Bayesian P-Splines in Structured Additive Regression Models

Andreas Brezger

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

München, 28. Dezember 2004

Bayesian P-Splines in Structured Additive Regression Models

Andreas Brezger

Dissertation
an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig–Maximilians–Universität
München

vorgelegt von
Andreas Brezger
aus München

München, 28. Dezember 2004

Erstgutachter: Prof. Dr. Ludwig Fahrmeir
Zweitgutachter: Prof. Dr. Leonhard Held
Drittgutachter: Prof. Brian D. Marx

Rigorosum: 10. Mai 2005

"Hütet euch vor der geraden und vor der betrunkenen Linie. Aber besonders vor der geraden Linie.
Die gerade Linie führt zum Untergang der Menschheit."

Friedensreich Hundertwasser

Vorwort

Diese Arbeit entstand während meiner Tätigkeit als Mitarbeiter am Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" am Department für Statistik an der Ludwig-Maximilians-Universität München und wurde somit durch Mittel der Deutschen Forschungsgemeinschaft (DFG) gefördert. Neben der finanziellen Unterstützung möchte ich eine Reihe von Personen hervorheben, die auf unterschiedliche Art und Weise einen wesentlichen Anteil an der vorliegenden Arbeit haben.

Zu allererst möchte ich meinem Doktorvater Prof. Dr. Ludwig Fahrmeir für seine hervorragende Betreuung danken. Er war während der Anfertigung der Dissertation stets für Fragen erreichbar und hat mit seiner unkomplizierten Art entscheidend zu dem angenehmen Arbeitsklima am Lehrstuhl beigetragen. Dafür möchte ich mich auch bei allen weiteren Mitarbeitern herzlich bedanken. Frau Schnabel vom Sekretariat und Frau Burger von der SFB-Geschäftsstelle möchte ich hier ebenfalls nicht unerwähnt lassen.

In gleichem Maße möchte ich mich bei Stefan Lang für die enge und produktive Zusammenarbeit bedanken. Außerdem gebührt mein Dank Thomas Kneib und dem gesamten übrigen *BayesX*-Team (Christiane Belitz, Alexander Jerak, Andrea Hennerfeind, Thomas Kneib, Stefan Lang, Leyre Osuna) für die fruchtbare Zusammenarbeit (und die Geduld, wenn wieder einmal etwas nicht funktionierte). Ein besonderer Dank geht auch an meinen weiteren Co-Autor Winni Steiner.

Weiterhin möchte ich mich bei Prof. Dr. Leonhard Held und Prof. Brian D. Marx bedanken, die freunlicherweise als Gutachter für meine Dissertation tätig waren.

Nicht zuletzt möchte ich meinen Eltern danken, die immer volles Vertrauen in meine universitäre Laufbahn gesetzt haben und diese überhaupt ermöglicht haben.

Entschuldigen möchte ich mich bei allen, die gelegentlichen Forschungsfrust ertragen mussten und zur Überwindung desselben beigetragen haben, ganz besonders bei Andrea Hennerfeind.

München, Juni 2005 *Andreas Brezger*

Zusammenfassung

Diese Arbeit beschäftigt sich mit der Entwicklung von Bayesianischen semiparametrischen Regressions-Modellen und deren Schätzung mit Hilfe von Markov chain Monte Carlo (MCMC) Verfahren. Es werden Modelle mit einem strukturierten additiven Prädiktor betrachtet. Dieser kann neben parametrisch und nichtparametrisch modellierten Effekten auch räumliche Effekte und zufällige Effekte zur Berücksichtigung von unbeobachteter Heterogenität sowie zeitlich oder räumlich variierende Effekte enthalten. Die am weitesten verbreiteten univariaten und multivariaten Verteilungen für die Zielgröße können behandelt werden.

Diese Arbeit konzentriert sich speziell auf die Modellierung metrischer Kovariablen durch Bayesianische P-Splines. Dabei werden eindimensionale sowie zweidimensionale Oberflächenschätzungen behandelt. Außerdem finden lokal adaptive Glättung und mögliche Monotonie-Restriktionen an die Schätzung Berücksichtigung. Ein wesentliches Ziel ist dabei die Entwicklung von effizienten MCMC Algorithmen für die Bayesiansche Inferenz und deren Implementierung in einem benutzerfreundlichen Programm-Paket.

Ein weiteres Kapitel beschäftigt sich mit der Berechnung von simultanen Wahrscheinlichkeitsaussagen über die geschätzten P-Splines. Damit kann beurteilt werden, ob eine nichtparametrische Modellierung erforderlich ist oder eine einfachere, parametrische Modellierung ausreicht. Die in dieser Arbeit entwickelten Methoden werden auf mehrere komplexe, reale Problemstellungen angewendet und erweisen sich in der Praxis als äußerst wirkungsvolles und flexibles Instrument.

Abstract

This thesis aims at developing Bayesian semiparametric regression models and making inference using Markov chain Monte Carlo (MCMC) simulation techniques. The focus is on models with structured additive predictor, which may comprise parametric and nonparametric effects as well as spatial effects and random effects to capture unobserved heterogeneity and spatially or temporally varying effects. The most common univariate and multivariate response distributions are considered.

This work concentrates especially on modeling continuous covariates by Bayesian P-splines. One-dimensional P-splines and two-dimensional surface estimations are considered. Additionally, locally adaptive smoothing and possible monotonicity restrictions regarding the estimations are taken into account. An important goal is to develop efficient MCMC

algorithms for Bayesian inference, and their implementation in an easy to use software package.

A further topic is the computation of simultaneous probability statements on the estimated P-spline to assess the necessity of a nonparametric estimate compared to a more parsimonious, parametric fit. The methodology developed in this thesis is applied to several complex real problems and proves to be a very flexible and powerful tool in statistical practice.

Contents

Chapter 1

Introduction

Regression analysis certainly is one of the central areas of statistical research. Nowadays, Generalized Additive Models (GAM, Hastie and Tibshirani 1990) can be considered as a well established tool in semi- and nonparametric regression. Many implementations in commercial (S-PLUS, SAS) and public domain (e.g. R) software have made GAMs to a widely used instrument in practical statistical analysis. Various extensions to GAMs have been made in the recent years. Among the most important ones are Generalized Additive Mixed Models (GAMM, Lin and Zhang 1999) for incorporation of unobserved heterogeneity, and Varying Coefficient Models (VCM) by Hastie and Tibshirani (1993) to account for interactions of covariates. Geoadditive models within a mixed model setting have been introduced by Kammann and Wand (2003).

For modeling the non-linear parts of a GAM there exist a variety of different approaches. While polynomials of a certain degree l are often not flexible enough for small l, estimates become more flexible but also rather unstable for large l, especially at the boundaries. Taking into account more sophisticated methods, we may distinguish mainly two approaches for nonparametric modeling. These are local polynomial regression and approaches based on basis functions. A good overview over recent developments in semiparametric regression can be found in Fahrmeir and Tutz (2001) or Ruppert, Wand and Carroll (2003). In this thesis we will focus on basis function methods. More specifically we use a specific form of polynomial regression splines which are parameterized in terms of B-spline basis functions together with a penalization of adjacent parameters, also known as P-splines (Eilers and Marx 1996).

Another key development in statistics is the rise of Markov Chain Monte Carlo methods in the last one and a half decades. First introduced by Metropolis, Rosenbluth, Teller and Teller (1953) and Hastings (1970) these methods had their breakthrough in statistics not before the 1990s when increasing computer power facilitated fast computation of previously intractable problems in Bayesian statistics. In the recent years contributions to Bayesian statistics have gained more and more weight in the literature and Bayesian inference is now a well established tool for complex statistical analysis. In the context of the GAMs and their extensions mentioned above, a great number of regression parameters have to be estimated together with additional hyperparameters, such as smoothing parameters. This

is typical situation where MCMC methods are of great importance. Inference in this thesis is based solely on fully Bayesian methodology via MCMC.

This thesis aims to develop a unified framework for GAMs in which a Bayesian version of P-splines is the main building block for modeling nonparametric effects. These models are imbedded in a very general class of models, which we call Structured Additive Regression (STAR) Models. Therein VCMs, GAMMs and other model classes well known from the literature are included as special cases. Inference relies on fully Bayesian methodology. Special attention has been drawn to computationally efficient implementation of the methods and to the development of an easy to use public domain software package that makes the methodology applicable for a wide range of problems for researchers and applied statisticians. The software developed and used in this work is available via internet at *http://www.stat.uni-muenchen.de/∼lang/bayesx/*.

In this introduction we first give an overview over different approaches in spline regression. In the second subsection we explain how Bayesian P-splines can be combined with a variety of other approaches for modeling of covariates within the framework of STAR models.

1.1 Spline regression

Consider the classical smoothing problem

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where (y_i, x_i) is the i-th observation from the continuous variables y and x. To approximate the unknown function $f(x)$ we restrict ourselves to the class of spline functions. A spline is a piecewise defined function that fulfills certain smoothness constraints at the interfaces, the so called *knots*. The most widely known spline function is the sub-class of polynomial splines, where the pieces consist of polynomials of a certain degree l.

Suppose the domain of x is partitioned by a set of knots

$$x_{min} = \xi_0 < \xi_1 < \ldots < \xi_{r-1} < \xi_r = x_{max},$$

then a polynomial spline has the following properties:

- A spline is a polynomial of degree l in each interval $[\xi_{\rho-1}, \xi_\rho]$, $\rho = 1, \ldots, r$.
- A spline is $l - 1$ times continuous differentiable at the knots ξ_ρ.

Defining a design matrix X, where the entry X_{ij} is the value of the j-th basis function evaluated at observation i, enables us to write the vector of function evaluations $f = (f(x_1), \ldots, f(x_n))'$ as

$$f = X\beta.$$

Here, $\beta = (\beta_1, \ldots, \beta_M)'$ is a M-dimensional vector of regression parameters. For example, De Boor (1978) shows that such a spline may be written in terms of a linear combination of $M = r + l$ B-spline basis functions

$$f(x) = \sum_{\rho=1}^{M} \beta_\rho B_{\rho l}(x),$$

where $B_{\rho l}(x)$ denotes a B-spline basis of degree l. Figure 1.1 (a) illustrates 6 basis functions of degree 2 covering the interval $[0, 1]$, which is partitioned by 5 equidistant knots. In graphs (b) and (c) weighted basis functions and the resulting spline function $f(x)$ are displayed.

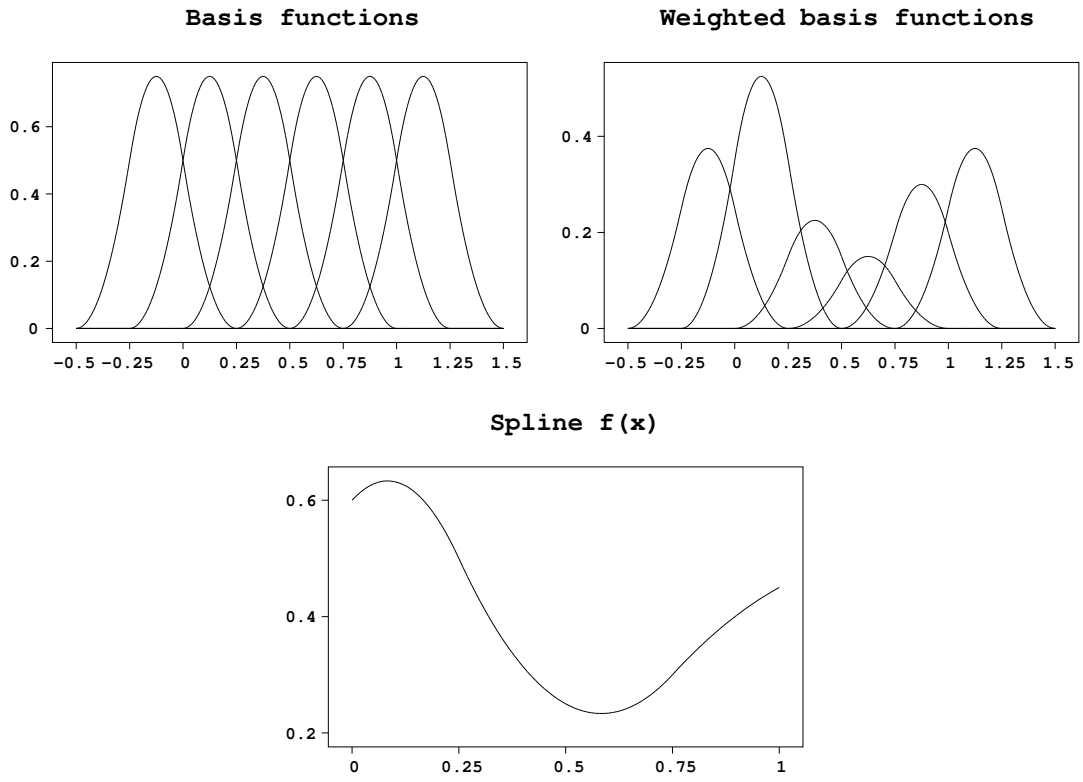


Figure 1.1: 6 B-spline basis function of degree $l = 2$ (5 knots at {0.0,0.25,0.5,0.75,1.0}) (a), weighted basis functions (b), and the resulting spline (c).

The crucial point of simple spline regression is selecting the number and the position of the knots. We may distinguish three different approaches that make use of basis functions. These are approaches based on adaptive knot selection, approaches based on roughness penalties, and a combination of both. We will sketch the three approaches in the following subsections.

1.1.1 Approaches based on adaptive knot selection

Approaches based on adaptive knot selection aim at a parsimonious selection of the number of basis functions and a careful choice of the position of the knots to obtain a smooth, yet sufficiently flexible estimation. Such methods have been presented in a frequentist setting for example by Friedman (1991), who introduced the software MARS (Multivariate Adaptive Regression Splines) and Stone, Hansen, Kooperberg and Truong (1997), who developed POLYMARS. Bayesian approaches have been considered by Smith and Kohn (1996), Denison, Mallick and Smith (1998), Biller (2000), Biller and Fahrmeir (2001) or Di Matteo, Genovese and Kass (2001).

MARS uses so called *reflected pairs* $\{(x_j - t)_+, (t - x_j)_+\}$ of piecewise linear basis functions for each input variable x_j. The set of candidate functions therefore is

$$\mathcal{C} = \{(x_j - t)_+, (t - x_j)_+\} \quad \begin{array}{l} t \in \{x_{1j}, \ldots, x_{nj}\} \\ j = 1, \ldots, p \end{array}$$

A basis function $h_\rho(x)$ can be any function of $\mathcal{C}$ or the product of one or more of them – under the restriction that each input variable can appear at most once in a product. The model is built in a forward procedure until some pre-specified maximum number of knots is reached, and then the model is pruned via a backward procedure. In each step of the forward procedure all products of a function h_ρ included in the existing model $\mathcal{M}$ with a reflected pair in $\mathcal{C}$ is considered as a new basis function pair. The term of the form

$$\beta_{M+1} h_\rho(x)(x_j - t)_+ + \beta_{M+2} h_\rho(x)(t - x_j)_+, \quad h_\rho \in \mathcal{M},$$

that gives the smallest GCV value is added to $\mathcal{M}$. Estimation is performed by standard linear regression. A very similar software specially designed for classification is the routine POLYMARS (Stone et al. 1997). The main difference is that POLYMARS works within a multiple logistic framework and uses a quadratic approximation of the log-likelihood to search for a new candidate function pair.

Smith and Kohn (1996) follow a Bayesian approach for the linear regression model $y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$. They introduce an indicator vector γ with the i-th element such that $\gamma_i = 0$ if $\beta_i = 0$ and $\gamma_i = 1$ if $\beta_i \neq 0$. A multivariate Gaussian

$$N(0, c\sigma^2 (X_\gamma' X_\gamma)^{-1}) \tag{1.1}$$

prior is assigned on β_γ and $p(\gamma_i = 1) = \pi_i = 0.5$ is assumed for $i = 1, \ldots, M$. Here X_γ denotes the submatrix of the truncated power basis regression spline design matrix X that corresponds to the basis functions present in the current model. β_γ denotes the corresponding subvector of β. The prior 1.1 is chosen since it is proportional to the variance of the least squares estimate of β_γ.

Smith and Kohn (1996) suggest to place a knot at every third to fifth observation, up to a maximum of 40. The main advantage of this model formulation is that a Gibbs sampler can be used for MCMC inference. Smith and Kohn (1997) generalize this approach to bivariate curve fitting and Kohn, Smith and Chan (2001) refine their approach by using

the very general class of radial basis functions. Therein, thin plate splines (compare Duchon 1977, Wood 2003) are included as a special case. Furthermore they place a hyperprior on the probability π_i and modify the mean of the prior for the regression coefficients by using a least square estimate $\hat{\gamma}$ instead of mean zero

$$\beta_\gamma \sim N(\hat{\beta}_\gamma, c\sigma^2(X'_\gamma X_\gamma)^{-1}).$$

This is the likelihood for β_γ conditional on y and γ, where the variance is scaled by a constant c. Kohn et al. set $c = n$, since they consider it likely that $c\sigma^2(X'_\gamma X_\gamma)^{-1}$ remains reasonably constant as n increases'.

The approach of Denison et al. (1998) deals with the same Gaussian model, but here the basis functions are not spline but piecewise polynomials

$$y_i = f(x_i) + \epsilon_i = \sum_{\rho=0}^{l} \beta_{\rho 0}(x_i - \xi_0)_+^\rho + \sum_{\nu=1}^{k} \sum_{\rho=l_0}^{l} \beta_{\rho\nu}(x_i - \xi_\nu)_+^\rho + \epsilon_i, \quad i = 1, \ldots, n$$

to allow for discontinuous functions. The authors place a prior on the number and on the position of the knots and employ a reversible jump MCMC algorithm (Green 1995) for inference of these parameters. However, the regression parameters are treated as fixed and are therefore not supplied with a prior distribution. Hence, estimation of the vector β is performed via a simple least squares minimization in each step. The candidate knot set is the set of all different observations $\{x_1, \ldots, x_n\}$.

Biller (2000) presents a fully Bayesian approach based on natural cubic regression splines, where he assigns either a poisson or a discrete uniform prior on the number of knots r and a multivariate normal distribution

$$\beta|\xi, r, \sigma^2 \sim N(0, cI_r)$$

on the regression coefficients. The candidate knot set is again the set of all different observations $\{x_1, \ldots, x_n\}$. Inference is based on a reversible jump MCMC algorithm. In every iteration the algorithm randomly chooses between adding (birth), deleting (death) or shifting a knot. In the birth step the new knot position is selected uniformly at random from the candidate knots not yet in the model, the death step is defined reversely. For a change in position of knots, first a knot ξ_ρ to be shifted is chosen uniformly at random, and then the new position is drawn uniformly from all candidate knots in the interval $(\xi_{\rho-1}, \xi_{\rho+1})$. Biller and Fahrmeir (2001) generalize this approach to VCMs. In contrast to the lastly described methods, Biller's (2000) and Biller and Fahrmeir's (2001) approach is not restricted to Gaussian models, but can be applied also to binomial or Poisson data.

A very similar method, where the candidate knots are not pre-specified, but are allowed to vary freely, is suggested by Di Matteo et al. (2001). They extend the work of Denison et al. (1998) to a fully Bayesian approach and to non-Gaussian models. In contrast to Denison et al. (1998) – but in accordance with Biller (2000) and Biller and Fahrmeir (2001) – they use natural cubic smoothing splines with an unknown number r and location of knots. However, they use the so called *unit-information* prior

$$\beta|\xi, r, \sigma \sim N(0, n\sigma^2(X'_\xi X_\xi)^{-1}),$$

where X_ξ is the natural cubic spline design matrix for the knot sequence $\xi = (\xi_0, \ldots, \xi_r)'$. The denomination unit-information is due to the fact, that the amount of information in the prior equals the information of *one* observation, as represented by the Fisher information matrix.

1.1.2 Approaches based on roughness penalties

Instead of carefully selecting knots and reducing the number of basis functions, the basic idea of roughness penalty approaches is to guarantee enough flexibility by a relatively large number of basis functions. In order to avoid overfitting and to reduce variability of the estimations appropriate restrictions on the parameter vector are imposed, that shrink the coefficients towards zero or penalize too abrupt jumps between adjacent parameters.

One of the most prominent representative of this approach are smoothing splines. This approach can be traced back to Reinsch (1967). Other key references are Wahba (1990) and Hastie and Tibshirani (1990). Here, the objective is to minimize the penalized least squares criterion

$$PL = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx, \tag{1.2}$$

where $f(x)$ is assumed to be a twice continuous differentiable function. The smoothing parameter λ controls the trade-off between data fit and smoothness, which is measured in terms of the integral over the curvature of $f(x)$. As it turns out, the solution is a natural cubic smoothing spline, with knots at every different observation point x_i and thus with as many regression parameters as different observations.

Another penalty based approach is presented by Shively, Kohn and Wood (1999) in conjunction with model selection. They use an integrated Wiener process W_j as prior for nonparametric effects

$$f_j(x_j) = \phi_j x_j + (\tau_j^2)^{1/2} \int_0^{x_j} W_j(\nu) d\nu, \quad j = 1, \ldots, p,$$

which results in cubic smoothing splines for the posterior means. In addition they use an indicator variable for each smoothing parameter τ_j^2 and for each of the parameters ϕ_j, enabling the algorithm to model an effect linearly or to drop a variable from the model. A two-step procedure is pursued, where in the first step a data based prior for $\Phi = (\phi_1, \ldots, \phi_p, \tau_1^2, \ldots, \tau_p^2)'$ in the second step, consisting of variable selection and model averaging, is provided. The first step uses a sampling scheme developed by Wong and Kohn (1996). The data based prior

$$\Phi_j \sim N(\hat{\Phi}_j, nB_j^{-1}),$$

in the second step is obtained from the posterior mean $\hat{\Phi}$ and the posterior covariance matrix A of the full model, where $B = A^{-1}$, and $\hat{\Phi}_j$ denotes the subvector of $\hat{\Phi}$ (the submatrix of B) consisting of the rows (rows and columns), corresponding to component j.

In the popular P-spline approach (Eilers and Marx 1996, Marx and Eilers 1998) $f(x)$ is approximated by a linear combination of a relatively large (usually 20 to 40) number of B-spline basis functions and smoothness is ensured by a penalty

$$P(\lambda) = \lambda \sum_{\rho=k+1}^{M} (\Delta^k \beta_\rho)^2, \tag{1.3}$$

based on k-th order differences of the parameter vector, which can be seen as a discrete approximation of the penalty in (1.2). Various articles about different usages of P-splines have been published by the authors since then. In recent papers Eilers and Marx (2003) and Marx and Eilers (2005) present a 2-dimensional surface estimator based on tensor-product P-splines. Figure 1.2 displays such basis functions with degree $l = 3$. To let the graphic not become too complex, instead of the full basis, we show only every fourth basis function in each direction.

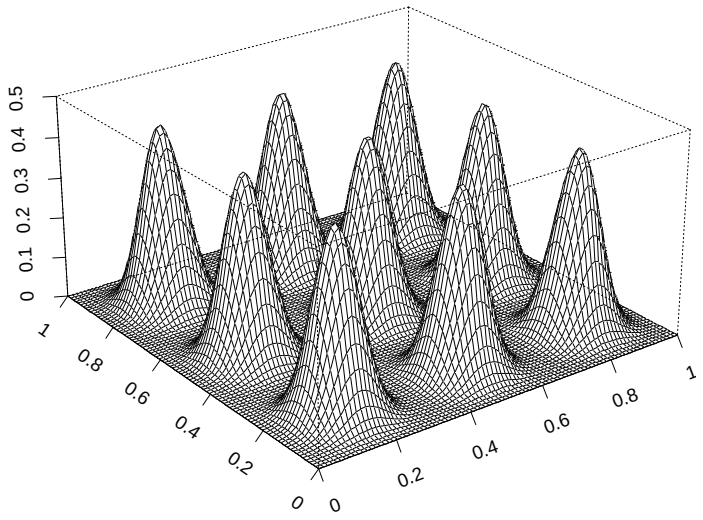


Figure 1.2: Tensor product B-spline basis functions of degree $l = 3$

In both approaches, smoothing splines and P-splines, the crucial point is the selection of the smoothing parameter λ. Traditionally the selection of λ is based on some goodness-of-fit criterion, e.g. the GCV criterion or the AIC and minimization is performed via a simple grid search algorithm.

In this thesis a Bayesian version of the P-spline approach of Eilers and Marx is developed and extended in several ways. The key idea is to use B-spline basis functions of a certain degree l to approximate the unknown function $f(x)$. The grid of knots is extended beyond the domain of x for l knots in each direction to guarantee enough flexibility at the boundaries. To demonstrate how the penalization works, in figure 1.3 we present a simulated data set which is fitted using a different amount of penalization. In the Bayesian version of P-splines we replace the penalty (1.3) by a stochastic version, namely a first

(RW1) or second (RW2) order random walk

$$\begin{aligned} \beta_\rho &= \beta_{\rho-1} + u_\rho, \quad u_\rho \sim N(0, \tau^2) \quad (RW1) \\ \beta_\rho &= 2\beta_{\rho-1} - \beta_{\rho-2} + u_\rho, \quad u_\rho \sim N(0, \tau^2) \quad (RW2) \end{aligned}$$

or in an equivalent two-sided formulation

$$\begin{aligned} \beta_\rho &= \frac{1}{2}\beta_{\rho-1} + \frac{1}{2}\beta_{\rho+1} + u_\rho, \quad u_\rho \sim N(0, \tau^2/2) \quad (RW1) \\ \beta_\rho &= -\frac{1}{6}\beta_{\rho-2} + \frac{2}{3}\beta_{\rho-1} + \frac{2}{3}\beta_{\rho+1} - \frac{1}{6}\beta_{\rho-2} + u_\rho, \quad u_\rho \sim N(0, \tau^2/6) \quad (RW2) \end{aligned}$$

The second representation can be interpreted as a least squares fit of the 2 (respectively 4) nearest neighbors of a coefficient β_ρ. Figure 1.4 illustrates this kind of penalization, figure 1.5 depicts a number of spline functions obtained from samples from the full conditional of β drawn during the MCMC simulation. The red line represents the actual estimation of the spline, which is taken to be the mean of a sufficiently high number of samples from the chain. Chapter 2 of this thesis is especially dedicated to the development of Bayesian P-splines.

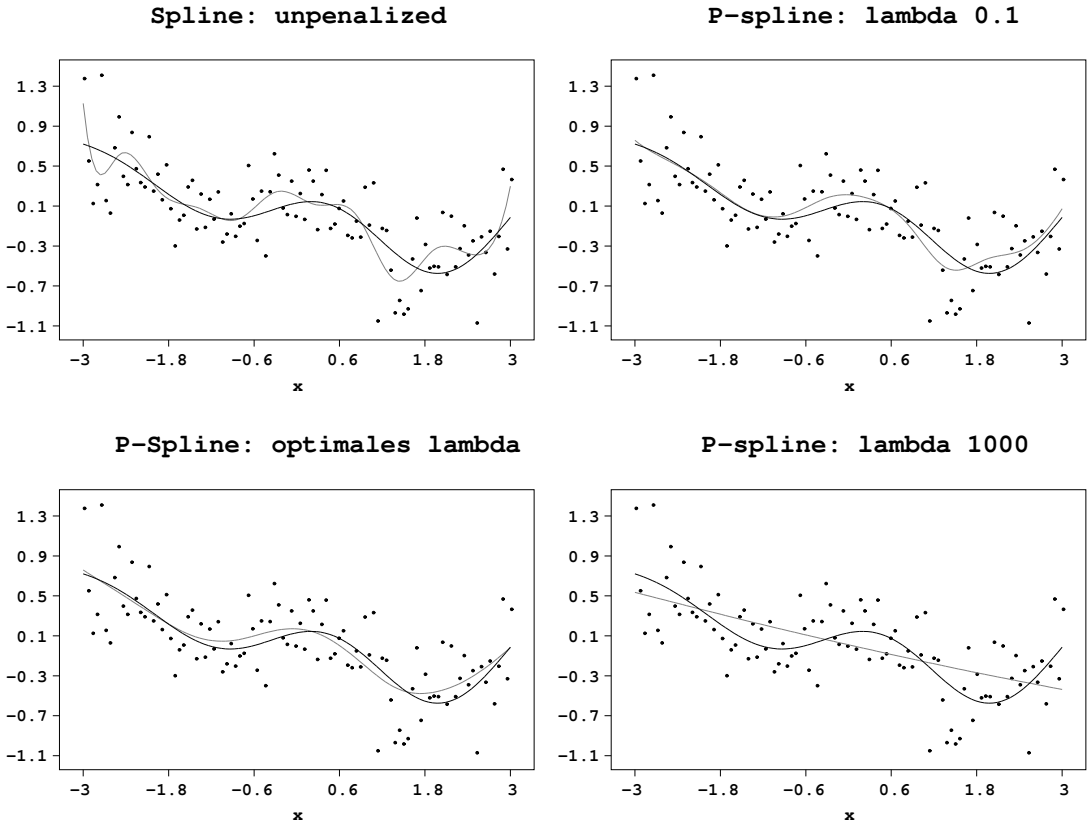


Figure 1.3: P-spline fit for different values of the smoothing parameter

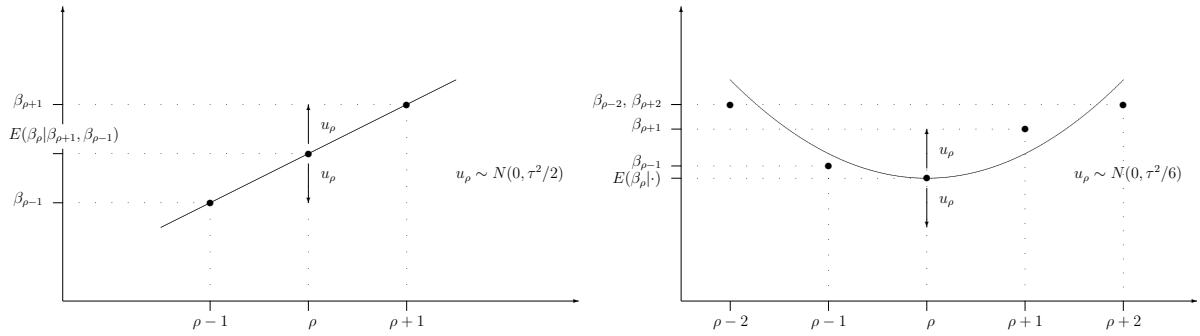


Figure 1.4: Illustration for first order (left) and second order (right) random walk penalty on regression coefficients.

A computationally very efficient implementation of penalized regression splines is provided by the package *mgcv* (Wood 2001) within the public domain software R. The original implementation is based on thin plate spline basis functions (Wood 2003) (in the latest version, however, there are different alternatives for choosing the basis functions). Thin plate splines can be seen as a multivariate generalization of smoothing splines. The objective is to find a minimizer of

$$\|y - f\|^2 + \lambda J_{kp}, \tag{1.4}$$

where

$$\begin{aligned} J_{kp} &= \int \cdots \int \|D^k f\|^2 dx_1 \ldots dx_p \\ &= \int \cdots \int \sum_{\nu_1+\ldots+\nu_p=k} \frac{k!}{\nu_1! \ldots \nu_p!} \left(\frac{\partial^k f}{\partial x_1^{\nu_1} \ldots \partial x_p^{\nu_p}} \right)^2 dx_1 \ldots dx_p \end{aligned} \tag{1.5}$$

and $\nu = (\nu_1, \ldots, \nu_p)'$ is a multi-index. Mostly, only the two dimensional case ($p = 2$) and second derivatives ($k = 2$) is of interest, in which case 1.5 simplifies to

$$\int \int \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2.$$

Note that for $k = 2, p = 1$ we retain expression (1.2) for one-dimensional smoothing splines. It can be shown (Duchon 1977) that the minimizing function of (1.4) has the expression

$$f(x) = \sum_{j=1}^{M} \alpha_j \phi_j(x) + \sum_{i=1}^{n} \beta_i \eta_{kp}(\|x - x_i\|)$$

provided that $2k > p$, where the $M = \begin{pmatrix} \text{k+p-1} \\ \text{p} \end{pmatrix}$ functions ϕ_j span the space of polynomials in $\mathbb{R}^p$. A thin plate spline basis is now the set

$$\{\phi_1, \ldots, \phi_M, \eta_{kp}(\|x - x_1\|), \ldots, \eta_{kp}(\|x - x_n\|)\}.$$

In the following we give some examples of one and two dimensional thin plate spline bases:

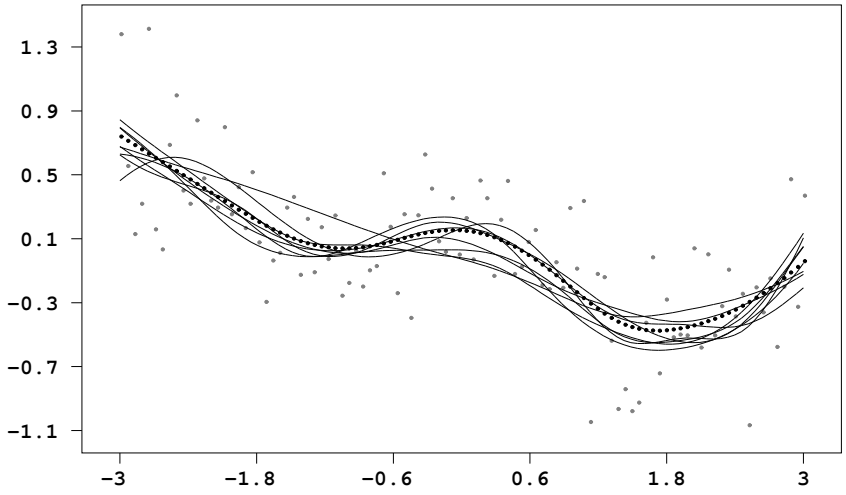


Figure 1.5: Visualization of the spline at different states of the Markov chain. The dotted line depicts the posterior mean.

$$p = 1,\ k = 1\text{: } \{1, x, |x - x_1|, \ldots, |x - x_n|\}$$

$$p = 1,\ k = 2\text{: } \{1, x, |x - x_1|^3, \ldots, |x - x_n|^3\}$$

$$p = 2,\ k = 2\text{: } \{1, x, x^2, \|x - x_1\|^2 \log(\|x - x_1\|), \ldots, \|x - x_n\|^2 \log(\|x - x_n\|)\}$$

The default value in *mgcv* is the lowest value satisfying $2k > p + 1$.

The characterizing feature of Wood's (2001) implementation is the simultaneous estimation of regression coefficients and smoothing parameters (compare also Wood 2000). The models are fitted by the common iteratively re-weighted least squares algorithm for GLMs, except that a penalized least squares problem is solved in each iteration instead of an ordinary least squares problem. Additionally, the smoothing parameters are estimated simultaneously by minimizing

$$\frac{\|W^{1/2}(z - X\hat{\beta}_\lambda)\|^2}{[tr(I - A)]^2}$$

with respect to λ, where z are the working observations and W is the weight matrix from the Fisher scoring algorithm. A is the hat matrix and $\hat{\beta}_\lambda$ the estimate of regression coefficients given λ.

Locally adaptive splines

Locally adaptive splines are an extension of the roughness penalty approaches in the sense that the global smoothing parameter λ is replaced by a locally varying smoothing parameter. This is particularly useful for highly oscillating functions and functions with changing curvature.

The most simple way of relaxing the assumption of a constant smoothing parameter to a locally adaptive one is to associate a different smoothing parameter with each spline

regression parameter, i.e. to replace $u_\rho \sim N(0, \tau^2)$ by $u_\rho \sim N(0, \tau^2/\delta_\rho)$, where δ_ρ are additional weight variables that have to be estimated. In a Bayesian approach the necessary prior for the additional parameters may be either independent or dependent. For simple random walk models Lang, Fronk and Fahrmeir (2002) use independent $IG(\nu/2, \nu/2)$ priors on $q_\rho^2 = 1/\delta_\rho$, which leads to a t-distribution with ν degrees of freedom for the marginal distribution of the errors. They also consider the alternative $u_\rho \sim N(0, exp(h_\rho))$ of dependent first or second order random walk priors

$$h_\rho = h_{\rho-1} + \tilde{u}_\rho \quad \text{or} \quad h_\rho = 2h_{\rho-1} - h_{\rho-2} + \tilde{u}_\rho, \quad \tilde{u}_\rho \sim N(0, \sigma_h^2).$$

Jerak and Lang (2005) generalize this approach to GAMs. Lang and Brezger (2004) (compare Chapter 2 of this thesis) propose to assign independent Gamma $G(1/2, 1/2)$ priors to the weights δ_ρ. This leads to a Cauchy distribution for the marginal distribution of the errors.

Ruppert and Carroll (2000) use a regression spline model

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_l x^l + \sum_{i=0}^{r} \beta_{l+i}(x - \xi_i)_+^l,$$

of degree $l \geq 1$ and define $\hat{\beta}$ as the minimizer of

$$\sum_{i=1}^{n} \{y_i - f(x)\} + \sum_{j=0}^{r} \lambda(\xi_j)\beta_{d+j}^2$$

The estimate for the smoothing parameter vector $\lambda = (\lambda(1), \ldots, \lambda(r))'$ is obtained by minimizing the GCV criterion using a smaller set of knots and smoothing parameters $\lambda^* = (\lambda^*(1), \ldots, \lambda^*(r^*))'$, $r^* < r$, where minimizing is done separately for each $\lambda^*(i)$ over a one dimensional grid centered at the optimal value for a global smoothing parameter (according to GCV) to avoid a full grid search. The estimate for the original λ vector is then obtained by linear interpolation. Ruppert and Carroll also give an algorithm for additive models.

Baladandayuthapani, Mallick and Carroll (2005) propose a fully Bayesian approach to locally adaptive (linear) Bayesian P-splines. They assign independent Gaussian priors on the regression parameters, with locally adaptive variances

$$\beta_i \sim N(0, \sigma_j^2(\xi_j)).$$

The functional form of $\sigma^2(\cdot)$ is again assumed to be a linear regression spline

$$-log(\sigma^2(x)) = \beta_0^* + \beta_1^* x + \sum_{i=1}^{r^*} \beta_{l+i}^*(x - \xi_i^*)_+,$$

on the log-scale to ensure positivity of $\sigma^2(\cdot)$.

1.1.3 Hybrid Splines

A procedure that combines the idea of adaptive regression splines and traditional smoothing splines is presented by Luo and Wahba (1997). They call their method hybrid splines to emphasize the conflating character.

Using the fact that the minimizer of (1.2) has a representation

$$f_\lambda(x) = \alpha_1\phi_1(x) + \alpha_2\phi_2(x) + \sum_{i=1}^{n} \beta_i R(x, x_i)$$

with certain functions ϕ_1, ϕ_2 and $R(x, x')$, they select the number of basis function by minimizing the GCV criterion in a first step. This is done by a forward procedure, where the basis function that gives the smallest GCV of all basis functions not yet included is added to the model. The second step of estimating the smoothing parameter λ is performed by the GCVPACK routine (Bates, Lindstrom, Wahba and Yandell 1987), which uses again the GCV. As Luo and Wahba (1997) point out, the first step is the dominating one for accounting for the bias-variance trade-off and the second one of choosing the smoothing parameter is merely a refinement of the result. The main benefit from the second step lies in the improved numerical stability of the procedure.

A Bayesian version of hybrid splines has been proposed recently by Dias and Gamerman (2002). In their approach they assign priors on the number of knots r on the smoothing parameter λ and on the variance parameter σ, but not on the spline parameters β. For inference they make use of the reversible jump MCMC algorithm of Green (1995) to sample from the full conditional distribution $p(r, \lambda, \sigma^2|y, \hat{\beta})$, while the estimate $\hat{\beta}$ is obtained by minimizing the penalized least squares criterion

$$PL(\lambda) = \|y - X_r\beta\|^2 + \lambda\beta'\Omega\beta$$

given a realization (r, λ, σ^2). Here, $f(x)$ is approximated by $X_r\beta$ and Ω is a $r \times r$ matrix with entries $\Omega_{ij} = \int B_i''(t)B_j''(t)dt$. The authors claim that a good guess of the hyperparameters of the priors for r and $\lambda|r$ makes the procedure quite fast, however, this might not hold for more complex functions, e.g. 2-dimensional surfaces.

1.2 P-Splines in Structured Additive Regression Models

Bayesian Generalized Linear Models (Fahrmeir and Tutz 2001) assume that, given covariates v and unknown parameters γ the distribution of the response y belongs to an exponential family with mean $\mu = E(y|v, \gamma)$ and a linear predictor η that is linked to the expectation of the response y by

$$\eta = v'\gamma = g(\mu),$$

where g is referred to as *link function*.

Table 1.1: Model classes included in STAR models as special cases. Expressions in brackets correspond to models for longitudinal data.

Model term	Predictor	Notation
fixed effects	$\ldots + v_i\gamma$ $[v_{it}\gamma]$	$r = i$ $[r = (i,t)]$
P-splines	$\ldots + f_1(x_{ij})$ $[f_1(x_{itj})]$	$r = i$, $\psi_{ij} = x_{ij}$ $[r = (i,t)$, $\psi_{itj} = x_{itj}]$
i.i.d. random effect	$\ldots + b_{ij}w_{itj}$	$r = (i,t)$, $\psi_{rj} = w_{ith}$, $f_j(\psi_{rj}) = b_{ji}w_{itj}$
VCM	$\ldots + g_1(x_{ij})z_{ij}$	$r = i$, $\psi_{rj} = (x_{ij}, z_{ij})$, $f_j(\psi_{rj}) = g_j(x_{ij})z_{ij}$
Geoadditive	$\ldots + f_{time}(t) + f_{str}(s_{it})$	$r = (i,t)$, $\psi_{rj} = t$ and $\psi_{rj'} = s_{it}$
2 dim. surface	$\ldots + f_{1\|2}(x_{i1}, x_{i2})$	$r = i$, $\psi_{rj} = (x_{i1}, x_{i2})$

In presence of features often found in practical applications, such as nonlinearity of continuous covariates, spatially or temporally correlated observation or unobserved unit or cluster specific heterogeneity, the above linear predictor is not appropriate for a comprehensive regression analysis. Therefore we replace the strictly linear predictor by an structured additive predictor

$$\eta_r = f_1(\psi_{r1}) + \ldots + f_p(\psi_{rp}) + v_r'\gamma \tag{1.6}$$

with a generic observation indicator r and a generic covariate notation ψ. Here, the functions f_j may comprise different types of functions of (not necessarily 1-dimensional) covariates. These may be nonlinear effects of continuous covariates, time trends and seasonal effects, varying coefficient terms, 2-dimensional surfaces, random slopes and intercepts, and spatially correlated random effects. Table 1.1 gives an overview of possible model terms together with the notation to cast the term into the general notation (1.6). This model formulation contains many regression models well known from the literature as special cases.

In a *Generalized additive model (GAM)* for cross-sectional data we have an additive predictor of the form

$$\eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + v_i'\gamma \tag{1.7}$$

for observation i, $i = 1, \ldots, n$. Here, the unknown functions f_j are smooth functions of a one-dimensional continuous covariate x_j, and are modeled by Bayesian P-splines in this thesis.

Given longitudinal data at time points $t \in \{t_1, t_2, \ldots\}$ for individuals i, $i = 1, \ldots, n$, we obtain a *Generalized additive mixed model (GAMM)* for longitudinal data from (1.7) by adding individual specific random effects, i.e.

$$\eta_{it} = f_1(x_{it1}) + \ldots + f_p(x_{itp}) + b_{i1}w_{it1} + \cdots + b_{iq}w_{itq} + v_{it}'\gamma.$$

Here, η_{it}, $x_{it1}, \ldots, x_{itp}$, $w_{it1}, \cdots, w_{itq}$, v_{it} are predictor and covariate values for individual i at time t, $b_{i1}, \ldots, b_{iq}$ are i.i.d. random intercepts (for $w_{tij} = 1$) or random slopes. Random effects are modeled by i.i.d. Gaussian priors. Note, that the assumption of the same time

points for each individual is only for simplicity of notation and may be generalized in a straightforward manner. GAMMs for cluster data can be written in the same general form.

The predictor for a *varying coefficient model (VCM)* as introduced by Hastie and Tibshirani (1993) is given by

$$\eta_i = g_1(x_{i1})z_{i1} + \ldots + g_p(x_{ip})z_{ip},$$

where the continuous covariates x_{ij} are the effect modifiers of the categorical or continuous interaction variables z_{ij}. Note, that in this work the effect modifiers need not to be continuous covariates, but may also comprise geographical locations (compare Fahrmeir, Lang, Wolff and Bender 2003). VCMs with spatially varying effect modifiers are well known under the synonym *geographically weighted regression*.

Consider the situation that we observe longitudinal data with additional geographic information for each observation. This spatio-temporal data can be accounted for by a *space-time main effect model* with the predictor

$$\eta_{it} = f_1(x_{it1}) + \ldots + f_p(x_{itp}) + f_{time}(t) + f_{str}(s_{it}) + v_{it}'\gamma,$$

see e.g. Fahrmeir and Lang (2001b). Here, $f_{time}(t)$ is a possibly nonlinear time trend that can be modeled by Bayesian P-splines, for example, and $f_{str}(s_{it})$ is modeled by Gaussian Markov random field, i.e. a correlated random effect for the spatial location of observation s_{it}, or by 2-dimensional P-splines based on the geographic coordinates of s_{it}. Kammann and Wand (2003) call models with predictors of this form *geoadditive models*.

If we are given two continuous covariates x_{i1} and x_{i2}, we may model an *ANOVA type interaction model* by a predictor of the form

$$\eta_i = f_1(x_{i1}) + f_2(x_{i2}) + f_{1|2}(x_{i1}, x_{i2}) + \cdots .$$

The main effects $f_1(x_{i1})$ and $f_2(x_{i2})$ are modeled by P-splines. The interaction effect $f_{1|2}(x_{i1}, x_{i2})$ is a 2-dimensional surface and may be modeled by tensor product P-splines. These will be introduced and discussed in the following chapter.

Throughout this thesis our main concern is on modeling continuous covariates with Bayesian P-splines. However, it should be emphasized that the scope of the thesis is always to embed the developed methodology in the rich class of the above described STAR models. This is facilitated through the implementation in the software package *BayesX* (Brezger, Kneib and Lang 2003). In the applications of this work we rely on many of the described features of STAR models.

Chapter 2

Bayesian P-Splines

In this chapter we introduce a Bayesian version of P-splines for the use as a smoothing technique in structured additive regression models. A Bayesian approach is chosen because of its enormous flexibility regarding different extensions already mentioned in the introduction. Furthermore, a Bayesian approach allows for simultaneous estimation of regression parameters and hyperparameters, e.g. smoothing parameters. P-splines are very attractive for several reasons. They facilitate a parsimonious yet flexible parametrization, and guarantee numerical stability and efficient implementation.

Part I of the chapter was originally published in a paper by Lang and Brezger (2004) under the title 'Bayesian P-Splines' in the *Journal of Computational and Graphical Statistics*. In this article only additive (Gaussian) models are considered. The focus is on developing Bayesian P-splines for one and two-dimensional smoothing, and for estimation of the functional form of the effect modifier in VCMs. Additionally, locally adaptive smoothing parameters are suggested. The employed MCMC inference methods are described in detail.

Part II consists of a preliminary version of the paper 'Generalized structured additive regression based on Bayesian P-Splines' by Brezger and Lang (2005) which is accepted for publication in the Journal of *Computational Statistics and Data Analysis*. In this paper the approach is generalized to non-Gaussian responses. A special emphasis lies on the elaborated MCMC sampling schemes, which provide fast convergence of the Markov chains. A data augmentation approach is used for binary and cumulative probit models. Iteratively weighted least squares proposals and joint updating of regression and smoothing parameters are presented for general exponential family responses.

In the presented version of the first article the full conditional distribution of the weights for the locally adaptive smoothing parameter in the 2-dimensional case is corrected. The second paper contains an additional section on a simulation study, that is not included in the version intended for publication. Note that both contributions are slightly revised to unify notation and to correct typos.

Bayesian P-Splines

Stefan Lang and Andreas Brezger
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

ABSTRACT

P-splines are an attractive approach for modeling nonlinear smooth effects of covariates within the additive and varying coefficient models framework. In this paper, we first develop a Bayesian version for P-splines and generalize in a second step the approach in various ways. First, the assumption of constant smoothing parameters can be replaced by allowing the smoothing parameters to be locally adaptive. This is particularly useful in situations with changing curvature of the underlying smooth function or with highly oscillating functions. In a second extension one dimensional P-splines are generalized to two dimensional surface fitting for modeling interactions between continuous covariates. In a last step the approach is extended to situations with spatially correlated responses allowing the estimation of geoadditive models. Inference is fully Bayesian and uses recent MCMC techniques for drawing random samples from the posterior. In a couple of simulation studies the performance of Bayesian P-splines is studied and compared to other approaches in the literature. We illustrate the approach by two complex applications on rents for flats in Munich and on human brain mapping.

Keywords: geoadditive models, locally adaptive smoothing parameters, MCMC, surface fitting, varying coefficient models

2.1 Introduction

Consider the *additive model* (AM) with predictor

$$E(y|x) = \eta = \gamma_0 + f_1(x_1) + \cdots + f_p(x_p)$$

where the mean of a continuous response variable y is assumed to be the sum of smooth functions f_j.

Several proposals are available for modeling and estimating the smooth functions f_j, see e.g. Fahrmeir and Tutz (2001, Ch. 5) and Hastie, Tibshirani and Friedman (2001) for an overview. An attractive approach, based on *penalized regression splines* (P-splines), has been presented by Eilers and Marx (1996). The approach assumes that the effect f of a covariate x can be approximated by a polynomial spline written in terms of a linear combination of B-spline basis functions. The crucial problem with such regression splines is the choice of the number and the position of the knots. A small number of knots may result in a function space which is not flexible enough to capture the variability of the data. A large number may lead to serious overfitting. Similarly, the position of the knots may potentially have a strong influence on estimation. A remedy can be based on a roughness penalty approach as proposed by Eilers and Marx (1996). To ensure enough flexibility a moderate number of equally spaced knots within the domain of x is chosen. Sufficient smoothness of the fitted curve is achieved through a difference penalty on adjacent B-spline coefficients. A different approach focuses on a parsimonious selection of basis functions and a careful selection of the position of the knots, see e.g. Friedman (1991).

This paper presents a Bayesian version of the P-splines approach by Eilers and Marx for AM's and extensions by replacing difference penalties with their stochastic analogues, i.e. Gaussian (intrinsic) random walk priors which serve as smoothness priors for the unknown regression coefficients. The approach generalizes work by Fahrmeir and Lang (2001a, b) based on simple random walk priors. A closely related approach based on a Bayesian version of smoothing splines can be found in Hastie and Tibshirani (2000), see also Carter and Kohn (1994) who choose state space representations of smoothing splines for Bayesian estimation with MCMC using the Kalman filter. Compared to smoothing splines, in a P-splines approach a more parsimonious parametrization is possible, which is of particular advantage in a Bayesian framework where inference is based on MCMC techniques. Other Bayesian approaches for nonparametric regression focus on adaptive knot selection and are close in spirit to the work by Friedman (1991). Denison et al. (1998) present an approach based on reversible jump MCMC for univariate curve fitting with continuous response which is extended to GAMs by Biller (2000) and Mallick, Denison and Smith (2000). A similar idea avoiding reversible jump MCMC is followed for Gaussian errors by Smith and Kohn (1996). Hansen and Kooperberg (2002) discuss adaptive knot selection for the very broad class of extended linear models. Di Matteo et al. (2001) present an approach for GAMs where knots are selected on a continuous proposal distribution rather than a discrete set of candidate knots as in the other approaches.

In further steps, we extend and generalize our approach in various ways. First, the assumption of global smoothing parameters can be replaced by *locally adaptive smoothing*

parameters to improve the estimation of functions with changing curvature. Such situations have attained considerable attention in the recent literature, see e.g. Luo and Wahba (1997) and Ruppert and Carroll (2000). Locally adaptive smoothing parameters are incorporated by replacing the usual Gaussian prior for the regression parameters by a Cauchy distribution. Such a prior has been already used in the context of dynamic models (Knorr–Held 1999) and for edge preserving spatial smoothing (e.g. Besag and Higdon 1999).

In a second step, we generalize the P-spline approach for one dimensional curves to two dimensional surface fitting by assuming that the unknown surface can be approximated by the tensor product of one dimensional B-splines. Smoothness is now achieved by smoothness priors common in spatial statistics, e.g. two dimensional generalizations of random walks. Once again, global smoothing parameters may be replaced by spatially adaptive smoothing parameters. We demonstrate the benefit of spatially adaptive smoothing parameters in our second application on human brain mapping. Another Bayesian approach for bivariate curve fitting based on adaptive knot selection has been developed by Smith and Kohn (1997).

In a last step, the classical AM is extended to *additive mixed models* to deal with unobserved heterogeneity among units or clusters. A main focus is thereby on *spatially correlated random effects*. Kammann and Wand (2003) call models with an additive predictor composed of nonlinear functions of continuous covariates and spatial effects *geoadditive models*. We will present an application of such a geoadditive model in our first data application on rents for flats in Munich. Additive mixed models (without spatially correlated random effects) have been considered in a Bayesian framework by Hastie and Tibshirani (2000), geoadditive models have also been developed by Fahrmeir and Lang (2001a, b).

Bayesian inference is based on a Gibbs sampler to update the full conditionals of the regression parameters and variances. Numerical efficiency is guaranteed by matrix operations for band matrices (Rue 2001) or sparse matrices (George and Liu 1981).

Most of the methodology of this paper is implemented in *BayesX* a software package for Bayesian inference based on MCMC techniques. The program is available free of charge at `http://www.stat.uni-muenchen.de/~lang/bayesx`.

The rest of this paper is organized as follows: Section 2.2 describes Bayesian AMs with P-splines and extensions. Section 2.3 gives details about MCMC inference for the proposed models. Section 2.4 contains extensive simulation studies in order to gain more insight into the practicability and the limitations of our approach and to compare it with other techniques in the literature. In Section 2.5, the methods of this paper are applied to complex data sets on rents for flats in Munich and on human brain mapping.

2.2 Bayesian AMs and extensions based on P-Splines

2.2.1 Additive models

Consider regression situations where observations (y_i, x_i, v_i), $i = 1, \ldots, n$, on a continuous response y, a vector of continuous covariates $x = (x_1, \ldots, x_p)'$ and a vector of further

covariates $v = (v_1, \ldots, v_q)'$ are given. Given covariates and unknown parameters, we assume that the responses y_i, $i = 1, \ldots, n$, are independent and Gaussian with mean or predictor

$$\eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + v_i'\gamma. \tag{2.1}$$

and a common variance σ^2 across subjects. Here $f_1, \ldots, f_p$ are unknown smooth functions of the continuous covariates. The linear combination $v_i'\gamma$ corresponds to the usual parametric part of the predictor.

Note that the mean levels of the unknown functions f_j are not identifiable. To ensure identifiability, the functions f_j are constrained to have zero means, i.e.

$$1/range(x_j) \int f_j(x_j)dx_j = 0.$$

This can be incorporated into estimation via MCMC by centering the functions f_j about their means in every iteration of the sampler. To avoid, that the posterior is changed the subtracted means are added to the intercept (included in $v_i'\gamma$).

In the P-splines approach by Eilers and Marx (1996), it is assumed that the unknown functions f_j can be approximated by a spline of degree l with equally spaced knots $x_{j,min} = \zeta_{j0} < \zeta_{j1} < \cdots < \zeta_{j,r_j-1} < \zeta_{jr_j} = x_{j,max}$ within the domain of x_j. It is well known that such a spline can be written in terms of a linear combination of $M_j = r_j + l$ B-spline basis functions $B_{j\rho}$, i.e.

$$f_j(x_j) = \sum_{\rho=1}^{M_j} \beta_{j\rho} B_{j\rho}(x_j).$$

For the ease of notation, we assume the same number of knots $M = M_j$ for every function f_j. The basis functions $B_{j\rho}$ are defined only locally in the sense that they are nonzero only on a domain spanned by $2 + l$ knots. It would be beyond the scope of this paper to go into the details of B-splines and their properties, see De Boor (1978) as a key reference. By defining the $n \times M$ design matrices X_j, where the element in row i and column ρ is given by $X_j(i, \rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor (2.1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V'\gamma. \tag{2.2}$$

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})'$, $j = 1, \ldots, p$, correspond to the vectors of unknown regression coefficients. The matrix V is the usual design matrix of fixed effects. In a simple regression spline approach the unknown regression coefficients are estimated using standard maximum likelihood algorithms for linear models. To overcome the difficulties of regression splines, already mentioned in the introduction, Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent B-Spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to penalized likelihood estimation where the penalized likelihood

$$L = l(y, \beta_1, \ldots, \beta_p, \gamma) - \lambda_1 \sum_{l=k+1}^{M} (\Delta^k \beta_{1l})^2 - \cdots - \lambda_p \sum_{l=k+1}^{M} (\Delta^k \beta_{pl})^2 \tag{2.3}$$

is maximized with respect to the unknown regression coefficients $\beta_1, \ldots, \beta_p$ and γ. In (2.3) Δ^k denotes the difference operator of order k. In this paper we restrict ourselves to penalties based on first and second differences, i.e. $k = 1$ or $k = 2$. Estimation can be carried out with backfitting (Hastie and Tibshirani 1990) or by direct maximization of the penalized likelihood (Marx and Eilers 1998). The trade off between flexibility and smoothness is determined by the smoothing parameters λ_j, $j = 1, \ldots, p$. Typically "optimal" smoothing parameters are estimated via cross validation or by minimizing the AIC with respect to the λ_j, $j = 1, \ldots, p$. However, these procedures often fail in practice since no optimal solutions for the λ_j can be found (see also Section 2.4.1). More severe is the fact that these criteria fail to work if the number of smooth functions in the model is large as then the computational effort to compute an optimal solution (if there is any) becomes intractable. However, a computational efficient algorithm for computing the smoothing parameters has been presented recently by Wood (2000), which seems to work at least for a moderate number of smoothing parameters.

In a Bayesian approach unknown parameters β_j, $j = 1, \ldots, p$, and γ are considered as random variables and have to be supplemented with appropriate prior distributions.

For the fixed effects parameters γ we assume independent diffuse priors, i.e. $\gamma_j \propto const$, $j = 1, \ldots, q$.

Priors for the regression parameters β_j of nonlinear functions are defined by replacing the difference penalties in (2.3) by their stochastic analogues. First differences correspond to a first order random walk and second differences to a second order random. Thus, we obtain

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{2.4}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. Note, that the priors in (2.4) could have been equivalently defined by specifying the conditional distributions of a particular parameter $\beta_{j\rho}$ given its *left* and *right* neighbors. Then, the conditional means may be interpreted as locally linear or quadratic fits at the knot positions $\zeta_{j\rho}$. The amount of smoothness is controlled by the additional variance parameters τ_j^2, which correspond to the smoothing parameters λ_j in the classical approach. The priors (2.4) can be equivalently written in the form of global smoothness priors

$$\beta_j | \tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right) \tag{2.5}$$

with appropriate penalty matrix K_j. Since K_j is rank deficient with $rank(K_j) = M - 1$ for a first order random walk and $rank(K_j) = M - 2$ for a second order random walk, the prior (2.5) is improper.

For full Bayesian inference, the unknown variance parameters τ_j^2 are also considered as random and estimated simultaneously with the unknown β_j. Therefore, hyperpriors are assigned to the variances τ_j^2 (and the overall variance parameter σ^2) in a further stage of the hierarchy by highly dispersed (but proper) inverse Gamma priors $p(\tau_j^2) \sim IG(a_j, b_j)$. The prior for τ_j^2 must not be diffuse in order to obtain a proper posterior for β_j, see Hobert and Casella (1996) for the case of linear mixed models. A common choice for

the hyperparameters is $a_j = 1$ and a small value for b_j, e.g. $b_j = 0.005$, $b_j = 0.0005$ or $b_j = 0.00005$, leading to almost diffuse priors for τ_j^2.

The amount of smoothness allowed by a particular prior specification depends (weakly) on the scale of the responses. To avoid the problem, we standardize the vector of responses y before estimation and retransform the results afterwards. Standardizing the responses is also important to avoid numerical difficulties with MCMC simulations.

In some situations, the estimated nonlinear functions f_j may considerably depend on the particular choice of hyperparameters a_j and b_j. This may be the case for very low signal to noise ratios or/and small sample sizes. It is therefore highly recommended to estimate all models under consideration using a (small) number of *different* choices for a_j and b_j to assess the dependence of results on minor changes in the model assumptions. In that sense, the variation of hyperparameters can be used as a tool for model diagnostics. More details on the dependency of results from the hyperparameters are given in our simulation studies in Section 2.4.

In some applications, the assumption of global variances τ_j^2 (or smoothing parameters) may be inappropriate, e.g. when the underlying functions are highly oscillating. In such situations, we can replace the errors $u_{j\rho} \sim N(0, \tau_j^2)$ in (2.4) by $u_{j\rho} \sim N(0, \frac{\tau_j^2}{\delta_{j\rho}})$ where the weights $\delta_{j\rho}$ are additional hyperparameters. We assume that the weights $\delta_{j\rho}$ are independent and Gamma distributed $\delta_{j\rho} \sim G(\frac{1}{2}, \frac{1}{2})$. This implies that $\beta_{j\rho}|\beta_{j\rho-1}$ or $\beta_{j\rho}|\beta_{j\rho-1}, \beta_{j\rho-2}$ follow a Cauchy distribution which has heavier tails than the normal distribution.

2.2.2 Modeling interactions

The models considered so far are not appropriate for modeling interactions between covariates. A common way to deal with interactions are varying coefficient models (VCM) introduced by Hastie and Tibshirani (1993). Here nonlinear terms $f_j(x_{ij})$ are generalized to $f_j(x_{ij})z_{ij}$, where z_j may be a component of x or v or a further covariate. The predictor (2.1) is replaced by

$$\eta_i = f_1(x_{i1})z_{i1} + \cdots + f_p(x_{ip})z_{ip} + v_i'\gamma.$$

Covariate x_j is called the effect modifier of z_j because the effect of z_j varies smoothly over the range of x_j. For $z_{ij} \equiv 1$ we obtain the AM as a special case. Estimation of VCMs poses no further difficulties, since only the design matrices X_j in (2.2) have to be redefined by multiplying each element in row i of X_j with z_{ij}.

VCMs are particularly useful if the interacting variable z_j is categorical. Consider now situations where both interacting covariates are continuous. In principal, interactions between continuous covariates could be modeled via VCMs as well. Note, however, that we model a very special kind of interaction since one of both covariates still enters linearly into the predictor. A more flexible approach is based on (nonparametric) two dimensional surface fitting. In this case, the interaction between two covariates x_j and x_s is modeled by a two dimensional smooth surface $f_{js}(x_j, x_s)$ leading to a predictor of the form

$$\eta_i = \cdots + f_j(x_{ij}) + f_s(x_{is}) + f_{js}(x_{ij}, x_{is}) + \cdots \quad .$$

Here we assume that the unknown surface can be approximated by the tensor product of the two one dimensional B-splines, i.e.

$$f_{js}(x_j, x_s) = \sum_{\rho=1}^{M} \sum_{\nu=1}^{M} \beta_{js\rho\nu} B_{j\rho}(x_j) B_{s\nu}(x_s).$$

Similar to the one dimensional case, additional identifiability constraints have to be imposed on the functions f_j, f_s and f_{js}. Following Chen (1993) or Stone et al. (1997), we impose the constraints

$$\begin{aligned}
\bar{f}_j &= \frac{1}{range(x_j)} \int f_j(x_j) dx_j = 0 \\
\bar{f}_s &= \frac{1}{range(x_s)} \int f_s(x_s) dx_s = 0, \\
\bar{f}_{js}(x_j) &= \frac{1}{range(x_s)} \int f_{js}(x_j, x_s) dx_s = 0 \text{ for all distinct values of } x_j, \\
\bar{f}_{js}(x_s) &= \frac{1}{range(x_j)} \int f_{js}(x_j, x_s) dx_j = 0 \text{ for all distinct values of } x_s, \text{ and} \\
\bar{f}_{js} &= \frac{1}{range(x_j) \cdot range(x_s)} \int \int f_{js}(x_j, x_s) dx_j dx_s = 0.
\end{aligned}$$

This is achieved in an MCMC sampling scheme by appropriately centering the functions in every iteration. More specifically, we first compute the centered function f_{js}^c by $f_{js}^c(x_{ij}, x_{is}) = f_{js}(x_{ij}, x_{is}) - \bar{f}_{js}(x_j) - \bar{f}_{js}(x_s) + \bar{f}_{js}$. In order to ensure that the posterior is unchanged, we proceed by adding $\bar{f}_{js}(x_j)$ and $\bar{f}_{js}(x_s)$ to the respective main effects and subtracting $\bar{f}_{js}$ from the intercept. In the last step, the main effects are centered in the same way as described above.

Priors for $\beta_{js} = (\beta_{js11}, \ldots, \beta_{jsMM})'$ are based on spatial smoothness priors common in spatial statistics (see e.g. Besag and Kooperberg 1995). Since there is no natural ordering of parameters, priors have to be defined by specifying the conditional distributions of $\beta_{js\rho\nu}$ given neighboring parameters and the variance component τ_{js}^2. The most commonly used prior specification based on the four nearest neighbors can be defined by

$$\beta_{js\rho\nu}|\cdot \sim N\left(\frac{1}{4}(\beta_{js\rho-1,\nu} + \beta_{js\rho+1,\nu} + \beta_{js\rho,\nu-1} + \beta_{js\rho,\nu+1}), \frac{\tau_{js}^2}{4}\right) \tag{2.6}$$

for $\rho, \nu = 2, \ldots, M-1$ and appropriate changes for corners and edges. For example, for the upper left corner we obtain $\beta_{js11}|\cdot \sim N(\frac{1}{2}(\beta_{js12} + \beta_{js21}), \frac{\tau_{js}^2}{2})$. For the left edge, we get $\beta_{js1\nu}|\cdot \sim N(\frac{1}{3}(\beta_{js1,\nu+1} + \beta_{js1,\nu-1} + \beta_{js2,\nu}), \frac{\tau_{js}^2}{3})$. This prior is a direct generalization of a first order random walk in one dimension. Its conditional mean can be interpreted as a least squares locally linear fit at knot position (ζ_ρ, ζ_ν) given the neighboring parameters.

Another choice for a prior for β_{js} can be based on the Kronecker product $K_{js} = K_j \otimes K_s$ of penalty matrices of the main effects, see Clayton (1996) for a justification. We prefer (2.6) because the priors based on Kronecker products tend to overfitting (at least in the context of spline smoothing). Note, that all priors for two dimensional smoothing can be easily brought into the general form (2.5).

Prior (2.6) can be generalized to allow for spatially adaptive variance parameters. For that reason, we introduce weights $\delta_{(\rho\nu)(kl)}$ with the requirement that $\delta_{(\rho\nu)(kl)} = \delta_{(kl)(\rho\nu)}$ and generalize (2.6) to

$$\beta_{js\rho\nu}|\cdot \sim N\left(\sum_{(kl)\in\partial_{(\rho\nu)}} \frac{\delta_{(\rho\nu)(kl)}}{\delta_{(\rho\nu)+}}\beta_{jskl}, \frac{\tau^2_{\rho\nu}}{\delta_{(\rho\nu)+}}\right). \tag{2.7}$$

Here, $\partial_{\rho\nu}$ corresponds to the set of neighboring knots to ζ_ρ, ζ_ν and $\delta_{(\rho\nu)+}$ denotes the sum of weights $\sum_{(kl)\in\partial_{(\rho\nu)}} \delta_{(\rho\nu)(kl)}$. For $\delta_{(\rho\nu)(kl)} = 1$, we obtain (2.6) as a special case. Introducing hyperpriors for the weights $\delta_{(\rho\nu)(kl)}$ in a further stage of the hierarchy we get a smoothness prior with spatially adaptive variances. In analogy to the one dimensional case, we assume that the $\delta_{(\rho\nu)(kl)}$ are independent and Gamma distributed $\delta_{(\rho\nu)(kl)} \sim G(\frac{1}{2}, \frac{1}{2})$.

2.2.3 Geoadditive models

In a number of applications responses depend not only on continuous and categorical covariates but also on the *spatial location* where they have been observed. For example, in our application on rents for flats in Munich, the monthly rent considerably depends on the location in the city. In this and various other applications, models are needed which are able to deal simultaneously with nonlinear effects of continuous covariates and nonlinear spatial effects.

To consider the spatial variation of responses, we can add an additional *spatial effect* f_{spat} to the predictor (2.2) leading to *geoadditive models* (Kammann and Wand 2003). Depending on the application, the spatial effect may be further split up into a spatially correlated (structured) and an uncorrelated (unstructured) effect, i.e. $f_{spat} = f_{str} + f_{unstr} = X_{str}\beta_{str} + X_{unstr}\beta_{unstr}$. A rational is that a spatial effect is usually a surrogate of many unobserved influential factors, some of them may obey a strong spatial structure while others may exist only locally. By estimating a structured and an unstructured effect, we aim at separating between the two kinds of influential factors. For data observed on a regular or irregular lattice a common approach for the correlated spatial effect f_{str} is based on Markov random field priors for the regression coefficients β_{str}, e.g. Besag, York and Mollie (1991). Let $s \in \{1, \ldots, S\}$ denote the pixels of a lattice or regions of a geographical map. Then, the most simple Markov random field prior for $\beta_{str} = (\beta_{str,1}, \ldots, \beta_{str,S})$ is defined by

$$\beta_{str,s}|\beta_{str,u}, u \neq s \sim N\left(\sum_{u\in\partial_s} \frac{1}{N_s}\beta_{str,u}, \frac{\tau^2_{str}}{N_s}\right), \tag{2.8}$$

where N_s is the number of adjacent regions or pixels, and ∂_s denotes the regions which are neighbors of region s. Hence, prior (2.8) can be seen as a 2 dimensional extension of a first

order random walk. More general priors than (2.8) are described in Besag et al. (1991). The design matrix X_{str} is a $n \times S$ incidence matrix whose entry in the i-th row and s-th column is equal to one if observation i has been observed at location s and zero otherwise.

Alternatively, we could use two dimensional surface estimators as described in Section 2.2.2 to model the structured spatial effect f_{str}.

For the uncorrelated effect, we assume i.i.d. Gaussian random effects for β_{unstr}, i.e.

$$\beta_{unstr}(s) \sim N(0, \tau^2_{unstr}), \quad s = 1, \ldots, S. \tag{2.9}$$

Formally, the priors for β_{str} and β_{unstr} can both be brought into the form (2.5). For β_{str}, the elements of K are given by $k_{ss} = N_s$, $k_{su} = -1$ if $u \in \partial_s$ and 0 else. For β_{unstr}, we may set $K = I$.

Again, for τ^2_{str} and τ^2_{unstr} we assume inverse Gamma priors $\tau^2_{str} \sim IG(a_{str}, b_{str})$ and $\tau^2_{unstr} \sim IG(a_{unstr}, b_{unstr})$.

2.3 Posterior inference via MCMC

Bayesian inference is based on the posterior of the model, which is analytically intractable. Therefore, inference is carried out by recent Markov chain Monte Carlo (MCMC) simulation techniques.

For the ease of notation, we subsume for the rest of this paper two dimensional surfaces f_{js} into the functions f_j, $j = 1, \ldots, p$, so that a function f_j may also be a two dimensional function of covariates x_j and x_s. For the following let α denote the vector of all parameters appearing in the model. Under usual conditional independence assumptions for the parameters the posterior is given by

$$\begin{aligned} p(\alpha) \quad \propto \quad & L(y, \beta_1, \ldots, \beta_p, \beta_{str}, \beta_{unstr}, \gamma, \sigma^2) \prod_{j=1}^{p} \left(p(\beta_j|\tau^2_j) p(\tau^2_j) \right) \\ & p(\beta_{str}|\tau^2_{str}) p(\tau^2_{str}) p(\beta_{unstr}|\tau^2_{unstr}) p(\tau^2_{unstr}) p(\gamma) p(\sigma^2) \end{aligned} \tag{2.10}$$

where $L(\cdot)$ denotes the likelihood which is the product of individual likelihood contributions. If a locally adaptive variance parameter is assumed for one of the smooth functions f_j, the term $p(\beta_j|\tau^2_j)p(\tau^2_j)$ in the first line of (2.10) must be replaced by $p(\beta_j|\delta_j, \tau^2_j)p(\delta_j)p(\tau^2_j)$. Because the individual weights are assumed to be independent, the prior $p(\delta_j)$ is a product of Gamma densities.

MCMC simulation is based on drawings from full conditionals of blocks of parameters given the other parameters and the data. It can be shown that the full conditionals for β_j, $j = 1, \ldots, p$, β_{str}, β_{unstr} and γ are multivariate Gaussian. Straightforward calculations show that the precision matrix P_j and the mean m_j of $\beta_j|\cdot$ are given by

$$P_j = \frac{1}{\sigma^2} X_j' X_j + \frac{1}{\tau^2_j} K_j, \quad m_j = P_j^{-1} \frac{1}{\sigma^2} X_j' (y - \tilde{\eta}), \tag{2.11}$$

where $\tilde{\eta}$ is the part of the predictor associated with all remaining effects in the model. Because of the special structure of the design matrices X_j and the penalty matrices K_j, the posterior precisions P_j are band matrices. For a one dimensional P-spline, the bandwidth of P_j is the maximum between the degree l of the spline and the order of the random walk. For a two dimensional P-spline, the bandwidth is $M \cdot l + l$.

Following Rue (2001), drawing random numbers from $p(\beta_j|\cdot)$ is as follows: We first compute the Cholesky decomposition $P_j = LL'$. We proceed by solving $L'\beta_j = z$, where z is a vector of independent standard Gaussians. It follows that $\beta_j \sim N(0, P_j^{-1})$. We then compute the mean m_j by solving $P_j m_j = \frac{1}{\sigma^2} X_j'(y - \tilde{\eta})$. This is achieved by first solving $L\nu = \frac{1}{\sigma^2} X_j'(y - \tilde{\eta})$ by forward substitution followed by backward substitution $L'm_j = \nu$. Finally, adding m_j to the previously simulated β_j yields $\beta_j \sim N(m_j, P_j^{-1})$. The algorithms involved take advantage of the band matrix structure of the posterior precision P_j.

The precision matrix and the mean of the full conditionals for the regression coefficients β_{str} and β_{unstr} of the spatial effect f_{spat} can be formally brought into the form (2.11). The posterior precision matrix for β_{unstr} is diagonal whereas the precision matrix for β_{str} is usually neither a diagonal nor a band matrix but a sparse matrix. However, the regions of a geographical map can be reordered using the *reverse Cuthill-McKee algorithm* (George and Liu 1981) to obtain a band matrix. In contrast to posterior precision matrices of P-splines, the band size usually differs from row to row. This can be exploited to further improve the computational efficiency. In our implementation, we use the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981). Our experience shows that the speed of computations improves up to 25% by using the envelope method rather than simple matrix operations for band matrices.

Regarding the fixed effects parameters γ, we obtain for the precision matrix and the mean

$$P_\gamma = \frac{1}{\sigma^2} V'V, \quad m_\gamma = (V'V)^{-1}V'(y - \tilde{\eta}).$$

The full conditionals for the variance parameters τ_j^2, $j = 1, \ldots, p$, τ_{str}^2, τ_{unstr}^2 and σ^2 are all inverse Gamma distributions with parameters

$$a_j' = a_j + \frac{rank(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\beta_j' K_j \beta_j$$

for τ_j^2, τ_{str}^2 and τ_{unstr}^2. For σ^2 we obtain

$$a_\sigma' = a_\sigma + \frac{n}{2} \quad \text{and} \quad b_\sigma' = b + \frac{1}{2}\epsilon'\epsilon$$

where ϵ is the usual vector of residuals. If for some of the functions f_j locally adaptive variances are assumed, we additionally need to compute the full conditionals for the weights $\delta_{j\rho}$, or $\delta_{(\rho\nu)(kl)}$. For one dimensional P-splines with a first or second order random walk penalty the full conditionals for the weights $\delta_{j\rho}$ are Gamma distributed with parameters

$$a_{\delta_{j\rho}}' = \frac{\nu}{2} + \frac{1}{2} \quad \text{and} \quad b_{\delta_{j\rho}}' = \frac{\nu}{2} + \frac{u_{j\rho}^2}{2\tau_j^2}$$

where $u_{j\rho}$ is the error term in (2.4). Because all full conditionals involved are known distributions, a simple Gibbs sampler can be used to successively update the parameters of the model.

In the case of a two dimensional P-spline, the full conditionals for the weights $\delta_{(\rho\nu)(kl)}$ are proportional to

$$\begin{aligned} p(\delta_{(\rho\nu)(kl)}|\cdot) &\propto p(\beta_{js}|\cdot)p(\delta_{(\rho\nu)(kl)}) \\ &\propto \left(\prod_{i=2}^{M^2}\lambda_i\right)^{1/2} \delta_{(\rho\nu)(kl)}^{\frac{\nu}{2}-1} \exp\left(-\delta_{(\rho\nu)(kl)}\left[\frac{\nu}{2}+\frac{(\beta_{js\rho\nu}-\beta_{jskl})^2}{2\tau_{js}^2}\right]\right) \end{aligned} \quad (2.12)$$

where λ_i, $i = 2, \ldots, M^2$, are the non-zero eigenvalues of $K_{js}(\delta)$. We explicitly denote the penalty matrix by $K_{js}(\delta)$ to emphasize its dependency on the weights δ. Note that (2.12) is a Gamma $G(a'_{\delta_{(\rho\nu)(kl)}}, b'_{\delta_{(\rho\nu)(kl)}})$ density with parameters

$$a'_{\delta_{(\rho\nu)(kl)}} = \frac{\nu}{2} \quad \text{and} \quad b'_{\delta_{(\rho\nu)(kl)}} = \frac{\nu}{2} + \frac{(\beta_{js\rho\nu}-\beta_{jskl})^2}{2\tau_{js}^2}, \quad (2.13)$$

multiplied by $\left(\prod_{i=2}^{M^2}\lambda_i\right)^{1/2}$. In order to sample from this distribution we employ a MH-step and use a Gamma distribution with the parameters in (2.13) as proposal density. Therefore the acceptance probability reduces to

$$\alpha = \min\left\{1, \left(\frac{\prod_{i=2}^{M^2}\lambda_i^*}{\prod_{i=2}^{M^2}\lambda_i}\right)^{1/2}\right\},$$

where the λ_i^* denote the non-zero eigenvalues of the penalty matrix $K_{js}(\delta^*_{(\rho\nu)(kl)})$ resulting from a proposed weight $\delta^*_{(\rho\nu)(kl)}$. Acceptance rates are usually quite high. In our implementation we use the fact that

$$\frac{\prod_{i=2}^{M^2}\lambda_i^*}{\prod_{i=2}^{M^2}\lambda_i} = \frac{|K_{11}^*|}{|K_{11}|}, \quad (2.14)$$

where $|K_{11}|$ and $|K_{11}^*|$ denote the determinant of the sub-matrices of $K_{js}(\delta_{(\rho\nu)(kl)})$ and $K_{js}(\delta^*_{(\rho\nu)(kl)})$, respectively, where the last row and the last column is deleted. The advantage arising from (2.14) is, that instead of an expensive computation of eigenvalues of order $O(n^3)$, the ratio can be obtained by the computationally much more efficient Cholesky decomposition of band matrices, which is of order $O(n)$. Additionally, we exploit the fact that it is sufficient to start the Cholesky decomposition in the row corresponding to the position where a proposed new weight is located. Block updating of several weights in one step speeds up computation considerably, since the ratio (2.14) has to be evaluated only once per block. Since the proposal densities $p(\delta_{(\rho\nu)(kl)})$ are independent no further difficulties are imposed by sampling from a block of weights. In our application to human brain mapping in Subsection 2.5.2 joint updating of 6 weights still yields an acceptance rate $> 50\%$.

A proof of (2.14) can be found in Appendix A.1.

2.4 Simulations

In this section we present a couple of simulation studies mainly to compare the proposed methodology with related approaches in the literature. The main focus of Section 2.4.1 lies on functions with low or moderate curvature while Section 2.4.2 deals with the estimation of highly oscillating functions. Finally, Section 2.4.3 compares some surface estimators where we mainly refer to Smith and Kohn (1997) who compare their approach with the most common surface estimators in the literature.

2.4.1 Functions with moderate curvature

The main focus of this section is on functions with low or moderate curvature. We considered three functions, a linear one ($f_1(x) = 1.0/1.758x$), a quadratic one ($f_2(x) = 1.0/2.75x^2 - 1.5$) and a sinusoidal one ($f_3(x) = 1.0/0.72sin(x)$). The values of x were chosen on an equidistant grid of $n = 100$ design points between -3 and 3. To assess the dependence of results on the curvature, we scaled the three functions such that the standard deviations $\sigma(f_j)$, $j = 1, 2, 3$, of f_j are all equal to one. For the overall variance parameter σ^2, we chose the values $\sigma = 1, 0.5, 0.33$ which corresponds to a very low, low and medium signal to noise ratio. Figure 2.1 a) - c) shows typical data sets for the sinusoidal function f_3 with the different signal to noise ratios. We simulated 250 replications for every function and variance σ^2 and applied and compared the following estimators:

- Bayesian cubic P-splines with second order random walk penalty and 20 knots. We estimated the models with three different choices for the hyperparameters a and b of the variance τ^2 to assess the dependence of results on the hyperparameters. We used $a = 1, b = 0.005$, $a = 1, b = 0.0005$ and $a = 1, b = 0.00005$.

- Classical (cubic) P-splines with second order difference penalty and 20 knots. Estimation was carried out using the GAM object of S-Plus 4.0 and the P-spline function for GAM objects provided by Brian Marx. The function is available at http://www.stat.lsu.edu/bmarx/. The smoothing parameters were estimated by cross validation where the optimal smoothing parameter was chosen on a geometrical grid of 30 knots between 10^4 and 10^{-4}.

- Adaptive Bayesian regression splines by Biller (2000) as an example of a competing Bayesian approach. Estimation was carried out using the program 'bvcm' which is available at http://www.stat.uni-muenchen.de/sfb386/. The number of knots k are assumed be Poisson distributed with mean $\bar{k}$ restricted to the set $k \in \{4, \ldots, k_{max} = 50\}$. We used $\bar{k} = 20$ which is the default in the program. Experiments with $\bar{k} = 10$ or $\bar{k} = 30$ showed no substantial differences to the findings below.

The performance of the estimators is measured by the empirical mean squared error given by $MSE(\hat{f}) = 1/n \sum_{i=1}^{n} (f(x_i) - \hat{f}(x_i))^2$.

Figure 2.2 displays boxplots of log(MSE) for very low (first column), low (second column) and medium (third column) signal to noise ratio (SNR), respectively. The first row

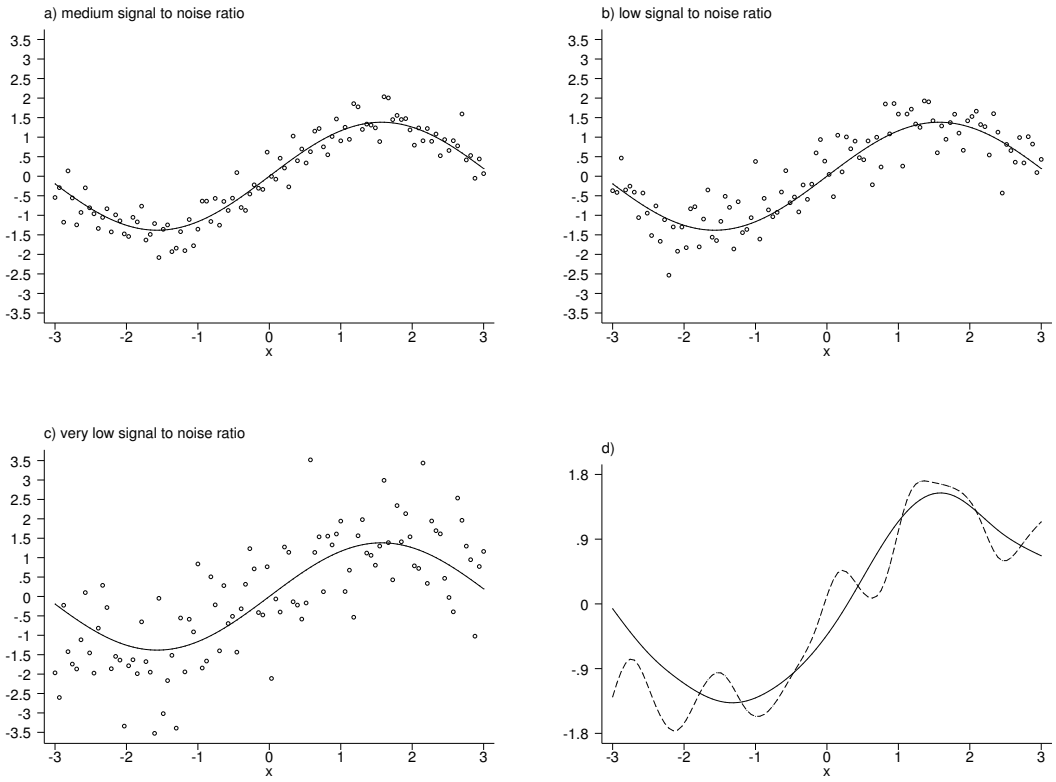


Figure 2.1: *Sinusoidal function of simulation study 1: The graphs a)-c) show typical data sets for medium, low and very low signal to noise ratio. The true function is included in the graphs (solid lines). Panel d) displays the classical (dashed line) and the Bayes estimator (solid line) for a particular replication where cross validation fails.*

refers to the linear function f_1, the second row to the quadratic function f_2, and the third row to the sinusoidal function f_3. From left to right, the boxplots in the graphs correspond to adaptive Bayesian regression splines, Bayesian P-splines with three different choices for the hyperparameters and the classical approach. Additionally, Table 2.1 summarizes the rankings of the various estimators (in terms of the MSE measure) for very low, low and medium SNR together with average rankings averaged over all SNR's. From Figure 2.2 and Table 2.1 we can draw the following conclusions:

- For Bayesian P-splines, the dependence of results on the hyperparameters is strongest for the linear function f_1 whereas for the quadratic and sinusoidal function it is relatively small. However, inspecting the individual estimates for the linear function f_1 shows that the estimates for the different choices of hyperparameters always suggest an underlying linear function but estimates become slightly more wiggled for increasing b.

- Biller's adaptive regression splines perform inferior compared to both P-splines approaches.

- Compared to the frequentist version, our fully Bayesian approach performs equally well or better for quadratic function f_2 and the sinusoidal function f_3. Regarding the linear function f_1, the performance of Bayesian P-splines depends on the choice of the hyperparameters. For $b = 0.0005$ and $b = 0.00005$ the Bayesian approach is superior, for $b = 0.005$ it performs inferior. In general, Table 2.1 suggests that the Bayesian approach with hyperparameters $b = 0.0005$ and $b = 0.00005$ performs superior while with hyperparameter $b = 0.005$ both approaches perform roughly equal.

We sometimes observed strange results for the frequentist version of P-splines. For a very low signal to noise ratio, approximately 3-5% of its estimates are quite unsmooth because the cross validation score function has no global minimum or a too small smoothing parameter was found as the optimum. For the Bayesian approaches we never observed these problems. As an example, compare Figure 2.1 d) which shows for f_3 the classical (dashed line) and the Bayesian P-spline (solid line) for a particular replication. For higher signal to noise ratios, however, the problem disappears.

For Bayesian P-splines, we also investigated the coverage of pointwise credible intervals. Using MCMC simulation techniques, credible intervals are estimated by computing the respective quantiles of the sampled function evaluations. For a nominal level of 80% the average coverage usually varies between 81 and 86% for all models and all choices for the hyperparameters. Taking a nominal level of 95% the average coverage varies between 95 and 97%. Only in the case of the sinusoidal function f_3 and a very low signal to noise ratio we observed with hyperparameters $b = 0.0005$ and $b = 0.00005$ average coverages slightly below the respective nominal levels. This implies that the credible intervals obtained by the fully Bayesian approach are rather conservative.

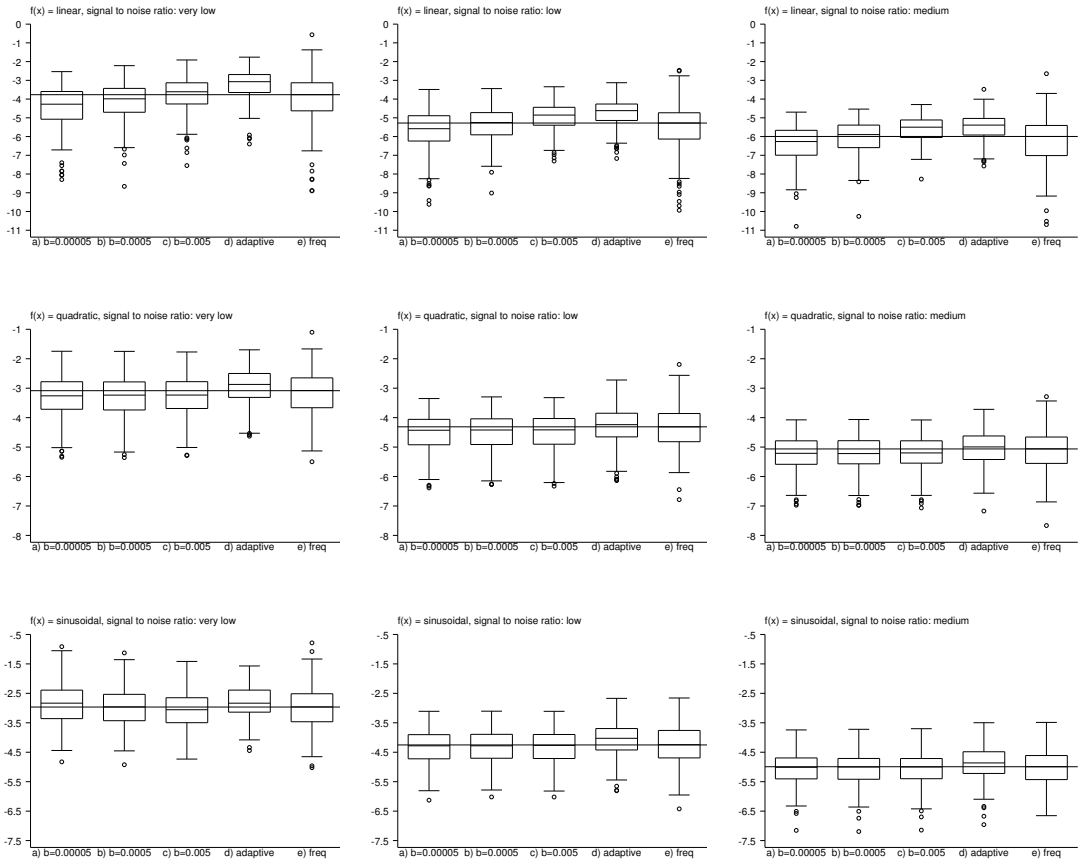


Figure 2.2: *Boxplots of log(MSE) for the various estimators of simulation study 1. The first row refers to the linear function f_1, the second row to the quadratic function f_2 and the third row to the sinusoidal function f_3. The left panel corresponds to a very low SNR ($\sigma = 1$), the medium panel to a low SNR ($\sigma = 0.5$) and the right panel to a medium SNR ($\sigma = 0.33$). From left to right the boxplots in the respective graphs refer to Bayesian P-splines with hyperparameters $b = 0.00005$, $b = 0.0005$, $b = 0.005$, Biller's adaptive Bayesian regression splines, and the frequentist version of P-splines.*

Table 2.1: *Average rankings from simulation study 1.*

	very low SNR	low SNR	medium SNR	average
Classical P-splines	2.9	2.9	2.8	2.9
Bayesian P-splines ($b = 0.00005$)	2.5	2.1	2.1	2.3
Bayesian P-splines ($b = 0.0005$)	2.6	2.5	2.5	2.5
Bayesian P-splines ($b = 0.005$)	2.8	3.1	3.3	3.1
adaptive regression splines	4.2	4.3	4.1	4.2

2.4.2 Highly oscillating functions

In order to compare our method for highly oscillating curves, we mainly refer to Ruppert and Carroll (2000) who propose P-splines based on a truncated power series basis and quadratic penalties on the regression coefficients with locally adaptive smoothing parameters. In their first simulation example they used the function

$$f_4(x) = \sqrt{x(1-x)}\sin\left(\frac{2\pi(1+2^{(9-4j)/5})}{x+2^{(9-4j)/5}}\right),$$

whose spatial variability depends on the additional parameter j. They used $j = 3$ which corresponds to low spatial variability and $j = 6$ which corresponds to severe spatial variability. We simulated 250 replications for both specifications and applied the following estimators:

- Bayesian cubic P-splines with a second order random walk penalty using a global variance and locally adaptive variances. We used both 40 and 80 knots and the same three different choices of hyperparameters as in Section 2.4.1.
- Classical (cubic) P-splines with second order difference penalty. Similar to Bayesian P-splines, we used both 40 and 80 knots.
- Adaptive Bayesian regression splines by Biller (2000) with $\bar{k} = 20$ as the mean number of knots (see also Section 2.4.1.) Experiments with $\bar{k} = 10$ and $\bar{k} = 30$ showed virtually no difference.
- Multivariate adaptive regression splines (MARS) of Friedman (1991) with a maximum number 150 of basis functions.

In order to compare results, we computed $log_{10}(\sqrt{(MSE)})$ as Ruppert and Carroll did. It turned out that the dependence of the results on the three choices for the hyperparameters is negligible. Therefore, the presentation of results is restricted to the choice of $a = 1$ and $b = 0.005$ for the hyperaparameters. Figure 2.3 displays boxplots of $log_{10}(\sqrt{(MSE)})$ for the various estimators. Figure a) corresponds to $j = 3$, i.e. low spatial variability, and Figure b) to $j = 6$, i.e. severe spatial variability. From left to right, the respective boxplots refer to Bayesian P-splines with a global variance (40 and 80 knots), Bayesian P-splines with locally adaptive variances (40 and 80 knots), adaptive Bayesian regression splines, classical P-splines (40 and 80 knots) and MARS.

From Figure 2.3 we can draw the following conclusions:

- For $j = 3$, i.e. low spatial variability, our estimators with global and locally adaptive variance perform almost equally well. If 80 knots are used we observe a slight loss in statistical efficiency. Hence, there is (almost) no loss of statistical efficiency when a locally adaptive estimator is used but not needed.
- For $j = 6$, i.e. severe spatial variability, our estimators with locally adaptive variance clearly outperform the estimators with global variance.

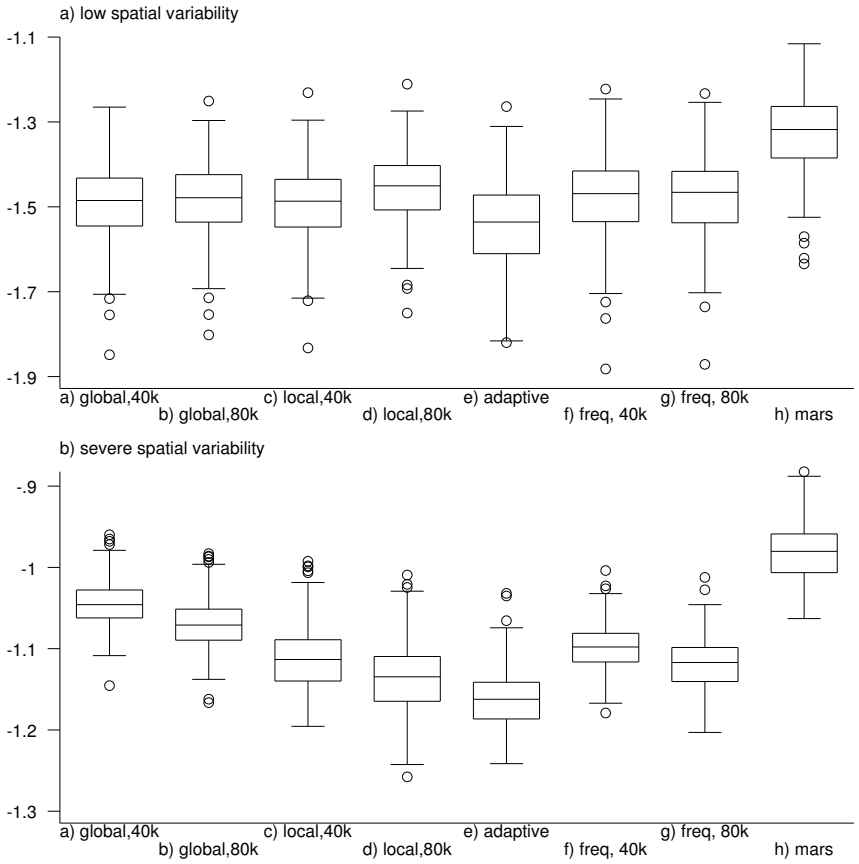


Figure 2.3: *Simulation study 2: Boxplots of* $log_{10}(\sqrt{(MSE)})$ *for function* f_4 *with low spatial variability (panel a)) and severe spatial variability (panel b)). From left to right the boxplots in the graphs correspond to Bayesian P-splines with a global variance (40 and 80 knots), Bayesian P-splines with locally adaptive variances (40 and 80 knots), adaptive Bayesian regression splines, classical P-splines (40 and 80 knots) and MARS.*

- For $j = 3$, i.e. low spatial variability Biller's adaptive Bayesian regression splines are slightly superior to the other approaches. Bayesian and classical P-splines perform almost equally well.

- For $j = 6$, i.e. severe spatial variability, the best results are obtained by Biller's adaptive Bayesian regression splines followed by our Bayesian P-splines approach with 80 knots and locally adaptive variances. Comparing Bayesian P-splines with a global variance and classical P-splines we observe that the frequentist approach performs superior to our Bayesian variant.

- The main reason for the poor performance of MARS is primarily because it uses linear splines. Therefore estimates are less smooth than for the other estimators. The crude functional form is, however, always detected. In fact, MARS was developed for problems with many covariates and interactions and it is not too surprising that it is less efficient for univariate problems.

To gain more insight into the differences of the various estimators, Figure 2.4 displays the respective 10th percent worst fit (in terms of MSE) for Biller's adaptive Bayesian regression splines, Bayesian P-splines with 80 knots and locally adaptive variances, Bayesian P-splines with 80 knots and global variance and classical P-splines with 80 knots. We see that both P-spline estimators with a global variance (or smoothing parameter) are relatively wiggled in the right part where the function is less oscillating. Classical P-splines are more wiggled than Bayesian P-splines in this part of the function which is quite typical. It is also typical that the adaption to the function in the highly oscillating part is better for classical P-splines (although the very well adaption in the example is accidently). This implies that Bayesian P-splines (with global variance) have a tendency to larger smoothing parameters than classical P-splines in this example. We also see how Bayesian P-splines with locally adaptive variance improve the estimator with global variance, as they are less wiggled in the less oscillating part of the function and adapt better to the function where it is highly oscillation. Biller's adaptive Bayesian regression splines, however, perform even better. Even the 10th percent worst fit shows a very well adaption to the underlying true function.

Although a direct comparison (using the same data) with Ruppert and Carroll's approach was not possible a rough comparison seams justified because they used exactly the same models. For $j = 3$, they obtained values of approximately -1.5 for the median of $log_{10}(\sqrt{(MSE)})$, i.e. Ruppert and Carroll's (2000) approach performs equally well as the estimators we compared. Both their global and local penalty estimator perform equally well in this situation. For $j = 6$, their local penalty estimator has superior performance compared to their global penalty estimator with a median value of approximately -1.25 for $log_{10}(\sqrt{(MSE)})$ implying that their approach performs even superior than Biller's adaptive Bayesian regression splines. Ruppert and Carroll compared their method also with results from a simulation study by Wand (2000) who compares POLYMARS of Stone et al. (1997), the Bayesian approach to nonparametric regression by Smith and Kohn (1996) and penalized shrinkage. Compared to our results, the Bayesian approach by Smith and

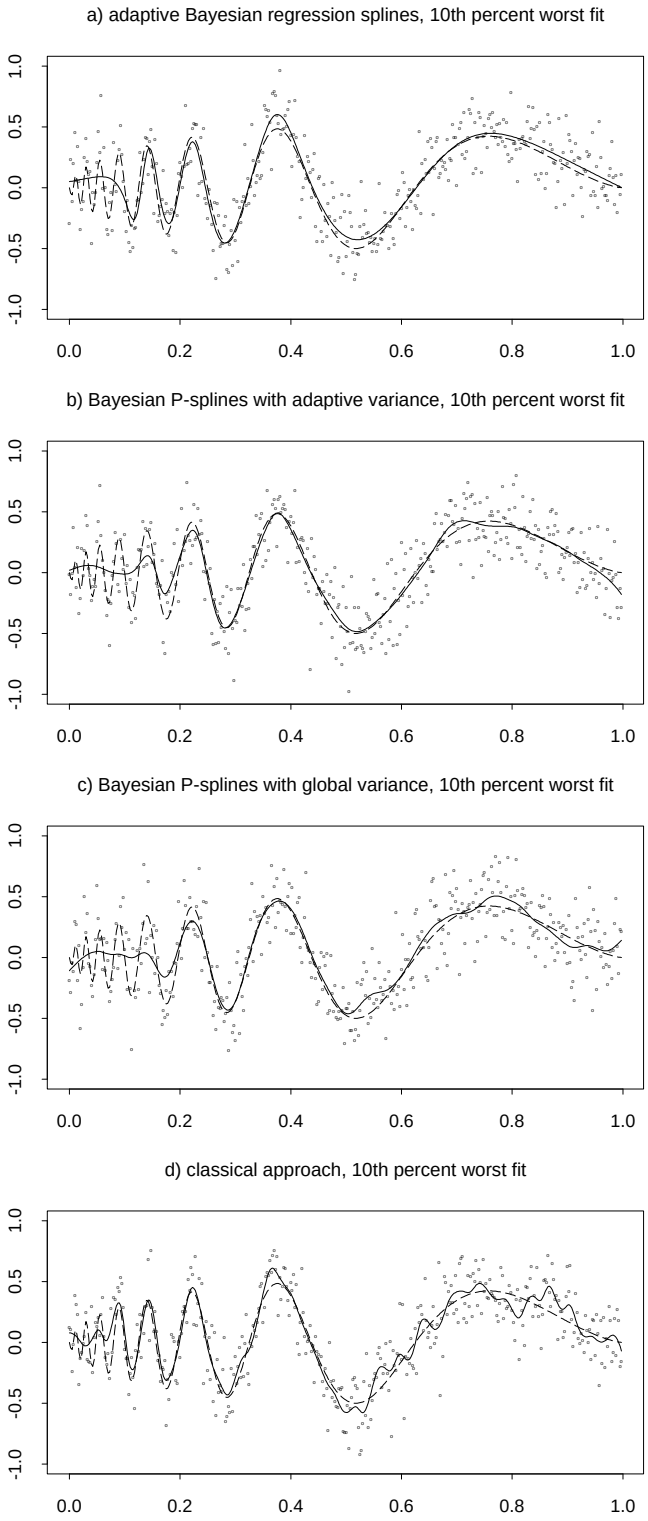


Figure 2.4: *Simulation study 2: The graphs show the respective 10th percent worst fits in terms of the MSE measure plotted in Figure 2.3 for Biller's adaptive Bayesian regression splines, Bayesian P-splines with 80 knots and adaptive variances, Bayesian P-splines with 80 knots and global variance and classical P-splines with 80 knots.*

Kohn (1996) performs roughly equally well than Bayesian P-splines with locally adaptive variances while POLYMARS and penalized shrinkage perform slightly inferior.

For both simulation examples, we also computed the coverage of pointwise credible intervals. For $j = 3$, i.e. low spatial variability, the average coverage for Bayesian P-splines with global variance as well as adaptive variance are always above the nominal levels of 80 and 95 percent. For a nominal level of 80 percent the average coverage varies (depending on the choice of the hyperparameters and the number of knots) between 83 and 84 percent and for a nominal level of 95 percent between 96 and 97 percent. Taking $j = 6$, i.e. severe spatial variability, the average coverage of the estimators is always below the nominal level mainly because of very low coverage rates for $x < 0.15$. However, using P-splines with locally adaptive variances clearly increases the average coverage. For P-splines with 40 knots the average coverage increases from 71.6 to 74.2% and from 84.5 to 87.5% for nominal levels of 80 and 95 percent. For P-splines with 80 knots the average coverage increases from 73.5 to 77% and from 80.5 to 90.1%.

2.4.3 Surface fitting

In our last simulation study we compare our approach for surface fitting with related methods in the literature. We mainly refer to Smith and Kohn (1997) who compared their Bayesian subset selection-based procedure with a variety of other approaches. Besides their own approach they included MARS of Friedman (1991), Clive Loader's "locfit" (see Cleveland and Grosse 1991), bivariate cubic thin plate splines with a single smoothing parameter (henceforth tps), tensor product cubic smoothing splines with five smoothing parameters, Breiman and Friedman's (1985) additive basis fitting routine and a parametric linear interaction model (henceforth lsp). They regarded the following three examples:

- $f_5(x_1, x_2) = 1/5 \exp(-8x_1^2) + 3/5 \exp(-8x_2^2)$ where x_1 and x_2 are distributed independently normal with mean 0.5 and variance 0.1.
- $f_6(x_1, x_2) = x_1 \sin(4\pi x_2)$ where x_1 and x_2 are distributed independently uniform on $[0, 1]$.
- $f_7(x_1, x_2) = x_1 x_2$ where x_1 and x_2 are bivariate normal with mean 0.5, variance 0.05 and correlation of 0.5.

Function f_5 represents a model with main effects only, and functions f_6 and f_7 correspond to a model with interactions. The sample size was $n = 300$ observations and $\sigma = 1/4\, range(f_j)$. We simulated 250 replications and considered the following estimators:

- Bayesian (cubic) P-splines on a 12 by 12 knots grid and the smoothness prior (2.6). We examined the same three choices for the hyperparameters of the variance as in the preceding subsections.
- Bayesian P-splines as described above but with main effects included. For the main effects we used cubic P-splines with 20 knots and a second order random walk penalty with global variance.

- For a direct comparison with other methods, we used MARS, locfit, tps and lsp of Smith and Kohn's simulation study with the same estimation parameters as described in their article.

We have also experimented with P-splines and locally adaptive variances, i.e the priors (2.6) and (2.4) replaced by their locally adaptive variants. Because the functions under consideration are not highly oscillating the results are more or less identical to those of P-splines with a global variance. An exception is function f_6 which is the only function under study with moderate spatial variability. Here, the locally adaptive variants perform slightly better.

Figure 2.5 shows boxplots of $\log(MSE)$ for the various estimators. Panel a) refers to function f_5, panel b) to function f_6 and panel c) to function f_7. From left to right the boxplots refer to Bayesian P-splines without main effects, Bayesian P-splines with main effects included, locfit, lsp, MARS and tps. Results are shown for the choice of $a = 1$ and $b = 0.005$ for hyperparameters only because for the other choices we obtained almost identical results. We also noticed that the results for MARS, locfit, tps and lsp are very close to Smith and Kohn's study. For that reason, it seems justified to include also those estimators in our comparison that have been considered in Smith and Kohn (1997) but not here. From Figure 2.5 and the results of Smith and Kohn we draw the following conclusions:

- Regarding function f_5 the best results are obtained by the estimators with main effects included which is not surprising because the true function consists of main effects only. Moreover, an inspection of single estimates shows that the estimated interaction effects are more or less zero which makes sense, too. For the functions f_6 and f_7 the estimators with and without main effects perform roughly equally well.

- From the comparison with other estimators we see that our approach is competitive. For function f_5 the estimator without main effects performs comparable to 'tps' and is among the three best in Smith and Kohn's study. The estimators with main effects included perform equally well (if not slightly better) than the best estimator in Smith and Kohn's study which is the cubic tensor product spline. For f_6 our estimators are comparable to 'tps' which is the third best estimator in Smith and Kohn's article. For function f_7 Smith and Kohn's Bayesian subset selection-based procedure clearly outperforms the other estimators in their study including the parametric linear fit 'lsp'. The performance of our estimator is once again comparable to 'tps'.

Furthermore, we investigated the coverage of pointwise credible intervals of our estimators. The average coverage of all estimators is within a range of 80 to 88% for a nominal level of 80% and within a range of 94 and 98% for a nominal level of 95% which confirms the findings of the previous sections that the fully Bayesian approach yields rather conservative credible intervals. An exception is the estimator with main effects included for f_6 where the average coverage is only 68% and 83%, respectively.

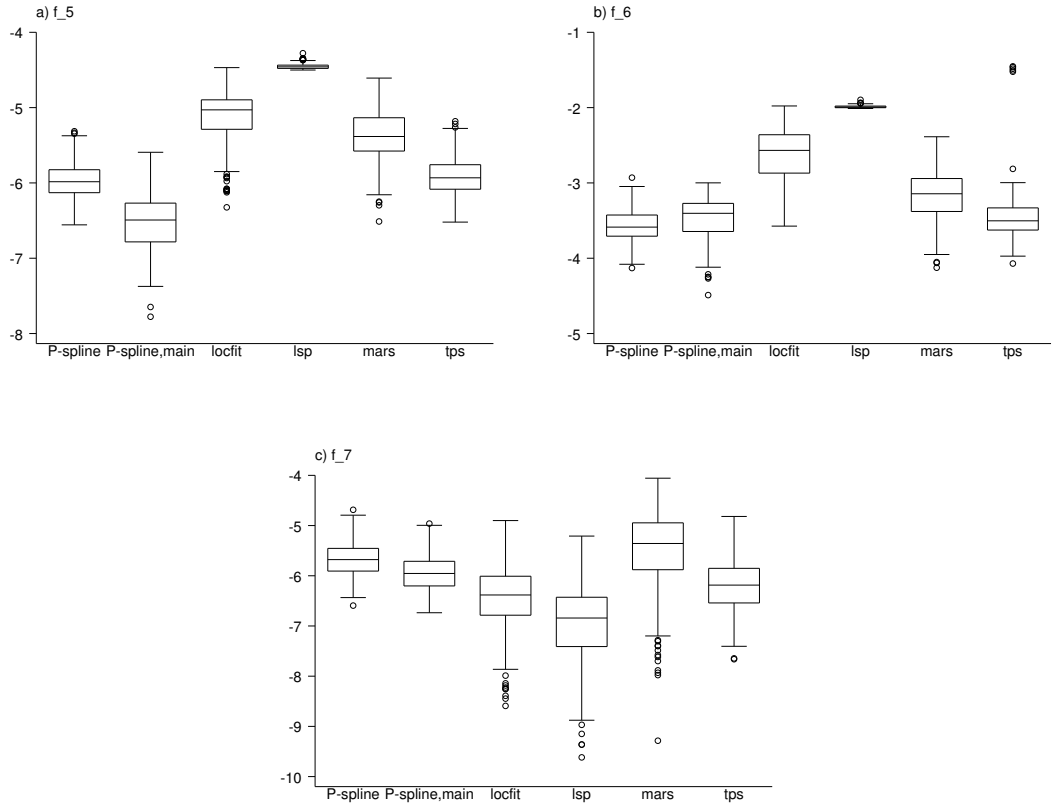


Figure 2.5: *Simulation study 3: Boxplots of log(MSE) for the various surface estimators. Panel a) corresponds to function f_5, panel b) to function f_6 and panel c) to function f_7. From left to right the respective boxplots refer to Bayesian P-splines without main effects, Bayesian P-splines with main effects, the parametric linear interaction model (lsp), Clive Loader's 'locfit', MARS and thin plate splines (tps).*

2.5 Applications

In this section we demonstrate the practicability of our approach with two applications. The first application on rents for flats in Munich is an example of a geoadditive model. The second application on human brain mapping demonstrates the usefulness of smoothing with spatially adaptive variances.

2.5.1 Rents for flats

According to the German rental law, owners of apartments or flats can base an increase in the amount that they charge for rent on "average rents" for flats comparable in type, size, equipment, quality and location in a community. To provide information about these "average rents", most larger cities publish "rental guides", which can be based on regression analysis with rent as the dependent variable. We use data from the City of Munich, collected in 1998 by Infratest Sozialforschung for a random sample of more than 3000 flats. As response variable we choose

R monthly net rent per square meter in German Marks, that is the monthly rent minus calculated or estimated utility costs.

Covariates characterizing the flat were constructed from almost 200 variables out of a questionnaire answered by tenants of flats. In our reanalysis we use the highly significant continuous covariates "floor space" (F) and "year of construction" (Y) and a vector v of 25 binary covariates characterizing the quality of the flat, e.g. the kitchen and bath equipment, the quality of the heating or the quality of the warm water system. Another important covariate is the location L of the flat in Munich. For the official Munich '99 rental guide, location in the city was assessed in three categories (average, good, top) by experts. In our reanalysis we focus on a more data driven assessment of the quality of location by including a spatial effect f_{spat} of the location L into the predictor. So we choose the geoadditive model with predictor

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + f_{12}(F,Y) + f^{str}(L) + f(L)^{unstr} + v'\gamma.$$

The main effects f_1 and f_2 of floor space and year of construction are modeled by cubic P-splines with 20 knots and a second order random walk penalty. For the interaction we choose a two dimensional P-spline on a grid of 12 by 12 knots with smoothness prior (2.6). We have also experimented with P-splines and locally adaptive variances but the differences were negligible. For the spatially structured effect $f^{str}(L)$ we choose the Markov random field prior (2.8), and (2.9) for the unstructured spatial effect $f^{unstr}(L)$.

To assess the dependence of results on the choice for the hyperparameters of variance parameters we estimated the model with three different choices, $a = 1, b = 0.005$, $a = 1, b = 0.0005$ and $a = 1, b = 0.00005$. Table 2.2 compares the relative changes of estimated posterior means for different hyperparameters with respect to the choice $a = 1, b = 0.005$. The following figures are based on the first choice of $a = 1$ and $b = 0.005$.

Figure 2.6 shows the effects of floor space and year of construction. Panels a) and b) show the posterior means together with 80% and 95% pointwise credible intervals of the main effects. Panel c) displays the posterior mean of the interaction term. Figure 2.6 a) shows the strong influence of floor space on rents: small flats and apartments are considerably more expensive than larger ones, but this nonlinear effect becomes smaller with increasing floor space. The effect of year of construction on rents in Figure b) is more or less constant until the '50s. It then distinctly increases until about 1990, and it stabilizes on a high level in the '90s. Although the interaction effect in Figure c) is not overwhelmingly large, we clearly see that old flats built before the second world war with a floor space below 45 square meters are cheaper than the average. On the other hand, modern flats built after 1972 (the year of the Olympic summer games) are somewhat more expensive than the average. Taking a look at Table 2.2, we see that both main effects are virtually unchanged by different choices of hyperparameters whereas the interaction effect changes considerably. There seems to be particularly doubt about the size of the effect, not so much about the functional form. However, there is justification not to remove the interaction effect because reestimating the model without considering the interaction effect leads for all choices of hyperparameters to a significant increase in the deviance information criteria DIC (Spiegelhalter, Best, Carlin and van der Linde 2002), which can be used as a tool for model comparison in complex hierarchical Bayesian models.

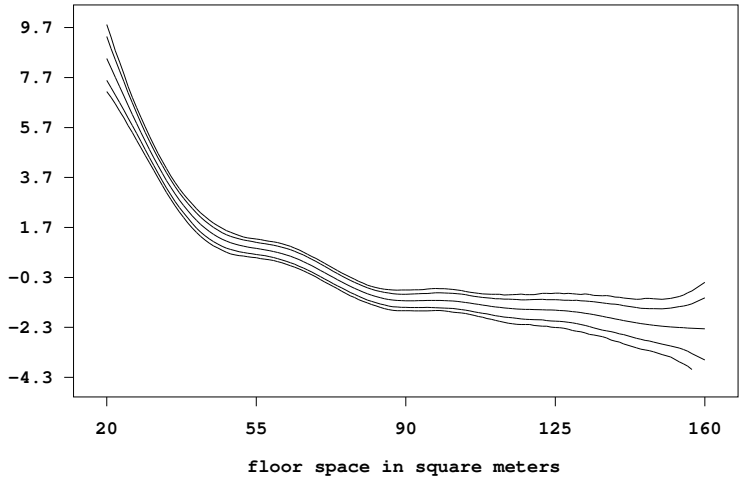
Figure 2.7 a) shows a map of Munich, displaying subquarters and the posterior mean estimates of the spatial effect f_{spat}. Note that the correlated effects clearly exceed the uncorrelated effects with a range approximately between -1.7 and 1.7. In contrast, the coefficients of the uncorrelated effects have only a range between -0.5 and 0.5. As can be seen in Table 2.2 the sensitivity of the spatial effect on the choice of hyperparameters is relatively small.

The inclusion of a spatial effect f_{spat} is a good opportunity to investigate empirically the validity of the experts assessment of the quality of location. In fact, we could reestimate the model with the experts assessment included in form of two additional dummy variables for good and top locations. If the experts assessment is valid the extra spatial variation measured by the spatial effect should considerably decrease. Figure 2.7 b) displays the spatial effect when the experts assessment is included. The effects of floor space, year of construction and the fixed effects are virtually unchanged and therefore omitted. We observe that the remaining variation in Figure b) is smoother although there is considerable spatial variation remaining. The reason for the small decrease is that the variation of the uncorrelated effects remains more or less stable. The variation of the correlated random effects, however, decreases considerably.

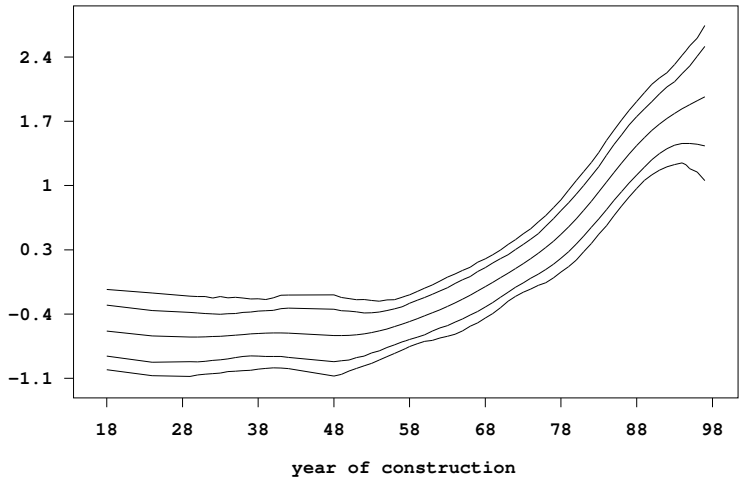
2.5.2 Human brain mapping

The purpose of human brain mapping is to detect regions of the brain that are activated if a certain stimulus (e.g. visual or acoustic) is present. Detecting areas in the brain that are responsible for the processing of certain stimuli is not only of pure scientific interest but also important in many practical disciplines, e.g. in surgery. The realization of human brain

a) Main effect of floor space



b) Main effect of year of construction



c) Interaction between floor space and year of construction

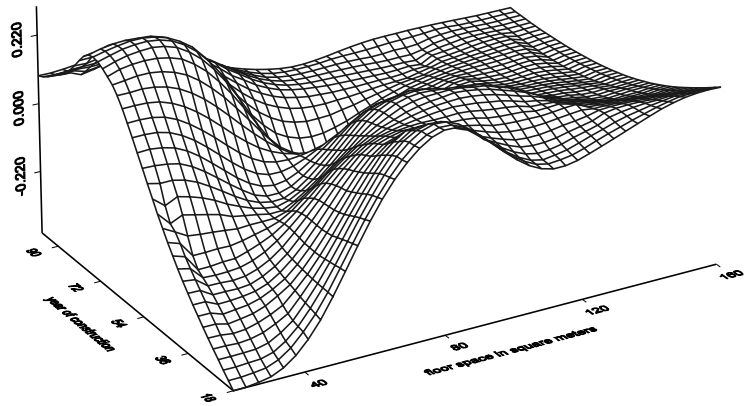


Figure 2.6: *Rents for flats: Effect of floor space and year of construction. Panel a) and b) show the main effects (posterior means, 80% and 95% pointwise credible intervals). The posterior means of the interaction effect is given in panel c).*

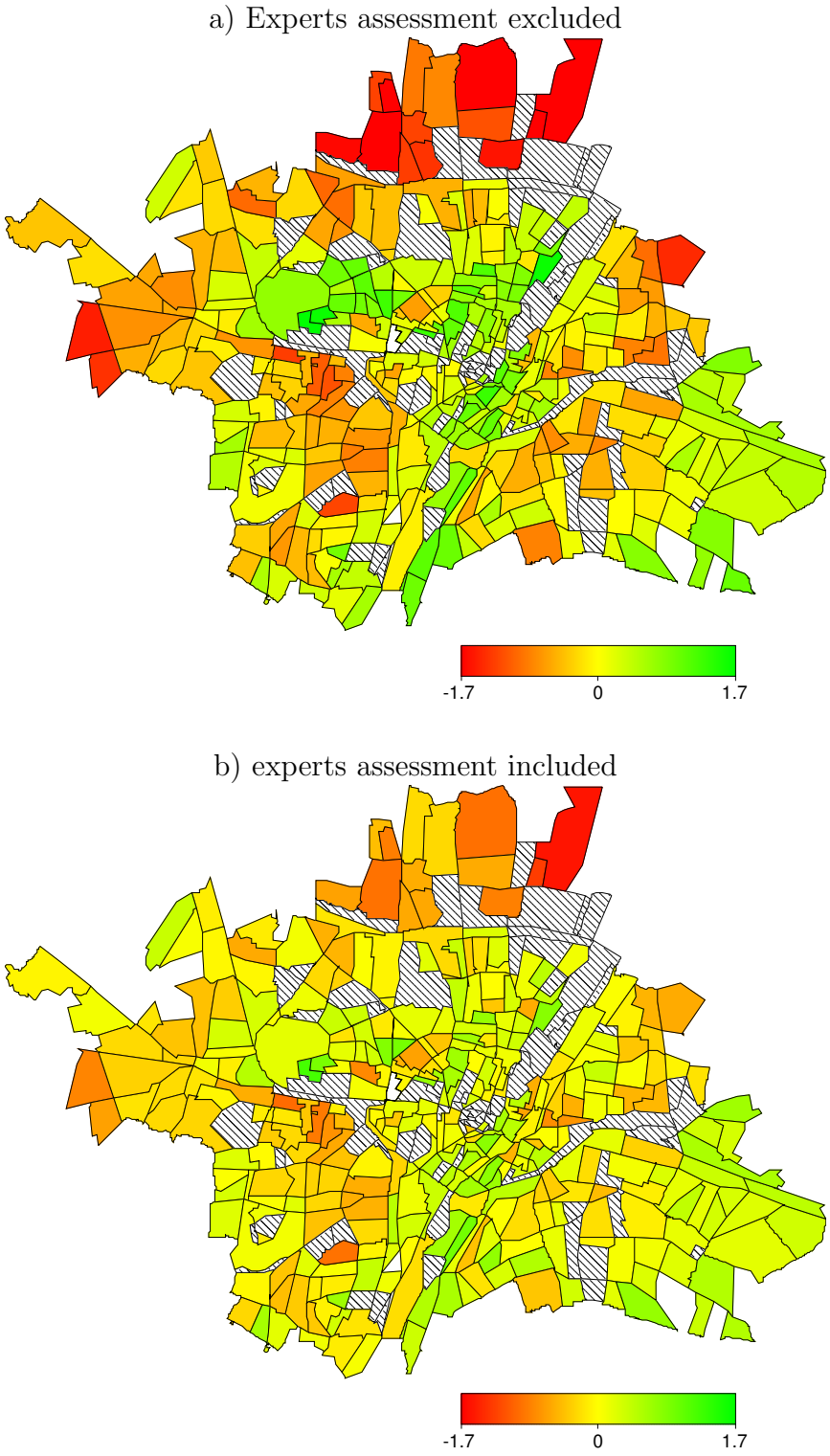


Figure 2.7: *Rents for flats: Posterior means of the spatial effect f_{spat}. Panel a) refers to the model that excludes the experts assessment of location, panel b) refers to the model that includes it.*

Table 2.2: *Rents for flats: Relative changes of estimated functions for different choices of hyperparameters.*

b	$f_1(F)$	$f_1(Y)$	$f_{12}(F,Y)$	spatial effect
0.005	0	0	0	0
0.0005	0.0002	0.0020	0.3600	0.0114
0.00005	0.0002	0.0066	0.7657	0.0251

mapping experiments has been considerably facilitated by the development of functional Magnetic Resonance Imaging (fMRI) which is the first non-invasive technique in this area. fMRI allows to determine the blood oxygenation level in the brain which can be used as a measure of brain activity, see e.g. Lange (1996) for details. In a typical fMRI experiment, the brain activity Y of a certain person is measured at a number of usually equidistant time points $t = 1, \ldots, T$. During the observation period, a kind of ON-OFF stimulus X (e.g. visual) is periodically presented (e.g. 30s of rest, 30s of stimulus, 30s of rest,...). At each of the T time points an MRI image Y_t consisting of I pixels or voxels i is measured, i.e. $Y_t = (Y_{t1}, \ldots, Y_{tI})$. The scientific question is now to determine which of the pixels i, $i = 1, \ldots, I$, is activated when the stimulus X_t is present. Unfortunately, the measurement of the level of activation Y_{ti} is subject to a number of non-negligible interferences. Hence, the Y_{ti}'s are measured with (considerable) noise and statistical methodology is required to remove the noise from the data. Typically, the statistical analysis of fMRI data can be divided into three parts. The first part consists mostly of preprocessing of the data, e.g. motion correction. In the second step, pixelwise statistical analysis is performed to remove a possible time trend in the data and to estimate the influence of the stimulus. Various competing approaches are currently discussed in the vast literature on this subject, see e.g. Gössl (2001) for an overview. The third step is concerned with spatial dependencies between voxels, i.e. the pixelwise estimated effect of the stimulus is spatially smoothed, mainly to overcome the multiple test problem which arises when analyzing several thousand time series non-simultaneously. Particulary in experiments with a visual stimulus, edge preserving spatial smoothing is required because of sudden jumps from non-activation to activation.

In this demonstrating example, we solely focus on the second and third step. We analyze data from a typical fMRI experiment where the level of activation of a volunteer was measured with a delay of 3 seconds at $T = 70$ time points. For simplicity the analysis is restricted to a particular horizontal slice of the brain which consists of $59 \times 64 = 3776$ pixels. From the 3776 pixels 826 pixels are known to lie outside the brain so that finally a total number of 2950 time series is analyzed. A visual stimulus was presented in three time periods of 30 seconds during the experiment. The first period was between $t = 11$ and $t = 30$. Each of the three stimulus periods was followed by a 30 seconds lasting period of rest. We first analyze the data pixelwise using Gaussian regression models with predictors

$$\eta_{ti} = \gamma_0 + f_1^i(t) + f_2^i(t) Z_{it}, \quad t = 1, \ldots, 70, \quad i = 1, \ldots, 2950. \tag{2.15}$$

Here, Z_{it} is a delayed and continuously modified stimulus which is routinely obtained from

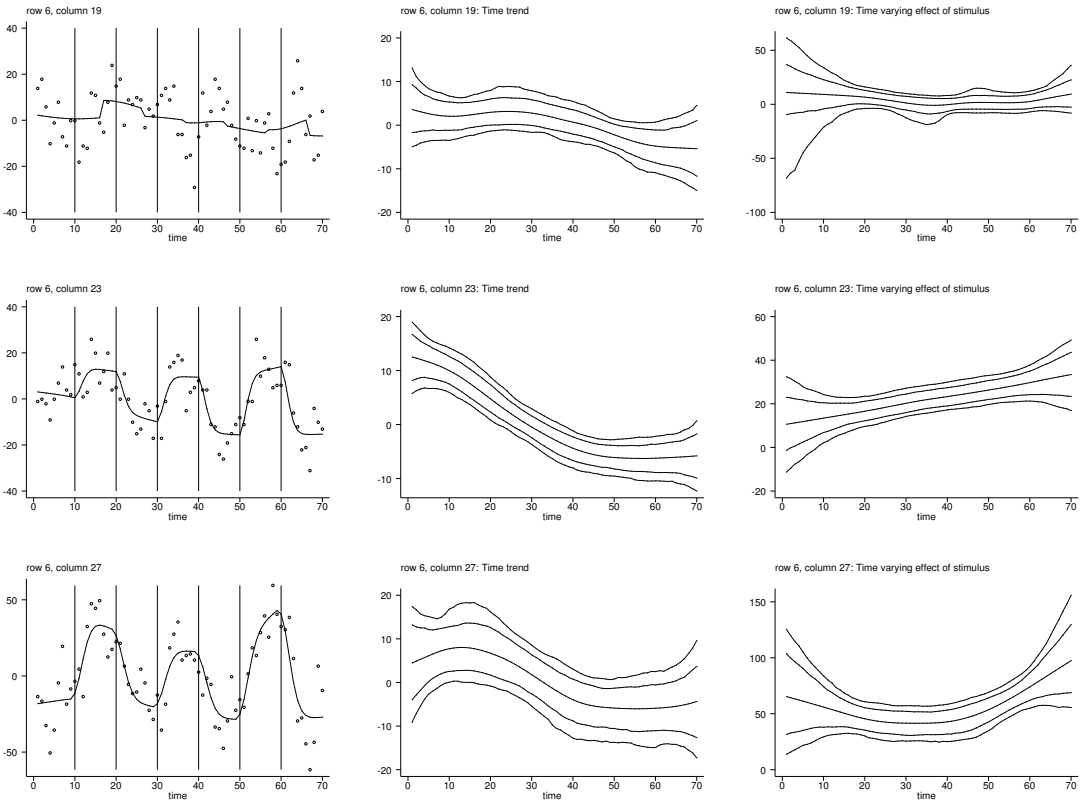


Figure 2.8: *Human brain mapping: The graphs display examples of the pixelwise analysis. The left panel shows the time series Y_{it} together with $\hat{Y}_{it}$ (solid line). The middle and the right panel display the estimated time trend and the time varying effect of the transformed stimulus. Shown is the posterior mean, 80% and 95% pointwise credible intervals.*

X_t in the preprocessing step, see Gössl (2001) for details. For f_1^i and f_2^i we assume Bayesian cubic P-splines with second order random walk penalty and 20 knots. The first function f_1^i models a nonlinear trend in the data. The second function f_2^i reflects a possibly time varying effect of the stimulus. The hypothesis is that the level of activation may vary over time, e.g. because it may take some time to get used to the experiment and to be fully concentrated. This approach has already been followed by Gössl, Auer and Fahrmeir (2000) who apply dynamic or state space models to estimate (2.15). Examples for the pixelwise analysis are given in Figure 2.8 where each row corresponds to a particular pixel. The left panel displays the unsmoothed time series Y_{ti}, $t = 1, \ldots, 70$, together with their reconstruction $\hat{Y}_{ti}$ according to (2.15). The middle and the right panel display the corresponding estimates for the functions f_1^i and f_2^i. Here, the posterior means together with 80% and 95% credible intervals are shown.

In a second step, we spatially smoothed the estimated function values $\hat{f}_2^i(t)$ at three distinct time points $t = 18, 38, 58$ which correspond to the end of the three stimulus periods. We used two dimensional P-splines on a grid of 25×25 knots with spatial smoothness prior

Table 2.3: *Human brain mapping: DIC of the six surface estimators.*

	$t = 18$	$t = 38$	$t = 58$
global variance	18117	17151	19749
spatially adaptive variance	17824	16716	19367

(2.6). Figure 2.9 shows the posterior mean of the two dimensional surfaces for $t = 18$ (first row), $t = 38$ (second row) and $t = 58$ (third row). The left panel corresponds to the estimators with a global variance and the right panel to the estimators with spatially adaptive variances. The DIC of the six estimates can be found in Table 2.3. Obviously, the use of spatially adaptive variances significantly reduces the DIC. Moreover, we observe that the usage of adaptive variances leads to slightly smoother estimates in areas which are less activated (mainly the right part of the graphs). On the other hand the peaks of the activation areas are much more pronounced.

2.6 Conclusions

In this paper we propose a fully Bayesian approach for P-splines and present a couple of extensions. Our approach covers additive models, varying coefficient models, geoadditive models, two dimensional surface fitting and improved estimation of functions with changing curvature. Our implementation (included in *BayesX*) allows a more or less arbitrary additive decomposition of the predictor using one or two dimensional nonlinear functions, interactions based on varying coefficient models, spatially correlated effects based on MRF priors or two dimensional surface estimators and i.i.d Gaussian random effects. In all cases, the amount of smoothing is estimated simultaneously with the unknown nonlinear functions. We consider this as a distinct advantage of our Bayesian approach as the estimation of smoothing parameters is still a problem in a frequentist approach at least when the predictor contains a moderate or large number of unknown functions. The competitiveness of our approach has been demonstrated through extensive simulation studies in Section 2.4.

There are, however, some remaining open problems. The following points will be investigated in future research:

- Although the usage of spatially adaptive variances rather than a global variance considerably improves estimation of highly oscillating functions our simulation study shows that Biller's adaptive Bayesian regression splines perform even better. A possible idea for further improvements could be to define the knots of the spline on a non-equidistant grid such that more knots are placed where the variability of the data is high.

- Estimation of surfaces via MCMC is relatively slow because the bandwidth of the posterior precision matrix is much larger than for univariate smoothers. A remedy might be to update the parameters row- or columnwise rather than all parameters

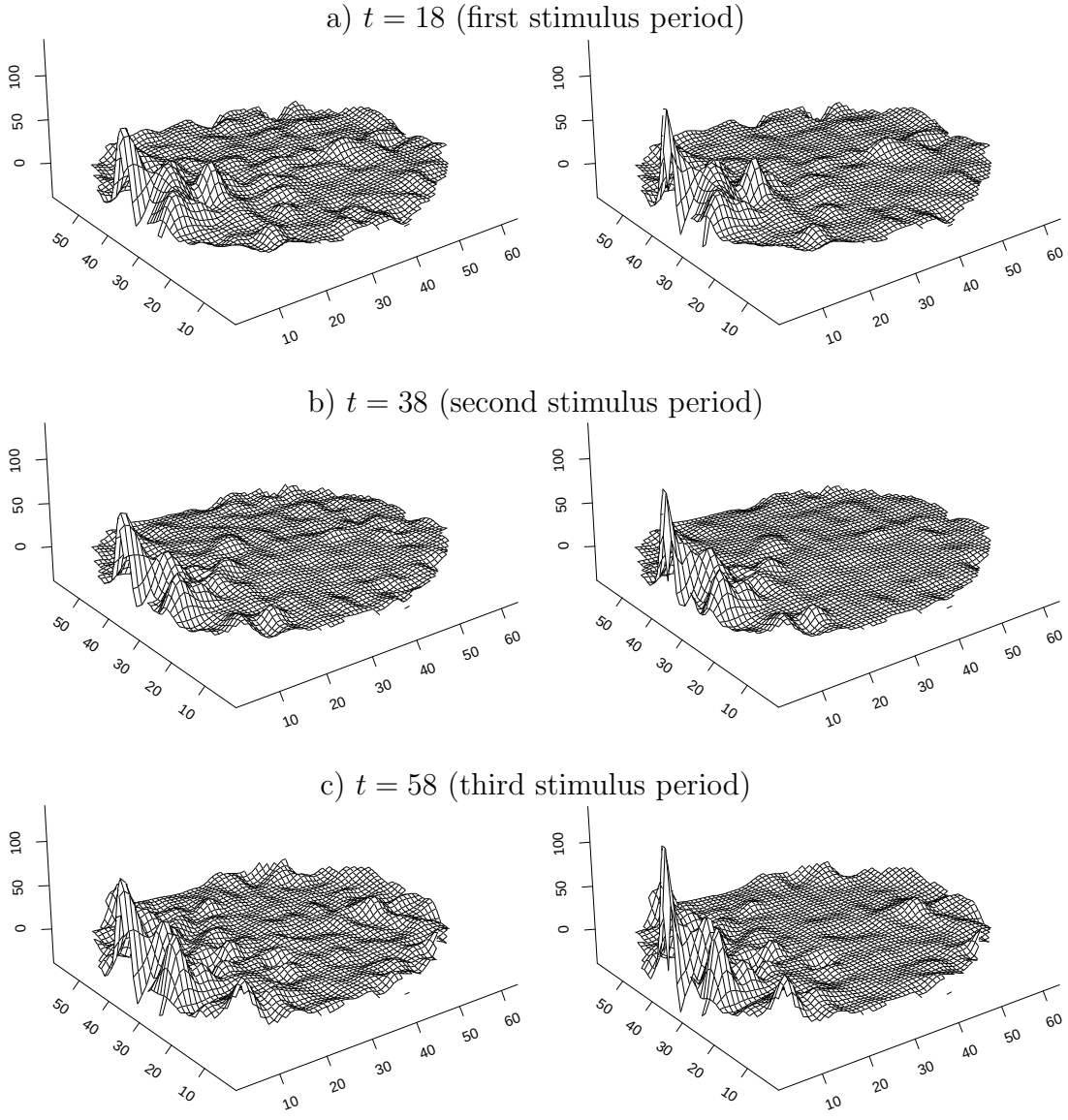


Figure 2.9: *Human brain mapping: The graphs show the spatially smoothed estimates of the effect of the transformed stimulus from the pixelwise analysis for different time points. The first row corresponds to* $t = 18$ *(first stimulus period), the second row to* $t = 38$ *(second stimulus period) and the third row to* $t = 58$ *(third stimulus period). The left panel shows the estimators with a global variance and the right panel with spatially adaptive variances.*

in one step. Then, the bandwidth of precision matrices of full conditionals reduces considerably.

- Finally, we intend to extend the approach to non-Gaussian errors e.g. by using similar sampling schemes as in Albert and Chib (1993) or Fahrmeir and Lang (2001b) for categorical probit models or as in Fahrmeir and Lang (2001a) for generalized additive models. First results are very promising.

Acknowledgement:
This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen". We thank Ludwig Fahrmeir for helpful discussions, Andrea Hennerfeind for support with the human brain mapping data and Brian Marx for providing the S-plus functions for P-splines. Last but not least we thank the editors and the three referees for their valuable suggestions to improve the first version of the paper.

Generalized structured additive regression based on Bayesian P-Splines

Andreas Brezger and Stefan Lang
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

ABSTRACT

Generalized additive models (GAM) for modeling nonlinear effects of continuous covariates are now well established tools for the applied statistician. In this paper we develop Bayesian GAM's and extensions to generalized structured additive regression (STAR) based on one or two dimensional P-splines as the main building block. The approach extends the work of Part I of this chapter for Gaussian responses. Inference relies on Markov chain Monte Carlo (MCMC) simulation techniques, and is either based on iteratively weighted least squares (IWLS) proposals or on latent utility representations of (multi)categorical regression models. Our approach covers the most common univariate response distributions, e.g. the binomial, Poisson or gamma distribution, as well as multicategorical responses. For the first time, we present Bayesian semiparametric inference for the widely used multinomial logit model. As we will demonstrate through two applications on the forest health status of trees and a space-time analysis of health insurance data, the approach allows realistic modeling of complex problems. We consider the enormous flexibility and extendability of our approach as a main advantage of Bayesian inference based on MCMC techniques compared to more traditional approaches. Software for the methodology presented in the paper is provided within the public domain package *BayesX*.

Keywords: geoadditive models, IWLS proposals, multicategorical response, structured additive predictors, surface smoothing

2.7 Introduction

Generalized additive models (GAM) provide a powerful class of models for modeling nonlinear effects of continuous covariates in regression models with non-Gaussian responses. A considerable number of competing approaches is now available for modeling and estimating nonlinear functions of continuous covariates. Prominent examples are smoothing splines (e.g. Hastie and Tibshirani 1990), local polynomials (e.g. Fan and Gijbels 1996), regression splines with adaptive knot selection (e.g. Friedman and Silverman 1989, Friedman 1991, Stone et al. 1997) and P-splines (Eilers and Marx 1996, Marx and Eilers 1998). Currently, smoothing based on mixed model representations of GAM's and extensions is extremely popular, see Lin and Zhang (1999), Currie and Durban (2002), Wand (2003) and the book by Ruppert et al. (2003). Indeed, the approach is very promising and has several distinct advantages, e.g. smoothing parameters can be estimated simultaneously with the regression functions.

Bayesian approaches are currently either based on regression splines with adaptive knot selection (e.g. Smith and Kohn 1996, Denison et al. 1998, Biller 2000, Di Matteo et al. 2001, Biller and Fahrmeir 2001, Hansen and Kooperberg 2002), or on smoothness priors (Hastie and Tibshirani 2000, Fahrmeir and Lang 2001a, Fahrmeir and Lang 2001b).

In this paper, we extend the work of Part I of this chapter for Gaussian responses based on one or two dimensional Bayesian P-splines as the main building block. Our approach covers univariate GAM's and extensions for the most common response distributions (binomial, Poisson, gamma) as well as models for multicategorical responses. For the first time, we present semiparametric Bayesian inference for multinomial logit models. Inference is fully Bayesian and is based on Markov chain Monte Carlo inference techniques (a nice introduction into MCMC can be found in Green 2001). We develop a number of highly efficient updating schemes with iteratively weighted least squares (IWLS) used for fitting generalized linear models as the main building block. Related algorithms have been proposed by Gamerman (1997) and Lenk and DeSarbo (2000) for estimating Bayesian generalized linear mixed models. Compare also Rue (2001) and Knorr–Held and Rue (2002) who develop efficient MCMC updating schemes for spatial smoothing of poisson responses. A simple alternative are conditional prior proposals (Knorr–Held 1999) which work surprisingly well in many situations. For categorical response models, alternative and in some cases more efficient sampling schemes are based on latent utility representations of such models, see Albert and Chib (1993), Chen and Dey (2000) and Fahrmeir and Lang (2001b) for (multicategorical) probit models and Holmes and Held (2004) for logit models. The advantage of such representations for MCMC inference is, that the full conditionals of the regression coefficients are (multivariate) Gaussian and sampling schemes developed for Gaussian responses in Part I of this chapter can be utilized with only minor changes. In all updating schemes, numerical efficiency is guaranteed by using matrix operations for band or sparse matrices (George and Liu 1981).

Our Bayesian approach for semiparametric regression has the following advantages compared to existing methodology:

- *Extendability to more complex formulations*
  A main advantage of a Bayesian approach for GAM's is its flexibility and extendability to more complex formulations. Our approach can be well extended (within a unified framework) to deal with unobserved unit- or cluster specific heterogeneity by incorporating random intercepts or slopes into the predictor. Spatial heterogeneity may be considered by incorporating spatial effects. We will discuss two alternatives, Gaussian Markov random fields (e.g. Besag et al. 1991) and two dimensional P-splines (compare Part I of this chapter). Models that can deal simultaneously with nonlinear effects of continuous covariates as well as spatial heterogeneity are called *geoadditive models* (Kammann and Wand 2003) and are of growing interest in the recent literature, see also Fahrmeir and Lang (2001a) and Fahrmeir and Lang (2001b). In general, we will use models with a *structured additive predictor* (STAR) including many well known model classes as special cases. Examples are generalized additive mixed models, geoadditive models, dynamic models, varying coefficient models and geographically weighted regression. The latter is well known in the geography literature (and less known to statisticians), see e.g. Fotheringham, Brunsdon and Charlton (2002).

  Our approach may also be used as a starting point for Bayesian inference in other model classes or more specialized settings. For example Hennerfeind, Brezger and Fahrmeir (2003) build on the inference techniques of this paper for developing geoadditive survival models.

- *Inference for functions of the parameters*
  Another important advantage of inference based on MCMC is easy prediction for unobserved covariate combinations including credible intervals, and the availability of inference for functions of the parameters (again including credible intervals). We will give specific examples in our second application.

- *Estimating models with a large number of parameters and observations*
  In Fahrmeir, Kneib and Lang (2004) we compare the relative merits of the full Bayesian approach presented here, and empirical Bayesian inference based on mixed model technology where the smoothing parameters are estimated via restricted maximum likelihood. Although the standard methodology from the literature has been improved the algorithms are still of the order p^3 where p is the total number of parameters. Similar problems arise if the smoothing parameters are estimated via GCV, see Wood (2000). In our Bayesian approach based on MCMC techniques we can use a divide and conquer strategy similar to backfitting. The difference to backfitting is, however, that we are able to estimate the smoothing parameters simultaneously with the regression parameters with almost negligible additional effort. We are therefore able to handle problems with more than 1000 parameters and 200000 observations.

The methodology of this paper is included in the public domain program *BayesX*, a software package for Bayesian inference. It may be downloaded including a detailed man-

ual from `http://www.stat.uni-muenchen.de/~lang/bayesx`. As a particular advantage *BayesX* can estimate reasonable complex models and handle fairly large data sets.

We will present examples of STAR models in two applications. In our first application we analyze longitudinal data on the health status of beeches in northern Bavaria. Important influential factors on the health state of trees are e.g. the age of the trees, the canopy density at the stand, calendar time as a surrogate for changing environmental conditions, and the location of the stand. The second application is a space-time analysis of hospital treatment costs based on data from a German private health insurance company.

The remainder of this paper is organized as follows: The next section describes Bayesian GAM's based on one or two dimensional P-splines and discusses extensions to STAR models. Section 2.9 gives details about MCMC inference. Section 2.10 contains some simulation studies in order to gain more insight into the properties of our approach. In Section 2.11 we present two applications on the health status of trees and hospital treatment costs. Section 2.12 concludes and discusses directions for future research.

2.8 Bayesian STAR models

We first describe usual GAM's based on Bayesian P-splines (Subsection 2.8.1). In Subsection 2.8.2 we include interactions into the predictor. Subsection 2.8.3 deals with unit- or cluster specific and spatial heterogeneity. We call a predictor with one or two dimensional nonlinear effects of continuous covariates, time scales, and unit- or cluster specific and spatial heterogeneity a *structured additive predictor* because it still retains an additive structure but is more flexible than the usual predictor in GAM's. Despite the complexity of the predictor we are able to develop a unified framework for the different priors (Subsection 2.8.4).

2.8.1 GAM's based on Bayesian P-Splines

Suppose that observations (y_i, x_i, v_i), $i = 1, \ldots, n$, are given, where y_i is a response variable, $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of continuous covariates and $v_i = (v_{i1}, \ldots, v_{iq})'$ are further (mostly categorical) covariates. Generalized additive models (Hastie and Tibshirani 1990) assume that, given x_i and v_i the distribution of y_i belongs to an exponential family, i.e.

$$p(y_i \,|\, x_i, v_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right) c(y_i, \phi) \tag{2.16}$$

where $b(\cdot)$, $c(\cdot)$, θ_i and ϕ determine the respective distributions. A list of the most common distributions and their specific parameters can be found e.g. in Fahrmeir and Tutz (2001), page 21. The mean $\mu_i = E(y_i \,|\, x_i, v_i)$ is linked to a semiparametric additive predictor η_i by

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + v_i'\gamma. \tag{2.17}$$

Here, h is a known response function and $f_1, \ldots, f_p$ are unknown smooth functions of the continuous covariates and $v_i'\gamma$ represents the strictly linear part of the predictor.

For modeling the unknown functions f_j we follow the approach of Part I of this chapter and use a Bayesian version of P-splines introduced in a frequentist setting by Eilers and Marx (1996) and Marx and Eilers (1998). The approach assumes that the unknown functions can be approximated by a polynomial spline of degree l and with equally spaced knots

$$\zeta_{j0} = x_{j,min} < \zeta_{j1} < \cdots < \zeta_{j,r_j-1} < \zeta_{jr_j} = x_{j,max}$$

over the domain of x_j. The spline can be written in terms of a linear combination of $M_j = r_j + l$ B-spline basis functions (De Boor 1978). Denoting the ρ-th basis function by $B_{j\rho}$, we obtain

$$f_j(x_j) = \sum_{\rho=1}^{M_j} \beta_{j\rho} B_{j\rho}(x_j).$$

By defining the $n \times M_j$ design matrices X_j with the elements in row i and column ρ given by $X_j(i,\rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor (2.17) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V\gamma. \tag{2.18}$$

Here, $\beta_j = (\beta_{j1}, \ldots, \beta_{jM_j})'$, $j = 1, \ldots, p$, correspond to the vectors of unknown regression coefficients. The matrix V is the usual design matrix for linear effects. To overcome the well known difficulties involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 to 40) to ensure enough flexibility, and to introduce a roughness penalty on adjacent regression coefficients to regularize the problem and avoid overfitting. In their frequentist approach they use penalties based on squared k-th order differences. Usually first or second order differences are enough. In our Bayesian approach, we replace first or second order differences with their stochastic analogues, i.e. first or second order random walks defined by

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{2.19}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The amount of smoothness is controlled by the variance parameter τ_j^2 which corresponds to the inverse smoothing parameter in the traditional approach. By defining an additional hyperprior for the variance parameters the amount of smoothness can be estimated simultaneously with the regression coefficients. We assign the conjugate prior for τ_j^2 which is an inverse gamma prior with hyperparameters a_j and b_j, i.e. $\tau_j^2 \sim IG(a_j, b_j)$. Common choices for a_j and b_j are $a_j = 1$ and b_j small, e.g. $b = 0.005$ or $b_j = 0.0005$. Alternatively we may set $a_j = b_j$, e.g. $a_j = b_j = 0.001$. Based on experience from extensive simulation studies we use $a_j = b_j = 0.001$ as our standard choice. Since the results may considerably depend on the choice of a_j and b_j some sort of sensitivity analysis is strongly recommended. For instance, the models under consideration could be re-estimated with (a small) number of different choices for a_j and b_j.

In some situations, a global variance parameter τ_j^2 may be not appropriate, for example if the underlying function is highly oscillating. In such cases the assumption of a global

variance parameter τ_j^2 may be relaxed by replacing the errors $u_{j\rho} \sim N(0, \tau_j^2)$ in (2.19) by $u_{j\rho} \sim N(0, \tau_j^2/\delta_{j\rho})$. The weights $\delta_{j\rho}$ are additional hyperparameters and assumed to follow independent gamma distributions $\delta_{j\rho} \sim G(\frac{\nu}{2}, \frac{\nu}{2})$. This is equivalent to a t-distribution with ν degrees of freedom for β_j (see e.g. Knorr–Held (1996) in the context of dynamic models). As an alternative, *locally adaptive dependent* variances as proposed in Lang et al. (2002) and Jerak and Lang (2005) could be used as well. Our software is capable of estimating such models, but we do not investigate them in the following. However, estimation is straightforward, see Part I of this chapter, Lang et al. (2002) and Jerak and Lang (2005) for details.

2.8.2 Modeling interactions

In many situations, the simple additive predictor (2.17) may be not appropriate because of interactions between covariates. In this section we describe interactions between categorical and continuous covariates, and between two continuous covariates. In the next section, we also discuss interactions between space and categorical covariates. For simplicity, we keep the notation of the predictor as in (2.17) and assume for the rest of the section that x_j is now two dimensional, i.e. $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})'$.

Interactions between categorical and continuous covariates can be conveniently modeled within the varying coefficient framework introduced by Hastie and Tibshirani (1993). Here, the effect of covariate $x_{ij}^{(1)}$ is assumed to vary smoothly over the range of the second covariate $x_{ij}^{(2)}$, i.e.

$$f_j(x_{ij}) = g\left(x_{ij}^{(2)}\right) x_{ij}^{(1)}. \tag{2.20}$$

The covariate $x_{ij}^{(2)}$ is called the effect modifier of $x_{ij}^{(1)}$. The design matrix X_j is given by $diag(x_{1j}^{(1)}, \ldots, x_{nj}^{(1)})X_j^{(2)}$ where $X_j^{(2)}$ is the usual design matrix for splines composed of the basis functions evaluated at the observations $x_{ij}^{(2)}$.

If both interacting covariates are continuous, a more flexible approach for modeling interactions can be based on two dimensional surface fitting. Here, we concentrate on two dimensional P-splines described in Part I of this chapter, see also Wood (2003) for a recent approach based on thin plate splines. We assume that the unknown surface $f_j(x_{ij})$ can be approximated by the tensor product of one dimensional B-splines, i.e.

$$f_j(x_{ij}^{(1)}, x_{ij}^{(2)}) = \sum_{\rho=1}^{M_{1j}} \sum_{\nu=1}^{M_{2j}} \beta_{j,\rho\nu} B_{j,\rho}\left(x_{ij}^{(1)}\right) B_{j,\nu}\left(x_{ij}^{(2)}\right). \tag{2.21}$$

The design matrix X_j is now $n \times (M_{1j} \cdot M_{2j})$ dimensional and consists of products of basis functions. Priors for $\beta_j = (\beta_{j,11}, \ldots, \beta_{j,M_{1j}M_{2j}})'$ are based on spatial smoothness priors common in spatial statistics (see e.g. Besag and Kooperberg 1995). Based on previous experience, we prefer a two dimensional first order random walk constructed from the four nearest neighbors. It is usually defined by specifying the conditional distributions of a

parameter given its neighbors, i.e.

$$\beta_{j\rho\nu} \,|\, \cdot \sim N\left(\frac{1}{4}(\beta_{j\rho-1,\nu} + \beta_{j\rho+1,\nu} + \beta_{j\rho,\nu-1} + \beta_{j\rho,\nu+1}), \frac{\tau_j^2}{4}\right) \tag{2.22}$$

for $\rho = 2, \ldots, M_{1j} - 1$, $\nu = 2, \ldots, M_{2j} - 1$ and appropriate changes for corners and edges. Again, we restrict the unknown function f_j to have mean zero to guarantee identifiability.

Sometimes it is desirable to decompose the effect of the two covariates $x_j^{(1)}$ and $x_j^{(2)}$ into two main effects modeled by one dimensional functions and a two dimensional interaction effect. Then, we obtain

$$f_j(x_{ij}) = f_j^{(1)}\left(x_{ij}^{(1)}\right) + f_j^{(2)}\left(x_{ij}^{(2)}\right) + f_j^{(1|2)}\left(x_{ij}^{(1)}, x_{ij}^{(2)}\right). \tag{2.23}$$

In this case, additional identifiability constraints have to be imposed on the three functions, see Part I of this chapter.

2.8.3 Unobserved heterogeneity

So far, we have considered only continuous and categorical covariates in the predictor. In this section, we relax this assumption by allowing that the covariates x_j in (2.17) or (2.18) are not necessarily continuous. We still pertain the assumption of the preceding section that covariates x_j may be one or two dimensional. Based on this assumptions the models can be considerably extended within a unified framework. We are particularly interested in the handling of unobserved unit- or cluster specific and spatial heterogeneity. Models that can deal with spatial heterogeneity are also called *geoadditive models* (Kammann and Wand 2003).

Unit- or cluster specific heterogeneity

Suppose that covariate x_j is an index variable that indicates the unit or cluster a particular observation belongs to. An example are longitudinal data where x_j is an individual index. In this case, it is common practice to introduce unit- or cluster specific i.i.d. Gaussian random intercepts or slopes, see e.g. Diggle, Haegerty, Liang and Zeger (2002). Suppose x_j can take the values $1, \ldots, M_j$. Then, an i.i.d. random intercept can be incorporated into our framework of structured additive regression by assuming $f_j(m) = \beta_{jm} \sim N(0, \tau_j^2)$, $m = 1, \ldots, M_j$. The design matrix X_j is now a 0/1 incidence matrix with dimension $n \times M_j$. In order to introduce random slopes we assume $x_j = \left(x_j^{(1)}, x_j^{(2)}\right)$ as in Section 2.8.2. Then, a random slope with respect to index variable $x_j^{(2)}$ is defined as $f_j(x_{ij}) = g\left(x_{ij}^{(2)}\right) x_{ij}^{(1)}$ with $g\left(x_{ij}^{(2)}\right) = \beta_{jm} \sim N(0, \tau_j^2)$. The design matrix X_j is given by $diag\left(x_{1j}^{(1)}, \ldots, x_{nj}^{(1)}\right) X_j^{(2)}$ where $X_j^{(2)}$ is again a 0/1 incidence matrix. Note the close similarity between random slopes and varying coefficient models. In fact, random slopes may be regarded as varying coefficient terms with unit- or cluster variable $x_j^{(2)}$ as the effect modifier.

Spatial heterogeneity

To consider spatial heterogeneity, we may introduce a *spatial effect* f_j of location x_j to the predictor. Depending on the application, the spatial effect may be further split up into a spatially correlated (structured) and an uncorrelated (unstructured) effect, i.e. $f_j = f_{str} + f_{unstr}$. The correlated effect f_{str} aims at capturing spatially dependent heterogeneity and the uncorrelated effect f_{unstr} local effects.

For data observed on a regular or irregular lattice a common approach for the correlated spatial effect f_{str} is based on Markov random field (MRF) priors, see e.g. Besag et al. (1991). Let $s \in \{1, \ldots, S_j\}$ denote the pixels of a lattice or the regions of a geographical map. Then, the most simple Markov random field prior for $f_{str}(s) = \beta_{str,s}$ is defined by

$$\beta_{str,s} \,|\, \beta_{str,u}, u \neq s \sim N\left(\sum_{u \in \partial_s} \frac{1}{N_s}\beta_{str,u}, \frac{\tau^2_{str}}{N_s}\right), \tag{2.24}$$

where N_s is the number of adjacent regions or pixels, and ∂_s denotes the regions which are neighbors of region s. Hence, prior (2.24) can be seen as a two dimensional extension of a first order random walk. More general priors than (2.24) are described in Besag et al. (1991). The design matrix X_{str} is a $n \times S_j$ incidence matrix whose entry in the i-th row and s-th column is equal to one if observation i has been observed at location s and zero otherwise.

Alternatively, the structured spatial effect f_{str} could be modeled by two dimensional surface estimators as described in Section 2.8.2. In most of our applications, however, the MRF proves to be superior in terms of model fit.

For the unstructured effect f_{unstr} we may again assume i.i.d. Gaussian random effects with the location as the index variable.

Similar to continuous covariates and index variables we can again define varying coefficient terms, now with the location index as the effect modifier, see e.g. Fahrmeir et al. (2003) and Gamerman, Moreira and Rue (2003) for applications. Models of this kind are known in the geography literature as *geographically weighted regression* (Fotheringham et al. 2002).

2.8.4 General structure of the priors

As we have pointed out, it is always possible to express the vector of function evaluations $f_j = (f_{j1}, \ldots, f_{jn})$ of a covariate effect as the matrix product of a design matrix X_j and a vector of regression coefficients β_j, i.e. $f_j = X_j\beta_j$. It turns out that the smoothness priors for the regression coefficients β_j can be cast into a general form as well. It is given by

$$\beta_j \,|\, \tau^2_j \propto \frac{1}{(\tau^2_j)^{rk(K_j)/2}} \exp\left(-\frac{1}{2\tau^2_j}\beta_j' K_j \beta_j\right), \tag{2.25}$$

where K_j is a *penalty matrix* which depends on the prior assumptions about *smoothness of* f_j and the *type of covariate*. E.g. for a P-spline with a first order random walk penalty

K_j is given by

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

For an i.i.d. random effect the penalty matrix is the identity matrix, i.e. $K_j = I$. For the variance parameter an inverse gamma prior (the conjugate prior) is assumed, i.e. $\tau_j^2 \sim IG(a_j, b_j)$.

The general structure of the priors particularly facilitates the description and implementation of MCMC inference in the next section.

2.9 Bayesian inference via MCMC

Bayesian inference is based on the posterior of the model which is given by

$$\begin{aligned} p(\alpha \,|\, y) \quad \propto \quad & L(y, \beta_1, \tau_1^2, \ldots, \beta_p, \tau_p^2, \gamma) \\ & \prod_{j=1}^{p} \frac{1}{(\tau_j^2)^{rk(K_j)/2}} \exp\left(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j\right) \prod_{j=1}^{p} (\tau_j^2)^{-a_j - 1} \exp\left(-\frac{b_j}{\tau_j^2}\right), \end{aligned}$$

where α is the vector of all parameters in the model. The likelihood $L(\cdot)$ is a product of the individual likelihoods (2.16). Since the posterior is analytically intractable we make use of Markov chain Monte Carlo (MCMC) simulation techniques. Models with Gaussian responses are already covered in Part I of this chapter. Here, the main focus is on methods applicable for general distributions from an exponential family. We first develop in Section 2.9.1 several sampling schemes based on iteratively weighted least squares (IWLS) used for estimating generalized linear models (Fahrmeir and Tutz 2001). For many models with (multi)categorical responses alternative sampling schemes can be developed by considering their latent utility representations (Section 2.9.2). In either case, MCMC simulation is based on drawings from full conditionals of blocks of parameters, given the rest and the data. We use the blocks $\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2, \gamma$.

An alternative to MCMC techniques is proposed in Fahrmeir et al. (2004). Here, mixed model representations and inference techniques are used for estimation. The drawback is that models with a large number of parameters and/or observations as well as multivariate responses can not be handled by the approach.

2.9.1 Updating by iteratively weighted least squares (IWLS) proposals

The basic idea is to combine Fisher scoring or IWLS (e.g. Fahrmeir and Tutz 2001) for estimating regression parameters in generalized linear models, and the Metropolis-Hastings algorithm. More precisely, the goal is to approximate the full conditionals of regression

parameters β_j and γ by a Gaussian distribution, obtained by accomplishing *one* Fisher scoring step in every iteration of the sampler. Suppose we want to update the regression coefficients β_j of the function f_j with current state β_j^c of the chain. Denote η^c the current predictor based on the current regression coefficients β_j^c. Then, according to IWLS, a new value β_j^p is proposed by drawing a random number from the multivariate Gaussian proposal distribution $q(\beta_j^c, \beta_j^p)$ with precision matrix and mean

$$P_j = X_j' W(\eta^c) X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' W(\eta^c)(\tilde{y}(\eta^c) - \tilde{\eta}^c). \tag{2.26}$$

The matrix $W(\eta^c) = diag(w_1(\eta^c), \ldots, w_n(\eta^c))$ and the vector $\tilde{y}(\eta^c) = (\tilde{y}_1(\eta^c), \ldots, \tilde{y}_1(\eta^c))'$ contain the usual weights and working observations for IWLS with $w_i^{-1}(\eta_i^c) = b''(\theta_i)\{g'(\mu_i)\}^2$ and $\tilde{y}(\eta_i^c) = \eta_i^c + (y_i - \mu_i)g'(\mu_i)$. The weights and the working observations depend on the current predictor η^c which in turn depends on the current state β_j^c. The vector $\tilde{\eta}^c$ is the part of the predictor associated with all remaining effects in the model. The proposed new value β_j^p is accepted with probability

$$\alpha(\beta_j^c, \beta_j^p) = \frac{L(y, \ldots, \beta_j^p, \ldots, \gamma^c) p(\beta_j^p \mid (\tau_j^2)^c) q(\beta_j^p, \beta_j^c)}{L(y, \ldots, \beta_j^c, \ldots, \gamma^c) p(\beta_j^c \mid (\tau_j^2)^c) q(\beta_j^c, \beta_j^p)}. \tag{2.27}$$

The computation of the likelihood $L(y, \ldots, \beta_j^p, \ldots, \gamma^c)$ and the proposal density $q(\beta_j^p, \beta_j^c)$ is based on the current predictor η^c where $X_j\beta_j^c$ is exchanged by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - \beta_j^c)$. In order to emphasize implementation aspects and details we use here and elsewhere a pseudo code like notation. Note that the computation of $q(\beta_j^p, \beta_j^c)$ requires to recompute P_j and m_j. If the proposal is accepted we set $\beta_j^c = \beta_j^p$, otherwise we keep the current β_j^c and exchange $X_j\beta_j^p$ in η^c by $X_j\beta_j^c$.

A slightly different sampling scheme uses the current posterior mode approximation m_j^c rather than β_j^c for computing the IWLS weight matrix W and the transformed responses $\tilde{y}$ in (2.26). More precisely, we first replace $X_j\beta_j^c$ in the current predictor η^c by $X_j m_j^c$, i.e. $\eta^c = \eta^c + X_j(m_j^c - \beta_j^c)$. The vector m_j^c is the mean of the proposal distribution used in the last iteration of the sampler. We proceed by drawing a proposal β_j^p from the Gaussian distribution with covariance and mean (2.26). The difference of using m_j^c rather than β_j^c in η^c is that the proposal is *independent* of the current state of the chain, i.e. $q(\beta_j^c, \beta_j^p) = q(\beta_j^p)$. Hence, it is not required to recompute P_j and m_j when computing the proposal density $q(\beta_j^p, \beta_j^c)$ in (2.27). The computation of the likelihood $L(y, \ldots, \beta_j^p, \ldots, \gamma^c)$ is again based on the current predictor η^c where $X_j m_j^c$ is exchanged by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - m_j^c)$. If the proposal is accepted we set $\beta_j^c = \beta_j^p$, otherwise we keep the current β_j^c and exchange $X_j\beta_j^p$ in η^c by $X_j\beta_j^c$. The last step is to set $m_j^c = m_j$.

The advantage of the updating scheme based on the current mode approximation m_j^c is that acceptance rates are considerably higher compared to the sampling scheme based on the current β_j^c. This is particularly important for updating spatial effects based on Markov random field priors because of the usually high dimensional parameter vector β_j.

It turns out that convergence to the stationary distribution can be slow for both algorithms because of inappropriate starting values for the β_j. As a remedy, we initialize

the Markov chain with posterior mode estimates which are obtained from a backfitting algorithm with fixed and usually large values for the variance parameters.

Note, that the posterior precision matrix P_j in (2.26) is a band matrix or can be at least transformed into a matrix with band structure. For one dimensional P-splines, the band size is $max\{\text{degree of spline}, \text{order of differences}\}$, for two dimensional P-splines the band size is $M_j \cdot l + l$, and for i.i.d. random effects the posterior precision matrix is diagonal. For a Markov random field, the precision matrix is not a priori a band matrix but sparse. It can be transformed into a band matrix (with differing band size in every row) by reordering the regions using the reverse Cuthill Mc-Kee algorithm (see George and Liu (1981) p. 58 ff). Hence, random numbers from the (high dimensional) proposal distributions can be efficiently drawn by using matrix operations for sparse matrices, in particular Cholesky decompositions. In our implementation we use the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981), see also Rue (2001) and Part I of this chapter.

Updating of the variance parameters τ_j^2 is straightforward because their full conditionals are inverse gamma distributions with parameters

$$a_j' = a_j + \frac{rank(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\beta_j' K_j \beta_j. \tag{2.28}$$

We finally summarize the second proposed IWLS updating scheme based on the current mode approximations m_j^c and m_γ^c. The first IWLS updating scheme based on the current β_j^c and γ_j^c is very similar and therefore omitted. In what follows, the order of evaluations is stressed because it is crucial for computational efficiency.

Sampling scheme 1 (IWLS proposals based on current mode):

For implementing the sampling scheme described below the quantities β_j^c, β_j^p, $(\tau_j^2)^c$, γ^c, m_j^c, X_j, m_j, P_j and η^c must be created and enough memory must be allocated to store them.

1. *Initialization:*
   Compute the posterior modes m_j^c and γ^c for $\beta_1, \ldots, \beta_p$ and γ given fixed variance parameters $\tau_j^2 = c_j$, (e.g. $c_j = 10$). The mode is computed via backfitting within Fisher scoring. Use the posterior mode estimates as the current state β_j^c, γ^c of the chain. Set $(\tau_j^2)^c = c_j$. Store the current predictor in the vector η^c.

2. *For $j = 1, \ldots, p$ update β_j:*

   - Compute the likelihood $L(y, \ldots, \beta_j^c, \ldots, \gamma^c)$.
   - Exchange $X_j\beta_j^c$ in the current predictor η^c by $X_j m_j^c$, i.e. $\eta^c = \eta^c + X_j(m_j^c - \beta_j^c)$.
   - Draw a proposal β_j^p from the Gaussian proposal density $q(\beta_j^c, \beta_j^p)$ with mean m_j and precision matrix P_j given in (2.26).
   - Exchange $X_j m_j^c$ in the current predictor η^c by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - m_j^c)$.

- Compute the likelihood $L(y, \ldots, \beta_j^p, \ldots, \gamma^c)$.
- Compute $q(\beta_j^p, \beta_j^c)$, $q(\beta_j^c, \beta_j^p)$, $p(\beta_j^c \mid (\tau_j^2)^c)$ and $p(\beta_j^p \mid (\tau_j^2)^c)$.
- Accept β_j^p as the new state of the chain β_j^c with probability (2.27). If the proposal is rejected exchange $X_j\beta_j^p$ in the current predictor η^c by $X_j\beta_j^c$, i.e. $\eta^c = \eta^c + X_j(\beta_j^c - \beta_j^p)$.
- Set $m_j^c = m_j$.

3. *Update fixed effects parameters:*
   Update fixed effects parameters by similar steps as for updating of β_j.

4. *For $j = 1, \ldots, p$ update variance parameters τ_j^2:*
   Variance parameters are updated by drawing from the inverse gamma full conditionals with hyperparameters given in (2.28). Obtain $(\tau_j^2)^c$.

Usually convergence and mixing of Markov chains is excellent with both variants of IWLS proposals. If, however, the effect of two covariates $x_j^{(1)}$ and $x_j^{(2)}$ is decomposed into main effects and a two dimensional interaction effect as in (2.23), severe convergence problems for the variance parameter of the interaction effect are the rule. To overcome the difficulties, we follow Knorr–Held and Rue (2002) who propose to construct a *joint proposal* for the parameter vector β_j and the corresponding variance parameter τ_j^2, and to simultaneously accept/reject (β_j, τ_j^2). We illustrate the updating scheme with IWLS proposals based on the current state of the chain β_j^c. We first sample $(\tau_j^2)^p$ from a proposal distribution for τ_j^2, and subsequently draw from the IWLS proposal for the corresponding regression parameters given the proposed $(\tau_j^2)^p$. The proposal distribution for τ_j^2 may depend on the current state $(\tau_j^2)^c$ of the variance, but *must be independent* of β_j^c. As suggested by Knorr–Held and Rue (2002), we construct the proposal by multiplying the current state $(\tau_j^2)^c$ by a random variable z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, where $f > 1$ is a tuning constant. The density is independent of the regression parameters and the joint proposal for (β_j, τ_j^2) is the product of the two proposal densities. We tune f in the burn in period to obtain acceptance probabilities of 30-60%. The acceptance probability is given by

$$\alpha(\beta_j^c, (\tau_j^2)^c, \beta_j^p, (\tau_j^2)^p) = \frac{L(y, \ldots, \beta_j^p, (\tau_j^2)^p, \ldots, \gamma^c)}{L(y, \ldots, \beta_j^c, (\tau_j^2)^c, \ldots, \gamma^c)} \frac{p(\beta_j^p \mid (\tau_j^2)^p) p((\tau_j^2)^p)}{p(\beta_j^c \mid (\tau_j^2)^c) p((\tau_j^2)^c)} \frac{q(\beta_j^p, \beta_j^c)}{q(\beta_j^c, \beta_j^p)}. \quad (2.29)$$

Computation of the acceptance probability requires the evaluation of the normalizing constant of the IWLS proposal which is given by $|P_j|^{0.5}$. The determinant of P_j can be computed without significant additional effort as a by-product of the Cholesky decomposition. Note also that the proposal ratio of the variance parameter cancels out. Summarizing, we obtain the following sampling scheme:

Sampling scheme 2: IWLS proposals, update β_j's and τ_j^2's in one block:

1. *Initialization:*
   Compute the posterior mode for $\beta_1, \ldots, \beta_p$ and γ given fixed variance parameters $\tau_j^2 = c_j$, (e.g. $c_j = 10$). Use the posterior mode estimates as the current state β_j^c, γ^c of the chain. Set $(\tau_j^2)^c = c_j$. Store the current predictor in the vector η^c.

2. *For $j = 1, \ldots, p$ update β_j, τ_j^2:*

   - Compute the likelihood $L(y, \ldots, \beta_j^c, (\tau_j^2)^c, \ldots, \gamma^c)$.
   - *Propose new τ_j^2:*
     Sample a random number z with density proportional to $1 + 1/z$ on the interval $[1/f, f]$, $f > 1$. Set $(\tau_j^2)^p = z \cdot (\tau_j^2)^c$ as the proposed new value for the jth variance parameter.
   - Draw a proposal β_j^p from $q(\beta_j^c, \beta_j^p)$ with mean $m_j((\tau_j^2)^p)$ and precision matrix $P_j((\tau_j^2)^p)$ defined in (2.26).
   - Exchange $X_j\beta_j^c$ in the current predictor η^c by $X_j\beta_j^p$, i.e. $\eta^c = \eta^c + X_j(\beta_j^p - \beta_j^c)$.
   - Compute the likelihood $L(y, \ldots, \beta_j^p, (\tau_j^2)^p, \ldots, \gamma^c)$.
   - Compute $q(\beta_j^c, \beta_j^p)$, $p(\beta_j^c \mid (\tau_j^2)^c)$, $p(\beta_j^p \mid (\tau_j^2)^p)$, $p((\tau_j^2)^c)$ and $p((\tau_j^2)^p)$.
   - Based on the current predictor η^c compute again P_j and m_j defined in (2.26) and use these quantities to compute $q(\beta_j^p, \beta_j^c)$.
   - Accept $\beta_j^p, (\tau_j^2)^p$ with probability (2.29). If the proposals are rejected exchange $X_j\beta_j^p$ in the current predictor η^c by $X_j\beta_j^c$, i.e. $\eta^c = \eta^c + X_j(\beta_j^c - \beta_j^p)$.

3. *Update fixed effects parameters*

We conclude this section with two remarks:

- *Suppressing the computation of weights:*
  A natural source for saving computing time is to avoid the (re)computation of the IWLS weight matrix W (and thereby the matrix $X'WX$) in every iteration of the sampler. A possible strategy is to recompute the weights only every t-th iteration. It is even possible to keep the weights fixed after the burn in period. Our experience suggests that for most distributions the acceptance rates and the mixing of the chains is almost unaffected by keeping the weights fixed.

- *Multinomial logit models:*
  Our sampling schemes for univariate response distributions can be readily extended to the widely used multinomial logit model. Suppose that the response is multicategorical with k categories, i.e. $y_i = (y_{i1}, \ldots, y_{ik})'$ where $y_{ir} = 1$ if the r-th category has been observed and zero otherwise. The multinomial logit model assumes that given covariates and parameters the responses y_i are multinomial distributed,

i.e. $y_i \,|\, x_i, v_i \sim MN(1, \pi_i)$ where $\pi_i = (\pi_{i1}, \ldots, \pi_{ik})'$. The covariates enter the model by assuming

$$\pi_{ir} = \frac{\exp(\eta_{ir})}{1 + \sum_{l=1}^{k-1} \exp(\eta_{il})}, \qquad r = 1, \ldots, k-1,$$

where

$$\eta_{ir} = f_{1r}(x_{i1r}) + \cdots + f_{pr}(x_{ipr}) + v_{ir}'\gamma_r, \qquad r = 1, \ldots, k-1,$$

are structured additive predictors of the covariates (as described in Section 2.8.3). Note that our formulation allows category specific covariates. For identifiability reasons one category must be chosen as the reference category, without loss of generality we use the last category, i.e. $\pi_{ik} = 1 - \sum_{r=1}^{k-1} \pi_{ir}$. Abe (1999) (see also Hastie and Tibshirani (1990), Ch. 8.1) describes IWLS in combination with backfitting to estimate a multinomial logit model with additive predictors. Here, the transformed responses

$$\tilde{y}_{ir} = \eta_{ir} + \frac{1}{\pi_{ir}(1 - \pi_{ir})}(y_{ir} - \pi_{ir})$$

and weights

$$w_{ir} = \pi_{ir}(1 - \pi_{ir})$$

are used for subsequent backfitting to obtain estimates of the unknown functions. We can use the same transformed responses and weights for our IWLS proposals and the sampling algorithms described above readily extend to multicategorical logit models. E.g. sampling scheme 2 successively updates parameters in the order $(\beta_{11}, \tau^2_{11})$, $(\beta_{12}, \tau^2_{12})$,..., $(\beta_{1p}, \tau^2_{1p})$,$\gamma_1$,...,$(\beta_{k-1,1}, \tau^2_{k-1,1})$,...,$(\beta_{k-1,p}, \tau^2_{k-1,p})$, γ_{k-1}, where β_{jr}, τ^2_{jr} correspond to the regression parameters and variance parameter of the j-th nonlinear function f_{jr} of category r. Again, computing time may be saved by avoiding the computation of the weights in every iteration of the sampler.

2.9.2 Inference based on latent utility representations of categorical regression models

For models with categorical responses alternative sampling schemes based on latent utility representations can be developed. The seminal paper by Albert and Chib (1993) develops algorithms for probit models with ordered categorical responses. The case of probit models with unordered multicategorical responses is dealt with e.g. in Chen and Dey (2000) or Fahrmeir and Lang (2001b). Recently, another important data augmentation approach for binary and multinomial logit models has been presented by Holmes and Held (2004). The adaption of these sampling schemes to the models discussed in this paper is more or less straightforward. We briefly illustrate the concept for binary data, i.e. y_i takes only the values 0 or 1. We first assume a probit model. Conditional on the covariates and the parameters, y_i follows a Bernoulli distribution $y_i \sim B(1, \mu_i)$ with conditional

mean $\mu_i = \Phi(\eta_i)$ where Φ is the cumulative distribution function of a standard normal distribution. Introducing latent variables

$$U_i = \eta_i + \epsilon_i, \tag{2.30}$$

with $\epsilon_i \sim N(0,1)$, we define $y_i = 1$ if $U_i \geq 0$ and $y_i = 0$ if $U_i < 0$. It is easy to show that this corresponds to a binary probit model for the y_i's. The posterior of the model augmented by the latent variables depends now on the extra parameters U_i. Correspondingly, additional sampling steps for updating the U_i's are required. Fortunately, sampling the U_i's is relatively easy and fast because the full conditionals are truncated normal distributions. More specifically, $U_i \,|\, \cdot \sim N(\eta_i, 1)$ truncated at the left by 0 if $y_i = 1$ and truncated at the right if $y_i = 0$. Efficient algorithms for drawing random numbers from a truncated normal distribution can be found in Geweke (1991) or Robert (1995). The advantage of defining a probit model through the latent variables U_i is that the full conditionals for the regression parameters β_j (and γ) are Gaussian with precision matrix and mean given by

$$P_j = X_j' X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' (U - \tilde{\eta}).$$

Hence, the efficient and faster sampling schemes developed for Gaussian responses can be used with slight modifications. Updating of β_j and γ can be done exactly as described in Part I of this chapter using the current values U_i^c of the latent utilities as (pseudo) responses.

For binary logit models, the sampling schemes become more complicated and less efficient (regarding computing time). A logit model can be expressed in terms of latent utilities by assuming $\epsilon_i \sim N(0, \lambda_i)$ in (2.30) with $\lambda_i = 4\psi_i^2$, where ψ_i follows a Kolmogorov-Smirnov distribution (Devroye 1986). Hence, ϵ_i is a scale mixture of normal form with a marginal logistic distribution (Andrews and Mallows (1974)). The main difference to the probit case is that additional parameters λ_i must be sampled. Holmes and Held (2004) propose to joint update U_i, λ_i by first drawing from the marginal distribution $p(U_i \,|\, \beta_1, \ldots, \beta_p, \gamma, y_i)$ of the $U_i's$ followed by drawing from $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p, \gamma)$. The marginal densities of the $U_i's$ are truncated logistic distributions while $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p)$ is not of standard form. Detailed algorithms for sampling from both distributions can be found in Holmes and Held (2004), appendix A3 and A4. Similar to probit models the full conditionals for the regression parameters β_j are Gaussian with precision matrix and mean given by

$$P_j = X_j' \Lambda^{-1} X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} X_j' \Lambda^{-1} (U - \tilde{\eta}) \tag{2.31}$$

with weight matrix $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$. This updating scheme is considerably slower than the scheme for probit models. It is also much slower than the IWLS schemes discussed in Section 2.9.1. The reason is that drawing random numbers from $p(\lambda_i \,|\, U_i, \beta_1, \ldots, \beta_p, \gamma)$ is based on rejection sampling and therefore time consuming. Moreover, the matrix products $X_j' \Lambda^{-1} X_j$ in (2.31) must be recomputed in every iteration of the sampler. The advantage of the updating scheme is, however, that the acceptance rates will always be unity regardless of the number of parameters. This may be a particular advantage when estimating high dimensional Markov random fields.

2.9.3 Future prediction with Bayesian P-Splines

In our second application on health insurance data it is necessary to get estimates of a function f_j outside the range of x_j. More specifically, we are interested in a one year ahead prediction of a time trend. Future prediction with Bayesian P-splines is obtained in a similar way as described in Besag, Green, Higdon and Mengersen (1995) for simple random walks. The spline can be defined outside the range of x_j by defining additional equidistant knots and by computing the corresponding B-spline basis functions. Samples of the additional regression parameters $\beta_{j,M_j+1}, \beta_{j,M_j+2}, \ldots$ are obtained by continuing the random walks in (2.19). E.g. for a second order random walk samples $\beta^{(t)}_{j,M_j+1}$, $t = 1, 2, 3, \ldots$ are obtained through $\beta^{(t)}_{j,M_j+1} \sim N(2\beta^{(t)}_{j,M_j} - \beta^{(t)}_{j,M_j-1}, (\tau^2_j)^{(t)})$, i.e. the samples of β_{j,M_j}, β_{j,M_j-1} and τ^2_j are inserted into (2.19). Samples of additional parameters $\beta_{j,M_j+2}, \ldots$ are computed accordingly.

2.10 Simulations

We have carried out several simulation studies to gain insight into the properties of our models and sampling schemes. We first considered usual functions (e.g. sine, linear, quadratic etc.) of continuous covariates, similar to Section 2.4 of Part I of this chapter. In turned out that our approach is competitive but surprising results did not occur. The presentation of results is therefore omitted. Section 2.10.1 aims at demonstrating that our sampling schemes for multinomial logit models work satisfactorily. A comparison with other approaches in the literature for two dimensional surface estimation is given in Section 2.10.2. In Fahrmeir et al. (2004) we carefully compare the full Bayesian approach presented here with empirical Bayesian inference based on mixed model technology. Here, complex models with STAR predictor are used.

2.10.1 Multinomial logit models

In order to demonstrate the practicability of our sampling schemes for multinomial logit models, we used a three categorical model, i.e.

$$y_i = (y_{i1}, y_{i2}, y_{i3})' \sim M_3(n_i, (\pi_{i1}, \pi_{i2}, \pi_{i3}))$$

with predictors

$$\eta_{i1} = f_{11}(x_{i1}) + f_{12}(x_{i2})$$

and

$$\eta_{i2} = f_{21}(x_{i1}) + f_{22}(x_{i2}).$$

For the functions we chose

$$f_{11}(x_{i1}) = \cos(\pi \cdot x_{i1}/3), \qquad f_{12}(x_{i2}) = \sin(\pi \cdot (x_{i2} \cdot 2 - 1))$$

and

$$f_{21}(x_{i1}) = x_{i1}/3, \qquad f_{22}(x_{i2}) = \sin(2 \cdot \pi \cdot (x_{i2} \cdot 2 - 1)).$$

The values of x_1 are chosen on an equidistant grid of 100 design points between -3 and 3, the values of x_2 are chosen between 0 and 1. To assess the dependence of results on the sample size we used $n_i = 5, 10, 20$ corresponding to sample sizes $n = 500, 1000, 2000$. We simulated 250 replications for every sample size. We are not aware of any other publicly available software for fitting semiparametric multinomial logit models. Hence, comparisons with competing approaches were not possible.

Estimates are based on cubic P-splines with 20 knots and second order random walk penalty. For the hyperparameters a_j and b_j we tested $a_j = 1, b_j = 0.005$ and $a_j = b_j = 0.001$. Estimates for both choices are relatively similar with slightly better results for the choice $a_j = b_j = 0.001$ which is in agreement with the findings in Fahrmeir et al. (2004). In the following the presentation of results is restricted to the choice $a_j = b_j = 0.001$.

Figure 2.10 shows boxplots of the empirical $log(MSE)$ for the four estimated functions. Figures 2.11 and 2.12 display function estimates for the four functions and different sample sizes averaged over the 250 replications. Finally, Table 2.4 investigates the coverage of pointwise credible intervals for nominal levels of 80 and 95 percent. Using MCMC simulation techniques, credible intervals are estimated by computing the respective quantiles of the sampled function evaluations. We can draw the following conclusions:

- The quality of estimates depends on the sample size with MSE's decreasing with increasing sample size. For $n = 500$ function estimates for functions f_{12} and particularly f_{22} are considerably biased. The bias decreases, however, with increasing sample size. For $n = 2000$ observations all functions are estimated with negligible bias.

- The quality of estimates depends on the curvature of the functions with increased bias for functions with higher curvature.

- The coverage of pointwise credible intervals depends on the bias of the point estimates. If the estimates are biased, as is the case for f_{12}, f_{22} and $n = 500$, the coverage rates are slightly below the nominal level. Otherwise, the coverage rates are close to or above the nominal levels.

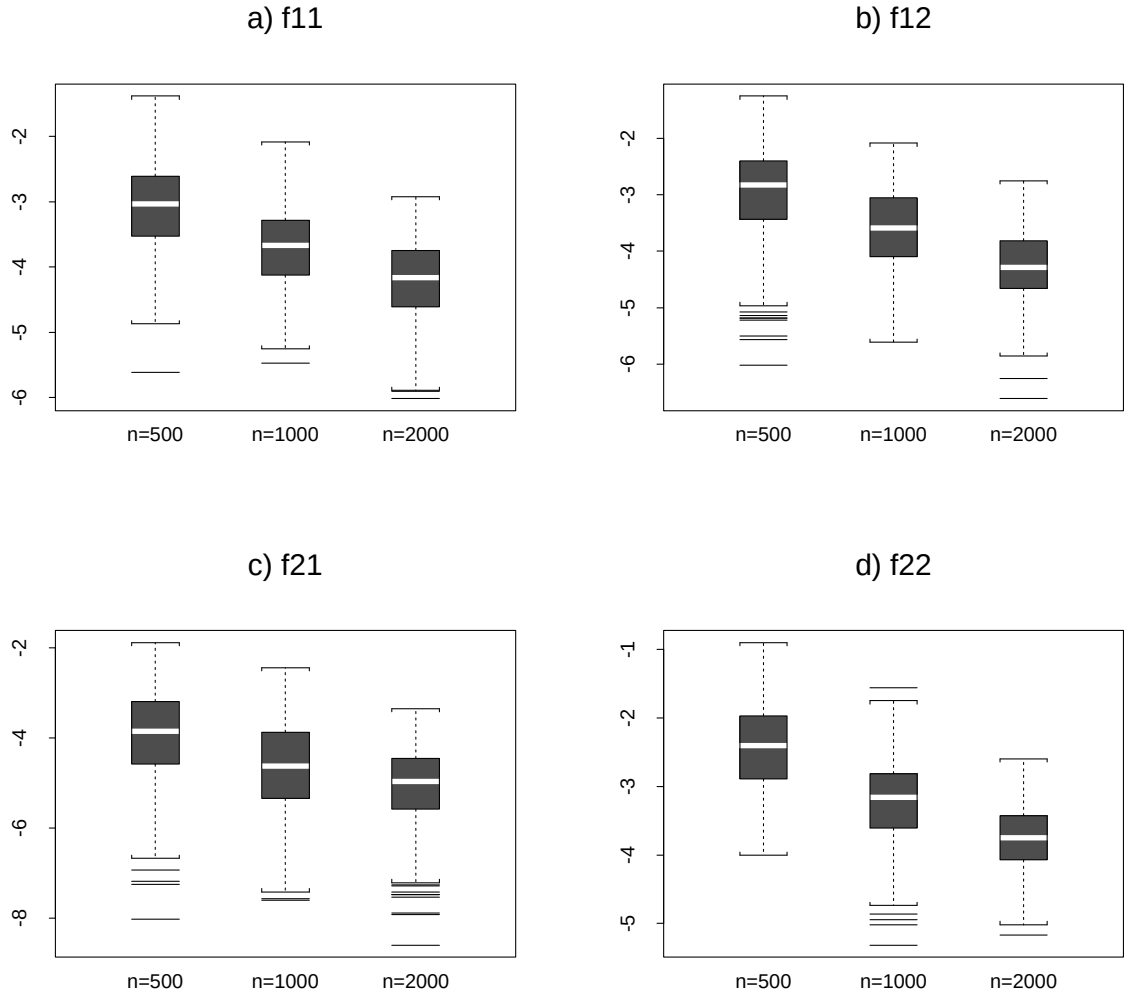


Figure 2.10: Simulation study on multinomial logit models. The panels a)-d) show boxplots of the empirical log(MSE) for the four functions f_{11}, f_{12}, f_{21} and f_{22}.

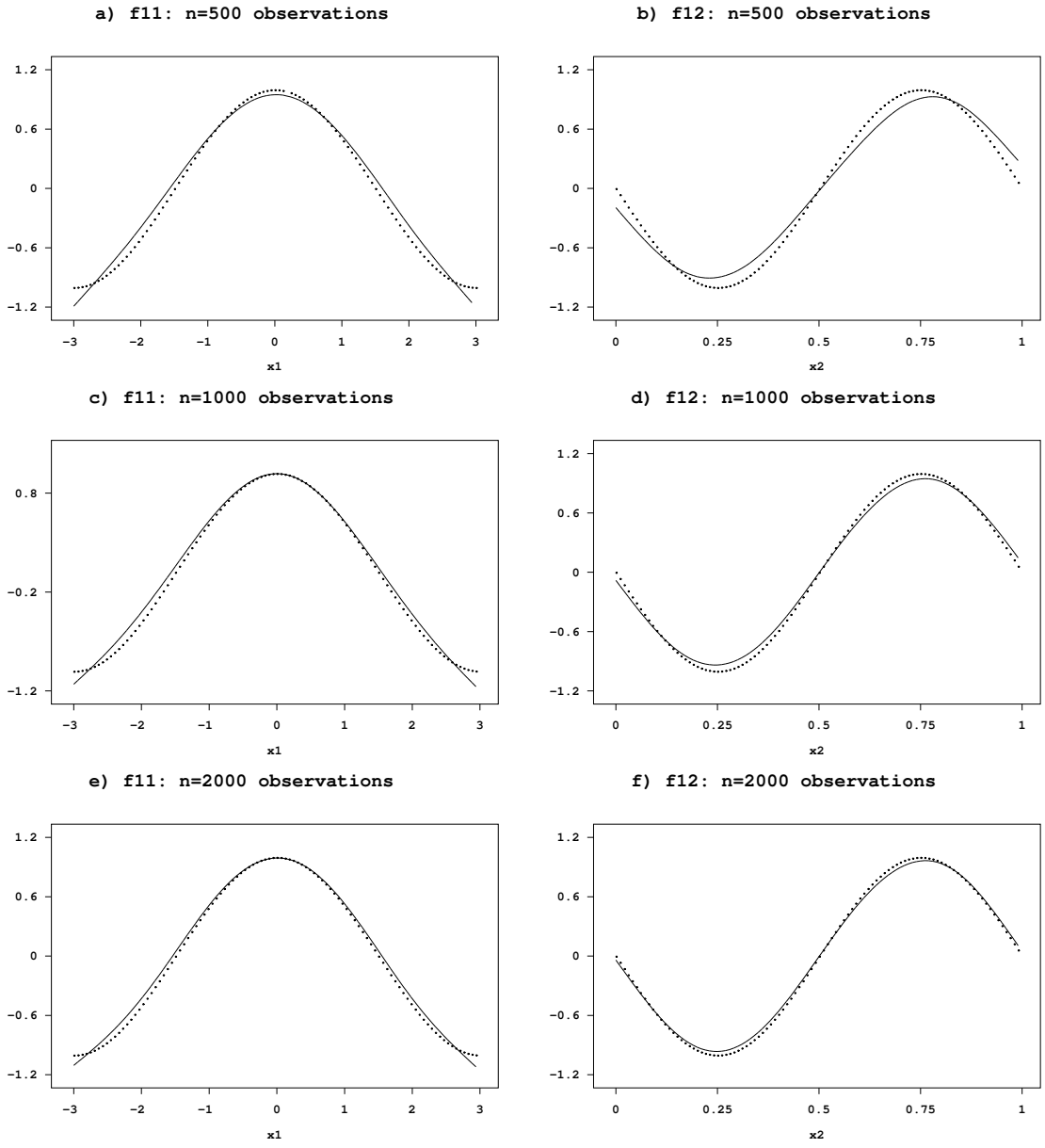


Figure 2.11: Simulation study on multinomial logit models. The left panels show function estimates for f_{11} averaged over the 250 replications. The right panels show function estimates for f_{12}. For comparison the true functions are additionally included (solid lines).

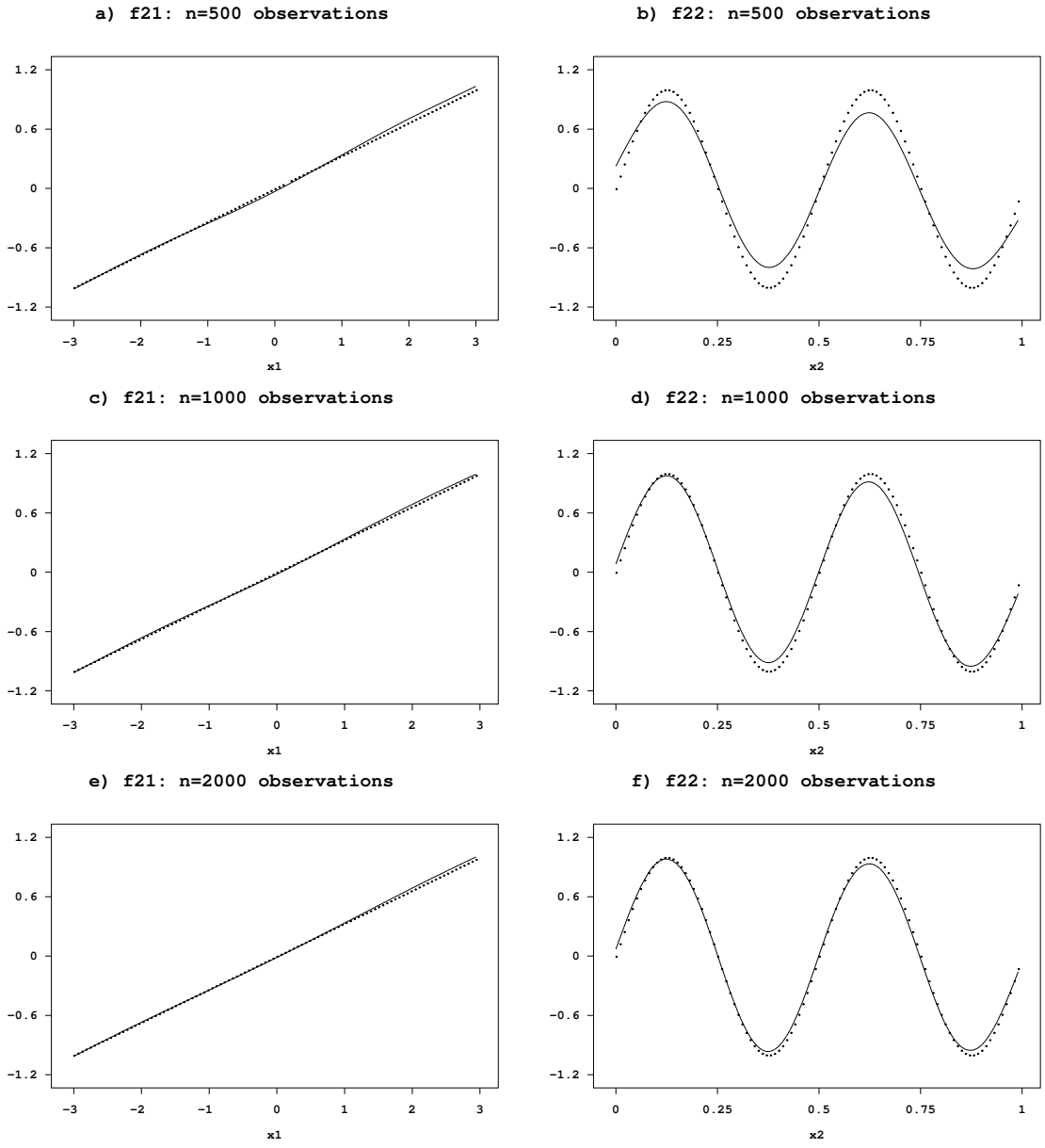


Figure 2.12: Simulation study on multinomial logit models. The left panels show function estimates for f_{21} averaged over the 250 replications. The right panels show function estimates for f_{22}. For comparison the true functions are additionally included (solid lines).

Table 2.4: Simulation study on multinomial logit models. Average coverage rates of pointwise 80% and 95% credible intervals.

		Average coverage 80%	Average coverage 95%
n=500	f_{11}	81	96.1
	f_{12}	79.6	94.5
	f_{21}	85	97.7
	f_{22}	78.8	93.8
n=1000	f_{11}	82.7	96
	f_{12}	81.6	95.2
	f_{21}	87.1	97.4
	f_{22}	83.2	96.2
n=2000	f_{11}	81	96
	f_{12}	84.1	96.7
	f_{21}	85.9	98.1
	f_{22}	83.6	96.6

2.10.2 Two dimensional surface estimation

In this section we study the performance of two dimensional P-splines and compare them with other approaches in the literature. We simulated data from two logit models. The respective predictors contain only a two dimensional surface. For the first model we used

$$f_1(x_1, x_2) = 1.5 \cdot (N(\mu_1, \Sigma_1, x_1, x_2) + N(\mu_2, \Sigma_2, x_1, x_2) - 1.406)$$

where $N(\mu, \Sigma, x_1, x_2)$ denotes a bivariate normal density with mean μ and covariance matrix Σ evaluated at x_1 and x_2. We set $\mu_1 = (0.25, 0.75)'$, $\Sigma_{1,11} = \Sigma_{1,22} = 0.05$, $\Sigma_{1,12} = \Sigma_{1,21} = 0.01$, $\mu_2 = (0.75, 0.25)'$, $\Sigma_{2,11} = \Sigma_{2,22} = 0.1$ and $\Sigma_{2,12} = \Sigma_{2,21} = 0.01$. A similar function has been used in Kohn et al. (2001). The covariates (x_1, x_2) were simulated uniformly over the unit square. The sample size is $n = 600$ observations.

In the second model we used a function from Wood, Kohn, Shively and Jiang (2002). It is given by

$$f_2(x_1, x_2) = 1.9 \cdot [1.35 + \exp(x_1) \cdot \sin(13 \cdot (x_1 - 0.6)^2) \cdot \exp(-x_2) \cdot \sin(7 \cdot x_2)] - 3.5,$$

where x_1, x_2 are uniformly distributed on a regular and equidistant 20 by 20 grid. The sample size is $n = 400$ observations. Both functions f_1 and f_2 are shown in Figure 2.13.

We simulated 250 replications and applied the following estimators:

- Two dimensional Bayesian cubic P-splines on a 12 by 12 grid of inner knots (henceforth BP). We examined both variants of IWLS proposals and the data augmentation scheme by Holmes and Held (2004) and obtained identical results as required. We tested the same choices for the hyperparameters of the variance as in Section 2.10.1. Again, the results are quite insensitive to the choice of hyperparameters but for $a_j = b_j = 0.001$ slightly better results are obtained. In the following, the presentation is restricted to this case.

- Thin plate splines proposed by Wood (2003) (henceforth TPS).

- Smoothing splines as implemented in the software package grkpack (available at http://lib.stat.cmu.edu), see Wang (1995).

- Loess as implemented in S-plus. The smoothing parameter is chosen by stepwise selection with the S-plus procedure step.gam.

- Clive Loader's locfit with fixed smoothing parameter (henceforth locfitfix).

- Clive Loader's locfit with adaptive smoothing parameter, see Cleveland and Grosse (1991) and Loader (1997).

Figure 2.14 shows boxplots of $log(MSE)$ for the various estimators. Figure a) corresponds to model one and Figure b) to model two. The average coverage rates of pointwise credible intervals can be found in Table 2.5. We draw the following conclusions:

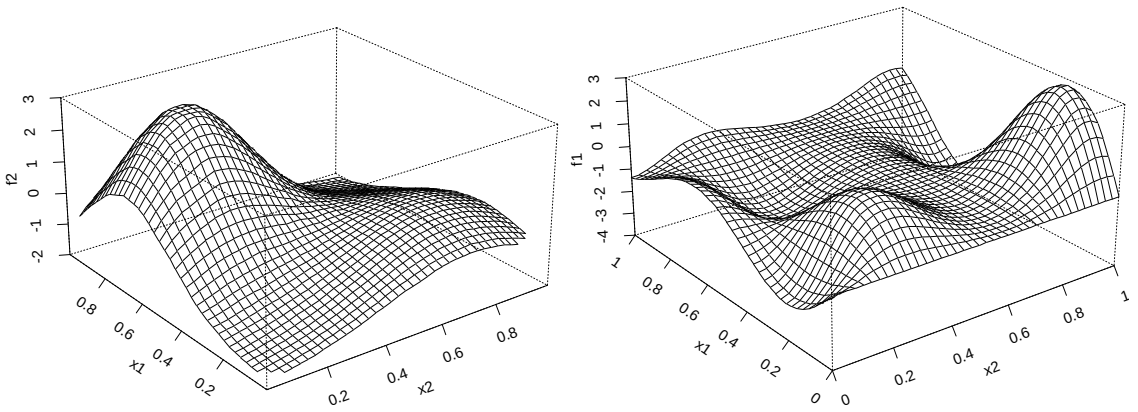


Figure 2.13: Simulation study on two dimensional surface estimation. True functions f_1 and f_2 used for simulations.

- In terms of the MSE measure Bayesian P-splines are competitive. Slightly better results are obtained with Wood's thin plate splines.
- The coverage of pointwise credible intervals is close to or slightly above the nominal level for BP. Acceptable coverage rates are also obtained with TPS. The coverage rates of the other estimators are usually far below the nominal level.

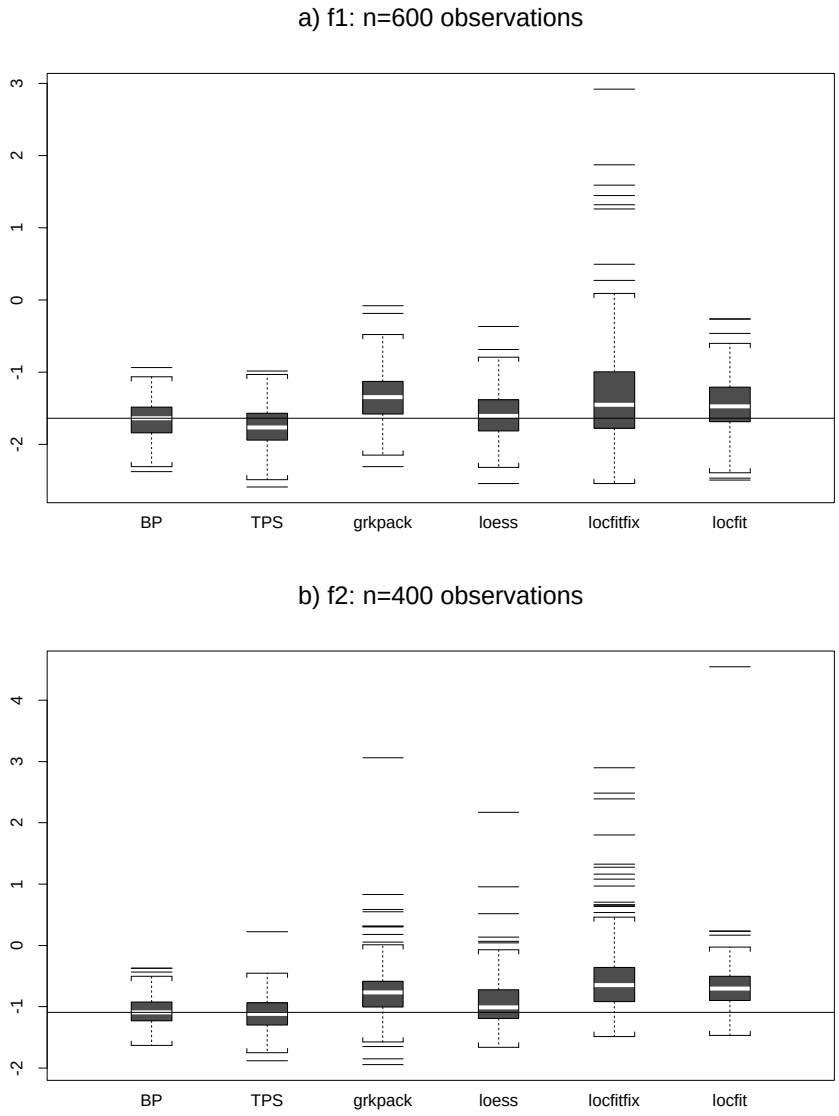


Figure 2.14: Simulation study on two dimensional surface estimation. Boxplots of $log(MSE)$ for the various surface estimators. From left to right the boxplots correspond to BP, TPS, grkpack, loess, locfitfix and locfit.

Table 2.5: Simulation study on two dimensional surface estimation. Average coverage rates of pointwise 80% and 95% credible intervals.

	Ave. coverage 80% (f_1)	Mean coverage 95% (f_1)
BP	83%	96.6%
TPS	75.1%	92.3%
grkpack	68.4%	87.4%
loess	69.5%	88.3%
locfitfix	63.9%	81.9%
locfit	56.9%	74.5%
	Ave. coverage 80% (f_2)	Mean coverage 95% (f_2)
BP	81.9%	95.5%
TPS	80.4%	94.3%
grkpack	70.9%	87.9%
loess	77.2%	92.5%
locfitfix	72.1%	88.8%
locfit	59.4%	77.2%

2.11 Applications

2.11.1 Longitudinal study on forest health

In this longitudinal study on the health status of trees, we analyze the influence of calendar time t, age of trees A (in years), canopy density CP (in percent) and location L of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 2001. There are 80 observation points with occurrence of beeches spread over an area extending about 15 km from east to west and 10 km from north to south. The degree of defoliation is used as an indicator for the state of a tree. It is measured in three ordered categories, with $y_{it} = 1$ for "bad" state of tree i in year t, $y_{it} = 2$ for "medium" and $y_{it} = 3$ for "good". A detailed data description can be found in Göttlein and Pruscha (1996).

We use a three-categorical ordered probit model based on a latent semiparametric model $U_{it} = \eta_{it} + \epsilon_{it}$ with predictor

$$\eta_{it} = f_1(t) + f_2(A_{it}) + f_{1|2}(t, A_{it}) + f_3(CP_{it}) + f_{str}(L_i). \tag{2.32}$$

The calendar time trend $f_1(t)$ and the age effect $f_2(A)$ are modeled by cubic P-splines with a second order random walk penalty. The interaction effect between calendar time and age $f_{1|2}(t, A)$ is modeled by a two dimensional cubic P-splines on a 12 by 12 grid of knots. Since canopy density is measured only in 11 different values (0%, 10%,...,100%) we use a simple second order random walk prior (i.e. a P-spline of degree 0) for $f_3(CP)$. For the spatial effect $f_{str}(L)$ we experimented with both a two dimensional P-spline (model 1) and a Markov random field prior (model 2). Following Fahrmeir and Lang (2001b), the neighborhood ∂_s of trees for the Markov random field includes all trees u with Euclidian distance $d(s, u) \leq 1.2$ km. In terms of the DIC (Spiegelhalter et al. 2002), the model based on the Markov random field is preferable. An unstructured spatial effect f_{unstr} is excluded from the predictor for the following two reasons. First, a look at the map of observation points (see Figure 2.17) reveals some sites with only one neighbor, making the identification of a structured and an unstructured effect difficult if not impossible. The second reason is that for each of the 80 sites only 19 observations on the same tree are available with only minor changes of the response category. In fact, there are only a couple of sites where all three response categories have been observed.

The data have been already analyzed in Fahrmeir and Lang (2001b) (for the years 1983-1997 only). Here, nonlinear functions have been modeled solely by random walk priors. Also, the modeling of the interaction between calendar time and age is less sophisticated.

Since the results for model 1 and 2 differ only for the spatial effect, we present for the remaining covariates only estimates based on model 2. All results are based on the choice $a_j = b_j = 0.001$ for the hyperparameters of the variances. A sensitivity analysis revealed that the results are robust to other choices of a_j and b_j. Figure 2.15 shows the nonlinear main effects of calendar time and age of the tree as well as the effect of canopy density. The interaction effect between calendar time and age is depicted in Figure 2.16. The spatial effect is shown in Figure 2.17. Results based on a two dimensional P-spline can be found

in the left panels, and for a Markov random field in the right panels. Shown are posterior probabilities based on a nominal level of 80% (top panels) and 95% (bottom panels).

As we might have expected younger trees are in healthier state than the older ones. We also see that trees recover after the bad years around 1986, but after 1994 health status declines to a lower level again. The interaction effect between time and age is remarkably strong. In the beginning of the observation period young trees are affected higher than the average from bad environmental conditions. Thereafter, however, they recover better than average. The distinct monotonic increase of the effect of canopy densities $\geq 30\%$ gives evidence that beeches get more shelter from bad environmental influences in stands with high canopy density. The spatial effect based on the two dimensional P-spline and the Markov random field are very similar. The Markov random field is slightly rougher (as could have been expected). Note that the spatial effect is quite strong and therefore not negligible.

2.11.2 Space-time analysis of health insurance data

In this section we analyze space-time data from a German private health insurance company. In a consulting case the main interest was on analyzing the dependence of treatment costs on covariates with a special emphasis on modeling the spatio-temporal development. The data set contains individual observations for a sample of 13.000 males (with about 160.000 observations) and 1.200 females (with about 130.000 observations) in West Germany for the years 1991-1997. The variable of primary interest is the treatment cost C in hospitals. Except some categorical covariates characterizing the insured person we analyzed the influence of the continuous covariates age (A) and calendar time (t) as well as the influence of the district (D) where the policy holder lives. We carried out separate analysis for men and women. We also distinguish between 3 types of health services, "accommodation", "treatment with operation" and "treatment without operation". In this demonstrating example, we present only results for males and "treatment with operation". Since the treatment costs are nonnegative and considerably skewed we assume that the costs for individual i at time t given covariates x_{it} are gamma distributed, i.e. $C_{it} \,|\, x_{it} \sim Ga(\mu_{it}, \phi)$ where ϕ is a scale parameter and the mean μ_{it} is defined as

$$\mu_{it} = \exp(\eta_{it}) = \exp(\gamma_0 + f_1(t) + f_2(A_{it}) + f_3(D_{it})).$$

For the effects of age and calendar time we assumed cubic P-splines with 20 knots and a second order random walk penalty. To distinguish between spatially smooth and small scale regional effects, we further split up the spatial effect f_3 into a spatially structured and a unstructured effect, i.e.

$$f_3(D_{it}) = f_{str}(D_{it}) + f_{unstr}(D_{it})$$

For the unstructured effect f_{unstr} we assume i.i.d. Gaussian random effects. For the spatially structured effect we tested both a Markov random field prior and a two dimensional P-spline on a 20 by 20 knots grid. Both sampling schemes 1 and 2 may be used for posterior inference in this situation.

The estimation of the scale parameter ϕ deserves special attention because MCMC inference is not trivial. In analogy to the variance parameter in Gaussian response models, we assume an inverse gamma prior with hyperparameters a_ϕ and b_ϕ for ϕ, i.e. $\phi \sim IG(a_\phi, b_\phi)$. Using this prior the full conditional for ϕ is given by

$$p(\phi \,|\, \cdot) \propto \left(\frac{1}{\Gamma(\phi)\phi^\phi}\right)^n \phi^{a_\phi - 1} \exp(-\phi b'_\phi)$$

with

$$b'_\phi = b_\phi + \sum_{i,t} (\log(\mu_{it}) - \log(C_{it}) + C_{it}/\mu_{it}).$$

This distribution is not of standard form. Hence, the scale parameter must be updated by Metropolis-Hastings steps. We update ϕ by drawing a random number ϕ^p from an inverse gamma proposal distribution with a variance s^2 and a mean equal to the current state of the chain ϕ^c. The variance s^2 is a tuning parameter and must be chosen appropriately to guarantee good mixing properties. We choose s^2 such that the acceptance rates are roughly between 30 and 60 percent.

It turns out that the results are unsensitive to the choice of hyperparameters a_j and b_j. The presentation of results is therefore restricted to the standard choice $a_j = b_j = 0.001$ for the hyperparameters. of the variances.

Figure 2.18 shows the time trend f_1 (panel a) and the age effect f_2 (panel b). Shown are the posterior means together with 80% and 95% pointwise credible intervals. The effect for the year 1999 is future prediction explaining the growing uncertainty for the time effect in this year. Note also the large credible intervals of the age effect for individuals of age 90 and above. The reason are small sample sizes for these age groups. To gain more insight into the size of the effects, panels c) and d) display the marginal effects $f_j^{marginal}$ which are defined as $f_j^{marginal}(x_j) = \exp(\gamma_0 + f_j(x_j))$, i.e. the mean of treatment costs with the values of the remaining covariates fixed such that their effect is zero. The marginal effects (including credible intervals) can be easily estimated in a MCMC sampling scheme by computing (and storing) $f_j^{marginal}(x_j)$ in every iteration of the sampler from the current value of $f_j(x_j)$ and the intercept γ_0. Posterior inference is then based on the samples of $f_j^{marginal}(x_j)$. For the ease of interpretation, a horizontal line is included in the graphs indicating the marginal effect for $f_j = 0$, i.e. $\exp(\gamma_0) \approx 940DM$. Finally, panels e) and f) show the first derivatives of both effects (again including credible intervals). They may be computed by the usual formulas for derivatives of polynomial splines, see De Boor (1978).

Figure 2.19 displays the structured spatial effect f_{str} based on a Markov random field prior. The posterior mean of f_{str} can be found in panel a), the marginal effect is depicted in panel b). Panels c) and d) show posterior probabilities based on nominal levels of 80% and 95%. Note the large size of the spatial effect with a marginal effect ranging from 730-1200 German marks. It is clear that it is of great interest for health insurance companies to detect regions with large deviations of treatment costs compared to the average. The unstructured spatial effect f_{unstr} is negligible compared to the structured effect and therefore omitted.

Figure 2.20 shows the respective estimates of f_{str} now based on two dimensional P-splines. The time trend and age effect for this model are almost identical to the effects displayed in Figure 2.18 and are therefore not displayed. The estimated effects are similar but smoother (as could have been expected) and therefore easier to interpret. However, in terms of the DIC the model based on the MRF prior is preferable.

2.12 Conclusions

This paper proposes semiparametric Bayesian inference for regression models with responses from an exponential family and with structured additive predictors. The paper can be seen as the final in a series of articles on Bayesian semiparametric regression based on smoothness priors, see Fahrmeir and Lang (2001a), Fahrmeir and Lang (2001b) and Part I of this chapter. It particularly extends the methodology for Gaussian responses in Part I of this chapter to situations with fundamentally non-Gaussian responses. Our approach allows estimation of nonlinear effects of continuous covariates and time scales as well as the appropriate consideration of unobserved unit- or cluster specific as well as spatial heterogeneity. Many well known regression models from the literature appear to be special cases of our approach, e.g. dynamic models, generalized additive mixed models, varying coefficient models, geoadditive models or the famous and widely used BYM-model for disease mapping (Besag et al. 1991). The proposed sampling schemes work well and automatically for the most common response distributions. Software is provided in the public domain package *BayesX*.

Our current research is mainly focused on model choice and variable selection. Presently, model choice is based primarily on pointwise credible intervals for regression parameters and the DIC. A first step for more sophisticated variable selection is to replace pointwise credible intervals by simultaneous probability statements as proposed by Besag et al. (1995) and more recently by Held (2004). For the future, we plan to develop Bayesian inference techniques that allow estimation and model choice (to some extent) simultaneously.

Acknowledgement:
This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

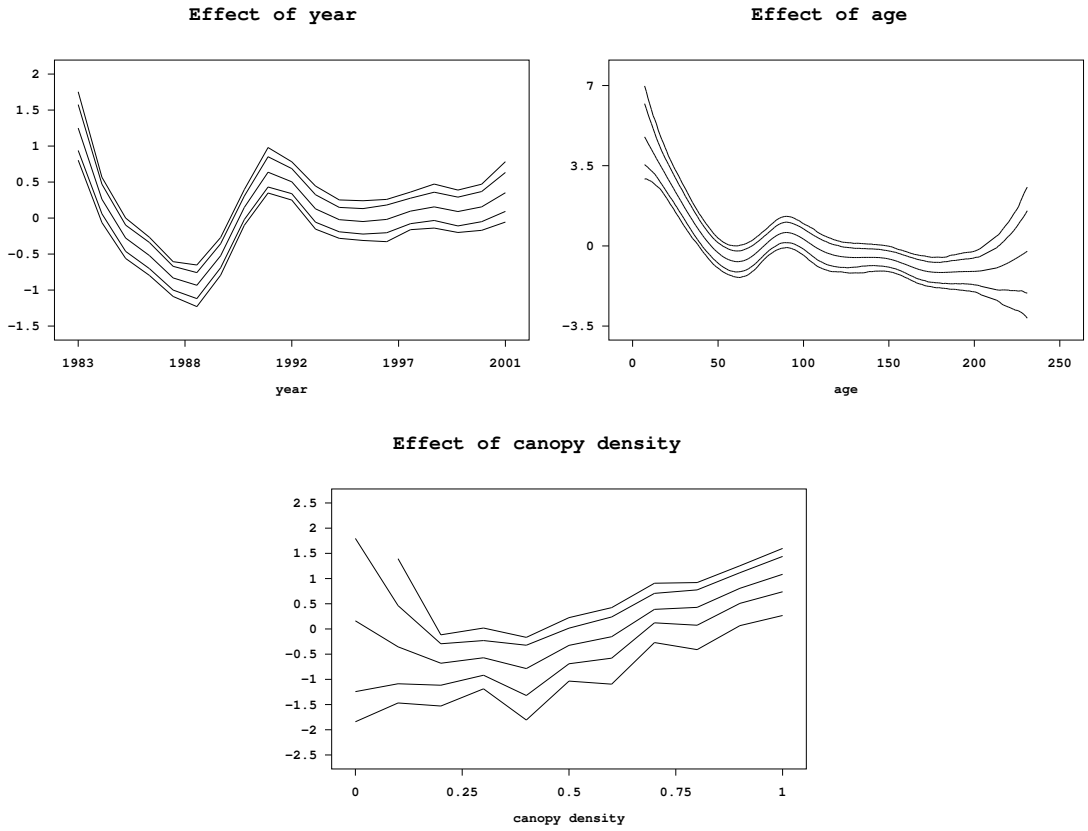


Figure 2.15: Forest health data. Nonlinear main effects of calendar time, age of the tree and canopy density. Shown are the posterior means together with 95% and 80% pointwise credible intervals.

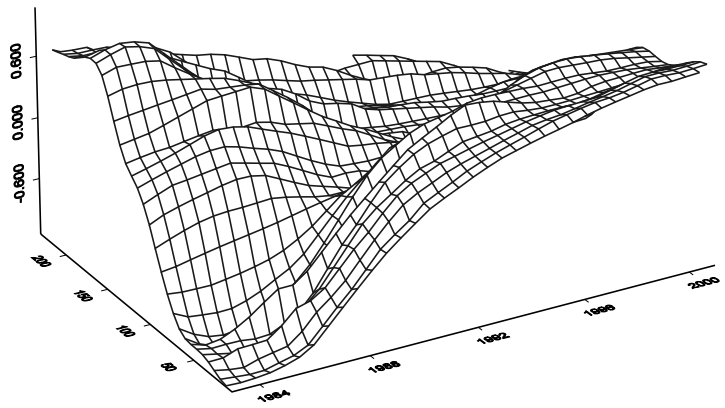


Figure 2.16: Forest health data. Nonlinear interaction between calendar time and age of the tree. Shown are the posterior means.

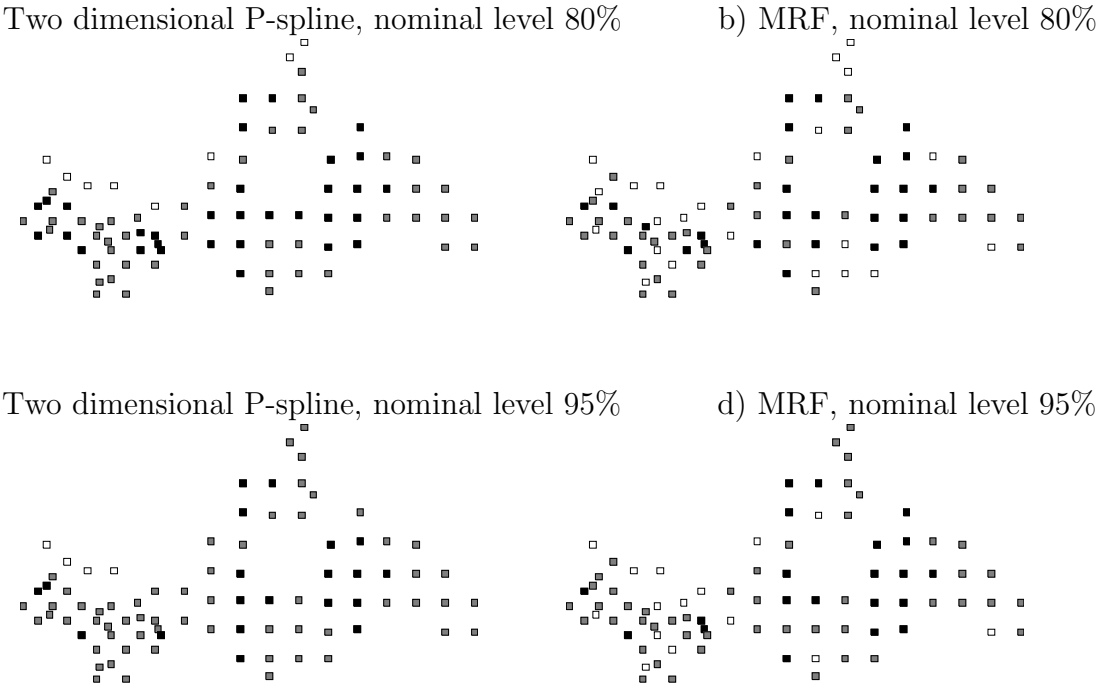


Figure 2.17: Forest health data. Panels a) and c) show the spatial effect based on two dimensional P-splines. Panels b) and d) display the spatial effect based on Markov random fields. Shown are posterior probabilities for a nominal level of 80% (top panels) and 95% (bottom panels). Black denotes locations with strictly negative credible intervals, white denotes locations with strictly positive credible intervals.

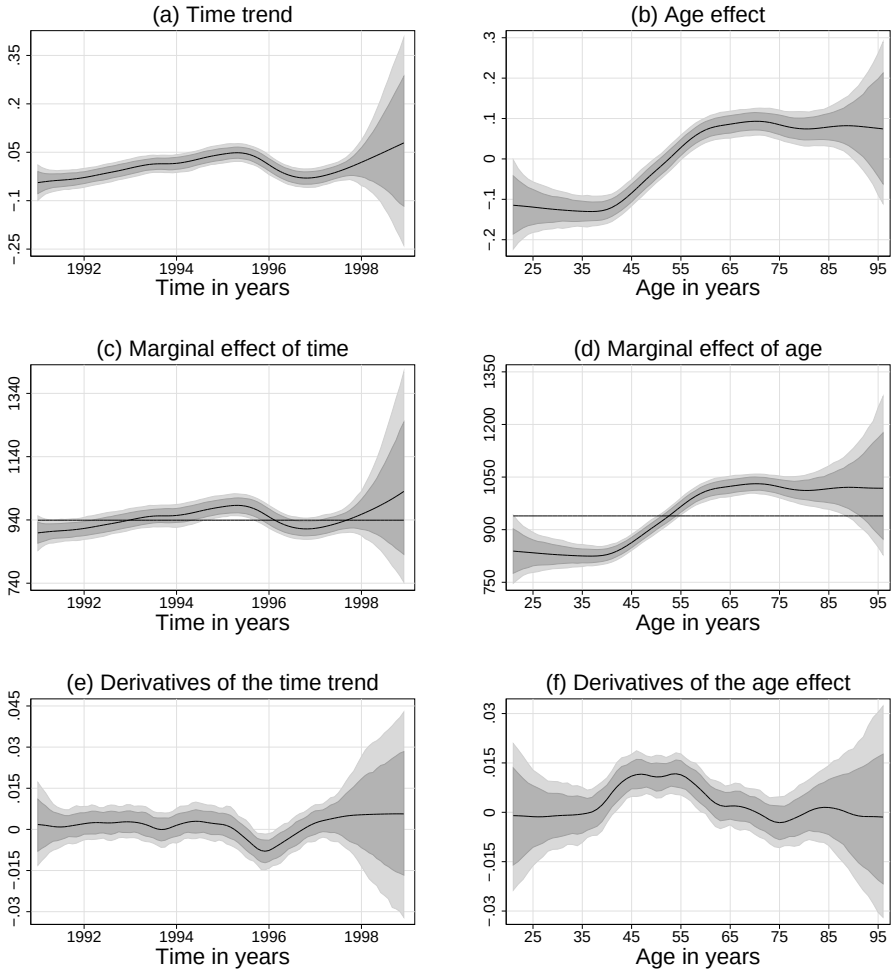


Figure 2.18: Health insurance data: Time trend and age effect. Panels a) and b) show the estimated posterior means of functions f_1 and f_2 together with pointwise 80% and 95% pointwise credible intervals. Panels c) and d) depict the respective marginal effects and panels e) and f) the first derivatives f_1' and f_2'.

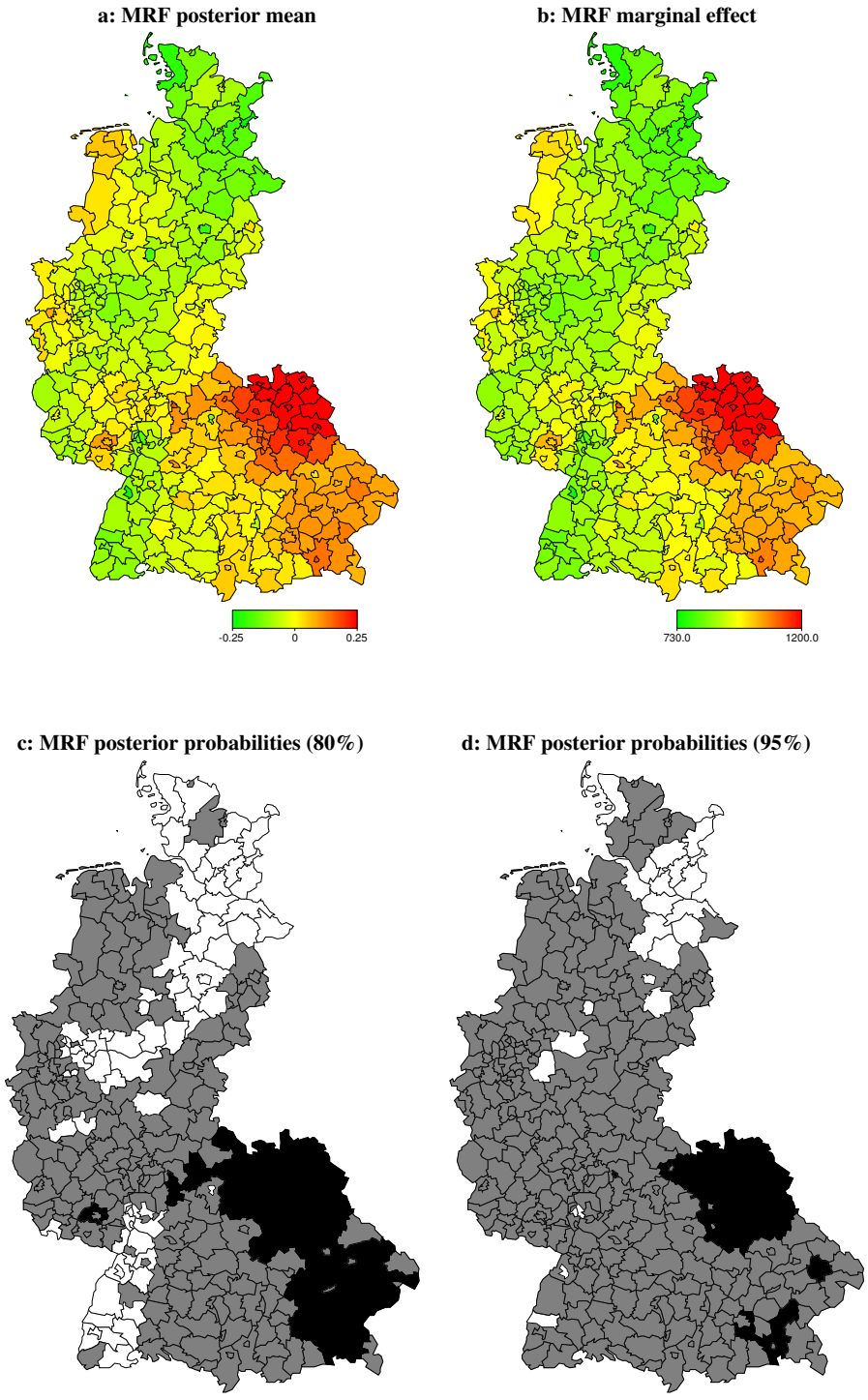


Figure 2.19: Health insurance data: Structured spatial effect f_{str} based on Markov random field priors. The posterior mean of f_{str} is shown in panel a) and the marginal effect in panel b). Panels c) and d) display posterior probabilities for nominal levels of 80% and 95%. Black denotes regions with strictly positive credible intervals and white regions with strictly negative credible intervals.

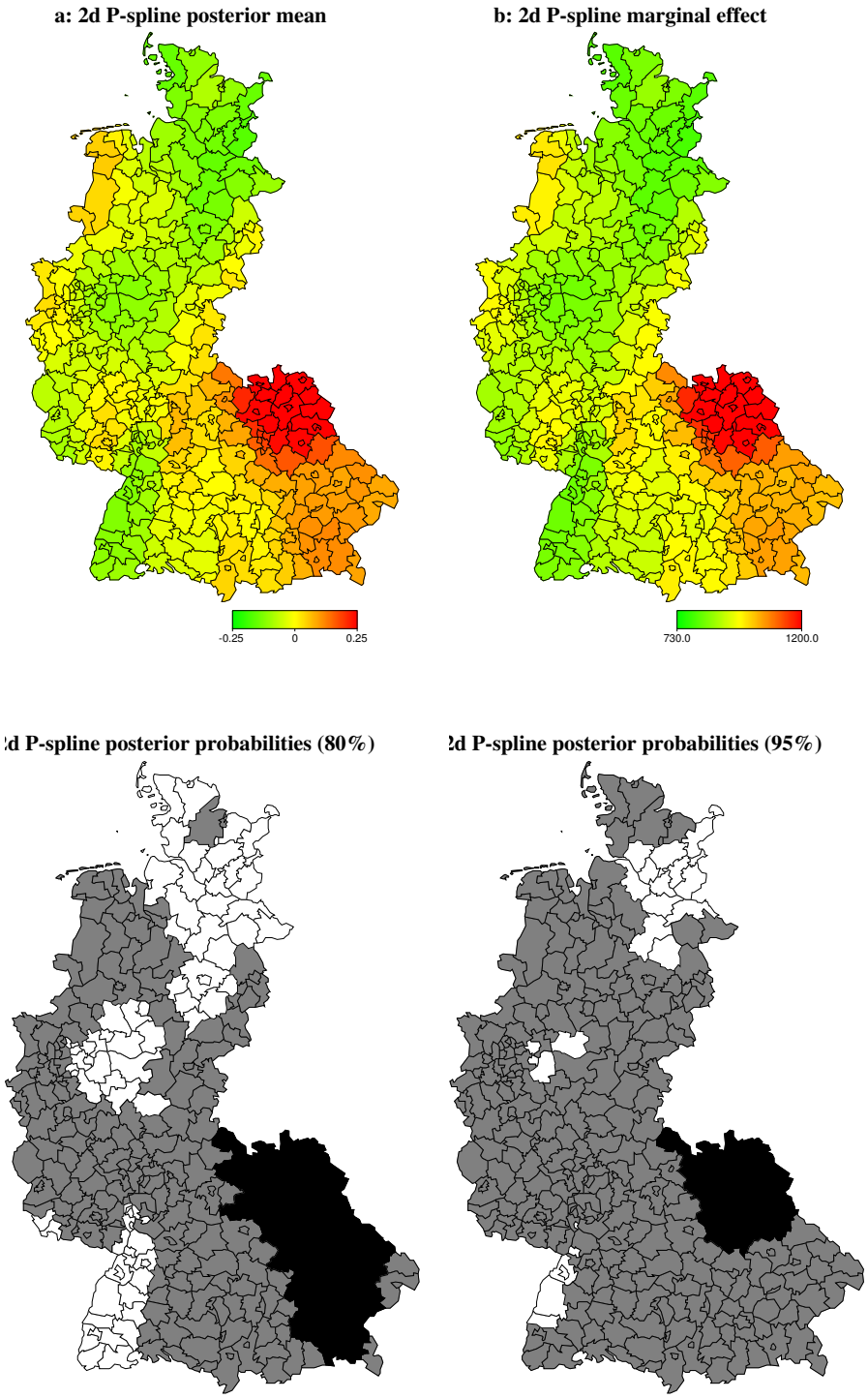


Figure 2.20: Health insurance data: Structured spatial effect f_{str} based on two dimensional P-splines. The posterior mean of f_{str} is shown in panel a) and the marginal effect in panel b). Panels c) and d) display posterior probabilities for nominal levels of 80% and 95%. Black denotes regions with strictly positive credible intervals and white regions with strictly negative credible intervals.

Chapter 3

Monotonic regression based on Bayesian P-Splines

So far we have introduced Bayesian P-splines as a very flexible method for modeling nonparametric effects in one or two dimensions within a structured additive regression framework. Flexibility was even enhanced by allowing for locally adaptive smoothing parameters in the one dimensional case as well as for two dimensional surface estimation. In this chapter the goal is not to further increase, but to restrict flexibility by imposing constraints on the shape of the functional form of nonparametric estimates. Specifically, we consider monotonicity constraints on one dimensional P-splines. This restriction is reasonable if one knows a priori that the relationship between a continuous covariate and the outcome is either increasing or decreasing. In this case a restriction is useful in order to avoid unreasonable results coming from noisy observations. Prior knowledged of this kind is given for example in many applications in statistical medicine, where a dose-response relationship is known to be monotonic. In our example, the relation between prices and sales of consumer goods can be assumed to behave monotonic from an ecological point of view.

This chapter develops Bayesian methodology in order to impose monotonicity constraints on Bayesian P-splines and demonstrates their usefulness by an application to estimating price response functions from store-level scanner data. Gaussian and non Gaussian responses are considered. The content of this chapter is also available as SFB 386 discussion paper 331 under the title 'Monotonic regression based on Bayesian P-Splines: an application to estimating price response functions from store-level scanner data' by Brezger and Steiner (2003). Note, that this chapter differs slightly from the original paper due to unification of notation and correction of typos.

Monotonic regression based on Bayesian P-Splines: an application to estimating price response functions from store-level scanner data

Andreas Brezger
Department of Statistics
University of Munich
Ludwigstr. 33
80539 Munich
Germany

Winfried J. Steiner
Department of Marketing
University of Regensburg
Universitätsstr. 31
93053 Regensburg
Germany

ABSTRACT

Generalized additive models have become a widely used instrument for flexible regression analysis. In many practical situations, however, it is desirable to restrict the flexibility of nonparametric estimation in order to accommodate a presumed monotonic relationship between a covariate and the response variable. For example, consumers usually will buy less of a brand if its price increases, and therefore one expects a brand's unit sales to be a decreasing function in own price. We follow a Bayesian approach using penalized B-splines and incorporate the assumption of monotonicity in a natural way by an appropriate specification of the respective prior distributions. We illustrate the methodology in an empirical application modeling demand for a brand of orange juice and show that imposing monotonicity constraints for own- and cross-item price effects improves the predictive validity of the estimated sales response function considerably.

Keywords: Generalized Additive Model, Markov Chain Monte Carlo, Sales Promotion, Own- and Cross-Item Price Effects, Asymmetric Quality Tier Competition

3.1 Introduction

Generalized additive models (GAM) are a powerful tool for modeling possibly nonlinear effects of multiple covariates. For continuous covariates, the variety of different approaches for nonlinear modeling comprises, for example, smoothing splines (e.g. Hastie and Tibshirani 1990), regression splines (e.g. Friedman and Silverman 1989, Friedman 1991, Stone et al. 1997), local methods (e.g. Fan and Gijbels 1996) as well as P-splines (Eilers and Marx 1996, Marx and Eilers 1998). Bayesian nonparametric approaches make use of adaptive knot selection (e.g. Smith and Kohn 1996, Denison et al. 1998, Biller 2000, Di Matteo et al. 2001, Biller and Fahrmeir 2001, Hansen and Kooperberg 2002) or smoothness priors (Hastie and Tibshirani 2000, Fahrmeir and Lang 2001a, Fahrmeir and Lang 2001b). In Part I of Chapter 2 the frequentist P-splines of Eilers and Marx (1996) is adopted for a Bayesian framework for additive models and in Part II of Chapter 2 this work is extended to GAMs.

While strictly parametric modeling is too restrictive in many cases, the flexibility of non- and semiparametric approaches may lead to implausible results on the other hand. Clearly, the problem of overfitting can be addressed by penalization of too rough functions or by adaptive knot selection. Much less discussed in the literature on nonparametric estimation is, however, the important case when theory and/or empirical evidence strongly suggest a monotonic relationship between a covariate and a response variable. For example, consumers usually will buy less of a brand as its price increases, and therefore one expects a brand's unit sales or market share to decrease monotonically in price. The downward slope of own price response functions is in accordance with economic theory (e.g. Rao 1993), and there is strong empirical support that own-price elasticities are negative and elastic (e.g. Tellis 1988, Hanssens, Parsons and Schultz 2001). Similarly, we generally expect cross-price effects on competitive items (i.e., brand substitutes) to be positive or at least nonnegative, implying that a price cut by a brand may decrease but by no means will increase the unit sales of competitive brands (Sethuraman, Srinivasan and Kim 1999). Examples for presumed monotonic relationships can also be found in disciplines other than business and economics, as it is the case for many dose-response relationships in medicine. For instance, the concentration of dust and the duration of exposition to it at working places is assumed to affect the occurrence of certain lung diseases in a monotonic way (Ulm and Salanti 2003). Monotonic effects are also referred to as isotonic if the respective function is nondecreasing, and antitonic if a function is nonincreasing.

The topic of monotonic regression has already been addressed in Ulm and Salanti (2003) and Salanti and Ulm (2003) in a frequentist setting. Dunson and Neelon (2003) and Holmes and Heard (2003) have presented Bayesian approaches to monotonic regression. The former, however, have considered only GLMs and modeling has been based on piecewise constant functions, while the latter have dealt with only a small number of level sets obtained from a categorization of continuous covariates.

In this paper, we propose to use Bayesian P-splines of an arbitrary degree and enforce monotonicity in a straightforward way by an additional restriction of the prior distribution via indicator functions. This restriction may be imposed either for one or an arbitrary num-

ber of the additive terms in the model, whereas other terms may be modeled unrestricted. MCMC inference involves sampling from multivariate truncated normal distributions. This is accomplished by an "internal" Gibbs sampler in each iteration, i.e., we employ a short Gibbs sampler in order to draw from the proposal density. In the non-Gaussian case, this procedure is used to draw from an iteratively weighted least squares (IWLS) proposal density in a Metropolis-Hastings step. Our methodology is implemented in the public domain software package *BayesX* (Chapter 5) and it is possible to combine monotonic regression with all types of response distributions supported by *BayesX*. These are the most common one dimensional distributions like Gaussian, Binomial, Poisson, Gamma and Negative Binomial, and multinomial logit and cumulative probit models for multivariate responses. *BayesX* also supports the use of random effects to account for unobserved heterogeneity, Gaussian Markov random field (GMRF) priors for spatial covariates, varying coefficient terms and surface smoothing for interactions of covariates.

The remainder of the paper is organized as follows: Section 2 briefly reviews GAMs and (Bayesian) P-splines, whereas section 3 provides details on the MCMC techniques employed. In section 4, we apply the proposed methodology to weekly store-level scanner data to relate unit sales of a particular brand of orange juice in a major supermarket chain to own and competing brands' promotional instruments. Using a log-normal model and a Gamma model, we illustrate for both Gaussian and non-Gaussian responses that imposing monotonicity constraints on the nonparametric terms for own-item and cross-item price effects improves the predictive validity of the estimated sales response functions considerably. We conclude with a summary of the most important contents and key findings in section 5.

3.2 Model Assumptions

3.2.1 Generalized additive models and P-Splines

Suppose we are given n observations (y_i, x_i, v_i), $i = 1, \ldots, n$, where y_i is a response variable, $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of continuous covariates and $v_i = (v_{i1}, \ldots, v_{iq})'$ is a vector of additional covariates. GAMs assume that, given x_i and v_i, the response y_i follows an exponential family distribution (Hastie and Tibshirani 1990, Fahrmeir and Tutz 2001)

$$p(y_i|x_i, v_i) = c(y_i, \theta_i) \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{\phi} \right\}$$

and that the mean $\mu_i = E(y_i|x_i, v_i)$ is linked to a semiparametric additive predictor η_i via a known link function g:

$$g(\mu_i) = \eta_i, \quad \eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + v_i'\gamma. \tag{3.1}$$

$f_1, \ldots, f_p$ are unknown smooth functions of the continuous covariates and $v_i'\gamma$ represents the parametric part of the predictor.

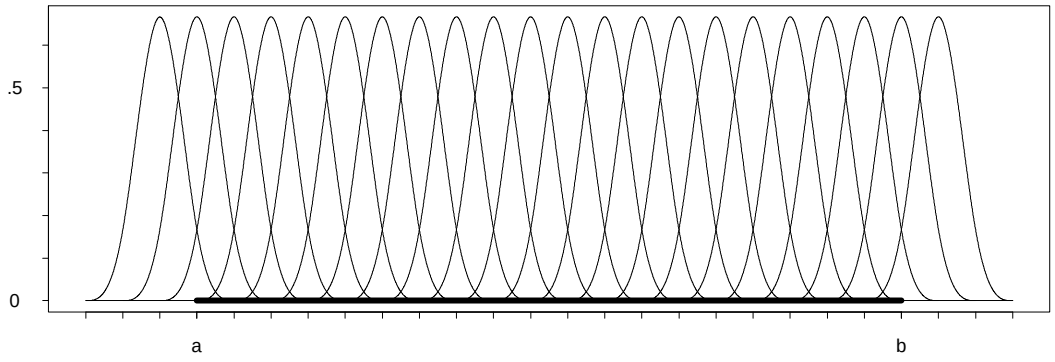


Figure 3.1: B-spline basis functions of degree three covering the interval $[a, b]$.

For modeling the unknown functions f_j, $j = 1, \ldots, p$, we follow Chapter 2 and propose a Bayesian version of the P-splines approach introduced in a frequentist setting by Eilers and Marx (1996). Accordingly, we assume that the unknown functions can be approximated by a polynomial spline of degree l and with $r + 1$ equally spaced knots

$$x_{j,min} = \zeta_{j0} < \zeta_{j1} < \cdots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$$

over the domain of x_j. The spline can be written in terms of a linear combination of $M = r + l$ B-spline basis functions (De Boor 1978). Figure 3.1 gives an illustration of B-spline basis functions of degree three, which are also referred to as cubic splines. Note that except at the boundaries each basis function overlaps with $2 \cdot l$ neighboring B-splines.

Denoting the ρ-th basis function by $B_{j\rho}$, we obtain

$$f_j(x_j) = \sum_{\rho=1}^{M} \beta_{j\rho} B_{j\rho}(x_j).$$

To keep notation simple, we assume an equal number of basis functions M for all functions f_j. By defining the $n \times M$ design matrices X_j where the element in row i and column ρ is given by $X_j(i, \rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor (3.1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V\gamma. \tag{3.2}$$

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jM})'$, $j = 1, \ldots, p$ corresponds to the vector of unknown regression coefficients. The matrix V is the usual design matrix for fixed effects. To overcome the difficulties in determining the position and the number of the knots involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 and 40) to ensure sufficient flexibility, and to introduce a roughness penalty of first or second order differences on adjacent regression coefficients to avoid overfitting. These penalized B-splines have also become known as P-splines. In our Bayesian approach, we replace first or second order differences used in this frequentist approach with their stochastic analogues, i.e., first or second order random walks defined by

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{3.3}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The amount of smoothness is controlled by the variance parameter τ_j^2 which corresponds to the inverse of the smoothing parameter in the frequentist approach. The amount of smoothness can be estimated simultaneously with the regression coefficients by defining an additional hyperprior for the variance parameters τ_j^2.

We assign the conjugate prior for τ_j^2 (and for the scale parameter σ^2 in the Gaussian case) which is an inverse Gamma distribution

$$\tau_j^2 \sim IG(a_j, b_j)$$

with hyperparameters a_j and b_j. A common choice for a_j and b_j leading to almost diffuse priors is $a_j = b_j$, e.g. $a_j = b_j = 0.001$, which is also our default choice. Alternatively, we may set $a_j = 1$ and b_j small, e.g. $b_j = 0.005$ or $b_j = 0.0005$. We estimated all models discussed in this paper with alternative settings for the hyperparameters. The results proofed to be almost insensitive regarding the specific choice of hyperparameters. All results presented in the remainder of the paper are obtained by the default choice.

Defining a penalty matrix K^k corresponding to a random walk of order k enables us to formulate the prior for a P-spline term as a joint prior distribution for β_j:

$$p(\beta_j) \propto \exp\left\{-0.5\frac{1}{\tau_j^2}\beta_j' K_j^k \beta_j\right\}, \tag{3.4}$$

see Part I of Chapter 2 for details. For example for $k = 1$ we have

$$K^1 = \begin{pmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & -1 & 2 & -1 & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 1 \end{pmatrix}$$

with zero elements outside the first off-diagonals.

3.2.2 Monotonicity constraints

To obtain monotonicity, i.e., $f_j'(x) \geq 0$ or $f_j'(x) \leq 0$, it is sufficient to guarantee that subsequent parameters are ordered, such that

$$\beta_{j1} \leq \cdots \leq \beta_{jM} \quad \text{or} \quad \beta_{j1} \geq \cdots \geq \beta_{jM},$$

respectively. A proof can be found in Appendix A.2. In our approach, these constraints are imposed by introducing indicator functions to truncate the prior appropriately to obtain

the desired support. This leads to

$$p(\beta_j) = c_1(\beta_j) \exp\left\{ -0.5 \frac{1}{\tau_j^2} \beta_j' K_j^k \beta_j \right\} \prod_{\rho=2}^{M} \mathbf{1}(\beta_{j\rho} \geq \beta_{j\rho-1}) \tag{3.5}$$

for nondecreasing functions (isotonic case) and

$$p(\beta_j) = c_1(\beta_j) \exp\left\{ -0.5 \frac{1}{\tau_j^2} \beta_j' K_j^k \beta_j \right\} \prod_{\rho=2}^{M} \mathbf{1}(\beta_{j\rho} \leq \beta_{j\rho-1})$$

for nonincreasing functions (antitonic case), respectively, where $c_1(\beta_j)$ is a normalizing function depending on β_j.

3.2.3 Extensions

Various extensions regarding the additive predictor (3.1) are possible. In order to account for unobserved heterogeneity between different groups or clusters of units, we may add an unstructured group-specific random effect. Suppose we are given a grouping variable that can take values in $\{1, \ldots, G\}$. Then, we can extend (3.1) to

$$\eta_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + v_i'\gamma + f_{unstr}(g_i)$$

and assume

$$f_{unstr}(g) = b_g \sim N(0, \tau_b^2), \quad g = 1, \ldots, G, \tag{3.6}$$

where $f_{unstr}(g_i) = f_{unstr}(g)$ if observation i belongs to group g. Using the penalty matrix $K^b = I$, we can write (3.6) in the general form

$$p(b_g|\tau_b^2) \propto \exp\{-\frac{1}{2} b_g' K^b b_g\}.$$

If we would presume a spatial correlation between groups, we may additionally introduce a spatial correlated GMRF. Further possible extensions are varying coefficient terms and interactions of covariates (see Chapter 2). In the remainder, we focus on models with random effects.

3.3 MCMC Inference

Let α be the vector of all parameters to be estimated in the model. Bayesian inference is based on the posterior distribution

$$\begin{aligned} p(\alpha|y) \quad \propto \quad & L(y, \beta_1, \ldots, \beta_p, \gamma, b_g, \phi) \prod_{j=1}^{p} \big(p(\beta_j|\tau_j^2) p(\tau_j^2)\big) \\ & p(b_g|\tau_b^2) p(\tau_b^2) p(\gamma) p(\phi) \end{aligned} \tag{3.7}$$

where $L(\cdot)$ consists of the product of all individual likelihood contributions. ϕ and $p(\phi)$ have to be omitted for response distributions without a scale parameter. Because (3.7) is analytically intractable in all but the most simple cases, we employ Markov Chain Monte Carlo (MCMC) techniques to obtain estimates for the parameters of interest. More specifically, we implement a block move, i.e. we subsequently draw from the full conditionals $p(\beta_j|\cdot)$, $j = 1, \ldots, p$, $p(\gamma|\cdot)$ and $p(b_g|\cdot)$ of the blocks of parameters β_j, $j = 1, \ldots, p$, γ and b_g. For Gaussian responses, these blocks can be updated by block move Gibbs sampling steps. In binary probit and cumulative probit models, we can rely on the same sampling scheme as building block, see Chen and Dey (2000) or Part II of Chapter 2 for details. In all other cases, we use Metropolis-Hastings steps with iteratively weighted least squares (IWLS) proposals. The variance parameters $\tau_1^2, \ldots, \tau_p^2$, τ_b^2 (and the scale parameter σ^2 in the Gaussian case) are updated by single move Gibbs sampling steps.

For posterior inference, we discard the draws from an initial burn-in period and take only every rth draw thereafter in order to minimize the autocorrelation of the samples. The formulas and algorithms in the following subsections are formulated with respect to isotonic constraints. The adjustments for antitonic constraints are straightforward.

3.3.1 Gaussian Response

For Gaussian response, the full conditional distribution for β_j is given by

$$p(\beta_j|\cdot) \propto c_1(\beta_j) \exp\{-0.5(\beta_j - m_j)' P_j (\beta_j - m_j)\} \prod_{\rho=2}^{M} \mathbf{1}(\beta_{j\rho} \geq \beta_{j,\rho-1}), \tag{3.8}$$

and

$$\begin{aligned} P_j &= \frac{1}{\sigma^2} X_j' X_j + \frac{1}{\tau_j^2} K_j^k \\ m_j &= \frac{1}{\sigma^2} P_j^{-1} X_j' (y - \eta + X_j \beta_j^c) \end{aligned}$$

where β_j^c is the current state of β_j.

In order to sample from this M-dimensional truncated Gaussian distribution (3.8), we adopt the method of Robert (1995) and run an extra (short) single move Gibbs sampler in each MCMC iteration. The algorithm is as follows:

(i) Set $\beta^{(0)} = \beta_j^c$.

(ii) For $t = 1, \ldots, T$, successively draw from the one-dimensional truncated Gaussian distributions

1. $\beta_1^{(t)} \sim N(\mu_1, \sigma_1^2, -\infty, \beta_2^{(t-1)})$
2. $\beta_2^{(t)} \sim N(\mu_2, \sigma_2^2, \beta_1^{(t)}, \beta_3^{(t-1)})$

3. $\beta_3^{(t)} \sim N(\mu_3, \sigma_3^2, \beta_2^{(t)}, \beta_4^{(t-1)})$

$\vdots$

M. $\beta_M^{(t)} \sim N(\mu_M, \sigma_M^2, \beta_{M-1}^{(t)}, \infty)$

where $N(\mu, \sigma^2, \mu_l, \mu_r)$ denotes a Gaussian distribution with mean μ, variance σ^2 and with left truncation point μ_l and right truncation point μ_r, respectively. The truncation points in the algorithm above are the current states of the adjacent parameters. Therefore, we have only right truncation for $\beta_1^{(t)}$ and left truncation for $\beta_M^{(t)}$. The parameters μ_ρ and σ_ρ^2, $\rho = 1, \ldots, M$, are the conditional means and variances of the (nontruncated) full conditional (3.8):

$$\begin{aligned} \mu_\rho &= \frac{1}{p_{\rho\rho}} \left\{ \sum_{\psi<\rho} (\beta_\psi^{(t)} - m_\psi) \cdot p_{\rho\psi} + \sum_{\psi>\rho} (\beta_\psi^{(t-1)} - m_\psi) \cdot p_{\rho\psi} \right\} \\ \sigma_\rho^2 &= \frac{1}{p_{\rho\rho}} \end{aligned}$$

where m_ψ is the ψ-th element of m_j and $p_{\rho\psi}$ is the element in row ρ and column ψ of the precision matrix P_j in (3.8). Note that the subscript j is suppressed in the formulae above.

(iii) Take $\beta^{(T)} = (\beta_1^{(T)}, \ldots, \beta_M^{(T)})'$ as a random sample from (3.8).

Usually, convergence is reached after 10-20 cycles. To reach convergence with considerable certainty, we set $T = 100$. Computation is very fast, as the mean m_j and the precision matrix P_j have to be computed only once. Moreover, m_j is obtained by sparse matrix operations exploiting the band structure of P_j. This involves a Cholesky decomposition and avoids expensive matrix inversions (compare Rue 2001).

Regarding the fixed effects we obtain a normal distribution with precision matrix and mean

$$P_\gamma = \frac{1}{\sigma^2} V'V, \quad m_\gamma = (V'V)^{-1} V'(y - \eta + V\gamma)$$

as full conditional.

The full conditionals for the variance parameters τ_j^2, $j = 1, \ldots, p$, τ_b^2 and the scale parameter σ^2 are all inverse Gamma distributions with parameters

$$a_j' = a_j + \frac{rank(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2} \beta_j' K_j \beta_j$$

for τ_j^2, $j = 1, \ldots, p$, τ_b^2 and

$$a_{\sigma^2}' = a_{\sigma^2} + \frac{n}{2} \quad \text{and} \quad b_{\sigma^2}' = b_{\sigma^2} + \frac{1}{2} \epsilon'\epsilon$$

for σ^2, where ϵ is the usual vector of residuals.

3.3.2 Non-Gaussian Response

For non-Gaussian response, the full conditional $p(\beta_j|\cdot)$ is

$$p(\beta_j|\cdot) \propto \prod_{i=1}^{n} c(y_i,\theta_i)\exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi}\right\} c_1(\beta_j)\exp\left\{-0.5\frac{1}{\tau_j^2}\beta_j' K_j^k \beta_j\right\} \prod_{\rho=2}^{M} \mathbf{1}(\beta_{j\rho} \geq \beta_{j,\rho-1}),$$

which has no longer standard form. Thus, we use a Metropolis-Hastings step with an IWLS proposal to update β_j. An IWLS proposal is obtained by a quadratic approximation of the likelihood via Taylor expansion around the current state β_j^c of β_j, compare Part II of Chapter 2 or, in the mixed model context, Gamerman (1997). This leads to a truncated multivariate Gaussian proposal

$$q(\beta_j) \propto c_1(\beta_j)\exp\{-0.5(\beta_j - m(\beta_j^c))' P(\beta_j^c)(\beta_j - m(\beta_j^c))\} \prod_{\rho=2}^{M} \mathbf{1}(\beta_{j\rho} \geq \beta_{j\rho-1}) \tag{3.9}$$

where

$$\begin{aligned} P(\beta_j^c) &= X_j' W(\beta_j^c) X_j + \frac{1}{\tau_j^2} K_j^k \\ m(\beta_j^c) &= P(\beta_j^c)^{-1} X_j' W(\beta_j^c) \tilde{y}(\beta_j^c) \\ W(\beta_j^c) &= diag(w_1,\ldots,w_n) \\ \tilde{y}(\beta_j^c) &= (y-\mu)g'(\mu) + X_j\beta_j^c \end{aligned}$$

with $w_i^{-1} = b''(\theta_i)\{g'(\mu_i)\}^2$. Alternatively, we could also use the current mode of $p(\beta_j|\cdot)$ rather than β_j^c to perform the Taylor expansion, which would simplify the calculation of the acceptance probability, compare Part II of Chapter 2. Generating a proposed value β_j^p from (3.9) is again accomplished by an extra Gibbs sampler as described in Subsection 3.3.1. It has been our experience that convergence in the non-Gaussian case is slower than in the Gaussian case. We therefore set the number of iterations for the single move Gibbs sampler to $T = 250$ (as opposed to $T = 100$ for Gaussian response) to ensure convergence, which implies that we take the 250th sample as a random sample from (3.9). The main difference to the Gaussian sampling scheme is that this sample can only be accepted with probability

$$\alpha(\beta_j^c, \beta_j^p) = \min\left\{1, \frac{L(\beta_j^p)p(\beta_j^p)q(\beta_j^p,\beta_j^c)}{L(\beta_j^c)p(\beta_j^c)q(\beta_j^c,\beta_j^p)}\right\}$$

as the new state of β_j. Note that the normalizing functions $c_1(\cdot)$ cancel out and we have the same acceptance probability as in the unrestricted case.

The full conditionals for the variance parameters τ_j^2, $j = 1,\ldots,p$ and τ_b^2 are again inverse Gamma distributions and therefore updated via Gibbs sampling. Fixed effects are updated by Metropolis-Hastings steps.

3.4 Empirical Application: Estimating price response from store-level scanner data

3.4.1 Background

It is important for both manufacturers and retailers to know how sales respond to price promotions. For example, if a brand's sales response to own price cuts shows increasing returns to scale, a firm will run deeper price discounts for the brand than in case of decreasing returns to scale. In the following, we apply the monotonic regression approach to estimating promotional price response functions from store-level scanner data.

It is well documented that sales promotions, especially in the form of temporary price reductions, substantially increase sales of promoted brands (e.g. Wilkinson, Mason and Paksoy 1982, Blattberg and Neslin 1990, Bemmaor and Mouchoux 1991, Blattberg, Briesch and Fox 1995). There is also empirical evidence that a temporary price cut by a brand may decrease sales of competitive items significantly (e.g. Mulherne and Leone 1991, Blattberg and Wisniewski 1989, Bemmaor and Mouchoux 1991). Cross-item price effects, however, are usually much lower than own-item price effects, see Hanssens et al. (2001) for an overview of empirical findings. In addition, there is strong empirical support that cross-promotional effects are asymmetric, implying that promoting higher-priced/higher quality brands generates more switching from lower-priced/lower quality brands than does the reverse (e.g. Blattberg and Wisniewski 1989, Allenby and Rossi 1991, Blattberg et al. 1995). This phenomenon has also become known as asymmetric quality tier competition (e.g. Sivakumar and Raj 1997). Moreover, a recent meta-analysis of cross-price elasticity estimates revealed strong neighborhood price effects, indicating that brands that are closer to each other in price have larger cross-price effects than brands priced farther apart (Sethuraman et al. 1999).

Despite the wealth of empirical findings on own- and cross-price effects, little was known about the shape of the promotional price response function until recently. Most studies addressing this issue employed strictly parametric functions, and came to different results from model comparisons. For example, Wisniewski and Blattberg (1983) found the own-item price effect curve to be modeled best by an s-shaped function, while Blattberg and Wisniewski (1987) found the curve to show increasing returns with deeper price discounts. The former, however, estimated own price response functions at the category level rather than the individual brand level, while the latter analyzed a limited range of price discounts. Today, multiplicative (log-log), exponential (semi-log) and log-reciprocal functional forms are the most widely used parametric specifications to represent nonlinearities in sales response to promotional instruments (e.g. Blattberg and Wisniewski 1989, Blattberg and George 1991, Montgomery 1997, Kopalle, Mela and Marsh 1999, Foekens, Leeflang and Wittink 1999, van Heerde, Leeflang and Wittink 2002). These functional forms are inherently monotonic (decreasing for own-price and increasing for cross-price effects) and all use a logarithmic transformation of brand sales to normalize the distribution of the dependent variable which typically is markedly skewed with promotional data (e.g Mulherne and

Leone 1991). However, there does not seem to exist a "best" parametric functional form generalizable across product categories or even across brands within a category. Therefore, nonparametric regression methods seem to be highly promising to explore the shape of the promotional price response curve more flexibly.

van Heerde, Leeflang and Wittink (2001) proposed a kernel-based semiparametric approach in which a brand's unit sales is modeled as a nonparametric function of own- and cross-item price variables and a parametric function of other predictors. The model can also accommodate flexible interaction effects between price cuts of different brands but may suffer from the curse of dimensionality as the number of competing items increases. van Heerde et al. (2001) obtained superior performance for the semiparametric model in both fit and predictive validity relative to two benchmark parametric models. Their results based on store-level scanner data for three product categories (tuna, beverage and a third packaged food product) indicate threshold and/or saturation effects for both own- and cross-item price cuts. Threshold effects are present if consumers do not change their purchase intentions unless a promotional price cut exceeds a certain threshold level, say, e.g., 15% (Gupta and Cooper 1992). A common argument for the existence of saturation effects is based on the belief that consumers can stockpile and/or consume only limited amounts of a promoted good, e.g., due to inventory constraints or perishability (Blattberg et al. 1995, van Heerde et al. 2001). About two-third of the nonparametric own-item and cross-item price response curves estimated by van Heerde et al. (2001) showed a (reverse) s-shape reflecting both threshold and saturation effects, with a wide range of different saturation points across brands. Some curves revealed a (reverse) L-shape with a strong kink at a certain level of price cut, while other curves do not show a threshold nor a saturation effect. These different results across individual brands strongly support the use of nonparametric estimators to let the data determine the shape of price response functions. However, two own-item price response curves indicated a decrease in unit sales as price discounts become very deep. Clearly, this nonmonotonicity is difficult to interpret from an economic point of view. One explanation may be that consumers associate a loss in quality with very deep price cuts, but this argument seems at least questionable with frequently purchased consumer nondurables.

In contrast to van Heerde et al. (2001), Kalyanam and Shively (1998) proposed a stochastic spline regression approach (Wahba 1978) in the context of a hierarchical Bayes model (Wong and Kohn 1996) and found much stronger irregularities in own-price response for some of the brands (tuna, margarine) examined. Especially, although overall downward sloping, the respective curves show local upturns and downturns with spikes at certain price points resulting in less smooth and nonmonotonic shapes. Kalyanam and Shively (1998) illustrated that these nonmonotonicities may be associated with odd pricing or a complex convolution of odd pricing with other effects like, e.g., the existence of segments with distinct reservation prices. Odd pricing refers to the practice of setting prices ending in odd numbers or just below a round number (e.g., 0.99 cents instead of 1.00 dollar). On the other hand, the curve plots also revealed that the estimates at the very strongest local sales peaks were based only on one or a few data points (see, e.g., the results for the Starkist brand in Kalyanam and Shively 1998, p. 26). Kalyanam and Shively (1998) themselves

point out that in case of an insufficient number of data points, the estimated functions may show irregularities where none exist. This problem also applies to another tuna brand (Bumble Bee) where the estimated curve indicated a (monotonic) increase in unit sales with increasing own-price beyond a certain price point (i.e., for higher price levels). This latter irregularity is not in accordance with economic theory and, as a consequence, would suggest an optimal price at infinity.

Besides the problem of inaccurate estimation due to sparse data in some cases, the findings of Kalyanam and Shively (1998) agree with those of van Heerde et al. (2001) with respect to the existence of threshold effects for several brands, i.e., flat own-price response around prices at the upper bound of the range of observed prices. In comparison to a parametric semilog specification, Kalyanam and Shively (1998) obtained a superior fit of their spline model in terms of adjusted R^2 values for each of the brands analyzed. Unfortunately, no model validation results were reported.

The monotonic nonparametric regression approach as proposed in this paper is our answer to resolve the problem whether nonmonotonic effects indeed exist when theory and/or empirical experience would rather suggest not. Our perspective is that an unconstrained estimation allowing for nonmonotonicities should be preferred only if it outperforms a constrained estimation in validation samples. Otherwise, nonmonotonic effects are likely to represent an artefact caused by sparse data or merely by too much flexibility of the nonparametric estimator. Importantly, imposing monotonicity constraints does not preclude the estimation of irregular pricing effects like steps and kinks at certain price points or threshold and saturation effects at the extremes of the observed price/price cut ranges.

3.4.2 An Illustration

For illustration, we use weekly store-level scanner data from Dominick's Finer Foods, a major supermarket chain in the Chicago metropolitan area. The data set includes unit sales, retail price and a deal code indicating the use of an in-store display for 11 brands of refrigerated orange juice (64 oz). The sample covers individual brand sales in 81 stores ($s = 1, \ldots, 81$) of the chain over a time span of 89 weeks ($t = 1, \ldots, 89$). Table 3.1 provides summary statistics pooled across the stores for average weekly prices, market shares and unit sales of the brands.

As table 3.1 reveals, the brands can be classified into three price-quality tiers: the premium brands (made from freshly squeezed oranges), the national brands (reconstituted from frozen orange juice concentrate) and the store brand (Dominick's private label brand). The differences in quality across the tiers are well represented by higher (lower) average prices for higher (lower) quality tier brands. Average weekly prices and market shares of all brands vary considerably reflecting the frequent use of promotions.

We now illustrate the usefulness of imposing monotonicity constraints to estimate price response functions considering as example the brand Florida Gold. We focus on two distributional models, namely a log-normal model

$$sales_{st} \sim LN(\eta_{st}, \sigma^2),$$

which can be equivalently written in terms of the assumption of a Gaussian distribution for the natural logarithm of the response as

$$\log(sales_{st}) \sim N(\eta_{st}, \sigma^2),$$

and a Gamma model

$$sales_{st} \sim G(\exp(\eta_{st}), \nu),$$

where $sales_{st}$ denotes the unit sales of Florida Gold in store s and week t. Note that the exponential function is the so called natural link function for a Gamma model. The scale parameter ν is supplied with a Gamma prior with parameters $a_\nu = 0.001$, $b_\nu = 0.001$ and estimated in a Metropolis-Hastings step.

As mentioned above, the use of a log-normal model is the standard approach in marketing to relate brand sales to promotional instruments. The Gamma model, on the other hand, provides high flexibility with respect to the shape of the distribution (e.g., it can take on a highly skewed distribution) and is used to demonstrate the applicability of our method in the non-Gaussian case. Like Kalyanam and Shively (1998) and van Heerde et al. (2001), we choose a semiparametric additive predictor to model sales response: with nonparametric terms for own- and cross-price effects as well as weekly effects, and parametric terms for own and competitive display and store-specific effects. According to economic theory and the empirical findings discussed in section 3.1, we expect the unit sales of Florida Gold to be an antitonic function in own promotional price and an isotonic function in competitive items' promotional prices rather than to show a nonmonotonic shape, respectively. Specifically, we estimate three variants of the semiparametric additive predictor for both the log-normal and the Gamma model:

$$\begin{aligned}
\eta_{st}^{(1)} &= f_{1_antitonic}^{RW1}(price_{st}) + f_{2_isotonic}^{RW1}(price_premium_{st}) + f_{3_isotonic}^{RW1}(price_national_{st}) \\
&\quad + f_{4_isotonic}^{RW1}(price_Dominicks_{st}) + f_5^{RW2}(week) + f_{random}(store) \\
&\quad + display_{st} + display_premium_{st} + display_national_{st} + display_Dominicks_{st}
\end{aligned}$$

$$\begin{aligned}
\eta_{st}^{(2)} &= f_{1_antitonic}^{RW2}(price_{st}) + f_{2_isotonic}^{RW2}(price_premium_{st}) + f_{3_isotonic}^{RW2}(price_national_{st}) \\
&\quad + f_{4_isotonic}^{RW2}(price_Dominicks_{st}) + f_5^{RW2}(week) + f_{random}(store) \\
&\quad + display_{st} + display_premium_{st} + display_national_{st} + display_Dominicks_{st}
\end{aligned}$$

and

$$\begin{aligned}
\eta_{st}^{(3)} &= f_1^{RW2}(price_{st}) + f_2^{RW2}(price_premium_{st}) + f_3^{RW2}(price_national_{st}) \\
&\quad + f_4^{RW2}(price_Dominicks_{st}) + f_5^{RW2}(week) + f_{random}(store) \\
&\quad + display_{st} + display_premium_{st} + display_national_{st} + display_Dominicks_{st}
\end{aligned}$$

The three variants differ in the specification of the unknown smooth functions f_1 to f_4 for own- and cross-price effects. These are estimated either by P-splines with monotonicity constraints, with first order random walk prior ($\eta^{(1)}$) or second order random walk prior ($\eta^{(2)}$), respectively, or by unconstrained P-splines with second order random walk prior ($\eta^{(3)}$) as a reference. The choice of the reference specification is based on a study conducted in Part I of Chapter 2 where superior results for P-splines with second order rather than first order random walk priors in the unrestricted case are reported. *price* denotes Florida Gold's actual price in store s and week t, and *display* is an indicator variable representing the usage (1) or nonusage (0) of an in-store display for Florida Gold in store s and week t. Similar to Blattberg and George (1991), we capture cross price effects in a more parsimonious way through the use of competitive variables at the tier level rather than the individual brand level: $price\_premium_{st}$ and $price\_national_{st}$ indicate the minimum price for competing brands within the premium brand and the national brand tier in store s and week t, respectively, whereas $price\_Dominicks_{st}$ is the actual price of Dominick's private label brand in store s and week t. It is important to note that the price of Florida Gold (which itself is a national brand) is excluded from computing $price\_national_{st}$. Accordingly, the indicator variables $display\_premium_{st}$ and $display\_national_{st}$ take the value '1' if a display is used for at least one brand within the respective tier in store s and week t, and '0' otherwise. $display\_Dominicks_{st}$ is the corresponding fixed effect for the private label brand.

The *week* covariate is incorporated to capture seasonal and missing variable (e.g., manufacturer advertising) effects, and the store covariate to accommodate differences in base sales of Florida Gold across the stores, e.g., due to their spatial location. The effect of *week* is modeled as a P-spline with second order random walk prior and *store* is incorporated as a random effect. We use cubic splines with 20 knots for all P-spline terms, except for the *week* effect, where we use 40 knots to be able to account for possibly strong time variability. The specification with 40 knots for the time effect, however, is still much less costly in terms of degrees of freedom lost than if we were to use weekly indicator variables. Finally, the hyperparameters σ^2 and ν are supplied with inverse Gamma priors $\sigma^2 \sim IG(0.001, 0.001)$ and $\nu \sim IG(0.001, 0.001)$, respectively, and are estimated simultaneously with the regression parameters. The resulting models are referred to as LN1-LN3 for the log-normal variants and G1-G3 for the Gamma model variants in the following. With regard to the sampling process, we store every 10th sample of a Markov chain of length 10,000 (after the burn-in period) to obtain 1,000 draws for each parameter and take the means as parameter estimates.

3.4.3 Model evaluation and interpretation of results

We evaluate the different models in terms of the Average Mean Squared Error (AMSE) in validation samples (also compare van Heerde et al. 2001). Specifically, we randomly split the data into nine equally-sized subsets and performed nine-fold cross-validation. For each subset, we fitted the respective model to the remaining eight subsets making up the

estimation sample and calculated the squared prediction errors of the fitted model when applied to the observations in this holdout subset (Efron and Tibshirani 1998). Let n denote the number of observations of the entire data set, and $k(i)$ the holdout subset containing observation i. Let further $\widehat{sales}_i^{-k(i)}$ indicate the fitted value of observation i computed from the estimation sample without subset $k(i)$, then the AMSE of prediction is:

$$AMSE = \frac{1}{n}\sum_{i=1}^{n}\big(sales_i - \widehat{sales}_i^{-k(i)}\big)^2.$$

Because we are interested in unit sales rather than log unit sales of Florida Gold, conditional mean predictions from the estimated log-normal models were obtained as follows (Goldberger 1968, Greene 1997):

$$\widehat{sales}_{st}^{-k(i)} = \frac{1}{1000}\sum_{k=1}^{1000}\exp\{\eta_{stk} + \sigma_k^2/2\}, \tag{3.10}$$

where η_{stk} is the additive predictor for store s, week t and stored iteration k and σ_k^2 denotes the residual variance of the respective log-normal model in iteration k. For the Gamma model, no correction factor $\sigma_k^2/2$ is required for the conditional mean predictions.

The validation results are displayed in table 3.2. Under both the log-normal and the Gamma distribution, the models with monotonicity constraints (LN1, LN2, G1, G2) clearly outperform the respective model without monotonicity constraints (LN3, G3). Interestingly, whereas in the unrestricted case the log-normal model (LN3) yields a smaller AMSE compared to the Gamma model (G3), the restricted Gamma models G1 and G2 provided the highest predictive validity. Furthermore, the differences between restricted models with first order and second order random walk priors for the nonparametric terms are virtually negligible. These results indicate that imposing monotonicity constraints on own- and cross-item price effects can substantially improve the predictive validity of a sales response model.

Figures 3.2 and 3.3 show the nonparametrically estimated own- and cross price effects for Florida Gold resulting from the log-normal models (LN1-LN3) and the Gamma models (G1-G3), respectively. Shown are the posterior means as well as 80% and 95% pointwise credible intervals. To ensure identifiability, the functions are centered to have mean zero, i.e. $1/range(x_j)\int f_j(x_j)\,dx_j = 0$. The subtracted means are added to an intercept term, which is not displayed here. As can be seen, the effects are very similar for corresponding model versions (LN1|G1, LN2|G2 and LN3|G3), except for the own price effect which reveals a stronger increase in unit sales for very low prices under the Gamma distribution. Probably, this difference in own-price response is responsible for the higher predictive validity of the Gamma models. As already indicated by the AMSE values, there is also not much difference in own- and cross-price effects between the restricted Gamma models G1 and G2. We therefore focus in the following on Gamma model G2, the model with the highest predictive validity, for interpretation of results. Importantly, the unrestricted models LN3 and G3 which are inferior in predictive validity show strong local nonmonotonicities

in both own- and cross price effects which indicates too much flexibility (strong overfitting) of an unconstrained estimation.

Our results are similar to the findings of van Heerde et al. (2001) with respect to the shape of price response functions. Specifically, the own price response curve for Florida Gold shows a reverse s-shape with an additional increase in sales for extremely low prices. This strong sales spike can be attributed to an odd pricing effect at 99 cents, the lowest observed price of Florida Gold (compare table 3.1). The cross-price response curve with respect to the premium tier brands reveals a reverse L-shape and a threshold effect for competitive prices over two dollars. In other words, only if one of the premium brands is priced lower than two dollars, unit sales of Florida Gold significantly decrease and consumers switch up to the low-priced premium brand. The cross price effect with respect to the national brand tier (the tier of Florida Gold) is s-shaped but by far less strong than the premium tier effect, which contradicts the hypothesis that brands which are priced closer to each other (like Florida Gold and the other national brands) are more competitive than brands priced farther apart (like Florida Gold and the premium brands). Finally, the cross price effect of Dominick's private label brand on Florida Gold's sales is almost negligible. Comparing the three cross price effects in magnitude, our results confirm previous empirical findings of asymmetric quality tier competition. Specifically, a price cut by a premium brand may draw substantial sales from Florida Gold, whereas a price cut by a private label brand does not. As expected, the own-price effect is much stronger than each of the cross-price effects.

Tables 3.3 and 3.4 provide parameter estimates for the display effects and the corresponding multiplier effects (Leeflang, Wittink, Wedel and Naert 2000). The multiplier effects are obtained from the transformation

$$\frac{1}{1000}\sum_{k=1}^{1000}\exp\{\gamma_{jk}\}, \quad j = 1, \ldots, 4.$$

Shown are the posterior means, posterior standard deviations and the corresponding 2.5% and 97.5% quantiles, respectively. Multipliers with values larger (smaller) than 1 indicate a positive (negative) effect on unit sales of Florida Gold. γ_{1k} denotes the own display effect of Florida Gold, and γ_{2k} to γ_{4k} refer to the tier-specific competitive display effects. k denotes the kth stored sample for the respective parameter. Except for the cross display effect of Dominick's private label brand, the display multipliers show the expected impact. For example, if a display is used for Florida Gold, its unit sales increase on average by a factor of 1.36, whereas a display for a premium brand causes a decrease in Florida Gold's unit sales of about 11% on average. The display effect with respect to the brands in the national tier (except Florida Gold) is not significant. One possible explanation for the positive cross display effect of Dominick's private label could be that promotion activities of Dominick's for its own store brand are especially distinct and not only stimulate own brand sales but also sales of some other brands in the category. As expected, the own display effect is much stronger than competitive display effects.

Finally, figure 3.4 shows estimated results for the store-specific random effect. The store effect is portrayed with a spatial map which represents the store locations of Dominick's

Finer Foods in the Chicago metropolitan area. There is a noticeable difference in base sales across stores, with an apparent drop from the coastline in the east, where we have a high concentration of stores, to the interior region in the west. We found (weak) positive correlations between the store effect and the percentage of the population under age nine (0.28) and the percentage of households with three or more members (0.24). Hence, one possible explanation for the east-west drop of base sales may be that more households with little children live in the east part of the Chicago area, and people buy more orange juice there because they are concerned with their childrens health. We abstain from depicting the estimated effect for the time covariate *week*, because it does not reveal any seasonal pattern nor a trend.

3.5 Discussion

We proposed a methodology to incorporate specific prior knowledge of a monotonic relationship between a response variable and one or more continuous covariates into (Bayesian) generalized additive models. Unlike other approaches to monotonic regression, our method offers the possibility of nonparametric monotonic modeling by penalized splines of arbitrary degree. Sampling is accomplished by block updates of nonparametric effects. An internal Gibbs sampler is employed for drawing random numbers from truncated multivariate normal densities. Convergence of the internal Gibbs sampler is fast in the Gaussian case, but might be improved for other response distributions. Our approach can also accommodate additional covariates modeled by appropriate other specifications, like fixed effects, unrestricted P-splines, random effects or spatial effects as well as varying coefficient terms and interactions of covariates. We illustrated the methodology and its practical relevance in an empirical application estimating sales response for a brand of refrigerated orange juice from store-level scanner data. Our results show that imposing monotonicity constraints for own- and cross-item price effects can considerably improve the predictive validity of a sales response model. The methodology is implemented in the public domain software package *BayesX*.

Acknowledgement:
This research has been partly financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures". We thank Ludwig Fahrmeir, Harald Hruschka and Stefan Lang for helpful discussion. The data is provided by the James M. Kilts Center, GSB, University of Chicago.

Table 3.1: Descriptive Statistics for Weekly Brand Prices, Market Shares and Unit Sales

Refrigerated Orange Juice Category (64 oz)								
	Retail Price			Market Share			Unit Sales	
Brand	Range ($)	Mean ($)	Std Dev ($)	Range (%)	Mean (%)	Std Dev (%)	Minimum	Maximum
Premium Brands:								
Tropicana Pure Premium	[1.60; 3.55]	2.95	0.53	[3;73]	15	15	6,388	100,712
Florida Natural Premium	[1.57; 3.16]	2.86	0.33	[1;53]	5	7	1,138	56,037
National Brands:								
Citrus Hill	[1.09; 2.82]	2.31	0.31	[1;78]	8	12	2,006	151,570
Minute Maid	[1.29; 2.92]	2.23	0.40	[3;87]	21	22	4,805	243,711
Tropicana	[1.49; 2.75]	2.20	0.35	[2;75]	21	23	3,041	102,629
Florida Gold	[0.99; 2.83]	2.17	0.39	[1;63]	4	8	325	150,945
Tree Fresh	[1.07; 2.48]	2.15	0.27	[1;42]	4	6	916	39,401
Store Brand:								
Dominick's	[0.99; 2.47]	1.75	0.4	[1;83]	22	22	2,170	189,462

Table 3.2: Evaluation of models in terms of AMSE.

Model specification	log-normal	Gamma
$\eta_{st}^{(1)}$ (restricted/RW1)	49930.6	49576.6
$\eta_{st}^{(2)}$ (restricted/RW2)	50045.7	49369.7
$\eta_{st}^{(3)}$ (unrestricted)	50799.5	52200.3

Table 3.3: Estimation results for the display effects (Model G2).

effect	posterior mean		2.5%-quantile	97.5%-quantile
γ_1 (display)	0.30	(0.04)	0.24	0.38
γ_2 (display_premium)	-0.12	(0.04)	-0.19	-0.05
γ_3 (display_national)	-0.02	(0.05)	-0.11	0.08
γ_4 (display_Dominicks)	0.07	(0.03)	0.00	0.14

Table 3.4: Estimation results for the display multiplier effects (Model G2).

effect	posterior mean		2.5%-quantile	97.5%-quantile
γ_1 (display)	1.36	(0.05)	1.27	1.45
γ_2 (display_premium)	0.89	(0.03)	0.83	0.95
γ_3 (display_national)	0.98	(0.05)	0.90	1.08
γ_4 (display_Dominicks)	1.07	(0.04)	1.00	1.15

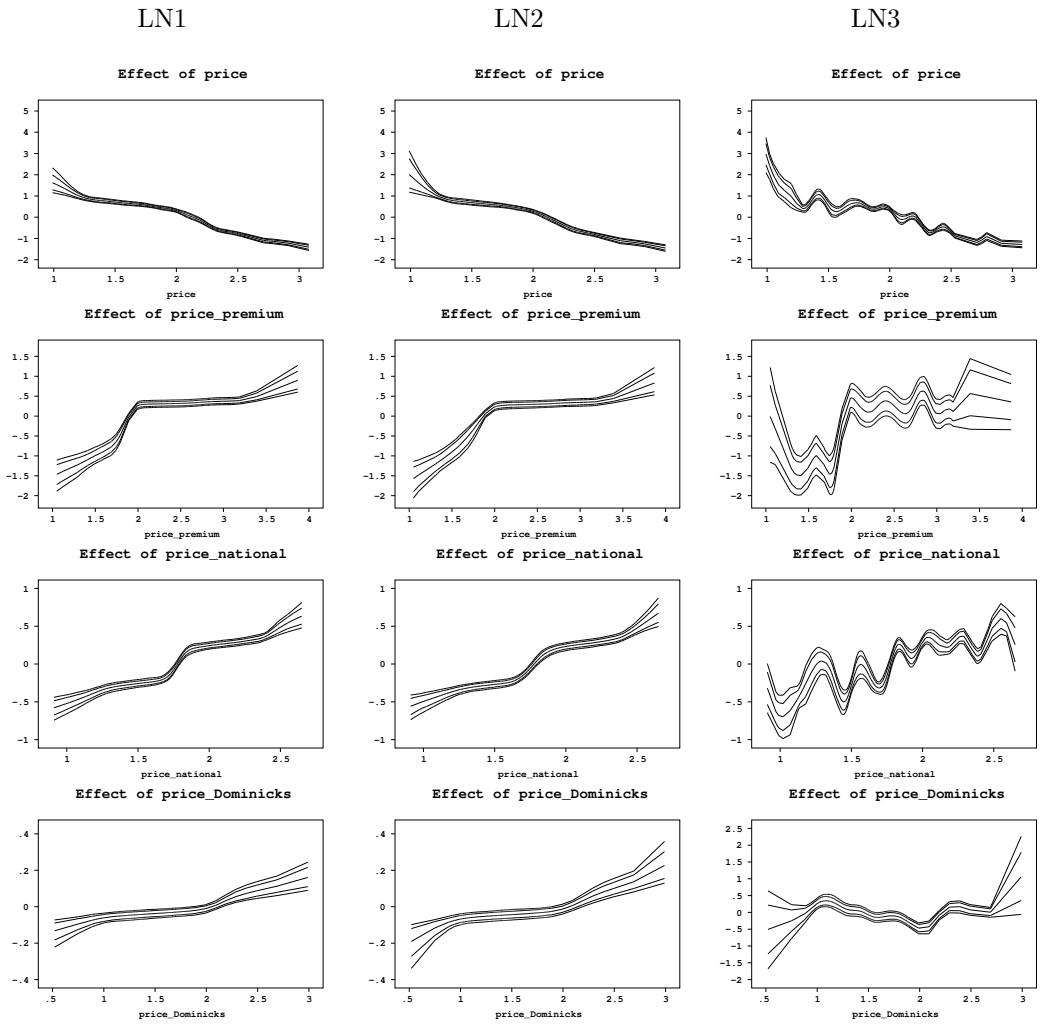


Figure 3.2: Estimated curves for own-price (*price*) and tier-specific cross-price (*price_premium*, *price_national*, *price_Dominicks*) effects on unit sales of Florida Gold. Columns 1-3 show the effects for the models LN1-LN3. Shown are the posterior means as well as 80% and 95% pointwise credible intervals.

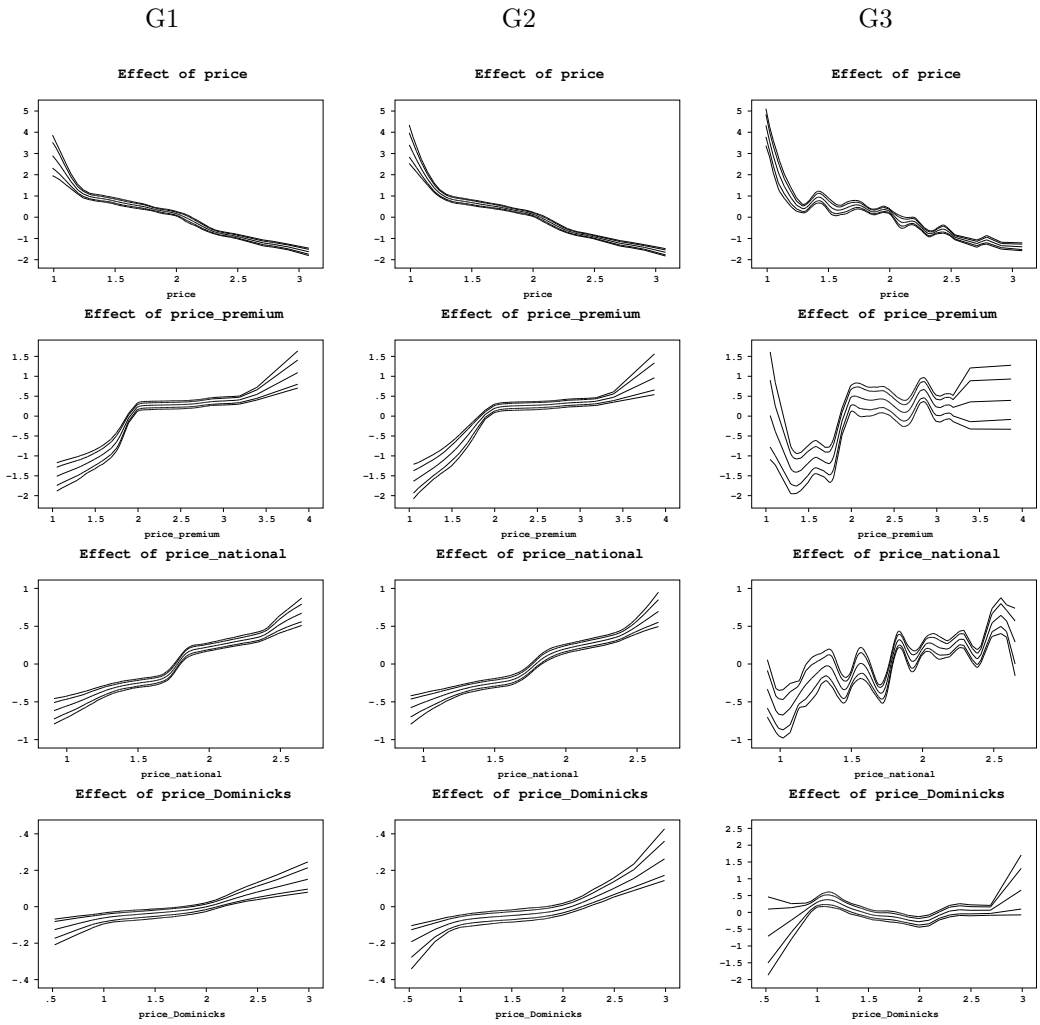


Figure 3.3: Estimated curves for own-price (*price*) and tier-specific cross-price (*price_premium*, *price_national*, *price_Dominicks*) effects on unit sales of Florida Gold. Columns 1-3 show the effects for the models G1-G3. Shown are the posterior means as well as 80% and 95% pointwise credible intervals.

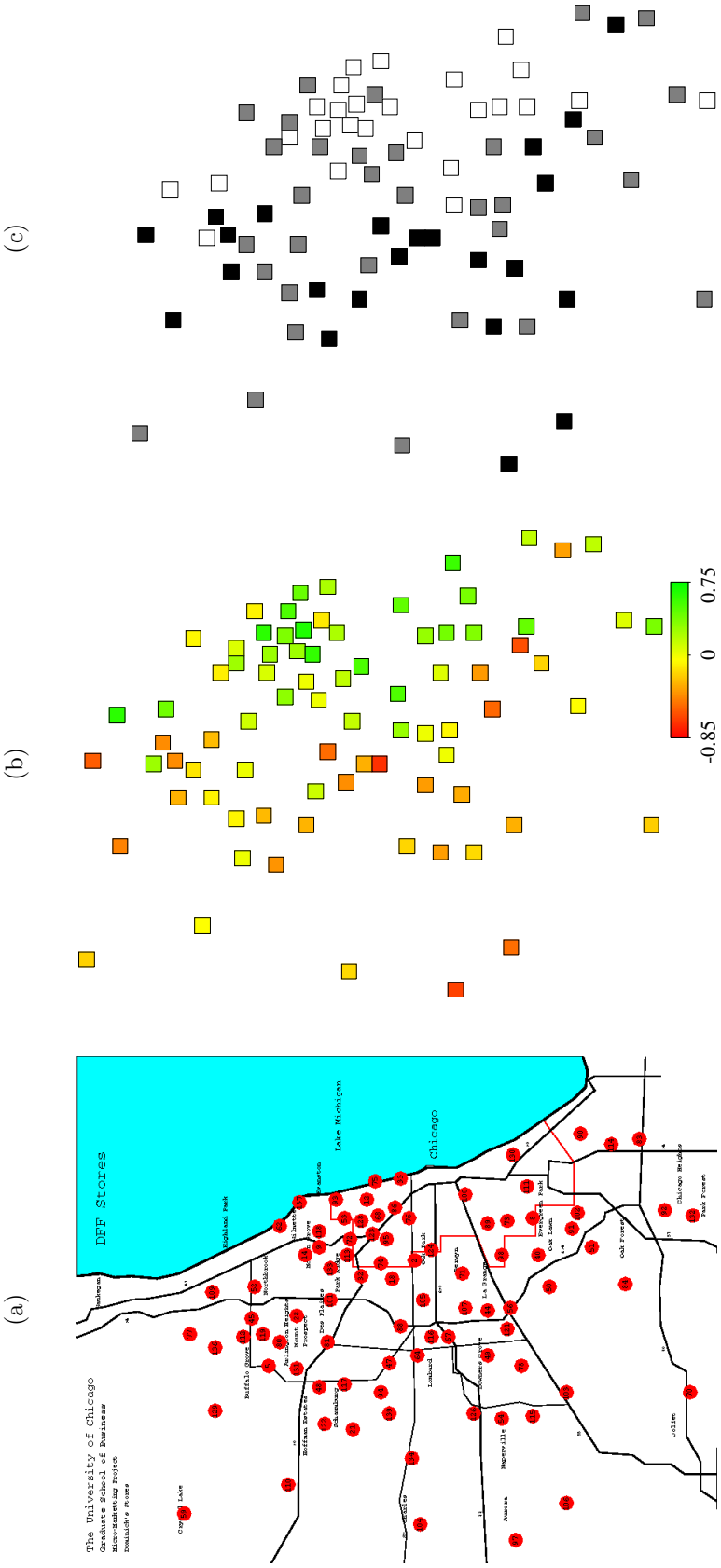


Figure 3.4: (a) Map of the Chicago metropolitan area with store locations of Dominick's Finer Foods. (b) Estimated random effect of *store* for the Gamma model (G2). (c) Posterior probabilities of *store*. White (black) indicates strictly positive (negative) 95% credible intervals, grey indicates that the 95% credible intervals contain zero.

Chapter 4

Simultaneous probability statements for Bayesian P-Splines

In Chapters 2 and 3 the focus was on modeling of effects of continuous covariates by Bayesian P-splines within structured additive regression models. However, a swelling model complexity as induced by the popularity of hierarchical Bayesian models brings along an increased demand for diagnostic tools for model selection to keep the results interpretable. In the preceding chapters we mainly used the Deviance Information Criterion (DIC) for model comparison. In addition, only pointwise credible intervals for the regression parameters and the resulting function evaluated at the observation points are available as interval estimates so far. However, for more elaborated model diagnostics simultaneous probability statements are desirable.

In this chapter we aim at computing simultaneous probability statements by two different methods, one based on the highest posterior density region and another based on simultaneous credible intervals. We derive conditions on the regression parameters of P-splines that result in a constant, linear or more generally a polynomial fit, which facilitates us to make statements on the probability for or against the suitability of a polynomial fit of a certain degree instead of a nonparametric P-spline.

Simultaneous probability statements for Bayesian P-Splines

Andreas Brezger and Stefan Lang
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

ABSTRACT

P-splines are a popular approach for fitting nonlinear effects of continuous covariates in semiparametric regression models. Recently, a Bayesian version for P-splines has been developed on the basis of Markov chain Monte Carlo simulation techniques for inference. In this work we adopt and generalize the concept of Bayesian contour probabilities to Bayesian P-splines within a generalized additive models framework. More specifically, we aim at computing the maximum credible level (sometimes called Bayesian p-value) for which a particular parameter vector of interest lies within the corresponding highest posterior density (HPD) region. We are particularly interested in parameter vectors that correspond to a constant, linear or more generally a polynomial fit. As an alternative to HPD regions simultaneous credible intervals could be used to define pseudo contour probabilities. Efficient algorithms for computing contour and pseudo contour probabilities are developed. The performance of the approach is assessed through simulation studies and applications to data for the Munich rental guide and on undernutrition in Zambia and Tanzania.

4.1 Introduction

Consider the additive model

$$y_i = \eta_i + \varepsilon_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.1}$$

where the mean of a continuous response variable y_i is the sum of nonlinear but sufficiently smooth functions $f_1, \ldots, f_p$ of the covariates $x_i = (x_{i1}, \ldots, x_{ip})'$.

Currently one of the most popular approaches for estimating the functions f_j is based on P(enalized)-splines as proposed by Eilers and Marx (1996), see also Marx and Eilers (1998) and Eilers and Marx (2004). The approach assumes that the unknown functions f_j can be approximated by a spline of degree l with equally spaced knots $x_{j,min} = \zeta_{j0} < \zeta_{j1} < \cdots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$ within the domain of x_j. The spline can be written in terms of a linear combination of $r + l$ B-spline basis functions $B_{j\rho}$, i.e.

$$f_j(x_j) = \sum_{\rho=1}^{r+l} \beta_{j\rho} B_{j\rho}(x_j). \tag{4.2}$$

By defining the design matrices X_j, where the element in row i and column ρ is given by $X_j(i, \rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor in (4.1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p.$$

Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on squared differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. In Chapter 2 a Bayesian version of P-splines is developed which is based on stochastic analogues of difference penalties as priors for the regression coefficients. More specifically, first or second order random walks are used as smoothness prior, i.e.

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{4.3}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The priors (4.3) can be equivalently written in the form of a global smoothness priors

$$\beta_j|\tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right)$$

with appropriate penalty matrix K_j. In a further stage of the hierarchy, inverse Gamma hyperpriors $p(\tau_j^2) \sim IG(a_j, b_j)$ are assigned to the variances τ_j^2 (and the overall variance parameter σ^2). Bayesian inference for the regression and variance parameters can be based on MCMC simulation. For Gaussian responses, as primarily considered in this paper, a Gibbs sampler can be used to successively update the parameters $\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2$, see Chapter 2 for details.

Currently, interval estimates are limited to *pointwise credible intervals* for the regression parameters β_j and the functions f_j evaluated at the observations. The primary goal of this paper is to develop techniques for obtaining *simultaneous probability statements* about the regression parameters and as a result about the unknown functions. More specifically, we aim at computing the *maximum credible level* (sometimes called Bayesian p-value) for which a particular parameter vector of interest lies within the corresponding highest posterior density (HPD) region. We are particularly interested in parameter vectors that

correspond to a constant, linear or more generally a polynomial fit. Since the functions f_j are centered around zero, a constant fit corresponds to $\beta_j = 0$, i.e. the particular covariate has no effect on the conditional mean of the response variable. The final goal is to assist the analyst in the model building process towards more parsimonious models. For instance, if the contour probability for a linear fit is small but relatively high for a quadratic fit, a more parsimonious model with a parametric linear fit could be used.

The plan of the paper is as follows:

- In Section 4.2.1 we review ideas recently proposed by Held (2004) for estimating and computing contour probabilities or Bayesian p-values. As an alternative to HPD regions, simultaneous credible intervals as proposed by Besag et al. (1995) could be used to define pseudo contour probabilities.

- We derive in Section 4.2.2 conditions on the regression parameters that lead to a constant, linear or in general a polynomial fit and develop efficient algorithms for computing the corresponding (pseudo) contour probabilities. So far, algorithms and software are restricted to models with Gaussian responses and models where latent Gaussian responses can be obtained through data augmentation. The latter is possible for most categorical regression models, see Albert and Chib (1993) for probit models and Holmes and Held (2004) for logit models.

- The performance of the different approaches is assessed through simulation studies (Section 4.3). We finally present in Section 4.4 applications to data for the Munich rental guide and on undernutrition in Zambia and Tanzania.

4.2 Contour probabilities for P-Splines

In order to keep the notation as simple as possible the development in this section is presented for a particular covariate x with regression parameters β. Hence the index j in (4.2) and everywhere else is suppressed.

4.2.1 Contour probabilities

Suppose we are interested in simultaneous posterior probability statements for a particular parameter vector $\beta = \beta^*$. The posterior *contour probability* $P(\beta^* \,|\, y)$ of β^* is defined as 1 minus the content of the HPD region of $p(\beta \,|\, y)$ which just covers β^*, i.e.

$$P(\beta^* \,|\, y) = P\{p(\beta \,|\, y) \leq p(\beta^* \,|\, y) \,|\, y\}, \tag{4.4}$$

see Box and Tiao (1973) and Held (2004). Note that $p(\beta \,|\, y)$ is treated here as a random variable. In the following we briefly review concepts for estimating the probability (4.4) from posterior samples $\beta^{(t)}$, $t = 1, \ldots, T$ obtained via MCMC simulation.

Held (2004) proposes to estimate (4.4) by

$$\widehat{P(\beta^* \,|\, y)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{p(\beta^{(t)} \,|\, y) \leq p(\beta^* \,|\, y)\}, \tag{4.5}$$

i.e. the proportion of the MCMC samples for which the posterior density is smaller than the density of the point of interest β^*.

Unfortunately the functional form of the marginal density $p(\beta \,|\, y)$ is unknown (otherwise MCMC would not be necessary) and we have to employ some method of density estimation to obtain estimates $\widehat{p(\beta^{(t)} \,|\, y)}, t = 1, \ldots, T$, and $\widehat{p(\beta^* \,|\, y)}$. For (latent) Gaussian responses the full conditionals $p(\beta \,|\, \cdot)$, i.e. the conditional densities of β given the data and the remaining parameters, are available and an approach based on Rao-Blackwellization seems natural (Held 2004), since the Rao-Blackwell estimate is more efficient than any other density estimate based on $\beta^{(1)}, \ldots, \beta^{(T)}$ and no smoothing parameter is involved. Estimates for the marginal density $p(\beta \,|\, y)$ can be obtained using the Rao-Blackwell theorem

$$\widehat{p(\beta \,|\, y)} = \frac{1}{T} \sum_{v=1}^{T} p(\beta \,|\, \alpha_{-}^{(v)}, y), \tag{4.6}$$

where $\alpha_{-}^{(v)}$ comprises all model parameters excluding β and hence $p(\beta \,|\, \alpha_{-}^{(v)}, y)$ denotes the full conditional density of β. As an alternative to the mean in (4.6) Held (2004) suggests to use the median, i.e.

$$\widehat{P(\beta \,|\, y)} = med_{1 \leq v \leq T} \left\{ p(\beta \,|\, \alpha_{-}^{(v)}, y) \right\}. \tag{4.7}$$

As an advantage, the estimated contour probabilities are invariant to monotonic transformations of $p(\beta \,|\, y)$ in (4.4). For instance, one could replace $p(\beta \,|\, \alpha_{-}^{(v)}, y)$ in (4.7) by the log density, i.e.

$$\log(\widehat{P(\beta \,|\, y)}) = med_{1 \leq v \leq T} \left\{ log(p(\beta \,|\, \alpha_{-}^{(v)}, y)) \right\}. \tag{4.8}$$

Usually, this is computationally more favorable than using the density directly (see Subsection 4.2.3) and also more robust against extreme samples.

Summarizing, the contour probability (4.4) is estimated by replacing the marginal densities with (4.6), (4.7), or (4.8) if log densities are used. Using (4.8), for instance, we obtain

$$\widehat{P(\beta^* \,|\, y)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1} \left\{ med_{1 \leq v \leq T} \left\{ log(p(\beta^{(t)} \,|\, \alpha_{-}^{(v)}, y)) \right\} \leq med_{1 \leq v \leq T} \left\{ log(p(\beta^* \,|\, \alpha_{-}^{(v)}, y)) \right\} \right\} \tag{4.9}$$

Pseudo contour probabilities based on credible intervals

As an alternative to the definition of contour probabilities via HPD regions, we could base the definition on simultaneous credible intervals for the parameter β^* of interest.

For instance, Besag et al. (1995) propose to define a simultaneous credible interval as the hyperrectangular defined by

$$[\beta_\rho^{[T+1-t^*]}, \beta_\rho^{[t^*]}] \quad \rho = 1, \ldots, r+l, \tag{4.10}$$

where $\beta_\rho^{[t]}$ denotes the ordered samples of the parameter β_ρ. The index t^* is the smallest integer such that the hyperrectangular (4.10) contains at least 100α percent of the samples $\beta^{(1)}, \ldots, \beta^{(T)}$ if α is the desired level of the credible interval.

The *(pseudo) contour probability* $P(\beta^* \,|\, y)$ for β^* can now be defined as 1 minus the smallest credible level, for which β^* is contained in the corresponding credible interval.

4.2.2 Contour probabilities for P-Splines

In the context of P-splines, we are particularly interested in parameters $\beta = \beta^*$ that lead to a constant, linear or in general a polynomial fit. Since P-splines are centered around zero a constant fit corresponds to $\beta^* = 0$, i.e. the corresponding covariate is excluded from the predictor. In this section we determine conditions on the regression parameters that lead to a polynomial fit rather than a piecewise polynomial as is generally the case.

It can be shown that a spline $f(x)$ reduces to a polynomial of degree $s \leq l$ if the $(s+1)$-th differences of the regression parameters are zero, i.e.

$$\Delta^{s+1}\beta_\rho = 0, \quad \rho = s+2, \ldots, r+l, \tag{4.11}$$

or in matrix notation

$$D_{s+1}\beta = 0,$$

where D_{s+1} is a difference matrix of order $s+1$. A proof can be found in Appendix A.3.

In order to compute (pseudo) contour probabilities the full conditional of $D_s\beta$ must be computed. The full conditional of β is multivariate Gaussian

$$\beta \,|\, \alpha_-, y \sim N(m, P^{-1}) \tag{4.12}$$

with

$$P = \frac{1}{\sigma^2}X'X + \frac{1}{\tau_j}K, \quad m = P^{-1}\frac{1}{\sigma^2}X'(y - \tilde{\eta}).$$

Here, $\tilde{\eta}$ is the part of the predictor associated with all remaining effects in the model. Thus $D_s\beta =: \tilde{\beta}$ is also multivariate Gaussian

$$\tilde{\beta} \,|\, \alpha_-, y \sim N(\tilde{m}, \tilde{P}^{-1}), \tag{4.13}$$

with $\tilde{m} = D_s m$ and $\tilde{P} = D_s P^{-1} D_s'$. Note that for the special case $s = 0$, i.e. $D_s = I$, we recover (4.12) as full conditional for $D_s\beta$.

4.2.3 Computational aspects

This section is concerned with computational aspects of the estimator (4.9). We will distinguish the two cases $s = 0$ and $s > 0$.

In the case $s = 0$ we have to evaluate

$$log(p(\beta^{(t)} \mid \alpha_-^{(v)}, y)) = \frac{1}{2}\log(|P^{(v)}|) - \frac{1}{2}(\beta^{(t)} - m^{(v)})'P^{(v)}(\beta^{(t)} - m^{(v)}) \tag{4.14}$$

for $t, v = 1, \ldots, T$ in order to estimate (4.9). Here, $P^{(v)}$ is the posterior precision matrix evaluated at the v-th sample of τ^2 and σ^2 and $m^{(v)}$ is the posterior mean evaluated at the v-th sample of P, σ^2 and $\tilde{\eta}$. It is useful to decompose the quadratic form in (4.14) by

$$\begin{aligned}(\beta^{(t)} - m^{(v)})'P^{(v)}(\beta^{(t)} - m^{(v)}) = \\ \frac{1}{(\sigma^2)^{(v)}}(\beta^{(t)})'X'X\beta^{(t)} + \frac{1}{(\tau^2)^{(v)}}(\beta^{(t)})'K\beta^{(t)} + (m^{(v)})'P^{(v)}m^{(v)} - 2(m^{(v)})'P^{(v)}\beta^{(t)},\end{aligned}$$

This shows that (4.9) can be evaluated by computing and storing the samples $\log(|P^{(t)}|)$, $(\beta^{(t)})'X'X\beta^{(t)}$, $(\beta^{(t)})'K\beta^{(t)}$, $(m^{(t)})'P^{(t)}m^{(t)}$ and $(m^{(v)})'P^{(v)}\beta^{(t)}$. Except $(m^{(v)})'P^{(v)}\beta^{(t)}$ these quantities are obtained as a by product of the MCMC simulation run. For $t \leq v$, $t, v = 1, \ldots, T$ it is also possible to store $(m^{(v)})'P^{(v)}\beta^{(t)}$. For $t > v$ the quantity $(m^{(v)})'P^{(v)}\beta^{(t)}$ must be computed after MCMC simulation. This is facilitated by storing $(m^{(v)})'P^{(v)}$ after every iteration of the MCMC sampler.

The case $s > 0$ is computationally more demanding. In this case the log densities

$$\log(p(\tilde{\beta}^{(t)} \mid \alpha_-^{(v)}, y)) = \frac{1}{2}\log(|\tilde{P}^{(v)}|) - \frac{1}{2}(\tilde{\beta}^{(t)} - \tilde{m}^{(v)})'\tilde{P}^{(v)}(\tilde{\beta}^{(t)} - \tilde{m}^{(v)})$$

must be computed. Evaluation of the quadratic form yields

$$(\tilde{\beta}^{(t)} - \tilde{m}^{(v)})'\tilde{P}^{(v)}(\tilde{\beta}^{(t)} - \tilde{m}^{(v)}) = (\tilde{\beta}^{(t)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)} + (\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{m}^{(v)} - 2(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}.$$

Hence the quantities $\log(|\tilde{P}^{(v)}|)$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{m}^{(v)}$ can be computed as a by product of the MCMC sampler and stored in every iteration. However, the quantities $(\tilde{\beta}^{(t)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ can only be stored for $t \leq v$. For $t > v$ both quantities must be computed *after* the MCMC run.

Now we can compute $med_v \left\{\log(p(\tilde{\beta}^{(t)}|\alpha_-^{(v)}, y))\right\}$ for all t in two ways which differ in the order of evaluations:

Algorithm 1:

For $t = 1, \ldots, T$:

1. For $v = 1, \ldots, T$:
   (a) If $t > v$:
   Compute $\tilde{P}^{(v)}$ and with it the quantities $(\tilde{\beta}^{(t)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$.
   (b) Compute $\log(p(\tilde{\beta}^{(t)}|\alpha_-^{(v)}, y))$.
2. Compute $med_v \left\{\log(p(\tilde{\beta}^{(t)}|\alpha_-^{(v)}, y))\right\}$.

This algorithm is very time consuming, because $\tilde{P}^{(v)}$ has to be computed $T(T-1)/2$ times.

The second algorithm is:

Algorithm 2:

1. For $v = 1, \ldots, T$:
   (a) Compute $\tilde{P}^{(v)}$.
   (b) For $t = 1, \ldots, T$:
   If $t \leq v$: Compute $\log(p(\tilde{\beta}^{(t)} \,|\, \alpha_-^{(v)}, y))$ based on the stored quantities.
   If $t > v$:
   Compute first $(\beta^{(t)})'\tilde{P}^{(v)}\beta^{(t)}$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ and then $\log(p(\tilde{\beta}^{(t)} \,|\, \alpha_-^{(v)}, y))$.
2. For $v = 1, \ldots, T$: Compute $med_v \left\{\log(p(\tilde{\beta}^{(t)} \,|\, \alpha_-^{(v)}, y))\right\}$.

The drawback of this algorithm is that it takes an enormous amount of memory space, because we have to create a $T \times T$ matrix to store all values $\log(p(\tilde{\beta}^{(t)} \,|\, \alpha_-^{(v)}, y))$, $t, v = 1, \ldots, T$, before computing the median. A remedy is to take only every k-th sample to estimate $\widehat{p(\tilde{\beta}^{(t)}|y)}$, but the memory requirement is still extraordinarily high.

As an alternative to the direct computation of the medians, we propose to use the method of stochastic approximation as described in Tierney (1983). The advantage is, that the quantiles can be estimated by a very space-efficient recursive procedure. Throughout this work we use Algorithm 2 together with stochastic approximation of quantiles to avoid extensive use of memory space.

4.3 Simulations

We realized an extensive simulation study in order to compare the performance of contour and pseudo contour probabilities. We investigated the functions

$$y_i = 1 + k \cdot \sin(2\pi x_i) + \epsilon_i, \quad k = 0.0, 1.0, 1.5, \tag{4.15}$$

and

$$y_i = 1 + x_i + k \cdot \sin(2\pi x_i) + \epsilon_i, \quad k = 0.0, 1.0, 1.5. \tag{4.16}$$

For x we chose 100 equidistant design points in the interval $[0, 1]$ and generated data sets with 250 replications of each of the models (4.15) and (4.16) with $\epsilon_i \sim N(0, 0.5)$. This corresponds to a signal to noise ratio (SNR) of $0.0, 1.0$ and 2.25 for $k = 0, 1.0, 1.5$ and model (4.15) and a SNR of approximately $0.17, 0.53$ and 1.47 for $k = 0, 1.0, 1.5$ and model (4.16), respectively. We used an IG(0.001,0.001) prior for the scale parameter σ^2 and the variance parameter τ^2. Figures 4.1 and 4.2 display the simulated functions as well as a typical replication from the generated response for $k = 0, 1.0, 1.5$.

We compare the results in terms of the 'p-values' obtained from contour probabilities based on the median and the mean of the log-density, and from pseudo contour probabilities. Figure 4.3 shows boxplots of p-values for model (4.15). Note that $\Delta^s\beta = 0$ corresponds to a constant fit (i.e. no effect) for both, $s = 0$ and $s = 1$. The results for model (4.16) exhibit mainly the same behavior and are displayed in Figure 4.4. The results of both models can be summarized as follows:

- *No effect (SNR=0.0)*
  As we could have expected, for a signal to noise ratio of 0.0 the contour probabilities are close to one for all difference orders considered, i.e. p-values give no evidence of any influence of the covariate at all. Pseudo contour probabilities do not suggest the existence of an influence of the covariate either, though they are considerably lower than the contour probabilities.

  It is striking that pseudo contour probabilities show a noticeable difference between difference orders $s = 0$ and $s = 1$, though both correspond to the probability for no effect of the covariate. Held (2004) reports severe underestimation for $s = 0$ and conjures that this comes from strong correlations between successive parameters. Since the correlation decreases when considering first differences of the parameters instead of the parameters directly the problem becomes less distinctive. This may explain the big differences between $s = 0$ and $s = 1$.

- *Very low to low signal to noise ratio (SNR=0.17, 0.53, 1.0)*
  For the very low and low signal to noise ratios (0.17, 0.53 in model 4.16, 1.0 in model 4.15) the p-values clearly decrease for all difference orders smaller than 4, i.e. the posterior probabilities for a (at least) cubic effect increase. However, neither contour probabilities nor pseudo contour probabilities give clear cut results and hence further investigation is advisable. An exception are p-values from pseudo contour probabilities based on 0-th order differences. Here, pseudo contour probabilities exhibit mainly very low p-values. However, this may be due to the underestimation mentioned by Held (2004).

- *Medium signal to noise ratio (SNR=1.47, 2.25)*
  For medium signal to noise ratios (1.47 in model 4.16, 2.25 in model 4.15) the contour probabilities for parametric fits with polynomials of degree smaller than three

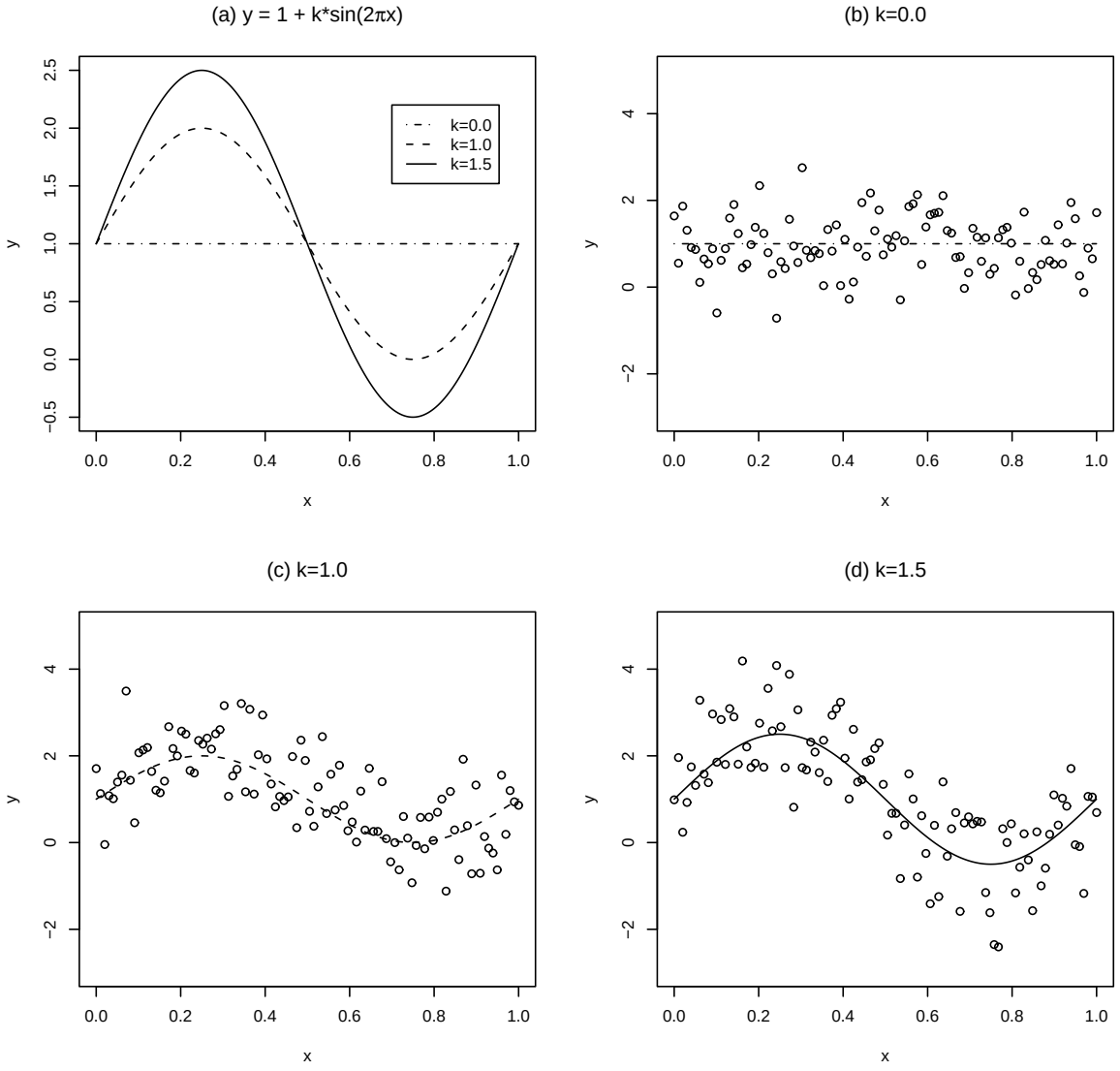


Figure 4.1: True functions for model (4.15) (a) and a typical replication of y for $k = 0, 1.0, 1.5$ (b)-(d).

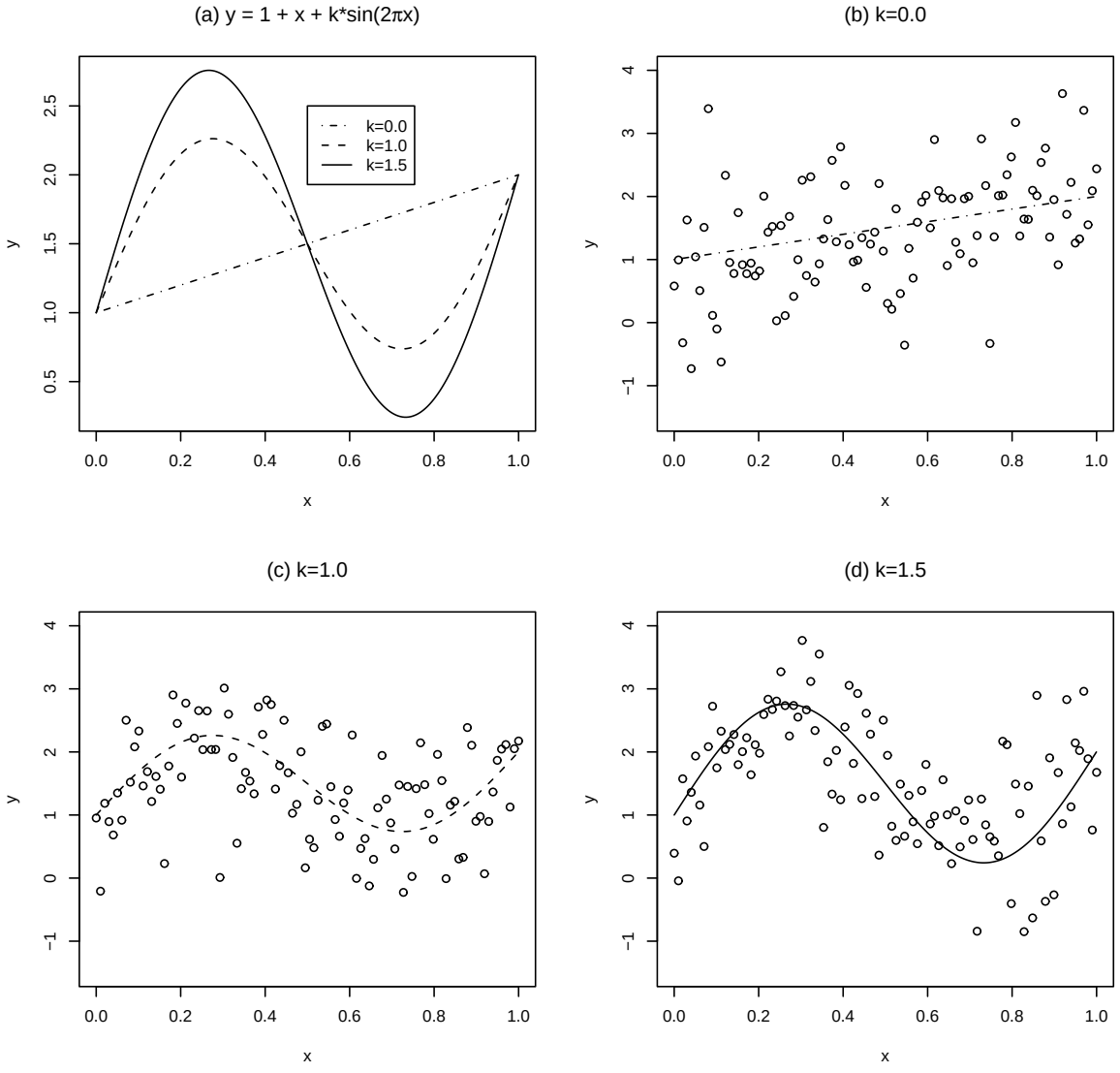


Figure 4.2: True functions for model (4.16) (a) and a typical replication of y for $k = 0, 1.0, 1.5$ (b)-(d).

(i.e. difference order smaller than 4) are very small, suggesting that a more flexible modeling is needed. However, the need of a polynomial of degree higher than 3 is rather unlikely a posteriori. This is in perfect agreement with the data, since a sine curve can be approximated by a polynomial of degree 3 without major deviations (compare Figure 4.5). Pseudo contour probabilities, on the other hand, perform very poorly for difference orders higher than 1.

- *Contour probabilities versus pseudo contour probabilities*
  It turns out that p-values based on pseudo contour probabilities are apparently smaller than that obtained from the contour probabilities for very low signal to noise ratios. This is in accordance with findings of Held (2004) who reported severe underestimation of p-values especially in the case of difference order $s = 0$, but also - to a smaller degree - when considering first differences.

  In contrast, pseudo contour probabilities behave rather conservative regarding higher differences compared to contour probabilities. For a SNR of 2.25 p-values in favor of a parametrization by polynomials of a degree higher than quadratic are still reasonably close to one.

- *Contour probabilities based on the median/mean of log-density*
  Estimated p-values may differ considerably regarding on which definition they are based. In our simulation study we compared p-values based on the median or on the mean of the log-density, respectively. We found p-values based on the mean of the log-density to be noticeably higher than the ones based on the median.

We conclude that pseudo contour probabilities seem to underestimate the p-values regarding the decision whether a covariate has an effect on the response or not, whereas for the decision for modeling an effect linearly (or by a polynomial of higher degree) they seem to behave too conservative. Contour probabilities seem to give the most reasonable results. However, it is difficult to base model selection solely on contour probabilities in cases when the obtained p-values lie in a medium range (i.e. between 0.1 and 0.4, approximately).

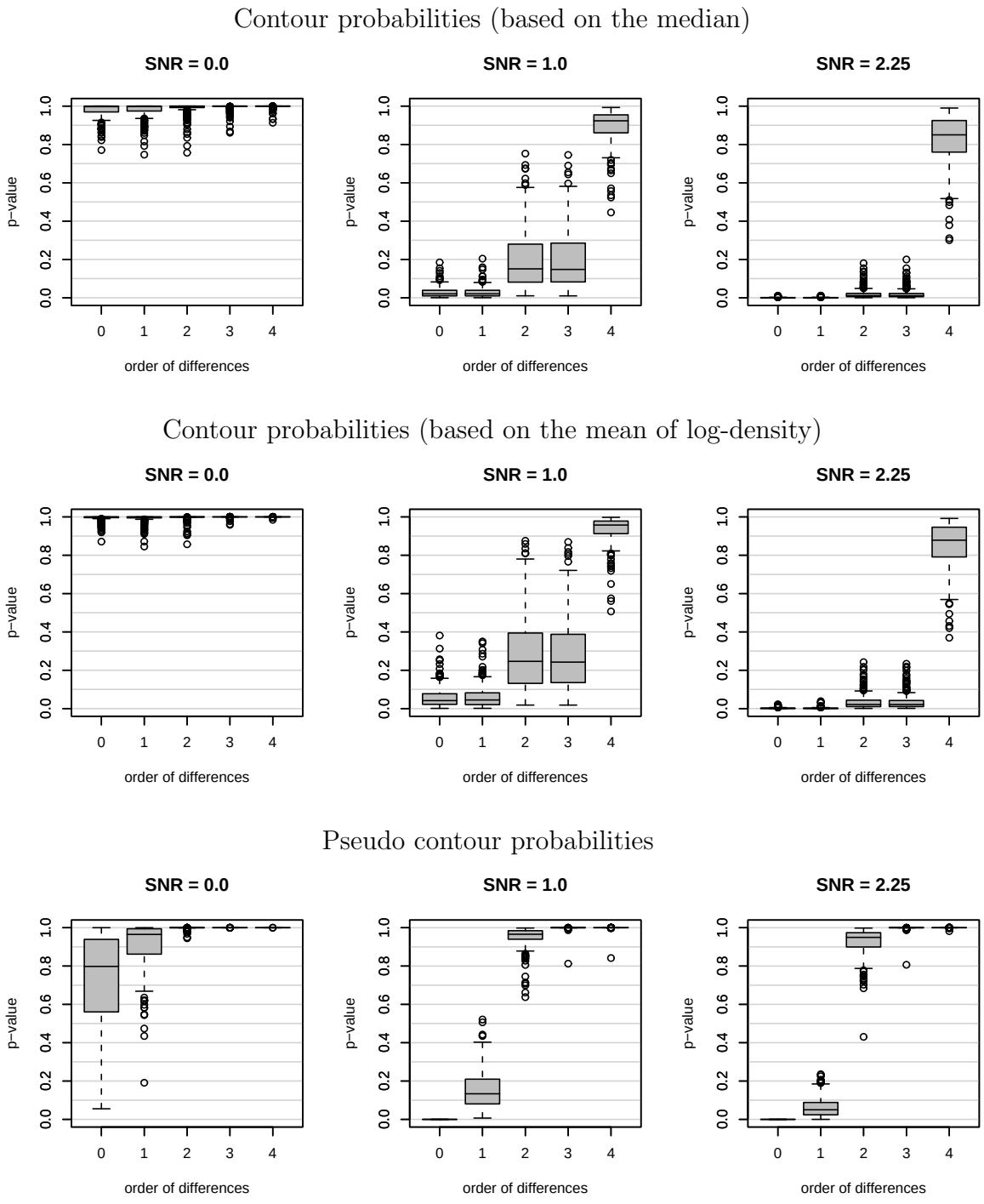


Figure 4.3: Boxplots of p-values obtained from contour probabilities based on the median (top), contour probabilities based on the mean of the log-density (middle), and pseudo contour probabilities (bottom) for different SNRs and difference orders (model 4.15). Difference order $s = 0$ and $s = 1$ corresponds to no effect, $s = 2$ $(3, 4)$ corresponds to a linear (quadratic, cubic) effect.

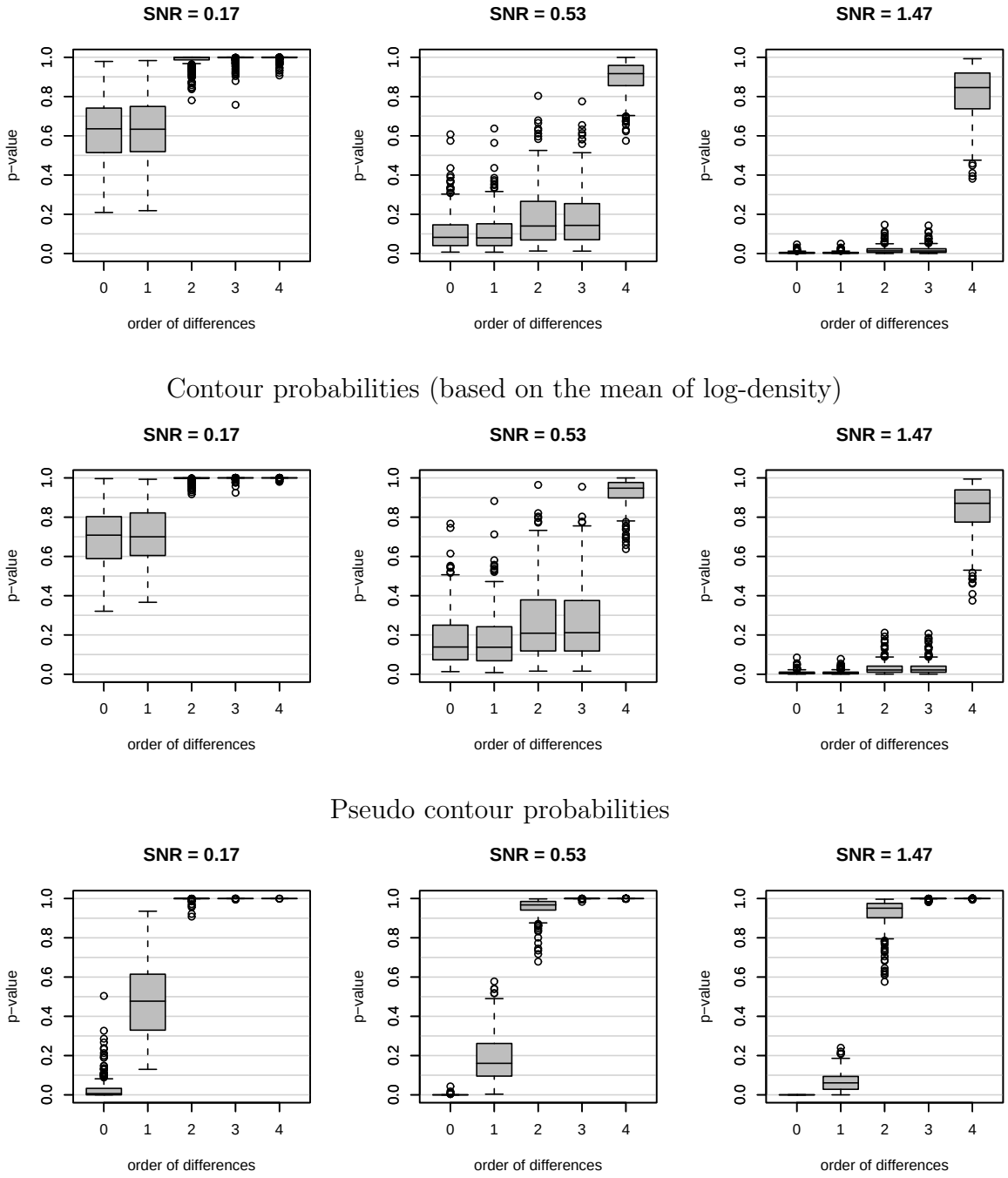


Figure 4.4: Boxplots of p-values obtained from contour probabilities based on the median (top), contour probabilities based on the mean of the log-density (middle), and pseudo contour probabilities (bottom) for different SNRs and difference orders (model 4.16). Difference order $s = 0$ and $s = 1$ corresponds to no effect, $s = 2$ $(3, 4)$ corresponds to a linear (quadratic, cubic) effect.

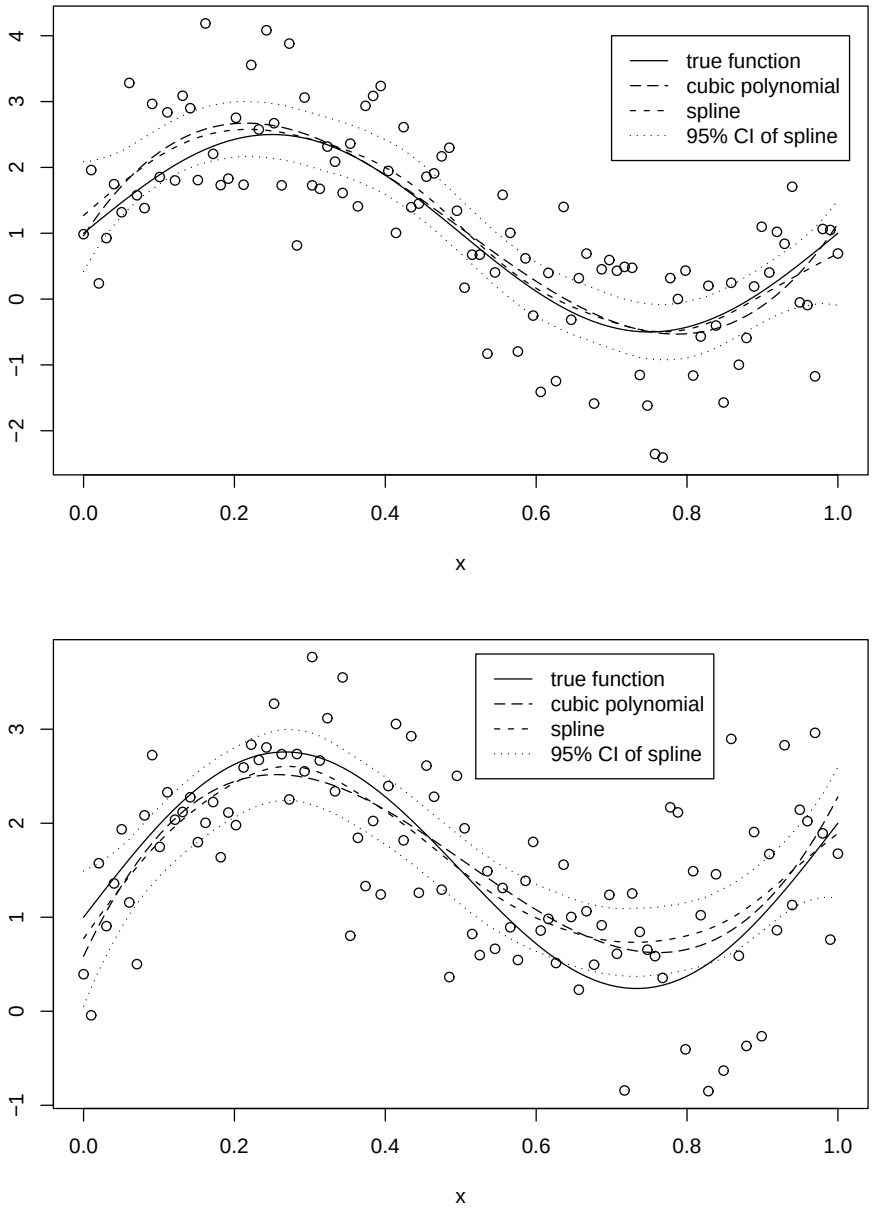


Figure 4.5: True function, cubic fit, spline fit and 95% pointwise credible levels for a selected replication of model (4.15) (top) and model (4.16) (bottom) for $k = 1.5$. Circles visualize the observation points.

4.4 Applications

In this section we illustrate the performance of the previously described model selection tools by applications to complex data sets. First, we reanalyze data from the official Munich rental guide of the year 2003 using the model developed by Fahrmeir, Biller, Brezger, Gieger, Hennerfeind, Jerak and Schmid (2003). We omit a detailed description of the data at this point and refer the reader back to Chapter 2, where we deal with data from the munich rental guide from 1999. Here, mostly the same covariates are involved compared to the analysis in Chapter 2.

Our second example investigates undernutrition of children in Zambia and Tanzania and is based on data already analyzed by Kandala, Lang, Klasen and Fahrmeir (2001). We give a brief description of the data in Subsection 4.4.2.

4.4.1 Rental guide

Compared to the model considered in Chapter 2 the official munich rental guide 2003 does not contain a nonparametric interaction effect nor a structured or unstructured random effect for spatial heterogeneity. Instead, dummy variables according to an experts assessment of the locations of flats in munich in three categories (average, good, top) are included. Thus, our model has a semiparametric predictor of the form

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + \gamma' x,$$

where F denotes the floor space in square meters, Y denotes the year of construction and γ contains all covariates to be modeled parametrically, e.g. the quality of the kitchen and bath equipment, location of the flat, etc. (compare Chapter 2). Since we aim at deciding whether the nonparametric modeling of the continuous covariates F and Y could possibly replaced by a more parsimonious polynomial fit, we do not compare the results for the fixed effects to the results of Fahrmeir et al. (2003).

By visual inspection, the nonparametric effects as depicted in Figure 4.6 seem justified and one presumes that they can not be adequately modeled by low degree polynomials. Table 4.1 displays the p-values obtained from contour probabilities and pseudo contour probabilities. Obviously, there is strong evidence for the need of nonparametric effects for F and Y regarding contour probabilities, as all p-values are either exactly or at least near zero. However, pseudo contour probabilities only suggest a linear effect of both covariates. Therefore we additionally estimate the model with only fixed effects included, i.e. we model F and Y linearly. Comparison of the DIC (Table 4.2) shows that the semiparametric model clearly outperforms the parametric model suggested by pseudo contour probabilities.

Figure 4.7 depicts the linear and nonparametric estimates for F and Y together with the partial residuals

$$r_F = y - \hat{\eta}^{-F} \quad \text{and} \quad r_Y = y - \hat{\eta}^{-Y},$$

where $\hat{\eta}^{-F}$, $\hat{\eta}^{-Y}$ is the estimated predictor with the estimated effect of F and Y, respectively, excluded. Figure 4.8 displays the mean and the standard deviation of the partial

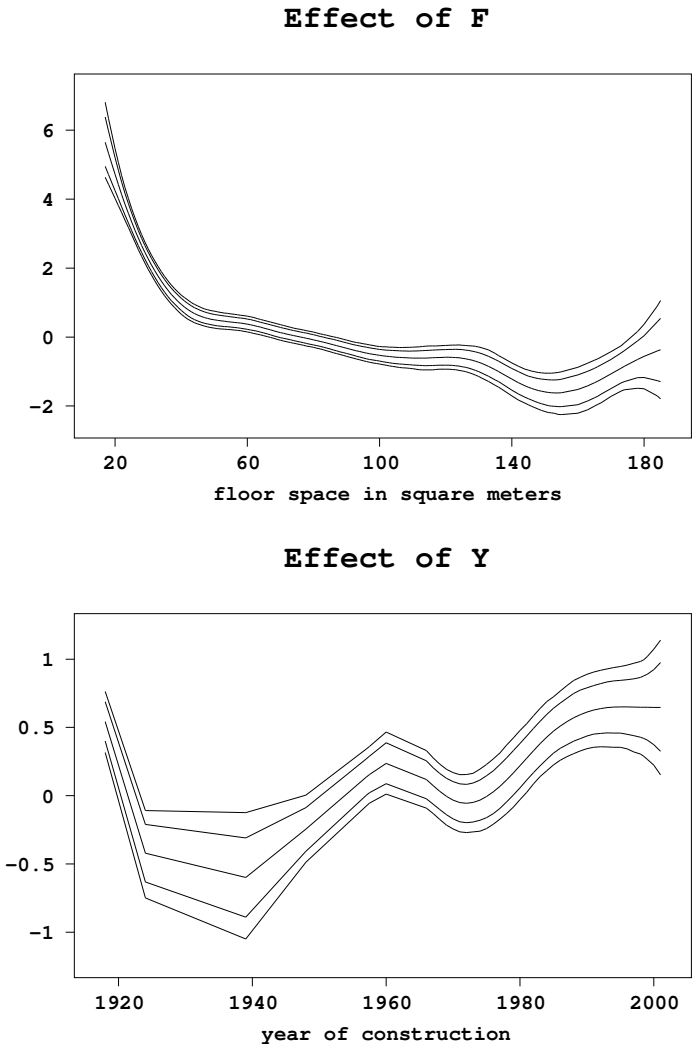


Figure 4.6: Nonparametric effects of `F` and `Y` for IG(0.001,0.001) priors on τ_F^2, τ_Y^2 and σ^2.

residuals. Note, that for F the data is rounded to integer values to compute means and standard deviations. A comparison between the figures for the linear and the nonparametric model gives further evidence for the superiority of the semiparametric model, since the nonparametric estimates show a clearly better adaptation to the partial residuals.

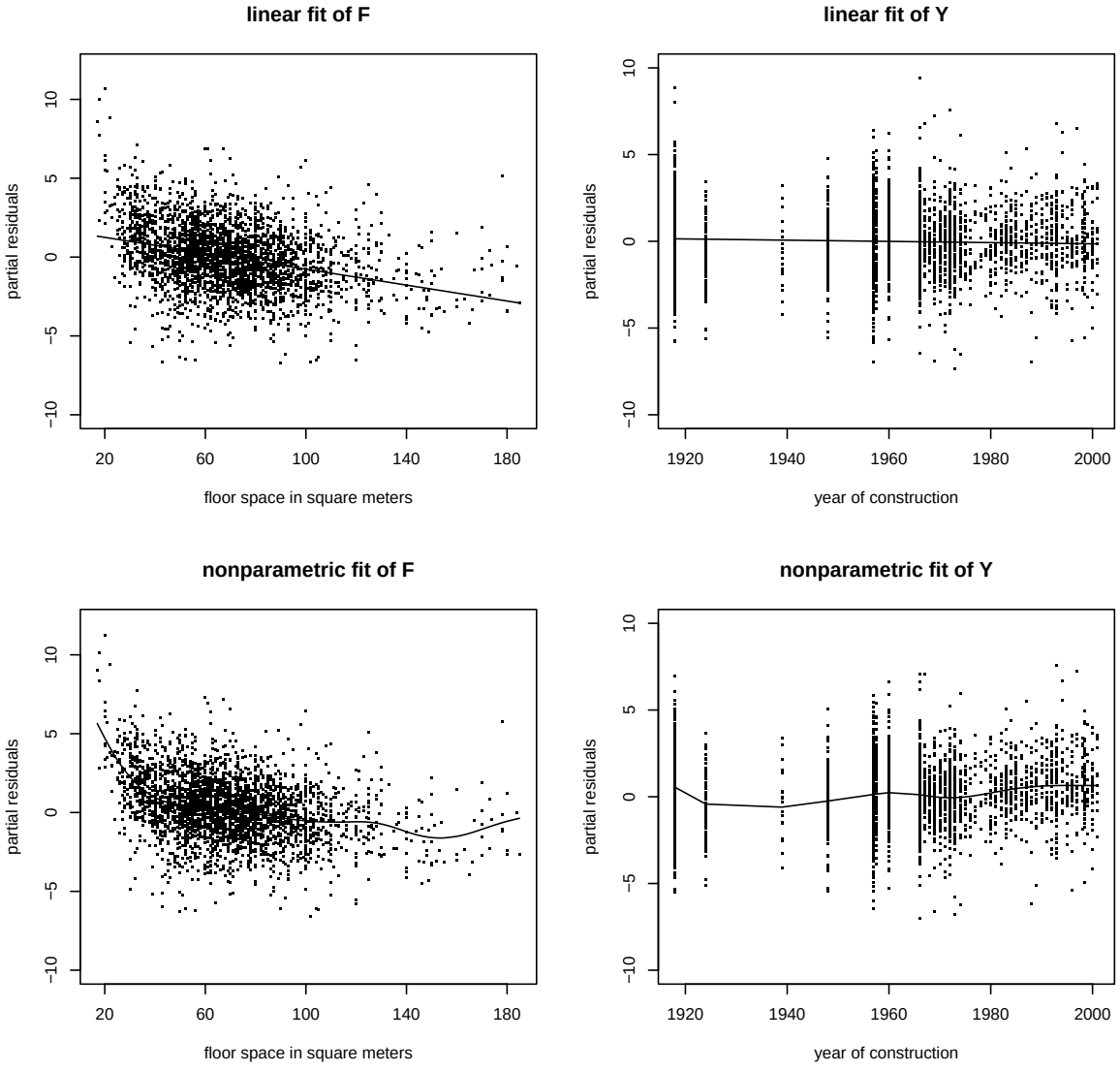


Figure 4.7: Estimated function (solid line) and partial residuals (dots) for F (left panel) and Y (right panel). Shown are the results for the parametric (top) and the nonparametric fit (bottom).

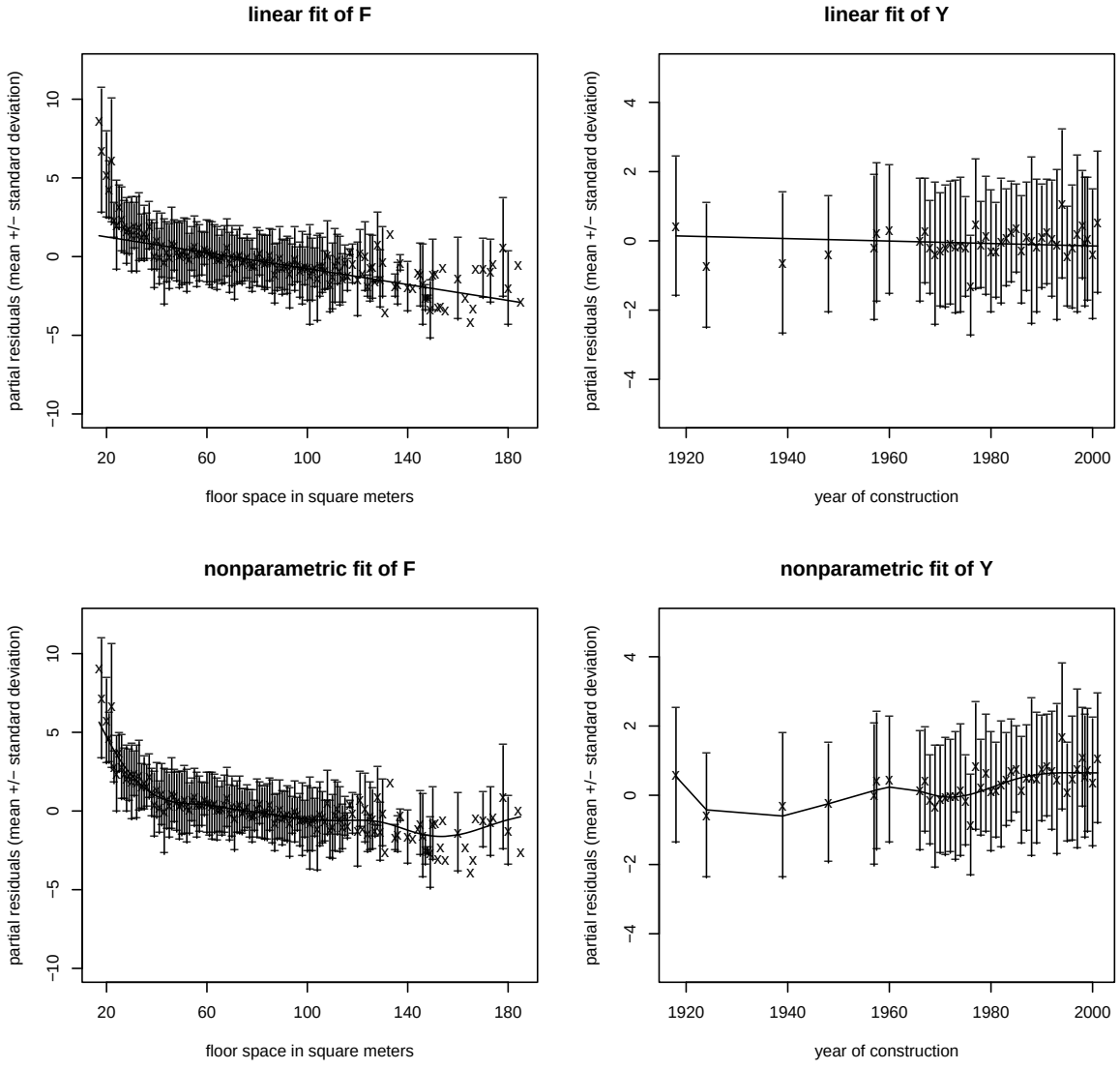


Figure 4.8: Estimated function (solid line), mean (symbol 'x') and mean ± standard deviation of partial residuals (symbol '-') for F (left panel) and Y (right panel). Shown are the results for the parametric (top) and the nonparametric fit (bottom).

Table 4.1: Contour probabilities based on median, mean of log-density and mean of density and pseudo contour probabilities for the effects of `F` and `Y` with IG(0.001,0.001) priors on τ_F^2, τ_Y^2 and σ^2.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`F` (based on median)	0.0	0.0	0.0	0.01	0.03
`F` (based on mean of log-density)	0.0	0.0	0.0	0.01	0.04
`F` (pseudo contour probabilities)	0.0	0.0	0.46	0.94	1.0
`Y` (based on median)	0.0	0.02	0.01	0.11	0.12
`Y` (based on mean of log-density)	0.01	0.03	0.03	0.17	0.19
`Y` (pseudo contour probabilities)	0.0	0.0	0.70	0.99	0.99

Table 4.2: Deviance, effective degrees of freedom (pD) and DIC for the parametric and the semiparametric model using IG(0.001,0.001) priors on τ_F^2, τ_Y^2 (in the semiparametric model) and σ^2.

	Deviance $D(\bar{\theta})$	pD	DIC
parametric model	12629.7	26.0	12681.7
semiparametric model	12400.0	41.8	12483.6

4.4.2 Undernutrition in Zambia and Tanzania

The Demographic Health Surveys (DHS) of Tanzania and Zambia, both conducted in 1992, draw a representative sample of women in reproductive age in the two countries. Thereafter they administer a questionnaire and an anthropometric assessment of themselves and their children that where born within the previous five years. The data contains 6299 cases in Zambia and 8138 cases in Tanzania. Kandala et al. (2001) use this data to explore determinants of undernutrition measured through stunting, which is insufficient height for age, indicating chronic undernutrition. Stunting for a child i is determined by a Z-score

$$Z_i = \frac{AI_i - MAI}{\sigma_R},$$

where AI refers to the childs height at a certain age, MAI refers to the median of a reference population, and σ_R denotes the standard deviation of the reference population.

Kandala et al. (2001) estimate separate additive models for each country with a predictor

$$\eta = \gamma_0 + f_1(bmi) + f_2(agc) + f_{spat}(d) + \gamma' x,$$

where the mothers body mass index bmi and the age of the child agc are modeled nonparametrically with Bayesian P-splines. The expression $f_{spat}(d)$ denotes the spatial effect associated with the district d the child lives in, and is modeled as sum of a structured and an unstructured random effect for Zambia. For Tanzania they exclude the unstructured effect from the model. The fixed effects γ include categorical variables concerning the education and employment situation of the mother, the gender of the child and the characteristic of the area (urban or rural), where the child resides. For more details on the analysis we refer the reader to Kandala et al. (2001).

Here, our aim is to investigate whether the nonparametric modeling of bmi and agc is necessary by employing contour probabilities and pseudo contour probabilities. Moreover, we compare different models in terms of the DIC. In a first attempt, we use the model developed by Kandala et al. (2001) and model both continuous covariates, bmi and agc, nonparametrically by P-splines. Based on the contour probabilities for the nonparametric effects obtained from this model, we investigate a number of different model specifications, where bmi and agc are modeled nonparametrically, or parametrically with polynomials of different degrees. Following Kandala et al. (2001), spatial heterogeneity is captured by adding an unstructured and a structured random effect for Zambia and a structured random effect for Tanzania and the remaining covariates are modeled parametrically throughout our analysis. The models under consideration are:

$$
\begin{aligned}
\eta_1 &= \gamma_0 + f_1(bmi) && + f_2(agc) && + f_{spat}(d) + \gamma' x \\
\eta_2 &= \gamma_0 + f_1(bmi) && + \beta_1 agc + \beta_2 agc^2 && + f_{spat}(d) + \gamma' x \\
\eta_3 &= \gamma_0 + f_1(bmi) && + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 && + f_{spat}(d) + \gamma' x \\
\eta_4 &= \gamma_0 + && + f_2(agc) && + f_{spat}(d) + \gamma' x \\
\eta_5 &= \gamma_0 + && + \beta_1 agc + \beta_2 agc^2 && + f_{spat}(d) + \gamma' x \\
\eta_6 &= \gamma_0 + && + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 && + f_{spat}(d) + \gamma' x \\
\eta_7 &= \gamma_0 + \alpha_1 bmi && + f_2(agc) && + f_{spat}(d) + \gamma' x \\
\eta_8 &= \gamma_0 + \alpha_1 bmi && + \beta_1 agc + \beta_2 agc^2 && + f_{spat}(d) + \gamma' x \\
\eta_9 &= \gamma_0 + \alpha_1 bmi && + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 && + f_{spat}(d) + \gamma' x \\
\eta_{10} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 && + f_2(agc) && + f_{spat}(d) + \gamma' x \\
\eta_{11} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 && + \beta_1 agc + \beta_2 agc^2 && + f_{spat}(d) + \gamma' x \\
\eta_{12} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 && + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 && + f_{spat}(d) + \gamma' x
\end{aligned}
$$

Table 4.3 shows the resulting contour probabilities and pseudo contour probabilities, respectively. In Table 4.4 values for the deviance, the effective degrees of freedom (pD) and the DIC are displayed. Models are ordered according to the DIC.

The best fit in terms of the DIC is achieved by model 7. The models 1, 9 and 10 perform almost equally well, however, the deviance is considerably smaller for the higher parameterized models 1, 7 and 10. Excluding *bmi* from the model does not give a satisfying fit (though the contour probability tends to favor this). Comparing the models with a linear effect for *bmi* (models 7, 8, 9) to those with a quadratic effect (models 10, 11, 12), there are no apparent differences, indicating that linear modeling is sufficient. For *agc* a quadratic fit does not perform very well. On the other hand, the differences between cubic and nonparametric modeling of *agc* are almost negligible. This means that nonparametric modeling could be replaced by parametric modeling using a polynomial of degree 3 without appreciable loss in terms of model fit.

The findings based on the DIC are in agreement with the low p-values obtained from contour probabilities based on the median or on the log-density for difference order 3 and lower for the effect of *agc*. Pseudo contour probabilities suggest a linear effect and therefore clearly contradict the results from the DIC. The p-values based on contour probabilities in favor of 'no effect' of *bmi* are in a medium region and allow no clear decision, though the DIC clearly prefers the linear fit. Pseudo contour probabilities allow no clear decision either since the p-value for difference order $s = 0$ pleads for an effect, whereas the p-value for $s = 1$ tends towards 'no effect'.

In Figure 4.9 we compare the nonparametric, the linear and the quadratic fit for *agc*, and the nonparametric, the quadratic and the cubic effect for *bmi*. The depicted effects correspond to the best model in terms of DIC that includes the corresponding type of modeling of an effect. Here we see that the functional form of the two effects can indeed satisfactorily be modeled by a quadratic and a cubic term, respectively.

The results for Tanzania are reported in Tables 4.5 and 4.6. A comparison of parametric and nonparametric estimations is displayed in Figure 4.10. The main differences compared

to the results for Zambia are a more curved estimation of the effect of *bmi* and an additional local maximum of the effect of *agc* at in the interval $[25, 30]$. However, while a quadratic fit for *bmi* (models 10, 11, 12) only slightly improves the model in terms of DIC compared to a linear fit (models 7, 8, 9), we observe a more distinct improvement when using a nonparametric fit for *agc* (models 1, 4, 7, 10) instead of a cubic fit (models 3, 6, 9, 12). This is confirmed by the obtained contour probabilities (based on the median and the mean of the log-density), which are somewhat smaller for difference order $s = 1$ for *bmi* and clearly smaller for 4th differences for *agc*. Figure 4.10 shows that a cubic fit totally misses the local maximum exhibited by the nonparametric estimate. Pseudo contour probabilities are smaller than for Zambia, too, but still fail to give results agreeing with that obtained from contour probabilities and the DIC.

4.5 Conclusion

We applied contour probabilities and pseudo contour probabilities in order to decide whether nonparametric modeling of continuous covariates is necessary or if parametric modeling by polynomials of small degree is sufficient. In a simulation study we found contour probabilities to perform clearly superior compared to pseudo contour probabilities. In two applications we highlight that contour probabilities qualify as a helpful instrument for model selection. We conclude that contour probabilities give useful hints for a careful model selection, but seem to behave somewhat conservative regarding the possible model fit in terms of the DIC. Therefore, we recommend not to rely solely on them, but to take into account other model selection tools as for example the DIC, especially when the resulting contour probabilities are not close to 0 or 1.

Estimating Bayesian p-values for general distributions from an exponential family is computationally much more expensive since the marginal distributions are no longer available by Rao-Blackwellization. Instead an approach of Chib and Jeliazkov (2001) could be used. This might be a challenge for future research.

Acknowledgement:
This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

Table 4.3: Contour probabilities for the effects of `bmi` and `agc` in Zambia. Displayed are the results for model 1.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`bmi` (based on median)	0.29	0.38	1.0	1.0	1.0
`bmi` (based on mean of log-density)	0.30	0.42	1.0	1.0	1.0
`bmi` (pseudo contour probabilities)	0.0	0.45	1.0	1.0	1.0
`agc` (based on median)	0.0	0.0	0.0	0.09	0.84
`agc` (based on mean of log-density)	0.0	0.0	0.0	0.12	0.87
`agc` (pseudo contour probabilities)	0.0	0.0	0.83	1.0	1.0

Table 4.4: Deviance, effective degrees of freedom (pD) and DIC for models 1-12 for Zambia using IG(0.001,0.001) priors on τ^2_{bmi}, τ^2_{agc} and σ^2.

	Deviance $D(\theta)$	pD	DIC
Model 7	12640.5	46.7	12733.9
Model 10	12640	47.5	12735
Model 9	12651.5	42.7	12736.9
Model 1	12639.1	49.4	12737.9
Model 12	12650.9	43.9	12738.7
Model 3	12650.9	45.3	12741.5
Model 4	12663	46.6	12756.2
Model 6	12676.5	42.6	12761.7
Model 8	12696.3	41.6	12779.5
Model 11	12695.2	42.5	12780.2
Model 2	12694.3	44.0	12782.3
Model 5	12722	41.1	12804.2

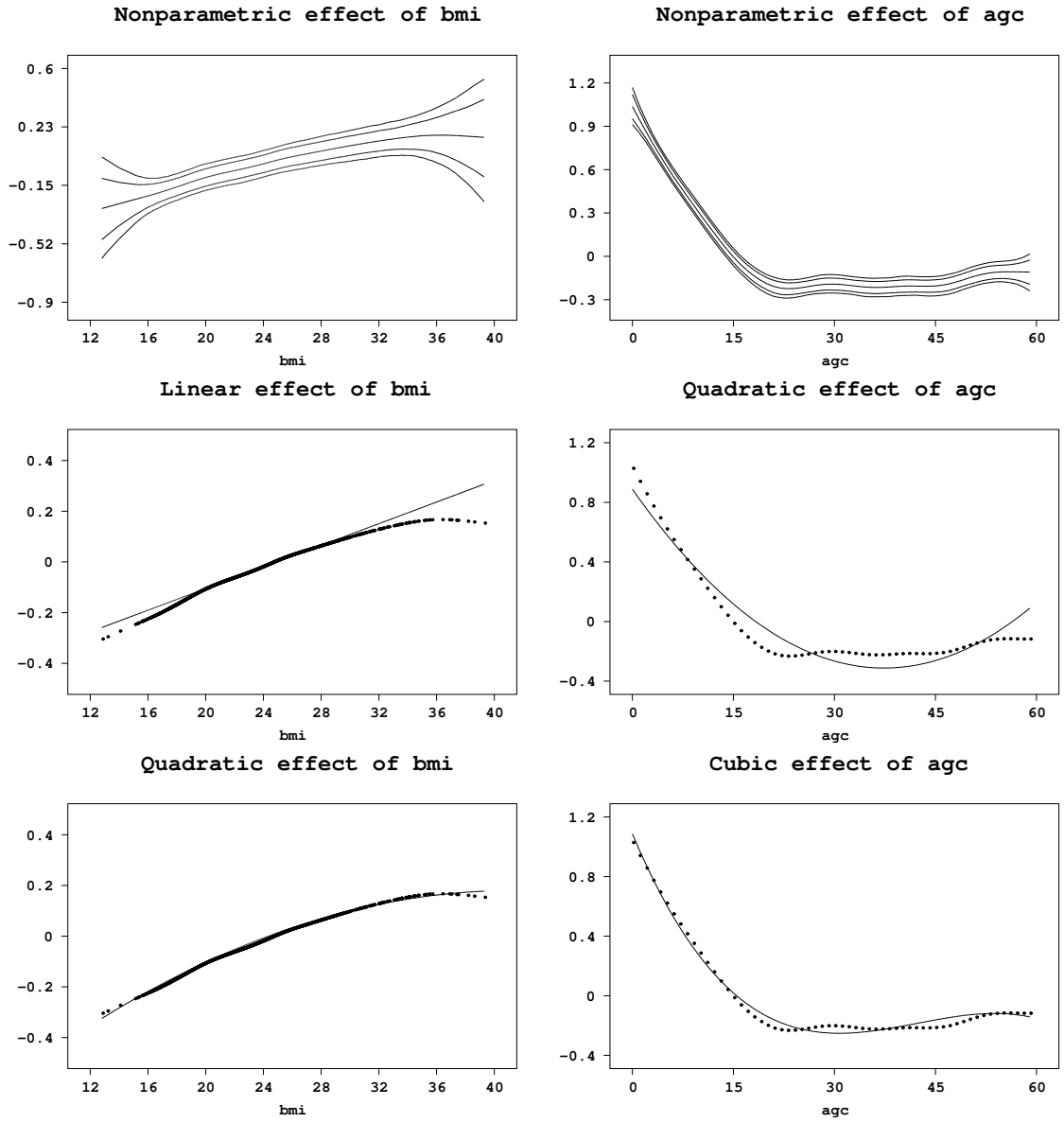


Figure 4.9: Effects of `bmi` (left panel) and `agc` (right panel) in Zambia for different model specifications. In the two lower panels the solid line corresponds to the parametric fit, the dotted curve displays the spline estimate.

Table 4.5: Contour probabilities for the effects of `bmi` and `agc` in Tanzania. Displayed are the results for model 1.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`bmi` (based on median)	0.33	0.14	0.79	0.93	0.93
`bmi` (based on mean of log-density)	0.42	0.25	0.86	0.97	0.97
`bmi` (pseudo contour probabilities)	0.02	0.04	0.89	0.84	0.97
`agc` (based on median)	0.0	0.0	0.0	0.0	0.09
`agc` (based on mean of log-density)	0.0	0.0	0.0	0.01	0.20
`agc` (pseudo contour probabilities)	0.0	0.0	0.60	0.77	0.99

Table 4.6: Deviance, effective degrees of freedom (pD) and DIC for models 1-12 for Tanzania using IG(0.001,0.001) priors on τ^2_{bmi}, τ^2_{agc} and σ^2.

	Deviance $D(\theta)$	pD	DIC
Model 1	15477.8	39.0	15555.8
Model 10	15489.3	35.0	15559.3
Model 7	15494.7	34.2	15563.1
Model 3	15521.2	32.8	15586.8
Model 12	15431.4	28.9	15589.6
Model 9	15537	27.9	15592.8
Model 4	15539.9	33.3	15606.5
Model 6	15583.4	27.0	15638.4
Model 2	15601	32.0	15665
Model 11	15612.5	27.8	15668.1
Model 8	15618.8	26.9	15672.6
Model 5	15671.8	25.9	15723.6

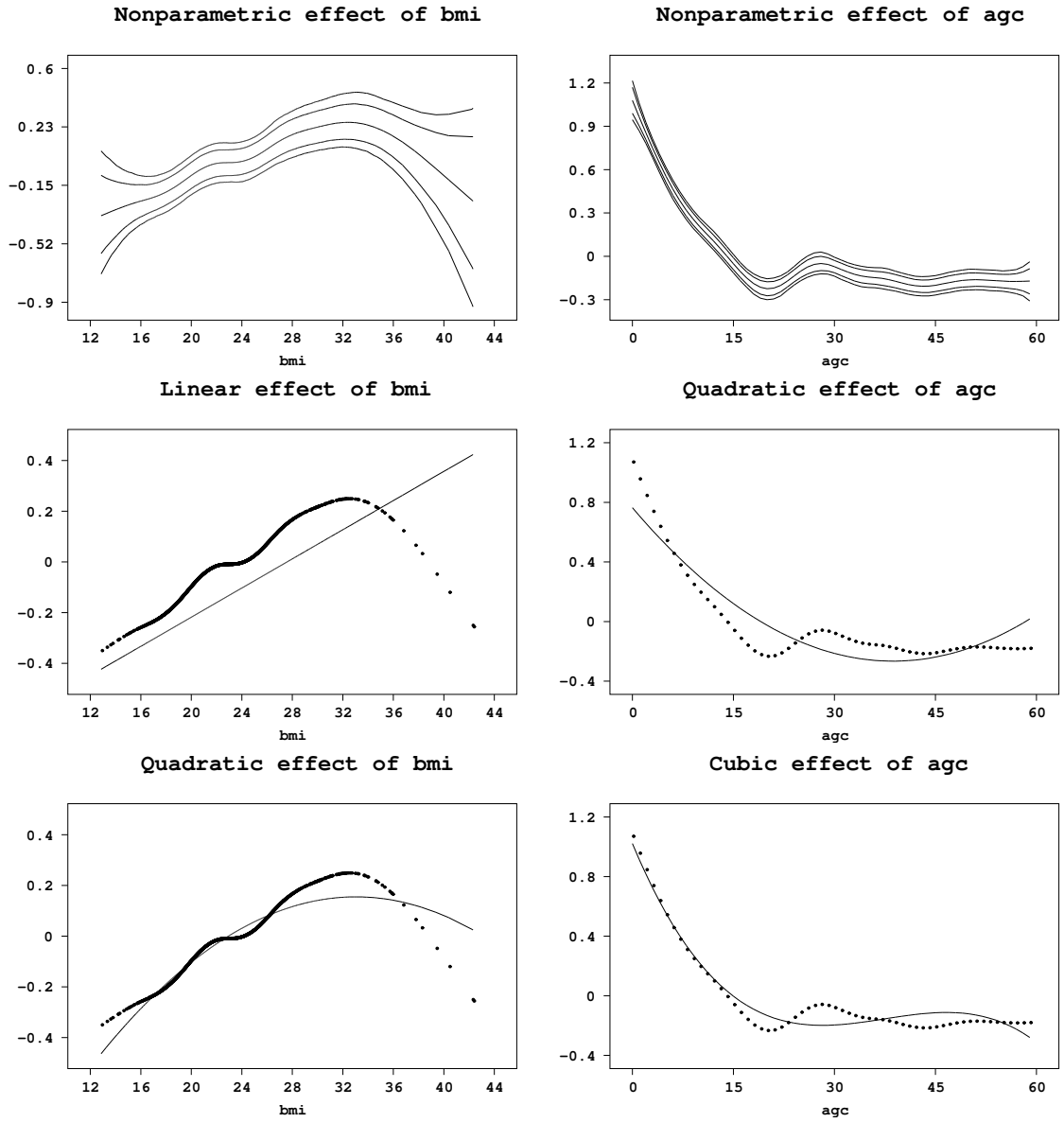


Figure 4.10: Effects of `bmi` (left panel) and `agc` (right panel) in Tanzania for different model specifications. In the two lower panels the solid line corresponds to the parametric fit, the dotted curve displays the spline estimate.

Chapter 5

BayesX: Analyzing Bayesian structured additive regression models

One of the main concerns of this work is not only to develop methodology for fitting Bayesian P-splines as a building block within a very general model class, but also to efficiently implement this methodology in an easy to use public domain software. All models discussed so far are implemented in the software *BayesX*, a program for Bayesian inference. Therefore, in this last chapter we describe the usage and the capabilities of *BayesX*, available via internet at `http://www.stat.uni-muenchen.de/~lang/bayesx/`. Although *BayesX* is not restricted to MCMC methods, we focus on such methods in this chapter, since in this work we solely rely on simulation based inference.

This chapter consists of the SFB 386 discussion paper 332 entitled 'BayesX: Analyzing Bayesian structured additive regression models' by Brezger, Lang and Kneib (2003). The article is intended to give the user an overview over the methodology implemented in *BayesX* and to give a basic introduction into practical working with the program. This is achieved by explaining the philosophy of the program and providing concrete examples consisting of executable commands for performing the analysis of a data set on undernutrition in countries of Zambia and Malawi. The data is shipped along with the software and was originally analyzed by Kandala et al. (2001). Therefore, the focus lies on performing the estimation and not on interpretation of the results. The article enables the user to do first steps with the program. Note that slight modifications regarding the notation have been made at some places to achieve consistency with the other chapters. A more detailed extensive user manual (Brezger et al. 2003) is available at `http://www.stat.uni-muenchen.de/~lang/bayesx/manual.pdf`.

BayesX: Analyzing Bayesian structured additive regression models

Andreas Brezger, Thomas Kneib and Stefan Lang
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

SUMMARY

There has been much recent interest in Bayesian inference for generalized additive and related models. The increasing popularity of Bayesian methods for these and other model classes is mainly caused by the introduction of Markov chain Monte Carlo (MCMC) simulation techniques which allow realistic modeling of complex problems. This paper describes the capabilities of the public domain software *BayesX* for estimating regression models with structured additive predictor based on MCMC inference. The program extends the capabilities of existing software for semiparametric regression like S-plus, SAS or R. Many model classes well known from the literature are special cases of the models supported by *BayesX*. Examples are generalized additive (mixed) models, dynamic models, varying coefficient models, geoadditive models, geographically weighted regression and models for space-time regression. *BayesX* supports the most common distributions for the response variable. For univariate responses these are Gaussian, Binomial, Poisson, Gamma and negative Binomial. For multicategorical responses, both multinomial logit and probit models for unordered categories of the response as well as cumulative threshold models for ordered categories can be estimated. Moreover, *BayesX* allows the estimation of complex continuous time survival and hazard rate models.

5.1 Introduction

BayesX is a public domain software package developed during the last seven years at the Department of Statistics, University of Munich. The program comprises a number

of powerful features and tools for full and empirical Bayesian inference. Functions for handling and manipulating data sets and geographical maps, and for visualizing results are added for convenient use.

In this paper, we describe a powerful tool for estimating regression models with structured additive predictor (see Section 5.2) based on recent MCMC simulation techniques. This paper may primarily serve as a starting point for getting an overview about the capabilities of this tool and as a guideline through the more detailed description in the *BayesX* manual, see Brezger et al. (2003). Besides the regression tool described in this paper, the current version of *BayesX* contains an alternative approach for inference based on mixed model methodology (Fahrmeir et al. (2004) and Ruppert et al. (2003)), and also allows for estimating *Bayesian dags* (Fronk and Giudici (2000) and Fronk (2002)).

The next section provides a brief introduction to the methodological background. In Section 5.3 we give an overview about the general usage of *BayesX* and show how Bayesian structured additive regression models are estimated. A complex example about childhood undernutrition in Zambia is discussed in Section 5.4. Instructions for downloading the program and recommendations for further reading are given in the concluding Section 5.5.

5.2 Methodological background

The model class supported by *BayesX* is based on the framework of Bayesian generalized linear models (GLM), e.g. Fahrmeir and Tutz (2001)). GLM's assume that, given covariates u and unknown parameters γ, the distribution of the response variable y belongs to an exponential family with mean $\mu = E(y \mid u, \gamma)$ linked to a linear predictor η by

$$\mu = h(\eta) \qquad \eta = u'\gamma. \tag{5.1}$$

Here h is a known response function, and γ are unknown regression parameters. *BayesX* is, however, able to estimate much more flexible models with *structured additive predictor* (see Part II of Chapter 2 and Fahrmeir et al. (2004))

$$\eta_r = f_1(\psi_{r1}) + \cdots + f_p(\psi_{rp}) + u_r'\gamma, \tag{5.2}$$

where r is a generic observation index, ψ_{rj} denote generic covariates of different types and dimension, and f_j are (not necessarily smooth) functions of the covariates. The functions f_j comprise usual nonlinear effects of continuous covariates, time trends and seasonal effects, two-dimensional surfaces, varying coefficient terms, i.i.d. random intercepts and slopes, spatially correlated effects, and geographically weighted regression. In order to demonstrate the generality of the model class supported by *BayesX* we point out some special cases of (5.2) well known from the literature:

- *Generalized additive model (GAM) for cross-sectional data*
  A GAM (Hastie and Tibshirani (1990)) is obtained if the ψ_j, $j = 1, \ldots, p$, are univariate and continuous and f_j are smooth functions. In *BayesX* the functions f_j are modeled either by random walk priors or P-splines, see Fahrmeir and Lang (2001a) and Chapter 2 for the methodological background.

- *Generalized additive mixed model (GAMM)*
  Consider longitudinal data for individuals $i = 1, \ldots, n$, observed at time points $t \in \{t_1, t_2, \ldots\}$. For notational simplicity we assume the same time points for every individual, but generalizations to individual-specific time points are obvious. A GAMM extends a GAM by introducing individual-specific random effects, i.e.

$$\eta_{it} = f_1(x_{it1}) + \cdots + f_k(x_{itk}) + b_{1i}w_{it1} + \cdots + b_{qi}w_{itq} + u'_{it}\gamma,$$

  where $\eta_{it}, x_{it1}, \ldots, x_{itk}, w_{it1}, \ldots, w_{itq}, u_{it}$ are predictor and covariate values for individual i at time t and $b_i = (b_{1i}, \ldots, b_{qi})$ is a vector of q i.i.d. random intercepts (if $w_{itj} = 1$) or random slopes. The random effects components are modeled by i.i.d. Gaussian priors, see e.g. Clayton (1996). GAMM's can be subsumed into (5.2) by defining $r = (i, t)$, $\psi_{rj} = x_{itj}$, $j = 1, \ldots, k$, $\psi_{r,k+h} = w_{ith}$, and $f_{k+h}(\psi_{r,k+h}) = b_{hi}w_{ith}$, $h = 1, \ldots, q$. Similarly, GAMM's for cluster data can be written in the general form (5.2).

- *Geoadditive models*
  In many situations additional geographic information for the observations in the data set is available. As an example compare our demonstrating example in Section 5.4 on the determinants of childhood undernutrition in Zambia. Here, the district where the mother of a child lives may be used as an indicator for regional differences in the health status of children. A reasonable predictor for such data is

$$\eta_r = f_1(x_{r1}) + \cdots + f_k(x_{rk}) + f_{spat}(s_r) + u'_r, \gamma \tag{5.3}$$

  where f_{spat} is an additional spatially correlated effect of the location s_r an observation pertains to. Models with a predictor that contains a spatial effect are also called geoadditive models, see Kammann and Wand (2003). In *BayesX*, the spatial effect may be modeled by Markov random fields (Besag et al. 1991) or two-dimensional P-splines (Chapter 2).

- *Varying coefficient model (VCM) - geographically weighted regression*
  A VCM as proposed by Hastie and Tibshirani (1993) is defined by

$$\eta_r = g_1(x_{r1})z_{r1} + \cdots + g_p(x_{rp})z_{rp},$$

  where the effect modifiers x_{rj} are continuous covariates or time scales and the interacting variables z_{rj} are either continuous or categorical. This model can be cast into (5.2) by $\psi_{rj} = (x_{rj}, z_{rj})$ and defining the special function $f_j(\psi_{rj}) = f_j(x_{rj}, z_{rj}) = g_j(x_{rj})z_{rj}$. Note that in *BayesX* the effect modifiers are not necessarily restricted to be continuous variables as in Hastie and Tibshirani (1993). E.g. the geographical location may be used as effect modifier as well, see Fahrmeir et al. (2003) for an example. VCM's with spatially varying regression coefficients are well known in the geography literature as *geographically weighted regression*, see e.g. Fotheringham et al. (2002).

- *ANOVA type interaction model*
  Suppose x_r and z_r are two continuous covariates. Then, the effect of x_r and z_r may be modeled by a predictor of the form

$$\eta_r = f_1(x_r) + f_2(z_r) + f_{1|2}(x_r, z_r) + \ldots,$$

  see e.g. Chen (1993). The functions f_1 and f_2 are the main effects of the two covariates and $f_{1|2}$ is a two-dimensional interaction surface which can be modeled e.g. by two-dimensional P-splines (Chapter 2). The interaction can be cast into the form (5.2) by defining $\psi_{r1} = x_r$, $\psi_{r2} = z_r$ and $\psi_{r3} = (x_r, z_r)$.

All regression models discussed above and arbitrary combinations can be estimated with *BayesX* in a Bayesian framework based on recent MCMC simulation techniques. The software provides a variety of different smoothness priors whose applicability depends on the type of covariate and the prior assumptions on smoothness. For continuous covariates *BayesX* supports random walk priors (Fahrmeir and Lang 2001a) and Bayesian P-splines (Part I of Chapter 2). For spatial effects a variety of Markov random field priors (Besag et al. 1991) and two-dimensional P-splines (Part II of Chapter 2) are available. Unobserved unit- or cluster specific heterogeneity may be considered by introducing random intercepts or slopes. Interactions may be modeled via varying coefficient terms or two-dimensional P-splines.

At first sight it may look strange to use one general notation for nonlinear functions of continuous covariates, i.i.d. random intercepts and slopes, and spatially correlated effects as in (5.2). However, the unified treatment of the different components in our model is justified because the priors for the different types of effects can be cast into a general form. The vector of function evaluations $f_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ of an unknown function f_j can be written as the product of a design matrix X_j and a vector of unknown parameters β_j, i.e.

$$f_j = X_j \beta_j. \tag{5.4}$$

Then, we obtain the predictor (5.2) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + U\gamma, \tag{5.5}$$

where U corresponds to the usual design matrix for fixed effects. A prior for a function f_j is now defined by specifying a suitable design matrix X_j and a prior distribution for the vector β_j of unknown parameters. The general form of the prior for β_j is

$$p(\beta_j|\tau_j^2) \propto \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{5.6}$$

where K_j is a *penalty matrix* that shrinks parameters towards zero, or penalizes too abrupt jumps between neighboring parameters. In most cases K_j will be rank deficient and therefore the prior for β_j is partially improper. Specific examples for X_j and K_j are given in Fahrmeir and Lang (2001a) and Chapter 2. The general form of the priors allows rather

general and unified estimation procedures, see particularly Part II of Chapter 2. As a side effect the implementation and description of these procedures is considerably facilitated. The variance parameter τ_j^2 in (5.6) is equivalent to the inverse smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness. Weakly informative inverse Gamma hyperprior $\tau_j^2 \sim IG(a_j, b_j)$ are assigned to τ_j^2, with $a_j = b_j = 0.001$ as a standard option.

BayesX supports the most common distributions for the response variable. Possible choices for univariate responses are Gaussian, Binomial, Poisson, Gamma and negative Binomial. For multicategorical responses, both multinomial logit and probit models for unordered categories of the response as well as cumulative threshold models for ordered categories are available. Note that models for categorical responses may also be used for estimating discrete time survival and competing risk models, see Fahrmeir and Tutz (2001), Ch. 9. The Poisson distribution allows the estimation of piecewise exponential survival models, see e.g. Ibrahim, Chen and Sinha (2001). Furthermore, extensions of continuous time Cox models have been added recently.

The goodness of fit is assessed by the deviance, deviance residuals, the deviance information criterion DIC (Spiegelhalter et al. 2002) and leverage statistics.

The methodology for univariate responses is described in full detail in Fahrmeir and Lang (2001a) and Chapter 2. Count data regression is covered in Fahrmeir and Osuna (2003). Models with multicategorical responses are dealt with in Fahrmeir and Lang (2001b) and Part II of Chapter 2. Survival models are treated in Hennerfeind et al. (2003) and Fahrmeir and Hennerfeind (2003). A thorough (and for most practical purposes sufficient) introduction into the regression models supported by the program is provided in the *BayesX* manual (Brezger et al. (2003), Ch. 7).

5.3 Usage of *BayesX*

After having started *BayesX*, a main window divided into four sub-windows appears on the screen. These are a *command window* for entering and executing code, an *output window* for displaying results, a *review window* for easy access to past commands, and an *object browser* that displays all objects currently available.

BayesX is object oriented although the concept is limited, i.e. inheritance and other concepts of object oriented languages like C++ or S-plus are not supported. For every object type a number of object-specific methods may be applied to a particular object. To estimate Bayesian regression models we need a *dataset object* to incorporate, handle and manipulate data, a *bayesreg object* to estimate semiparametric regression models, and a *graph object* to visualize estimation results. If spatial effects are to be estimated, we additionally need *map objects*. *Map objects* serve as auxiliary objects for *bayesreg objects* and are used to read the boundary information of geographical maps and to compute the neighborhood matrix and weights associated with the neighbors. The syntax for generating a new object in *BayesX* is

> *objecttype objectname*

where *objecttype* is the type of the object, e.g. `dataset`, and *objectname* is the arbitrarily chosen name of the new object. In the following subsections we give an overview about the most important methods of the object types required to estimate Bayesian structured additive regression models.

5.3.1 dataset objects

Data (in form of external ASCII files) are read into *BayesX* with the `infile` command. The general syntax is:

> *objectname*`.infile` [*varlist*] [, *options*] `using` *filename*

Here, *varlist* denotes a list of variable names separated by blanks (or tabs), and *filename* is the name (including full path) of the external ASCII file storing the data. The variable list may be omitted if the first line of the file already contains the variable names. *BayesX* assumes that the variables are stored column wise, that is one column per variable. Two options may be passed, the `missing` option to indicate missing values and the `maxobs` option for reading in large data sets. Specifying for example '`missing = M`' defines the letter '`M`' as an indicator for a missing value. The default values are a period '.' or '`NA`' (which remain valid indicators for missing values even if an additional indicator is defined). The `maxobs` option may be used to speed up the reading of large data sets. Its usage is strongly recommended if the number of observations exceeds 10000. For instance, '`maxobs=100000`' indicates that the data set has 100000 or less observations. Having read in the data, the data set may be inspected by double clicking on the respective object in the *object browser*.

Besides the `infile` command many more methods for handling and manipulating data are available, e.g. the `generate` command to create new variables, the `drop` command to drop observations and variables or the `descriptive` command to obtain summary statistics for the variables.

5.3.2 map objects

The boundary information of a geographical map containing connected regions is read into *BayesX* using the `infile` command of *map objects*. The current version supports two file formats, *boundary files* and *graph files*. A *boundary file* stores the boundaries of every region in form of closed polygons. Having read in a boundary file, *BayesX* automatically computes the neighbors and associated weights of each region. By double clicking on the respective object in the *object browser* the map may be inspected visually. A *graph file* simply stores the nodes N and edges E of a graph $G = (N, E)$, which is a convenient way of representing the neighborhood structure of a geographical map. The nodes of the graph correspond to the region codes. The neighborhood structure is represented by the edges of the graph. Weights associated with the edges may be given in a graph file as well. For the detailed structure of *boundary* and *graph files* we refer to the *BayesX* manual, Ch. 5.

Examples of boundary and graph files for different countries and regions are available at the *BayesX* homepage, see Section 5.5 for the internet address. The syntax for reading boundary or graph files is

> *objectname*.`infile` [, `weightdef=` *wd*] [`graph`] `using` *filename*

where option '`weigthdef`' specifies how the weights associated with each pair of neighbors are computed. Currently, there are three weight specifications available, '`weightdef = adjacency`', '`weightdef = centroid`' and '`weightdef = combnd`'. If '`weightdef = adjacency`' is specified, the weights for each pair of neighbors are set equal to one. Specifying '`weightdef=centroid`' results in weights inverse proportional to the distance of the centroids of neighboring regions and '`weightdef=combnd`' results in weights proportional to the length of the common boundary. If '`graph`' is specified as an additional option *BayesX* expects a *graph file* rather than a *boundary file*.

5.3.3 bayesreg objects

Bayesian regression models are estimated using the `regress` command of *bayesreg objects*. The general syntax is

> *objectname*.`regress` *model* [`weight` *weightvar*] [`if` *expression*] [, *options*] `using` *dataset*

Executing this command estimates the regression model specified in *model* using the data specified in *dataset*, where *dataset* is the name of a *dataset object* created previously. An `if` statement may be included to analyze only a part of the data and a weight variable *weightvar* to estimate weighted regression models. Options may be passed to specify the response distribution, details of the MCMC algorithm (for example the number of iterations or the thinning parameter), etc. The syntax of models is:

$$depvar = term_1 + term_2 + \cdots + term_r$$

Here, '*depvar*' specifies the dependent variable in the model and $term_1$,...,$term_r$ define the way the covariates influence the response variable. The different terms must be separated by '+' signs. In the following we give some examples. An overview about the capabilities of *BayesX* is given in Table 5.1. Table 5.2 shows how interactions between covariates are specified. More details can be found in the *BayesX* manual Ch. 8.

Suppose we want to model the effect of three covariates X1, X2 and X3 on the response variable Y. Traditionally a strictly linear predictor is assumed which can be specified in *BayesX* by:

```
Y = X1 + X2 + X3
```

Note that a constant intercept is automatically included into the models and must not be specified. If we assume possibly nonlinear effects of the continuous variables X1 and X2, for instance quadratic P-splines with second order random walk smoothness priors, we obtain:

```
Y = X1(psplinerw2,degree=2) + X2(psplinerw2,degree=2) + X3
```

The second argument in the model formula above is optional. If omitted, a cubic spline will be estimated by default. Moreover, some more optional arguments may be passed, e.g. to define the number of knots. For details we refer to the *BayesX* manual.

Suppose now that we observe an additional variable L which provides information about the geographical location an observation pertains to. A spatial effect based on a Markov random field prior is added by:

```
Y = X1(psplinerw2,degree=2) + X2(psplinerw2,degree=2) + X3 +
 L(spatial,map=m)
```

The option '`map`' specifies the *map object* that contains the boundaries of the regions and the neighborhood information required to estimate a spatial effect.

The distribution of the response is specified by adding the option '`family`' to the options list. For instance, '`family=gaussian`' defines the responses to be Gaussian. Other valid specifications are found in Table 5.3.

Table 5.1: Overview over different model terms in BayesX.

Prior/Effect	**Syntax example**	**Description**
Linear effect	X1	Linear effect of X1.
First or second order random walk	X1(rw1) X1(rw2)	Nonlinear effect of X1.
P-spline	X1(psplinerw1) X1(psplinerw2)	Nonlinear effect of X1.
Seasonal prior	X1(season,period=12)	Time varying seasonal effect of X1 with period 12.
Markov random field	X1(spatial,map=m)	Spatial effect of X1 where X1 indicates the region an observation pertains to. The boundary information and the neighborhood structure is stored in the map object 'm'.
Two-dimensional P-spline	X1(geospline,map=m)	Spatial effect of X1. Estimates a two-dimensional P-spline based on the centroids of the regions. The centroids are stored in the map object 'm'.
Random intercept	X1(random)	I.i.d. Gaussian (random) effect of the group indicator X1, e.g. X1 may be an individuum indicator when analyzing longitudinal data.
Baseline in Cox models	X1(baseline)	Nonlinear shape of the baseline effect $\lambda_0(X1)$ of a Cox model. $\log(\lambda_0(X1))$ is modeled by a P-spline with second order penalty.

5.3.4 graph objects

graph objects are used to visualize data and estimation results obtained by other objects in *BayesX*. Currently *graph objects* may be used to draw scatter plots between variables (method '`plot`'), or to draw and color geographical maps stored in *map objects* (method '`drawmap`'). We illustrate the usage of *graph objects* with method '`drawmap`' which is used to color the regions of a map according to some numerical characteristics. The syntax is:

Table 5.2: Possible interaction terms in BayesX.

Type of interaction	Syntax example	Description
Varying coefficient term	X1*X2(rw1) X1*X2(rw2) X1*X2(psplinerw1) X1*X2(psplinerw2)	Effect of X1 varies smoothly over the range of the continuous covariate X2.
Random slope	X1*X2(random)	The regression coefficient of X1 varies with respect to the unit- or cluster index variable X2.
Geographically weighted regression	X1*X2(spatial,map=m)	Effect of X1 varies geographically. Covariate X2 indicates the region an observation pertains to.
Two-dimensional surface	X1*X2(pspline2dimrw1)	Two-dimensional surface for the continuous covariates X1 and X2.

> *objectname*.`drawmap` *plotvar regionvar* [`if` *expression*] , `map=`*mapname* [*options*] `using` *dataset*

Method '`drawmap`' draws the map stored in the *map object* 'mapname' and prints the graph either on the screen or stores it as a postscript file (if option '`outfile`' is specified). The regions with regioncode 'regionvar' are colored according to the values of the variable 'plotvar'. The variables 'plotvar' and 'regionvar' are supposed to be stored in the *dataset object* 'dataset'. Several options are available for customizing the graph, e.g. for changing from grey scale to color scale or storing the map as a postscript file, see the *BayesX* manual Ch. 6. A typical graph obtained with method '`drawmap`' is given in Figure 5.2.

5.4 A complex example: childhood undernutrition in Zambia

In this example we demonstrate the usage of *BayesX* by an analysis of data on undernutrition of children in Zambia. This data set has already been analyzed in Kandala et al. (2001). Here, we apply the same model as developed in their paper. Since our focus is on demonstrating how a regression model can be specified and estimated using *BayesX* we do not discuss or interpret the estimation results.

Undernutrition among children is usually determined by assessing the anthropometric status of a child relative to a reference standard. In our example undernutrition is measured through stunting or insufficient height for age, indicating chronic undernutrition. Stunting for a child i is determined using a Z-score which is defined as

$$Z_i = \frac{AI_i - MAI}{\sigma},$$

where AI refers to the child's anthropometric indicator (height at a certain age in our

Table 5.3: Response distributions in BayesX.

Family	Link	Description
gaussian	identity	Gaussian responses. Details about MCMC inference in Part I of Chapter 2.
binomial	logit	Binomial responses. Inference is based on conditional prior or IWLS proposals, see Fahrmeir and Lang (2001a) and Part II of Chapter 2.
bernoullilogit	logit	Models with binary responses and logit link. Estimation is based on latent utility representations, see Holmes and Held (2004).
binomialprobit	probit	Models with binary responses and probit link. Estimation is based on latent utility representations, see Albert and Chib (1993).
multinomial	logit	Multinomial logit model, see Part II of Chapter 2.
multinomialprobit	probit	Multinomial probit model. Estimation is based on latent utility representations, see Fahrmeir and Lang (2001b).
cumprobit	probit	Cumulative threshold model for ordered responses with three categories. Estimation is based on latent utility representations, see Fahrmeir and Lang (2001b).
poisson	log	Poisson distribution. Inference is based on conditional prior or IWLS proposals, see Fahrmeir and Lang (2001a) and Part II of Chapter 2.
negbin	log	Negative Binomial responses. Details in Fahrmeir and Osuna (2003).
gamma	log	Gamma distribution. Inference is based on conditional prior or IWLS proposals, see Fahrmeir and Lang (2001a) and Part II of Chapter 2.
cox	–	Cox model. Details in Hennerfeind et al. (2003) and Fahrmeir and Hennerfeind (2003).

example), MAI refers to the median of the reference population and σ refers to the standard deviation of the reference population.

The main interest is on modeling the dependence of undernutrition on covariates including the age of the child, the body mass index of the child's mother, the district the child lives in and some further categorical covariates. Table 5.4 gives a description of the variables used in our model.

The data is analyzed in largely five steps: We first read in the data into *BayesX* using a *dataset object*. Since we want to estimate a spatial effect of the district in which the child lives, we need the boundaries of the districts to compute the neighborhood information of the map of Zambia. Therefore, we create a *map object* which contains the required information in the second step. A regression model is estimated in the third step followed by visualizing results. Since our analysis is based on MCMC techniques it is important to investigate the sampling paths and the autocorrelation functions of the estimated parameters in a last step.

In the following, we assume that the data set and the map of Zambia are stored in

Table 5.4: Variables in the data set on childhood undernutrition.

Variable	Description
hazstd	Standardized Z-score of stunting.
bmi	Body mass index of the mother.
agc	Age of the child.
district	District where the child lives.
rcw	Mother's employment status with categories "working" (= 1) and "not working" (= −1).
*edu*1 *edu*2	Mother's educational status with categories "complete primary but incomplete secondary" (*edu*1 = 1), "complete secondary or higher" (*edu*2 = 1) and "no education or incomplete primary" (*edu*1 = *edu*2 = −1).
tpr	Locality of the domicile with categories "urban" (= 1) and "rural" (= −1).
sex	Gender of the child with categories "male" (= 1) and "female" (= −1).

c:\data\zambia.raw and *c:\data\mapzambia.raw*, respectively.

1. Reading data set information

To read the data into *BayesX*, we create a *dataset object* and use the `infile` command of *dataset objects*:

```
> dataset d
> d.infile using c:\data\zambia.raw
```

2. Compute neighborhood information

The neighborhood information of the map of Zambia is computed and stored in *BayesX* by creating a *map object* and using the `infile` command:

```
> map m
> m.infile using c:\data\mapzambia.raw
```

Having read in the boundary information, *BayesX* automatically computes the neighborhood matrix of the map. In our example, two regions are assumed to be neighbors if they share a common boundary.

3. Regression analysis

Kandala et al. (2001) estimated a Gaussian regression model with predictor

$$\begin{aligned}\eta &= \gamma_0 + \gamma_1 rcw + \gamma_2 edu1 + \gamma_3 edu2 + \gamma_4 tpr + \gamma_5 sex + f_1(bmi) + f_2(agc) + \\ &\quad f_{str}(district) + f_{unstr}(district)\end{aligned} \tag{5.7}$$

The two continuous covariates *bmi* and *agc* are assumed to have a possibly nonlinear effect on the Z-score and are therefore modeled nonparametrically (as cubic P-splines with second order random walk prior in our example). The spatial effect of the district is split up into a spatially correlated part $f_{str}(district)$ and an uncorrelated part $f_{unstr}(district)$. The former is modeled by a Markov random field prior, where the neighborhood matrix and possible weights associated with the neighbors are obtained from the *map object* `m`. The latter is modeled by an i.i.d. Gaussian effect.

We now estimate model (5.7) using *bayesreg objects*. We create a *bayesreg object* and estimate the model using the `regress` command:

```
> bayesreg b
> b.regress hazstd = rcw + edu1 + edu2 + tpr + sex + bmi(psplinerw2)
  + agc(psplinerw2) + district(spatial,map=m) + district(random),
  family=gaussian iterations=12000 burnin=2000 step=10 predict using d
```

The options `iterations`, `burnin` and `step` define the number of iterations, the burn in period and the thinning parameter of the MCMC simulation run. Specifying `step=10` as above forces *BayesX* to store only every 10th sampled parameter which leads to a random sample of length 1000 for every parameter in our example.

If option `predict` is specified, samples of the deviance, the effective number of parameters p_D and the deviance information criterion DIC of the model are computed and stored, see Spiegelhalter et al. (2002). In addition, estimates for the additive predictor and the posterior expectations are computed for every observation.

On a 2.4 GHz personal computer estimation of the model is carried out in about 1 minute and 5 seconds.

After estimation, results for each effect are written to an external ASCII file. These files contain the posterior mean and median, the posterior 2.5%, 10%, 90% and 97.5% quantiles and the corresponding 95% and 80% posterior probabilities of the estimated effects. For example, the beginning of the file for the effect of *bmi* looks like this:

	intnr	bmi	pmean	pqu2p5	pqu10	pmed	pqu90	pqu97p5	pcat95	pcat80
1	12.8	-0.284065	-0.660801	-0.51678	-0.283909	-0.0585753	0.085998	0	-1	
2	13.15	-0.276772	-0.609989	-0.483848	-0.275156	-0.070517	0.0572406	0	-1	
3	14.01	-0.258674	-0.515628	-0.416837	-0.257793	-0.10009	-0.00289024	-1	-1	

The numbers 1 and -1 for the variables pcat95 and pcat80 indicate that the corresponding credible intervals are either strictly positive or negative. Zero indicates credible intervals containing zero.

4. Visualizing estimation results

Estimation results for nonlinear effects of *bmi* and *agc* and the spatial effect of the *district* are best summarized by visualization. *BayesX* automatically creates appropriate plots of the effects and stores the graphs as postscript files. The file names are given in the *output*

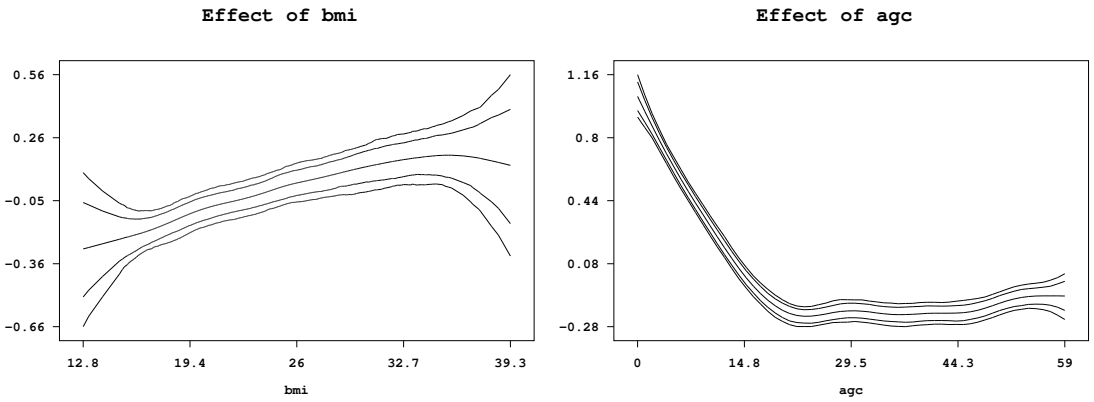


Figure 5.1: Example on childhood undernutrition: Effect of the body mass index of the child's mother and of the age of the child together with pointwise 80% and 95% credible intervals.

window for each effect. Figures 5.1 and 5.2 show the content of these files. Moreover, a batch-file is created that contains all commands necessary to reproduce the plots. The advantage is that additional options may be added by the user to customize the graphs (e.g. to change the title or axis labels).

It is also possible to visualize effects on the screen immediately after estimation. For the nonlinear effects of the two continuous covariates such plots are obtained by executing the commands

```
> b.plotnonp 1
```

and

```
> b.plotnonp 3
```

The numbers following the `plotnonp` command depend on the order in which the model terms have been specified. They are supplied in the *output window* after estimation.

Results for spatial effects are best visualized by drawing the respective map and coloring the regions of the map according to some characteristic of the posterior, e.g. the posterior mean. For instance, the structured spatial effect is visualized by typing

```
> b.drawmap 5, color
```

The additional option '`color`' forces *BayesX* to use colors instead of grey shades for visualization.

5. Post estimation commands

In addition to the `regress` command, *bayesreg objects* provide some post estimation commands to get sampled parameters or to compute autocorrelation functions of sampled parameters. For example

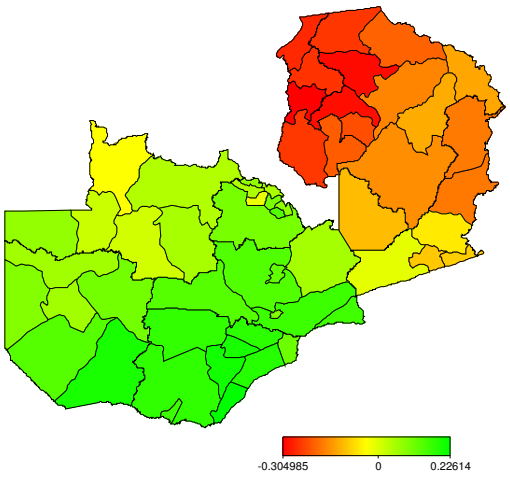


Figure 5.2: Example on childhood undernutrition: Structured spatial effect.

```
> b.getsample
```

stores sampled parameters in ASCII files and plots the sampling paths. The resulting graphs are stored in postscript format leading e.g. to the plots shown in Figure 5.3 for the scale parameter and the intercept. To avoid too large files, the samples are typically partitioned into several files.

Autocorrelation functions may be drawn e.g. by typing

```
> b.plotautocor , maxlag=150
```

where '`maxlag`' specifies the maximum lag number. The default is '`maxlag=250`'. Executing the `plotautocor` command also stores the autocorrelation functions in an ASCII file. Figure 5.4 shows the autocorrelation function for the scale parameter and the intercept.

5.5 Download and recommendations for further reading

The latest version of *BayesX* including a detailed 200 pages manual is available at

`http://www.stat.uni-muenchen.de/~lang/bayesx/`.

The *BayesX* homepage also contains all files required to reproduce the results presented in the example on childhood undernutrition in Zambia. In addition, a more detailed tutorial based on the Zambia data set is available, click on *Tutorials* at the homepage. Finally, to download the boundary and graph files for a number of countries and regions, click on *Maps*.

For users not familiar with MCMC simulation techniques, it is strongly recommended to read at least one of the introductions into MCMC. A very nice and thorough introduction is given in Green (2001). To get an overview about the methodology *BayesX* is based

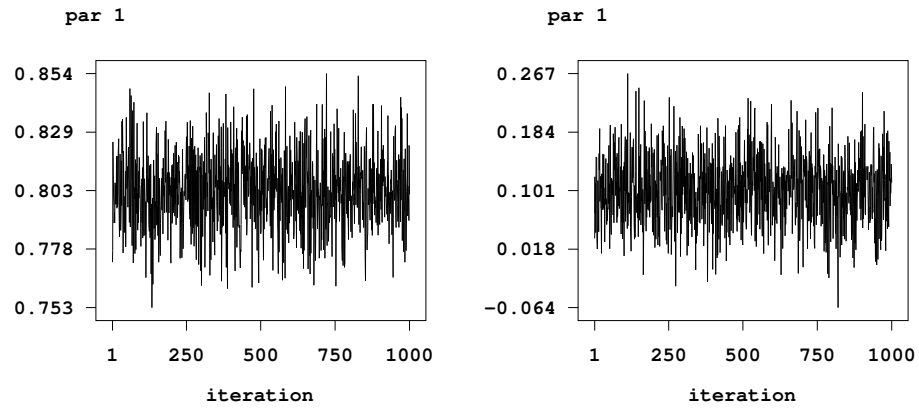


Figure 5.3: Example on childhood undernutrition: Sampling paths for the scale parameter and the intercept.

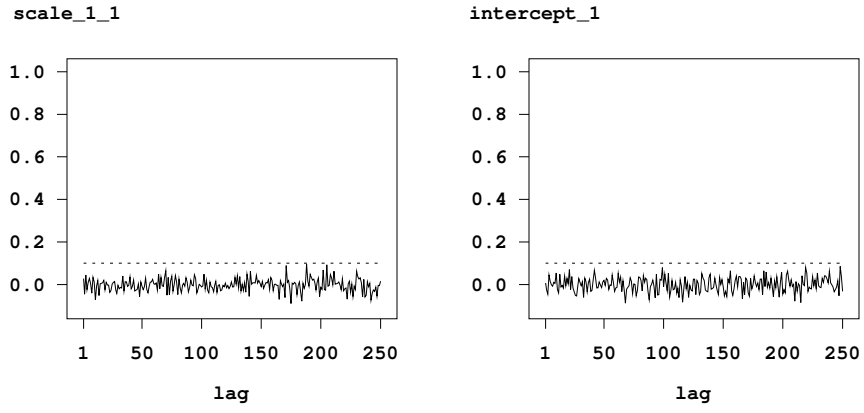


Figure 5.4: Example on childhood undernutrition: Autocorrelation functions for the scale parameter and the intercept.

on, we consider it sufficient to read Chapter 7 of the manual. More details may be found in the references cited therein and in this paper. First steps with *BayesX* can be done with the example of this paper and the tutorial on childhood undernutrition in Zambia.

Acknowledgement:

We thank Ludwig Fahrmeir, Eva-Maria Fronk and Andrea Hennerfeind for helpful comments and discussions. This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures".

Appendix A

Proofs

A.1 Proof of equation (2.14)

The penalty matrix for a 2-dimensional P-spline prior with locally adaptive variances can be written in terms of

$$K = (\tilde{D}_1' \tilde{D}_2') \Delta \begin{pmatrix} \tilde{D}_1 \\ \tilde{D}_2 \end{pmatrix}.$$

Here, $\tilde{D}_1 = I \otimes D_1$ and $\tilde{D}_2 = D_1 \otimes I$, and D_1 denotes a first order difference matrix of dimension $M \times (M-1)$. The diagonal matrix Δ contains in its i-th diagonal element the weight $\delta_{(\rho\nu)(kl)}$ that is associated with the difference formed by the i-th row of $\begin{pmatrix} \tilde{D}_1 \\ \tilde{D}_2 \end{pmatrix}$.

The spectral decomposition of K gives

$$K = \Gamma \Lambda \Gamma' = (\Gamma_1 \Gamma_2) \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \Gamma_1' \\ \Gamma_2' \end{pmatrix} = \Gamma_1 \Lambda_{11} \Gamma_1'.$$

Here, let K and Γ be partitioned as

$$K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & k_{22} \end{pmatrix}, \quad \text{and} \quad \Gamma = (\Gamma_1 \Gamma_2) = \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \gamma_{22} \end{pmatrix},$$

where k_{22} and γ_{22} are scalars and Γ_2 is a column vector. The matrix Γ_1 contains the eigenvectors of K corresponding to the non-zero eigenvalues of K and $|\Lambda_{11}| = \prod_{i=1}^{M^2-1} \lambda_i$ is the product of the non-zero eigenvalues of K. Furthermore, Γ is orthonormal and $\Gamma'\Gamma = \Gamma\Gamma' = I$, i.e. $\Gamma^{-1} = \Gamma'$.

For a proof of (2.14) the following statements are required:

(i) $|K_{11}| = |\Gamma_{11}|^2 |\Lambda_{11}|$.

Proof:

Since

$$\begin{pmatrix} K_{11} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} K \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \Gamma_1 \Lambda_{11} \Gamma_1' \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} \Gamma_{11} \\ 0 \end{pmatrix} \Lambda_{11} \left(\Gamma_{11}' 0\right)$$

it follows that

$$K_{11} = \Gamma_{11} \Lambda_{11} \Gamma_{11}'$$

and therefore

$$|K_{11}| = |\Gamma_{11} \Lambda_{11} \Gamma_{11}'| = |\Gamma_{11}||\Lambda_{11}||\Gamma_{11}'| = |\Gamma_{11}|^2 |\Lambda_{11}|.$$

(ii) $|K_{11}| > 0$.

Proof:

Since $|\Lambda_{11}| > 0$, it is sufficient to show that $|\Gamma_{11}| > 0$. Since $I = \Gamma\Gamma' = \Gamma_1\Gamma_1' + \Gamma_2\Gamma_2'$ it holds that

$$\underbrace{rg(\Gamma_1\Gamma_1' + \Gamma_2\Gamma_2')}_{=M^2} \leq rg(\Gamma_1\Gamma_1') + \underbrace{rg(\Gamma_2\Gamma_2')}_{=1}.$$

It follows that

$$M^2 - 1 \leq rg(\Gamma_1\Gamma_1') = rg(\Gamma_1) \leq M^2 - 1$$

and hence

$$rg(\Gamma_1) = M^2 - 1.$$

The rows of Γ_1 are a generating system of R^{M^2-1}. Since the column-sums of Γ_1 are zero it follows

$$-\sum_{i=1}^{M^2-1} \gamma_i = \gamma_{M^2},$$

where γ_i, $i = 1, \ldots, M^2$, denote the rows of Γ_1. Therefore the rows of Γ_{11} are a generating system of R^{M^2-1} and hence

$$rg(\Gamma_{11}) = M^2 - 1,$$

which implies that $|\Gamma_{11}| > 0$.

(iii) $|\Gamma_{11}|^2 = \gamma_{22}^2$.

Proof:

Since $|\Gamma_{11}| > 0$, it holds that

$$|\Gamma| = |\Gamma_{11}|(\gamma_{22} - \Gamma_{21}\Gamma_{11}^{-1}\Gamma_{12}) = \frac{1}{\gamma_{22}}|\Gamma_{11}|. \tag{A.1}$$

The second equality can be derived from the fact that if A is a non-singular quadratic matrix and

$$A = \begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix}, \quad \text{and} \quad A^{-1} = B = \begin{pmatrix} B_{11} & B_{21} \\ B_{12} & B_{22} \end{pmatrix}$$

then

$$A_{22}^{-1} = B_{22} - B_{21}B_{11}^{-1}B_{12},$$

where

$$B = \Gamma \quad \text{and} \quad A = \Gamma^{-1} = \Gamma' = \begin{pmatrix} \Gamma_{11}' & \Gamma_{21}' \\ \Gamma_{12}' & \gamma_{22} \end{pmatrix}.$$

From (A.1) it follows

$$|\Gamma_{11}|^2 = \gamma_{22}^2 |\Gamma|^2 = \gamma_{22}^2$$

since the orthonormality of Γ implies $|\Gamma|^2$=1.

(iv) $\gamma_{22} = 1/M$.

Proof:

The row-sums of K equal zero, i.e.

$$K(1, \ldots, 1)' = \Gamma_1 \Lambda_{11} \Gamma_1'(1, \ldots, 1)' = 0.$$

Since the columns of $\Gamma_1\Lambda_{11}$ are linearly independent, it follows that $\Gamma_1'(1, \ldots, 1)' = 0$. This means that the elements of each column of Γ_1 sum up to zero and therefore any vector with constant elements is orthogonal to all columns of Γ_1. Since Γ_2 must be orthogonal to all columns of Γ_1, i.e. $\Gamma_1'\Gamma_2 = 0$, and additionally fulfill $\Gamma_2'\Gamma_2 = 1$, it follows that $\Gamma_2 = (1/M, \ldots, 1/M)'$ and thus $\gamma_{22} = 1/M$.

Now, combining (i), (iii) and (iv) gives

$$|\Lambda_{11}| = M^2 |K_{11}|,$$

and statement (2.14) follows immediately.

A.2 Conditions for Monotonicity

To ensure that $f_j'(x) \geq 0$ or $f_j'(x) \leq 0$, it is sufficient to guarantee that subsequent parameters are ordered, such that

$$\beta_{j1} \leq \cdots \leq \beta_{jM} \quad \text{or} \quad \beta_{j1} \geq \cdots \geq \beta_{jM}, \tag{A.2}$$

respectively.

Proof: Letting the superscript $l-1$ denote basis functions of degree $l-1$, we can write $f_j'(x)$ in terms of

$$\begin{aligned} f_j'(x) &= \frac{1}{h}\sum_{\rho=1}^{M}\beta_{j\rho}\big(B_{j\rho}^{l-1}(x) - B_{j,\rho+1}^{l-1}(x)\big) \\ &= \frac{1}{h}\sum_{\rho=2}^{M}(\beta_{j\rho}-\beta_{j,\rho-1})B_{j\rho}^{l-1}(x), \end{aligned} \tag{A.3}$$

where h denotes the distance between two adjacent knots. The second equivalence in (A.3) holds, because $B_{j1}^{l-1}(x) = 0$ and $B_{j,M+1}^{l-1}(x) = 0$ for $x \in [x_{j,min}, x_{j,max}]$. Since $h > 0$ and $B_{j\rho}^{l-1}(x) \geq 0$, it follows that $f_j'(x) \geq 0$ if $\beta_{j\rho} - \beta_{j,\rho-1} \geq 0$ for all $\rho \in \{2,\ldots,M\}$. Correspondingly, from $\beta_{j\rho} - \beta_{j,\rho-1} \leq 0$ for all $\rho \in \{2,\ldots,M\}$ it follows that $f_j'(x) \leq 0$.

A.3 Proof of equation (4.11)

For a proof of (4.11) we exploit the fact that the B-spline basis functions in (4.2) for representing the spline can be computed as differences of truncated power functions (e.g. Eilers and Marx 2004), i.e.

$$B_\rho(x) = -1^{l+1}\Delta^{l+1}t(x,\rho)/(h^l l!), \quad \rho = 1,\ldots,r+l \tag{A.4}$$

where h is the distance between two neighboring knots and $t(x,\rho) := (x-(\zeta_0+\rho h))_+^l$ is the truncated power function that corresponds to the knot $\zeta_\rho = \zeta_0 + \rho h$.

Assume first that $s = 0$, which corresponds to a constant fit. Then we get

$$\frac{(h^l l!)}{-1^{l+1}}f(x) = \frac{(h^l l!)}{-1^{l+1}}\sum_{\rho=1}^{r+l}B_\rho(x)\beta_\rho = \sum_{\rho=1}^{r+l}\Delta\Delta^l t(x,\rho)\beta_\rho = \sum_{\rho=1}^{r+l}\Delta^l t(x,\rho)\beta_\rho - \sum_{\rho=1}^{r+l}\Delta^l t(x,\rho-1)\beta_\rho$$

Rearranging the two sums by combining the respective ρ-th summand of the first sum and the $(\rho+1)$-th summand of the second sum yields

$$\frac{(h^l l!)}{-1^{l+1}}f(x) = -\sum_{\rho=1}^{r+l-1}\Delta^l t(x,\rho)\Delta\beta_{\rho+1} + \Delta^l t(x,r+l)\beta_{r+l} - \Delta^l t(x,0)\beta_1. \tag{A.5}$$

Provided that $\Delta\beta_\rho = 0$, the summands in the first term are all zero. The second term in (A.5) is zero within the range $[x_{min}, x_{max}]$ of x because the polynomial part of $t(x, r+l)$ starts at x_{max}. In the third term the truncated power function $t(x,0)$ is a polynomial of degree l within the range of x. Since the l-th difference of a polynomial of degree l is a constant (compare, e.g. Schlittgen and Streitberg, p. 39f), the spline $f(x)$ reduces to a constant as claimed in (4.11).

For an arbitrary degree $s \leq l$ the proof is based on analogous arguments. Using again relationship (A.4) we get

$$\begin{aligned} \frac{(h^l l!)}{-1^{l+1}} f(x) &= \sum_{\rho=1}^{r+l} \Delta^{s+1}\Delta^{l-s} t(x,\rho)\beta_\rho \\ &= a_1 \sum_{\rho=1}^{r+l} \Delta^{l-s} t(x,\rho)\beta_\rho + \cdots + a_{s+2} \sum_{\rho=1}^{r+l} \Delta^{l-s} t(x,\rho-(s+1))\beta_\rho \end{aligned} \tag{A.6}$$

with constants $a_1, \ldots, a_{s+2}$ given by

$$a_j = (-1)^{s+j} \binom{s+1}{j-1}, \quad j = 1, \ldots, s+2.$$

Combining the ρ-th summand of the first sum, $(\rho+1)$-th summand of the second sum, to the $(\rho+s+1)$-th summand of the $(s+2)$-th sum, $\rho = 1, \ldots, r+l-s-1$, we obtain

$$\frac{(h^l l!)}{-1^{l+1}} f(x) = (-1)^{s+1} \sum_{\rho=1}^{r+l-s-1} \Delta^{l-s} t(x,\rho)\Delta^{s+1}\beta_{\rho+s+1} + R_1 + R_2 \tag{A.7}$$

with

$$\begin{aligned} R_1 &= a_1 \left(\Delta^{l-s} t(x, r+l-s)\beta_{r+l-s} + \cdots + \Delta^{l-s} t(x, r+l)\beta_{r+l}\right) \\ &\quad + \cdots + a_{s+1}\Delta^{l-s} t(x, r+l)\beta_{r+l-s} \end{aligned}$$

and

$$R_2 = a_2\Delta^{l-s} t(x,0)\beta_1 + \cdots + a_{s+2}\left(\Delta^{l-s} t(x,-s)\beta_1 + \cdots + \Delta^{l-s} t(x,0)\beta_{s+1}\right).$$

Provided that $\Delta^{s+1}\beta_\rho = 0$, the sum in (A.7) is zero. The expression R_1 is zero within the range $[x_{min}, x_{max}]$ of x. Since the $(l-s)$-th difference of a polynomial of degree l is a polynomial of degree s (compare Schlittgen and Streitberg, p. 39f) all differences of the truncated power functions appearing in R_2 are polynomials of degree $l-s$ within the range of x. Hence R_2, and therefore the spline $f(x)$, is a polynomial of degree s.

Bibliography

Abe, M. (1999), A generalized additive model for discrete–choice data, *Journal of Business & Economic Statistics*, 17, 271–284.

Albert, J. and Chib, S. (1993), Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, 88, 669–679.

Allenby, G.M. and Rossi, P.E. (1991), Quality Perceptions and Asymmetric Switching Between Brands, *Marketing Science*, 10(3), 185–204.

Andrews, D.F. and Mallows, C.L. (1974), Scale mixtures of normal distributions, *Journal of the Royal Statistical Society B*, 36, 99–102.

Baladandayuthapani, V., Mallick, B.K. and Carroll, R.J. (2005), Spatially Adaptive Bayesian Penalized Regression Splines (P-splines), *Journal of Computational and Graphical Statistics*, to appear.

Bates, D., Lindstrom, M., Wahba, G. and Yandell, B. (1987), GCVPACK – Routines for Generalized Cross-Validation, *Communication in Statistics, Part B – Simulation and Computation*, 16, 263–297.

Biller, C. (2000), Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models, *Journal of Computational and Graphical Statistics*, 9, 122–140.

Biller, C. and Fahrmeir, L. (2001), Bayesian Varying–coefficient Models using Adaptive Regression Splines, *Statistical Modeling*, 1(3), 195–211.

Bemmaor, A.C. and Mouchoux, D. (1991), Measuring the Short–Term Effect of In–Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment, *Journal of Marketing Research*, 28(2), 202–214.

Besag, J. E., Green, P. J., Higdon, D. and Mengersen, K. (1995), Bayesian computation and stochastic systems (with discussion), *Statistical Science*, 10, 3–66.

Besag, J. E., and Higdon, D. (1999), Bayesian Analysis of Agricultural Field Experiments, *Journal of the Royal Statistical Society B*, 61, 691–746.

Besag, J. and Kooperberg, C. (1995), On conditional and intrinsic autoregressions, *Biometrika*, 82, 733–746.

Besag, J., York, J. and Mollie, A. (1991), Bayesian image restoration with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Blattberg, R.C. and Neslin, S.A. (1990), Sales Promotion: Concepts, Methods, and Strategies, *Englewood Cliffs*, New Jersey.

Blattberg, R.C. and George, E.I. (1991), Shrinkage Estimation of Price and Promotional Elasticities, *Journal of the American Statistical Association*, 86(414), 304–315.

Blattberg, R.C., Briesch, R. and Fox, E.J. (1995), How Promotions Work, *Marketing Science*, 14(3)(Part 2), G122–G132.

Blattberg, R.C. and Wisniewski, K.J. (1987), How Retail Price Promotions Work, Marketing Working Paper 42, University of Chicago.

Blattberg, R.C. and Wisniewski, K.J. (1989), Price–Induced Patterns of Competition, *Marketing Science*, 8(4), 291–309.

Box, G.E.P. and Tiao, G.C. (1973): *Bayesian Inference in Statistical Analysis.* Reading, MA: Addision-Wiley. Reprint by Wiley in 1992 in the Wiley Classics Library Edition.

Breiman, L. and Friedman, J. (1985), Estimating Optimal Transformations for Multiple Regression and Correlation, *Journal of the American Statistical Association*, 80, 580–598.

Brezger, A., Kneib, T. and Lang, S. (2003), BayesX manual. Available at: `http://www.stat.uni-muenchen.de/~lang/bayesx/manual.pdf`

Brezger, A. and Lang, S. (2005), Generalized structured additive regression based on Bayesian P–splines, *Computational Statistics and Data Analysis*, to appear.

Brezger, A., Lang, S. and Kneib, T. (2003), BayesX: Analysing Bayesian Semiparametric Regression Models. SFB 386 Discussion paper 332, Department of Statistics, University of Munich.

Brezger, A. and Steiner, W.J. (2003), Monotonic regression based on Bayesian P–splines: an application to estimating price response functions from store–level scanner data. SFB 386 Discussion paper 331, Department of Statistics, University of Munich.

Carter, C. and Kohn, R. (1994), On Gibbs Sampling for State Space Models, *Biometrika*, 81, 541–553.

Chen, Z. (1993), Fitting multivariate regression functions by interaction spline models, *Journal of the Royal Statistical Society B*, 55, 473–491.

Chen, M. H. and Dey, D. K. (2000), Bayesian analysis for correlated ordinal data models. In: Dey, D. K., Ghosh, S. K. and Mallick, B. K. (2000), *Generalized linear models: A Bayesian perspective.* Marcel Dekker, New York.

Chib, S., and Jeliazkov, I. (2001), Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association*, 96, 270-281.

Clayton, D. (1996), Generalized linear mixed models. In: Gilks, W.R., Richardson, S. and Spiegelhalter, D.J.: *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

Cleveland, W. and Grosse, E. (1991), Computational methods for local regression, *Statistics and Computing*, 1991, 1, 47–62.

Currie, I. and Durban, M. (2002), Flexible smoothing with P–splines: a unified approach, *Statistical Modelling*, 4, 333–349.

De Boor, C. (1978), *A Practical Guide to Splines*, Springer, New York.

Denison, D.G.T., Mallick, B.K. and Smith, A.F.M. (1998), Automatic Bayesian curve fitting, *Journal of the Royal Statistical Society B*, 60, 333–350.

Devroye, L. (1986), *Non–uniform random variate generation.* Springer–Verlag, New York.

Dias, R. and Gamerman, D. (2002), A Bayesian approach to hybrid splines non-parametric regression, *Journal of Statistical Computation and Simulation*, 72(4), 285–297.

Diggle, P.J., Haegerty, P., Liang, K.Y. and Zeger, S.L. (2002), *Analysis of longitudinal data*, Clarendon Press, Oxford.

Di Matteo, I., Genovese, C.R. and Kass, R.E. (2001), Bayesian curve–fitting with free–knot splines, *Biometrika*, 88, 1055–1071.

Duchon, J. (1977), Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: *Construction Theory of Functions of Several Variables.* Springer, Berlin.

Dunson, D.B. and Neelon, B. (2003), Bayesian Inference on Order Constrained Parameters in Generalized Linear Models, *Biometrics*, 59(2), 286–295.

Efron, B. and Tibshirani, R.J. (1998), *An Introduction to the Bootstrap*, Chapman and Hall/CRC, Boca Raton.

Eilers, P.H.C. and Marx, B.D. (1996), Flexible smoothing using B–splines and penalized likelihood (with comments and rejoinder), *Statistical 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. Available at `http://www.stat.lsu.edu/bmarx/`.

Fahrmeir, L., Biller, C., Brezger, A., Gieger, C., Hennerfeind, A., Jerak, A. and Schmid, V. (2003), *Teil 2: Statistische Analyse der Nettomieten*, Gutachten zur Erstellung des Mietspiegels für München© 2003, Landeshauptstadt München, Sozialreferat – Amt für Wohnungswesen, *in German*.

Fahrmeir, L. and Hennerfeind, A. (2003), Nonparametric Bayesian hazard rate models based on penalized splines. SFB 386 Discussion paper 361, University of Munich.

Fahrmeir, L., Kneib, T. and Lang, S. (2004), Penalized additive regression for space-time data: a Bayesian perspective. Revised for *Statistica Sinica*.

Fahrmeir, L. and Lang, S. (2001a), Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors, *Journal of the Royal Statistical Society C (Applied Statistics)*, 50, 201–220.

Fahrmeir, L. and Lang, S. (2001b), Bayesian Semiparametric Regression Analysis of Multicategorical Time–Space Data, *Annals of the Institute of Statistical Mathematics*, 53, 10–30.

Fahrmeir, L., Lang, S., Wolff, J. and Bender, S. (2003), Semiparametric Bayesian time–space analysis of unemployment duration, *Journal of the German Statistical Society*, 87, 281–307.

Fahrmeir, L. and Osuna, L. (2003), Structured count data regression. SFB 386 Discussion paper 334, University of Munich.

Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer, New York.

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

Foekens, E.W., Leeflang, P.S.H. and Wittink, D.R. (1999), Varying Parameter Models to Accomodate Dynamic Promotion Effects, *Journal of Econometrics*, 89, 249–268.

Fotheringham, A.S., Brunsdon, C. and Charlton, M.E. (2002), *Geographically weighted regression: The analysis of spatially varying relationships.* Chichester: Wiley.

Friedman, J. H. (1991), Multivariate Adaptive Regression Splines (with discussion), *Annals of Statistics*, 19, 1–141.

Friedman, J. H. and Silverman, B. L. (1989), Flexible Parsimonious Smoothing and Additive Modeling (with discussion), *Technometrics*, 31, 3–39.

Fronk, E.M. (2002), Model Selection for dags via RJMCMC for the discrete and mixed case. SFB 386 Discussion Paper 271, Department of Statistics, University of Munich.

Fronk, E.M. and Giudici, P. (2000), Markov chain Monte Carlo model selection for dag models. SFB 386 Discussion paper 221, Department of Statistics, University of Munich.

Gamerman, D. (1997), Efficient sampling from the posterior distribution in generalized linear models, *Statistics and Computing*, 7, 57–68.

Gamerman, D., Moreira, A.R.B., and Rue, H. (2003), Space–varying regression models: specifications and simulation, *Computational Statistics and Data Analysis*, 42, 513–533.

George, A. and Liu, J.W. (1981), *Computer solution of large sparse positive definite systems.* Prentice–Hall.

Geweke, J. (1991), Efficient Simulation From the Multivariate Normal and Student–t Distribution Subject to Linear Constraints, in: *Computing Science and Statistics: Proceedings of the Twenty–Third Symposium on the Interface*, 571–578, Alexandria.

Gössl, C. (2001), *Bayesian Models in Functional Magnetic Resonance Imaging: Approaches for Human Brain Mapping*, Aachen: Shaker–Verlag.

Gössl, C., Auer, D. and Fahrmeir, L. (2000), Dynamic Models in fMRI, *Magnetic Resonace in Medicine*, 42, 72–81.

Göttlein, A. and Pruscha, H. (1996), Der Einfluß von Bestandskenngrößen, Topographie, Standord und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch, *Forstwissenschaftliches Centralblatt*, 114, 146–162.

Goldberger, A. (1968), The Interpretation and Estimation of Cobb–Douglas Functions, *Econometrica*, 35, 464–472.

Green, P.J. (2001), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82(4), 711–732.

Green, P.J. (2001), A Primer in Markov Chain Monte Carlo. In: Barndorff–Nielsen, O.E., Cox, D.R. and Klüppelberg, C. (eds.), *Complex Stochastic Systems*. Chapmann and Hall, London, 1–62.

Greene, W. (1997), *Econometric Analysis*, Prentice Hall, New Jersey.

Gupta, S. and Cooper, L. (1992), The Discounting of Discounts and Promotion Thresholds, *Journal of Consumer Research*, 19, 401–411.

Hansen, M. H. and Kooperberg, C. (2002), Spline adaptation in extended linear models, *Statistical Science*, 17, 2–51.

Hanssens, D.M., Parsons L.J. and Schultz, R.L. (2001), *Market Response Models: Econometric and Time Series Analysis*, Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1990), *Generalized additive models.* Chapman and Hall, London.

Hastie, T. and Tibshirani, R. (1993), Varying–coefficient models, *Journal of the Royal Statistical Society B*, 55, 757–796.

Hastie, T. and Tibshirani, R. (2000), Bayesian Backfitting, *Statistical Science*, 15, 193–223.

Hastie, T. and Tibshirani, R. and Friedman J.H. (2001), *The Elements of Statistical Learning*, Springer, New York.

Hastings, W.K. (1970), Monte–Carlo Sampling Methods using Markov Chains and their Applications, *Biometrika*, 57, 97–109.

Held, L. (2004), Simultaneous posterior probability statements from Monte Carlo output, *Journal of Computational and Graphical Statistics*, 13, 20–35.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2003), Geoadditive survival models. *Revised for JASA*.

Hobert, J. and Casella, G. (1996), The Effect of Improper priors on Gibbs Sampling in Hierarchical Linear Mixed Models, *Journal of the American Statistical Association*, 91, 1461–1473.

Holmes, C.C. and Heard, N.A. (2003), Generalized monotonic regression using random change points, *Statistics in Medicine*, 22, 623–638.

Holmes, C.C., and Held, L. (2004), Bayesian auxiliary variable models for binary and multinomial regression. Technical report. Available at: `http://www.stat.uni-muenchen.de/~leo`

Ibrahim, J.G., Chen, M.H. and Sinha, D. (2001), *Bayesian survival analysis.* Springer–Verlag, New York.

Jerak, A. and Lang, S. (2005), Locally adaptive function estimation for binary regression models, *Biometrical Journal*, to appear.

Kalyanam, K. and Shively, T.S. (1998), Estimating Irregular Pricing Effects: A Stochastic Spline Regression Approach, *Journal of Marketing Research*, 35(1), 16–29.

Kammann, E. E. and Wand, M. P. (2003), Geoadditive models, *Journal of the Royal Statistical Society C*, 52, 1–18.

Kandala, N. B., Lang, S., Klasen, S. and Fahrmeir, L. (2001), Semiparametric Analysis of the Socio–Demographic and Spatial Determinants of Undernutrition in Two African Countries, *Research in Official Statistics*, 1, 81–100.

Knorr–Held, L. (1996), *Hierarchical modelling of discrete longitudinal data.* Shaker Verlag.

Knorr–Held, L. (1999), Conditional prior proposals in dynamic models, *Scandinavian Journal of Statistics*, 26, 129–144.

Knorr–Held, L. and Rue, H. (2002), On block updating in Markov random field models for disease mapping, *Scandinavian Journal of Statistics*, 29, 597–614.

Kohn, R., Smith, M. and Chan, D. (2001), Nonparametric regression using linear combinations of basis functions, *Statistics and Computing*, 11, 313–322.

Kopalle, P.K., Mela, C.F. and Marsh, L. (1999), The Dynamic Effect of Discounting on Sales: Empirical Analysis and Normative Pricing Implications, *Marketing Science*, 18(3), 317–332.

Lang, S. and Brezger, A. (2004), Bayesian P–splines, *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lang, S., Fronk, E.–M. and Fahrmeir, L. (2002), Function estimation with locally adaptive dynamic models, *Computational Statistics*, 17, 479–500.

Lange, N. (1996), Statistical Approaches to Human Brain Mapping by Functional Magnetic Resonance Imaging , *Statistics in Medicine*, 15, 389–428.

Lenk, P. and DeSarbo, W.S. (2000), Bayesian inference for finite mixtures of generalized linear models with random effects, *Psychometrika*, 65, 93–119.

Leeflang, P.S.H., Wittink, D.R., Wedel, M. and Naert, P.A. (2000), *Building Models for Marketing Decisions*, Kluwer Academic Publishers, Boston.

Lin, X. and Zhang, D. (1999), Inference in generalized additive mixed models by using smoothing splines, *Journal auf the Royal Statistical Society B*, 61, 381–400.

Loader, C. (1997), Locfit: An introduction, *Statistical Computing and Graphics Newsletter*, 8(1), 11–17.

Luo, Z. and Wahba, Z. (1997), Hybrid Adaptive Splines, *Journal of the American Statistical Association*, 92, 107–116.

Mallick, B., Denison, D. and Smith, A. (1998), Semiparametric Generalized Linear Models: Bayesian Approaches, in: *Generalized Linear Models: A Bayesian Perspective*, eds. D. Dey, S. Gosh, and B.K. Mallick, Marcel–Dekker, New York.

Marx, B.D. and Eilers, P.H.C. (1998), Direct Generalized Additive Modeling with Penalized Likelihood, *Computational Statistics and Data Analysis*, 28, 193–209.

Marx, B.D. and Eilers, P.H.C. (2005), Multidimensional Penalized Signal Regression, *Technometrics*, 47 (1).

Montgomery, A.L. (1997), Creating Micro–Marketing Pricing Strategies Using Supermarket Scanner Data, *Marketing Science*, 16(4), 315–337.

Metropolis, N., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), Equations of State Calculations by Fast Computing Machines, *Journal of Chemical Physics*, 21, 1087–1091.

Mulherne, F.J. and Leone, R.P. (1991), Implicit Price Bundling of Retail Products: A Multiproduct Approach to Maximizing Store Profitability, *Journal of Marketing*, 55(4), 63–76.

Rao, V. (1993), Pricing Models in Marketing, in: Handbooks in Operations Research and Management Science, Vol. 5, Marketing, 517–552, eds. Eliashberg, J. and Lilien, G.L., Elsevier Science Publishers B.V., Amsterdam.

Reinsch, C. (1967), Smoothing by spline functions. *Numerische Mathematik*, 10, 177–183.

Rue, H. (2001), Fast sampling of Gaussian Markov random fields with applications. *Journal of the Royal Statistical Society B*, 63, 325–338.

Ruppert, D. and Carroll, R.J. (2000), Spatially Adaptive Penalties for Spline Fitting, *Australian and New Zealand Journal of Statistics*, 42, 205–223.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003), *Semiparametric Regression.* Cambridge University Press.

Robert, C.P. (1995), Simulation of truncated normal variables, *Statistics and Computing*, 5, 121–125.

Salanti, G. and Ulm, K. (2003), The analysis of dose–response relationship for binary data using monotonic regression, *American Journal of Epidemiology*, 157, 273–291.

Schlittgen, R. and Streitberg, B.H.J. (1999), Zeitreihenanalyse, *Oldenbourg*, München Wien, in German.

Sethuraman, R., Srinivasan, V. and Kim, D. (1999), Asymmetric and Neighborhood Cross–Price Effects: Some Empirical Generalizations, *Marketing Science*, 18(1), 23–41.

Shively, T.S., Kohn, R. and Wood, S. (1999), Variable Selection and Function Estimation: an Additive Nonparametric Regression Using a Data Based Prior (with discussion), *Journal of the American Statistical Society*, 94, 777–807.

Sivakumar, K. and Raj, S.P. (1997), Quality Tier Competition: How Price Change Influences Brand Choice and Category Choice, *Journal of Marketing*, 61(3), 71–84.

Smith, M. and Kohn, R. (1996), Nonparametric regression using Bayesian variable selection, *Journal of Econometrics*, 75, 317–343.

Smith, M. and Kohn, R. (1997), A Bayesian Approach to Nonparametric Bivariate Regression, *Journal of the American Statistical Association*, 92, 1522–1535.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002), Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society B*, 65, 583–639.

Stone, C.J., Hansen, M., Kooperberg, C. and Truong, Y.K. (1997), Polynomial Splines and their Tensor Products in Extended Linear Modeling (with discussion), *Annals of Statistics*, 25, 1371–1470.

Tellis, G.J. (1988), The Price–Elasticity of Selective Demand, *Journal of Marketing Research*, 25, 331–341.

Tierney, L. (1983), A space-efficient recursive Procedure for estimating a Quantile of an unknown Distribution, *SIAM J. Sci. Stat. Comput.*, 4 (4), 706-711.

Ulm, K. and Salanti, G. (2003), Estimation of general threshold limit values for dust, *Int Arch Occup Environ Health*, 76, 233–240.

van Heerde, H.J., Leeflang, P.S.H. and Wittink, D.R. (2001), Semiparametric Analysis to Estimate the Deal Effect Curve, *Journal of Marketing Research*, 38(2), 197–215.

van Heerde, H.J., Leeflang, P.S.H. and Wittink, D.R. (2002), How Promotions Work: SCAN*PRO–Based Evolutionary Model Building, *Schmalenbach Business Review*, 54, 198–220.

Wahba, G. (1978), Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression, *Journal of the Royal Statistical Society B*, 40, 364–372.

Wand, M.P. (2000), A Comparison of Regression Spline Smoothing Procedures, *Computational Statistics*, 15, 443–462.

Wand, M.P. (2003), Smoothing and mixed models, *Computational Statistics*, 18, 223–249.

Wilkinson, J.B., Mason, J.B. and Paksoy, C.H. (1982), Assessing the Impact of Short–Term Supermarket Strategy Variables, *Journal of Marketing Research*, 19(1), 72–86.

Wisniewski, K.J. and Blattberg, R.C. (1983), Response Function Estimation Using UPC Scanner Data, in: *Proceedings of ORSA–TIMS Marketing Science Conference*, 300–311, ed. F.S. Zufryden.

Wahba, G. (1990), Spline models for observational data, *CBMS-NSF Regl. Conf. Ser. appl. Math.*, 59.

Wang, Y. (1995), GRKPACK: Fitting smoothing spline ANOVA models for exponential families. Technical Report No. 942, Department of Statistics, University of Wisconsin.

Wong, C.M. and Kohn, R. (1996), A Bayesian Approach to Additive Semiparametric Regression, *Journal of Econometrics*, 74, 209–235.

Wood, S.N. (2000), Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society B*, 62, 413–428.

Wood, S.N. (2001), mgcv: GAMs and Generalized Ridge Regression for R, *R News*, 1, 20–25.

Wood, S.N. (2003), Thin plate regression splines, *Journal of the Royal Statistical Society B*, 65, 95–114.

Wood, S.N., Kohn, R., Shively, T. and Jiang, W. (2002), Model selection in spline nonparametric regression, *Journal of the Royal Statistical Society B*, 64, 119–139.

Lebenslauf

Andreas Brezger

geboren am 10. März 1975 in Heilbronn

Familienstand:	ledig
Schulbildung:	
	Grundschule in Heilbronn–Biberach
	Elly–Heuss–Knapp Gymnasium in Heilbronn (mathematisch–naturwissenschaftlich)
Zivildienst:	
1994-1995	Zivildienst in der Klinik für Gefäß– und Thorax–Chirurgie in Löwenstein
Studium:	
1995-2000	Studium der Statistik an der Ludwig–Maximilians–Universität München mit den Anwendungsgebieten Betriebswirtschaftslehre und Psychologie und der speziellen Ausrichtung mathematische Statistik
Sep. 1997	Diplom–Vorprüfung
Nov. 2000	Diplom–Hauptprüfung
Beruf:	
seit Jan. 2001	vollbeschäftigter wissenschaftlicher Mitarbeiter im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" bei Prof. Dr. L. Fahrmeir am Institut für Statistik der Ludwig–Maximilians–Universität München