Lang, Brezger:

Bayesian P-Splines

Sonderforschungsbereich 386, Paper 236 (2001)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

gsf

MAX-PLANCK-GESELLSCHAFT

TUM

Bayesian P-Splines

Stefan Lang and Andreas Brezger
University of Munich, Ludwigstr. 33, 80539 Munich
email: lang@stat.uni-muenchen.de and andib@stat.uni-muenchen.de

Abstract

P-splines are an attractive approach for modelling nonlinear smooth effects of covariates within the generalized additive and varying coefficient models framework. In this paper we propose a Bayesian version for P-splines and generalize the approach for one dimensional curves to two dimensional surface fitting for modelling interactions between metrical covariates. A Bayesian approach to P-splines has the advantage of allowing for simultaneous estimation of smooth functions and smoothing parameters. Moreover, it can easily be extended to more complex formulations, for example to mixed models with random effects for serially or spatially correlated response. Additionally, 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 where the function is highly oscillating. 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 a complex application on rents for flats in Munich.

Keywords: generalized additive models, locally adaptive smoothing parameters, MCMC, P-Splines, surface fitting, varying coefficient models

1 Introduction

Consider the *additive model* (AM) with predictor

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

where the mean of a metrical response variable y is assumed to be the sum of smooth functions f_j. The response y is assumed to follow a Gaussian distribution $y \sim N(\eta, \sigma^2)$. To allow for non-Gaussian responses the AM is extended to *generalized additive models* (GAM) (Hastie and Tibshirani, 1990) by assuming that the distribution of the response belongs to an exponential family and that the mean μ is related to the predictor by a response function $h(\eta) = \mu$. A more general class of models including the GAM as a special case are *varying coefficient models* (VCM) introduced by Hastie and Tibshirani (1993). Here, the effects of additional covariates $z = (z_1, \ldots, z_p)'$ are assumed to vary smoothly over the range of the metrical covariates x leading to the predictor

$$\eta = \gamma_0 + f_1(x_1)z_1 + \cdots + f_p(x_p)z_p.$$

For $z_j \equiv 1$, $j = 1, \ldots, p$, the VCM reduces to a GAM.

Several proposals are available for modelling and estimating the smooth functions f_j, see e.g. Hastie and Tibshirani (1990) for an overview. An attractive approach, based on *penalized regression splines* (P-splines), has been presented by Eilers and Marx (1996) for unvariate scatterplot smoothing and has been extended to GAMs by Marx and Eilers (1998). 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 of knots 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 and Silverman (1989), Friedman (1991) and Stone et al. (1997).

This paper presents a Bayesian version of the P-splines approach by Eilers and Marx for GAMs and VCMs. This is achieved by replacing difference penalties by their stochastic analogues, i.e. Gaussian (intrinsic) random walk priors which serve as smoothness priors for the unknown regression coefficients. Bayesian P-splines are a generalization of an approach for GAMs and VCMs presented by Fahrmeir and Lang (2001a, b) based on simple random walk priors. 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 through smoothness priors common in spatial statistics, e.g. two dimensional generalizations of random walks. A closely related approach (for one dimensional curve fitting) based on a Bayesian version of smoothing splines can be found in Hastie and Tibshirani (2000), see also Carter and Kohn (1994) who use 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 parameterization is possible, which is a particular advantage in a Bayesian approach where inference is based on MCMC techniques. Other Bayesian approaches for nonparametric regression are based on adaptive knot selection and are close in spirit to the work by Friedman and coauthors. Denison et al. (1998) present an approach based on reversible jump MCMC for univariate curve fitting with metrical response which is extended to GAMs by Mallick et al. (2000) and Biller (2000). A similar approach avoiding reversible jump MCMC is followed by Smith and Kohn (1996) for univariate curve fitting and by Smith and Kohn (1997) for bivariate regression.

One advantage of a Bayesian approach for P-splines, as the one we follow, is that extensions to more complex situations are comparably easy. For example, there may be situations where the curvature of an underlying function f is rapidly changing or the function is highly oscillating. In that case the assumption of a constant

smoothing parameter is inapproriate and must be replaced by a locally adaptive smoothing parameter. Such situations have attained considerable attention in the recent literature, see e.g. Fronk and Fahrmeir (1998), Luo and Whaba (1997) and Ruppert and Carroll (2000). In this paper, locally adaptive smoothing parameters are incorporated through a hierarchical t-formulation, i.e. the usual Gaussian prior for the regression parameters is replaced by a t-distribution. Such a prior has been already used in the context of dynamic models (Knorr-Held, 1996) and for edge preserving spatial smoothing (e.g. Higdon, 1994, Besag and Higdon, 1999).

The classical GAM or VCM might be inappropriate for longitudinal data if heterogeneity among units or clusters is not sufficiently described by covariates. In that case, the introduction of an additional unit- or cluster specific random effect to account for heterogeneity is necessary. Such *generalized additive mixed models* (GAMM) or *varying coefficient mixed models* (VCMM) have been considered in a Bayesian framework by Hastie and Tibshirani (2000), and Fahrmeir and Lang (2001a, b) who extend the approach to situations with spatially correlated response. We will present an application of the GAMM with spatially correlated response in our data application on rents for flats or apartments in Munich.

Bayesian inference for GAMMs and VCMMs with P-splines is based on recent MCMC simulation techniques. For Gaussian responses a Gibbs sampler can be used to update the full conditionals of the regression parameters. Drawing random numbers from the high dimensional distributions is, however, not trivial and is considerably facilitated by matrix operations for band matrices (Rue, 2000). Updating of the full conditionals in categorical probit models is facilitated by considering latent utility representations of such models. In this case, additional drawings from the latent utilities are necessary but, as an advantage, the full conditionals of the regression parameters are Gaussian allowing to use the efficient algorithms for updating regression coefficients for Gaussian responses. The updating schemes for categorical probit models in this paper are based on Albert and Chib (1993) and particularly on Fahrmeir and Lang (2001b). For general exponential family distributions we propose MH-algorithms based on conditional prior proposals (Knorr-Held, 1999 and Fahrmeir and Lang, 2001a) and on iteratively weighted least squares proposals as suggested by Gamerman (1997) for generalized linear mixed models.

Software for fitting the models in this paper is included in the program *BayesX* for Bayesian inference via MCMC. The program is available via internet under `http://www.stat.uni-muenchen.de/~lang/`.

The rest of this paper is organized as follows: Section 2 describes Bayesian GAMMs and VCMMs with P-splines. Section 3 gives details about MCMC inference for the proposed models. Section 4 contains a couple of simulation studies in order to gain more insight into the practicability and the limitations of our approach and to compare it with other approaches in the literature. In Section 5 the methods of this paper are applied to a complex dataset on rents for flats in Munich.

2 Bayesian GAMMs and VCMMs based on P-Splines

2.1 GAMs

Consider regression situations, where observations (y_i, x_i, v_i), $i = 1, \ldots, n$ on a response y, a vector of metrical covariates $x = (x_1, \ldots, x_p)'$ and a vector of further covariates $v = (v_1, \ldots, v_q)'$ are given. Generalized additive models (Hastie, Tibshirani 1990) assume that, given x_i and v_i the distribution of y_i belongs to an exponential family, i.e.

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

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

$$\mu_i = h(\eta_i), \quad \eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + v_i'\gamma. \tag{1}$$

Here h is a known response function and $f_1, \ldots, f_p$ are unknown smooth functions of the covariates. The inverse $g = h^{-1}$ of h is called the link function. 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 ensure that the posterior is not changed the substracted means are added to the intercept.

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 $\zeta_{j0} = x_{j,min} < \zeta_{j1} < \ldots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$ within the domain of x_j. It is well known that such a spline can be written in terms of a linear combination of $m = r + l$ B-spline basis functions $B_{j\rho}$, i.e.

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

For the ease of notation we assume the same number of knots m 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 (1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + V'\gamma. \tag{2}$$

Here $\beta_j = (\beta_{j1}, \ldots, \beta_{jm})'$, $j = 1, \ldots, p$, correspond to the vectors of unknown regression coefficients. The matrix V is the usual design matrix for fixed effects. In a

simple regression spline approach the unknown regression coefficients are estimated using standard maximum likelihood algorithms for generalized 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

$$L = l(y, \beta_1, \ldots, \beta_p, \gamma) - \lambda_1 \sum_{l=k+1}^{m} (\Delta^k \beta_{1l})^2 - \cdots - \lambda_p \sum_{l=k+1}^{m} (\Delta^k \beta_{pl})^2 \qquad (3)$$

is maximized with respect to the unknown regression coefficients $\beta_1, \ldots, \beta_p$ and γ. 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 critertia with respect to the λ_j, $j = 1, \ldots, p$. A major problem is, however, that the procedures for choosing the smoothness parameters often fail in practice since no optimal solutions for the λ_j can be found (see also Section 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, as considered in this paper, unknown parameters β_j, $j = 1, \ldots, p$, and γ are considered as random variables and have to be supplemented with appropriate prior distributions. We replace the difference penalties in (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

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \qquad (4)$$

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 (4) could have been equivalently defined by specifying the conditional distributions of a particular parameter $\beta_{j\rho}$ given the *left* and *right* neighbours. 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 (4) can be equivalently written in the form of a global smoothness prior

$$\beta_j | \tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2} \beta_j' K_j \beta_j\right) \qquad (5)$$

with appropriate penalty matrix K_j. For example, for a first order random walk the

penalty matrix K_j is given by

$$K_j = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \tag{6}$$

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 (5) is improper. Note also that the penalty matrices are not stochastic.

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 them 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. Throughout this paper we will set $a_j = 1$ and $b_j = 0.005$ for the hyperparameters.

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 (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 $\delta_{j\rho} \sim G(\frac{\nu}{2}, \frac{\nu}{2})$, $\nu = 1, 2, \ldots$. This implies that $\beta_{j\rho}|\beta_{j\rho-1}$ or $\beta_{j\rho}|\beta_{j\rho-1}, \beta_{j\rho-2}$ follow a t-distribution with ν degrees of freedom. We will see in our simulation study in Section 4.2 that the best results are obtained for $\nu = 1$, which corresponds to a Cauchy distribution. In principal, we could estimate ν from the data as well, see Knorr-Held (1996) for an approach. The penalty matrices are now stochastic, e.g. the corresponding penalty matrix to (6) is given by

$$K_j = \begin{pmatrix} \delta_{j2} & -\delta_{j2} & & & & & \\ -\delta_{j2} & \delta_{j2} + \delta_{j3} & -\delta_{j3} & & & & \\ & -\delta_{j3} & \delta_{j3} + \delta_{j4} & & -\delta_{j4} & & \\ & & \ddots & & \ddots & & \\ & & -\delta_{j,m-2} & \delta_{j,m-2} + \delta_{j,m-1} & & -\delta_{j,m-1} & \\ & & & & -\delta_{j,m-1} & \delta_{j,m-1} + \delta_{j,m} & -\delta_{j,m} \\ & & & & & -\delta_{j,m} & \delta_{j,m} \end{pmatrix}.$$

A similar approach for locally adaptive smoothing parameters, but with correlated $\delta_{j\rho}$, is followed by Fronk and Fahrmeir (1998).

2.2 Modelling interactions

The models considered so far are not appropriate for modelling 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 (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. Estimation of VCMs poses no further difficulties, since only the design matrices X_j in (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 metrical. In principal, interactions between metrical covariates could be modelled through 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 modelled 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) we impose the constraints

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

This is achieved in an MCMC sampling scheme by appropriately centering the functions in every iteration of the sampler. 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 substracting $\bar{f}_{js}$ from the intercept. In the last step the main effects are centered in the same way as described before.

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 neighbouring parameters and the variance component τ_{js}^2. The most simplest prior specification based on the 4 nearest neighbours can be

defined by

$$\beta_{js\rho\nu}|\cdot \sim N\left(\frac{1}{4}(\beta_{js\rho-1,\nu}+\beta_{js\rho+1,\nu}+\beta_{js\rho,\nu-1}+\beta_{js\rho,\nu+1}),\frac{\tau_{js}^2}{4}\right) \tag{7}$$

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 neighbouring 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, however, prior (7) 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 (5).

Prior (7) 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 (7) to

$$\beta_{js\rho\nu}|\cdot \sim N\left(\sum_{(kl)\in\partial_{(\rho\nu)}}\frac{\delta_{(\rho\nu)(kl)}}{\delta_{(\rho\nu)+}}\beta_{kl},\frac{\tau_{\rho\nu}^2}{\delta_{(\rho\nu)+}}\right). \tag{8}$$

Here, $\partial_{\rho\nu}$ corresponds to the set of neighouring 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 (7) 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 $\delta_{(\rho\nu)(kl)} \sim G(\frac{\nu}{2},\frac{\nu}{2})$.

2.3 Unobserved heterogeneity

Suppose now that we have for each individual i repeated observations y_{it},x_{it},v_{it}, $t = 1,\ldots,T$ over time. For simplicity, we assume that we have the same number of observations for each individual although this is of course not a necessary requirement. In such situations we often observe the problem of heterogeneity among units caused by unobserved covariates. Neglecting unobserved heterogeneity may lead to considerably biased estimates for the nonlinear and fixed effects. A common approach to overcome these difficulties is the introduction of additional random effects b_i, $i = 1,\ldots,n$, into the predictors leading to GAMMs or VCMMs. In a Bayesian framework we assume that the b_i's are i.i.d. Gaussian, i.e.

$$b_i \sim N(0,\tau_{re}^2). \tag{9}$$

Formally, the prior for the vector $b = (b_1,\ldots,b_n)'$ can be brought into the general form (5) by simply setting $K = I$. For τ_{re}^2 we assume an inverse Gamma prior $\tau_{re}^2 \sim IG(a_{re},b_{re})$.

Another interpretation of the random effects b_i is that of an unstructured unsmooth covariate effect. In fact, we could replace some of the smoothness priors for the functions f_j in GAMMs or VCMMs by unstructured random effects if the assumption of a smooth effect is not justified. For a particular covariate x, we simply have to define parameters b_x for every observed covariate value and assign the prior (9) to them. In VCMMs the b_x's can then be interpreted as random slope parameters. In some situations it may be even necessary to split up the effect of a particular covariate x into a smooth and an unsmooth component. Such models have been used in the context of spatial smoothing by Besag et al. (1991) and Fahrmeir and Lang (2001b).

There may be also situations with spatially correlated responses. For example, in our application on rents for flats in Munich the monthly rent for a flat or apartment considerably depends on the location in the city. In this application, spatially correlated random effects based on Markov random field priors as decribed in detail in Fahrmeir and Lang (2001a, b) are incorporated into the predictor. Additionally, an unstructured random effect (9) is included.

2.4 Additional prior assumptions

We conclude this section with some additional prior assumptions. 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.

i) In the case of Gaussian responses an additional (overall) variance parameter σ^2 for the errors must be taken into consideration. In analogy to the variance parameters τ_j^2 we assign an inverse Gamma prior, i.e. $\sigma^2 \sim IG(a_\sigma, b_\sigma)$.

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

iii) For given covariates and parameters observations y_i (or y_{it}) are conditionally independent.

iv) Priors $p(\beta_j|\tau_j^2)$, or $p(\beta_j|\delta_j, \tau_j^2)$ in the case of locally adaptive variances, are conditionally independent. Here, δ_j denotes the vector of weights.

v) Priors for fixed and random effects, hyperpriors $\tau_j^2, j = 1, \ldots, p$, τ_{re}^2 and δ_j, are mutually independent.

3 Posterior inference via MCMC

Bayesian inference is based on the posterior of the model, which is in all cases analytically intractable. Therefore, inference is carried out by recent Markov Chain Monte Carlo (MCMC) simulation techniques. We first consider the case of Gaussian responses:

3.1 Gaussian responses

For the following let α denote the vector of all parameters appearing in the model. According to our prior assumptions the posterior is given by

$$\begin{aligned} p(\alpha) \quad \propto \quad & L(y, \beta_1, \ldots, \beta_p, \gamma, b, \sigma^2) \\ & \prod_{j=1}^{p} \Big(p(\beta_j|\tau_j^2) p(\tau_j^2) \Big) \\ & \prod_{i=1}^{n} p(b_i|\tau_{re}^2) p(\tau_{re}^2) p(\gamma) p(\sigma^2) \end{aligned} \tag{10}$$

where $L(\cdot)$ denotes the likelihood. By assumption ii) the likelihood is the product of individual likelihood contributions. If for one of the smooth functions f_j a locally adaptive variance parameter is assumed, the term $p(\beta_j|\tau_j^2)p(\tau_j^2)$ in the second line of (10) must be replaced by $p(\beta_j|\delta_j, \tau_j^2)p(\delta_j)p(\tau_j^2)$. 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 rest and the data. For Gaussian responses it can be shown that the full conditionals for β_j, $j = 1, \ldots, p$, b and γ are multivariate Gaussian. Straightforward calculations show that the precision matrix P_j and the mean m_j of $\beta_j|\cdot$ is given by

$$P_j = \frac{1}{\sigma^2} X_j' X_j + \frac{1}{\tau_j^2} K_j, \quad m_j = P_j^{-1} \frac{1}{\sigma^2} X_j'(y - \tilde{\eta}), \tag{11}$$

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 bandmatrices. 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 tensor product two dimensional P-spline the bandwidth is $m \cdot l + l$. Following Rue (2000) drawing random numbers from $p(\beta_j|\cdot)$ is as follows:

(i) Compute the Cholesky decomposition $P_j = LL'$.

(ii) Solve $L'\beta_j = z$, where z is a vector of independent standard Gaussians. It follows that $\beta_j \sim N(0, P_j^{-1})$.

(iii) 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 by forward substitution $L\nu = \frac{1}{\sigma^2} X_j'(y - \tilde{\eta})$ followed by backward substitution $L'm_j = \nu$.

(iv) Set $\beta_j = \beta_j + m_j$, then $\beta_j \sim N(m_j, P_j^{-1})$.

The algorithms involved take advantage of the bandmatrix structure of the posterior precision P_j. A detailed description of the bandmatrix operations used in this paper can be found in George and Liu (1981).

The precision matrix and the mean of the full conditional for the random effects b can be formally brought into the form in (11) where τ_j^2 is replaced by τ_{re}^2 and K_j by

the identity matrix I. For the fixed effects parameters γ we obtain for the precision matrix and the mean

$$P_\gamma = \frac{1}{\sigma^2} V'V, \quad m_\gamma = (V'V)^{-1}V'(y - \tilde{\eta}). \tag{12}$$

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

$$a_j' = a_j + \frac{rank(K_j)}{2} \quad \text{and} \quad b_j' = b_j + \frac{1}{2}\beta_j' K_j \beta_j \tag{13}$$

for τ_j^2 and τ_{re}^2. For σ^2 we obtain

$$a_\sigma' = a_\sigma + \frac{n}{2} \quad \text{and} \quad b_\sigma' = b + \frac{1}{2}\epsilon'\epsilon \tag{14}$$

where ϵ is the usual vector of residuals. In the case of replicated observations for each individual, n in (14) must be replaced by $n \cdot T$. 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

$$a_{\delta_{j\rho}}' = \frac{\nu}{2} + \frac{1}{2} \quad \text{and} \quad b_{\delta_{j\rho}}' = \frac{\nu}{2} + \frac{u_{j\rho}^2}{2\tau_j^2} \tag{15}$$

where $u_{j\rho}$ is the error term in (4). In the case of a two dimensional P-spline the full conditionals for the weights $\delta_{(\rho\nu)(kl)}$ are Gamma distributed with

$$a_{\delta_{(\rho\nu)(kl)}}' = \frac{\nu}{2} + \frac{1}{2} \quad \text{and} \quad b_{\delta_{(\rho\nu)(kl)}}' = \frac{\nu}{2} + \frac{(\beta_{\rho\nu} - \beta_{kl})^2}{2\tau_{\rho\nu}^2}. \tag{16}$$

Since all full conditionals involved are known distributions, a simple Gibbs sampler can be used to successively update the parameters of the model. The resulting sampling scheme can be summarized as follows:

i) Update β_j $(j = 1, \ldots, p)$ and b by drawing random numbers from the multivariate Gaussian distribution with precision matrix and mean given by (11).

ii) Update γ by drawing random numbers from a Gaussian distribution with mean and precision matrix in (12).

iii) Update τ_j^2 $(j = 1, \ldots, p)$ and τ_{re}^2 by drawing random numbers from inverse Gamma full conditionals with parameters given by (13).

iv) Update the weights δ_j (if for f_j locally adaptive variances are assumed) by drawing random numbers from the Gamma distributions with parameters in (15) for one dimensional P-splines and (16) for two dimensional P-splines. Recompute the penalty matrix K_j.

v) Update σ^2 by drawing random numbers from the inverse Gamma distribution specified in (14).

3.2 Categorical probit models

MCMC inference for a probit model can be considerably facilitated by equivalently rewriting the model in terms of latent utilities U_i (or U_{it}). We illustrate the concept for cross sectional binary data, i.e. y_i takes only the values 0 or 1. 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 $U_i \sim N(\eta_i, 1)$ we define $y_i = 1$ if $U_i > 0$ 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 additional parameters U_i. Thus, an additional sampling step for updating the U_i's is 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). The advantage of defining a probit model through the latent variables U_i is that the full conditionals for β_j, b and γ are Gaussian with only slightly modified precison matrices and means in (11) and (12). Since the variance of U_i is one, we have to fix σ^2 in the Gibbs sampler and replace it in (11) and (12) by 1. Additionally, y must be replaced by the current vector of latent variables U. Thus the sampling scheme for Gaussian responses can used with slight modifications including an additional sampling step for updating the U_i's. More details on Gibbs sampling for binary probit models can be found in Albert and Chib (1993).

The concept of MCMC sampling through data augemtation as illustrated for binary data can be extended to multicategorical probit models. Detailed updating schemes for the cumulative threshold model and the multinomial probit model with independent errors can be found in Fahrmeir and Lang (2001b). For these models an implementation of Bayesian P-splines is already available in *BayesX*. A further extension to multinomial probit models with correlated errors can be found in Chen and Dey (2000). Their approach is, however, restricted to purely parametric predictors.

3.3 General distributions from an exponential family

For general distributions from an exponential family Fahrmeir and Lang (2001a) propose an MH-algorithm for updating unknown regression parameters based on *conditional prior proposals*. As an advantage this approach is distribution free in the sense that for updating of parameters only the likelihood is required but no approximations of characteristics of the posterior (e.g. the mode). For one dimensional P-splines conditional prior proposals work rather satisfying. For two dimensional P-splines in some cases the mixing of Markov chains is relatively slow, thus making rather long MCMC runs necessary.

In the following we focus on *iteratively weighted least squares* (IWLS) proposals introduced by Gamerman (1997) in the context of generalized linear mixed models. Similar proposals have been made by Hastie and Tibshirani (2000) and Rue (2000). In generalized linear models parameter estimates are obtained by Fisher scoring or

in other words IWLS (see Fahrmeir and Tutz, 2001). In order to simulate from the posterior Gamerman suggests to combine IWLS and MH-updating of parameters. Suppose we want to update the regression coefficients β_j of the smooth function f_j with current value β_j^c of the chain. 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

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

Here, $W(\beta_j^c) = diag(w_1, \ldots, w_n)$ is the usual weight matrix for IWLS with weights $w_i^{-1} = b''(\theta_i)\{g'(\mu_i)\}^2$. The transformed observations $\tilde{y}_i$ are defined by $\tilde{y}_i = \eta_i + (y_i - \mu_i)g'(\mu_i)$. For simplicity we assumed that the scale parameter ϕ is known. Thus, we may use again the updating scheme for Gaussian responses with slight modifications. Steps iii) and iv) remain unchanged, and step v) is completely omitted. In steps i) and ii), however, the proposed new values are not accepted in any case as for Gaussian responses but with an acceptance propability α. The acceptance probability is a ratio of the full conditional and the proposal density at the current state β_j^c and the proposed state β_j^p. More specifically, we obtain

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

In some situations, particulary if β_j is high dimensional, the acceptance probabilities may be too small to guarantee satisfying mixing properties. Then the parameter vector β_j must be divided into smaller blocks, see Fahrmeir and Lang (2001a) for a possible blocking strategy. Random and fixed effects parameters b and γ can be updated in an analogous way.

4 Simulations

In this section we present a couple of simulation studies mainly to compare the proposed methodology with related approaches in the literature. Section 4.1 compares the Bayesian approach for P-splines with the frequentist version by Eilers and Marx (1996). In Section 4.2 we compare our approach for estimating highly oscillating functions with a variety of other recent approaches, particulary with Ruppert and Carroll (2000). Finally, Section 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.

4.1 Comparison of the Bayesian and the classical approach

In order to compare our Bayesian approach with the frequentist version by Eilers and Marx (1996), we considered the three functions $f_1(x) = 0.5x$, $f_2(x) = x^2/3 - 1.5$ and $f_3(x) = sin(x)$, i.e. a linear, a quadratic and a sinusoidal one. The values of x were chosen on an equidistant grid of $k = 100$ knots between -3 and 3. We used the sample sizes $n = 100$, $n = 500$ and $n = 1000$ and simulated 250 replications

for every model and sample size. For estimation we applied cubic P-splines with 20 and 40 knots. For Bayesian P-splines we used second order random walk penalties and for the frequentist version the corresponding second order difference penalties. For the frequentist versions of P-splines estimation was carried out with 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}. The performance of the estimators was measured by the empirical mean squared error given by $MSE(\hat{f}) = 1/k \sum_{i=1}^{k} (f(x_i) - \hat{f}(x_i))^2$

The comparison is restricted to Gaussian responses (with $\sigma^2 = 1$). We also indended to compare both approaches for binary probit models but we had too many problems with S-plus. It was almost impossible to run simulations automatically because the GAM routine of S-plus occassionally crashed for small smoothing parameters due to numerical problems (approximately every 50th estimation). We also made the strange experience that estimation required more and more computing time the longer S-plus was running.

In general, the differences between the two approaches are relatively small. For sample sizes of $n = 500$ and $n = 1000$ both estimators are more or less unbiased, i.e. function estimates averaged over the 250 replications are close to the true values. Also the obtained MSE's are almost identical. We therefore focus primarily on results for sample size $n = 100$.

Figure 1 displays boxplots of log(MSE) for the various estimators. Panel a) refers to f_1, panel b) to f_2 and panel c) to f_3. Both estimators perform more or less equally well with a slightly better performance of the frequentist versions for the linear function f_1. Our fully Bayesian approach performs slightly better for the sinusoidal function f_2. However, inspection of the individual estimates showes some strange results. Approximately 3-5% of the frequentist estimates are quite unsmooth because the cross validation score function has no global minimum or because a too small smoothing parameter was found as the optimum. For the Bayesian version we never observed these problems. As an example, compare Figure 1 d) which shows for f_3 the classical (dashed line) and the Bayesian estimate (solid line) for a particular replication. For comparison the true function is additionally included. We should stress that this is not the most severe example that we found. For sample sizes $n = 500$ and $n = 1000$, however, the problem disappeares.

For Bayesian P-splines we also investigated the coverage of pointwise credible intervals. In a Bayesian approach based on 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 varies between 80 and 86% for all models and sample sizes. This result holds also for binary probit models for which the same simulation study was carried out. This implies that the credible intervals obtained by the fully Bayesian approach are rather conservative. Similar findings have been obtained in a simulation study on generalized additive mixed models, see Lang and Fahrmeir (2001).

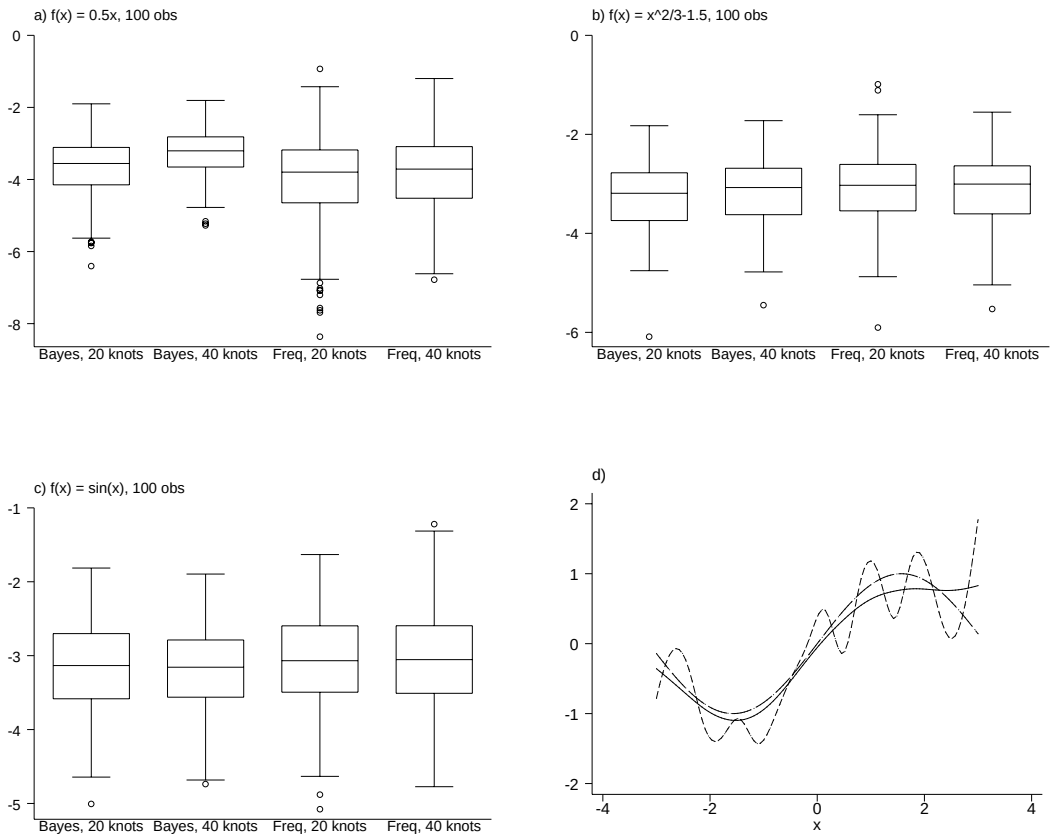


Figure 1: *Boxplots of log(MSE) for the various estimators. Panel a) refers to the linear function f_1, panel b) refers to the quadratic function f_2 and panel c) to the sinusoidal function f_3. Panel d) displays the classical (dashed line) and the Bayes estimator (solid line) for f_3 for a particular replication where cross validation failed. For comparison the true function is included.*

4.2 Locally adaptive smoothing parameters

In order to compare our approach for estimating 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 cubic P-splines with a second order random walk penalty for estimation. We used both 40 and 80 knots. Besides the estimator with a global variance we applied three estimators with locally adaptive variance which differ in the degrees of freedom ν of the hierarchical t-formulation. We set $\nu = 1$, $\nu = 2$ and $\nu = 4$.

In order to compare results we computed squared errors on a log_{10} scale as Ruppert and Carroll did. Figure 2 displays boxplots of $log_{10}(MSE)$. Additionally, Figure 3

shows, for the case of 40 knots, function estimates averaged over the 250 replications (dashed lines) together with the true functions (solid lines). From Figures 2 and 3 we can draw the following conclusions:

- For $j = 3$, i.e. low spatial variability, the estimators with global and locally adaptive variance perform more or less equally well. Hence, there is no loss of statistical efficiency when a locally adaptive estimator is used but not needed.
- For $j = 6$, i.e. severe spatial variability, the estimators with locally adaptive variance clearly outperform the estimator with global variance. The best results are obtained for $\nu = 1$. This is, however, not surprising because for growing ν the estimators approach more and more the estimator with global variance.
- The difference between using 40 and 80 knots is more or less negligible.

Similar findings have been reported by Ruppert and Carroll. For $j = 3$, they obtained values of approximately -1.5 for the median of log10(MSE). 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 $log_{10}(MSE)$. They claim that their estimator performs slightly better than the Bayesian method of Smith and Kohn (1996) and the stepwise selection method of Stone et al. (1997). To our own surprise the Bayesian P-splines approach seems to outperform Ruppert and Carrolls results by far. Even the estimators with a global variance perform better than their local penalty approach. An explanation might be that the local B-spline basis used in this paper is more suitable for the highly oscillating curve under study than the truncated power series basis they used.

In their second simulation example Ruppert and Carroll used the function

$$f_5(x) = \sin(2(4x - 2)) + 2\exp(-16^2(x - 0.5)^2).$$

This function has also been considered by Luo and Wahba (1997), who compared their h-splines approach with smoothing splines, SureShrink of Donoho and Johnstone (1995) and MARS of Friedman (1991). They all used the same value of $\sigma = 0.3$ and sample size $n = 256$ and used equally spaced x's on $[0, 1]$. We simulated 250 replications from this model. For estimation we applied cubic P-splines with 40 knots with a global variance and with locally adaptive variance ($\nu = 1$). Both Ruppert and Carroll (2000) and Luo and Wahba (1997) compared their estimators by the median and the interquartile range of squared errors. The best results were obtained with the local penalty approach by Ruppert and Carroll with a median squared error of 0.0053. For the global penalty estimator they obtained a value of 0.0061 for the median squared error which is still smaller or at least equal to the values reported by Luo and Wahba (1997) for the estimators they compared. Our results for the median squared errors are comparable to Ruppert and Carroll. We obtained a value of 0.0062 for P-splines with global variance and a value of 0.0052 for P-splines with locally adaptive variance. The interquartile range of our estimators is, however, slightly smaller. We obtained values of 0.0027 for estimators with

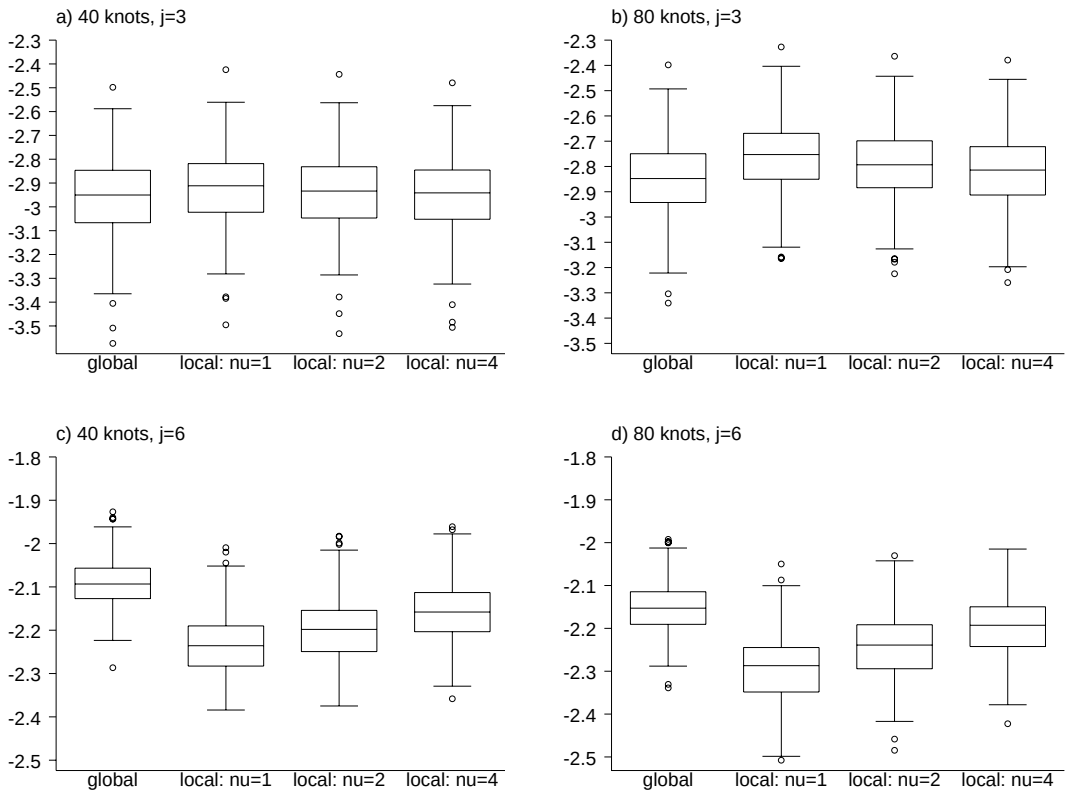


Figure 2: *Boxplots of* $log_{10}(MSE)$ *for function* f_4 *with* $j = 3$ *(panels a) and b)) and with* $j = 6$ *(panels c) and d)).*

global variance and 0.0023 for estimators with locally adaptive variance. Ruppert and Carroll reported values of 0.0029 for the global penalty estimator and 0.0035 for the local penalty estimator. Surprisingly, the interquartile range of their local penalty estimator has a higher interquartile range than the global penalty estimator.

For both simulation examples we also computed the coverage of pointwise credible intervals. Figure 4 compares the coverage for function f_4 ($j = 6$ only) and function f_5 for the estimators with global and locally adaptive variance ($\nu = 1$). In areas with strong spatial variability of the functions, the coverage of the estimator with locally adaptive variance is closer to the nominal level than of the estimator with global variance. In areas with low spatial variability for both estimators the coverage is close to the nominal level. This is another demonstration of the superiority of the estimator with locally adaptive variance.

4.3 Surface fitting

In our last simulation study we compare our approach for surface fitting with related approaches 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 in the literature. In their simulation study 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, tensor product cubic smoothing splines with five smoothing parameters, Breiman and Friedman's (1985)

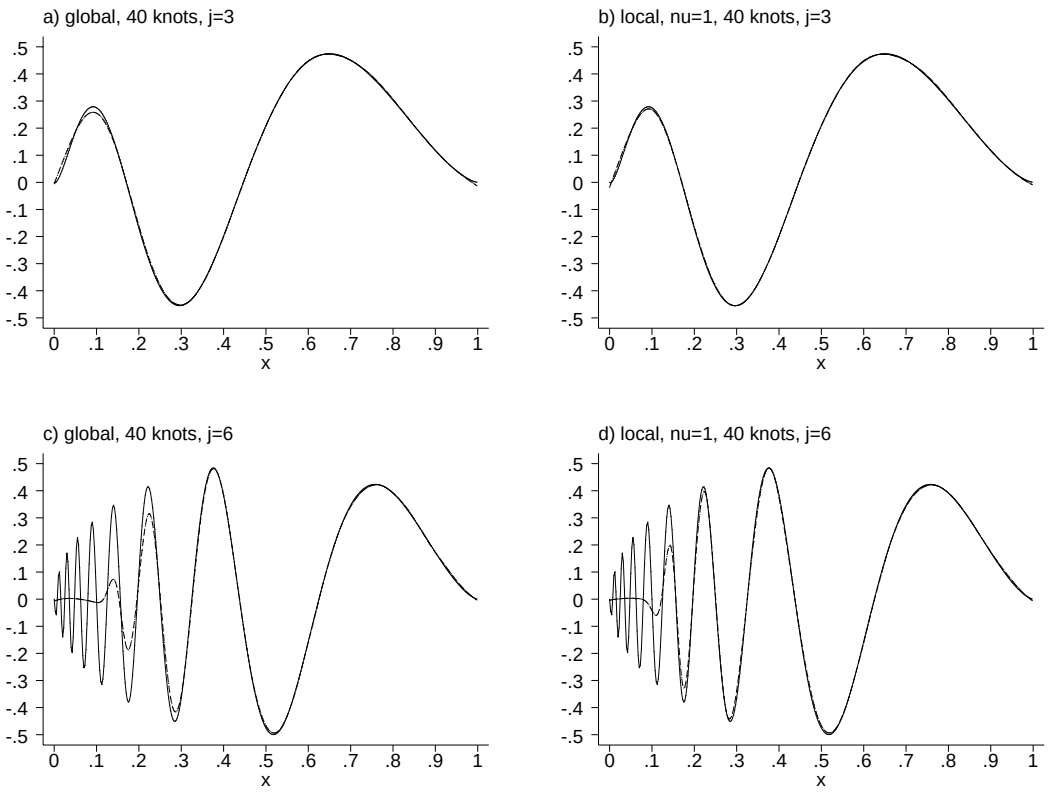


Figure 3: *Average estimates for function f_4 (dashed lines) with $j = 3$ (panels a) and b)) and with $j = 6$ (panels c) and d)). For comparison the true curves are added to the plots (solid lines).*

additive basis fitting routine and a parametric linear interaction model. For thin plate splines and tensor product smoothing splines the smoothing parameters were estimated by CGV as in Gu et al. (1989). They used the following three examples:

- $f_6(x_1, x_2) = 1/5 \exp(-8x_1^2) + 3/5 \exp(-8x_2^2)$ where x_1 and x_2 are distributed independently normal with mean 0.5 and variance 0.1.
- $f_7(x_1, x_2) = x_1 \sin(4\pi x_2)$ where x_1 and x_2 are distributed independently uniform on $[0, 1]$.
- $f_8(x_1, x_2) = x_1 x_2$ where x_1 and x_2 are bivariate normal with mean 0.5, variance 0.05 and correlation of 0.5.

Function f_6 represents a model with main effects only, and functions f_7 and f_8 correspond to a model with interactions. The sample size was $n = 300$ observations and $\sigma = 1/4\, range(f_j)$. We simulated 250 replications from the three models.

For estimation we considered both a simple bivariate surface estimator without main effects and a model with main effects and interactions. For the surface estimator we applied cubic tensor product P-splines on a 12 by 12 knots grid and the smoothness prior (7). For the main effects (if included) we used cubic P-splines with 20 knots and a second order random walk penalty with global variance. Additionally, we considered P-splines with locally adaptive variances, i.e the priors (7) and (4) were replaced by their locally adaptive variants.

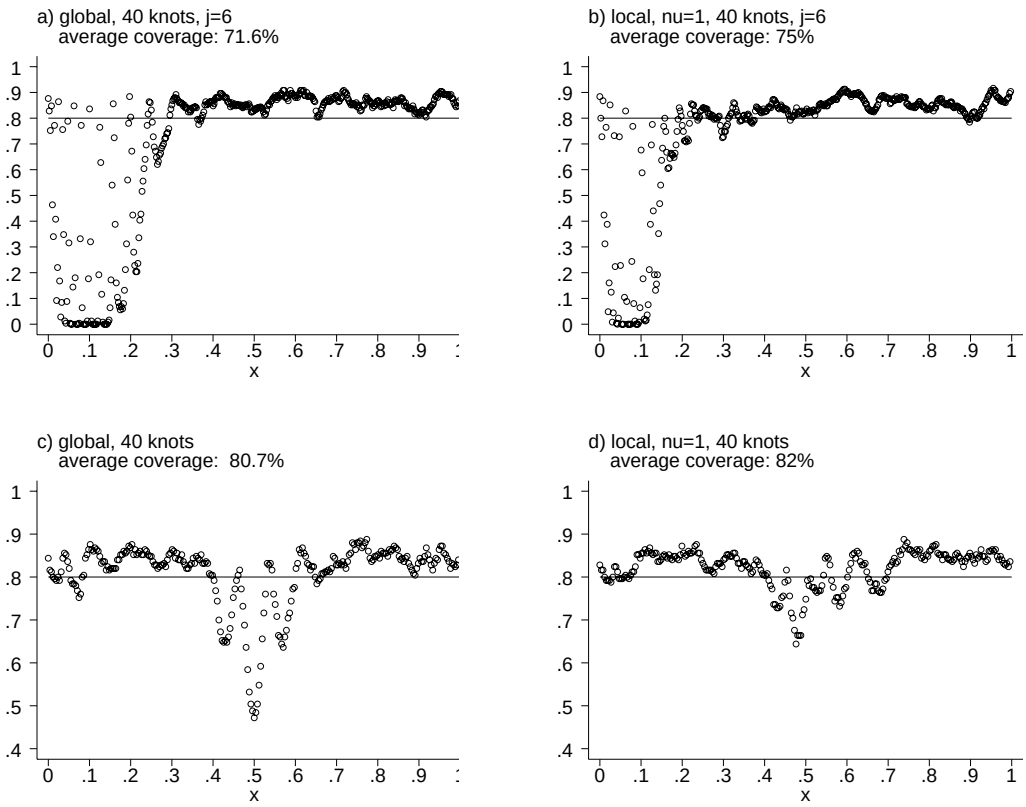


Figure 4: *Coverage of pointwise 80 % credible intervals for function f_4 and $j = 6$ (panels a) and b)) and function f_5 (panels c) and d)). Left panel: Estimators with global variance; right panel: Estimators with locally adaptive variance.*

Figure 5 shows boxplots of log(MSE) for the various estimators. Panel a) refers to function f_6, panel b) to function f_7 and panel c) to function f_8. For function f_6 the best results were 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_7 and f_8 the estimators without main effects perform slightly better although the differences are small. In all cases the estimators with global variance and locally adaptive variance perform almost equally well. An exception is function f_7 which is the only function under study with moderate spatial variability. Here the locally adaptive variants perform slightly better. Compared to the results of Smith and Kohn (1997) our approach is competitive. For function f_6 the estimators without main effects are among the three best in Smith and Kohn's study. The estimators with main effects included perform equally well (if not slightly better) than their best estimator for f_6 which is the cubic tensor product spline. For f_7 our estimators are comparable to the cubic thin plate splines which is the third best estimator. For function f_8 Smith and Kohn's Bayesian subset selection-based procedure cleary outperforms the other estimators in their study including the parametric linear fit. The performance of our estimator is comparable to that of the other estimators which are relatively close.

We also investigated the coverage of pointwise credible intervals of our estimators. The average coverage of all estimators is within a range of 82 to 88% which confirms the findings of the previous sections that the fully Bayesian approach yields rather conservative credible intervals. An exception are the estimators with main effects included for f_2 where the average coverage is only 68 %.

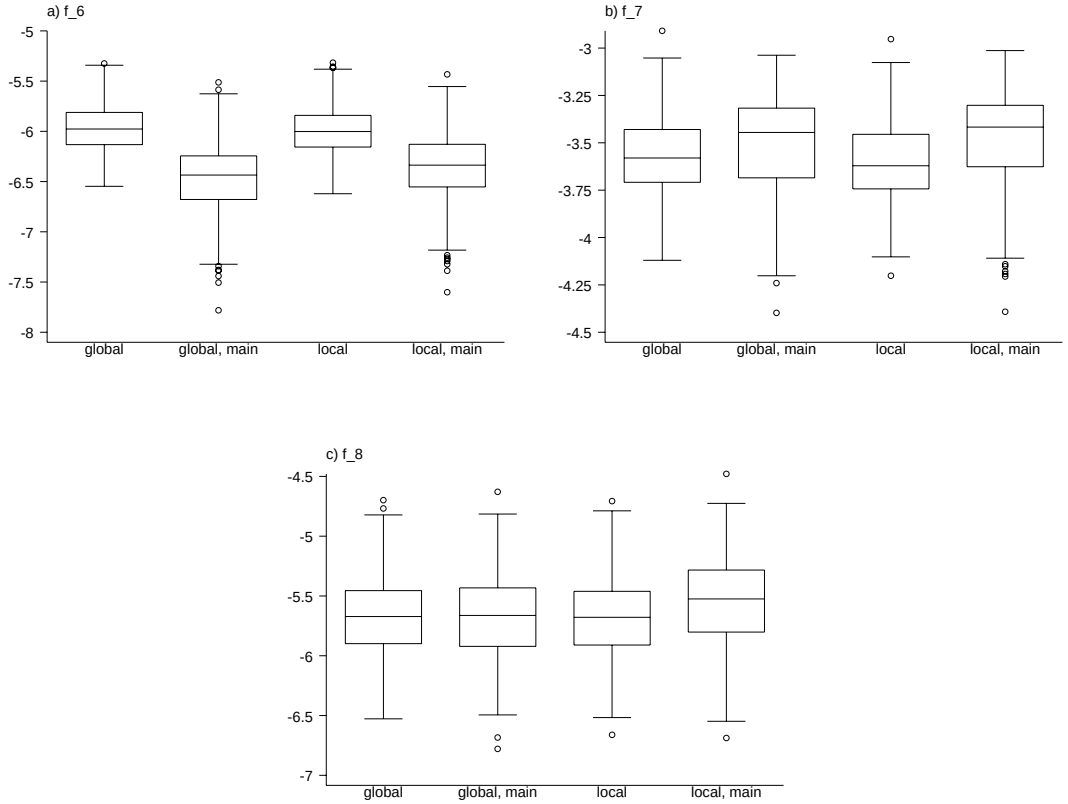


Figure 5: *Boxplots of log(MSE) for the various surface estimators.*

5 Application to rents for flats

In this section, we demonstrate the practicability of our approach with a careful analysis of rent data that have been collected to create a "rental guide" for flats in Munich. An application with binary responses to unemployment data based on the methodology of this paper can be found in Fahrmeir et al. (2001). There the focus is on reemployment chances of unemployed workers given covariates like age, calendar time, duration of unemployment and the district in which the unemployed live. An analysis of treatment costs in hospitals can be found Lang et al. (2001). Both applications show the practicability of our approach even for very large datasets.

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 metrical 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. More specifically, we include additional spatially correlated random effects b_L^{corr} into our model that are specific for subquarters ("Bezirksviertel") in Munich to account for extra spatial variation. Additionally, unstructured (uncorrelated) random effects b_L^{uncorr} with L as the grouping variable are incorporated in order to capture local extra variation. So we choose the Gaussian additive mixed model with predictor

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + f_{12}(F,Y) + b_L^{corr} + b_L^{uncorr} + v'\gamma.$$

The main effects f_1 and f_2 of floor space and year of construction are modelled 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 and with the smoothness prior (7). We have also experimented with P-splines and locally adaptive variances but the differences were negligible. For the spatially correlated random effects b_L^{corr} we choose a Markov random field prior with adjacency weights (see Fahrmeir and Lang, 2000a, b for details), and for the uncorrelated random effects the prior (9).

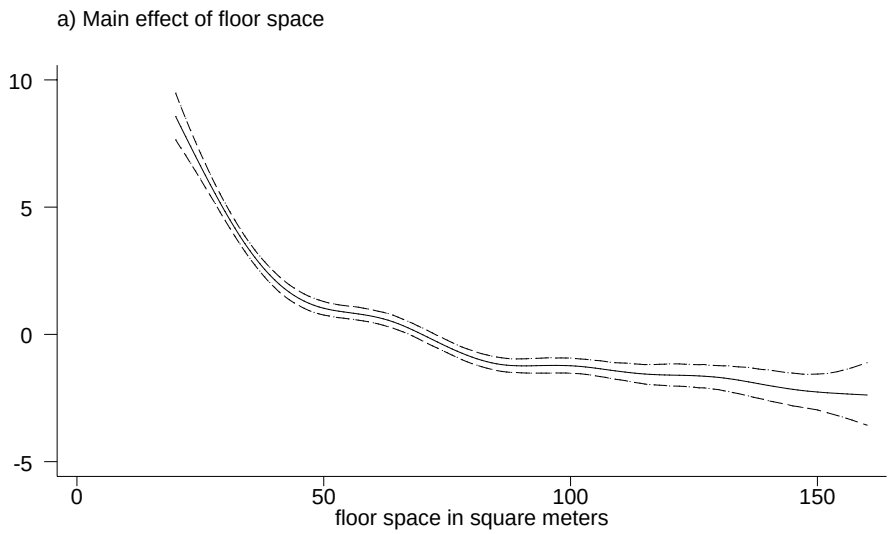
Figure 6 shows the effects of floor space and year of construction. Panels a) and b) show the posterior means together with 80 % pointwise credible intervals of the main effects. Panel c) displays the posterior mean of the interaction term. Figure 6 a) shows the strong influence of floor space on rents: 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 untill the '50s. It then distinctly increases untill 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.

Figure 7 a) shows a map of Munich, displaying subquarters and the posterior mean estimates of the *sum* of the spatially correlated and uncorrelated random effects. Note that the correlated effects clearly exceed the uncorrelated effects with a range approximately between -1.7 to 1.7. The coefficients of the uncorrelated effects have only a range between -0.5 and 0.5. The inclusion of spatially correlated and uncorrelated random effects 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 random effects should considerably decrease. Figure 7 b) displays the sum of both random effects 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 remained more or less stable. The variation of the correlated random effects, however, decreased considerably.

6 Conclusions

In this paper we propose a fully Bayesian approach for P-splines by replacing the difference penalties in the classical approach by random walk priors. The approach is extended to surface fitting using the tensor product of one dimensional B-splines and spatial generalizations of random walk priors. Highly oscillating curves can be estimated by replacing Gaussian priors by a hierarchical t-formulation. Although the simulation studies and the data example in this paper covers primarily the case of Gaussian responses our approach is already implemented and works well for non-Gaussian responses including models for multicategorical data. An application to unemployment data with binary responses can be found in Fahrmeir et al. (2001) showing the practicability of Bayesian P-splines even for very large datasets. The comparison of Bayesian P-splines with the classical approach in Section 4.1 shows that our approach is (at least) competitive. In the classical approach the choice of the smoothing parameters via cross validation is sometimes difficult whereas the simulatenous estimation of functions and smoothing parameters in the Bayesian approach works quite well in all situations. We consider this as a distinct advantage of our approach particularly in situations with a moderate or large number of smoothing parameters as in our data example on rents for flats. As has been shown in Section 4.2, our approach for estimating highly oscillating curves outperforms many of the recent approaches in the literature. Surface estimating via P-splines is competitive with other smoothers, although there are some open problems remaining. Particularly, estimation of surfaces via MCMC is relatively slow because the bandwidth of the posterior precisison matrix is much larger than for univariate smoothers. A simple remedy might be to update the parameters row- or columnwise rather than all parameters in one step. Then, the bandwidth of precision matrices of full conditionals reduces considerably. Another problem with surface fitting is that the values of the covariates might be irregularly distributed. In such situations the performance of our estimator might be improved by allowing the knot pairs to adapt to the data as has been suggested by Smith and Kohn (1997). We will investigate this in future research.



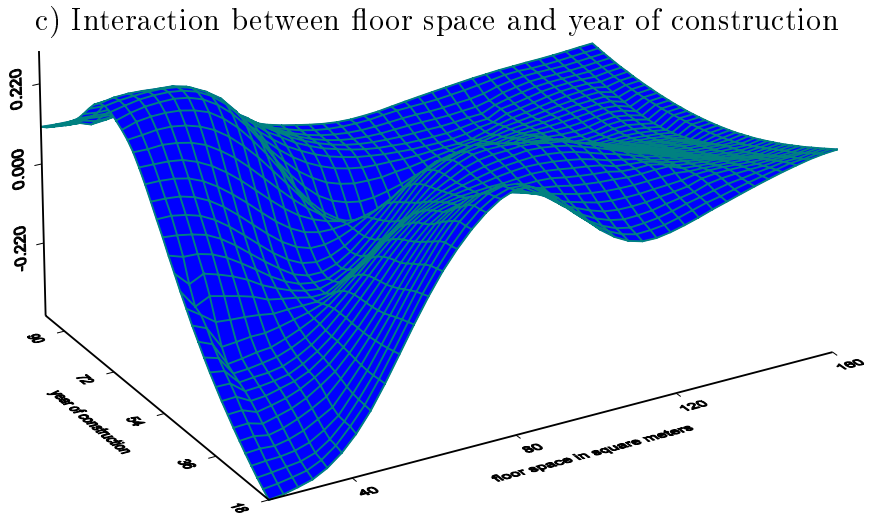


Figure 6: *Effect of floor space and year of construction. Panel a) and b) show the main effects (posterior means and 80 % pointwise credible intervals). The intercation is shown in panel c) (posterior mean only).*

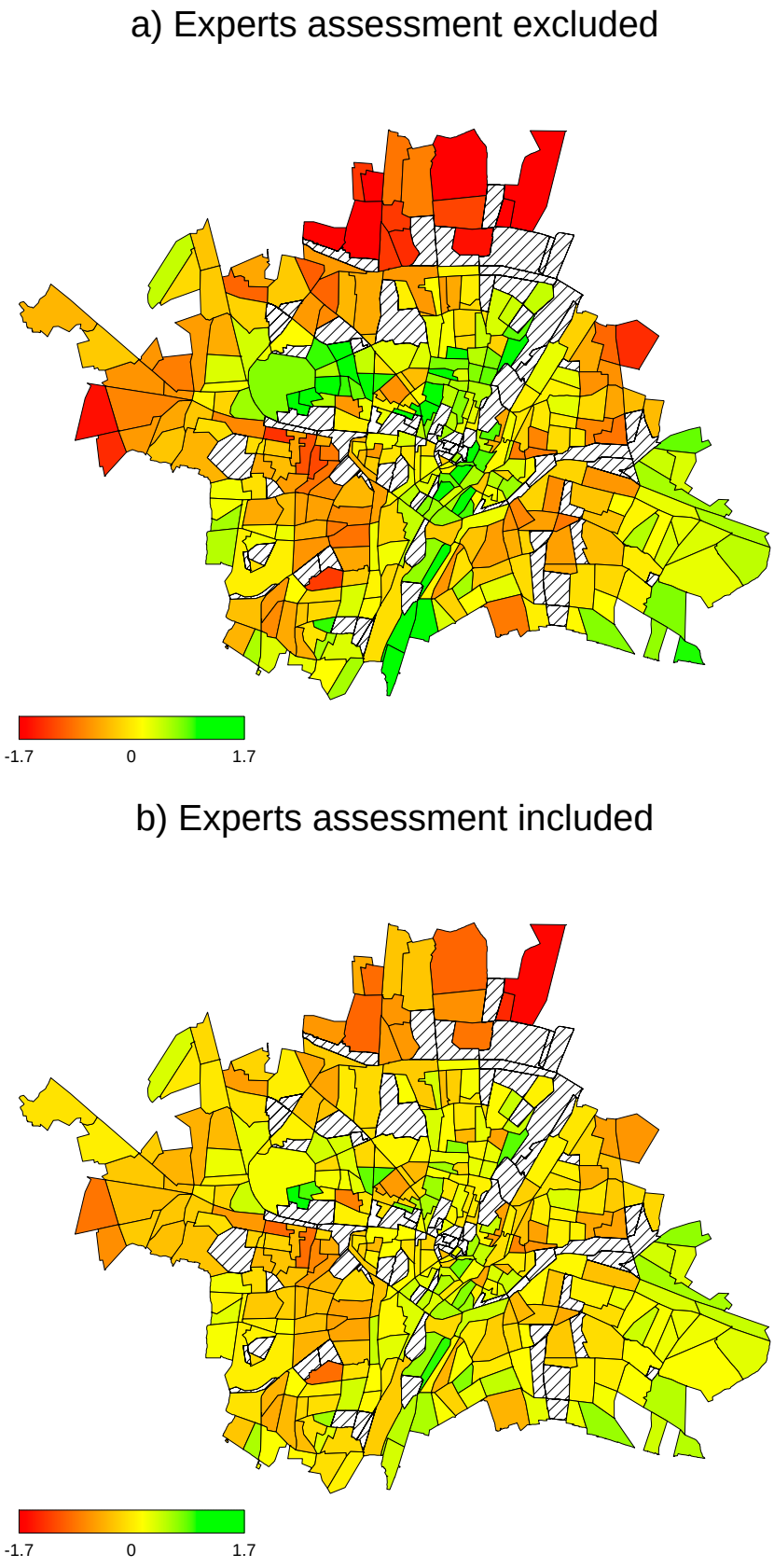


Figure 7: *Posterior means of the sum of spatially correlated and uncorrelated random effects. Panel a) refers to the model where the experts assessment of location is exluded, panel b) refers to the model where it is included.*

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary Polychotomous Response Data. *J. Am. Statist. Ass. 88*, 669–679.

Besag, J. and D. Higdon (1999). Bayesian Analysis of Agricultural Field Experiments. *J. R. Statist. Soc. 61*, 691–746.

Besag, J. and C. Kooperberg (1995). On Conditional and Intrinsic Autoregressions. *Biometrika 82*, 733–746.

Besag, J., Y. York, and A. Mollie (1991). Bayesian Image Restoration with two Applications in Spatial Statistics (with discussion). *Ann. Inst. Statist. Math. 43*, 1–59.

Biller, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *J. Comp. Stat. and Graph. Stat. 12*, 122–140.

Breiman, L. and J. Friedman (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Statist. Ass. 80*, 580–598.

Carter, C. and R. Kohn (1994). On Gibbs Sampling for State Space Models. *Biometrika 81*, 541–553.

Chen, M. and D. Dey (2000). Bayesian Analysis for Correlated Ordinal Data Models. In D. Dey, S. Ghosh, and B. Mallick (Eds.), *Generalized linear models: A Bayesian perspective*, pp. 133–159. Marcel Dekker, New York.

Chen, Z. (1993). Fitting Multivariate Regression Functions by Interaction Spline Models. *J. R. Statist. Soc. B 55*, 473–491.

Clayton, D. (1996). Generalized Linear Mixed Models. In R. S. Gilks, W. and S. D. (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 275 – 301. Chapman and Hall, London.

Cleveland, W. and E. Grosse (1991). Computational Methods for Local Regression. *Statistics and Computing 1*, 47–62.

De Boor, C. (1978). *A Practical Guide to Splines*. Springer–Verlag, New York.

Denison, D., B. Mallick, and A. Smith (1998). Automatic Bayesian Curve Fitting. *J. R. Statist. Soc. 60*, 333–350.

Donoho, D. and I. Johnstone (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *J. Am. Statist. Ass. 90*, 1200–1224.

Eilers, P. and B. Marx (1996). Flexible Smoothing using B-splines and Penalized Likelihood (with comments and rejoinder). *Statist. Sci. 11*, 89–121.

Fahrmeir, L. and S. Lang (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Appl. Statist. (JRSS C) (to appear)*.

Fahrmeir, L. and S. Lang (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Ann. Inst. Statist. Math. 53*, 11–30.

Fahrmeir, L., S. Lang, J. Wolff, and S. Bender (2001). Semiparametric Bayesian Time-Space Analysis of Unemployment Duration. SFB 386 Discussion Paper 211, University of Munich.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (3 ed.). Springer, New York.

Friedman, J. (1991). Multivariate Adaptive Regression Splines (with discussion). *Ann. Statist. 19*, 1–141.

Friedman, J. and B. Silverman (1989). Flexible Parsimonious Smoothing and Additive Modeling (with discussion). *Technometrics 31*, 3–39.

Fronk, E. M. and L. Fahrmeir (1998). Function Estimation with Locally Adaptive Dynamic Models. SFB 386 Discussion Paper 135, University of Munich.

Gamerman, D. (1997). Efficient Sampling from the Posterior Distribution in Generalized Linear Models. *Statistics and Computing 7*, 57–68.

George, A. and J. W. Liu (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*, Alexandria, pp. 571–578.

Gu, C., D. Bates, Z. Chen, and G. Wahba (1989). The Computation of CGV Functions Through Householder Tridiagonalization With Application to the Fitting of Interaction Spline Models. *SIAM Journal of Matrix Analysis and Applications 10*, 457–480.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient Models. *J. R. Statist. Soc. B 55*, 757–796.

Hastie, T. and R. Tibshirani (2000). Bayesian Backfitting. *Statist. Sci. (to appear)*.

Higdon, D. (1994). *Spatial Applications of Markov Chain Monte Carlo for Bayesian Inference*. Ph. D. thesis, University of Washington.

Hobert, J. and G. Casella (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *J. Am. Statist. Ass. 91*, 1461–1473.

Knorr-Held, L. (1996). *Hierarchical Modelling of Discrete Longitudinal Data*. Ph. D. thesis, University of Munich.

Knorr-Held, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scand. J. Statist. 26*, 129–144.

Lang, S. and L. Fahrmeir (2001). Bayesian Generalized Additive Mixed Models. A Simulation Study. SFB 386 Discussion Paper 230, University of Munich.

Lang, S., P. Kragler, G. Haybach, and L. Fahrmeir (2001). Bayesian Space-Time Analysis of Health Insurance Data. SFB 386 Discussion Paper 237, University of Munich.

Luo, Z. and G. Wahba (1997). Hybrid Adaptive Splines. *J. Am. Statist. Ass. 92*, 107–116.

Mallick, B., D. Denison, and A. Smith (2000). Semiparametric Generalized Linear Models: Bayesian Approaches. In D. Dey, S. Ghosh, and B. K. Mallick (Eds.), *Generalized linear models: A Bayesian perspective*. Marcel–Dekker.

Marx, B. D. and H. C. Eilers, P. (1998). Direct Generalized Additive Modeling with Penalized Likelihood. *Computational Statistics and Data Analysis 28*, 193–209.

Rue, H. (2000). Fast Sampling of Gaussian Markov Random Fields with Applications. Technical report, University of Trondheim, Norway.

Ruppert, D. and R. J. Carroll (2000). Spatially Adaptive Penalties for Spline Fitting. *Australian and New Zealand Journal of Statistics, to appear*.

Smith, M. and R. Kohn (1996). Nonparametric Regression using Bayesian Variable Selection. *Journal of Econometrics 75*, 317–343.

Smith, M. and R. Kohn (1997). A Bayesian Approach to Nonparametric Bivariate Regression. *J. Am. Statist. Ass. 92*, 1522–1535.

Stone, C., M. Hansen, C. Kooperberg, and Y. Troung (1997). Polynomial Splines and their Tensor Products in Extended Linear Modeling (with discussion). *Ann. Statist. 25*, 1371–1470.

Wood, S. N. (2000). Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *J. R. Statist. Soc. B 62*, 413–428.