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

Appeared in Journal of Computational and Graphical Statistics, 13, 183-212

Abstract

P-splines are an attractive approach for modelling 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 modelling interactions between metrical 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 application on rents for flats in Munich and on human brain mapping.

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

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 metrical response variable y is assumed to be the sum of smooth functions f_j. Several proposals are available for modelling and estimating the smooth functions f_j, see e.g. Fahrmeir and Tutz (2001, Ch. 5) and Hastie et al. (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 parameterization 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 metrical response which is extended to GAMs by Biller (2000) and Mallick et al. (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 selcection 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 Whaba (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*. Kamman and Wand (2001) calls models with an additive predictor composed of nonlinear functions of metrical 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/`.

The rest of this paper is organized as follows: Section 2 describes Bayesian AMs with P-splines and extensions. Section 3 gives details about MCMC inference for the proposed models. Section 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 5, the methods of this paper are applied to complex datasets on rents for flats in Munich and on human brain mapping.

2 Bayesian AMs and extensions based on P-Splines

2.1 Additive models

Consider regression situations where observations (y_i, x_i, v_i), $i = 1, \ldots, n$, on a metrical 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. Given covariates and unknown parameters, we assume that the responses y_i, $i = 1, \ldots, n$, are independent and Gaussian with mean or predictor

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

and a common variance σ^2 across subjects. Here $f_1, \ldots, f_p$ are unknown smooth functions of the metrical 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 substracted 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 $x_{j,min} = \zeta_{j0} < \zeta_{j1} < \ldots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$ 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 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

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

is maximized with respect to the unknown regression coefficients $\beta_1, \ldots, \beta_p$ and γ. In (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 critertia 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 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 (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

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

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 its *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 global smoothness priors

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

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 (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 patricular 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 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 (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 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. 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) 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 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 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) or Stone et al. (1997), 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. 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 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 neighbouring parameters and the variance component τ_{js}^2. The most commonly used prior specification based on the four nearest neighbours can be defined by

$$\beta_{js\rho\nu}|\cdot \sim N\left(\frac{1}{4}(\beta_{js\rho-1,\nu} + \beta_{js\rho+1,\nu} + \beta_{js\rho,\nu-1} + \beta_{js\rho,\nu+1}), \frac{\tau_{js}^2}{4}\right) \tag{6}$$

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 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 (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 (5).

Prior (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 (6) to

$$\beta_{js\rho\nu}|\cdot \sim N\left(\sum_{(kl)\in\partial_{(\rho\nu)}} \frac{\delta_{(\rho\nu)(kl)}}{\delta_{(\rho\nu)+}} \beta_{jskl}, \frac{\tau_{\rho\nu}^2}{\delta_{(\rho\nu)+}}\right). \tag{7}$$

Here, $\partial_{\rho\nu}$ corresponds to the set of neighbouring 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 (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 $\delta_{(\rho\nu)(kl)} \sim G(\frac{1}{2}, \frac{1}{2})$.

2.3 Geoadditive models

In a number of applications responses depend not only on metrical 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 metrical 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) leading to *geoadditive models* (Kamman and Wand, 2001). 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 et al. (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

$$\beta_{str,s}|\beta_{str,u}, u \neq s \sim N\left(\sum_{u\in\partial_s} \frac{1}{N_s}\beta_{str,u}, \frac{\tau^2_{str}}{N_s}\right), \tag{8}$$

where N_s is the number of adjacent regions or pixels, and ∂_s denotes the regions which are neighbours of region s. Hence, prior (8) can be seen as a 2 dimensional extension of a first order random walk. More general priors than (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 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.

$$\beta_{unstr}(s) \sim N(0, \tau^2_{unstr}), \quad s = 1, \ldots, S. \tag{9}$$

Formally, the priors for β_{str} and β_{unstr} can both be brought into the form (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 $\tau^2_{unstr} \sim IG(a_{unstr}, b_{unstr})$.

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

$$\begin{aligned} p(\alpha) \quad \propto \quad & L(y, \beta_1, \ldots, \beta_p, \beta_{str}, \beta_{unstr}, \gamma, \sigma^2) \prod_{j=1}^{p} \left(p(\beta_j|\tau^2_j) p(\tau^2_j) \right) \\ & p(\beta_{str}|\tau^2_{str}) p(\tau^2_{str}) p(\beta_{unstr}|\tau^2_{unstr}) p(\tau^2_{unstr}) p(\gamma) p(\sigma^2) \end{aligned} \tag{10}$$

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 (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

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

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 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 bandmatrix 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 (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 bandsize usually differs from row to row. This can be exploited to further improve the computational effiency. 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 (4). In the case of a two dimensional P-spline, the full conditionals for the weights $\delta_{(\rho\nu)(kl)}$ are Gamma distributed with

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

Since all full conditionals involved are known distributions, a simple Gibbs sampler can be used to successively update the parameters of the model.

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 4.1 lies on functions with low or moderate curvature while Section 4.2 deals with the estimation of highly oscillating functions. 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 in the literature.

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 a) - c) shows typical datasets 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 1 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 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 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 1 and Table 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 underyling 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 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 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 hyperaparameters $b = 0.0005$ 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.

Table 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

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 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 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 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 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.

- 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 developped 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 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-pslines 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 funtion 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 adption 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 $log10(\sqrt{(}MSE))$, i.e. Ruppert and Carroll's 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 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%.

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 (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 (6) and (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 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 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 cleary 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.

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.

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 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 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 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 with smoothness prior (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 (8), and (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 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 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 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. Taking a look at Table 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 et al., 2002), which can be used as a tool for model comparison in complex hierarchical Bayesian models.

Figure 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 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 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.

Table 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

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 scientifical interest but also important in many practical disciplines, e.g. in surgery. The realization of human brain 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 perodically 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-neglectable 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 analysing 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 analyse 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 ly outside the brain so that finally a total number of 2950 time series is analysed. 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 analyse the data pixelwise using Gaussian regression models with predictors

$$\eta_{ti} = \gamma_0 + f_1^i(t) + f_2^i(t) Z_{it}, \quad t = 1, \ldots, 70, \quad i = 1, \ldots, 2950. \tag{12}$$

Here, Z_{it} is a delayed and continuously modified stimulus which is routinely obtained from 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 et al. (2000) who apply dynamic or state space models to estimate (12). Examples for the pixelwise analysis are given in Figure 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 (12). 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 % credibel 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 (6) and (7). Figure 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 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.

Table 3: *Human brain mapping: DIC of the six surface estimators.*

	$t = 18$	$t = 38$	$t = 58$
global variance	18117	17151	19749
spatially adaptive variance	17879	16738	19400

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 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 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.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary Polychotomous Response Data. *Journal of the American Statistical Association 88*, 669–679.

Besag, J. and D. Higdon (1999). Bayesian Analysis of Agricultural Field Experiments. *Journal of the Royal Statistical Society B 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). *Annals of the Institute of Statistical Mathematics 43*, 1–59.

Biller, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics 12*, 122–140.

Breiman, L. and J. Friedman (1985). Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association 80*, 580–598.

Carter, C. and R. Kohn (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.

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. *Journal of the Royal Statistical Society B 60*, 333–350.

Di Matteo, I., C. R. Genovese, and R. E. Kass (2001). Bayesian curve-fitting with free-knot splines. *Biometrika 88*, 1055–1071.

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

Fahrmeir, L. and S. Lang (2001a). Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. *Applied Statistics (JRSS C) 50*, 201–220.

Fahrmeir, L. and S. Lang (2001b). Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics 53*, 11–30.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (2 ed.). Springer, New York.

Friedman, J. (1991). Multivariate Adaptive Regression Splines (with discussion). *Annals of Statistics 19*, 1–141.

George, A. and J. W. Liu (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice–Hall.

Gössl, C. (2001). *Bayesian Models in functional Magnetic Resonance Imaging: Approaches for Human Brain Mapping*. Shaker Verlag, Aachen.

Gössl, C., D. Auer, and L. Fahrmeir (2000). Dynamic models in fMRI. *Magnetic Resonance in Medicine 43*, 72 – 81.

Hansen, M. H. and C. Kooperberg (2002). Spline Adaptation in Extended Linear Models. *Statistical Science 17*, 2–51.

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

Hastie, T. and R. Tibshirani (1993). Varying-coefficient Models. *Journal of the Royal Statistical Society B 55*, 757–796.

Hastie, T. and R. Tibshirani (2000). Bayesian Backfitting. *Statistical Science 15*, 193–223.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

Hobert, J. and G. Casella (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association 91*, 1461–1473.

Kamman, E. and M. Wand (2001). Geoadditive Models. Technical report, Harvard School of public Health.

Knorr-Held, L. (1999). Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics 26*, 129–144.

Lange, N. (1996). Statistical Approaches To Human Brain Mapping By Functional Magnetic Resonance Imaging. *Statistics in Medicine 15*, 389 – 428.

Luo, Z. and G. Wahba (1997). Hybrid Adaptive Splines. *Journal of the American Statistical Association 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. (2001). Fast Sampling of Gaussian Markov Random Fields with Applications. *Journal of the Royal Statistical Society B 63*, 325–338.

Ruppert, D. and R. J. Carroll (2000). Spatially Adaptive Penalties for Spline Fitting. *Australian and New Zealand Journal of Statistics 42*, 205–223.

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. *Journal of the American Statistical Association 92*, 1522–1535.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B, to appear*.

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

Wand, M. (2000). A Comparision of Regression Spline Smoothing Procedures. *Computational Statistics 15*, 443 – 462.

Wood, S. N. (2000). Modelling and Smoothing Parameter Estimation with Multiple Quadratic Penalties. *Journal of the Royal Statistical Society B 62*, 413–428.

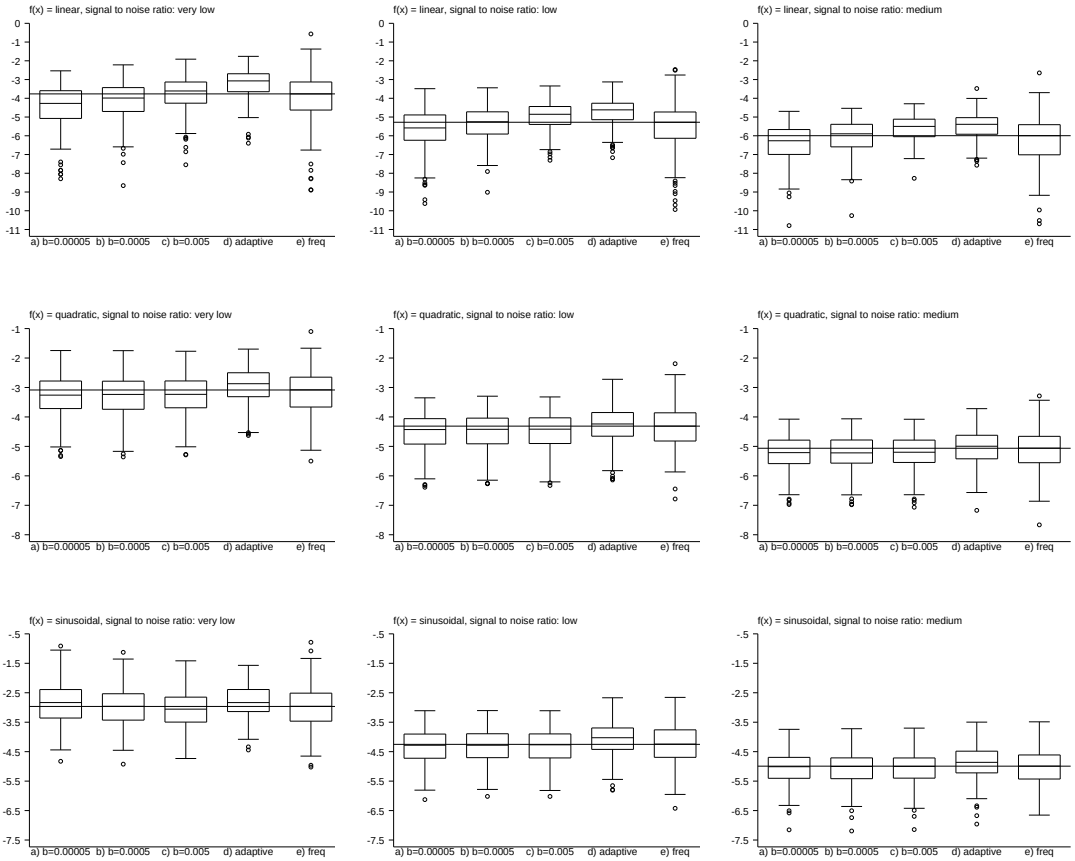

Figure 1: *Boxplots of log(MSE) for the various estimators of simulation study 1. The first row refers to the linear function f_1, the second row to the quadratic function f_2 and the third row to the sinusoidal function f_3. The left panel corresponds to a very low SNR ($\sigma = 1$), the medium panel 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.*

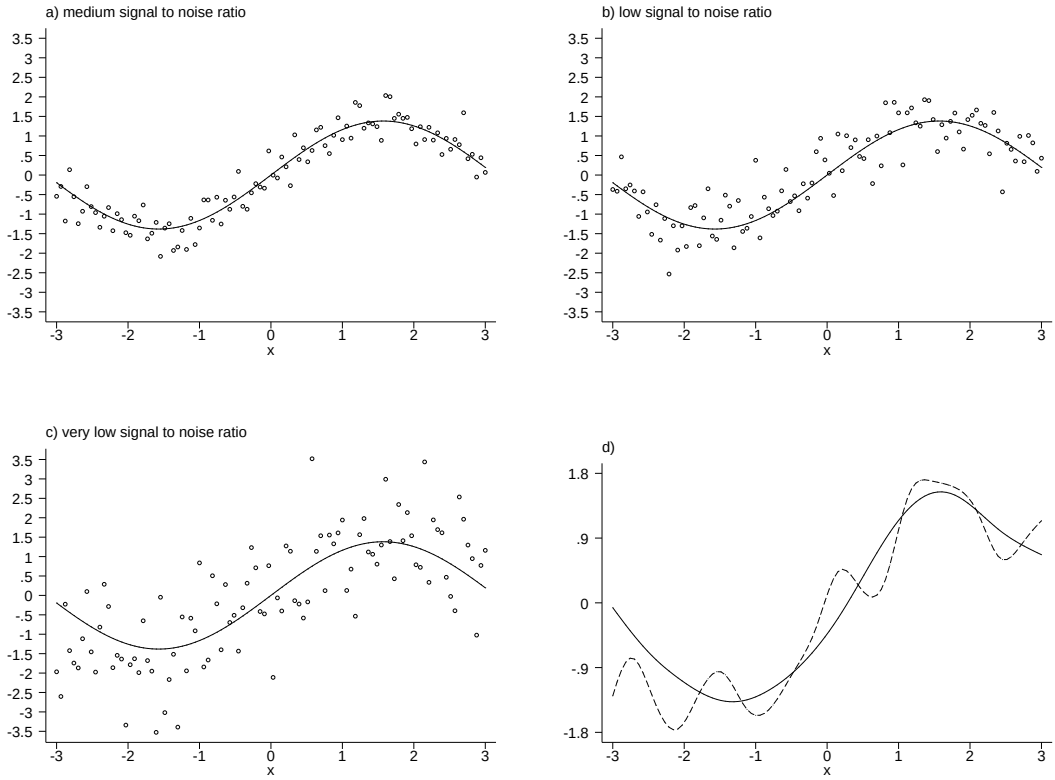

Figure 2: *Sinusoidal function of simulation study 1: The graphs a)-c) show typical datasets for medium, low and very low signal to noise ratio. The true function is included in the graphs (solid lines). Panel d) displays the classical (dashed line) and the Bayes estimator (solid line) for a particular replication where cross validation fails.*

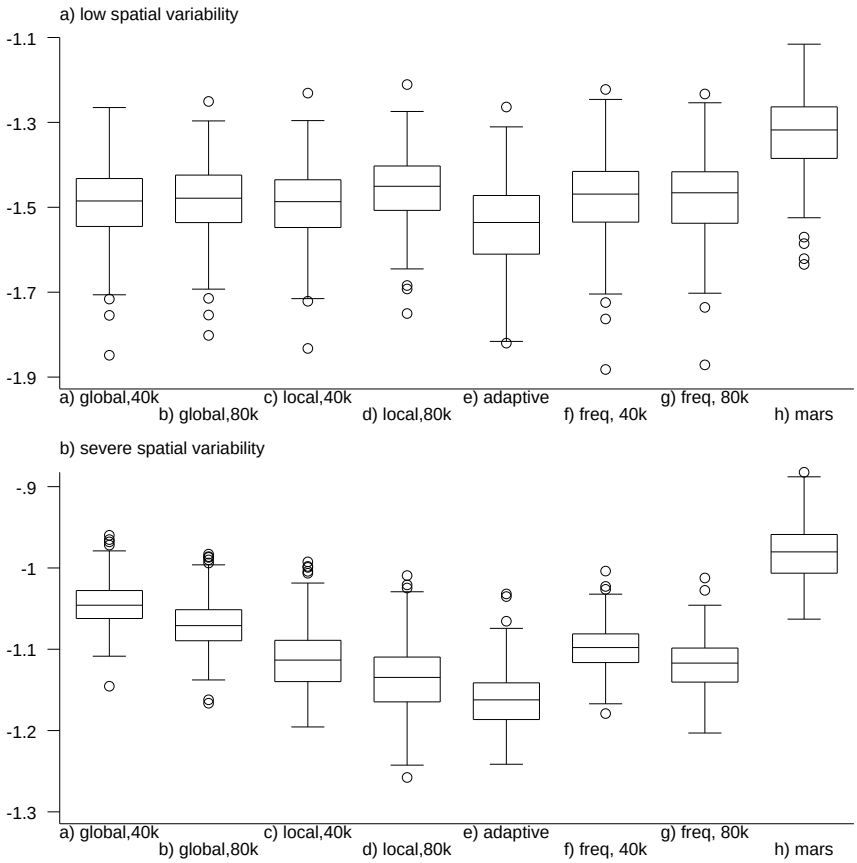

Figure 3: *Simulation study 2: Boxplots of* $log_{10}(\sqrt{(}MSE))$ *for function* f_4 *with low spatial variability (panel a)) and severe spatial variability (panel b)). From left to right the boxplots in the graphs correspond to Bayesian P-splines with a global variance (40 and 80 knots), Bayesian P-splines with locally adaptive variances (40 and 80 knots), adaptive Bayesian regression splines, classical P-splines (40 and 80 knots) and MARS.*

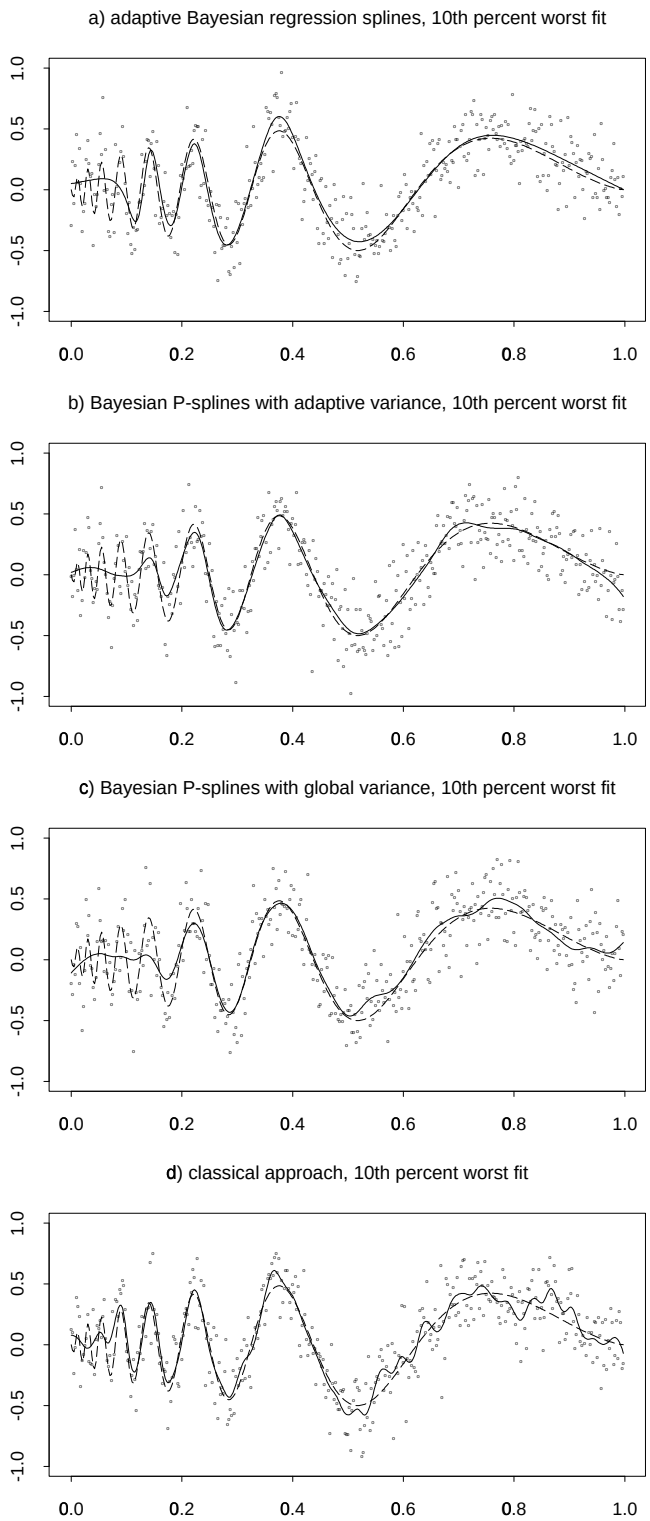

Figure 4: *Simulation study 2: The graphs show the respective 10th percent worst fits in terms of the MSE measure plotted in Figure 3 for Biller's adaptive Bayesian regression splines, Bayesian P-splines with 80 knots and adaptive variances, Bayesian P-splines with 80 knots and global variance and classical P-splines with 80 knots.*

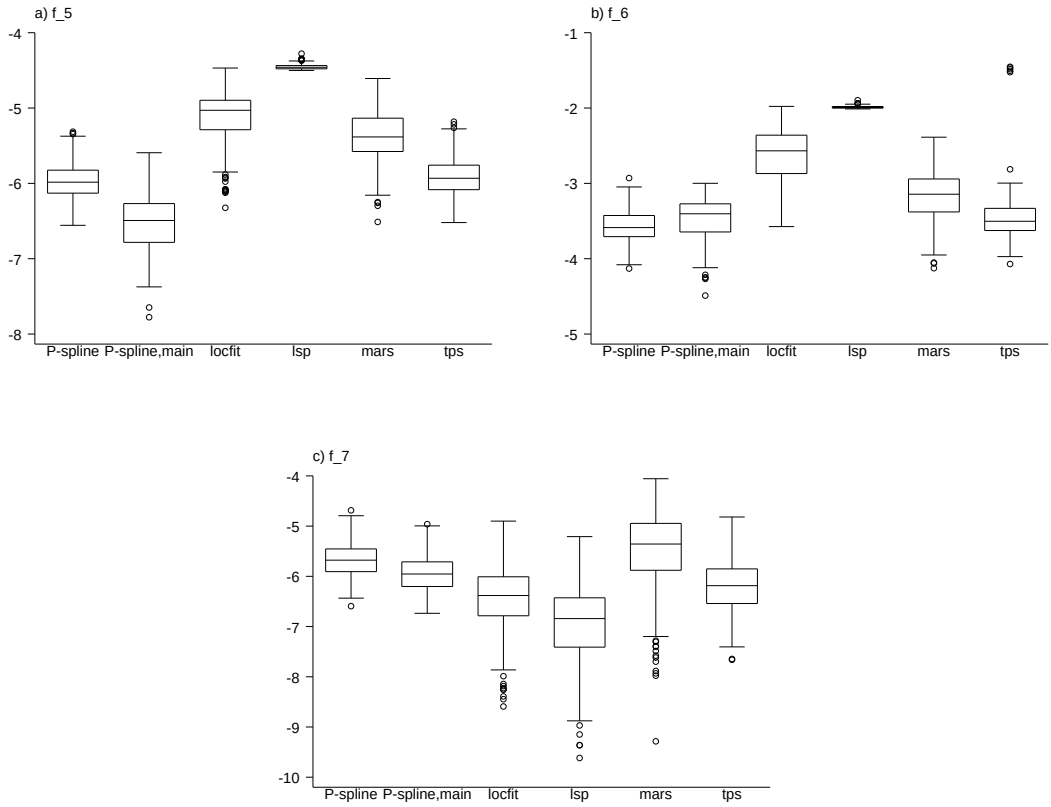

Figure 5: *Simulation study 3: Boxplots of log(MSE) for the various surface estimators. Panel a) corresponds to function f_5, panel b) to function f_6 and panel c) to function f_7. From left to right the respective boxplots refer to Bayesian P-splines without main effects, Bayesian P-splines with main effects, the parametric linear interaction model (lsp), Clive Loader's 'locfit', MARS and thin plate splines (tps).*

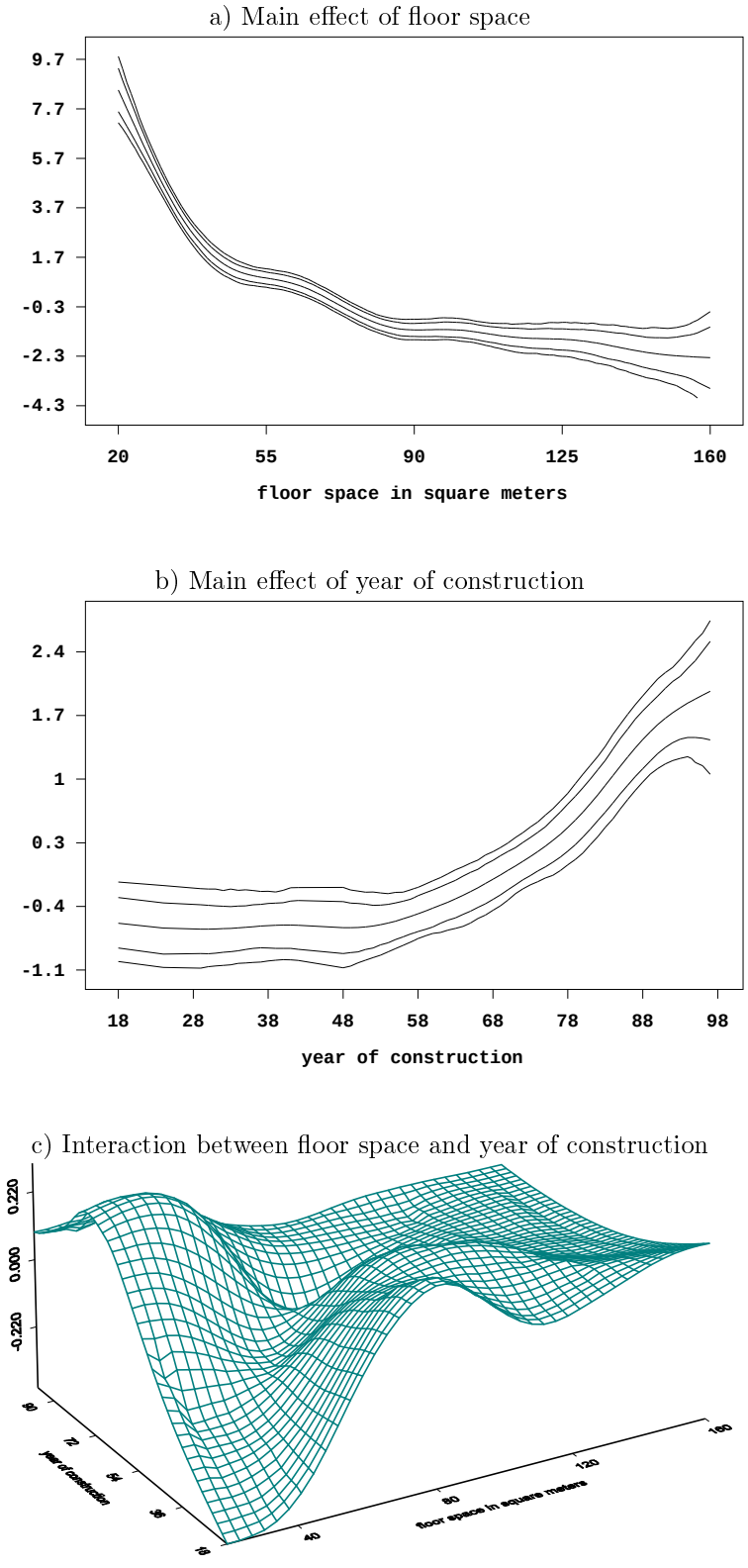

Figure 6: *Rents for flats: Effect of floor space and year of construction. Panel a) and b) show the main effects (posterior means, 80% and 95% pointwise credible intervals). The posterior means of the interaction effect is given in panel c).*

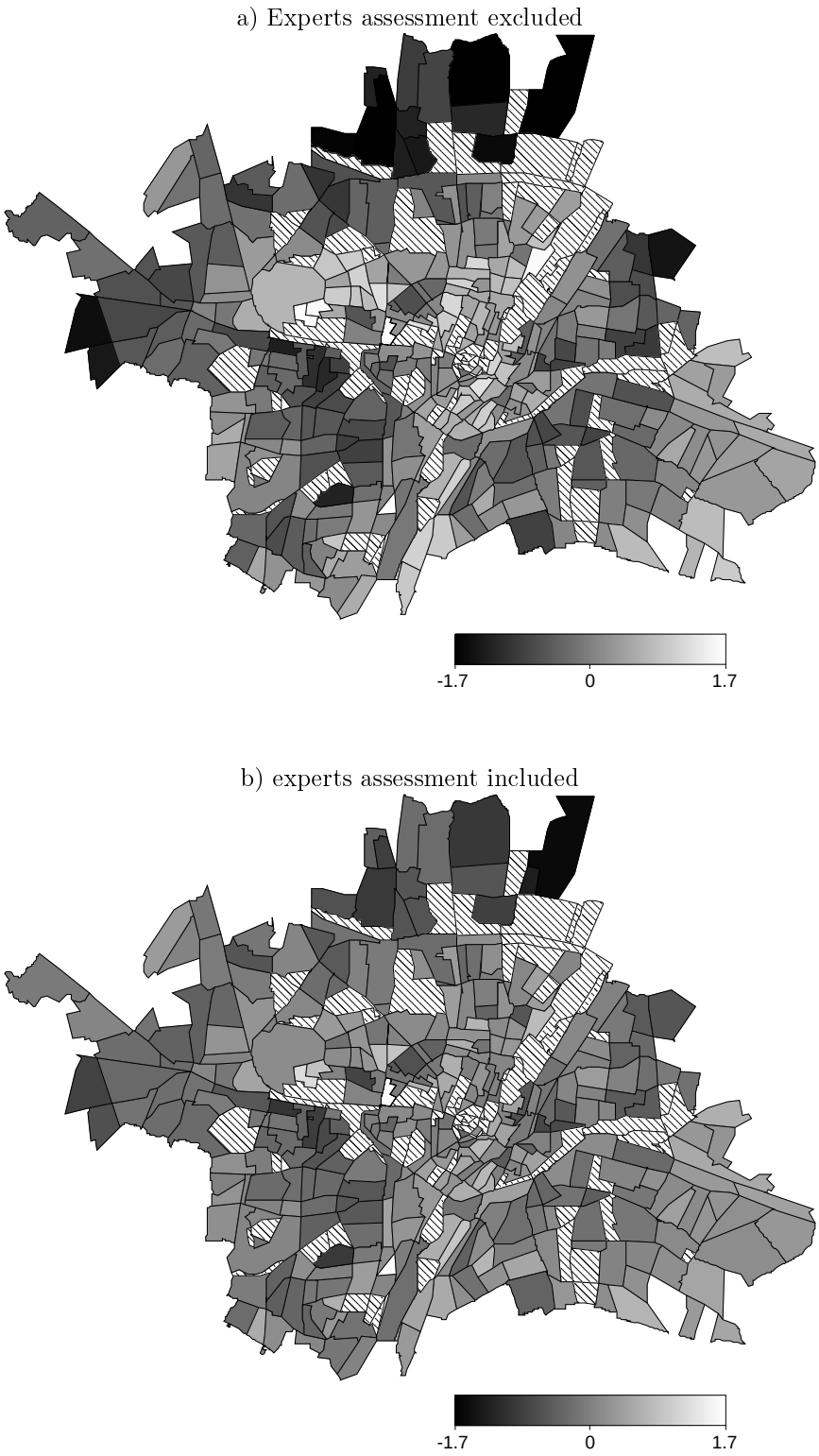

Figure 7: *Rents for flats: Posterior means of the spatial effect* f_{spat}*. Panel a) refers to the model that excludes the experts assessment of location, panel b) refers to the model that includes it.*

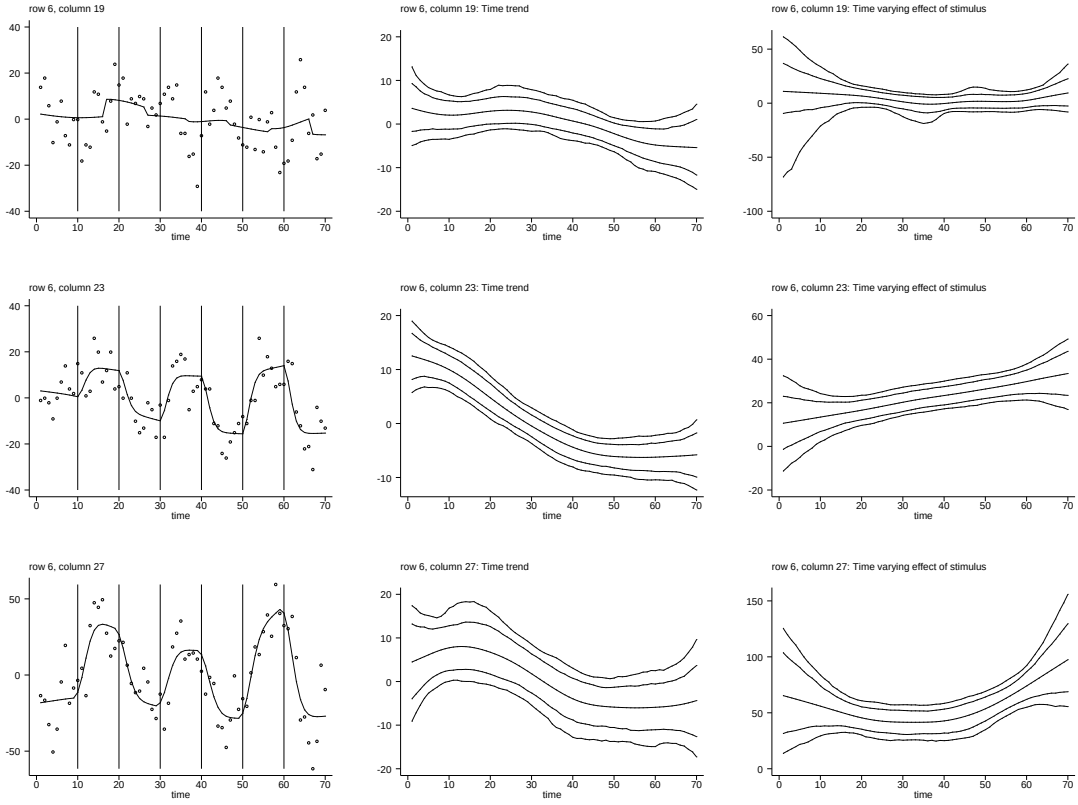

Figure 8: *Human brain mapping: The graphs display examples of the pixelwise analysis. The left panel shows the time series Y_{it} together with $\hat{Y}_{it}$ (solid line). The middle and the right panel display the estimated time trend and the time varying effect of the transformed stimulus. Shown is the posterior mean, 80% and 95% pointwise credible intervals.*

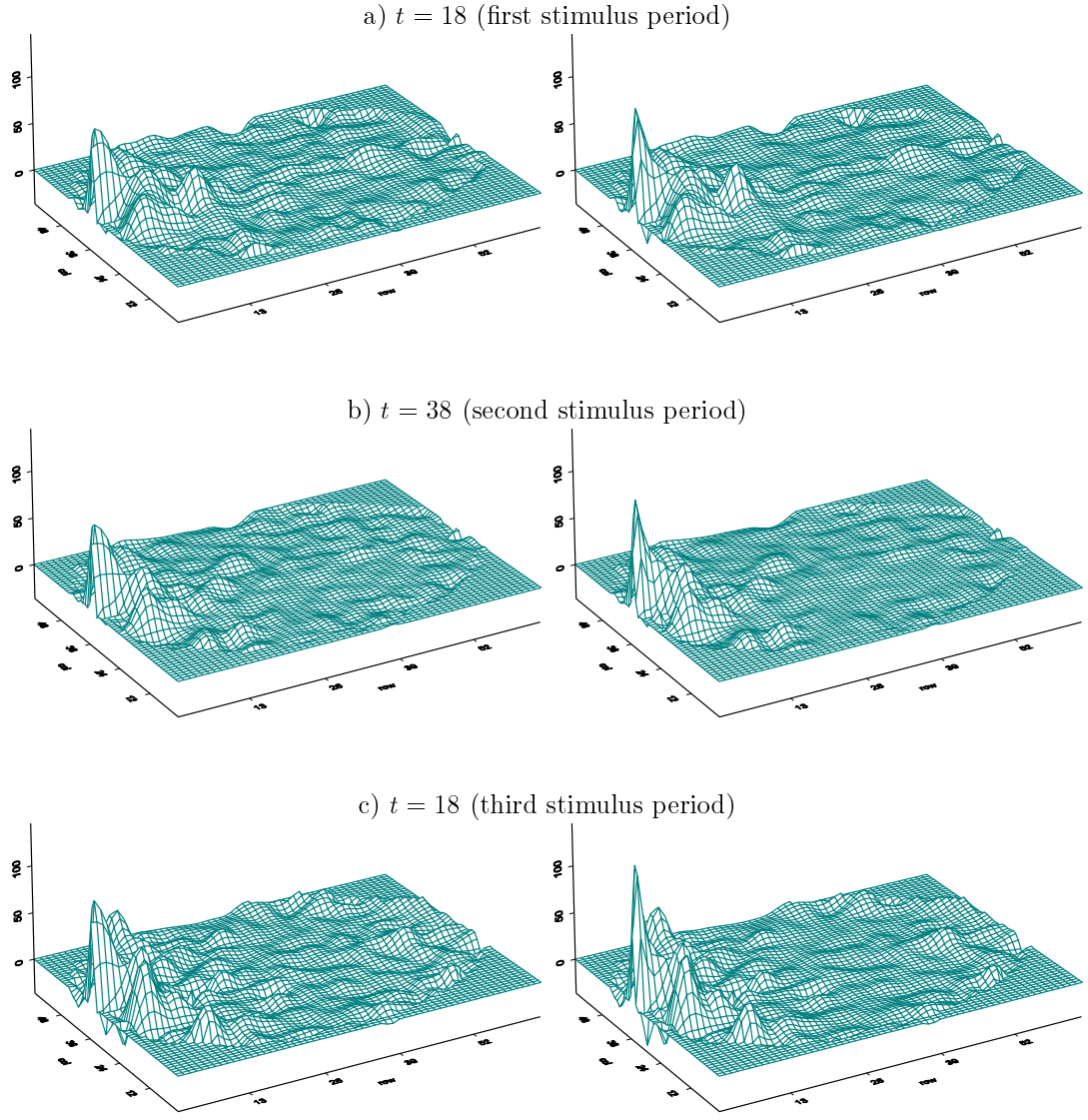

Figure 9: *Human brain mapping: The graphs show the spatially smoothed estimates of the effect of the transformed stimulus from the pixelwise analysis for different time points. The first row corresponds to* $t = 18$ *(first stimulus period), the second row to* $t = 38$ *(second stimulus period) and 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.*