta(X\beta + Zu)\right]\right\} - \frac{\lambda}{2}u^T u \tag{2.26}$$

or from the generalized linear mixed model

$$E(y|x,u) = \mu^u(x) = h(X\beta + Zu), \quad u \sim N(0, \sigma_u^2 I_K). \tag{2.27}$$

Other bases and penalties are easily applicable. We consider parameter estimation from (2.26) and (2.27) consecutively.

Estimation with penalized log-likelihood

Estimating parameters of the model (2.26) by applying the Fisher scoring with the score function

$$\frac{\partial l(\theta)}{\partial \theta} = C^T W[h^{-1}(\mu)]'(y-\mu) - \lambda D\theta$$

and information matrix

$$E\left(-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right) = C^T WC + \lambda D$$

is equivalent to the *iterated weighted least squares* (IWLS)

$$\hat{y} = C\theta = C(C^T WC + \lambda D)^{-1}C^T Wz = S_{W,\lambda}z, \tag{2.28}$$

with the working vector $z = C\theta + [h^{-1}(\mu)]'(y-\mu)$ and W as the $n \times n$ diagonal matrix with diagonal elements $w_i = (\phi v_i[h^{-1}(\mu_i)']^2)^{-1}$. The smoothing parameter selection can be carried out with GCV criterion, with the usual residuals replaced by Pearson residuals (see O'Sullivan, Yandell & Raynor, 1986 or Green & Silverman, 1994)

$$GCV_G = \frac{1}{n}\sum_{i=1}^{n} w_i \left\{\frac{y_i - \mu_i}{1 - \mathrm{tr}(S_{W,\lambda})/n}\right\}^2,$$

or with Akaike criterion

$$AIC_G = \frac{1}{n}\sum_{i=1}^{n} D_i(y_i, \hat{\mu}_i) + 2\mathrm{tr}(S_{W,\lambda})/n,$$

where D is the model deviance and $S_{W,\lambda}$ is the smoothing matrix resulting from the last iteration.

Smoothing parameter selection with the above criteria is carried out with the grid search. One fits models with the Fisher scoring (or equivalently with IWLS) for different values of smoothing parameters and chooses the smoothing parameter with the smallest GCV_G or AIC_G value. An alternative approach discussed in Green & Silverman (1994) is to invert the order of (i) iteration to update parameters and (ii) minimization of GCV_G (or AIC_G). That is, within each cycle of the Fisher scoring algorithm the smoothing parameter is chosen as if the current linear problem were the original model. There is some evidence (see e.g. Gu, 1992) that the second approach gives better results. Moreover, this method is more efficient numerically. In Section 3.6 we provide the implementation of the latter approach in R using Demmler-Reinsch orthogonalisation, discussed in Sec-

tion 2.5.2.

Generalized linear mixed models

Parameter estimation of the model (2.27) has to be performed from the likelihood

$$L(\beta, \sigma_u^2) = \int_{R^K} f(y|u)f(u)du = (2\pi)^{-K/2} \exp[1_n^T c(y, \phi)] J(\beta, \sigma_u^2),$$

with

$$\begin{aligned} J(\beta, \sigma_u^2) &= \int_{R^K} \exp\left\{ y^T \vartheta(X\beta + Zu)/\phi - 1_n^T b[\vartheta(X\beta + Zu)]/\phi - \frac{u^T u}{2\sigma_u^2} - \frac{K}{2} \log \sigma_u^2 \right\} du \\ &= \int_{R^K} \exp\{-g(u)\}\, du. \end{aligned} \tag{2.29}$$

Such intractable integrals can either be solved with MCMC based techniques or approximated with Laplace's method, leading to the penalized quasi likelihood (PQL). The latter approach is presented in Breslow & Clayton (1993). Laplace approximation applied to (2.29) results in the following log-likelihood

$$-2l(\beta, \sigma_u^2) \approx -\frac{2}{\phi} \left\{ y^T \vartheta(X\beta + Z\hat{u}) - 1_n^T b[\vartheta(X\beta + Z\hat{u})] \right\} + \frac{\hat{u}^T \hat{u}}{\sigma_u^2} + K \log(\sigma_u^2) + \log |I_{uu}|, \tag{2.30}$$

where $\hat{u}$ solves $\partial g(u)/\partial u = 0$ and $I_{uu} = E[\partial^2 g(\hat{u})/\partial u \partial u^T] = (C^T W C + D/\sigma_u^2)$. Minimization of (2.30) over β and u with the Fisher scoring is identical to the iterated weighted least squares (2.28) with $\lambda = 1/\sigma_u^2$. The variance parameter estimate results in

$$\hat{\sigma}_u^2 = \frac{\hat{u}^T \hat{u}}{\operatorname{tr}\left[C^T W C (C^T W C + D/\sigma_u^2)^{-1}\right]}. \tag{2.31}$$

Note, that (2.31), as well as (2.28), are not explicit solutions. Thus, the parameter estimation is carried out by iterating back and forth, getting $(\hat{\beta}, \hat{u})$ from (2.28) and the variance parameter from (2.31). More details are available in Breslow & Clayton (1993) and Searle, Casella, & McCulloch (1992).
It should be noted, Shun & McCullagh (1995) showed that the Laplace approximation can be applied to the integrals of the form (2.29) without correction only if the order of random effects dimension is at most $K = o(n^{1/3})$. More detailed treatment of this problem is provided in Section 6.3.1.

Note, that the extensions to the generalized additive model

$$E(y|x_1, ..., x_d) = \mu(x_1, ..., x_d) = h\left[m_1(x_1) + ... + m_l(x_d)\right]$$

as well as to generalized bivariate smoothing

$$E(y|x_1, x_2) = \mu(x_1, x_2) = h\left[m(x_1, x_2)\right]$$

are straightforward. Redefining model matrices X and Z and penalties as described in Sections 2.4.1 and 2.4.2 immediately delivers the corresponding estimates.

2.5 Computational Issues

Flexibility of penalized smoothing and its ties to mixed and Baysian models lead to a vast choice of numerical techniques for estimation and inference. Some of them are the subject of this section.

2.5.1 Choice of Knots and Basis

As was mentioned in Section 2.1.3, one has to make the following choice when applying penalized splines:

- amount and location of knots;
- spline basis functions;
- degree of the spline and the penalty matrix.

Penalized splines are low-rank smoothers, i.e. amount of knots used for estimation is far less than the number of observations, which significantly reduces the numerical effort. The default choice $\min\{n/4, 40\}$ is suggested in Ruppert (2002) and is commonly used. Based on a simulation study Ruppert (2002) stated that "there must be enough knots to fit the features in the data, but after this minimum necessary number of knots has been reached, further increases in K often have little effect on the fit". The asymptotic order of spline basis dimension K, so that the mean squared error is optimized will be investigated in Chapter 6.
The best type of knot spacing for scatterplot smoothing - equidistant or quantile - is

still controversial. Eilers & Marx (1996) and Eilers & Marx (2004) stress that "equally-spaced knots are always to be preferred", while Ruppert, Wand & Carroll (2003) emphasize utilization of quantile-based knots. Eilers & Marx (2004) provided an example where equally spaced knots were superior to quantile spaced. Crainiceanu, Ruppert & Carroll (2005) doubted that "any type of spacing is always best" and presented an example where quantile spacing outperformed equidistant knots noticeably. In general, both approaches do the work equally well for most of the examples. For bivariate smoothing there are several approaches to knot spacing available, e.g. space filling algorithm by Nychka & Saltzman (1998) or *clara* algorithm of Kaufman & Rousseeuw (1990). The latter is implemented in R package *cluster*. Equidistant knots for bivariate smoothing are less used due to their inefficiency.

One also needs to make a choice about spline basis. Truncated polynomials are useful for understanding spline regression, but direct estimation with this basis can lead to numerical problems. However, transformation to a more stable version is available (see next section), resulting, in fact, in another basis. In contrast, B-splines seem to be a superior alternative - they are numerically stable, easy to calculate and as noticed Eilers & Marx (2004) "allow informative visualization" (see their Figure 3). Radial basis functions are computationally efficient and stable as well. In the Bayesian framework this basis also provided a better MCMC convergence, as noted by Crainiceanu, Ruppert & Carroll (2005).

Finally, the degree of the spline and the penalty matrix need to be determined. Truncated polynomials are less flexible in this respect. The penalty matrix is just an identity matrix and polynomial degree defines, at the same time, the limiting fit. In contrast, B-splines allow for separate choice of the spline degree p and the order of the penalty q. The latter is the discrete approximation of integrated squared qth derivative and defines the limiting estimate: large smoothing parameter with $q+1$ order penalty shrink the fit toward a polynomial of degree q. This property of B-splines allows for a greater flexibility of the smoothing model. Different versions of radial basis are established in the most R packages, e.g. in *mgcv* and *SemiPar*. While *SemiPar* allows varying the spline degree only, the *mgcv* version lets one choose the penalty order q as well (with the condition $2q \geq d$ to be fulfilled, where d is the number of covariates).

In general, in most examples the knots' location, basis and penalty types have no noticeable effect on the fit, in so far as the amount of knots is sufficiently large. However, for complex problems like smoothing of regression functions with strong varying local variability or with sparse data in some regions, some care is needed for knots and basis selection.

2.5.2 Fast and Stable Penalized Smoothing

An optimal smoothing parameter, if not chosen in the mixed model or Bayesian framework, has to be identified by grid searching, which can be numerically exhaustive, especially for large datasets. Moreover, some basis functions, like truncated polynomials, can be numerically unstable. The Demmler-Reinsch orthogonalisation allows one not only to speed up computations for smoothing parameter selection, but also to stabilize it numerically. A variation of this algorithm is used to compute smoothing splines (see Eubank, 1988 or Green & Silverman, 1994). Following Ruppert, Wand & Carroll (2003) we give here the algorithm for some basis matrix C:

1. Obtain Cholesky decomposition of $C^TC = K^TK$ with square and invertible K;
2. Obtain singular value decomposition of $K^{-T}DK^{-1} = U\text{diag}(b)U^T$;
3. Get the smoothing matrix as $S_\lambda = A\,\text{diag}(1+\lambda b)^{-1}A^T$ with $A = CK^{-1}U$.

Note that $A^TA = U^TK^{-T}C^TCK^{-1}U = I_n$, yielding $df = \text{tr}(S_\lambda) = 1_n^T(1+\lambda b)^{-1}$. The beauty of this approach is that matrix A and vector b have to be calculated just once and then these quantities can be used for all values of the smoothing parameter λ. For justification of this algorithm one just needs to note that

$$C^TC + \lambda D = K^TK + \lambda D = K^T(I + \lambda K^{-T}DK^{-1})K = K^TU(I + \lambda\text{diag}[b])U^TK.$$

Now the calculation of GCV or AIC criteria as given in (2.10) or (2.12) can be made for all values of λ at once. For implementation of Demmler-Reinsch algorithm in R and MATLAB see Appendix B of Ruppert, Wand & Carroll (2003).
This approach can also be used to optimize calculations in the BRUTO algorithm of Hastie & Tibshirani (1990) for smoothing parameter choice in additive models, as well as for bivariate smoothing. The optimization in this case is performed iteratively for one smoothing parameter at a time. Moreover, since generalized penalized spline smoothing can be represented as iterated weighted penalized least squares, it is straightforward to adjust the Demmler-Reinsch approach for generalized penalized spline smoothing. Implementation in R is provided in Section 3.6.

2.5.3 R Packages *mgcv*, *nlme* and *SemiPar*

This section demonstrates how penalized spline smoothing can be performed with packages available in R.

Function `gam`

Function `gam` of R package *mgcv*, written by S.N. Wood, implements (generalized additive) penalized spline smoothing with a smoothing parameter chosen with GCV. It handles (generalized) scatterplot, additive and multivariate smoothing with flexible choice of knots and bases. For example, estimating the parameters of penalized likelihood $\|y - C(x)\theta\|^2 + \lambda\theta^T D\theta$ with thin plate penalized splines based on $K = 40$ knots and second order penalty using automatic smoothing parameter choice is carried out with

```
> library(mgcv)
> y.fit <- gam(y~s(x,k=40,bs="ts",m=2)).
```

More details are given in Wood (2004) and in the documentation to the *mgcv* package, available at CRAN pages (http://cran.r-project.org/doc/packages/mgcv.pdf). In general, function `gam` is fast and time-tested.
Note that R function `gam` is fundamentally different from the Splus function `gam` based on the work Hastie & Tibshirani (1990). Splus function uses smoothing splines or loess-smoother (local polynomial regression) for fitting and has no option for automatic choice of smoothing parameters.

R functions for mixed models

R package *nlme* with its function `lme` allows for the fitting of linear mixed models of any complexity. To use this function for smoothing one needs first to define the model matrices X and Z in accordance with the smoother type. For example, to fit the model

$$y|u \sim N(X\beta + Zu, \sigma_\epsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_K)$$

with squared truncated polynomials based on $K = 40$ equidistant knots, we first set up the model matrices

```
> K <- 40
> n <- length(y)
> st <- (max(x)-min(x))/(K+1)
> knot.x <- seq(min(x)+st,max(st)-st,by=st)
> Z <- outer(x,knot.x,"-")
> Z <- (Z*(Z>0))^2
```

```
> X <- cbind(rep(1,n),x,x^2)
```

and perform the penalized fit as follows

```
> library(nlme)
> all <- rep(1,n)
> y.fit <- lme(y~X-1,random=list(all=pdIdent(~Z-1))
```

Note that the `lme` functions in R and Splus have different formats, e.g. fitting the same model in Splus has to be performed with

```
> y.fit <- lme(y~X-1,random=pdIdent(~Z-1))
```

Moreover, Splus version of `lme` makes use of a different optimizer, which is faster and more robust. John Fox in the online appendix to his book Fox (1997) (see http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf) also notices that "comparing the maximized log-likelihoods for the two programs suggests that the R version of `lme` has not converged fully to the REML solution".
The estimates for β, u, σ_ϵ^2, σ_u^2 from both (Splus and R) fits can be obtained with

```
> beta.hat <- y.fit$coef$fixed
> u <- as.vector(unlist(y.fit$coef$random))
> sigma.sq.eps.hat <- y.fit$sigma^2
> sigma.sq.u.hat <- sigma.sq.eps.hat*exp(2*unlist(y.fit$modelStruct)).
```

For fitting the models with generalized response, e.g.

$$E(y|x,u) = h(X\beta + Zu), \quad u \sim N(0, \sigma_u^2 I_K),$$

function `glmmPQL` of package *MASS* can be used. This function fits generalized mixed models by sequential calls of the function `lme` and consequently has the same format and output. Having y, for example, drawn from the Poisson distribution and model matrices defined as above we can call

```
> library(MASS)
> y.fit <- glmmPQL(y~X-1,random=list(all=pdIdent(~Z-1),family=poisson)
```

The estimated object is of `lme`-class.

Package SemiPar

The R package *SemiPar* was written by M.P. Wand to accompany the book Ruppert, Wand & Carroll (2003). In general its function `spm` facilitates the interface with mixed model functions. Based on `lme` (or `glmmPQL`) function `spm` performs (generalized) scatterplot, additive and bivariate smoothing with two possible basis functions (truncated polynomials and low-rank thin plate splines) and flexible choice of knots. A comprehensive user's manual is available at http://www.maths.unsw.edu.au/∼wand/SPmanu.pdf. For example, a scatterplot fit with a squared truncated polynomial basis based on 40 knots can be performed with

```
> knot.x <- default.knots(x,40)
> y.fit <- spm(y~f(x,knots=knot.x,basis="trunc.poly",degree=2))
```

The parameter estimates can be extracted from `y.fit$fit`, which is the `lme`-class object as above.
There are two more functions in R which make use of `lme` (or `glmmPQL`) - these are `lmeSplines` and `gamm`. Both functions are not as general as `spm`, however `gamm` is the most similar to it. Function `gamm` is a part of the S.N. Wood's *mgcv* package and has the same format as `gam`, which makes it convenient for comparison.
Finally, one should be aware that the function `gam` and all `lme`-based approaches estimate in two different frameworks, which also results in a different inference and in particular different confidence intervals, as pointed out in Section 2.3.

2.6 Bayesian Model for Smoothing

Specifying in the mixed model

$$y = X\beta + Zu + \epsilon, \qquad \begin{bmatrix} u \\ \epsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 I_K & 0 \\ 0 & \sigma_\epsilon^2 I_n \end{bmatrix} \right)$$

prior distributions on $(\beta, \sigma_u^2, \sigma_\epsilon^2)$ we get a complete Bayesian model for smoothing. It is standard to specify prior distribution for β either $p(\beta) \propto 1$ or $p(\beta) = N(0, \sigma_\beta^2 I_{p+1})$ with a large σ_β^2. Priors for variances σ_u^2 and σ_ϵ^2 are taken from gamma family of priors $\sigma_u^2 \sim IG(A_u, B_u)$ and $\sigma_\epsilon^2 \sim IG(A_\epsilon, B_\epsilon)$ with hyperparameters chosen small to provide noninformative proper prior. With this we get a hierarchical Bayes model.

Estimation

Since $p(\beta, u|y, \sigma_u^2, \sigma_\epsilon^2) \propto p(y|\beta, u, \sigma_u^2, \sigma_\epsilon^2)p(u|\sigma_u^2, \sigma_\epsilon^2)p(\beta|\sigma_u^2, \sigma_\epsilon^2)$, we find that conditional on $(y, \sigma_u^2, \sigma_\epsilon^2)$ the posterior distribution of parameters β and u is proportional to

$$\exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\|y - X\beta - Zu\|^2 + \frac{\sigma_\epsilon^2}{\sigma_u^2}\|u\|^2\right]\right\} = \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\|y - C\theta\|^2 + \frac{\sigma_\epsilon^2}{\sigma_u^2}\theta^T D\theta\right]\right\},$$

which is minimized by (2.21). Denoting with $\Sigma_\theta = (C^TC + \sigma_\epsilon^2 D/\sigma_u^2)^{-1}$ and representing

$$p(\beta, u|y, \sigma_u^2, \sigma_\epsilon^2) \propto \exp\left[-\frac{1}{2\sigma_\epsilon^2}\left\{\left[\theta - \Sigma_\theta C^T y\right]^T \Sigma_\theta^{-1}\left[\theta - \Sigma_\theta C^T y\right]\right\}\right]$$

we find

$$\theta|y, \sigma_u^2, \sigma_\epsilon^2 \sim N\left(\Sigma_\theta C^T y, \sigma_\epsilon^2 \Sigma_\theta\right). \tag{2.32}$$

Thus, the posterior distribution of $m(x) = C\theta$ given $(y, \sigma_u^2, \sigma_\epsilon^2)$ is normal with

$$\begin{aligned} E\left[m(x)|y, \sigma_u^2, \sigma_\epsilon^2\right] &= C(C^TC + \frac{\sigma_\epsilon^2}{\sigma_u^2}D)^{-1}C^Ty = Sy = C\tilde{\theta}, \\ \mathrm{Cov}\left[m(x)|y, \sigma_u^2, \sigma_\epsilon^2\right] &= \sigma_\epsilon^2 C(C^TC + \frac{\sigma_\epsilon^2}{\sigma_u^2}D)^{-1}C^T = \sigma_\epsilon^2 S. \end{aligned} \tag{2.33}$$

Now from

$$\begin{aligned} p(\sigma_\epsilon^2|y, \beta, u, \sigma_u^2) &\propto p(y|\beta, u, \sigma_u^2\sigma_\epsilon^2)p(\sigma_\epsilon^2|\beta, u, \sigma_u^2), \\ p(\sigma_u^2|y, \beta, u, \sigma_\epsilon^2) &\propto p(y|\beta, u, \sigma_u^2\sigma_\epsilon^2)p(\sigma_u^2|\beta, u, \sigma_\epsilon^2) \end{aligned}$$

we get that posterior distributions of variance parameters σ_ϵ^2 and σ_u^2 given (y, u, β) are respectively proportional to

$$\sigma_\epsilon^{-2(n/2+A_\epsilon+1)} \exp\left\{-\frac{1}{\sigma_\epsilon^2}\left(\|y - X\beta - Zu\|^2 + B_\epsilon\right)\right\}$$

$$\sigma_u^{-2(K/2+A_u+1)} \exp\left\{-\frac{1}{\sigma_u^2}\left(\|u\|^2 + B_u\right)\right\}.$$

Comparing these distributions with the inverse gamma density

$$IG(x; a, b) = b^a x^{(a+1)} \exp(-b/x)/\Gamma(a)$$

we determine

$$\begin{aligned}\sigma_\epsilon^2|y,\beta,u,\sigma_u^2 &\sim IG\left(A_\epsilon+\frac{n}{2},B_\epsilon+\frac{1}{2}\left\|y-X\beta-Zu\right\|^2\right)\\ \sigma_u^2|y,\beta,u,\sigma_\epsilon^2 &\sim IG\left(A_u+\frac{K}{2},B_u+\frac{1}{2}\left\|u\right\|^2\right).\end{aligned}$$

Estimation is carried out in that the MCMC algorithm iterates between sampling conditional on the data y the regression coefficients (β, u) from (2.32) given the variance components $(\sigma_u^2, \sigma_\epsilon^2)$, and vice a versa. Numerically this procedure can be implemented, e.g. in WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs) or with BayesX (http://www.stat.uni-muenchen.de/∼bayesx). While the latter is designed primarily for handling Bayesian semiparametric regression based on MCMC simulation techniques (see e.g. Lang & Brezger, 2004), the former is more general.
Compared to the mixed model framework the full Bayesian approach can assess the variability due to hyperparameter estimation, which can make a difference in inference for small datasets. On the other hand, one faces the problems specific to the MCMC inference such as determination of burn-in or inspection of the Markov chains' convergence to their stationary distribution. Moreover, the estimate resulting in the full Bayesian framework can be sensitive to the variance hyperpriors.

3 Smoothing with Correlated Errors

3.1 Motivation

It is well known that in the presence of correlated errors standard smoothing parameter selectors fail to work and overfit the data (see for instance Altman, 1990 or Hart, 1991). This has been nicely exposed in Opsomer, Wang & Yang (2001) for a number of smoothing techniques. Overfitting can be avoided by taking the correlation structure explicitly into account for bandwidth selection. This has been demonstrated among others in Wang (1998) for spline smoothing and in Altman (1990), Hart (1991), Beran & Feng (2001) or Ray & Tsay (1997) for local smoothing, see also McMullan, Bowman & Scott (2003) for an applied approach. For penalized spline fitting Currie & Durban (2002) and Durban & Currie (2003) present a strategy for smoothing with correlated errors and selecting the correlation structure based on the likelihood. A Bayesian approach for fitting with correlated errors is found, for instance, in Smith, Wong & Kohn (1998). In general, the correlation structure is unknown in advance and estimation of the correlation structure requires a sufficiently good fit of the mean function. Hence, one is faced with a dilemma in practice. In fact, even small misspecification of the correlation structure can result in serious over (or under) fitting as demonstrated in Opsomer, Wang & Yang (2001). This exhibits an undesirable sensitivity of MSE-based smoothing parameter selectors. The problem of smoothing with correlated errors is most prominent in a time series setting where $x = t$ gives the time and adjacent observations y_t and y_{t+1} are correlated. Typical examples are macro economic time series like inflation or GDP. In this case $m(t)$ gives the (long term) trend which has to be estimated in the presence of correlated residuals. An overview about common trend estimates is provided, for instance, in Fan & Yao (2003). A traditional method for long term trend estimation in time series is the Hodrick & Prescott (1997) (HP) filter, which incorporates a penalization. The latter clearly demands the specification of a penalty (smoothing) parameter. However, to our knowledge, no data driven routine for choosing the penalty parameter in the HP filter has been suggested yet and instead the choice "$\lambda = 1600$" as heuristically suggested by Hodrick & Prescott (1997) is usually used.

We investigate penalized spline smoothing using two different smoothing parameter selectors. First, a classical MSE minimizer, based on the Akaike criterion is used. Secondly, a restricted maximum likelihood (REML) smoothing parameter estimate is used by considering the smoothing model as a linear mixed model with random spline coefficient (see for instance Wand, 2003 or Kauermann, 2004). It is shown in theory and simulations that the latter approach is more recommendable, since REML based smoothing parameter selection is less sensitive to misspecification of the correlation structure compared to MSE based choices. This means, for instance, if data have been mistakenly considered as independent when they are (not too strongly) correlated, a serious overfit using a MSE smoothing parameter selector appears, while the REML estimate is robust and features a satisfactory behaviour. This performance is demonstrated for simulated data in Figure 3.1 upper row, where the two smoothing parameter selectors are applied to autocorrelated data, but mistakenly assumed no correlation when fitting it. Of course, any fit using a misspecified correlation structure is inferior to one which considers the true correlation, regardless of the smoothing parameter selection used. However, the true correlation is typically unknown (unless in simulations) so that the reported superiority of the REML provides a practical advantage when the correlation is not known.

3.2 Smoothing Parameter Selection

3.2.1 Akaike and REML

We consider the smoothing model

$$y_i \sim N(m(x_i), \sigma_\varepsilon^2), \quad i = 1, ..., n, \tag{3.1}$$

with $m(.)$ is a smooth, but unknown function. Estimation of $m(x)$ is carried out by penalized spline (P-spline) smoothing, replacing $m(x)$ in model (3.1) by some high dimensional parametric structure $m(x) = X\beta + Zu$. Here, X is a low dimensional basis, e.g. with rows $X_i^T = (1, x_i)$, while $Z = Z(x)$ is high dimensional, e.g. truncated lines with rows $Z_i = [(x_i - k_1)_+, \ldots, (x_i - k_K)_+]$, where k_j are fixed knots, $j = 1, \ldots, K$. Clearly, other bases can be used. For theoretical investigation the use of truncated polynomials proves, however, to be simpler and is therefore preferred here. With respect to the dimension K we follow the reported results by Ruppert (2002) and assume that the location and number of knots are fixed in advance.

Coefficients β and u are estimated from the penalized least squares

$$(y - C\theta)^T(y - C\theta)/\sigma_\varepsilon^2 + \lambda u^T u, \tag{3.2}$$

with $y = (y_1, ..., y_n)^T$, $C = (X, Z)$ and coefficient $\theta = (\beta^T, u^T)^T$, resulting in the estimate

$$\hat{m}(x) = C\hat{\theta} = C\left(C^T C + \lambda D\right)^{-1} C^T y = S_\lambda y, \tag{3.3}$$

with D as block diagonal matrix built from 0 and I_K with matching dimensions. With S_λ we denote the resulting smoothing matrix. The penalty parameter λ thereby steers the amount of smoothness. A data driven choice for λ is available by minimizing the Akaike criterion

$$AIC(\lambda) = n \log\left[RSS(\lambda)\right] + 2df(\lambda), \tag{3.4}$$

where $RSS(\lambda) = (y - \hat{m}(x))^T (y - \hat{m}(x))$ and $df(\lambda) = \mathrm{tr}(S_\lambda)$. Alternatively, a modified version of the criterion can be used (see Simonoff & Tsay, 1999), but for the sake of simplicity we stay with the simple case here.

The penalized fit can also be motivated by treating u as random coefficient leading to the linear mixed model

$$y|b \sim N(X\beta + Zu, \sigma_\varepsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_K). \tag{3.5}$$

In this case, $\hat{m}(x)$ as given in (3.3) results in a posterior Bayes estimate or Best Linear Unbiased Predictor (BLUP) with $\lambda = \sigma_\varepsilon^2/\sigma_u^2$. Model (3.5) thereby allows one to estimate the smoothing parameter λ by maximizing the likelihood resulting from the linear mixed model. In practice, an adjusted residual maximum likelihood (REML, see Section 2.2.1 or Harville, 1977) shows advantages. In this case λ is chosen by minimizing the negative REML function.

$$-2\, REML(\lambda) = (n - p)\log(\hat{\sigma}^2_{\varepsilon,MM}) + \log|V_\lambda| + \log|X^T\, V_\lambda^{-1}\, X|, \tag{3.6}$$

with p as dimension of β, $\hat{\sigma}^2_{\varepsilon,MM} = (y - X\hat{\beta})^T\, V_\lambda^{-1}\, (y - X\hat{\beta})/(n - p)$ as variance estimate in the mixed model (3.5) and $V_\lambda = I_n + ZD^- Z^T/\lambda$.
Differentiating (3.4) and (3.6) with respect to λ and solving resulting equations we can

define $\hat{\lambda}_{REML}$ as

$$\hat{\lambda}_{REML} = \hat{\sigma}^2_{\varepsilon,MM} \frac{\text{tr}\,(S_\lambda) - p}{\hat{\theta}^T\, D\, \hat{\theta}}, \tag{3.7}$$

while the AIC minimizer can be written as

$$\hat{\lambda}_{AIC} = \hat{\sigma}^2_{\varepsilon} \frac{\text{tr}(S_\lambda - S_\lambda S_\lambda)}{\hat{\theta}^T D(I - \tilde{S}_\lambda)\hat{\theta}}, \tag{3.8}$$

with $\hat{\sigma}^2_\varepsilon = RSS(\lambda)/n$ and $\tilde{S}_\lambda = (C^TC + \lambda D)^{-1}C^TC$. Note that both (3.7) and (3.8) are not explicit solutions, since the right hand sides of the equations depend on $\hat{\lambda}$. We can however use (3.7) iteratively as fixed point iteration by inserting the current estimate in the right hand side of the equation to obtain an update for the left hand side. This approach performs weakly for (3.8). Thus, even though equation (3.8) defines the estimate for our theoretical investigation with $\hat{\lambda}_{AIC}$ inserted on both sides of the equation, we would recommend to minimize (3.4) not using (3.8), but by grid searching. More details on estimates (3.7) and (3.8) are provided in Section 3.5.

3.2.2 Smoothing with misspecified correlation

When the data are in fact correlated with some, typically unknown, correlation structure, one has to incorporate a "working correlation" R in the smoothing parameter selection. Our objective is to explore which of the two above smoothing parameter selectors is more sensitive with respect to misspecification of such a working correlation. Without loss of generality, we explore this point using working zero correlation, that is $R = I_n$, with I_n as identity matrix. Note that if a different working correlation R is used, then observations $y^* = R^{-1/2}y$ show working zero correlation with mean function $m^*(x) = R^{-1/2}m(x)$. This implies that the results derived for zero correlation can directly be transferred to more general settings. Moreover, with $\tilde{R}$ we denote the true unknown correlation of y. Let now $\hat{\lambda}_{AIC}$ and $\hat{\lambda}_{REML}$ be smoothing parameter estimates calculated assuming uncorrelated residuals. We are interested in $E(\hat{\sigma}^2_\varepsilon/\hat{\lambda}_{AIC}|\tilde{R})$ and $E(\hat{\sigma}^2_{\varepsilon,MM}/\hat{\lambda}_{REML}|\tilde{R})$. We show subsequently that $E(\hat{\sigma}^2_{\varepsilon,MM}/\hat{\lambda}_{REML}|\tilde{R})$ is less dependent on $\tilde{R}$ than $E(\hat{\sigma}^2_\varepsilon/\hat{\lambda}_{AIC}|\tilde{R})$. This means that $\hat{\lambda}_{REML}$ varies less if the true (unknown) correlation changes. We consider the smoothing parameters in this form for technical reasons, since their expectations are more tractable. Moreover, in the mixed model framework $\hat{\sigma}^2_{\varepsilon,MM}/\hat{\lambda}_{REML} = \hat{\sigma}^2_u$, which in

fact steers the amount of smoothing. Using (3.7) and (3.8) we find

$$E_{y,u}\left(\frac{\hat{\sigma}^2_{\varepsilon,MM}}{\hat{\lambda}_{REML}}|\tilde{R}\right) = \frac{\sigma^2_\varepsilon \mathrm{tr}[\tilde{R}(S_\lambda - S_\lambda S_\lambda)]}{\lambda_{REML}(\mathrm{tr}(S_\lambda) - p)} + \frac{\mathrm{tr}[E_u(m(x)m(x)^T)(S_\lambda - S_\lambda S_\lambda)]}{\lambda_{REML}(\mathrm{tr}(S_\lambda) - p)}, \tag{3.9}$$

$$E_{y|u}\left(\frac{\hat{\sigma}^2_{\varepsilon}}{\hat{\lambda}_{AIC}}|\tilde{R}\right) = \frac{\sigma^2_\varepsilon \mathrm{tr}[\tilde{R}(S_\lambda - S_\lambda S_\lambda)(I_n - S_\lambda)]}{\lambda_{AIC}\mathrm{tr}(S_\lambda - S_\lambda S_\lambda)} + \frac{\mathrm{tr}[m(x)m(x)^T(S_\lambda - S_\lambda S_\lambda)(I_n - S_\lambda)]}{\lambda_{AIC}\mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}. \tag{3.10}$$

Note that in (3.9) we take expectation with respect to y and u, while in (3.10) coefficient u is treated as given, as it is also indicated by subscripts at the expectation symbols. This means that $\hat{\lambda}_{AIC}$ and (3.10) exist in the smooth model (3.1) with $m(x) = X\beta + Zu$ and u as unknown parameter while $\hat{\lambda}_{REML}$ and (3.9), respectively, are defined in the mixed model (3.5) with u as random coefficient.

Apparently, the mean values (3.9) and (3.10) depend on the unknown correlation structure $\tilde{R}$. Our theoretical investigation is embedded in the following framework. We assume equidistant and ordered covariates x_i with support $[0, 1]$ for simplicity. The correlation matrix $\tilde{R}$ has the form $\tilde{R}_{ij} = r(|i - j|)$ with r as some stationary positive correlation function, descending to zero for $|i - j|$ growing. Note that this implies that the correlation between two fixed points in $[0, 1]$ is decreasing as $n \to \infty$. We parametrise $\tilde{R}$ by some finite dimensional parameter vector ϱ, that is $\tilde{R} = \tilde{R}(\varrho)$, where $\varrho = 0$ stands for independence such that $\tilde{R}(\varrho = 0) = I_n$. The vector ϱ consists of the elements $\varrho = \{r(1), r(2), \ldots, r(s)\}$, taking the correlation function $r(d)$ to be zero or of ignorable size for $d > s$. The idea is now to approximate $E(\cdot|\tilde{R})$ by a first order Taylor series around $E(\cdot|\tilde{R} = I_n)$, that is

$$E(\cdot|\tilde{R}) = E(\cdot|\tilde{R} = I_n) + \left.\frac{\partial E(\cdot|\tilde{R})}{\partial \varrho}\right|_{\varrho=0} \varrho. \tag{3.11}$$

Note that the further terms in Taylor expansion (3.11) are essentially zero which is readily seen from (3.9) and (3.10) and the chosen parametrisation of $\tilde{R}$. The problem of smoothing parameter selection in presence of correlated errors is then mirrored in the

derivate of the expectation. Parameterising $\tilde{R}$ by $\varrho_d = r(d)$, $d = 1, \ldots, s$ allows us to approximate

$$C^T \frac{\partial R}{\partial \varrho_d}\bigg|_{\varrho_d=0} C \approx 2C^T C,$$

This follows since $\left[\frac{\partial R}{\partial \varrho_d}\right]_{ij}\Big|_{\varrho_d=0}$ takes value 1 for $i = j+d$ and $j = i+d$ and is 0 otherwise. Moreover, since basis C is built from continuous functions we get by denoting with C_i the i-th row of C: $C_i = C_j + O(n^{-1})$, assuming $i - j$ to be bounded (e.g. by s). Hence, $C_i^T C_j = C_i^T C_i\{1 + O(n^{-1})\}$. Since $C^T C$ is of finite dimension, we can proceed with (3.11) and differentiate (3.9) and (3.10) with respect to ρ, which yields

$$\nabla_{REML} = \frac{\lambda_{REML}}{\sigma_\varepsilon^2} \frac{\partial E(\hat{\sigma}^2_{\varepsilon,MM}/\hat{\lambda}_{REML}|\tilde{R})}{\partial \varrho_d}\bigg|_{\varrho_d=0} \approx 2\frac{\mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}{\mathrm{tr}(S_\lambda) - p}, \tag{3.12}$$

$$\nabla_{AIC} = \frac{\lambda_{AIC}}{\sigma_\varepsilon^2} \frac{\partial E(\hat{\sigma}^2_{\varepsilon}/\hat{\lambda}_{AIC}|\tilde{R})}{\partial \varrho_d}\bigg|_{\varrho_d=0} \approx 2\frac{\mathrm{tr}[(S_\lambda - S_\lambda S_\lambda)(I - S_\lambda)]}{\mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}, \tag{3.13}$$

with $\lambda_{REML} = E_u(\lambda_{AIC})$ as shown in Section 3.5.3. It remains to show that $\nabla_{AIC} \geq \nabla_{REML}$. We apply first the Demmler-Reinsch decomposition discussed in Section 2.5.2. To do so we write $C^T C = B^T B$, where B is a square and invertible matrix obtained by a Cholesky decomposition and apply a singular value decomposition $B^{-T} D B^{-1} = U\mathrm{diag}(e_l)U^T$, with U as a matrix of eigenvectors and $e = (e_1, e_2, \ldots)$ as corresponding eigenvalues and thus represent the smoothing matrix as $S_\lambda = CB^{-1}U\mathrm{diag}(b_l)(CB^{-1}U)^T$, with $b_l = 1/(1 + \lambda e_l)$. In this notation (3.12) and (3.13) become respectively

$$\nabla_{REML} \approx 2\frac{\sum_i b_i - \sum_i b_i^2}{\sum_j b_j - p},$$

$$\nabla_{AIC} \approx 2\frac{\sum_i b_i - 2\sum_i b_i^2 + \sum_i b_i^3}{\sum_j b_j - \sum_j b_j^2}.$$

Noting further that $b_l \geq 0$ by definition, one finds $\nabla_{AIC} \geq \nabla_{REML}$ since

$$\begin{aligned}
&\left(\sum_i b_i^3\right)\left(\sum_j b_j\right) - p\left(\sum_i b_i(b_i-1)^2\right) \quad \geq \quad \left(\sum_i b_i^2\right)\left(\sum_j b_j^2\right) \\
\Leftrightarrow \quad &\sum_i \sum_{j>i} b_i b_j (b_i - b_j)^2 - p\sum_i b_i(b_i-1)^2 \quad \geq \quad 0 \\
\Leftrightarrow \quad &\sum_{i=1}^{K} \sum_{i<j\leq K} b_i b_j (b_i - b_j)^2 \quad \geq \quad 0.
\end{aligned}$$

The last inequality follows from the fact, that due to the structure of D, the last p eigenvalues $e_{K+1} = .. = e_{K+p} = 0$ and thus $b_{K+1} = ... = b_{K+p} = 1$, with K as dimension of u and p as dimension of β. This proves that the first order derivative in (3.11) is more pronounced for the AIC smoothing parameter choice than for the REML.

3.2.3 Simulation

To illustrate the theoretical findings we ran a number of simulation studies some of each are reported here. Following Wang (1998) and Currie & Durban (2002), we generate

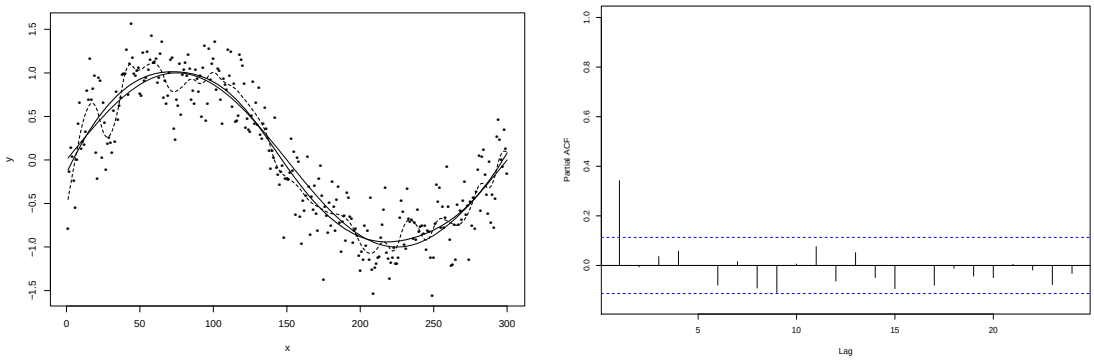

Figure 3.1: Estimated curves with AIC (dashed) and REML (bold) based smoothing parameter choice (left) and partial autocorrelation function corresponding to the true function (right).

$n = 300$ data points with $y_i = \sin(2\pi i/n) + 0.3\epsilon_i$, where ϵ_i, $i = 1, ..., n$ are drawn from a first-order autoregressive process with mean zero, standard deviation one and first-order autocorrelation equal to 0.3. Figure 3.1 shows an exemplary simulation. The smooth fit is based on a quadratic polynomial basis with $K = 40$ knots placed equidistantly over the observed x values. The smoothing parameters are selected assuming independence.

Clearly, the AIC based choice fails to estimate the function properly while the REML estimated smoothing parameter behaves well.
We rerun the simulation with different values for the autocorrelation, ranging from 0 to 0.5 with step size 0.1. Figure 3.2 shows the resulting simulated smoothing parameters $\hat{\lambda}_{AIC}$ and $\hat{\lambda}_{REML}$ on a log scale in a boxplot, each based on 100 simulations. The

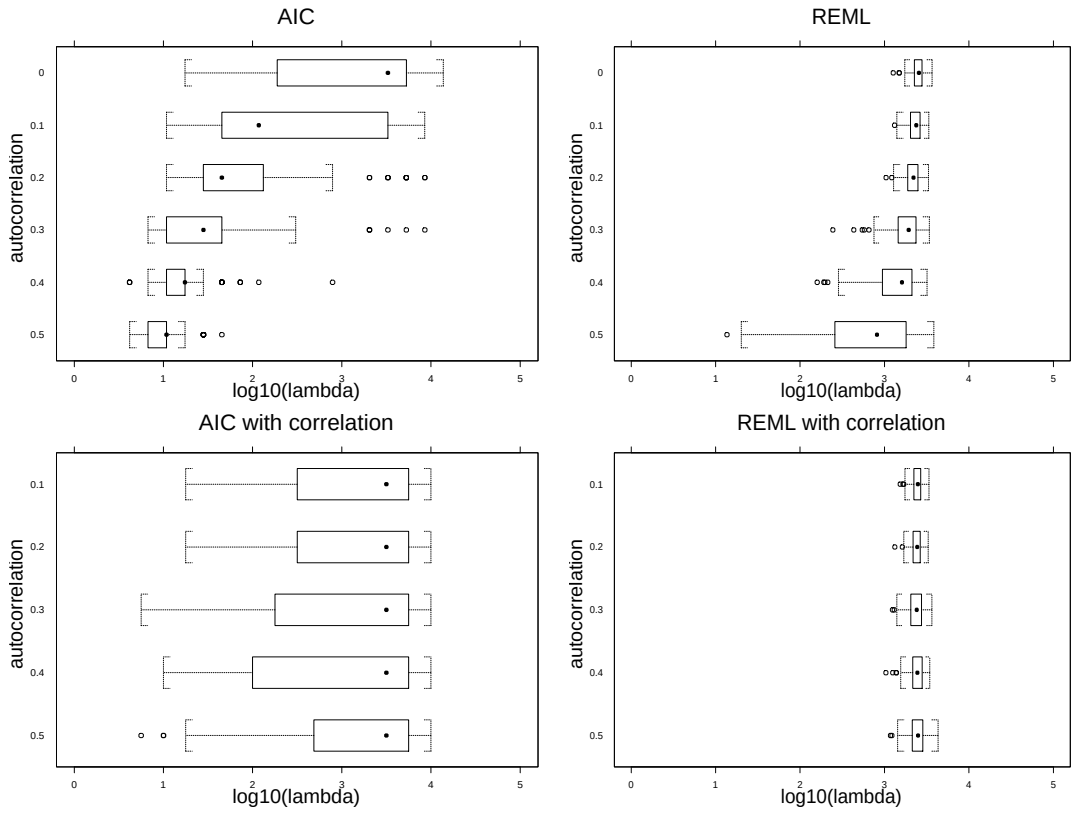

Figure 3.2: Boxplots for log-transformed smoothing parameter choice in 100 simulations. Upper row shows $\log \hat{\lambda}_{AIC}$ and $\log \hat{\lambda}_{REML}$ without accounting for correlation. The two bottom plots correspond to AIC and REML smoothing parameters when the true correlation structure is explicitly taken into account.

upper row shows the selected smoothing parameter when the correlation is ignored, that is, we use working independence. It appears that even for small correlations, the AIC tends to pick a small smoothing parameter, which in turn leads to overfitting. In general the dependence of $\hat{\lambda}_{AIC}$ on the correlation is strong. In contrast, the REML based λ behaves clearly more inertially and $\hat{\lambda}_{REML}$ picks a reasonable bandwidth even for (weakly) misspecified correlation. The deficit of the AIC choice is corrected if the true, but unknown, correlation is taken into account; that is, if we use the true (but

apparently unknown) correlation matrix R in the AIC choice (3.4) as well as for the REML criterion (3.6). The so selected smoothing parameters are plotted in the bottom row plots in Figure 3.2. Note that this second approach is only a theoretical exercise and not available in practice, since the true correlation structure is unknown. Therefore, the two bottom plots can serve as reference only. Even though our focus on the behaviour of the smoothing parameters λ_{AIC} and λ_{REML}, it is also practically of great interest to see the effect on the Mean Squared Error of the resulting estimates. This is visualized for the above simulations in Figure 3.3, where we show the term $\sum_{i=1}^{n}(\hat{m}(x_i)-m(x_i))^2/n$ for the different simulation scenarios. The organisation of the plot is the same as in Figure

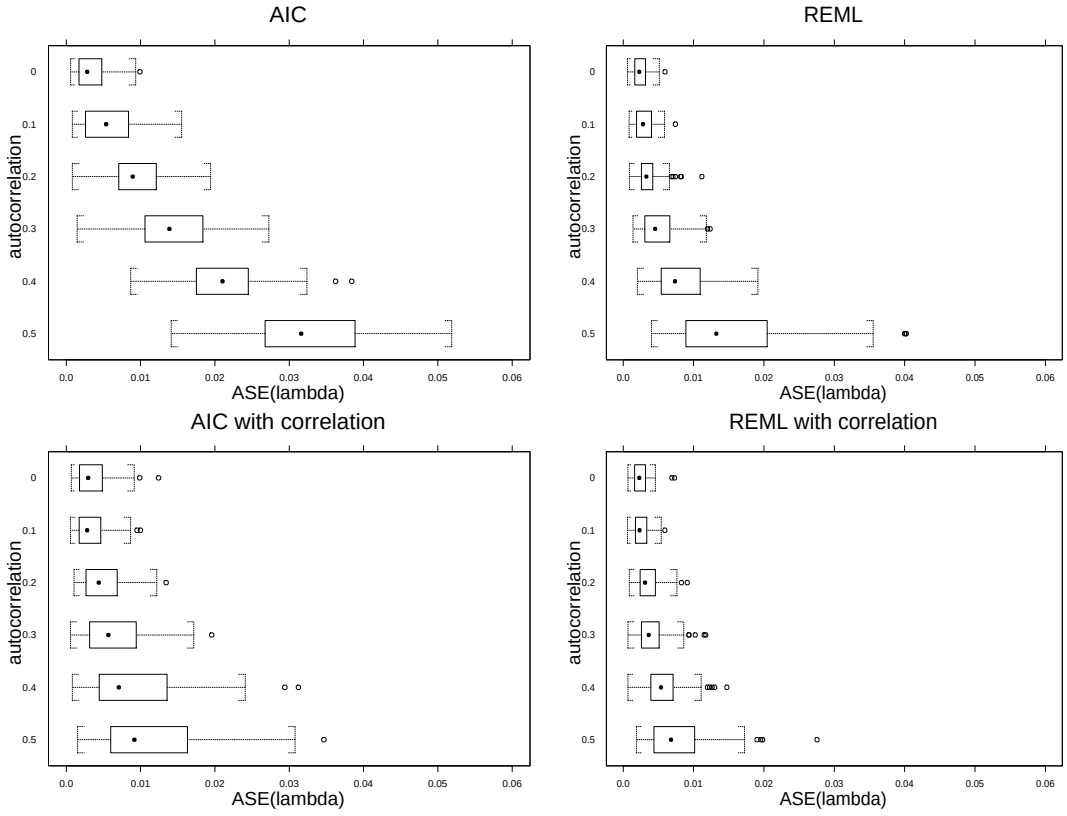

Figure 3.3: Boxplots for average squared error in 100 simulations. Upper row shows ASE($\hat{\lambda}_{AIC}$) and ASE($\hat{\lambda}_{REML}$) without accounting for correlation. The two bottom plots correspond to ASE of fits with AIC and REML smoothing parameters correspondingly, when the true correlation structure is explicitly taken into account.

3.2 and so is the resulting interpretation. Clearly, even small omitted correlations among the residuals have a dramatic effect on the Mean Squared Error of the AIC based fit,

while the REML based choice has a more stable behaviour.
We ran a number of other simulations with (i) different numbers of knots, (ii) different functional forms, (iii) different residual variability (i.e signal to noise ratio) and (iv) different basis functions (e.g. B-splines). The findings were the same as those reported here and these factors did not change the general behaviour. The superiority of the REML approach was always clearly seen.
The behaviour of the REML estimate transfers to more complex correlation scenarios. To demonstrate this we simulated data from an AR(2) process with first and second order autocorrelation 0.4 and 0.3, respectively. For fitting we employed a (misspecified) AR(1) correlation structure with first order autocorrelation estimated from the data. Note that this is easily accommodated in the linear mixed model framework and implemented, for instance, in the Splus or R `lme(.)` function (see Section 3.6 or Pinheiro & Bates, 2002). The resulting fit is shown in Figure 3.4, top row plots. The data clearly exhibit a correlation structure, if however this is not correctly specified, the AIC smoothing parameter choice suffers from overfitting. This is in contrast to the REML selected λ which works fine even for misspecified correlation. We rerun the simulation for different values of the second order autocorrelation, ranging from 0 (which equals AR(1)) to 0.5. The two plots in bottom row of Figure 3.4 show the resulting smoothing parameter estimates and Mean Squared Errors, respectively, if the data are in fact fitted with a misspecified AR(1) structure. The weak dependence of the REML estimate on the correlation structure is again visible and confirms our theoretical findings.

3.3 Examples and Applications

To illustrate the applicability of the described property of the REML estimator for the smoothing parameter selection we first consider data from Box & Jenkins (1970). 197 measurements of the "uncontrolled" concentration in a continuous chemical process are sampled at intervals of two hours and shown in Figure 3.5. The true correlation structure of the data is unknown. Diggle & Hutchinson (1989) made use of the data to demonstrate their smoothing parameter selection criterion which incorporates an AR(1) correlation structure. They suggest to estimate the smoothing and correlation parameter simultaneously, and receive an estimated first order correlation of 0.368. For the smoothing parameter choice a modified cross validation criterion that accommodates autocorrelated errors was applied. The resulting mean estimate is shown in the left upper plot of Figure 3.5 (see bold line). The dashed line is the fit assuming independence of the residuals, which is clearly not satisfactory. The right upper plot presents the fits

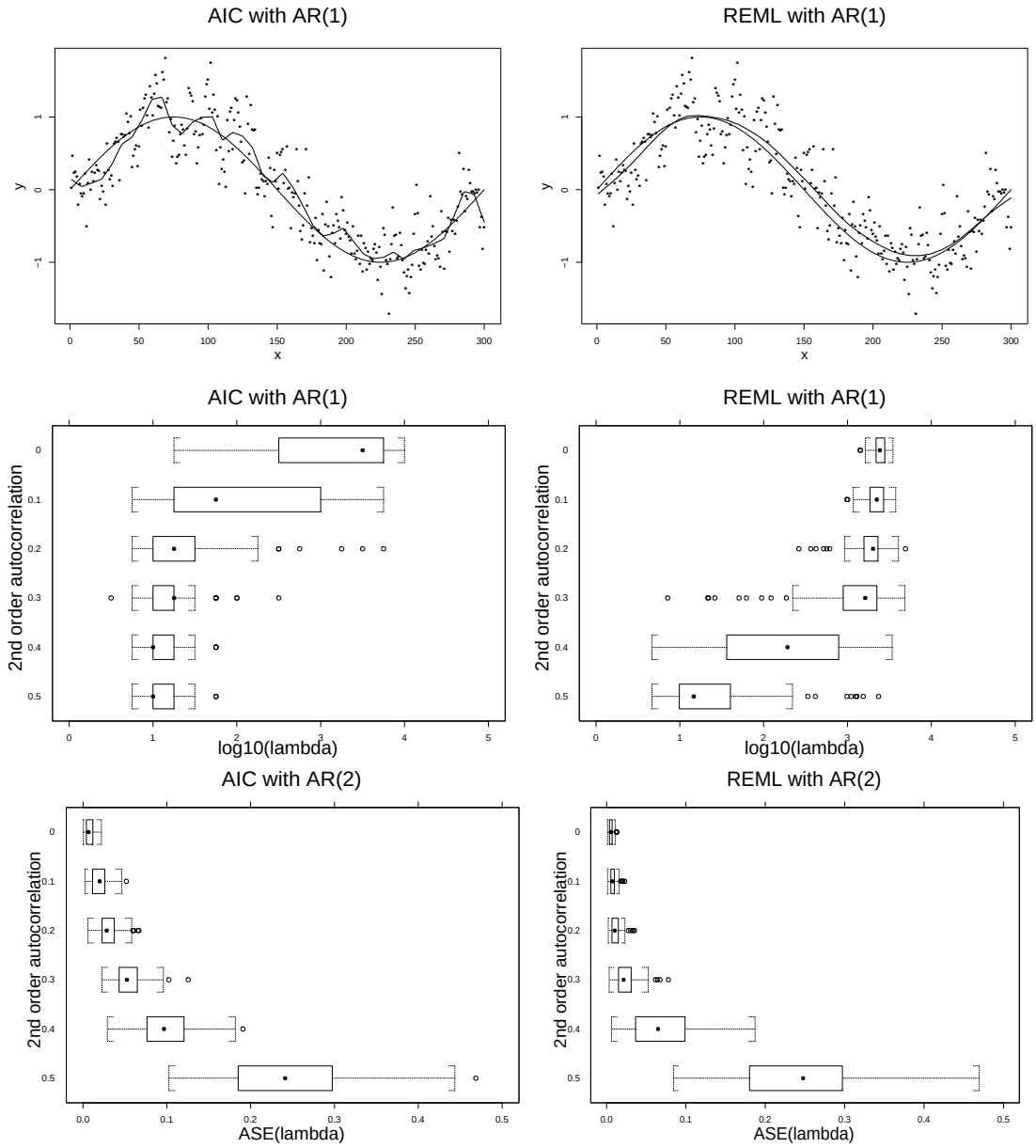

Figure 3.4: Estimated curves (bold) with AIC and REML based smoothing parameter choice (top row) with boxplots for log-transformed smoothing parameter choice in 100 simulations (middle row) and boxplots for average squared error in 100 simulations (bottom row).

with the proposed REML smoothing parameter choice, performed in the same way with (see bold line) and without (see dashed line) accounting for correlation. Both estimates are nearly indistinguishable, although the variance structure is wrongly specified in the

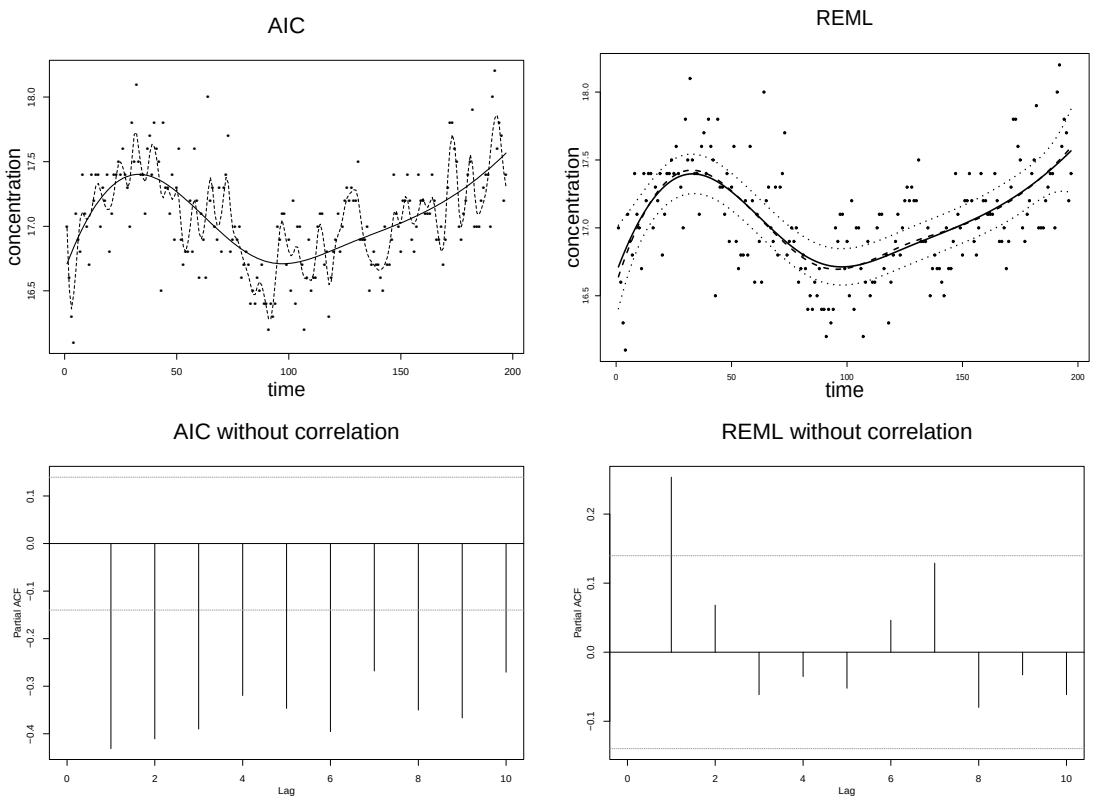

Figure 3.5: Estimated curves with (bold) and without (dashed) accounting for correlation using AIC (left) and REML based (right) choice of smoothing parameter (top row) with partial autocorrelation functions for the AIC (left) and REML based (right) fits without accounting for correlation (bottom row). Dotted lines show confidence bands for the REML fit which accounts for correlation.

second case. Thus, the REML based mean estimate with moderately misspecified correlation helps here to make a conclusion about unknown correlation structure of the data. This is visible from the bottom plots of Figure 3.5 where the estimated partial autocorrelation function based on the fitted residuals is shown. The REML estimate assuming independence provides a reasonable estimate for the autocorrelation structure (bottom right hand side plot) which can be used to modify the REML estimate by incorporating the estimated correlation structure (AR(1) in our example) in the smoothing parameter selection. The residual autocorrelation function for the latter fit looks the same as for the REML estimate ignoring correlation. Such behaviour does not exist for the MSE based choice, as can be seen from the bottom plot on the left hand side, which show the residual auto correlation function using an AIC choice without taking correlation into

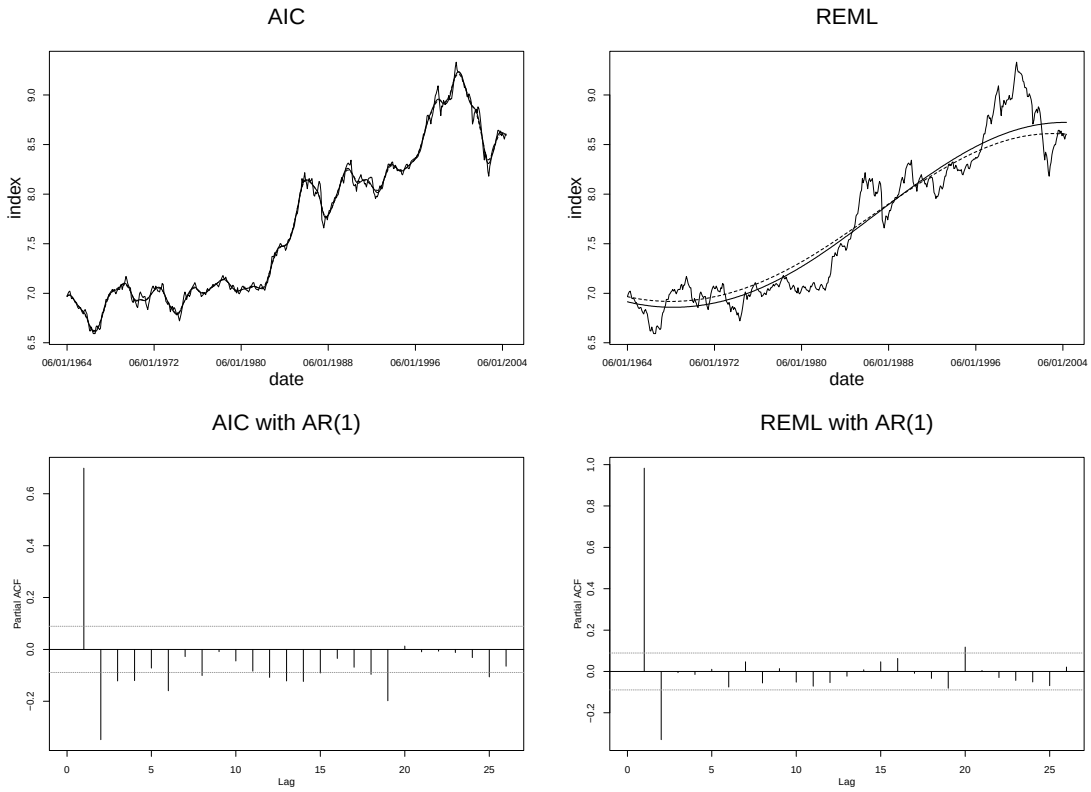

Figure 3.6: Estimated curves with an AR(2) (bold) and AR(1) (dashed) correlation structure using AIC (left) and REML based (right) choice of smoothing parameter (top row) with partial autocorrelation functions for the AIC (left) and REML based (right) fits accounting for an AR(1) correlation structure (bottom row).

account.

Our second example considers monthly averaged data of the German stock price index (CDAX) obtained from OECD *Main Economics Indicators*. We analyse 485 log observations from the period June 1964 to June 2004 as shown in Figure 3.6. We first fit the data with REML (upper right plot), assuming an AR(1) correlation structure (dashed). The plot of partial autocorrelation functions (bottom right plot) provides however evidence of the AR(2) structure of the residuals. We refit the data with a REML based smoothing parameter but now assuming an AR(2) structure. The resulting fit is shown in Figure 3.6 (top right plot) as solid line and the resulting autocorrelation function exhibits no changes. In contrast to the REML based estimate, the AIC criterion is not able to discover the mean structure at all, even if the covariance matrix of an AR(2)

process is explicitly taken into account in the estimation. This phenomenon can be explained by the fact that our time series is nearly non-stationary (with very large first order correlation), which however does not influence the REML estimate. Similar deficit of the MSE-based approach was mentioned e.g. in Diggle & Hutchinson (1989).
A further application is discussed in Chapter 4 where a two dimensional fit of the term structure of interest rates with non-standard correlation structure was performed. Again, it was the REML based smoothing parameter which made the fit possible.

3.4 Extensions

3.4.1 Additive Models

The result can be easily extended to other models, for instance additive models of the type

$$y_i \sim N(\beta_0 + m_1(x_{i1}) + ... + m_d(x_{id}), \sigma_\varepsilon^2), \quad i = 1, ..., n.$$

As in the univariate case we can represent each function as $m_l(x_i) = x_l\beta_l + Z_l u_l$, $l = 1, ..., d$ and estimate parameters from the penalized least squares

$$(y - X\beta - Zu)^T(y - X\beta - Zu)/\sigma_\varepsilon^2 + \lambda_1 u_1^T u_1 + ... + \lambda_d u_d^T u_d$$

or from the linear mixed model

$$y|u \sim N(X\beta + Zu, \sigma_\varepsilon^2 I_n), \quad u \sim N(0, \text{blockdiag}(\sigma_{u_1}^2 I_{K_1}, ..., \sigma_{u_d}^2 I_{K_d})),$$

with $\beta = (\beta_0, ..., \beta_d)$, $u = (u_1^T, ..., u_d^T)^T$, $X = [1, x_{i1}, ..., x_{id}]_{1 \le i \le n}$ and $Z = [Z_1, ..., Z_d]$, where Z_l is a basis matrix of dimension $n \times K_l$. Both models result in the estimate $\hat{y} = C(C^T C + \bar{D})^{-1} C^T y$ with $C = (X, Z)$ and $\bar{D} = \text{blockdiag}[0_{p\times p}, \lambda_1 I_{K_1}, ..., \lambda_d I_{K_d}]$, where p is the dimension of X. The expressions for $\hat{\lambda}_l$ in both frameworks can be easily derived similar to (3.7) and (3.8) and result in

$$\hat{\lambda}_{l,REML} = \hat{\sigma}^2_{\varepsilon,MM} \frac{\text{tr}\,(S_\lambda G_l) - p}{\hat{\theta}^T\, \bar{D} G_l/\lambda_l\, \hat{\theta}}$$

and

$$\hat{\lambda}_{l,AIC} = \hat{\sigma}^2_\varepsilon \frac{\text{tr}[(S_\lambda - S_\lambda S_\lambda) G_l]}{\hat{\theta}^T \bar{D} G_l/\lambda_l (I - \tilde{S}_\lambda)\hat{\theta}},$$

with $G_l = \text{blockdiag}[0_{(p+K_1+\ldots+K_{l-1})}, I_{K_l}, 0_{(K_{l+1}+\ldots+K_d)}]$. Now the results of the Section 2.2 are directly applicable. We ran a number of simulations to check the performance of the routine. Whenever correlation is associated with one of the covariates, the findings of the previous section are reproduced for the additive case.

3.4.2 Non-normal Response

The advantages of REML estimates of smoothing parameter remain valid in a generalized setting. As illustration we consider Westgren's gold series. Figure 3.7 shows the data from experiment described in Westgren (1916), as given in Guttorp (1991). Total 1 598 observations are the numbers of gold particles observed in the same volume of a solution every 1.39 seconds. Particles may move in and out of the volume. These data were considered by the numbers of authors - Chandrasekhar (1954), Heyde & Seneta (1971), Guttorp (1991), Grunwald & Hyndman (1998) - as an example of a Poisson branching process or of a Poisson AR(1) model. Particularly Grunwald & Hyndman (1998) considered smoothing parameter choice with AIC and BIC criteria for smoothing of this correlated Poisson series. We used REML approach to fit these data with $K = 120$ knots, taking an AR(1) correlation structure into account. Left plot of the Figure 3.7 shows the resulting REML fit (bold), while the right plot represents the estimated autocorrelation function for this estimate. The letter has evidence of an AR(2) correlation.

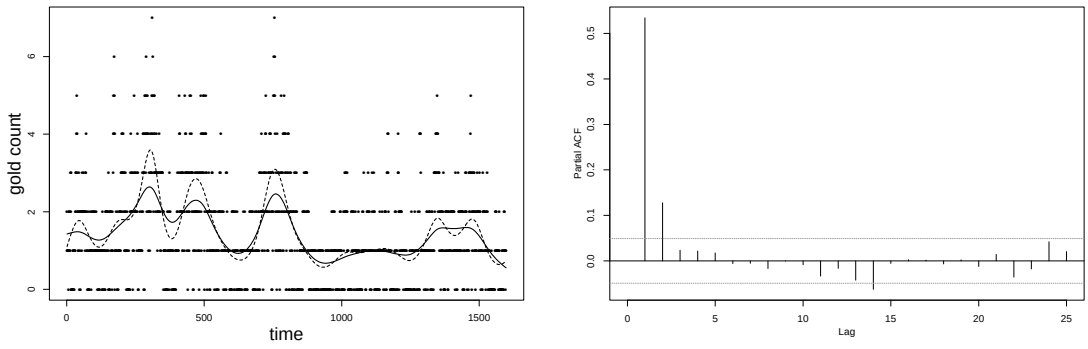

Figure 3.7: Right: Estimated curves with AIC (dashed) and REML (bold) based smoothing parameter choice. Left: Partial autocorrelation function corresponding to the REML estimate.

This is consistent with the findings of Grunwald & Hyndman (1998), which estimated the first order correlation to be 0.583 and noted a substantial additional variation. The

corresponding fit is shown in Figure 3.7 left hand side as dashed line. Obviously, AIC estimate is affected by the misspecified correlation structure, while REML fit remains stable. In fact, the estimate which takes an AR(2) correlation structure is indistinguishable to that of REML fit with an AR(1) structure.

3.5 Smoothing Parameter Estimation

3.5.1 REML estimate

Let us consider the mixed model

$$y|u \sim N(X\beta + Zu, \sigma_\varepsilon^2 I_n), \quad u \sim N(0, \sigma_u^2 I_K).$$

The restricted log-likelihood for this model is given by

$$-2l_R = (n-p)\log(\sigma^2_{\varepsilon,MM}) + \log|V_\lambda| + \log|X^T V_\lambda^{-1} X| + (y - X\beta)^T V_\lambda^{-1}(y - X\beta)/\sigma^2_{\varepsilon,MM},$$

yielding the following estimates for the fixed effects and $\sigma^2_{\varepsilon,MM}$

$$\begin{aligned} \hat{\beta} &= (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1} y, \\ \hat{\sigma}^2_{\varepsilon,MM} &= \frac{(y - X\hat{\beta})^T V_\lambda^{-1}(y - X\hat{\beta})}{(n-p)} = \frac{y^T(I_n - S_\lambda)y}{(n-p)}. \end{aligned} \tag{3.14}$$

With this the profile restricted log-likelihood for the smoothing parameter λ results to (3.6). Now $\hat{\lambda}_{REML}$ can be obtained by minimization of (3.6). Using the relationships

$$\frac{\partial S_\lambda}{\partial \lambda} = -\frac{1}{\lambda}(S_\lambda - S_\lambda S_\lambda), \quad \hat{\theta}^T D \hat{\theta} = \frac{1}{\lambda} y^T (S_\lambda - S_\lambda S_\lambda) y,$$

$$\frac{\partial(\log|V_\lambda|)}{\partial \lambda} = -\frac{1}{\lambda}\mathrm{tr}[Z(Z^T Z + \lambda D)^{-1} Z^T]$$

and

$$\frac{\partial(\log|X^T V_\lambda^{-1} X|)}{\partial \lambda} = \frac{1}{\lambda}\mathrm{tr}[Z(Z^T Z + \lambda D)^{-1} Z^T X (X^T V_\lambda^{-1} X)^{-1} X^T V_\lambda^{-1}],$$

we get the REML score equation

$$-2\frac{\partial REML(\lambda)}{\partial \lambda} = \frac{\hat{\theta}^T D \hat{\theta}}{\hat{\sigma}^2_{\varepsilon,MM}} - \frac{1}{\lambda}(\mathrm{tr}(S_\lambda) - p) = 0. \tag{3.15}$$

There are a number of approaches for solving (3.15) - EM, fixed point algorithm, Fisher scoring or Newton-Raphson algorithm (see e.g. Searle, Casella, & McCulloch, 1992 or Demidenko, 2004).

Fixed point estimation

Equation (3.15) invites for the solution with the fixed point algorithm using the representation

$$\lambda = \hat{\sigma}^2_{\varepsilon,MM} \frac{\mathrm{tr}(S_\lambda) - p}{\hat{\theta}^T D \hat{\theta}}. \tag{3.16}$$

The estimation procedure is now the following

1. Define some initial value λ;
2. Estimate $\hat{\beta}$ and $\hat{\sigma}^2_{\varepsilon,MM}$ from (3.14);
3. Update $\hat{\lambda}$ from (3.16);
4. Iterate between 2 and 3 until convergence.

This approach is in fact a version of EM algorithm as pointed out in Demidenko (2004) and is motivated and justified in the mixed model literature (see e.g. Searle, Casella, & McCulloch, 1992). Thus, we can use (3.16) not only as a definition of λ_{REML} with the true parameter inserted in both sides, but also for the estimation.

Fisher scoring

For technical reasons we solve (3.15) with the Fisher scoring algorithm (see Harville, 1977) after multiplication with $\lambda\hat{\sigma}^2_{\varepsilon,MM}$. That is, the iteration procedure $\lambda^{(r+1)} = \lambda^{(r)} - \tilde{s}_R / E[\partial \tilde{s}_R / \partial \lambda]$ is defined for $\tilde{s}_R(\lambda) := -2\lambda\, \hat{\sigma}^2_{\varepsilon,MM}\, \partial REML / \partial \lambda$. It remains to determine the first derivative of $\tilde{s}_R(\lambda)$, which can be written as

$$\frac{\partial \tilde{s}_R(\lambda)}{\partial \lambda} = \hat{\theta}^T D \hat{\theta} - 2\hat{\theta} D (I - \tilde{S}_\lambda)\hat{\theta} + \hat{\sigma}^2_{\varepsilon,MM} \frac{\mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}{\lambda} - \frac{\hat{\theta}^T D \hat{\theta}}{n-p}(\mathrm{tr}(S_\lambda) - p) \tag{3.17}$$

so that its expectation equals

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \tilde{s}_R(\lambda)}{\partial \lambda}\right] &= \frac{\sigma_\varepsilon^2(\mathrm{tr}(S_\lambda)-p)}{\lambda} - \frac{2\sigma_\varepsilon^2 \mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}{\lambda} \\ &+ \frac{\sigma_\varepsilon^2 \mathrm{tr}(S_\lambda - S_\lambda S_\lambda)}{\lambda} - \frac{\sigma_\varepsilon^2(\mathrm{tr}(S_\lambda)-p)^2}{\lambda(n-p)} \\ &= \frac{\sigma_\varepsilon^2}{\lambda}\left[\mathrm{tr}(S_\lambda S_\lambda) - p - \frac{(\mathrm{tr}(S_\lambda)-p)^2}{n-p}\right]. \end{aligned}$$

This type procedure is implemented in any standard mixed model software, which handles more complex models with general covariance structure of y and u as well.

Newton-Raphson algorithm

Under some additional assumptions we can interpret (3.16) also as a Newton procedure. Namely, we assume for the proof that $K << n$, that is the dimension of the spline is small compared to an increasing sample size. Moreover, for technical reasons we assume λ to be bounded and bounded away from zero. This means that $0 < \sigma_u^2 < \infty$. For simplicity we keep K bounded, that is θ is of finite dimension. For $\tilde{s}_R$ as score equation we can see that all components in its derivative (3.17) are of negligible order except of $\hat{\theta}^T D\hat{\theta}$. This follows since $H_\lambda^{-1} = (C^T C + \lambda D)^{-1}$ has order $O(n^{-1})$ for λ being bounded. This can be motivated by recognising that according to (2.19) in the mixed model framework $\mathrm{Cov}(\hat{\theta} - \theta) = \sigma_\varepsilon^2 H_\lambda^{-1} = O(n^{-1})$, yielding $\mathrm{tr}(S_\lambda - S_\lambda S_\lambda) = \mathrm{tr}(H_\lambda^{-1} D H_\lambda^{-1} C^T C) = O(n^{-1})$. Moreover, since $\hat{\theta}$ is assumed as finite dimensional we get $\hat{\theta}^T D(I - \tilde{S}_\lambda)\hat{\theta} = \lambda \hat{\theta}^T D H_\lambda^{-1} D\hat{\theta} = O_p(n^{-1})$. This simplifies the rth step of the Newton procedure to

$$\lambda^{(r+1)} = \lambda^{(r)} - \left(\frac{\partial \tilde{s}_R(\lambda)}{\partial \lambda}\right)^{-1} \tilde{s}_R(\lambda) = \hat{\sigma}^2_{\varepsilon,MM} \frac{\mathrm{tr}(S_\lambda) - p}{\hat{\theta}^T D\hat{\theta}} + O_p(n^{-1}),$$

with $\lambda = \lambda^{(r)}$ on the right hand side.

3.5.2 AIC estimate

Similar to the REML-estimate, we proceed for the AIC choice of smoothing parameter. We consider the model

$$y \sim N(X\beta + Zu, \sigma_\varepsilon^2 I_n).$$

In this framework parameter estimates are given with

$$\begin{aligned}\hat{\theta} &= (C^TC + \lambda_{AIC}D)^{-1}C^Ty,\\ \hat{\sigma}^2_\varepsilon &= \frac{(y - C\hat{\theta})^T(y - C\hat{\theta})}{n} = \frac{y^T(I_n - S_\lambda)^T(I_n - S_\lambda)y}{n}.\end{aligned}$$

The smoothing parameter λ_{AIC} is chosen by minimizing (3.4). Using the results from above, we obtain

$$\frac{\partial AIC(\lambda)}{\partial\lambda} = \frac{2\hat{\theta}^TD(I - \tilde{S}_\lambda)\hat{\theta}}{\hat{\sigma}^2_\varepsilon} - \frac{2}{\lambda}\mathrm{tr}(S_\lambda - S_\lambda S_\lambda) = 0. \tag{3.18}$$

Solving (3.18) for λ yields estimate (3.8). We can now, at least in principle, also use (3.8) iteratively, which is justified as Newton procedure. The arguments are comparable to the proof above and, therefore, not explicitly listed here. We stress however that even though (3.18) or (3.8) serves the purpose of theoretical investigation, the practical use of (3.8) as iterative formula is not recommended since Newton method in general has poor global convergence properties. Instead, grid search should be used. Clearly, Fisher scoring or EM algorithm are not applicable here.

3.5.3 Relation of AIC and REML based smoothing parameters

For comparison of $\lambda_{REML} = \sigma^2_\varepsilon/\sigma^2_u$ and λ_{AIC} we stress that they are defined in two different models. While λ_{REML} exists in the linear mixed model (3.5), λ_{AIC} is defined in the smoothing model (3.1). In particular, in (3.1) we assume that $\mu(x) = X\beta + Zu$ so that λ_{AIC} depends on the unknown coefficient u (and β), subsequently denoted by $\lambda_{AIC}(u)$. To relate λ_{REML} and $\lambda_{AIC}(u)$ we consider the mixed model (3.5) and show that

$$\lambda_{REML} = E_u[\lambda_{AIC}(u)]. \tag{3.19}$$

Note, that the Mean Squared Error in the smoothing model (3.1) equals

$$MSE(\lambda) = \sigma^2_\varepsilon\mathrm{tr}(S_\lambda S_\lambda) + \lambda^2_{AIC}\theta^TDH^{-1}_\lambda C^TCH^{-1}_\lambda D\theta. \tag{3.20}$$

Differentiating (3.20) defines the optimal Mean Squared Error smoothing parameter λ_{AIC} as a solution of

$$0 = \frac{\partial MSE(\lambda)}{\partial\lambda} = -\frac{2\sigma^2_\varepsilon}{\lambda_{AIC}}\mathrm{tr}(S_\lambda S_\lambda - S_\lambda S_\lambda S_\lambda) + 2\lambda_{AIC}\theta^TDH^{-1}_\lambda C^TS_\lambda CH^{-1}_\lambda D\theta.$$

Considering now u (and hence $\theta = (\beta^T, u^T)^T$) as random, that is assuming the mixed model (3.5) we find by taking expectation

$$0 = E_u\left(\frac{\partial MSE(\lambda)}{\partial \lambda}\right) = 2(\sigma_u^2 - \frac{\sigma_\varepsilon^2}{\lambda_{AIC}})\mathrm{tr}(S_\lambda S_\lambda - S_\lambda S_\lambda S_\lambda),$$

which is solved for $\lambda_{AIC} = \lambda_{REML} = \sigma_\varepsilon^2/\sigma_u^2$ and proves (3.19).

3.6 Computational Issues

To demonstrate the simplicity and numerical feasibility of the REML estimate we present the implementation in R (www.r-project.org, R Development Core Team, 2005) using the example of the German stock price index (CDAX) data. We take advantage of the `lme` function of package *nlme* (see also Pinheiro & Bates, 2002 or Ngo & Wand, 2004 for more details in smoothing using `lme(.)`). The dataset has 485 observation for "day" and "cdax". First we define our model matrices for $k = 80$ knots. We use squared truncated lines here.

```
> st <- 484/(k+1)
> kn <- seq(1+st,485-st,by=st) #set equidistant knots
> Z <- outer(1:485,kn,"-")
> Z <- (Z*(Z>0))^2
```

With this spline matrix we can call the `lme` function, using

```
> library(nlme)
> all <- rep(1,485)
> cdax.fit <- fitted(lme(cdax~day+day^2,random=list(all=pdIdent(~Z-1)),
              correlation=corAR1()))
> plot(day,cdax.fit)
```

Here, we allow for an AR(1) process in the residuals. The plot of the partial autocorrelation function in Figure 3.6 suggests refitting the model with an AR(2) correlation structure as follows

```
> cdax.fit1 <- fitted(lme(cdax~day+day^2,random=list(all=pdIdent(~Z-1)),
               correlation=corARMA(p=2)))
```

The confidence bands can be obtained according to (2.19) from the variance estimate $\mathrm{Var}[\hat{\mu}_\lambda(x)] = \hat{\sigma}_\varepsilon^2\mathrm{tr}[C(C^T R^{-1} C + \lambda D)^{-1} C^T]$, with R as estimated correlation matrix. In

Figure 3.5 the confidence bands are shown exemplary for the REML fit which accounts for correlation.
For generalized responses the above code looks the same, with `lme` being replaced by the `glmmPQL` function from library *MASS*, see also Section 2.5.3.
An efficient algorithm for the penalized smoothing using AIC/GCV criterion is given in Section 2.5.2. The R implementation of this algorithm can be found in Appendix B of Ruppert, Wand & Carroll (2003). We adjusted this implementation for the smoothing with generalized response and demonstrate it here on the example of Westgren's gold series with `time` as covariate and `gold` as Poisson distributed response variable.

```
# Set up model matrices
x <- time
y <- gold
n <- length(x)
k <- 120
st <- (max(x)-min(x))/(k+1)
kn <- seq(min(x)+st,max(x)-st,by=st)
Z <- (outer(x,kn,"-"))
Z <- (Z*(Z>0))^2
C <- cbind(rep(1,n),x,x^2,Z)
D <- diag(c(rep(0,3),rep(1,k)))

# Set up logarithmic grid of smoothing parameter lambda values

lambda.low <- 10
lambda.upp <- 10^6
num.lambda <- 150
lambda.vec <- 10^(seq(log10(lambda.low),log10(lambda.upp),
                length=num.lambda))

# Get initial values from simple regression spline model

theta <- coef(glm(y~C-1,family=poisson))
eta <- as.vector(C%*%theta)
mu <- exp(eta)
weights <- c(mu)

# Iterate for the mean estimate

for (i in 1:20)
```

```
{
  u <- eta+(y-mu)/weights

  # Take into account current iterative weights

  u <- sqrt(weights)*u
  CV <- sqrt(weights)*C
  CVC <- t(CV)%*%CV

  # Carry out Demmler-Reinsh algorithm

  K <- chol(CVC)
  svd.mat <- t(solve(t(K), t(solve(t(K), D))))
  svd.out <- eigen(svd.mat, symmetric = T)
  s.vec <- svd.out$values
  U <- svd.out$vectors
  A.mat <- CV %*% backsolve(K, U)
  b.vec <- as.vector(t(A.mat)%*% u)
  r.mat <- 1/(1 + outer(s.vec, lambda.vec))
  y.hats <- A.mat %*% (b.vec * r.mat)
  y.vec <- matrix(rep(u, num.lambda), n, num.lambda)
  RSS <- apply(((y.vec-y.hats)^2),2,sum)
  df.vec <- apply(r.mat, 2, sum)

  # Determine AIC ...

  AIC <- log(RSS)+2*df.vec/n

  # or GCV criterion  GCV <- RSS/(1-df.vec/n)^2

  # Find minimum ...

  ind.min <- order(AIC)[1]

  if(ind.min == 1)
    stop("make lambda.low smaller")
  if(ind.min == num.lambda)
    stop("make lambda.upp bigger")
  lambda.aic <- lambda.vec[ind.min]

  # and next approximation
```

```
    eta1 <- (1/sqrt(weights))*y.hats[,ind.min]
    epsilon.eta <- sum((eta-eta1)^2)/sum(eta^2)
    eta <- eta1
    mu <- exp(eta)
    weights <- c(mu)
    if(max(epsilon.eta)<=1e-05) break
    if(i==20) stop ("Iteration limit reached without convergence")
  }

    mu
```

The implementation can easily be adjusted for other distributions of exponential family. Known correlation structure can be taken into account by standardizing the working vector.

3.7 Discussion

We investigated the sensitivity of misspecified correlation for two data-driven smoothing parameter selectors for penalized spline smoothing - Akaike and REML. It has been shown that the AIC chosen smoothing parameter is (on average) more affected by the presence of correlated errors than the REML based smoothing parameter. Theoretical investigation based on a Taylor series and a simulation study illustrated that the REML chosen smoothing parameter possesses a kind of robustness against misspecification of the correlation structure, while AIC fails even for weak correlation. The findings were supplemented by real data examples.

4 Estimation of the Term Structure of Interest Rates

4.1 Motivation

Modelling the term structure of interest rates has become an active field of research in finance in the last years. Based on historical developments, a primary area of application is pricing and hedging of different contracts and options written on bonds. Numerous approaches have been proposed concerning the underlying time structure and the stochastic framework of term structure models. Furthermore, different perspectives on term structure modelling have stimulated the development of an enormous variety of models and methods used to study them.

Generally, term structure models can be divided into two main categories: *equilibrium* and *arbitrage-free* models. Note that both kinds of models are constructed under the assumption of no-arbitrage, therefore the term "arbitrage-free" may be a little misleading. Within the first category a state variable that determines the term structure is identified and both, the yield curve and the dynamic behaviour of interest rates are determined endogenously. Therefore, one has to estimate or choose parameter values to approximate the average yield curve as well as the short rate. Pioneering models of this kind are Vasicek (1977) and Cox, Ingersoll & Ross (1985). In the second category, the currently observed yield curve is used as an input to model the changes of the term structure over time. The basic model here is proposed in Ho & Lee (1986). Their approach differs from Vasicek in so far as it contains additional time dependent adjustment parameters to calibrate the initial yield curve with the goal to match the observed yield curve exactly (see Backus, Foresi & Zin, 1998).

Despite the widespread use and application of these "theoretical" models, the extraction or estimation of the complete term structure of interest rates from empirically observed bond prices is of less common use. In statistical terms this corresponds to an exploratory analysis of the term structure. In practice it is not possible to obtain the values of the term structure for all horizons since their number exceeds the number of available bonds.

To overcome this problem one may use smoothing as an interpolation technique. This is the task of a different stream in the term structure literature and emphasis of the following paper.

In recent work, Ioannides (2003) compares seven estimation methods for the term structure applied to UK data. The methods used in his paper can roughly be categorized as (i) parametric and (ii) nonparametric. For the first, a low dimensional basis is used for fitting the term structure to observed data. This approach traces back to McCulloch (1971) and is further explored and discussed for instance in Chambers, Carleton & Waldman (1984) or Nelson & Siegel (1987). The latter paper, in contrast to McCulloch, uses a parsimonious parametric function, with only a small number of unknown parameters, that is flexible enough to represent the shapes generally associated with yield curves (see also Steeley, 1990). In nonparametric estimation, a restrictive parametric term structure modelling is abandoned and replaced by unspecified, unknown functions. The idea is that the functional form should be estimated from the data and not pre-specified in advance. This approach was pursued in Fisher, Nychka & Zervos (1995) and is further employed in this paper. Since the term structure has implications both for the cross section and time series dimension of yields, we use a two-dimensional smoothing that leads to a more efficient estimation.

Nonparametric fitting in general has seen a considerable amount of research in the last two decades. Nonetheless, it has been just recently that nonparametric techniques have found their way to term structure modelling. Linton, Mammen, Nielsen & Tanggaard (2000) and Jeffrey, Linton & Nguyen (2001) concentrate on kernel smoothing while Jarrow, Ruppert & Yu (2004) employ penalized spline estimation (P-spline). Comparing the two fitting routines, P-spline smoothing features a considerably reduced numerical effort. This is an important issue, in particular if the number of observations is large, about 126 000 in our application. In P-spline smoothing the unknown term structure is replaced by a high dimensional basis (30-200 dimensional) which is then fitted in a penalized manner, that is spline coefficients are shrunk towards zero. This guarantees a smooth fit by prevailing all necessary structure in the function.

The term structure thereby depends on two components, the time left to maturity m and the calendar time t. We take this into account by denoting the term structure function as $f(t, m)$. To explore the term structure at a given time-point, one can fix t at some specific value t_0, say, and fit $f(t_0, m)$ as a function of m only. This is the approach used in the above cited papers. Fixing now the time left to maturity m to m_0, say, the development of $f(t, m_0)$ for a given m_0 is traditionally understood as a stochastic process. This is useful if the focus is on prediction of the yield based on data

(and history) available at the current data point. From an exploratory point of view one might however also be interested in describing or visualising the smooth trend in $f(t, m_0)$. This is what could be interpreted as long term development, which is visual from the raw data in a crude way only. In previous applied work this approach is mostly done for the short-rate, since it can be seen as an important state variable for the term structure (see e.g. Chan, Karolyi, Longstaff & Sanders, 1992).
Surprisingly, it has been only quite recently that papers in both streams of the term structure literature, i.e. theoretical modelling and smoothing techniques, try to model the complete panel of yield data simultaneously (see e.g. Brandt & Yaron, 2003 or Diebold & Li, 2003). In this paper we combine the two approaches by fitting $f(t, m)$ simultaneously as a function of both covariates, time t and time left to maturity m. This means we are using smoothing in two ways. First, as interpolation tool for showing $f(t, m)$ for a fixed time as function of m. Secondly, with smoothing we visualise long term trends in $f(t, m)$ taken as function of time for fixed m. This allows us to explore the term structure and its temporal variation simultaneously.
There are two challenges arising in this modelling exercise. First, one is faced with additional numerical effort, as the dataset has more than 126 000 points. Using the link of penalized splines to the linear mixed models and, thus, fitting our data with standard linear mixed models software makes it possible to overcome this problem. The second challenge occurs since bond prices are correlated over time which has to be taken into account. Ignoring correlation among observations typically leads to serious under-smoothing, that is overfitting, as discussed in previous Chapter. Research on smoothing correlated errors has nearly exclusively discussed univariate or spatial correlation. We here, however, observe correlation only along yield price development over time of single bonds. To handle this problem we offer a simple procedure, based on accounting for correlation of single yield strips. This means the smoothing parameter is chosen using one-dimensional estimates of correlation structure.

4.2 Data

Our investigation is based on daily ask quotations of US Treasury STRIPS (Separate Trading of Registered Interest and Principal of Securities). These are securities and synthetic zero-coupon bonds, which are constructed from coupon bearing Treasury bonds and issued by the US Federal Reserve Bank. The sample runs from July 1998 to July 2003 and contains 107 different US Treasury Strip coupon securities with maturities

from one month to 30 years, a total of 126 251 observations. The data are collected on

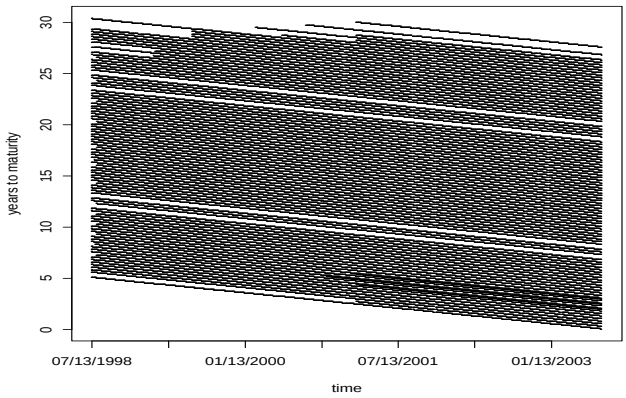

Figure 4.1: Design of independent covariates time t and maturity m

July 11, 2003 using the Reuters 3000 Xtra information service and include bond prices with maturity dates from August 2003 to May 2033. Figure 4.1 shows the observed time point and maturity pairs. Table 4.1 shows some specific properties of the data set. To provide a useful representation, the daily quotes are summarized in classes of different years to maturity. It should be clear that the average yield curve has an increasing, concave shape.

Maturity	Obs	Mean	St Dev	Autocorr
0.0986-0.25	36	1.0064	0.0365	0.0316
0.25-0.5	95	0.9866	0.1062	0.9202
0.5-1	376	1.1212	0.1419	0.9430
1-3	4255	2.5199	1.1757	0.9878
3-6	13882	4.4737	1.3132	0.9938
6-9	14301	5.2139	0.9423	0.9932
9-12	13457	5.4037	0.7329	0.9907
12-15	14655	5.6697	0.5945	0.9882
15-20	24614	5.8301	0.4677	0.9873
20-25	22776	5.8194	0.3920	0.9873
25-30.3616	17894	5.6992	0.3630	0.9863

Table 4.1: Properties of US Treasury STRIPS Yields

4.3 Spline Models for the Term Structure

4.3.1 Bivariate spline smoothing

We denote with $P_{t,m}$ the price of a zero bond at time point t with m years left to maturity and consider the continuously compounded yield obtained as

$$y_{t,m} = -\frac{\log(P_{t,m})}{m}.$$

We model

$$y_{t,m} = f(t,m) + \epsilon_{t,m}, \quad \epsilon_{t,m} \sim N(0, \sigma_\epsilon^2 R), \tag{4.1}$$

with $f(t,m)$ as an unknown smooth function of both time and years left to maturity. The structure of matrix R will be discussed in the next section and is assumed known for the moment.

Now we face the problem of the spline basis choice. There are three alternatives: tensor product of one dimensional either truncated polynomials or B-splines and low-rank radial basis as discussed in Section 2.4.2. Thereby, we have to take into account the non-standard structure of our data. Apart from the extremely large dimension (more than 126 000 points) we observe that the functional complexity over time is much more exposed than that over time left to maturity. Smoothing with the low-rank radial basis seems to be the least preferable, since it is controlled by a single smoothing parameter, implying same amount of smoothing in both directions. Moreover, we experimented with different basis dimensions and found that capturing the function complexity requires more radial basis functions than is numerically feasible. A more attractive approach appears to be the tensor product of either truncated polynomials or B-splines. By choosing more knots over time than over years left to maturity and having separate smoothing parameters for each dimension, we can obtain an adequate estimate with less numerical effort. Due to the correlation structure of our data, described in the next section, we need to estimate the model in the mixed model framework. Since mixed model representation for B-spline basis requires additional adjustment, which is undesirable for our very large dataset, we found B-spline basis less attractive. In contrast mixed model representation of the smoothing model based on truncated polynomials is straightforward. Thus, for estimation we choose K_t and K_m knots placed over time and time left to maturity and replace $f(t,m)$ with

$$f(t,m) = (X_t, Z_t) \otimes (X_m, Z_m)\theta.$$

Here $X_t = [1, t, t^2]$ and $X_m = [1, m, m^2]$ are low dimensional bases in time t and maturity m respectively, and $Z_t = [(t - k_1^t)_+^2, ..., (t - k_{K_t}^t)_+^2]$, $Z_m = [(m - k_1^m)_+^2, ..., (m - k_{K_m}^m)_+^2]$ are the high dimensional supplements. We rearrange the basis matrix to

$$C = [X_t \otimes X_m, X_t \otimes Z_m, Z_t \otimes X_m, Z_t \otimes Z_m],$$

with θ decomposing to (β, u_m, u_t, u_c). The different components in C and θ capture different aspects of the function. Coefficient β is the overall parametric fit, u_m models the dependence on maturity, while u_t mirrors the temporal variation. Finally, u_c captures the interactive influence of t and m. The disadvantage of tensor product matrices is that their dimension increases rapidly. For instance, if Z_t and Z_m are 30 dimensional $Z_t \otimes Z_m$ is 900 dimensional, which is at the limit of numerical applicability when it comes to matrix inversion. We, therefore, replace the last component in C by $Z_c = \tilde{Z}_t \otimes \tilde{Z}_m$, where $\tilde{Z}_t$ and $\tilde{Z}_m$ are of some lower dimension K_c.
Denoting now with Y the n-dimensional vector of observations $y_{t,m}$, we write the penalized least squares to be maximized as

$$-\frac{1}{2}(Y - C\theta)^T R^{-1}(Y - C\theta) - \frac{1}{2}\lambda_t u_t^T u_t - \frac{1}{2}\lambda_m u_m^T u_m - \frac{1}{2}\lambda_c u_c^T u_c.$$

Assumption $u_m \sim N(0, \sigma_m^2 I_{K_m})$, $u_t \sim N(0, \sigma_t^2 I_{K_t})$, $u_c \sim N(0, \sigma_c^2 I_{K_c})$ leads to the linear mixed model. Both approaches result in the estimate $\hat{Y} = C\hat{\theta}$ with $\hat{\theta} = (C^T R^{-1} C + \tilde{D})^{-1} C^T R^{-1} Y$, with $\tilde{D} = \text{blockdiad}[0_{9\times 9}, \lambda_s I_{K_s}, \lambda_t I_{K_t}, \lambda_c I_{K_c}]$. In the mixed models framework $\lambda_t = \sigma_\epsilon^2/\sigma_t^2$, $\lambda_m = \sigma_\epsilon^2/\sigma_m^2$, $\lambda_c = \sigma_\epsilon^2/\sigma_c^2$.

4.3.2 Spline Smoothing with correlated errors

The smoothing parameters $\lambda = (\lambda_t, \lambda_m, \lambda_c)$ steer the amount of penalization in each direction and therewith the smoothness of the fit. As extensively discussed in Chapter 3, smoothing parameter selection with any data driven method (cross validation, AIC or (RE)ML) fails in the case of correlated errors and typically leads to serious undersmoothing, that is overfitting of the data. However, (RE)ML estimate of the smoothing parameter has two advantages. First, the estimate resulting in mixed model framework is more robust to the correlation misspecification. Second, once the correlation structure is specified, estimation of regression and correlation parameters can be carried out simultaneously. Implementation of this procedure with `lme` function is shown in Section 3.6. We use these properties of the (RE)ML estimate in the following way. Let $Y \sim N(C\theta, \sigma_\varepsilon^2 R)$ with correlation matrix R assumed to be known. This yields $Y^* = R^{-1/2}Y$, as uncor-

related observations, and with $C^* = R^{-1/2}C$, one gets that $Y^* \sim N(C^*\theta, \sigma_\epsilon^2 I)$. Hence, knowing the correlation structure we can simplify the estimation to uncorrelated residuals. The idea is now as follows. We will develop a rough estimate for the correlation structure considering data along calendar time only. The estimate is then used to derive Y^*. Even though the estimated correlation might not equal the true correlation exactly, it has been shown in Chapter 3 that the REML estimate still provides reasonable variance estimates even if the correlation is moderately misspecified. The correlation structure of

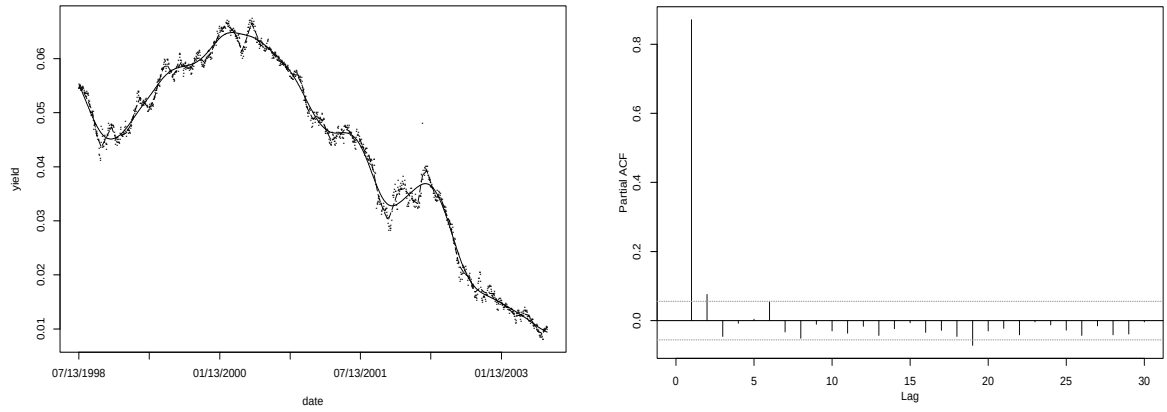

Figure 4.2: Yield development of a 6 years bond, smoothed with (solid line) and without (dashed line) accounting for correlation, with the corresponding partial autocorrelation function of residuals.

Y is however not standard. We observe correlation along calendar time t, but for given t it is not reasonable to assume that residuals $\varepsilon_{t,m}$ along maturity m are correlated. Figure 4.2 shows the yield development of a bond with 6 years left to maturity on the July 1998. The dashed line shows a penalized spline fit if autocorrelation is ignored, the solid line shows the fit if residuals are assumed to have an AR(1) correlation structure. For the latter we plot in Figure 4.2 the autocorrelation structure of the residuals. The AR(1) assumption seems plausible. Both fits are univariate smoothers and are calculated using the standard linear mixed models software, as described above without any modification. To account for correlation in the two dimensional fit (4.1), we now employ the idea of standardising the response variable in the following way. As visible in Figure 4.1 our data are observed in time series $Y_{t,m_c-(t-t_0)}$ where m_c is the maturity at t_0 in July 1998. All together there are 107 series, one of which is shown in Figure 4.2. Fitting these series in the same line as that in Figure 4.2 provides autocorrelation estimates ranging from

about 0.8 to 0.9. We therefore use an autocorrelation of 0.85 and let R_{m_c} denote the corresponding correlation matrix. Rearranging Y as $(Y^T_{t,m_1-(t-t_0)}, \ldots, Y^T_{t,m_{107}-(t-t_0)})$ with t taking all observed time points allows us to get $Y^* = R^{-1/2}Y$ with R as block diagonal matrix built from the 107 matrices R_{m_c}. Accordingly we get after rearrangement our matrix C^*.
It remains to fit a linear mixed model for independent errors $Y^* \sim N(C^*\theta, \sigma^2_{\epsilon_{t,m}} I_n)$, $u_m \sim N(0, \sigma^2_m I_{K_m})$, $u_t \sim N(0, \sigma^2_t I_{K_t})$ and $u_c \sim N(0, \sigma^2_c I_{K_c})$. Even though the dimension is large, due to independence and the lush, but finite, dimension of θ linear mixed models software can be applied to obtain the estimates $\hat{\theta} = (\hat{\beta}, \hat{u}_m, \hat{u}_t, \hat{u}_c)^T$ and the resulting fitted response $\hat{Y} = C\hat{\theta}$. This is a numerically handy version to cope with the complex correlation structure in the data. Based on this fit we also analyzed the residuals of the 107 series but did not find obvious violations from the model. There was a slight indication of heteroskedasticity which, however, was not too evident and for simplicity is ignored.

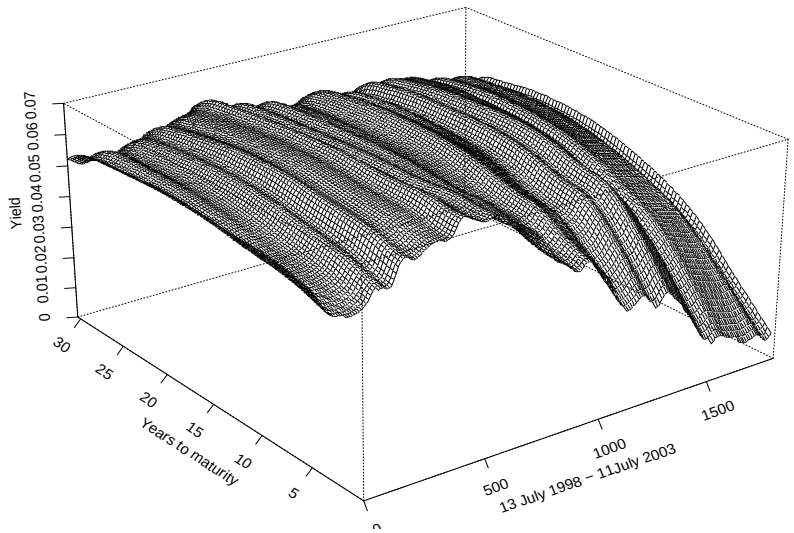

Figure 4.3: Bivariate fit of the term structure

4.3.3 Empirical Results

In Figure 4.3 we show the bivariate fit of the term structure. The fit is performed using matrix C constructed from truncated squared bases with equidistant knots. The dimensions are thereby chosen as $K_t = 40$, $K_m = 10$ and $K_c = 10 \times 10$. There are two

things apparent from the plot. First, the functional complexity over time is clearly more exposed than the functional complexity over maturity. Secondly, time and maturity have an interactive effect on the bond price, that is u_c can not be penalized to zero. To better understand the interactive effect we consider a number of plots by slicing the bivariate fit in Figure 4.3 time-wise and maturity-wise.
Figure 4.4 shows the estimated term structure for different time points. Beside the fit we have included prediction intervals based on $\pm 2\hat{\sigma}$. Prediction intervals are more useful than confidence intervals in this setting, since the latter are due to the large number of observations so small that they are visually indistinguishable from the fitted curve. From the plots there is a clear dynamic visible over time. The term structure during July

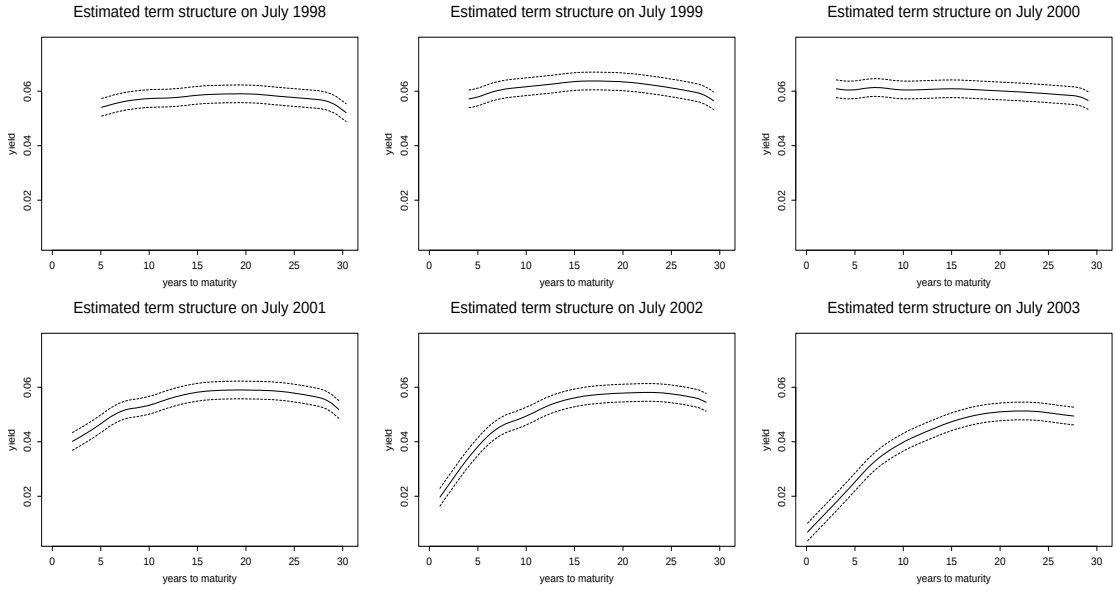

Figure 4.4: Estimated term structure with prediction intervals

1998 shows up as a typical "flat" yield curve, that is the case when the interest rates are about on the same level for different types of bonds. Two years later, in July 2000, after the stock market crash in USA (March 2000) the fitted term structure demonstrates a fully "flat" shape on a high level of about 0.06. Subsequently, the term structure becomes curved again, showing a "normal" shape with higher yield for a longer lending time and with more expressed differences between shorter and longer term yields in the most recent years. It is also obvious that long term bonds remain on an interest rate of about 0.06 while yield of short term bonds decreases with time.
Next, we consider the yield development for a given maturity over time. This is shown

in Figure 4.5. The figures visualize the stock market crash in 2000 in that yields for all maturities increase until about spring 2000 and decrease afterwards. In this respect we see that the yield for short maturity bonds is decreasing more rapidly than that for long term bonds.

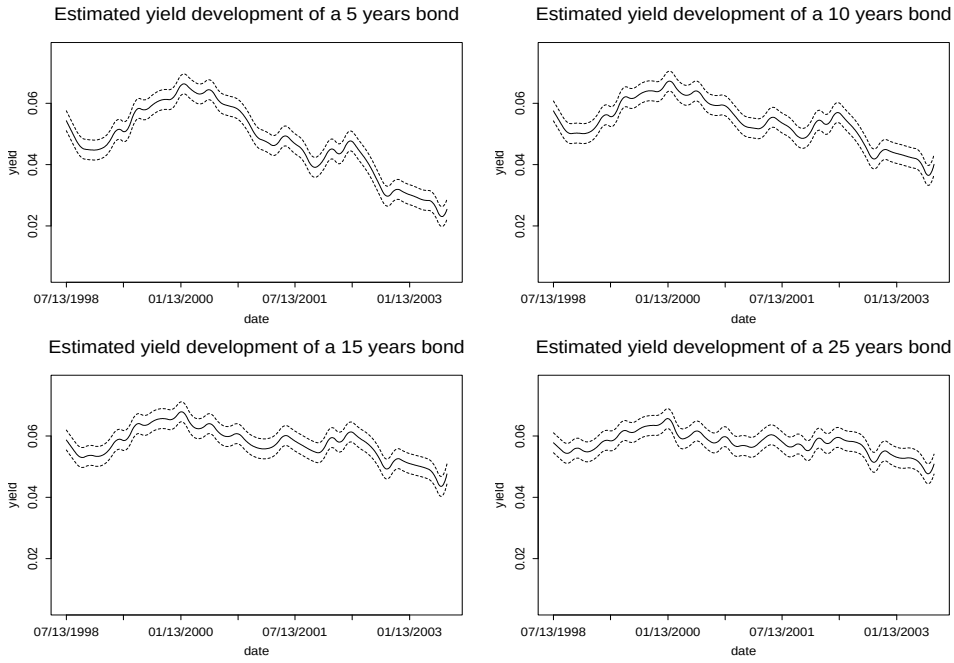

Figure 4.5: Estimated yield development with prediction intervals

4.4 Discussion

We pursued the exercise of fitting zero bonds yield as a bivariate function over time and maturity. We demonstrated the numerical efficiency of penalized spline smoothing as smoothing technique. The modelling exercise allowed to look in the term structure function and to study the dynamic effects. The approach exposed an interesting pattern in the term structure during the stock exchange crash. The empirical and exploratory approach can confirm theoretical investigations and stimulate new insights.

5 Fast Adaptive Penalized Smoothing

5.1 Motivation

Even though P-spline smoothing is easy and practical, the standard setting with a single penalization parameter fails if the function to be estimated is locally of varying complexity, that is if the function is changing rapidly in some regions while in other regions the function is very smooth. This is the general problem of spatially adaptive smoothing which has been treated by a number of authors. For kernel based methods Fan & Gijbels (1995) or Herrmann (1997) may serve as references. For spline smoothing Luo & Wahba (1997) suggest what they call hybrid adaptive splines. The idea is to replace the n dimensional spline basis, where n is the sample size, by a subset of the basis functions with the spline basis functions chosen adaptively. This idea has similarities to adaptive knot selection for regression splines as suggested in Friedman & Silverman (1989). An alternative approach is to allow the smoothing parameter to vary locally adaptive. Using a reproducing Hilbert space formulation has been suggested in Pintore, Speckman & Holmes (2005), where piecewise constant smoothing parameters are used. Similarly, making use of the P-spline idea, as also discussed in this paper, Ruppert & Carroll (2000) allow the penalty to act differently for each locally defined spline basis, where the smoothing parameters are then selected using a multivariate generalized cross validation. A similar approach is suggested in Wood, Jiang & Tanner (2002) working with mixtures of splines in a fully Bayesian framework. Lang & Brezger (2004) achieve a local adaptive P-spline by pursuing a Bayesian model of P-splines where spline coefficients trace from a heterogeneous random walk.
In this work we follow the idea of Baladandayuthapani, Mallick & Carroll (2005) and Crainiceanu, Ruppert & Carroll (2005) who achieve spatial adaptivity by imposing a functional structure on the smoothing parameters. However, these papers make use of full Bayesian framework and require the use of MCMC methods to obtain an estimate. We demonstrate how the MCMC techniques can be easily circumvented by simple Laplace approximation. Even though this is a step back in terms of the technical features we have nowadays, it is a step forward in terms of simplicity of numerics and therewith

allowing for fast calculation.

5.2 Smoothly varying local penalties for P-spline regression

5.2.1 Hierarchical penalty model

We introduce our approach with the simple scatterplot smoothing model

$$y_i \sim N(m(x_i), \sigma_\epsilon^2), \quad i = 1, ..., n, \tag{5.1}$$

where $m(x)$ is a smooth function in the univariate metrical quantity x. We assume that $m(x)$ can be of locally varying complexity and replace $m(x)$ for fitting by the penalized truncated polynomials

$$m(x) = \beta_0 + x\beta_1 + ... + x^q\beta_q + \sum_{s=1}^{K_b}(x - k_s^{(b)})_+^q b_s, \tag{5.2}$$

where $k_1^{(b)}, ..., k_{K_b}^{(b)}$ are knots covering the range of x and $(x - k_s^{(b)})_+^q$ is the truncated q-th order polynomial. The dimension K_b of the basis is chosen in a lush and generous manner and knots $k_s^{(b)}$ are placed over the range of x, e.g. using the quantiles of x. For fitting we impose a penalty term $\lambda b^T b$ on spline coefficients b. We present our routine for truncated polynomials for simplicity of notation, but any other basis functions described above can be used.
Mixed model representation arises from the assumption $b \sim N(0, \sigma_b^2 I_{K_b})$. The restriction explicitly occurring with this setting is that all coefficients have the same *a priori* variance and therewith undergo the same penalization. This is a critical point if the underlying function is of locally varying complexity. Like Crainiceanu, Ruppert & Carroll (2005) or Baladandayuthapani, Mallick & Carroll (2005) we therefore allow coefficients $b_1, ..., b_{K_b}$ to have locally varying variability which is accommodated by

$$b_s \sim N(0, \sigma_{bs}^2), \quad s = 1, ..., K_b.$$

We assume next that the variance components σ_{bs}^2 change smoothly over the (ordered) spline coefficients, meaning that the complexity of function $m(x)$ varies smoothly over x and does not change rapidly. A typical example for such function is the Doppler curve

(see left plot of Figure 5.1). We accommodate this assumption by setting $\sigma_{bs}^2 = \sigma_b^2(k_s^{(b)})$, where $\sigma_b^2(.)$ is a function smoothly varying over the knots of the basis. In a hierarchical manner the smooth structure is again modelled by P-splines. To do so we set

$$\sigma_b^2(k^{(b)}) = \exp[\gamma_0 + k^{(b)}\gamma_1 + ... + k^{(b)^p}\gamma_p + \sum_{t=1}^{K_c}(k^{(b)} - k_t^{(c)})_+^p c_t], \tag{5.3}$$

where $k_1^{(c)}, ..., k_{K_c}^{(c)}$ is a second layer of knots covering the range of $k_1^{(b)}, \ldots, k_{K_b}^{(b)}$. Note that $K_c < K_b$ is a restriction to be held and practically K_c is chosen far smaller than K_b. Extending now the smooth estimation, we fit $\sigma_b^2(\cdot)$ in a penalized form by imposing a penalty on coefficients c_t. From a Bayesian viewpoint this can be expressed as *a priori* distribution in the form

$$c_t \sim N(0, \sigma_c^2), \quad t = 1, ..., K_c.$$

Note that the variance σ_c^2 is set to be constant and serves as hyper parameter in our model construction.
For notational simplicity we rewrite the model in matrix form. Let, therefore,

$$y = (y_1, ..., y_n)^T,\ X_b = [1, x_i, \ldots, x_i^q]_{1\le i\le n},\ Z_b = [(x_i - k_1^{(b)})_+^q, \ldots, (x_i - k_{K_b}^{(b)})_+^q]_{1\le i\le n}$$

and write $\beta = (\beta_0, \ldots, \beta_q)^T$ and $b = (b_1, \ldots, b_{K_b})^T$. In an analogous way we define

$$X_c = [1, k_j^{(b)}, \ldots k_j^{(b)^p}]_{1\le j\le K_b},\ Z_c = [(k_j^{(b)} - k_1^{(c)})_+^p, \ldots, (k_j^{(b)} - k_{K_c}^{(c)})_+^p]_{1\le j\le K_b}$$

which gives the hierarchical model

$$\begin{aligned} y|b, c &= X_b\beta + Z_b b + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2 I_n), \\ b|c &\sim N(0, \Sigma_b),\ \Sigma_b = \mathrm{diag}[\exp(X_c\gamma + Z_c c)], \\ c &\sim N(0, \sigma_c^2 I_{K_c}). \end{aligned} \tag{5.4}$$

The corresponding likelihood results in

$$\begin{aligned} L(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) &= f(y; \beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) \\ &= (2\pi)^{-\frac{(n+K_c)}{2}} \sigma_\epsilon^{-n} \sigma_c^{-K_c} \int_{R^{K_c}} \exp[-g(c)]dc, \end{aligned} \tag{5.5}$$

with

$$g(c) = \frac{1}{2}\log|V_\epsilon| + \frac{c^T c}{2\sigma_c^2} + \frac{(y - X_b\beta)^T V_\epsilon^{-1}(y - X_b\beta)}{2\sigma_\epsilon^2}$$

and $V_\epsilon = I_n + Z_b\Sigma_b Z_b^T/\sigma_\epsilon^2$. Note that both, V_ϵ as well as Σ_b, depend on c and γ which is omitted throughout the paper for notational simplicity. The integral in (5.5) is not available analytically, which motivates a solution based on MCMC techniques as pursued in the previously cited papers. We, however, go a different route via Laplace approximation, which is justifiable for two reasons. First, the hierarchical model (5.4) is used as a vehicle for estimation only and has no specific data generating justification. This means finding the exact marginal likelihood by extensive numerics is not necessary, if an approximate version fulfills the task of estimation properly. Secondly, since K_c (and K_b) are assumed to be bounded while sample size n is growing, i.e. $K_c < K_b \ll n$, one finds function $g(\cdot)$ to be of order n. This implies that the Laplace approximation has an error of order $O(n^{-1})$ (see Severini, 2000). Therefore, the Laplace approximation appears as attractive alternative to simulation based techniques. The log-likelihood is then approximated, up to a constant, by

$$\begin{aligned} -2l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) &\approx n\log\sigma_\epsilon^2 + K_c\log\sigma_c^2 + \log|V_\epsilon(\hat{c})| + \log|I_{cc}(\hat{c})| \\ &+ \hat{c}^T\hat{c}/\sigma_c^2 + (y - X_b\beta)^T V_\epsilon^{-1}(\hat{c})(y - X_b\beta)/\sigma_\epsilon^2, \end{aligned} \tag{5.6}$$

where $\hat{c}_t,\ t = 1, ..., K_c$ is the solution to

$$\frac{\partial g(\hat{c})}{\partial c_i} = \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}\right) + \frac{c_i}{\sigma_c^2} - \frac{1}{2\sigma_\epsilon^2}(y - X_b\beta)^T V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}(y - X_b\beta) = 0 \tag{5.7}$$

and

$$[I_{cc}(c)]_{ij} = E\left(\frac{\partial^2 g(c)}{\partial c_i \partial c_j}\middle| c\right) = \frac{\delta_{ij}}{\sigma_c^2} + \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_j}\right), \tag{5.8}$$

with δ_{ij} as the Kronecker delta. It is not difficult to see that the derivative appearing in the above equations results in

$$\frac{\partial V_\epsilon}{\partial c_i} = Z_b\mathrm{diag}(Z_{c,i})\Sigma_b Z_b^T/\sigma_\epsilon^2,$$

where $Z_{c,i}$ stands for the ith column of the matrix Z_c. Moreover, noting that the prediction of b is defined through

$$Z_b^T V_\epsilon^{-1}(y - X_b\beta) = \sigma_\epsilon^2\Sigma_b^{-1}\hat{b}.$$

and $\mathrm{tr}\,(V_\epsilon^{-1}\partial V_\epsilon/\partial c) = Z_c^T w_{df}$, with w_{df} as K_b dimensional vector containing the diagonal elements of $A = Z_b^T Z_b(\sigma_\epsilon^2\Sigma_b^{-1} + Z_b^T Z_b)^{-1}$, we can represent (5.7) and (5.8) as

$$\frac{\partial g(c)}{\partial c} = -\frac{1}{2}Z_c^T\left\{\Sigma_b^{-1}\hat{b}^2 - w_{df}\right\} + \frac{c}{\sigma_c^2} = 0,$$

and

$$I_{cc}(c) = E\left(\frac{\partial^2 g(c)}{\partial c\partial c^T}\middle| c\right) = \frac{1}{2}Z_c^T\mathrm{diag}(v_{df})Z_c + \frac{I_{K_c}}{\sigma_c^2},$$

with v_{df} as K_b dimensional vector containing the diagonal elements of AA. Note that $df_b = \sum_{s=1}^{K_b} w_{df} = 1_{K_b}^T w_{df}$ measures the degree of freedom used for fitting b. In particular, for K_b assumed to be fixed we find $df_b \to K_b$ as n tends to infinity and both w_{df} and v_{df} tend to 1_{K_b}.

Assuming that weights v_{df} vary slowly or not at all as a function of γ (which is readily seen from $\partial v_{df}/\partial\gamma_i = 2\mathrm{diag}[(AA - AAA)\mathrm{diag}(X_{c,i})]$ with $X_{c,i}$ as the ith column of the matrix X_c) we can estimate γ and c simultaneously, resulting in the following iterated weighted least squares (IWLS) for estimation of parameter $\theta = (\gamma^T, c^T)^T$

$$\hat{\theta} = \left(W_c^T\mathrm{diag}\left(\frac{v_{df}}{2}\right)W_c + \frac{D_c}{\sigma_c^2}\right)^{-1} W_c^T\mathrm{diag}\left(\frac{v_{df}}{2}\right)u, \tag{5.9}$$

with $W_c = (X_c, Z_c)$, $D_c = \mathrm{diag}(0_{(p+1)\times(p+1)}, I_{K_c})$ and $u = W_c\theta + \mathrm{diag}(v_{df}^{-1})(\Sigma_b^{-1}\hat{b}^2 - w_{df})$ as a working vector. Fixing now parameter $\hat{\theta}$ provides, with the above log-likelihood (5.6), the following parameter estimates

$$\begin{aligned}\hat{\sigma}_c^2 &= \hat{c}^T\hat{c}/w_{df}^c \\ \hat{\beta} &= (X_b^T V_\epsilon^{-1}(\hat{\theta})X_b)^{-1}X_b^T V_\epsilon^{-1}(\hat{\theta})y, \\ \hat{\sigma}_\epsilon^2 &= (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/n,\end{aligned} \tag{5.10}$$

with $w_{df}^c = \mathrm{tr}(Z_c\mathrm{diag}(v_{df})Z_c^T I_{cc}^{-1}/2)$ and obvious definition for $V_\epsilon(\theta)$. Finally, we obtain the estimated best linear unbiased predictor (EBLUP) via

$$\hat{b} = \hat{\Sigma}_b Z_b^T\hat{V}_\epsilon^{-1}(y - X_b\hat{\beta})/\hat{\sigma}_\epsilon^2.$$

The latter steps are standard and available from linear mixed models technology. Estimation can now be carried out with the standard in mixed models framework EM type algorithm (see e.g. Searle, Casella, & McCulloch, 1992 or Breslow & Clayton, 1993) by iterating between (5.9) and (5.10) until convergence. It should be noted that the

estimation consists of two simple steps and is, therefore, numerically very fast.

5.2.2 Restricted maximum likelihood

The above results are presented for maximum likelihood estimates. The use of restricted maximum likelihood (REML) is, however, more common in mixed models. This approach makes a small sample adjustment for the estimation of the unpenalized parameters β, see Section 2.2.1. The restricted maximum log-likelihood for the model (5.4) takes the form

$$l_R(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) = l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2) - \frac{1}{2} \log |X_b^T V_\epsilon^{-1}(\hat{c}) X_b / \sigma_\epsilon^2|,$$

with $l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2)$ as given in (5.6). The further estimation procedure is identical to that described in the previous section, with the matrix A in w_{df} and v_{df} being replaced by

$$A_R = A - Z_b^T V_\epsilon^{-1} X_b (X_b^T V_\epsilon^{-1} X_b)^{-1} X_b^T Z_b (Z_b^T Z_b + \Sigma_b^{-1} \sigma_\epsilon^2)^{-1}$$

and the variance estimate defined as $\hat{\sigma}_\epsilon^2 = (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/n - q - 1$. We have compared the performance of both procedures in a simulation study presented in Section 5.2.5 and found little difference in estimates. However, REML estimates demonstrated a slightly better numerical stability for implementation in R.

5.2.3 Variance estimation

We denote with $\tilde{m}(x)|c = X_b\tilde{\beta} + Z_b\tilde{b}|c$ the best linear unbiased predictor (BLUP) of the function $m(x)|c = X_b\beta + Z_b b|c$, where $\tilde{\beta} = (X_b^T V_\epsilon^{-1} X_b)^{-1} X_b^T V_\epsilon^{-1} y$ and $\tilde{b}|c = \Sigma_b Z_b^T V_\epsilon^{-1}(y - X_b\tilde{\beta})/\sigma_\epsilon^2$. Note that within the linear mixed model framework the function $m(x)|c$ is random due to randomness of parameter b. Since $\tilde{m}(x)|c$ is unbiased for $m(x)|c$, the confidence intervals for $m(x)|c$ can be obtained from

$$[\tilde{m}(x) - m(x)]|c \sim N(0, \mathrm{Var}[\tilde{m}(x) - m(x)|c]),$$

where $\mathrm{Var}[\tilde{m}(x) - m(x)|c] = \sigma_\epsilon^2 S(\theta) = \sigma_\epsilon^2 W_b (W_b^T W_b + \sigma_\epsilon^2 D_b(\theta))^{-1} W_b^T$ with $W_b = (X_b, Z_b)$ and $D_b(\theta) = \mathrm{diag}(0_{(q+1)\times(q+1)}, \Sigma_b^{-1})$. Using the delta method and unbiasedness of $\tilde{m}(x)|c$ one can approximate the unconditional variance with

$$\mathrm{Var}[\tilde{m}(x) - m(x)] = E[\mathrm{Var}(\tilde{m}(x) - m(x)|c)] + \mathrm{Var}[E(\tilde{m}(x) - m(x)|c)] \approx \sigma_\epsilon^2 S(\hat{c}).$$

Let now $\hat{m}(x)|c = X_b\hat{\beta} + Z_b\hat{b}|c$ denote the estimated best linear unbiased predictor (EBLUP), obtained from $\tilde{m}(x)|c$ by plugging in the estimates of variance parameters. This can be used to obtain a plug in estimate $\widehat{\text{Var}}[\hat{m}(x) - m(x)] \approx \hat{\sigma}_\epsilon^2 S(\hat{\theta})$.
The variance estimate can also be calculated and justified within the Bayesian framework. Assuming parameters $\Sigma_b = \text{diag}[\exp(W_c\theta)]$ and σ_ϵ^2 are known, the posterior distribution of $m(x)$ is $N(\hat{m}(x,\theta), \sigma_\epsilon^2 S(\theta))$, where $\hat{m}(x,\theta) = S(\theta)y$. An empirical Bayes approach would now replace the unknown values Σ_b and σ_ϵ^2 in the prior by estimates and then treat these parameters as if they were known and given in advance. Thus, the approximate posterior distribution of $m(x)$ results in $N(\hat{m}(x,\hat{\theta}), \hat{\sigma}_\epsilon^2 S(\hat{\theta}))$, yielding the same confidence intervals as in the linear mixed model framework.
Even though the variance formula has the advantage of being simple it does not, however, account for the extra variability due to estimation of θ, that is the local varying penalty. This is the price to pay when using Laplace's method instead of a full Bayesian approach. For further discussion we refer to Morris (1983), Laird & Louis (1987), Kass & Steffey (1989) or Ruppert & Carroll (2000). To correct for this we now estimate the posterior variance of $m(x)$ calculated from the joint posterior distribution of b and θ. We, therefore, use the delta-method correction from Kass & Steffey (1989) and obtain

$$\begin{aligned}\text{Var}(m(x)|y) &= E[\text{Var}(\hat{m}(x)|\hat{\theta}, y)] + \text{Var}[E(\hat{m}(x)|\hat{\theta}, y)] \\ &\approx \hat{\sigma}_\epsilon^2 S(\hat{\theta}) + \left(\left.\frac{\partial \hat{m}(x,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right)^T \text{Var}(\hat{\theta}) \left(\left.\frac{\partial \hat{m}(x,\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}}\right).\end{aligned}$$

As estimate of $\text{Var}(\hat{\theta})$ one can use the inverse of the Fisher matrix $I_{\theta\theta}(\hat{\theta})$ resulting from the last iteration as by-product. The derivative in the last term, ignoring the dependence of $\hat{\sigma}_\epsilon^2$ on θ, results in

$$\left.\frac{\partial \hat{m}(x,\theta)}{\partial\theta_i}\right|_{\theta_i=\hat{\theta}_i} = \hat{\sigma}_\epsilon^2 W_b (W_b^T W_b + \hat{\sigma}_\epsilon^2 \hat{D}_b)^{-1} \tilde{W}_{c,i} \hat{D}_b (W_b^T W_b + \hat{\sigma}_\epsilon^2 \hat{D}_b)^{-1} W_b^T y,$$

with $\hat{D}_b = D_b(\hat{\theta})$ and $\tilde{W}_{c,i} = \text{diag}(0_{(q+1)\times(q+1)}, W_{c,i})$, where $W_{c,i}$ stands for the i-th column of matrix W_c.

5.2.4 Numerical implementation

For the numerical implementation one can make use of any standard mixed models software. More precisely, we use the following algorithm:

1. Obtain initial estimates for all parameters from a non-adaptive fit, using any mixed

model software;

2. Get next estimates for $\hat{\theta}$ and $\hat{\sigma}_c^2$ from (5.9) and (5.10);

3. Update estimates for the remaining parameters with a mixed model software, taking the estimated variance matrix $\hat{\Sigma}_b = \text{diag}[\exp(W_c\hat{\theta})]$ into account;

4. Iterate between 2 and 3 until convergence.

We implemented this algorithm in the package *AdaptFit* described below. With respect to the splines we experimented with a number of spline basis functions, such as B-splines of different degree and penalty order, quadratic and cubic truncated polynomials as well as cubic thin plate splines. Although all basis functions produced very similar, in fact almost indistinguishable, results, the cubic thin plate splines demonstrated a slightly better numerical stability and were preferred for the simulation study. Knots dimensions K_b and K_c need also to be chosen carefully to ensure capturing a complex function structure in the regions of a higher variability.

5.2.5 Simulations and comparisons with other univariate smoothers

We performed a number of simulations. A particular focus is to compare our results with those reported in Ruppert & Carroll (2000) and Baladandayuthapani, Mallick & Carroll (2005). First, for $n = 400$ x equally spaced on $[0, 1]$ and independent $\epsilon_i \sim N(0, 0.2^2)$ we

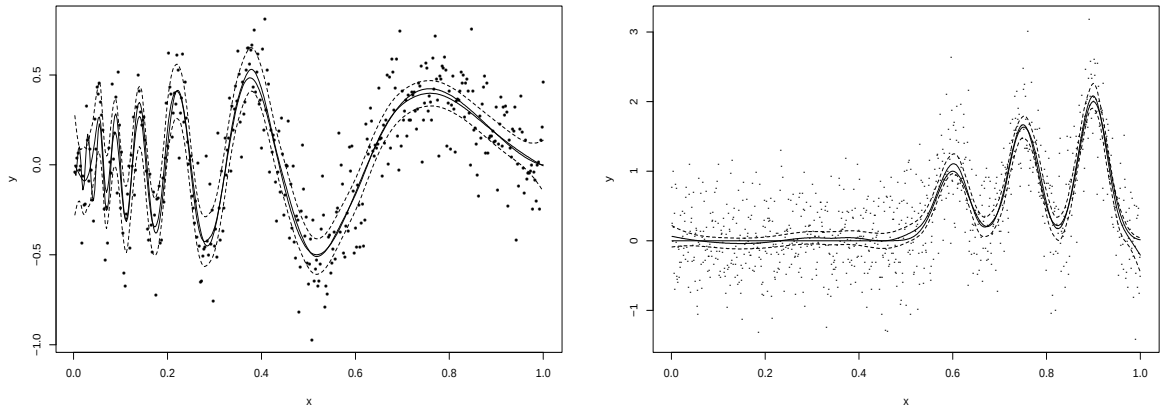

Figure 5.1: Estimated regression functions $m_1(x)$ (left) and $m_2(x)$ (right) with confidence intervals (dashed) and true function.

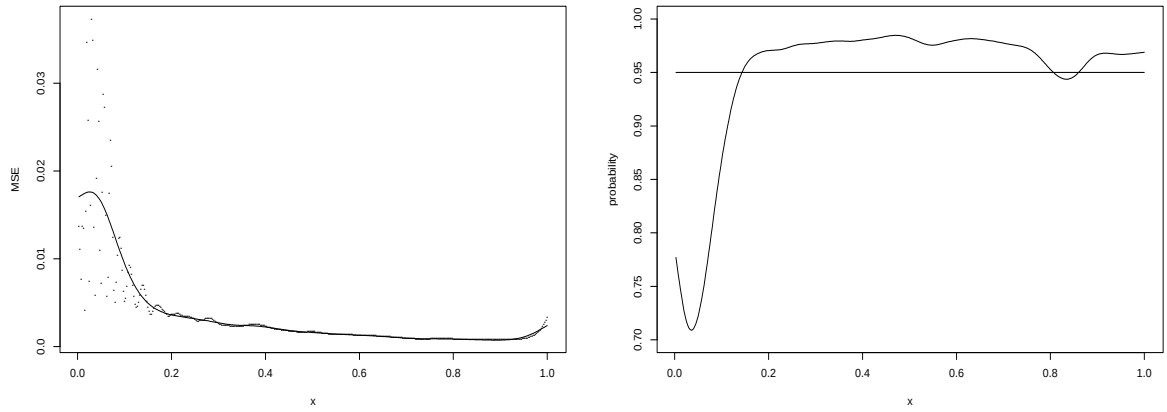

Figure 5.2: Pointwise MSE with a smoother of the points (left) and smoothed pointwise coverage probabilities of 95% confidence intervals (right) for 500 simulated datasets with function $m_1(x)$

examined the regression function

$$m_1(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right),$$

with $j = 6$. We performed 500 simulations with $K_b = 80$ and $K_c = 20$. An exemplary fit (bold) together with confidence intervals (dashed) is shown in the left plot of Figure 5.1. The left plot of Figure 5.2 displays the pointwise Mean squared error $E(\{\hat{m}(x)-m(x)\}^2)$ with the expectation being replaced by the mean of the simulations. For better visual impact, we show a simple smoother (thick line) for the latter. The average MSE over all x's (AMSE) equals 0.0033, which is comparable with 0.0027 reported in Baladandayuthapani, Mallick & Carroll (2005) and 0.0026 of Ruppert & Carroll (2000). We also computed the coverage probabilities of the 95% confidence intervals over all 500 simulated datasets. The right plot of Figure 5.2 shows smoothed pointwise coverage probabilities. For small values of $x \leq 0.1$, i.e. in the region with low signal-to-noise ratio, there is clear undercoverage, but beyond 0.1 the coverage probability exceeds 95% being slightly conservative. The average coverage probability is 94.95%. Next, we consider the heterogeneous regression function

$$m_2(x) = \exp(-400(x-0.6)^2) + \frac{5}{3}\exp(-500(x-0.75)^2) + 2\exp(-500(x-0.9)^2).$$

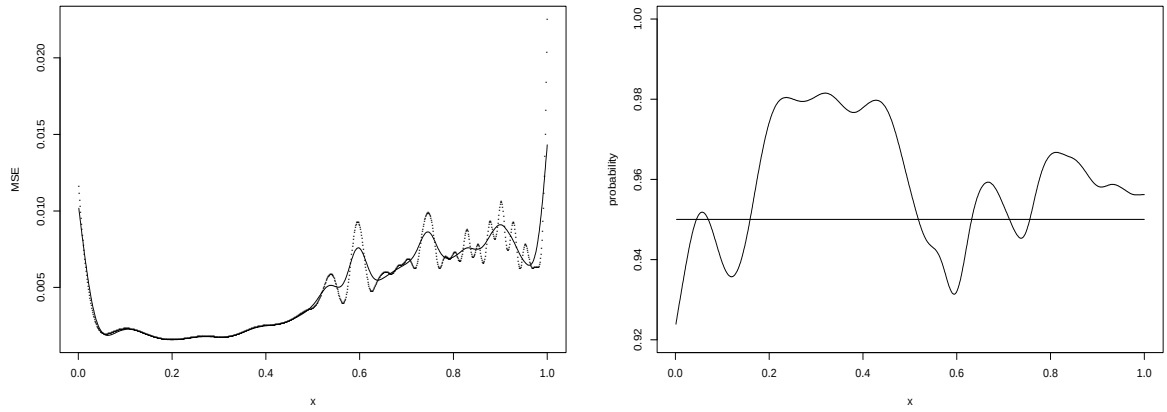

Figure 5.3: Pointwise MSE with a smoother of the points (left) and smoothed pointwise coverage probabilities of 95% confidence intervals (right) for 500 simulated datasets with function $m_2(x)$.

Now $n = 1000$ x values are equally spaced on $[0, 1]$ and $\epsilon_i \sim N(0, 0.5^2)$. We apply our approach to 500 simulated datasets, using $K_b = 40$ and $K_c = 4$. Figures 5.1 (right) and 5.3 (left) represent one of the simulated fits and pointwise MSE respectively. The resulted AMSE is equal 0.0049, which is somewhat smaller than 0.0061 and 0.0065, obtained by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000) respectively. The smoothed pointwise coverage probabilities can be seen in the right plot of Figure 5.3. The average coverage probability for this function is 95.94%, which is comparable with 95.22% and 96.28% reported by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000) respectively. Overall, our method provides comparable results to the other approaches, but with less numerical effort.

5.3 Spatial smoothing

5.3.1 Hierarchical modelling

We now generalize the ideas of the previous section with regards to spatial smoothing

$$y_i \sim N(m(\boldsymbol{x}_i), \sigma_\epsilon^2), \quad i = 1, ..., n,$$

with $\boldsymbol{x}_i \in R^2$ and $m(.)$ as a smooth function of two covariates. Following Crainiceanu, Ruppert & Carroll (2005) we use radial basis functions and choose K_b knots $\boldsymbol{k}_1^{(b)}, ..., \boldsymbol{k}_{K_b}^{(b)} \in R^2$. This defines the model matrices $X_b = [1, \boldsymbol{x}_i^T]_{1\leq i\leq n}$ and $Z_b = Z_{K_b}\Omega_{K_b}^{-1/2}$ where $Z_{K_b} = [\|\boldsymbol{x}_i - \boldsymbol{k}_s^{(b)}\|^2 \log\|\boldsymbol{x}_i - \boldsymbol{k}_s^{(b)}\|]_{1\leq s\leq K_b, 1\leq i\leq n}$ and $\Omega_{K_b} = [\|\boldsymbol{k}_t^{(b)} - \boldsymbol{k}_s^{(b)}\|^2 \log\|\boldsymbol{k}_t^{(b)} - \boldsymbol{k}_s^{(b)}\|]_{1\leq s,t\leq K_b}$ with $\|.\|$ denoting the Euclidean norm in R^2. Including penalties and using the link to linear mixed model we get

$$\begin{aligned} y|b &= X_b\beta + Z_b b + \epsilon,\ \epsilon \sim N(0, \sigma_\epsilon^2 I_n), \\ b &\sim N(0, \Sigma_b). \end{aligned} \tag{5.11}$$

Local adaptive smoothing is now implemented by allowing coefficients b to have locally varying variability. Like above we set subknots $\boldsymbol{k}_1^{(c)}, ..., \boldsymbol{k}_{K_c}^{(c)} \in R^2$, $K_c < K_b$ and define matrices X_c and Z_c similarly to the corresponding definition of matrices X_b and Z_b, that is $X_c^s = [1, (\boldsymbol{k}_s^{(b)})^T]_{1\leq s\leq K_b}$, $Z_c = Z_{K_c}\Omega_{K_c}^{-1/2}$ with $Z_{K_c} = [\|\boldsymbol{k}_s^{(b)} - \boldsymbol{k}_t^{(c)}\|^2 \log\|\boldsymbol{k}_s^{(b)} - \boldsymbol{k}_t^{(c)}\|]_{1\leq s\leq K_b, 1\leq t\leq K_c}$ and $\Omega_{K_c} = [\|\boldsymbol{k}_t^{(c)} - \boldsymbol{k}_s^{(c)}\|^2 \log\|\boldsymbol{k}_t^{(c)} - \boldsymbol{k}_s^{(c)}\|]_{1\leq s,t\leq K_c}$ where the $\boldsymbol{x}$ covariates are replaced by knots $\boldsymbol{k}^{(b)}$ and the knots are replaced with subknots $\boldsymbol{k}^{(c)}$. The model is completed by adding to (5.11) the hierarchical structure

$$\Sigma_b = \text{diag}[\exp(X_c\gamma + Z_c c)], \quad c \sim N(0, \sigma_c^2 I_{Kc}).$$

Estimation can now be carried out analogously to above. The knots can be selected with *clara* algorithm described in Kaufman & Rousseeuw (1990) and implemented in the R package "cluster".

5.3.2 Simulations and comparisons with other surface fitting methods

For comparison with Crainiceanu, Ruppert & Carroll (2005) and Lang & Brezger (2004) we consider the following regression function with moderate spatial variability

$$m_3(x_1, x_2) = x_1 \sin(4\pi x_2),$$

with x_1 and x_2 independently, uniformly distributed on $[0, 1]$. We used $n = 300$, $\sigma = 1/4\text{range}(m_3)$ and equally-spaced 12×12 and 5×5 knot grids for $k_i^{(b)}$ and $k_j^{(c)}$, respectively. Figure 5.4 displays the true function, an adaptive fit for one simulation and the same data fitted non-adaptively. We simulated 500 datasets to compare log(MSE) of our estimator with values reported in Crainiceanu, Ruppert & Carroll (2005) and

Lang & Brezger (2004). Our simulations provide a median of log(MSE) of -4.13 with

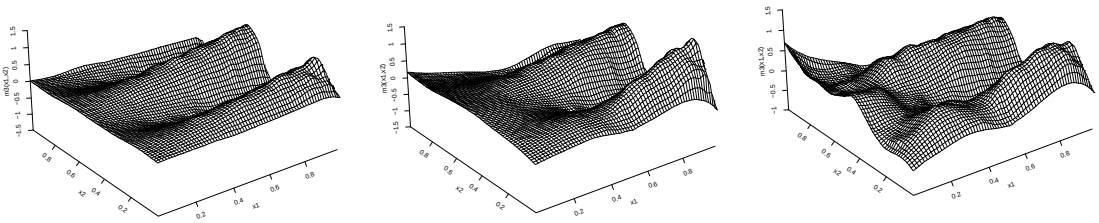

Figure 5.4: True regression function $m_3(x_1, x_2)$, adaptive and non-adaptive estimates

an interquartile range [-4.78,-3.82] and a range [-5.46,-2.7]. This is comparable with the results in Crainiceanu, Ruppert & Carroll (2005) (median -3.67, interquartile range [-3.80,-3.53] and a range [-4.21,-3.13]) which outperform the findings of Lang & Brezger (2004). The obtained AMSE equals 0.0176. The average coverage probability of the 95% confidence intervals is 94.31%. The smoothed coverage probabilities are displayed in Figure 5.5. Similarly to the Crainiceanu, Ruppert & Carroll (2005), the coverage

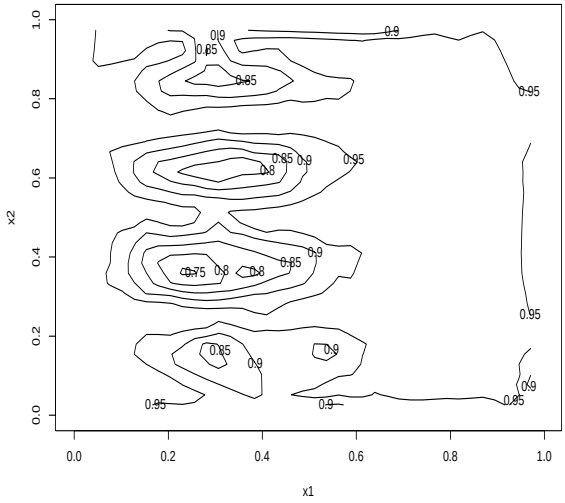

Figure 5.5: Smoothed coverage probability of 95% confidence intervals for 500 simulated datasets with function $m_3(x_1, x_2)$.

probability is lowest for $x_1 \in [0.2, 0.5]$. This is explained by the low signal-to-noise ratio in this region.

5.4 Non-normal response model

5.4.1 Hierarchical modelling

The technique is now extended to non normal response models by considering the following generalized linear hierarchical mixed model

$$\begin{aligned} E(y|b,c) &= \mu^{b,c} = h(X_b\beta + Z_b b),\ \mathrm{Var}(y|b,c) = \phi v(\mu^{b,c}), \\ b|c &\sim N(0,\Sigma_b),\ \Sigma_b = \mathrm{diag}[\exp(X_c\gamma + Z_c c)], \\ c &\sim N(0,\sigma_c^2 I_{K_c}), \end{aligned}$$

with function $h(.)$ as the inverse of link function $\tilde{g}(.)$, $v(.)$ as some specified variance function and ϕ as dispersion parameter. We follow Breslow & Clayton (1993) and estimate the parameters from the quasi-likelihood

$$\exp[ql(\beta,\gamma,\sigma_c^2)] = (2\pi)^{-\frac{(K_b+K_c)}{2}}\sigma_c^{-K_c}\int_{R^{K_b}}\int_{R^{K_c}}\exp[-k_1(b,c)]db\,dc, \qquad (5.12)$$

with

$$k_1(b,c) = \frac{1}{2\phi}\sum q_i(y_i,\mu_i^{b,c}) + \frac{1}{2}b^T\Sigma_b^{-1}b + \frac{1}{2}\log|\Sigma_b| + \frac{1}{2\sigma_c^2}c^Tc$$

and

$$q_i(y,\mu) = -2\int_y^\mu \frac{y-t}{v(t)}dt,$$

as deviance measure of the fit. Assuming that conditionally on b and c the observations are drawn from the exponential family $y|b,c \sim \exp[(y\vartheta(x) - b(\vartheta(x)))/\phi + c(y,\phi)]$, the quasi-likelihood (5.12) represents the true likelihood of the data. Using Laplace's method for approximation of the integral over b, one gets

$$\exp[ql(\beta,\gamma,\sigma_c^2)] \approx (2\pi)^{-\frac{K_c}{2}}\sigma_c^{-K_c}\int_{R^{K_c}}\exp[-k_2(c)]dc, \qquad (5.13)$$

with

$$k_2(c) = \frac{1}{2}\log|I_n + Z_b^TWZ_b\Sigma_b| + \frac{1}{2\phi}\sum q_i(y_i,\mu_i^{b,c}) + \frac{1}{2}\hat{b}^T\Sigma_b^{-1}\hat{b} + \frac{1}{2\sigma_c^2}c^Tc,$$

where $\hat{b}$ is the solution to

$$\frac{\partial k_1(b,c)}{\partial b} = -Z_b^TW\mathrm{diag}[\tilde{g}'(\mu^{b,c})](y-\mu^{b,c}) + \Sigma_b^{-1}b = 0,$$

with W as the $n \times n$ diagonal matrix of GLM iterated weights with diagonal elements $w_i = (\phi v(\mu_i^{b,c})[\tilde{g}'(\mu_i^{b,c})]^2)^{-1}$, using the simplifying assumption that the iterative weights w_i vary only slowly (or not at all) with the mean. Substituting the current estimate $\hat{b}$ into (5.13) and replacing the deviance $\sum q_i(y_i, \mu_i^{b,c})$ in $k_2(.)$ by the Pearson chi-squared statistic $\sum(y_i - \mu_i^{b,c})^2/v_i(\mu_i^{b,c})$ results in

$$\exp[ql(\beta, \gamma, \sigma_c^2)] \approx (2\pi)^{-\frac{K_c}{2}} \sigma_c^{-K_c} |W|^{-1/2} \int_{R^{K_c}} \exp[-k_3(c)]dc,$$

with

$$k_3(c) = \frac{1}{2} \log |V| + \frac{c^T c}{2\sigma_c^2} + (U - X_b\beta)^T V^{-1}(U - X_b\beta),$$

where $V = W^{-1} + Z_b \Sigma_b Z_b^T$ and $U = X_b\beta + Z_b\hat{b} + \text{diag}[\tilde{g}'(\mu^{b,c})](y - \mu^{b,c})$. Applying again Laplace's method, we end up with the following quasi-log-likelihood for the remaining parameters

$$\begin{aligned} -2l(\beta, \gamma, \sigma_c^2) &\approx K_c \log \sigma_c^2 + \log |V| + \log |k_3^{cc}| \\ &+ \hat{c}^T \hat{c}/\sigma_c^2 + (U - X_b\beta)^T V^{-1}(U - X_b\beta), \end{aligned}$$

with $k_3^{cc} = \partial^2 k_3(c)/\partial c \partial c^T$. In complete analogy to Section 5.2.1 the estimation of parameter $\theta = (\gamma^T, c^T)^T$ can be carried out from the score equation

$$\frac{\partial k_3(\hat{\theta})}{\partial \theta} = -\frac{1}{2} W_c^T \Sigma_b^{-1} \left\{ \hat{b}^2 - w_{df}\sigma_b^2 \right\} + D_c\theta/\sigma_c^2 = 0. \tag{5.14}$$

Numerically this procedure can be implemented by iterating between estimation of $\hat{\theta}$, and thus $\hat{\Sigma}_b$ from (5.14), and calls of any generalized linear mixed models software.

5.4.2 Simulations for the logistic regression example

We consider the following model for binomial data $y_i \sim B(n_i, \pi_i)$ with canonical link

$$\begin{aligned} \text{logit}[E(y_i|x_i)/n_i] &= \log\left(\frac{\pi_i}{1 - \pi_i}\right) = X_b^i\beta + Z_b^i b,\ i = 1, ..., n, \\ b|c &\sim N(0, \Sigma_b),\ \Sigma_b = \text{diag}[\exp(X_c\gamma + Z_c c)], \\ c &\sim N(0, \sigma_c^2 I_{K_c}). \end{aligned}$$

The diagonal elements of the iterative weights in matrix W for this model equal $w_i = 1/n_i\pi_i(1-\pi_i)$. We simulated data with probabilities $\pi = \text{logit}^{-1}[m_2(x)]$, where function $m_2(.)$ is the same as in Section 5.2.5. Figure 5.6 represents exemplarily the fit for the grouped data with $n_i = 5$ and $n = 1000$ (left) and the fit for $n = 5000$ binary data (right). For comparison the fit with global smoothing parameter is also presented (dashed). The benefit of local adaptivity is obvious.

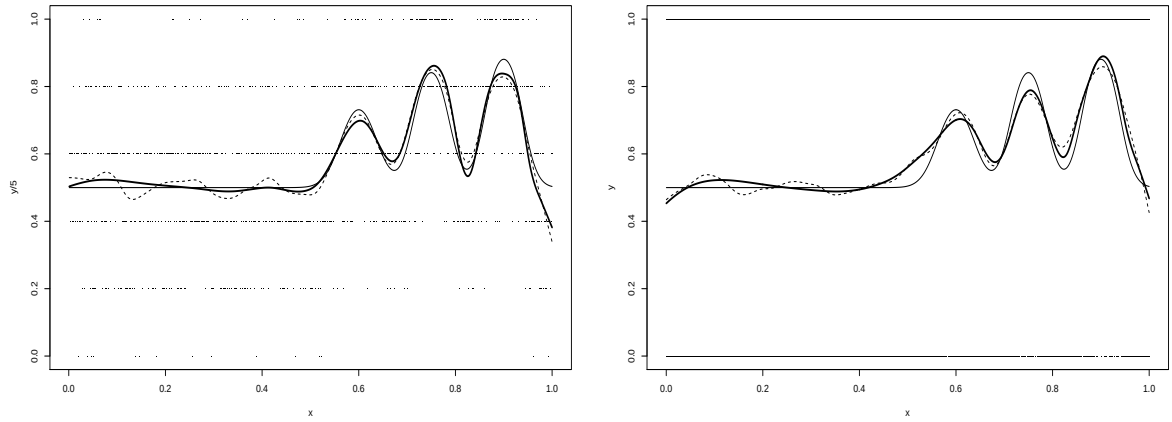

Figure 5.6: Estimated regression function $\pi = \text{logit}^{-1}[m_2(x)]$ with adaptive penalty (bold), with global smoothing parameter (dashed) and true function for 1000 grouped binomial data ($n_i = 5$) (left) and 5000 binary data (right).

5.5 Example

For demonstrational purposes we apply the above spatially adaptive smoothing technique to a dataset on the absenteeism of workers of a company in Germany. Parts of the data have been analysed before in Kauermann & Ortlieb (2004), though with a different focus. We consider absenteeism spells and model the probability of returning to work after a sick leave. With d denoting the duration of such a leave we model the discrete hazard rate

$$P(d = t|d \geq t) = h(t), \tag{5.15}$$

where $t = 1, 2, ...$ The duration is thereby measured in days and the event of interest is the recovery that allows workers to return to work. If the worker has been reported sick on one day, say Tuesday, but returns to work on a consecutive working day thereafter,

we count this as an event and the duration is the number of working days the worker has been absent. If, in contrast, the last days of absenteeism and the first day of returning to work are not consecutive working days, we consider the duration as censored observation, and d gives the number of consecutive working days of absenteeism. To make this more explicit, assume that a worker reports himself sick on Friday, but returns to work the Monday after. It is unclear when the worker actually recovered, either already Friday, Saturday or Sunday. It is however known, that the worker was called sick at least one day and the observation is, therefore, $d = 1$ with censoring indicated. Let now δ denote the censoring indicator which is either zero, for censoring, or 1, otherwise. For each absence spell we transform d to the binary variables $y_1, ..., y_d$ with $y_l = 0$ for $l < d$ and $y_d = \delta$. The hazard function is then the probability $P(y_t = 1|y_l = 0, l < t)$. We concentrate on short term absenteeism spells truncated at $d = 10$ and take longer spells as censored observations. Besides the explicit duration time we allow the hazard function to depend on calendar time c as well, where c is the first day of the worker's absenteeism spell. The final model is then

$$\text{logit}P(d = t|d \geq t, c) = m(t, c), \tag{5.16}$$

which is fitted in a locally adaptive way below.

The data were collected in company in Southern Germany and we analyse the data of 378 employees. Not all of them were employed at the same time with the observation period ranging from 1981 to 1998. On average, about 3/4 of the employees have been reported sick at least once per calender year. We assume that the durations of different sick leaves of the same worker are independent and even though it might be argued whether this is an appropriate assumption, for sake of simplicity we leave this issue aside for now. Figure 5.7 shows the fit of the model (5.16) using non-adaptive (left) and adaptive (right) smoothing. Both fits were carried out using 12 knots for each dimension and low-rank thin spline basis as defined in Section 5.3.1 The variance structure for the adaptive fit was modelled with 9 knots for each dimension. The differences in the plots are quite obvious. Both fits expose a bump at 1992 and 1993 and day 3, which becomes even more peaked for the spatially adaptive fit. Beyond this peak, particularly for longer absenteeism time, the non-adaptive fit is quite wiggled while the adaptive approach selects a smooth, flat behaviour. The latter fit appears preferable and once more demonstrates the benefits of spatial adaptivity. The peak at year 1993 and duration time at day 3 allows for an interesting economic interpretation. In 1992/93 the company went through a major downsizing process with more than 50% of the workers being dismissed. While this economic situation has hardly any effect on the hazard function for days $d \geq 5$, it does

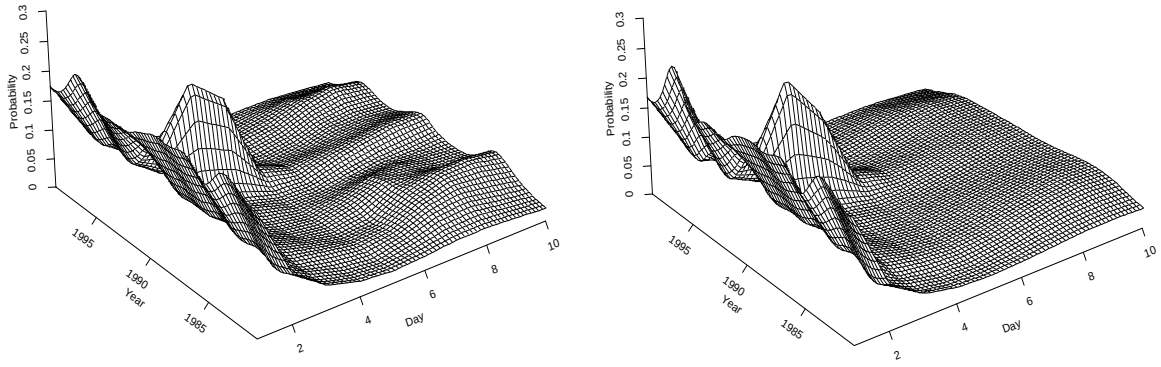

Figure 5.7: Estimated regression function $P(d = t|d \geq t, c) = \text{logit}^{-1}(m(t, c))$ with global (left) and local (right) smoothing parameter.

affect the hazard rate for short absenteeism times, particularly for $d = 3$. Due to the German law, workers reporting themselves sick for more than 3 consecutive days have to provide a medical certificate, at the latest by the third day of their sick leave, while for shorter periods no special medical documentation is required. Apparently, during the downsizing period the duration of sick leaves is clearly shorter with more employees returning after 3 days. This provides indication that economically critical conditions of a company have a direct influence on the absenteeism of employees. Looking further into the data it can be seen that it is mainly employees who are being dismissed who tend to change their absenteeism behaviour (see also Kauermann & Ortlieb, 2004), Figure 5.7 shows how this is changed. Moreover, the locally adaptive smoothing exposes the peak more clearly without overfitting the remaining regions and, therefore, justifies the additional modelling effort.

5.6 R Package *AdaptFit*

To implement our approach we developed an R package. We took advantage of the R package "SemiPar". To perform adaptive smoothing we integrate the Fisher scoring procedure (5.9) for θ with updates of the remaining parameters by subsequent calls of function "spm". The current version of our package "AdaptFit" with the function "asp" is available at http://cran.r-project.org. In general, the usage of "asp" is similar to that of function "spm". For example, the simulation of function $m_1(x)$, as described in Section

5.2.5, can be performed as

```
> x <- 1:400/400
> mu <- sqrt(x*(1-x))*sin((2*pi*(1+2^((9-4*6)/5)))/(x+2^((9-4*6)/5)))
> y <- mu+0.2*rnorm(400)
> kn <- default.knots(x,80)
> kn.var <- default.knots(kn,20)
> y.fit <- asp(y~f(x,knots=kn,var.knot=kn.var))
> plot(y.fit)
```

Switching between maximum likelihood and restricted maximum likelihood estimation can be done by specifying `spar.method="ML"`. Other examples are provided within the package. The algorithm used is given in Section 5.2.4. After defining the initial estimates from the simple non-adaptive fit we get estimates $\hat{\theta}$ and $\hat{\sigma}_c^2$ as defined in (5.9) and (5.10), standardize the random effects b with the current estimate of $\hat{\Sigma}_b = \text{diag}[\exp(W_c\hat{\theta})]$ and call the function `spm` of *SemiPar* package to obtain the remaining estimates. Usually convergence is achieved after 3-6 iterations.

5.7 Discussion

We demonstrated how locally adaptive smoothing can be easily carried out by formulating penalties on a spline coefficient as hierarchical mixed model. The major contribution was to show how simple Laplace approximation of the marginal likelihood allows for the fitting of such models relatively easily without MCMC methods. This also applies to more general settings like spatial smoothing or generalized response models. In an empirical Bayes style the method also allows one to estimate the remaining hyper parameters, so that the procedure is, in fact, fully adaptive. Developed R package *AdaptFit* allows for fast and convenient application of this technique.

6 Some Asymptotics on Penalized Splines

6.1 Introduction

Even though P-splines are practically convincing, theoretical investigations of their performance and properties are less explored. A recent investigation is found in Opsomer & Hall (2005) who reformulate the approach as white noise representation. Some first results were provided in Wand (1999) with subsequent work in Aerts, Claeskens & Wand (2002). The latter two papers are based on the simplifying assumption that the dimension of the spline basis is fixed. This is, apparently, a stringent assumption in theoretical terms, which however has proven to be of little practical impact if the dimension of the spline basis is chosen in lush and generous manner. A suggested rule of thumb is to select $\min(n/4, 40)$ spline basis functions with n as sample size. Also, knot selection for the spline basis is of secondary importance as investigated in Ruppert (2002). The theoretical advantage of fixing the number of spline functions in advance is that asymptotically one achieves a parametric model and penalization loses its influence. This is, however, based on the assumption that the true underlying function is, in fact, representable by the (finite dimensional) basis. The contrary approach is to assume that the spline basis grows at the same rate as the number of observations. This leads to classical spline theory with one spline basis per observation, see e.g. Wahba (1990) or Eubank (1999). Letting the spline basis grow with the sample size induces numerical problems, in particular when the sample size is large.

We start from a penalized spline approach, but allow the number of spline basis functions to depend on the sample size. We explore the asymptotic rate at which the dimension of the spline basis is supposed to grow such that the mean squared error of the estimate is minimized. The particular focus is thereby on the bias component which decomposes into two parts, one part occurring due to the penalized estimation, the second due to working with a spline basis of smaller dimension than the sample size. Our work thereby relates to Huang & Stone (2003) who consider unpenalized estimation and Cardot (2002)

who presents similar results but works with a different penalty. The framework for our theoretical investigation is based on generalized smoothing models of the form

$$\mu(x) = \mathrm{E}\{y(x)\} = h\{\eta(x)\}, \tag{6.1}$$

with x as a continuous covariate and y as response, assumed to be distributed according to an exponential family distribution. Function $h(\cdot)$ is a known invertible (inverse) link function while function $\eta(x)$ is supposed to be smooth and will be estimated via penalized spline smoothing.

The connection of penalized spline smoothing to mixed models results if the penalty imposed on the spline coefficients is written as a Gaussian prior. In case of a normally distributed response and a canonical link function $h(\cdot)$, the penalized spline estimate for the smoothing model is equivalent to a posteriori Bayes predictor in the linear mixed model. Moreover, the smoothing parameter steering the amount of penalization becomes the ratio of the dispersion parameter over the a priori variance of the random spline effect. This has a practical implication, since smoothing parameter selection can now be carried out by maximum likelihood (ML) or restricted maximum likelihood (REML) estimation. The correspondence is investigated analytically in Kauermann (2004) for penalized splines by keeping the dimension of the spline basis fixed. If the response is not normally distributed, but a generalized smoothing model like (6.1) is assumed, penalized spline fitting can be linked to generalized linear mixed models (GLMM). This results analogously in the normal response case by imposing an a priori normal distribution on the spline coefficients. Integrating out the random spline coefficients using a Laplace approximation is then equivalent to a penalized spline fit. This in turn implies that Penalized spline fitting is connected to generalized linear mixed models only if the Laplace approximation provides asymptotically exact results. It has been shown in Breslow & Lin (1995), however, as well as more generally in Shun & McCullagh (1995), that Laplace approximation can fail in generalized linear mixed models. For clustered data this occurs if the number of observations within a cluster is small and does not increase with the sample size. It should be noted that the classical literature on generalized linear mixed models assumes that the number of (independent) clusters increases while the number of observations within a cluster is fixed. The asymptotic scenario for penalized spline smoothing is, however, conceptually different. Here, spline coefficients play the role of clusters and the number of spline bases functions is small compared to the sample size n, while the number of observations for each spline is increasing with the sample size. This, however, is exactly the condition in which Laplace approximation

works. We investigate, therefore, how the number of spline coefficients may increase without disturbing the accuracy of the Laplace approximation.

6.2 Generalized P-Spline Smoothing

We consider the generalized smoothing model (6.1) where y for given x is assumed to follow an exponential family distribution with notation

$$y|x \sim \exp\left\{\frac{y\vartheta(x) - b\{\vartheta(x)\}}{\phi} + c(y,\phi)\right\}, \tag{6.2}$$

with $\vartheta(x) = \vartheta\{\eta(x)\}$ as the natural parameter of the underlying exponential family and ϕ as dispersion parameter. Functions $b(\cdot)$ and $c(\cdot)$ are determined by the distribution. For simplicity we ignore the role of the dispersion parameters in (6.2) for the moment and set $\phi \equiv 1$. Functions $\vartheta(x)$ and $\mu(x)$ stand in the unique relationship $b'(\vartheta) = \mu$, so that $\vartheta[\eta(x)] = b'^{-1}\{h[\eta(x)]\}$. Choosing the link function $h(\cdot) = b'(\cdot)$ provides the natural link. For simplicity we assume that x is distributed with density having compact support $[0,1]$ and we observe the independent pairs $(x_i, y_i), i = 1, \ldots, n$. Function $\eta(x)$ is assumed to be smooth in x and we decompose $\eta(x)$ to

$$\eta(x) = X^T(x)\beta + Z^T(x)u + \delta(x), \tag{6.3}$$

where $\delta(x) = \eta(x) - \left\{X^T(x)\beta + Z^T(x)u\right\}$ will be called *approximation bias* subsequently. Matrix $X(x)$ is thereby a low dimensional polynomial basis, e.g. $X^T(x) = (1, x, x^2/2, \ldots, x^q/q!)$, while $Z(x)$ is high dimensional, built for instance from truncated polynomials, i.e.

$$Z^T(x) = \left(\frac{(x-\tau_1)_+^q}{q!}, \ldots, \frac{(x-\tau_{k-1})_+^q}{q!}\right),$$

where $(x)_+^q = x^q$ for $x > 0$ and zero otherwise. With $k-1$ we get the dimension of $Z(x)$ and we assume that the (ordered) knots are placed such that $|\tau_j - \tau_{j-1}| = O(k^{-1})$ with $0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = 1$.
Ignoring the approximation bias $\delta(x)$ we obtain the log likelihood

$$l(\theta) = \sum_{i=1}^{n} y_i \vartheta(P_i^T\theta) - b\left\{\vartheta(P_i^T\theta)\right\}, \tag{6.4}$$

where $P_i^T = P^T(x_i) = (X^T(x_i), Z^T(x_i)) = (X_i^T, Z_i^T)$ and $\theta = (\beta^T, u^T)^T$. Maximizing $l(\theta)$ will lead to a wiggled estimate if the spline dimension k is large. Therefore, a penalty is imposed on θ, that is we consider the penalized likelihood

$$l_p(\theta, \lambda) = l(\theta) - \frac{\lambda k}{2} u^T u, \tag{6.5}$$

where λ is the smoothing or penalty parameter. The penalty in (6.5) can also be written as $\theta^T D_k \theta$, where D_k is a block diagonal with zero entries in the upper left $(q+1)\times(q+1)$ block and identity matrix I_{k-1} in the bottom right block. The explicit listing of k in (6.5) is done for practical reasons and possibly slightly awkward in theoretical terms. We will see that both, λ as well as $u^T u$ depend on k as well, so that explicit listing of k does not seem necessary. Practical experience, however, shows that implicit incorporation of the spline dimension in the penalty in (6.5) accounts for different spline dimensions in finite sample situations, which motivates us to list k explicitly in (6.5). Increasing λ to infinity leads to a purely parametric and, therefore, smooth fit. Decreasing λ induces more complexity into the fit. Clearly, λ has to be chosen data driven in order to minimize the mean squared error.
Instead of truncated polynomials the use of B-splines can be numerically more advisable, even though both approaches are equivalent in the following sense. We define with $P_{q,k}$ the n by $(q+k)$ dimensional truncated spline basis with rows

$$P_i^T = P_{q,k}(x_i) = \left(1, x_i, \frac{x_i^2}{2!}, \ldots, \frac{x_i^q}{q!}, \frac{(x_i - \tau_1)_+^q}{q!}, \ldots, \frac{(x_i - \tau_{k-1})_+^q}{q!}\right),$$

$i = 1, \ldots, n$. From $P_{q,k}$ we can construct the normed B-spline basis $B_{q,k}$ via $B_{q,k} = k^q P_{q,k} L_{q,k}$ where $L_{q,k}$ is a $(q+k)\times(q+k)$ dimensional invertible matrix constructed from the $q+1$ order difference matrix as discussed in Section 2.1.2. The spline representation can now be written as $P_{q,k}\theta = B_{q,k}\omega$ with $\omega = k^{-q} L_{q,k}^{-1}\theta$ as coefficient vector for the B-spline basis. Note that the coefficient vectors θ and ω depend on k as well so that we should write θ_k instead of θ. We suppress however this extra index to simplify the subsequent formulas. We can now formulate the penalized likelihood (6.5) in terms of parameter vector ω leading to

$$l_p(\omega, \lambda) = l(\omega) - \frac{\lambda k^{2q+1}}{2} \omega^T \tilde{D}_k \omega, \tag{6.6}$$

where $\tilde{D}_k = L_{q,k}^T D_k L_{q,k}$. Both, $l_p(.)$ as well as $l(.)$ depend on the sample size n which is not mirrored in our notation for simplicity of presentation. We will now investigate how

the spline dimension k should grow with increasing sample size. In particular, we will derive the asymptotic order $k = O(n^{1/(2q+3)})$, which guarantees that all components in the mean squared error are of the same asymptotic order. Let $P_{q,k}\theta_0 = B_{q,k}\omega_0$ be the best spline approximation of the unknown function $\eta(x)$ based on a Kullback Leibler measure, that is

$$\theta_0 = \text{argmax}\, \mathrm{E}\left\{l(\theta)|\eta\right\}$$

or equivalently $\omega_0 = k^{-1}L_{q,k}^{-1}\theta_0 = \text{argmax}\, \mathrm{E}\{l(\omega)|\eta\}$, where the expectation is calculated with the unknown predictor $\eta(x)$. Accordingly, we define with $\delta_0(x) = \eta(x) - P_{q,k}^T(x)\theta_0 = \eta(x) - B_{q,k}^T(x)\omega_0$ the smallest approximation bias with $B_{q,k}(x)$ as B-spline basis evaluated at x. We can now decompose the Mean Squared Error to

$$\begin{aligned}\mathrm{MSE}\left\{\hat{\eta}(x)\right\} &= \mathrm{E}\left[\left\{\hat{\eta}(x) - \eta(x)\right\}^2\right] \\ &= \mathrm{E}\left[\left\{\hat{\eta}(x) - B_{q,k}^T(x)\omega_0\right\}^2\right] + \delta_0^2(x) + 2\delta_0(x)\mathrm{E}\left\{\hat{\eta}(x) - B_{q,k}^T(x)\omega_0\right\}.\end{aligned}$$

The first component mirrors a conventional mean squared error in penalized parametric regression, while the remaining two components include the approximation bias. The central result of this chapter can now be stated as follows.

Theorem 1 With the assumptions listed in (A1) to (A5) in the Section 6.4 we find that the penalized estimate $\hat{\eta}(x) = B_{q,k}^T(x)\hat{\omega}$ obtained from (6.6) is consistent with the mean squared error of order

$$\mathrm{MSE}\{\hat{\eta}(x)\} = O\left(n^{-\frac{2q+2}{2q+3}}\right).$$

In particular, we can expand the estimate $\hat{\eta}(x)$ as

$$\hat{\eta}(x) - \eta(x) = \left[B_{q,k}^T F(\lambda)^{-1}\left\{\frac{\partial l(\omega)}{\partial \omega} - \lambda k^{2q+1}\tilde{D}_k\omega\right\} - \delta(x)\right]\left\{1 + o_p(1)\right\}, \tag{6.7}$$

with $F(\lambda) = \mathrm{E}\left(-\partial^2 l(\omega)/\partial\omega\partial\omega^T + \lambda k^{2q+1}\tilde{D}_k\right)$. The leading stochastic component in (6.7) has the same asymptotic order $O_p\left(n^{-\frac{1}{2}\frac{2q+2}{2q+3}}\right)$.

Based on (6.7) the central limit theorem applies in the form

$$\hat{\eta}(x) - \eta(x) \overset{a}{\sim} \mathcal{N}\left\{\mathrm{bias}\left(\hat{\eta}(x)\right), \mathrm{Var}\left(\hat{\eta}(x)\right)\right\}, \tag{6.8}$$

with

$$\text{bias}(\hat{\eta}(x)) = B_{q,k}^T(x)F(\lambda)^{-1}\lambda k^{2q+1}D_k\omega_0 - \delta(x)$$

and

$$\text{Var}(\hat{\eta}(x)) = B_{q,k}^T(x)F^{-1}(\lambda)F(\lambda=0)F^{-1}(\lambda)B_{q,k}(x). \tag{6.9}$$

The proof of the theorem is provided in Section 6.4.

Remarks

1. One of the central theoretical questions for P-spline smoothing is how fast should the spline dimension grow with the sample size. It is shown in the Section 6.4 that $k = O(n^{1/(2q+3)})$ is a recommended choice with respect to the above mean squared error criterion. With truncated quadratic functions or B-splines of second order, we get $k = O(n^{1/7})$, so that the spline basis grows clearly at a slower rate than the sample size. This motivates one to choose $k \ll n$ in practice.

2. As demonstrated in the Section 6.4, the mean squared error decomposes into two parts. The first is the mean squared error if we assume that the unknown function $\eta(x)$ is in fact representable by the parametric shape $X(x)^T\beta + Z(x)^Tu$. In this case, the penalty parameter λ has to be chosen such that it balances the variability and the squared smoothing bias. This is achieved if we set $\lambda = O(n^\kappa)$ with $\kappa = 1/(2q+3)$, assuming that the parametric model is correct.
   The remaining terms in the mean squared error are driven by the approximation bias $\delta(x)$ which is with Taylor approximation of order $\delta(x) = O(k^{-(q+1)})$. It results then, that both components in the mean squared error have the same asymptotic order if k grows with order $O(n^{1/(2q+3)})$. This means that the practical guideline for P-spline smoothing to choose a number of knots, such that the approximation bias $\delta(x)$ can be ignored compared to the variability of the estimates, does not have an asymptotic justification. In fact, the approximation bias is playing a non-ignorable role, at least in asymptotic terms.

3. The variance of $\hat{\eta}(x)$ is built in a sandwich form from Fisher type matrices. Due to the fact that the dimension k of ω grows with the sample size, the dimension of the Fisher matrix grows as well. It is shown in Section 6.4 that for B-Splines $F(\lambda)$ is a band diagonal matrix of order q resulting from properties of B-splines.

Moreover, the elements in the bands of the matrix grow with order $O(n/k)$, as long as $\lambda = O(n^{1/(2q+3)})$, as postulated. This motivates that the sandwich type variance in (6.9) is decreasing to zero with order $O(k/n)$.

4. The obtained optimal rate of convergence (as well as the order of the spline basis dimension k) corresponds exactly the results reported in Agarwal & Studden (1980) for regression splines under normality.

6.3 P-Spline Smoothing and Mixed Models

6.3.1 Laplace Approximation

P-spline smoothing can be linked to mixed models by comprehending the penalty as "a priori" normal distribution on the spline coefficients. The penalized estimate is then asymptotically equivalent to the posterior Bayes estimate resulting in the mixed model. This equivalence holds exactly in the normal response model with identity link and normal distribution imposed on the spline coefficients. In this case, the marginal likelihood is available analytically by integrating out the spline coefficients. The smoothing parameter λ now plays the role of the ratio of the residual variance and the "a priori" variance of the spline coefficients. Consequently, based on the mixed model, the smoothing parameter can be estimated by maximizing the marginal likelihood, or an adjusted version of it, yielding a restricted maximum likelihood estimate (REML). This is a practical benefit, since smoothing parameter selection can now be carried out by maximum likelihood estimation adopting the mixed model approach. For generalized response models, however, integration over the spline coefficients is not available analytically and alternative methods have to be used. The link to penalized spline estimation results by pursuing a Laplace approximation. The latter is justified asymptotically only, if the remaining correction terms converge to zero with growing sample size. In the following section we want to focus on the question whether the Laplace approximation is justified when we assume the spline dimension to grow with the previously proposed order $k = O(n^{1/(2q+3)})$.
We now model spline coefficient vector u as a priori normally distributed. Moreover, we assume in this section for the sake of simplicity that link function $h(.)$ is the canonical link. This leads to the generalized linear mixed model

$$E(y|u) = h(X\beta + Zu), \quad u \sim N\left(0, \frac{\sigma_u^2}{k} I_{k-1}\right), \tag{6.10}$$

with $y = (y_1, \ldots, y_n)^T$ and X and Z as defined above. Integrating out the random spline effects leads to the marginal likelihood (up to a constant)

$$L(\beta, \sigma_u^2) = \sigma_u^{-(k-1)} \int_{R^{k-1}} \exp[-g(u)]du, \tag{6.11}$$

with $g(u) = -y^T(X\beta + Zu) + 1_n^T b(X\beta + Zu) + ku^T u/2\sigma_u^2$, with $1_n = (1, \ldots, 1)^T$, The integral in (6.11) does not generally have an analytic solution. We, therefore, make use of a Laplace approximation to obtain the marginal likelihood (6.11). This means we expand $g(u)$ around its minimum. Note that $g(u) = g(\beta, u, \sigma_u^2)$, that is $g(u)$ depends as well on other quantities which are omitted in (6.11). It is not difficult to see that $\partial g(\hat{\beta}, \hat{u}, \sigma_u^2)/\partial(\beta, u) = 0$ defines the penalized estimating equation $\partial l_p(\theta, \lambda)/\partial\theta = 0$ with $l_p(\theta, \lambda)$ as defined in (6.5) and $\lambda = \sigma_u^{-2}$ playing the role of the penalization parameter. Instead of deriving a Laplace approximation for the integral (6.11) directly, we use a B-spline formulation for technical reasons. Let, therefore, the difference matrix $L_{q,k}$ from above be decomposed as

$$L_{q,k} = \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix},$$

according to the dimension of β and u, i.e. $L_{11} \in \mathbb{R}^{(q+1)\times(q+1)}$. Since the elements of L_{12} are all equal to zero it is easy to see that $P_{k,q}\theta = B_{q,k}\omega$ can be represented as

$$X\beta + Zu = k^q(XL_{11} + ZL_{21})\omega_1 + k^q Z L_{22}\omega_2 = B_{q,k,1}\omega_1 + B_{q,k,2}\omega_2, \tag{6.12}$$

with $\omega_1 := k^{-q}L_{11}^{-1}\beta$ and $\omega_2 := k^{-q}L_{22}^{-1}(u - L_{21}L_{11}^{-1}\beta)$. In this notation the integral (6.11) takes the form

$$L(\beta, \sigma_u^2) = \sigma_u^{-(k-1)} k^{(k-1)q} |L_{22}| \int_{R^{k-1}} \exp[-\tilde{g}(\omega_2)]d\omega_2, \tag{6.13}$$

where $\tilde{g}(\omega_2) := \tilde{g}(\omega_1, \omega_2) = g(\theta(\omega)) = g(\beta, u)$. The integral in (6.13) is approximated using a Laplace approximation by

$$\int_{R^{k-1}} \exp[-\tilde{g}(\omega_2)]d\omega_2 = |\tilde{G}|^{-1/2}(2\pi)^{\frac{(k-1)}{2}} \exp[-\tilde{g}(\hat{\omega}_2)] \left\{1 + O(\varepsilon_0)\right\}, \tag{6.14}$$

where $\tilde{G} = \tilde{G}(\hat{\omega}_2)$ denotes the second order derivative $\partial^2\tilde{g}(\hat{\omega}_2)/\partial\omega_2\partial\omega_2^T$, evaluated at $\hat{\omega}_2$ which minimizes $\tilde{g}(.)$. The objective is now to evaluate the asymptotic order of the correction term ε_0. Let $\hat{\tilde{g}}_{jl}$ denote the (j, l)-th element of $\tilde{G}$. Accordingly, third and

forth order derivatives of $\tilde{g}(\cdot)$ are denoted by $\hat{\tilde{g}}_{jlr}$ and $\hat{\tilde{g}}_{jlrs}$, respectively. Moreover, with $\hat{\tilde{g}}^{jl}$ we refer to the (i,j)-th element of the inverse of $\tilde{G}$. Following the results provided in Barndorff-Nielsen & Cox (1989) or Shun & McCullagh (1995) and using Einstein's summation convention we can then write the correction term in (6.14) as

$$\begin{aligned} \varepsilon_0 &= -\hat{\tilde{g}}_{jlrs}\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{rs}[3]/24 \\ &\quad +\hat{\tilde{g}}_{jlr}\hat{\tilde{g}}_{stv}\left(\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{rs}\hat{\tilde{g}}^{tv}[9]+\hat{\tilde{g}}^{js}\hat{\tilde{g}}^{lt}\hat{\tilde{g}}^{rv}[6]\right)/72. \end{aligned} \tag{6.15}$$

In (6.15) equal super and subscript imply a summation over the corresponding indices and the bracketed terms denote possible permutations over the indices, e.g. the first component in (6.15) is a short form for $1/24\hat{\tilde{g}}_{jlrs}\left(\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{rs}+\ \hat{\tilde{g}}^{jr}\hat{\tilde{g}}^{ls}+\hat{\tilde{g}}^{js}\hat{\tilde{g}}^{rl}\right)$. The objective is now to show that the term ε_0 vanishes asymptotically with the sample size n increasing. Note that

$$\hat{\tilde{g}}_{jl} = B^T_{q,k,j}WB_{q,k,l} + \frac{k^{2q+1}}{\sigma_u^2}(L_{22}^TL_{22})_{jl}, \tag{6.16}$$

with $B_{q,k,l}$ denoting the l-th column of $B_{q,k,2}$ defined in (6.12), W as diagonal matrix with variance elements $\mathrm{Var}(y_i)$, $i=1,...,n$ on the diagonal and $(L_{22}^TL_{22})_{jl}$ as (j,l)-th element of $L_{22}^TL_{22}$. Assuming the knots $\tau_j, j=0,\ldots,k$, to be distributed according to quantiles of covariate x (see assumptions (A1) and (A2) in the Section 6.4) we obtain that the number of non zero elements for each column of spline basis $B_{q,k}$ is of order $O(n/k)$. Consequently, the first element in (6.16) equals 0 if $|j-l|>q$ and is of order $O(n/k)$ otherwise. Considering the definition of $L_{22}^TL_{22}$ we find that the second component in (6.16) takes values 0 if $|j-l|>q+1$ and has order $O(k^{2q+1}/\sigma_u^2)$ otherwise. Hence,

$$\hat{\tilde{g}}_{jl} = \begin{cases} O\left(n/k+k^{2q+1}/\sigma_n^2\right), & |j-l|\le q+1 \\ 0, & \text{otherwise.} \end{cases}$$

In the same way we obtain

$$\hat{\tilde{g}}_{jlr} = \begin{cases} O\left(n/k\right), & |j-l|\le q \text{ and } |j-r|\le q \text{ and } |l-r|\le q \\ 0, & \text{otherwise} \end{cases}$$

and accordingly $\hat{\tilde{g}}_{jlrs} = O(n/k)$ if the maximum of absolute pairwise differences of the subscript indices is smaller than or equal to q, and otherwise zero.
This implies that ε_0 has the order

$$\varepsilon_0 = O\left\{n\left(\frac{n}{k}+\frac{k^{2q+1}}{\sigma_u^2}\right)^{-2}\right\} + O\left\{n^2\left(\frac{n}{k}+\frac{k^{2q+1}}{\sigma_u^2}\right)^{-3}\right\}. \tag{6.17}$$

Letting now k grow with order $O\left(n^{1/(2q+3)}\right)$, see (6.25), allows one to rewrite the asymptotic order of (6.17) to

$$O\left\{n^{-\frac{2q}{2q+3}}\left(1+n^{-\frac{1}{2q+3}}\sigma_u^{-2}\right)^{-3}\right\},$$

which decreases to zero if we set $\sigma_u^2 = O(n^{-1/(2q+3)})$. The latter is formulated in condition (A5) in the Section 6.4 and was postulated in the same way in the previous section as condition imposed on smoothing parameter λ. As a result we find that the Laplace approximation is justified asymptotically and the correction term ε_0 may be omitted even for a growing dimension of the spline basis. It remains, however, to show that the assumption imposed on σ_u^2 is sound, which will be discussed in detail in Section 6.3.3 below. The validity of the Laplace approximation was derived for the B-spline basis. Due to the equivalence of B-splines and truncated polynomials, the result transfers to the original formulation (6.10) with truncated polynomials as well. The latter is formulated in the following theorem:

Theorem 2 With assumptions (A1), (A2), (A3) and (A5) we find that the marginal likelihood function of the General Linear Mixed Model (6.10) can be approximated using Laplace approximation, that is

$$L(\beta,\sigma_u^2) = \left[\sigma_u^{-(k-1)}|G|^{-1/2}\exp\left\{-g(\beta,\hat{u},\sigma_u^2)\right\}\right]\left\{1+o(1)\right\},$$

with $g(\hat{u},\beta,\sigma_u^2) = -y^T(X\beta+Z\hat{u})+1_n^Tb(X\beta+Z\hat{u})+k\hat{u}^T\hat{u}/2\sigma_u^2$ where $y=(y_1,\ldots,y_n)^T$ and $\hat{u}$ as minimizer of $g(\beta,u,\sigma_u^2)$. Matrix G is defined through $\partial^2 g(\beta,\hat{u},\sigma_u^2)/\partial u\partial u^T = Z^TWZ + kI_{k-1}/\sigma_u^2$.

Remarks

1. Using the normal assumption for u, we get with $u \sim N(0,\sigma_u^2/kI_{k-1})$ and assumptions (A5) $\sigma_u^2 = O(n^{-1/(2q+3)})$ and (A3) $k = O(n^{1/(2q+3)})$ that

$$u^Tu = O_p(\sigma_u^2) = O_p(k^{-1}). \tag{6.18}$$

   This is the stochastic formulation of condition (6.27). Hence, in Theorem 2 we omitted assumption (A4) but imposed a normality on coefficients u which induces (A6.18) as stochastic version of assumption (A4).

2. Shun & McCullagh (1995) showed that the Laplace approximation of a likelihood for some k dimensional parameter based on the n data points (from exponential

family) is reliable provided that $k = o(n^{1/3})$. This is clearly satisfied for our choice $k = O(n^{1/(2q+3)})$, given $q > 0$.

6.3.2 Posterior Cumulants

We will now generally look at posterior cumulants of the spline coefficients based on the generalized mixed model (6.10). Based on (6.11), the corresponding moment generating function is defined through

$$M_{\hat{\omega}_2|y}(t) = \frac{\int_{R^{k-1}} \exp(t^T\omega_2)\exp[-\tilde{g}(\omega_2)]d\omega_2}{\int_{R^{k-1}} \exp[-\tilde{g}(\omega_2)]d\omega_2}. \tag{6.19}$$

Following the results from above, the denominator in (6.19) can be approximated by (6.14). Applying Laplace approximation in the same style to the nominator of (6.19) (see Barndorff-Nielsen & Cox, 1989) we obtain

$$(2\pi)^{(k-1)/2}|\tilde{G}|^{-1/2}\exp[-\tilde{g}(\hat{\omega}_2)]\left(M_z(t) + \exp(t^T\hat{\omega}_2)O\left\{\varepsilon_0 + \varepsilon_1(t)\right\}\right),$$

where $M_z(t)$ denotes the moment generating function of the normally distributed random variable $z \sim N(\hat{\omega}_2, \tilde{V})$ with $\tilde{V} = \tilde{G}^{-1}$ and $\varepsilon_1(t) = \hat{\tilde{g}}_{jlr}t_s\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{rs}[3]/6$. Note that ε_0 and $\varepsilon_1(t)$ are of the same asymptotic order for a given fixed t. With this we get

$$M_{\hat{\omega}_2|y}(t) = M_z(t)\frac{1 + \exp(-t^T\tilde{V}t/2)O\left\{\varepsilon_0 + \varepsilon_1(t)\right\}}{1 + O(\varepsilon_0)}.$$

The corresponding cumulant generating function can be written as

$$K_{\hat{\omega}_2|y}(t) = K_z(t) + \tilde{H}(t) + O(\varepsilon_0),$$

with $\tilde{H}(t) = \exp(-t^T\tilde{V}t/2)O\left\{\varepsilon_0 + \varepsilon_1(t)\right\}$. This shows that the sth derivative of the $\tilde{H}(t)$ with respect to $t_{i_1}, ..., t_{i_s}$ evaluated at $t = 0$ defines the difference of the sth posterior cumulant of the $\hat{\omega}_2$ to the corresponding cumulant of the normally distributed random variable z. Since the derivatives of $O\left\{\varepsilon_0 + \varepsilon_1(t)\right\}$ are negligible for $s > 2$, we obtain

$$\begin{aligned}\tilde{H}_{i_1,...,i_s}(t) &= [\exp(-t^T\tilde{V}t/2)]_{i_1,...,i_s}O\left\{\varepsilon_0 + \varepsilon_1(t)\right\} \\ &+ s[\exp(-t^T\tilde{V}t/2)]_{i_1,...,i_{s-1}}O(\varepsilon_{i_s}),\end{aligned}$$

with $\varepsilon_{i_s} = \hat{\tilde{g}}_{jlr}\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{ri_s}[3]/6$. From standard results on multivariate normal distribution we find

$$[\exp(-t^T\tilde{V}t/2)]_{i_1,\ldots,i_s} = \tilde{h}_{i_1,\ldots,i_s}\exp(-t^T\tilde{V}t/2),$$

where $\tilde{h}_{i_1,\ldots,i_s}$ are known as Hermite tensor (more details can be found in McCullagh, 1987, pages 149-151). The first six are given as follows

$$\begin{aligned}
\tilde{h}_i &= \tilde{v}_{ij}t^j \\
\tilde{h}_{ij} &= \tilde{h}_i\tilde{h}_j - \tilde{v}_{ij} \\
\tilde{h}_{ijk} &= \tilde{h}_i\tilde{h}_j\tilde{h}_k - \tilde{h}_i\tilde{v}_{ij}[3] \\
\tilde{h}_{ijkl} &= \tilde{h}_i\tilde{h}_j\tilde{h}_k\tilde{h}_l - \tilde{h}_i\tilde{h}_j\tilde{v}_{kl}[6] + \tilde{v}_{ij}\tilde{v}_{kl}[3] \\
\tilde{h}_{ijklm} &= \tilde{h}_i\tilde{h}_j\tilde{h}_k\tilde{h}_l\tilde{h}_m - \tilde{h}_i\tilde{h}_j\tilde{h}_k\tilde{v}_{lm}[10] + \tilde{h}_i\tilde{v}_{jk}\tilde{v}_{lm}[15] \\
\tilde{h}_{ijklmn} &= \tilde{h}_i\tilde{h}_j\tilde{h}_k\tilde{h}_l\tilde{h}_m\tilde{h}_n - \tilde{h}_i\tilde{h}_j\tilde{h}_k\tilde{h}_l\tilde{v}_{mn}[15] + \tilde{h}_i\tilde{h}_j\tilde{v}_{kl}\tilde{v}_{mn}[45] - \tilde{v}_{ij}\tilde{v}_{kl}\tilde{v}_{mn}[15],
\end{aligned}$$

where $\tilde{v}_{ij}$ denotes the ij element of the matrix $\tilde{V} = \tilde{G}^{-1}$. The general pattern is as follows: the summation involves unit blocks and double blocks with $\tilde{v}_{ij}$, blocks having three or more elements are ignored.
We are now interested in $\tilde{H}_{i_1,\ldots,i_s}(t=0)$. Since $\tilde{h}_i(t=0) = 0$ we immediately obtain

$$\begin{aligned}
\tilde{H}_i(0) &= O(\varepsilon_i) \\
\tilde{H}_{ij}(0) &= O(-\tilde{v}_{ij}\varepsilon_0) \\
\tilde{H}_{ijk}(0) &= O(-3\tilde{v}_{ij}\varepsilon_k) \\
\tilde{H}_{ijkl}(0) &= O(\tilde{v}_{ij}\tilde{v}_{kl}[3]\varepsilon_0) \\
\tilde{H}_{ijklm}(0) &= O(5\tilde{v}_{ij}\tilde{v}_{kl}[3]\varepsilon_m) \\
\tilde{H}_{ijklmn}(0) &= O(-\tilde{v}_{ij}\tilde{v}_{kl}\tilde{v}_{mn}[15]\varepsilon_0).
\end{aligned}$$

This in turn provides

$$\begin{aligned}
E(\hat{\omega}_{2,i}|y) &= \hat{\omega}_{2,i} + O(\varepsilon_i) \\
\mathrm{Cov}(\hat{\omega}_{2,i}, \hat{\omega}_{2,j}|y) &= \tilde{v}_{ij}\left\{1 + O(\varepsilon_0)\right\}.
\end{aligned}$$

Noting that the ij element of matrix $\tilde{V} = \tilde{G}^{-1}$ is of order $O(k/n) = O(n^{-(2q+2)/(2q+3)})$, provided $\sigma_u^2 = O(n^{-1/(2q+3)})$, we easily find the convergence rate of the higher order cumulants to zero.
Due to the definition of ω_2 the above results transfer directly to the cumulants of $\hat{u}|y$.

6.3.3 Maximum Likelihood Estimation

Above we have assumed that the *a priori* variance σ_u^2 converges to zero with order $O(n^{-1/(2q+3)})$. We will now demonstrate that this rate of convergence is sound. The mixed model formulation of penalized spline smoothing is commonly used to allow for a simple estimation of the smoothing parameter utilizing mixed models technology. In particular, since the smoothing parameter λ plays the role of $1/\sigma_u^2$ in the mixed model formulation, we can use maximum likelihood estimation for σ_u^2 to obtain an estimate for λ. For fixed spline dimension in the linear mixed model with normal response this has been investigated theoretically in Kauermann (2004). Here, we let the spline dimension grow with the sample size and focus on asymptotic rates. Instead of maximum likelihood (ML) estimation, the use of restricted maximum likelihood (REML) is more common in mixed models. The latter makes a small sample adjustment for the estimation of the unpenalized parameters β. The practical difference is usually minor (see Ruppert, Wand & Carroll, 2003), in particular if the dimension of the spline basis grows with increasing sample size. In this case, the difference between REML and ML becomes negligible since the dimension of β is fixed and therewith ignorable compared to the growing dimension of the spline basis. For simplicity we consider here therefore ML estimation of σ_u^2 only. The leading term in the Laplace approximated log likelihood is written as

$$l(\beta, \sigma_u^2) \approx -\frac{1}{2}\log|G| - g(\hat{u}) - \frac{k-1}{2}\log\sigma_u^2, \tag{6.20}$$

with $G = Z^T W Z + kI_{k-1}/\sigma_u^2$. Inserting the estimate for β and differentiating (6.20) with respect to σ_u^2 yields

$$\begin{aligned} \frac{\partial l(\hat{\beta}, \sigma_u^2)}{\partial \sigma_u^2} &= -\frac{1}{2}\mathrm{tr}\left(G^{-1}\frac{\partial G}{\partial \sigma_u^2}\right) + \frac{k\hat{u}^T\hat{u}}{2\sigma_u^4} - \frac{k-1}{2\sigma_u^2} && (6.21) \\ &= \frac{1}{2\sigma_u^2}\left\{\frac{k\hat{u}^T\hat{u}}{\sigma_u^2} - df(\sigma_u^2)\right\}, && (6.22) \end{aligned}$$

where $df(\sigma_u^2) = \mathrm{tr}\left\{G^{-1}Z^T W Z\right\}$. We get from (6.21) to (6.22) by reflecting the definition of G and using the fact that $\mathrm{tr}(G^{-1}G) = k-1$. The estimate is now defined through

$$\hat{\sigma}_u^2 = k\frac{\hat{\theta}^T D_k \hat{\theta}}{df(\sigma_u^2)} = k^{2q+1}\frac{\hat{\omega}^T \tilde{D}_k \hat{\omega}}{df(\sigma_u^2)}. \tag{6.23}$$

It should be remarked that (6.23) is not an analytic formula, since the right hand side of the equation contains the unknown parameter as well. For our analytic investigation we

can, however, make use of (6.23) by treating σ_u^2 on the right hand side as true "a priori" variance. We now show, that the estimate $\hat{\sigma}_u^2$ is efficient if $\sigma_u^2 = O(n^{-1/(2q+3)})$. It is not difficult to show that $\mathrm{E}(\hat{\sigma}_u^2) = \sigma_u^2$, so that we investigate the variance, expressed here as Fisher matrix. Note that

$$\begin{aligned} \frac{\partial^2 l(\beta, \sigma_u^2)}{\partial \sigma_u^2 \partial \sigma_u^2} &= \left[\frac{1}{2\sigma_u^4} - \frac{1}{2\sigma_u^2}\right] \frac{\partial l(\beta, \sigma_u^2)}{\partial \sigma_u^2} \\ &+ \frac{1}{2\sigma_u^4}\left[\mathrm{tr}(G^{-1}Z^TWZG^{-1}Z^TWZ) - \frac{2k}{\sigma_u^2}\hat{u}^TG^{-1}Z^TWZ\hat{u}\right] \end{aligned} \tag{6.24}$$

Our intention is to show that the $\mathrm{Var}(\hat{\sigma}_u^2)$ is decreasing to zero if σ_u^2 has the above listed order. This guarantees efficiency of the ML estimate. To do so we look at the Fisher information. Note that the first component in (6.24) has expectation zero and $E(k\hat{u}^TG^{-1}Z^TWZ\hat{u}/\sigma_u^2) = \mathrm{tr}(G^{-1}Z^TWZG^{-1}Z^TWZ)$. Using the relationship $k^qZL_{22} = B_{q,k,2}$ as defined in (6.12) we can represent

$$\mathrm{tr}(G^{-1}Z^TWZG^{-1}Z^TWZ) = \mathrm{tr}(\tilde{G}^{-1}B_{q,k,2}^TWB_{q,k,2}\tilde{G}^{-1}B_{q,k,2}^TWB_{q,k,2}),$$

with $\tilde{G}$ defined in (6.16) and find that the expected second component in (6.24) has the order

$$\begin{aligned} \mathrm{E}\left\{-\frac{\partial^2 l(\beta, \sigma_u^2)}{\partial \sigma_u^2 \partial \sigma_u^2}\right\} &= O\left\{\frac{n^2}{\sigma_u^4}\left(\frac{n}{k} + \frac{k^{2q+1}}{\sigma_u^2}\right)^{-2}\right\} \\ &= O\left\{\sigma_u^{-4}n^{\frac{2}{2q+3}}\left(1 + \sigma_u^{-2}n^{-\frac{1}{2q+3}}\right)^{-2}\right\}. \end{aligned}$$

This yields to

$$\mathrm{Var}(\hat{\sigma}_u^2) = O\left\{\sigma_u^4 n^{-\frac{2}{2q+3}}\left(1 + \sigma_u^{-2}n^{-\frac{1}{2q+3}}\right)^2\right\},$$

which is minimized if σ_u^2 is of order $O(n^{-1/(2q+3)})$.

6.4 Asymptotic Behavior

Proof of Theorem 1

The subsequent derivations relate to Cardot (2002) and Agarwal & Studden (1980). Unlike these two authors we do not restrict the proof to normal response model. Also, in contrast to Cardot we impose the penalty directly on the spline coefficients. This is advantageous since we can link the results described below to mixed model theory. Our

asymptotic scenario is built on the following assumptions.

(A1) We assume that design points $x_i \in [0,1]$ become dense with order n, such that two adjacent values x_i and x_j converge to zero with order $O(n^{-1})$. In other words, x_i are drawn independently from density $f(x)$ having the compact support $[0,1]$.

(A2) The knots for the spline basis are placed according to quantiles of the distribution of x and it is assumed that $0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = 1$ with $\tau_j - \tau_{j-1} = O(k^{-1})$ for $j = 1, \ldots, k$.

(A3) We assume that the dimension of the spline basis grows with the sample size with order

$$k = O(n^{\frac{1}{2q+3}}). \tag{6.25}$$

(A4) Function $\eta(x)$ is assumed to be $q+1$ times continuously differentiable and the penalty on coefficient u relating to the truncated polynomial basis (see 6.3) fulfills

$$\|u\|_\infty = \max_i |u_i| = O(k^{-1}). \tag{6.26}$$

In particular this yields

$$u^T u = O(k^{-1}). \tag{6.27}$$

Moreover, $\eta(x)$ is bounded so that $\mu(x) = h(\eta(x))$ is in the interior of the mean parameter space for all x.

(A5) The penalty parameter λ is assumed to grow with sample size with order

$$\lambda = O(n^{\frac{1}{2q+3}}). \tag{6.28}$$

Setting $\sigma_u^2 = 1/\lambda$ as used in Section 3, the order (6.28) is also formulated as $\sigma_u^2 = O(n^{-1/(2q+3)})$.

Remarks

1. The order (6.25) will be derived in the proof. It turns out that with the proposed choice the mean squared error is asymptotically optimal. The equivalence between truncated polynomials and B-splines leads to the formulation $P_{q,k}\theta = B_{q,k}\omega$ with $\omega = k^{-q} L_{q,k}^{-1}\theta$. Writing the penalty in $\theta = (\beta^T, u^T)^T$ in the form $\theta^T D_k \theta$ with

$D_k = \text{diag}(0_{q+1}, I_{k-1})$, where 0_{q+1} is the zero matrix and I_{k-1} the diagonal matrix with dimensions given as subscript, yields the equivalent penalty for ω in the form

$$\theta^T D_k \theta = k^{2q} \omega^T \tilde{D}_k \omega$$

where $\tilde{D}_k = L_{q,k}^T D_k L_{q,k}$.

2. The asymptotic order of the penalty (6.27) can be motivated as follows. Note that for x as inner point in [0, 1] and some $\nu > 0$ we have

$$\begin{aligned} P_{q,k}(x+\nu)\theta &- P_{q,k}(x)\theta \\ &= O(\nu) \left\{ \sum_{r=1}^{q} \frac{x^{r-1}}{(r-1)!} \beta_r + \sum_{l:\tau_l \le x} \frac{(x-\tau_l)^{q-1}}{(q-1)!} u_l \right\} \{1 + O(\nu)\} \\ &\quad + \sum_{l:x<\tau_l \le x+\nu} (x-\tau_l)^q u_l / q! \qquad (6.29) \\ &\le O(\nu) \left\{ O(1) + O(k) \cdot O(\|u\|_\infty) \right\} \{1 + O(\nu)\}, \qquad (6.30) \end{aligned}$$

Differentiability in the limit as k tends to infinity is achieved only if we postulate $O(\|u\|_\infty) = O(k^{-1})$. This implies $\theta^T D_k \theta = O(k^{-1})$. Writing $P_{q,k}\theta$ with penalty $\theta^T D_k \theta$ as B-spline formulation yields

$$k\theta^T D_k \theta = k^{2q+1} \omega_k^T \tilde{D} \omega = O(1)$$

For the proof of Theorem 1 we use the following notation. Let $l(\vartheta) = \sum_{i=1}^n y_i^T \vartheta(x_i) - b(\vartheta(x_i)) + c(y_i, 1)$ define the log-likelihood function and denote the derivative with respect to the vector $\vartheta = (\vartheta(x_1), \ldots, \vartheta(x_n))$ as $l_\vartheta(\vartheta) := \partial l(\vartheta)/\partial\vartheta = \big(y_i - \mu[\vartheta(x_i)]\big)_{i=1,\ldots,n}$. Accordingly we write $l_\eta(\eta)$ for the n dimensional column vector

$$l_\eta(\eta) := \frac{\partial \vartheta^T}{\partial \eta} \cdot l_\vartheta(\vartheta) = \left(\frac{\partial \vartheta(x_i)}{\partial \eta(x_i)} \big\{ y_i - \mu\left[\vartheta(x_i)\right] \big\} \right)_{i=1,\ldots,n} \qquad (6.31)$$

Let $\eta_0 = B_{q,k}\omega_0$, where ω_0 is the best coefficient in the sense that ω_0 minimizes the Kullback-Leibler distance, that is $\mathrm{E}\{B_{q,k}^T, l_\eta\left[\vartheta(B_{q,k}\omega_0)\right]\} = 0$, where the expectation is carried out with respect to the true function $\eta(x)$. Coefficient ω_0 defines the optimal approximation bias $\delta(x)$ in (6.3) through

$$\delta_0(x) = \eta(x) - B_{q,k}(x)\omega_0 \qquad (6.32)$$

with $B_{q,k}(x)$ as spline basis evaluated at x. The proof of the theorem follows now by decomposing

$$\begin{aligned} \mathrm{E}\left\{(\hat{\eta}(x)-\eta(x))^2\right\} &= \underbrace{\mathrm{E}\left\{(\hat{\eta}(x)-\eta_0(x))^2\right\}}_{1} \\ &\quad + \underbrace{\delta_0^2(x)}_{2} + \underbrace{2\mathrm{E}\left(\hat{\eta}(x)-\eta_0(x)\right)\delta_0(x)}_{3}. \end{aligned} \tag{6.33}$$

We gradually consider the separate components in (6.33).
We show first convergence of $\hat{\omega}$ to ω_0. Note that the penalized estimating equation for $\hat{\omega}$ results in

$$0 = B_{q,k}^T\, l_\eta\left[\vartheta(B_{q,k}\hat{\omega})\right] - \lambda\, k^{2q+1}\, \tilde{D}_k\hat{\omega}. \tag{6.34}$$

The subsequent proof will make use of Einstein's summation convention (see McCullagh, 1987 or Barndorff-Nielsen & Cox, 1989). This allows one to consider higher dimensional arrays, beyond vectors and matrices. To apply the technique we need some additional notation. Let the j-th component of vector ω be denoted with a super, instead of subscript, that is $\omega = (\omega^1, \omega^2, \ldots, \omega^{k+q})$. With $0 = l_{p,j}(\hat{\omega})$ we denote the j-th component of equation (6.34), that is $l_{p,j}(\hat{\omega}) = \partial l_p(\omega)/\partial\omega^j|_{\omega=\hat{\omega}}$ with $l_p(\cdot)$ as defined in (6.6). The objective is now to expand $l_{p,j}(\hat{\omega})$ around $l_{p,j}(\omega_0)$. We use the convention of omitting the explicit listing of parameters if the best coefficient ω_0 is used, that is $l_{p,j} = l_{p,j}(\omega_0)$. Moreover, the hat notation $\hat{l}_{p,j}$ is used for $l_{p,j}(\hat{\omega})$. Finally, higher order derivatives are notated by multiple subscripts, e.g. $l_{p,jl} = \partial^2 l_p(\omega_0)/\partial\omega^j\partial\omega^l$. We are now able to expand $\hat{l}_{p,j}$ around $l_{p,j}$. Using the Einstein summation convention implies that equal sub and superscripts are being summed over. This allows one to write the expansion as

$$0 = \hat{l}_{p,j} = l_{p,j} + l_{p,jl}(\hat{\omega}^l - \omega_0^l) + \frac{1}{2}l_{p,jlr}(\hat{\omega}^l - \omega_0^l)(\hat{\omega}^r - \omega_0^r) + \ldots \tag{6.35}$$

Solving (6.35) for $\hat{\omega}^l - \omega_0^l$ is done with series inversion (see Barndorff-Nielsen & Cox, 1989), and we get

$$\hat{\omega}^j - \omega_0^j = -l_p^{jl}l_{p,l} - \frac{1}{2}l_p^{jlr}l_{p,l}l_{p,r} + \ldots \tag{6.36}$$

with l_p^{jl} as (j,l)-th element of the matrix inverse of $l_{p,jl}, j,l = 1,\ldots,q+k$, and $l_p^{jlr} = l_p^{js}l_p^{lt}l_p^{ru}l_{p,stu}$. The remaining components not explicitly listed in (6.36) are of lower asymptotic order and, therefore, omitted. In the style of classical Maximum Likelihood Theory

(see McCullagh, 1987) we simplify (6.36) using the following arguments. First, we decompose $l_{p,jl} = w_{jl} + \lambda k^{2q+1} \tilde{D}_{jl} + s_{jl}$ where w_{jl} is the weight or Fisher matrix contribution $-\mathrm{E}(\partial^2 l(\omega_0)/\partial\omega^j \partial\omega^l)$ and s_{jl} is the stochastic component of the second order derivative without the penalty, i.e. $s_{jl} = l_{jl} - w_{jl}$. Finally, $\tilde{D}_{jl}$ is the (j,l)-th element of $\tilde{D}$. The technical idea of our proof is now to look at the form of matrix w_{jl}. Note that $w_{jl}, j, l = 1, \ldots, k+q$, written as matrix takes the form $B_{q,k}^T W B_{q,k}$ with W as weight matrix, which in case of canonical link simplifies to diagonal matrix with diagonal elements $\mathrm{Var}(y_i), i = 1, \ldots, n$. Using properties of B-splines we find that $B_{q,k}^T W B_{q,k}$ is a band diagonal matrix with bandwidth $q+1$ and with elements increasing with order n/k. Moreover $\tilde{D}_k$ is also a band diagonal matrix with bandwidth $q+2$ and elements of order $O(1)$. Hence, matrix $B_{q,k}^T W B_{q,k} + \lambda k^{2q+1} \tilde{D}_k$ is band-diagonal with elements of order $O(n/k + \lambda k^{2q+1})$. Similarly, writing s_{jl} as matrix yields $B_{q,k}^T S B_{q,k}$ with S as diagonal matrix, where the diagonal elements are stochastically independent random variables each having order $O_p(1)$. Hence, matrix s_{jl} is block diagonal as well, with elements of order $O_p\{(n/k)\}$. The first component in (A6.36) can then be simplified using

$$l_p^{jl} = f^{jl}(\lambda) \left[1 + O_p \left\{ \left(\frac{n}{k} + \lambda k^{2q+1} \right)^{-2} \frac{n}{k} \right\} \right]$$

with $f^{jl}(\lambda)$ as the matrix inverse of $f_{jl}(\lambda) = w_{jl} + \lambda k^{2q+1} \tilde{D}_{jl}$, or written as matrix $F(\lambda) = B_{q,k}^T W B_{q,k} + \lambda k^{2q+1} \tilde{D}_k$. With the same arguments we see that $l_{p,stu}$ is of diagonal structure, meaning that $l_{p,stu}$ is zero if $\max\{|s-t|, |s-u|, |t-u|\} > q+1$, otherwise the element has order $O(n/k)$. This allows for the quantifying of the remaining components in (A6.36) and we get with tedious, but simple, calculations

$$\begin{aligned} \hat{\omega}^j - \omega_0^j &= f^{jl}(\lambda) l_{p,l} + \left[O_p \left\{ \left(\frac{n}{k} + \lambda\, k^{2q+1} \right)^{-2} \frac{n}{k} \right\} \right. \\ &\quad \left. + O \left\{ \left(\frac{n}{k} + \lambda\, k^{2q+1} \right)^{-3} \frac{n}{k} \right\} \left\{ O_p \left(\frac{n}{k} \right) + O \left(\lambda^2\, k^{4q+2} \right) \right\} \right] \end{aligned} \tag{6.37}$$

Rewriting the leading component in matrix notation yields

$$\hat{\omega} - \omega_0 = F^{-1}(\lambda) \left\{ B_{q,k}^T l_\eta - \lambda\, k^{2q+1} \tilde{D}_k \omega_0 \right\} + \ldots \tag{6.38}$$

with correction terms of the asymptotic order listed in (6.37), where $l_\eta = l_\eta[\vartheta(B_{q,k}\omega_0)]$, that is we follow the convention of dropping the parameter argument if the function is evaluated at the best spline coefficient ω_0. Note that we postulated $\|u_0\|_\infty = O(k^{-1})$ (see (6.26)) or equivalently $\|\omega_0\|_\infty = O(k^{-(q+1)})$. A sufficient condition for this to hold

is that $L_{q,k}\omega_0$ is a k dimensional vector with elements of order $O(k^{-(q+1)})$. Accordingly $L_{q,k}^T L_{q,k}\omega = \tilde{D}_k\omega_0$ has elements of order $O(k^{-(q+1)})$. Consequently, the Mean Squared Error for $\hat{\omega}$ has the leading terms

$$\begin{aligned} E\left[\hat{\omega} - \omega_0\right] &= -F^{-1}(\lambda)\lambda\, k^{2q+1}\, \tilde{D}_k\omega_0 \left\{1 + o(1)\right\} \\ &= O\left\{\left(\frac{n}{k} + \lambda\, k^{2q+1}\right)^{-1} \lambda\, k^q\right\}, \end{aligned} \tag{6.39}$$

$$\begin{aligned} \mathrm{Var}(\hat{\omega}) &= F^{-1}(\lambda)F(\lambda = 0)F^{-1}(\lambda)\left\{1 + o(1)\right\} \\ &= O\left\{\left(\frac{n}{k} + \lambda\, k^{2q+1}\right)^{-2} \frac{n}{k}\right\} \end{aligned} \tag{6.40}$$

Finally, with (6.38) we see that the dominant stochastic part of $\hat{\omega} - \omega_0$ is $F^{-1}(\lambda)B_{q,l}^T l_\eta$. Since l_η is a vector of independent random variables the central limit theorem applies so that with (6.39) and (6.40) we get (6.8).
Setting $k = O(n^\kappa)$ and using (6.39) and (6.40), with the mean squared error taking ω_0 as true coefficient, has the order

$$\begin{aligned} \mathrm{MSE}(\hat{\omega}|\omega_0) &= \mathrm{E}(\hat{\omega} - \omega_0)^2 + \mathrm{Var}(\hat{\omega}) \\ &= O\left\{\left(\frac{n}{k} + \lambda\, k^{2q+1}\right)^{-2}\right\}\left\{O\left(\lambda^2\, k^{2q}\right) + O\left(\frac{n}{k}\right)\right\} \\ &= O\left((n^{1-\kappa} + \lambda\, n^{\kappa(2q+1)})^{-2}\right)\left(O\left(\lambda^2\, n^{2q\kappa}\right) + O\left(n^{1-\kappa}\right)\right) \end{aligned} \tag{6.41}$$

Apparently, if we now choose (a) $\lambda = O\left(n^{1-(2q+2)\kappa}\right)$ we find that asymptotically the penalization does not dominate the Fisher matrix and the mean squared error simplifies to

$$\mathrm{MSE}(\hat{\omega}|\omega_0) = O(n^{2(\kappa-1)}) \cdot \left\{O(\lambda^2\, n^{2q\kappa}) + O(n^{1-\kappa})\right\} \tag{6.42}$$

Moreover, choosing (b) $\lambda = O(n^{(1-(2q+1)\kappa)/2})$ provides the optimal choice for the smoothing parameter λ in an asymptotic sense so that the mean squared error simplifies to order $O(n^{\kappa-1})$. Simple calculation shows that the above postulated asymptotic orders (a) and (b) hold simultaneously if $\kappa = (2q+3)^{-1}$, which justifies (6.25) and (6.28). The mean squared error for $\hat{\omega}$ is then of order $O\left(n^{-(2q+2)/(2q+3)}\right)$.
The asymptotic orders are derived for the spline coefficients ω. It is however easy to directly extend them to the functional estimate $\hat{\eta}(x) = B_{q,k}^T(x)\hat{\omega}$. In particular the mean squared errors for $\hat{\eta}(x)$ and $\hat{\omega}$ are of the same asymptotic order. This follows since vec-

tor $B_{q,k}(x)$ is zero except of $q+1$ elements so that the asymptotic order of $\hat{\omega}$ directly transfers to $\hat{\eta}(x)$. This in turn provides the order of component 1 in (6.33).
In the second part of the proof we focus the approximation bias $\delta_0(x)$ given in (6.32). Since $\eta(x)$ is approximated in each interval $[\tau_j,\, \tau_{j+1}]$ by a polynomial of order q we find for $\eta(.) \in \mathcal{C}^{q+1}([0,\, 1])$ by Taylor series an approximation bias $\delta_0(x)$ of order $O(k^{-(q+1)})$. Observing the order of k given in (6.25) we see that the squared approximation bias, that is component 2 in (6.33), has order $O\left(k^{-2(q+1)}\right) = O\left(n^{-(2q+2)/(2q+3)}\right)$ which is the same as for component 1 in (6.33).
Finally, component 3 in (6.33) results by multiplication of the bias (6.39) and the approximation bias. Keeping the above results in mind we find with the same arguments as used above, that this component is also of order $O\left(n^{-(2q+2)/(2q+3)}\right)$ so that (6.33) is a decomposition with elements having all the same asymptotic order. Combining the results we get the final expansion

$$\begin{aligned} \hat{\eta}(x) - \eta(x) &= B_{q,k}(x)\hat{\omega} - \eta(x) && (6.43) \\ &= B_{q,k}(x)\left\{B_{q,k}^T W B_{q,k} + \lambda\, n^{\frac{2q+1}{2q+3}} \tilde{D}_k\right\}^{-1} \\ &\quad \times \left\{B_{q,k}^T l_\eta - \lambda\, n^{\frac{2q+1}{2q+3}} \tilde{D}_k \omega\right\} + O_p\left(n^{-\frac{q+1}{2q+3}}\right). && (6.44) \end{aligned}$$

6.5 Discussion

We studied the asymptotic order with which the spline basis dimension in the generalized penalized smoothing model is supposed to grow to optimize the mean squared error. The main results about optimal rate of convergence $n^{-(2q+2)/(2q+3)}$ and spline dimension $k = O(n^{1/(2q+3)})$ correspond exactly to the results reported under normality assumption for regression splines in Agarwal & Studden (1980). We showed that in the generalized mixed model framework the error of the Laplace approximation remains to be of negligible order even if the spline dimension grows with the above order.

7 Summary

In this work we considered some theoretical and practical aspects of penalized spline smoothing - a smoothing technique which gained much popularity over the last decade. In Chapter 2 we presented penalized splines as smoothing technique. P-spline smoothing is a very flexible concept - combination of different splines basis, penalties and knots provides a wide spectrum of smoothers. Mixed model representation and Bayes model for smoothing widen the possibilities for estimation and inference and deliver new insights. For example, some special problems, which are either impossible or difficult to handle with standard nonparametric techniques, like smoothing in presence of correlated errors or estimation of locally varying functions, can be treated successfully in the mixed model (and thus Bayes) framework for penalized splines, as shown in Chapters 3 to 5.
The problem of smoothing with correlated errors, discussed in Chapter 3, is complex and prominent in nonparametric statistics. Smoothing parameter choice with MSE-based criteria fails in presence of correlated errors leading to the serious overfitting unless the correlation structure is correctly specified. In general, the correlation structure is unknown in practice and its estimation requires a sufficiently good estimate of the mean function. Thus one faces a dilemma in practice. Mixed model representation of penalized splines has two advantages. First, in the mixed model framework the correlation matrix parameters can be estimated along with the other parameters if the correlation structure is specified (e.g. AR(1)). Second, the smoothing parameter estimate (which is just a (RE)ML estimate of two variances) is less sensitive towards misspecification of the correlation structure compared to MSE based choices. These two features of the (RE)ML estimate can help discovering the true correlation structure much more efficiently. We demonstrated on a number of examples a simple strategy. First, fit the model with the mixed model software, assuming the most probable correlation structure and inspect whether the residuals behave in accordance with the assumption about the covariance matrix. If the correlation structure is only moderately misspecified it can be visible in the plot of (partial) autocorrelation functions, which helps to determine the true correlation structure of the data. In Chapter 4 we made use of the above property of the (RE)ML estimate to perform the two-dimensional fit of the term structure

of interest rates (over calendar time and time left to maturity). Apart from the very large sample size, our data possesses an untypical correlation structure: while the single bonds are correlated over calendar time we observed no correlation over the second covariate (time left to maturity). We fitted the model in the mixed model framework in two stages. First, we determined the correlation structure over calendar time of the single bonds with the above described strategy. Afterwards, we standardized the data with the obtained correlation matrix and performed standard bivariate estimation with independent errors.

In Chapter 5 we approached the problem of smoothing a function of locally varying complexity, that is if the regression function is changing rapidly in some regions while in other regions it is very smooth. Smoothing of such functions with a single smoothing parameter, implying that all spline coefficients undergo the same penalization, is not efficient. We achieved spatial adaptivity by imposing a functional structure on the smoothing parameter and representing adaptive smoothing as a hierarchical (generalized) mixed model. Intractable integral in the corresponding likelihood we approximated with the Laplace's method, resulting in the fast and simple adaptive smoothing technique, which can be readily extended for the fitting of the (generalized) additive and bivariate models. We provided the R package *AdaptFit*, which is available at CRAN, to make the approach accessible. Adaptive smoothing in many application is superior to the smoothing with the global smoothing parameter and our fast and handy technique allows for more efficient smoothing with little additional numerical effort. We illustrated our method with the example on absenteeism of workers of a medium-sized German industrial company. We were interested in estimation of the discrete hazard function in two dimensions - calendar time and duration of absenteeism in days. Representing duration of absenteeism as a set of binary variables we modelled the hazard as a binomial probability and fitted two-dimensional logit-model. Our dataset was, however, not standard - during the observation period the company went through a major downsizing process which increased the probability of returning to work after a sick leave noticeably in this period. We showed that the generalized bivariate adaptive smoothing delivered more adequate estimate of such non-standard data structure than the non-adaptive methods.

Finally, in Chapter 6 we dealt with some asymptotics issues. In particular we were interested how fast spline basis dimension is supposed to grow to provide an optimal rate of convergence. We based our investigation on generalized penalized spline smoothing model. Since smoothing in generalized mixed model framework involves Laplace approximation, we investigated additionally the order of the corrections terms in the Laplace approximation, given that the spline dimension grows with the sample size. The ob-

tained optimal order of the spline dimension corresponds exactly to the results reported for the regression splines under normality assumption. Moreover, we showed that with this order of the spline basis dimensions the error term in the Laplace approximation remains of negligible order.

References

Aerts, M., Claeskens, G., and Wand, M. (2002). Some theory for penalized spline additive models. *Journal if Statistical Planing and Inference* **103**, 455–470.

Agarwal, G. and Studden, W. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Annals of Statistics* **8**, 1307–1325.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institue of Statistical Mathematics* **21**, 243 – 47.

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85**, 749–759.

Backus, D., Foresi, S., and Zin, S. (1998). Discrete-time models of bond pricing. *NBER Working Paper* **w6736**, www.nber.org/papers/w6736.

Baladandayuthapani, V., Mallick, B., and Carroll, R. (2005). Spatially adaptive Bayesian penalized regression splines (P-splines). *Journal of Computaional and Graphical Statistics* **14**, 378–394.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for use in Statistics*. Chapman & Hall.

Beran, J. and Feng, Y. (2001). Local polynomial estimation with FARIMA-GARCH error process. *Bernoulli* **7**, 733–750.

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

de Boor, C. (2001). *A practical guide to splines*. Revised edition. New York: Springer.

Box, G. and Jenkins, G. (1970). *Time Series Analysis, Forcasting and Control*. San Fransisco: Holden-Day.

Brandt, M. and Yaron, A. (2003). Time-consistent no-arbitrage models of the term structure. *NBER Working Paper* **w9458**, www.nber.org/papers/w9458.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association.* **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Cardot, H. (2002). Local roughness penalties for regression splines. *Computational Statistics* **17**, 89–102.

Chambers, D., Carleton, W., and Waldman, D. (1984). A new approach to estimation of the term structure of interest rates. *Journal of Finance and Quantitative Analysis* **19**, 233 – 253.

Chan, K., Karolyi, G. A., Longstaff, F., and Sanders, A. B. (1992). An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance 47*(3).

Chandrasekhar, S. (1954). *Selected papers on noise and stochastic processes. In: Wax, N (Ed.), Stochastic Problems in Physics and Astronomy*. Dover, New York.

Cox, J., Ingersoll, J., and Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica* **53**, 385–408.

Crainiceanu, C., Ruppert, D., and Carroll, R. (2005). Spatially adaptive Bayesian P-splines with heteroscedastic errors. *working paper*.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling* **2**, 333–349.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Hoboken, New Jersey: Wiley.

Diebold, F. and Li, C. (2003). Forecasting the term structure of government bond yields. *NBER Working Paper* **w10048**, www.nber.org/papers/w10048.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Clarendon Press.

Diggle, P. and Hutchinson, M. (1989). On spline smoothing with autocorrelated errors. *Austral. J. Statist* **31(1)**, 166–182.

Draper, N. and Smith, H. (1998). *Applied regression analysis*. Wiley.

Durban, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics* **18**, 251–262.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Stat. Science 11*(2), 89–121.

Eilers, P. H. C. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems* **66**, 159–174.

Eilers, P. H. C. and Marx, B. D. (2004). Splines, knots and penalties. *technical report*.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*. New York: Dekker.

Eubank, R. L. (1988). *Spline smoothing and nonparametric regression*. Dekker: Ney York.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 715–745.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B* **57**, 371–394.

Fan, J. and Yao, Q. (2003). *Nonlinear time series*. New York: Sringer.

Fisher, M., Nychka, D., and Zervos, D. (1995). Fitting the term structure of interest rates with smoothing splines. 1.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications.

Fried, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics* **31**, 3–21.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics* **31**, 3–39.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and generalized linear models*. Chapman & Hall.

Green, P. (1987). Penalized likelihood for general semiparametric regression models. *Internat. Statist. Rev.* **55**, 245–259.

Grunwald, G. and Hyndman, R. (1998). Smoothing non-Gaussian time series with autoregressive errors. *Computational Statistics and Data Analysis* **28**, 171–191.

Gu, C. (1992). Cross-validating non-Gaussian data. *Journal of Computational and Graphical Statistics* **1**, 169–179.

Gu, C. and Wahba, G. (1991). Minimizing GCV/GLM scores with multiple smoothing parameters via the newton method. *SIAM J. Sci. Statist. Computing* **12**, 383–398.

Guttorp, R. (1991). *Statistical inference in branching processes*. Wiley, New York.

Hart, J. D. (1991). Kernel regression estimation with time series error. *Journal of the Royal Statistical Society, Series B* **53**, 173–187.

Harville, D. (1974). Optimal procedures for some constrained selection problems. *Journal of the American Statistical Association.* **69**, 446–452.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association.* **72**, 320–338.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Henderson, C. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* **21**, 309–310.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* **6**, 35–54.

Heyde, C. and Seneta, E. (1971). Estimation theory for growth and immigration rates in a multiplicative process. *Journal of Applied Probability* **9**, 235–256.

Ho, T. and Lee, S.-B. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance* **41 (5)**, 1011 – 1029.

Hodrick, R. and Prescott, E. (1997). Postwar U.S. business cycles: An empirical approach. *Journal of Money, Credit ad Banking* **29[1]**, 1–16.

Huang, J. Z. and Stone, C. J. (2003). Statistical modeling of diffusion processes with free knot splines. *Journal of Statistical Planning and Inference* **116**, 451–474.

Hurvich, C., Simonoff, J., and Tsai, C. (19958). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271–93.

Ioannides, M. (2003). A comparison of yield curve estimation techniques using UK data. *Journal of Banking & Finance* **27**, 1 – 26.

Jarrow, R., Ruppert, D., and Yu, Y. (2004). Estimating the term structure of corporate debt with a semiparametric penalized spline model. *Journal of the American Statistical Association.* **99**, 57 – 66.

Jeffrey, A., Linton, O., and Nguyen, T. (2001). Flexible term structure estimation: Which method is preferred. Technical Report, London School of Economics.

Kass, R. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayesian models). *Journal of the American Statistical Association.* **84**, 717–726.

Kauermann, G. (2004). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planing and Inference* **127**, 53–69.

Kauermann, G. and Ortlieb, R. (2004). Temporal pattern in the number of staff on sick leave: the effect of downsizing. *Journal of the Royal Statistical Society, Series C* **53**, 353–367.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Laird, N. and Louis, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association.* **82**, 739–757.

Laird, N. and Ware, J. (1982). Random-effects models for longitudial data. *Biometrics* **38**, 963–974.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Lindstrom, M. and Bates, D. M. (1988). Newton-Raphson and EM algorithmus for linear mixed-effects models for repeated-measures data. *Journal of the Royal Statistical Society, Series B* **61**, 381–400.

Linton, O., Mammen, M., Nielsen, J., and Tanggaard, C. (2000). Estimating yield curves by kernel smoothing methods. *Journal of Econometrics* **105**, 185 – 223.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–116.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear, and Mixed Models*. New York: Wiley.

McCulloch, J. (1971). Measuring the term structure of interest rates. *Journal of Business* **44**, 19 – 31.

McMullan, A., Bowman, A., and Scott, E. (2003). Non-linear and nonparametric modelling of seasonal enviromental data. *Computational Statistics* **18**, 167–184.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association.* **78**, 47–65.

Nelson, C. and Siegel, A. (1987). Parsimonious modelling of yield curves. *Journal of Business 60*(4), 473 – 489.

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of statistical software* **9(1)**.

Nychka, D. (2000). Spatial process estimates as smoothers. *Smoothing and Regression*, Springer: Heidelberg.

Nychka, D. and Saltzman, N. (1998). *Case Studies in Enviromental Statistics*. New York: Springer.

Opsomer, J., Wang, Y., and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

Opsomer, J. D. and Hall, P. (2005). Theory for penalised spline regression. *Biometrika* **92**, 105–118.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**, 502–518.

O'Sullivan, F., Yandell, B., and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association.* **81**, 96–103.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554.

Pinheiro, J. and Bates, D. (2002). *Mixed-Effects Models in S and Splus*. New York: Springer Verlag.

Pintore, A., Speckman, P., and Holmes, C. C. (2005). Spatially adaptive smoothing splines. *Biometrika*, (to appear).

R Development Core Team (2005). *R: A Language and Enviroment for Statistical Computing*. R foundation for statistical computing, Wien.

Ray, B. and Tsay, R. (1997). Bandwidth selection for kernel regression with long-range dependent errors. *Biometrika* **84(4)**, 791–802.

Reinsch, C. (1967). Smoothing by spline functions. *Numer. Math.* **10**, 177–183.

Robinson, G. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15–32.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–224.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press.

Searle, S., Casella, G., , and McCulloch, C. (1992). *Variance Components*. Wiley.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Simonoff, J. and Tsay, C. (1999). Semiparametric and additive model selection using an improved akaike criterion. *Journal of Computational and Graphical Statistics* **8**, 22–40.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317–344.

Smith, M., Wong, C., and Kohn, R. (1998). Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society, Series B* **60**, 311–331.

Steeley, J. (1990). Modelling the dynamics of the term structure interest rates. *The Economic and Social Review 21*(4), 337 – 361.

Stone, C., Hansen, M., Kooperberg, C., and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics* **25**, 1371–1425.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics* **5**, 177–188.

Wahba, G. (1990). *Spline Models for observational data*. Philadelphia: SIAM.

Wand, M. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika* **86**, 936–940.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association.* **93**, 341–348.

Westgren, A. (1916). Die Veranderungsgeschwindigkeit der lokalen Teilchenkonzentration in kolloiden Systemen (erste mittelung). *Ark. Mat. Astron. Fys.* **11**, 1–25.

Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *Journal of the Royal Statistical Society, Series B*, 413–428.

Wood, S. (2003). Thin plate regresson splines. *Journal of the Royal Statistical Society, Series B* **65**.

Wood, S. (2004). Stable and efficient multiple smoothing parameter for generalized additive models. *Journal of the American Statistical Association*, 637–686.

Wood, S., Jiang, W., and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–528.