View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

Johns Hopkins University, Dept. of Biostatistics Working Papers

11-1-2004

Spatially Adaptive Bayesian P-Splines with Heteroscedastic Errors

Ciprian M. Crainiceanu
Johns Hokins Bloomberg School of Public Health, Department of Biostatistics, ccrainic@jhsph.edu

David Ruppert
School of Operational Research & Industrial Engineering, Cornell University, dr24@cornell.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, rcarroll@tamu.edu

Suggested Citation

Crainiceanu, Ciprian M.; Ruppert, David; and Carroll, Raymond J., "Spatially Adaptive Bayesian P-Splines with Heteroscedastic Errors" (November 2004). *Johns Hopkins University, Dept. of Biostatistics Working Papers.* Working Paper 61.
http://biostats.bepress.com/jhubiostat/paper61

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.
Copyright © 2011 by the authors

Spatially Adaptive Bayesian P-Splines With Heteroscedastic Errors

Ciprian M. Crainiceanu* David Ruppert† Raymond J. Carroll‡

November 9, 2004

Abstract

An increasingly popular tool for nonparametric smoothing are penalized splines (P-splines) which use low-rank spline bases to make computations tractable while maintaining accuracy as good as smoothing splines. This paper extends penalized spline methodology by both modeling the variance function nonparametrically and using a spatially adaptive smoothing parameter. These extensions have been studied before, but never together and never in the multivariate case. This combination is needed for satisfactory inference and can be implemented effectively by Bayesian MCMC. The variance process controlling the spatially-adaptive shrinkage of the mean and the variance of the heteroscedastic error process are modeled as log-penalized splines. We discuss the choice of priors and extensions of the methodology, in particular, to multivariate smoothing using low-rank thin plate splines. A fully Bayesian approach provides the joint posterior distribution of all parameters, in particular, of the error standard deviation and penalty functions. In the multivariate case we produce maps of the standard deviation and penalty functions. Our methodology can be implemented using the Bayesian software WinBUGS.

Keywords: Knot selection; MCMC; Mixed models; Multivariate smoothing; Spatially adaptive penalty; Thin-plate splines.

*Assistant Professor, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St. E3037 Baltimore, MD 21205 USA. E-mail: ccrainic@jhsph.edu

†Andrew Schultz Jr. Professor of Engineering and Professor of Statistical Science, School of Operational Research and Industrial Engineering, Cornell University, Rhodes Hall, NY 14853, USA. E-mail: dr24@cornell.edu

‡Distinguished Professor of Statistics and Professor of Nutrition and Toxicology, Department of Statistics, Texas A&M University College Station, 447 Blocker Building, TX 77843–3143 USA. E-mail: carroll@stat.tamu.edu

COBRA BEPRESS REPOSITORY Collection of Biostatistics Research Archive

Hosted by The Berkeley Electronic Press

1 Introduction

P-splines (Eilers and Marx, 1996; Ruppert, Wand, and Carroll, 2003) have become a popular nonparametric tool. Their success is due mainly to the use of low rank bases, which makes computations tractable. Also P-splines can be viewed as mixed models and fit with widely available statistical software (Ngo and Wand, 2003; Crainiceanu, Ruppert and Wand 2004).

Bayesian penalized splines (Ruppert, Wand, and Carroll, 2003; Lang and Brezger, 2004) use a stochastic process model as a prior for the regression function. The usual Bayesian assumes that both this processes and the errors are homoscedastic.

The P-spline methodology has been extended to heteroscedastic errors (Ruppert, Wand, and Carroll, 2004) and also to spatially adaptive penalty parameters (Ruppert and Carroll, 2000; Baladandayuthapani, Mallick, and Carroll, 2004; Lang and Brezger, 2004). However, this is the first paper to combine these features. We show that this combination is important. Since the penalty parameter is the ratio of the error variance to the prior variance on the mean function, it is true that spatially varying penalties can adapt to both spatial heterogeneity of both the mean function and the error variance, at least for the purpose of estimation. However, for correct inference it is necessary to separate the spatial heterogeneity of the mean and of the error variance, and the innovation of this paper is to do that. Implementation of this extension was not straightforward because of technical problems such as slow MCMC mixing and sensitivity to the choice of prior, but we were able develop an algorithm that works satisfactorily in practice.

Our methodology can be extended to almost any of the P-spline model, for example, those in Ruppert, Wand, and Carroll (2003) such as additive models, varying coefficient models, interaction models, and multivariate smoothing. As an illustration, we also study low rank thin plate (multivariate) splines. As in the univariate case we model the mean, the variance and the smoothing functions nonparametrically and estimate them from the data using a fully Bayesian approach. The computational advantage of low rank over full rank smoothers becomes even greater in more than one dimension.

Section 2 provides a quick introduction to P-splines and their connections with mixed models. Section 3 discusses the choice of priors for P-spline regression. Section 4 provides simultaneous credible bounds for mean and variance functions and their derivatives. Sections 5 provides an example and comparisons with simpler techniques and in Section 9 we compare our proposed methodology to the one proposed by Baladandayuthapani et al. (2004) for adaptive univariate P–spline regression. Section 7 describes the extension to multivariate smoothing and in Section 8 we present an example of bivariate smoothing with heteroscedastic errors and spatially adaptive smoothing. Section 9 compares our multivariate methodology with other adaptive surface fitting methods. Section 10 presents the full conditional distributions and discusses our implementation of models in WinBUGS and

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

MATLAB.

2 P-Splines regression and mixed models

Consider the regression equation $y_i = m(x_i) + \epsilon_i,\ i = 1, \ldots, n$, where the ϵ_i are independent $N(0, \sigma^2_{\epsilon,i})$ and the mean is modeled as

$$m(x) = m(x, \boldsymbol{\theta}_X) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{k=1}^{K_m} b_k (x - \kappa_k^m)_+^p \ ,$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$, $\boldsymbol{b} = (b_1, \ldots, b_{K_m})^T$, $\boldsymbol{\theta}_X = (\boldsymbol{\beta}^T, \boldsymbol{b}^T)^T$, $\kappa_1^m < \kappa_2^m < \ldots < \kappa_{K_m}^m$ are fixed knots, and a_+^p denotes $\{\max(a, 0)\}^p$. Following Gray (1994) and Ruppert (2002), we take K_m large enough (e.g., 20) to ensure the desired flexibility.

To avoid overfitting the $b_k \sim N\{0, \sigma_b^2(\kappa_k^m)\}$ and $\epsilon_i \sim N\{0, \sigma_\epsilon^2(x_i)\}$ are shrunk towards zero by an amount controlled by $\sigma_b^2(\cdot)$ and $\sigma_\epsilon^2(\cdot)$, which vary across the range of x. In Sections 2.1 and 2.2 $\sigma_\epsilon^2(x_i)$ and $\sigma_b^2(\kappa_k^m)$ are modeled using log-spline models.

The P-spline model can be written in linear mixed model form as

$$\left\{ \begin{array}{lcl} y_i & = & \beta_0 + \beta_1 x_i + \ldots + \beta_p x_i^p + \sum_{k=1}^{K_m} b_k (x_i - \kappa_k^m)_+^p + \epsilon_i \ ; \\ b_k & \sim & N\{0, \sigma_b^2(\kappa_k^m)\}, \ \ k = 1, \ldots, K_m \ ; \\ \epsilon_i & \sim & N\{0, \sigma_\epsilon^2(x_i)\}, \ \ i = 1, \ldots, n \ , \end{array} \right. \tag{1}$$

where b_k and ϵ_i are mutually independent given $(\sigma_\epsilon^2, \sigma_b^2)$, and σ_ϵ^2 and σ_b^2 are smooth functions that will be modeled as logsplines. Denote by $\boldsymbol{X}$ the $n \times (p+1)$ matrix with the ith row equal to $\boldsymbol{X}_i = (1, x_i, \ldots, x_i^p)$ and by $\boldsymbol{Z}$ the $n \times K_m$ matrix with ith row equal to $\boldsymbol{Z}_i = \{(x_i - \kappa_1^m)_+^p, \ldots, (x_i - \kappa_{K_m}^m)_+^p\}$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Model (1) can be rewritten in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \ \ \mathrm{E}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \ \ \mathrm{Cov}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_b & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_\epsilon \end{pmatrix} , \tag{2}$$

where the joint conditional distribution of $\boldsymbol{b}$ and $\boldsymbol{\epsilon}$ given $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_\epsilon$ is assumed normal. The $\boldsymbol{\beta}$ parameters are treated as fixed effects. The covariance matrices $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_\epsilon$ are diagonal with the vectors $\{\sigma_b^2(\kappa_1^m), \ldots, \sigma_b^2(\kappa_{K_m}^m)\}$ and $\{\sigma_\epsilon^2(x_1), \ldots, \sigma_\epsilon^2(x_n)\}$ as main diagonals respectively. For this model $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ and $\mathrm{cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} + \boldsymbol{Z}\boldsymbol{\Sigma}_b\boldsymbol{Z}^T$.

2.1 Error variance function estimation

Ignoring heteroscedasticity may lead to incorrect inferences and inefficient estimation, especially when the response standard deviation varies over several orders of magnitude. Moreover, understanding how variability changes with the predictor may be of intrinsic interest (Carroll, 2003).

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Transformation can stabilize the response variance when the conditional response variance is a function of the conditional expectation (Carroll and Ruppert, 1988), but in other cases one needs to estimate it by modeling the variance as a function of the predictors.

We model the variance function as a loglinear mixed model

$$\begin{cases} \log\{\sigma_\epsilon^2(x_i)\} & = & \gamma_0 + \ldots + \gamma_p x_i^p + \sum_{s=1}^{K_\epsilon} c_s (x_i - \kappa_s^\epsilon)_+^p \\ c_s & \sim & N(0, \sigma_c^2), \quad s = 1, \ldots, K_\epsilon \end{cases} , \tag{3}$$

where the γ are fixed effects and $\kappa_1^\epsilon < \cdots < \kappa_{K^\epsilon}^\epsilon$ are knots. The normal distribution of the c_s parameters shrinks them towards 0 and ensures stable estimation.

2.2 Smoothly varying local penalties

Following Baladandayuthapani et al. (2004), we model $\sigma_b^2(\cdot)$ as

$$\begin{cases} \log\{\sigma_b^2(x)\} & = & \delta_0 + \ldots + \delta_q x^q + \sum_{s=1}^{K_b} d_s (x - \kappa_s^b)_+^q \\ d_s & \sim & N(0, \sigma_d^2), \quad s = 1, \ldots, K_b \end{cases} , \tag{4}$$

where the δ's are fixed effects, the d_s's are mutually independent, and $\{\kappa_s^b\}_{s=1}^{K_b}$ are knots. The particular case of a global smoothing parameter corresponds to the case when the spline function is a constant, that is $\log\{\sigma_b^2(\kappa_s^m)\} = \delta_0$.

Ruppert and Carroll (2000) proposed a "local penalty" model similar to (4). Baladandayuthapani, Mallick and Carroll (2003) developed a Bayesian version of Ruppert and Carroll's (2000) estimator and showed that their estimator is similar to Ruppert and Carroll's in terms of mean square error and outperforms other Bayesian methods. Lang, Fronk and Fahrmeir (2004) consider locally adaptive dynamic models and find that their method is roughly comparable to the method of Ruppert and Carroll (2000) in terms of MSE and coverage probability.

In Lang and Brezger's (2004) prior for the b_k, the nonconstant variance is independent from knot to knot, since the variances of the b_k are assumed to be independent inverse-gammas. In contrast, with our model there is dependence so that if the variance is high at one knot then it is high at neighboring knots. Stated differently, the Lang and Brezger model is one of random heteroscedasticity and ours of systematic heteroscedasticity. Both types of priors are sensible and will have applications, but in any specific application it is likely that one of the two will be better. For example, Lang and Brezger find that for "doppler type" functions, e.g., the severe spatial heterogeneity function in Ruppert and Carroll (2001), their estimator is not quite as good as Ruppert and Carroll's and that the coverage of their credible intervals are not so accurate either. It is not hard to understand why. Doppler type functions that oscillate more slowly going from left to right are consistent with a prior variance for the b_k that is monotonically decreasing, exactly the type of prior included in our model but not Lang and Brezger's.

UCB Division of Biostatistics Collection of Biostatistics Research Archive

http://biostats.bepress.com/jhubiostat/paper61

2.3 Choice of spline basis, number and spacing of knots, and penalties

For concreteness, we have made some specific choices about the spline basis, number and knot spacings, and form of the penalty. In terms of model formulation, the choice of spline basis in the model is not important since an equivalent basis gives the same model and the basis used in computations need not be the same as the one used to express the model. However, spline basis are very different in terms of computational stability. For example, cubic thin plate spline (Ruppert, Wand and Carroll 2003, pp. 72–74) provide much better MCMC convergence and mixing properties than the truncated polynomial basis. When one uses good starting points for the MCMC simulation the truncated polynomial and the cubic thin plate spline bases produced very similar results. In more than one dimension, the tensor product of truncated polynomials proved even more unstable and we preferred using low rank radial smoothers (see Section 7).

The penalty we use is somewhat different than that of Eilers and Marx (1996) and also somewhat different from the penalties used by smoothing splines. We believe that the penalty parameter, not the form of the penalty, is the crucial choice, so we have concentrated on the former, in particular on spatially-adaptive modeling of the penalty parameter.

In this paper we use knots to model the mean, variance and spatially adaptive smoothing parameter. The methodology described here does not depend on the number of knots. For example one could use a knot at each observation for each of the three functions.

Although we use quantile knot spacings, the best type of knot spacings is controversial. Eilers and Marx (1996) state that "Equally-spaced knots are always to be preferred." In contrast, Ruppert, Wand, and Carroll (2003) used quantile-spacing in all of their examples, though they did not make any categorical statement that quantile–spacing is always to be preferred. Although the main focus of this paper is not the choice of knot-spacings, we felt it was necessary to discuss this issue here.

When $\sigma_b^2(x)$ is a constant, then the knot spacings determine the form of the prior on m. To appreciate this, note that the prior for m has a simple form: $m^{(p)}$ is a random walk taking jumps at the knots. The jumps are independent $N(0, \sigma_b^2)$. If the knots are equally spaced, then the prior makes $m^{(p)}$ spatially homogeneous in that the variance of the sum of its jumps in an interval is nearly proportional to the length of the interval. The prior can be viewed as a discrete approximation to the model

$$dm^{(p)}(x) = \sigma dW(x), \tag{5}$$

where W is a standard Wiener process. Model (5) is the prior for smoothing splines (Wahba, 1990). If the knots are not equally spaced, then $m^{(p)}$ changes more rapidly in regions where the x_i (and therefore the knots) are relatively dense so our prior is a discrete approximation

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

to the model

$$dm^{(p)}(x) = \sigma_b f(x) dW(x) \tag{6}$$

where $f(x)$ is some measure of the density of $\{x_1, \ldots, x_n\}$, e.g., is the probability density function of the x_i if they are random and iid. For non-random x_i, $f(x)$ might be a kernel density estimator computed from the x_i. There is no compelling reason to assume spatial homogeneity, that is, model (5), especially since it is not invariant to nonlinear transformations of the x_i. However, there is also no compelling reason to assume that $m^{(p)}$ changes most rapidly where the data are most dense, that is, that (6) holds.

Fortunately, when σ_b^2 is not constant, the knot spacings do not determine the form of the prior because the effect of spacing on the prior can be subsumed into the form of σ_b^2. Stately differently, if σ_b^2 depends on x, then (5) and (6) can both hold but with different $\sigma_b(x)$.

Our conclusion is that quantile knot spacing works well and can be recommended in practice.

2.3.1 A Monte Carlo study of knot spacing

The motivation behind quantile knot spacing is to use more densely spaced knots where there is more information (more data). Suppose that the regression function is spatially heterogenous and, in fact, has more "features" in regions where the data are dense. This is not an unreasonable assumption, especially when the x_i are chosen by design and there is some prior knowledge of where m will have the most "features." In such cases, quantile-spacing allows the penalized spline to fit fine detail where the features occur without undersmoothing elsewhere. The following simulation example illustrates this behavior. The purpose of the example is not to argue that quantile spacing is always to be preferred. We doubt that any type of spacing is always best, and Eilers and Marx (2004) provide an interesting example where equally-spaced knots seem superior to quantile spacing; see their Figure 13. The purpose of the example is, rather, to show that quantile spacing can work better than equal-spacing knots in some problems. In this example, we used non-Bayesian estimators since they are much faster to compute and the relative performance was not expected to depend on whether the estimator was Bayesian or not.

In this example, there are $n = 100$ observations with covariate values $x_i = (i - 1/99)^4$, $i = 1, \ldots, 100$ and response $y_i = 10 \sin\{60x_i/(1 + 3x_i)\} + \epsilon_i$ with $\epsilon_i \sim N(0, 9)$. The x_i are much more dense near 0 than near 1 and the regression function oscillates faster in this region. There were 1000 simulated data sets. For each estimator, the regression function was estimated on a 1000–point grid. The squared error was averaged over this grid and then averaged over the 1000 data sets to produce a MASE (mean average squared error). We used quadratic splines and the number of knots was varied as 10, 15, 20, 30, and 40. There were four estimators. "Equal, not adaptive" is the Eilers and Marx estimator computed using the MATLAB program in Eilers and Marx (1996) with the order of the penalty equal

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

to 2. This estimator uses B-splines and equally spaced knots. The second estimator, called "Quantile, not adaptive," used quantile spacing of the knots and a penalty of the sum of the squared jumps in the second derivative of the spline; this is equivalent to penalizing the sum of the squared coefficients of the truncated power functions. "Quantile, adaptive" and "Equal, adaptive" are the Ruppert and Carroll (2000) estimators with quantile and equal knot spacings, respectively. In all cases the penalty parameter was selected by generalized cross validation. In this example, quantile spacings outperformed equal knot spacings noticeably, especially for non-adaptive penalties; see Figure 1.

This example is an extreme case, as is the example of Eilers and Marx (2004) mentioned previously. We find in the majority of examples that both equal and quantile spacings work perfectly well. There is a long history of success with the quantile spacings, since smoothing splines are a special case of quantile spacing with a knot at each data point.

All univariate examples in this paper will use quantile knot spacing.

3 Prior Specification

Any smoother depends heavily on the choice of smoothing parameter, and for P-splines in a mixed model framework, the smoothing parameter is the ratio of the error variance to the prior variance on the mean (Ruppert, Wand and Carroll, 2003). The smoothness of the fit depends on how these variances are estimated. For example, Crainiceanu and Ruppert (2004) showed that, in finite samples, the (RE)ML estimator of the smoothing parameter is biased towards oversmoothing and Kauermann (2002) obtained corresponding asymptotic results for smoothing splines.

In Bayesian mixed models, the estimates of the variance components are known to be sensitivity to the prior specification, e.g., see Gelman (2004). To study the effect of this sensitivity upon Bayesian P-splines, consider model (1) with one smoothing parameter and homoscedastic errors so that σ_b^2 and σ_ϵ^2 are constant. In terms of the precision parameters $\tau_b = 1/\sigma_b^2$ and $\tau_\epsilon = 1/\sigma_\epsilon^2$, the smoothing parameter is $\lambda = \tau_\epsilon/\tau_b = \sigma_b^2/\sigma_\epsilon^2$ and a small (large) λ corresponds to oversmoothing (undersmoothing).

3.1 Priors on the fixed effects parameters

It is standard to assume that the fixed effects parameters, β_i, are apriori independent, with prior distributions either $[\beta_i] \propto 1$ or $\beta_i \propto N(0, \sigma_\beta^2)$, where σ_β^2 is very large. In our applications we used $\sigma_\beta^2 = 10^6$, which we recommend if x and y have been standardized or at least have standard deviations with order of magnitude one.

For the fixed effects γ and δ used in the log-spline models (3) and (4) we also used independent $N(0, 10^6)$ priors. When this prior is not consistent with the true value of the parameter, a possible strategy is to fit the model using a given set of priors and obtain

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

the estimators $\widehat{\gamma}$, $\widehat{\sigma}_{\widehat{\gamma}}$, $\widehat{\delta}$ and $\widehat{\sigma}_{\widehat{\delta}}$. We could then use independent priors $N(\widehat{\gamma}, 10^6\widehat{\sigma}^2_{\widehat{\gamma}})$ and $N(\widehat{\delta}, 10^6\widehat{\sigma}^2_{\widehat{\delta}})$ for γ and δ respectively.

3.2 Priors on the precision parameters

As just mentioned, the priors for the precisions τ_b and τ_ϵ are crucial. We now show how critically the choice of τ_b may depend upon the scaling of the variables. The gamma family of priors for the precisions is conjugate. If $[\tau_b] \sim \text{Gamma}(A_b, B_b)$ and, independently of τ_b, $[\tau_\epsilon] \sim \text{Gamma}(A_\epsilon, B_\epsilon)$ where $\text{Gamma}(A, B)$ has mean A/B and variance A/B^2, then

$$[\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon] \sim \text{Gamma}\left(A_b + \frac{K_m}{2}, B_b + \frac{||\boldsymbol{b}||^2}{2}\right) \tag{7}$$

and

$$[\tau_\epsilon|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon] \propto \text{Gamma}\left(A_\epsilon + \frac{n}{2}, B_\epsilon + \frac{||\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zb}||^2}{2}\right).$$

Also,

$$\text{E}(\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon) = \frac{A_b + K_m/2}{B_b + ||\boldsymbol{b}||^2/2}, \quad \text{Var}(\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon) = \frac{A_b + K_m/2}{(B_b + ||\boldsymbol{b}||^2/2)^2},$$

and similarly for $\tau_{|\epsilon}$.

The prior does not influence the posterior distribution of τ_ϵ when both A_b and B_b are small compared to $K_m/2$ and $||\boldsymbol{b}||^2/2$ respectively. Since the number of knots is $K_m \geq 1$ and in most problems considered $K_m \geq 5$, it is safe to choose $A_b \leq 0.01$. When $B_b << ||\boldsymbol{b}||^2/2$ the posterior distribution is practically unaffected by the prior assumptions. When B_b increases compared to $||\boldsymbol{b}||^2/2$, the conditional distribution is increasingly affected by the prior assumptions. $\text{E}(\tau_b|\boldsymbol{Y}, \boldsymbol{\beta}, \boldsymbol{b}, \tau_\epsilon)$ is decreasing in B_b so large B_b compared to $||\boldsymbol{b}||^2/2$ correspond to undersmoothing. Since the posterior variance of τ_b is also decreasing in B_b a poor choice of B_b will likely result in underestimating the variability of the smoothing parameter $\lambda = \tau_\epsilon/\tau_b$ causing too narrow confidence intervals for m. The condition $B_b << ||\boldsymbol{b}||^2/2$ shows that the "noninformativeness" of the gamma prior depends essentially on the scale of the problem.

To show the possible severity of these effects consider the LIDAR example in Section 5. We consider model (1) with a global smoothing parameter and homoscedastic error. We used quadratic splines with 30 knots. Figure 2 shows the effect of four mean-one Gamma priors for the precision of the truncated polynomial parameters. The variances of these priors are 10, 10^3, 10^6, and 10^{10} respectively. Obviously, the first two inferences provide severely under smoothed, almost indistinguishable, posterior means. The third graph is much smoother but still exhibits roughness especially in the right hand side of the plot, while the fourth graph displays a pleasing smooth pattern, consistent with our frequentist inference. Using either prior distribution one obtains that the posterior mean of $||\boldsymbol{b}||^2/2$ is of order 10^{-6} to 10^{-5}. This explains why values of B_b larger than 10^{-6} proved inappropriate for this problem.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

The size of $||\boldsymbol{b}||^2/2$ depends upon the scaling of the x and y variables and in the case of the LIDAR data $||\boldsymbol{b}||^2/2$ is small because the standard deviation of x is much larger than the standard deviation of y. If y is rescaled to $a_y y$ and x to $a_x x$, then the regression function becomes $a_y m(a_x x)$ whose p-th derivative is $a_y a_x^p m^{(p)}(a_x x)$ so that $||\boldsymbol{b}||^2/2$ is rescaled by the factor $a_y^2 a_x^{2p}$. Thus, $||\boldsymbol{b}||^2/2$ is particularly sensitive to the scaling of x. The size of $||\boldsymbol{b}||^2/2$ also depends on K_m. The integral of $m^{(p)}$ over the range of x will be approximately $\sum_{k=1}^{K_m} b_k \approx \sqrt{K_m}\sigma_b$, we can expect that $\sigma_b^2 \propto (K_m)^{-1}$ and the smoothing parameter should be proportional to K_m. For the LIDAR data, the GCV chosen smoothing parameter is 0.0095, 0.0205, 0.0440, and 0.0831 for K_m equal to 15, 30, 60, and 120, respectively, so as expected the smoothing parameter approximately doubles as K_m doubles.

Practical experience with LMMs for longitudinal or clustered data should be applied with caution to P-splines. In a mixed effects model $||\boldsymbol{b}||^2$ is an estimator of $K_m\sigma_b^2$. For longitudinal data $K_m\sigma_b^2$ would generally be large because K_m is the number of subjects with constant subject effect variance σ_b^2. As just discussed, for a P-spline $K_m\sigma_b^2$ should be nearly independent of K_m and could be quite small.

Figure 3 presents the same type of results as Figure 2 for Gamma priors for the precision parameter τ_b with the mean held fixed at 10^{-6} and variances equal to 10, 10^3, 10^6, and 10^{10} respectively. These prior distributions have a much smaller effect on the posterior mean of the regression function. The fit seems to be undersmooth when the variance is 10. Clearly, when the variance increases the fit becomes smooth indicating that a value of the variance larger than 10^3 will produce a reasonable fit.

It is sometimes believed that a Gamma(A, B) prior is non-informative if both A and B are sufficiently small. However, such non-informative priors are not flat. To illustrate this, consider a small right neighborhood of zero $I_0 = (0, 10^{-6}]$ and denote by P_A the probability distribution of a Gamma$(1/A, 1/A)$ distribution. Then $P_{10}(I_0) = 0.21$, $P_{10^3}(I_0) = 0.98$, $P_{10^6}(I_0) = 0.99997$, $P_{10^{10}}(I_0) \approx 1$. For this type of distributions the large variance is not due to the "flatness" of the prior but to the extremely rare very large values.

A similar discussion holds true for τ_ϵ but now large B_ϵ corresponds to oversmoothing and τ_ϵ does not depend on the scaling of x. In applications it is less likely that B_ϵ is comparable in size to $||\boldsymbol{Y} - \boldsymbol{X\beta} - \boldsymbol{Zb}||^2$, because the latter is an estimator of $n\sigma_\epsilon^2$. If $\hat{\sigma}_\epsilon^2$ is an estimator of σ_ϵ^2 a good rule of thumb is to use values of B_ϵ smaller than $n\hat{\sigma}_\epsilon^2/100$. This rule should work well when $\hat{\sigma}_\epsilon^2$ does not have an extremely large variance.

Alternative to gamma priors are discussed by, for example, Natarajan and Kass (2000) and Gelman (2004). These have the advantage of requiring less care in the choice of the hyperparameters. However, we find that with reasonable care, the conjugate gamma priors can be used in practice. Nonetheless, exploration of other prior families for P-splines would be well worthwhile, though beyond the scope of this paper.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

4 Simultaneous Credible Bounds

Let $f(\cdot)$ be either $m(\cdot)$, $\log\{\sigma_\epsilon^2(\cdot)\}$, $\log\{\sigma_b^2(\cdot)\}$, or a derivative of order q, $1 \leq q \leq p$, of one of these functions. It is straightforward to use MCMC output to construct simultaneous credible bounds on f over an arbitrary finite interval $[x^1, x^N]$. Typically, x^1 and x^N would be the smallest and largest observed values of x.

Let $x^1 < x^2 < \cdots < x^N$ be a fine grid of points on this interval. Let $E\{f(x^i)\}$ and $SD\{f(x^i)\}$ be the posterior mean and standard deviation of $f(x^i)$ estimated from a MCMC sample. Let M_α be the $(1-\alpha)$ sample quantile of $\max_{1\leq i\leq N}\left|[f(x^i) - E\{f(x^i)\}]/SD\{f(x^i)\}\right|$ computed from the realizations of f in the MCMC sample. Then $I(x^i) = E\{f(x^i)\} \pm M_\alpha SD\{f(x^i)\}$, $i = 1, \ldots, N$, are simultaneous credible intervals. For N large, the upper and lower limits of these intervals can be connected to form simultaneous credible bands. In the remainder of this paper we only report simultaneous credible bounds for the functions and their derivatives. In the examples considered in the following these bounds tend to be roughly 30–50% wider than pointwise credible bounds.

5 The LIDAR example

The data displayed in Figure 6-(a) are taken from Holst et al. (1996), who estimated the concentration of atmospheric mercury measured with LIDAR (Sigrist, 1994). The concentration is proportional to the derivative of the mean (with a negative and known constant of proportionality). Because of sizeable heteroscedasticity, the variance function must be estimated to obtain satisfactory credible intervals for the function and its derivative.

Three increasingly complex models were fit to the data. Model I uses a 30-knot quadratic P-spline for m and assumes that σ_b^2 and σ_ϵ^2 are constant. Model II has the same structure for m and σ_ϵ^2 as Model I but uses a linear log-P-spline with $K_b = 4$ knots to model $\log\{\sigma_b^2\}$. Model III differs from Model II in using a 30-knot quadratic P-spline for $\log\{\sigma_\epsilon^2(x)\}$.

To minimize the scale problems discussed in Section 3 we centered and standardized the covariate. The response was not standardized because its range is between -1 and 0, but in general we recommend standardizing the response. For simplicity we describe only the priors used for Model III. We assumed independent normal priors for the coefficients of the monomials $\beta_i \sim N(0, 10^6)$, $\gamma_i \sim N(0, 10^6)$, $i = 0, \ldots, p$, $\delta_i \sim N(0, 10^6), i = 0, \ldots, q$, and independent inverse Gamma priors for the variance components $\sigma_c^2 \sim \text{IGamma}(10^{-3}, 10^{-3})$ and $\sigma_d^2 \sim \text{IGamma}(10^{-3}, 10^{-3})$. The full conditional distributions are in Section 10. Given our discussion in Section 3 we investigated whether this choice is non-informative enough and found that much smaller values of a and b did not affect inference, probably due to standardization of x.

Figure 6 shows the results for model I. Given the homoscedasticity assumption, the

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

standard deviation function is a constant and the credible intervals do not change from left to right. However, this assumption is contradicted by the data. Because $m'(x, \boldsymbol{\theta}) = \beta_1 + 2\beta_2 x + 2\sum_{k=1}^{K} b_k(x - \kappa_k)_+$ is an explicit function of the parameters, its posterior distribution for each x can be easily estimated from the simulations.

Figure 7 presents the results for model II. While the posterior means $E(m|\boldsymbol{Y})$ from Models I and II are visually indistinguishable, $E(m'|\boldsymbol{Y})$ has a sharper peak and wider 95% credible intervals in Model II which has spatial adaptivity. An obvious question is whether differences are real. A rigorous test exceeds the scope of this paper, but we address a simpler related problem at the end of this section.

Figure 8 displays the same type of results for Model III. Figure 8-(b) shows that the simultaneous credible intervals for m are much narrower for small values of x than for large x, in accordance with the non-constant variance seen in the data. This pattern is not present in Figures 6-(d) and 7-(d). Similar differences were reported by Ruppert, Wand and Carroll (2003) (see Section 14.3) who used a frequentist approach to estimate the error variance, but ignored the possible spatial variability of the smoothing parameter.

Figure 8-(c) shows that the standard deviation of the error process increases nonlinearly with the covariate and the variability around the standard deviation increases as well. As we mentioned, the object of inference is $m'(\cdot)$. Comparing the posterior mean of $m(\cdot)$ for models I–III, one see little difference. However, there are noticable differences in the posterior means of $-m'(\cdot)$. For the local-penalty, Figures 7-(d) and 8-(d), have sharper peaks, which suggests that the global penalty model corresponds to oversmoothing $m'(\cdot)$ in the middle of the covariate range and undersmoothing $m'(\cdot)$ in the lower range. Second, when one accounts for heteroscedasticity, Figure 8-(d), the function $-m'(\cdot)$ is smoother and the 95% credible intervals are much narrower in the lower range of the covariate. This is probably due to the lower variability of the data in that region. Third, the credible intervals in the upper range of the covariate are slightly wider in 8-(d) compared to Figure 7-(d) and much wider compared to Figure 6-(d). These results are different from the results of Ruppert and Carroll, 2000, who did not account for the effect of heteroscedastic errors.

Figure 9 displays for model I with heteroscedastic errors and spatially adaptive smoothing parameter, the posterior mean and credible intervals for the logarithm of the shrinkage parameters, $\log(\sigma_b^2)$. The shape of the posterior mean of $\log(\sigma_b^2)$ suggests that a simpler linear trend may be suitable.

We fit a simplified model where the mean and log-variance of the errors are modeled as quadratic splines with $K_m = K_\epsilon = 30$ knots and $\log\{\sigma_b^2(\kappa_k^b)\} = \delta_0 + \delta_1 \kappa_k^b$. Inference for this model produced plots very similar Figure 8 and are not reported here. To test whether $\delta_1 > 0$ we used $500,000$ simulations to obtain $P(\delta_1 < 0) = 0.04$, indicating that the differences between the posterior means of $-m'(\cdot)$ using a global and a spatially adaptive smoothing parameter might be real because the global smoothing parameter is not supported

Research Archive

Hosted by The Berkeley Electronic Press

by the data.

6 Comparison with other univariate smoothers

Baladandayuthapani et al. (2004) present a comprehensive simulation study of their Bayesian spatially adaptive P–spline model, which is obtained as a particular case of our full model with constant error variance. The results of this study indicate that the method is comparable to or better than several other spatially adaptive Bayesian methods on a variety of data sets.

Since we are unaware of any other estimation methodology that estimates jointly the nonparametric adaptive model for the mean and the nonparametric model for the error process, we will compare our methodology with that of Baladandayuthapani et al. (2004).

Consider the regression model

$$y_i = m(x_i) + \epsilon_i$$

where ϵ_i are independent mean zero normal errors and $m(\cdot)$ is the spatially heterogeneous regression function

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

This function was also considered by Ruppert and Carroll (2000) and Baladandayuthapani et al. (2004) and it is roughly constant between $[0, 0.5]$ and has three sharp peaks at 0.6, 0.75 and 0.9. We consider $n = 1,000$ equally spaced xs on $[0, 1]$ and two different scenarios for the standard error of the error process: (a) $\sigma_\epsilon(x) = 0.5$ for homoscedastic errors and (b) $\sigma_\epsilon(x_i) = 0.5 - 0.8x + 1.6(x - 0.5)_+$. We used 500 simulations from these models and fit two spatially adaptive models: one that uses a constant variance and one that uses a log–P–spline model for the variance. For the mean and log–variance functions we used quadratic P-splines with $K^m = K^\epsilon = 40$ knots. For the log-variance corresponding to the shrinkage process we used $K^b = 4$ knots. We calculated the mean square error for each simulated data set as $\text{MSE} = \sum_{i=1}^n \left\{\widehat{m}(x_i) - m(x_i)\right\}^2/n$, where $\widehat{m}(x)$ is the posterior mean of $m(x)$ using a given model.

For case (a), $\sigma_\epsilon(x) = 0.5$, the two models produced practically indistinguishable MSEs. The first boxplot ("ho-ho") in Figure 4 corresponds to MSEs for the model using homoscedastic errors, while the second boxplot ("ho–he") corresponds to heteroscedastic errors. The average MSE over all x's was AMSE = 0.0054, which is smaller than 0.0061 reported by Baladandayuthapani et al. (2004). The coverage probabilities of the 95% credible intervals were very similar for the two models for each value of the covariate, with a slight advantage in favor of the model using homoscedastic errors. The average coverage probability was 94.7% for the model with homoscedastic error and 93.5% for the model with heteroscedastic errors. These coverage probabilities are very similar to the ones reported by Baladandayuthapani et al. (2004) in their Figure 3.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

For case (b), $\sigma_\epsilon(x_i) = 0.5 - 0.8x + 1.6(x - 0.5)_+$, the heteroscedastic model substantially outperformed the homoscedastic model both in terms of MSE and coverage probabilities. The third boxplot ("he-ho") in Figure 4 corresponds to MSEs for the model using homoscedastic errors with AMSE= 0.003. The fourth boxplot ("he–he") corresponds to heteroscedastic errors with AMSE= 0.0026. In this situation it would be misleading to compare only the average coverage probability for the 95% credible intervals. Indeed, these averages are very close, 94.1% for the heteroscedastic and 93.0% for the homoscedastic method, but they are obtained from very different sets of pointwise coverage probabilities. Figure 5 displays the coverage probabilities for these two methods. Note that for the heteroscedastic method the coverage probability in the $[0, 0.5]$ interval is close to 95%. In the same interval the coverage probability for the homoscedastic method starts from around 0.8 and increases until it crosses the 95% target around 0.2. Moreover, in the interval $[0.3, 0.5]$ this coverage probability is estimated to be 1. This is due to the fact that on this interval $\sigma_\epsilon(x)$ decreases linearly from 0.5 to 0.1. Since the homoscedastic method assumes a constant variance, the credible intervals will tend to be shorter than nominal in regions of higher variability (x close to zero), thus producing lower coverage probabilities. In regions of smaller variability (x close to 0.5) the credible intervals will tend to be much wider, thus producing extremely large coverage probabilities.

In the interval $[0.55, 0.65]$ the homoscedastic method slightly outperforms the heteroscedastic method. This seems to be the effect of a "lucky" combination of two factors. As we discussed, the size of the credible intervals for the homoscedastic method in a neighborhood of 0.5 is much larger than nominal. However, at the same point the mean function changes from a constant to a rapidly oscillating function and the coverage probabilities drop roughly at the same rate. In the interval $[0.8, 1]$ one can notice a phenomenon very similar to the one described for the interval $[0, 0.5]$. While the function oscillates more rapidly in this region, the heteroscedastic adaptive method produces credible intervals with coverage probabilities close to the nominal 95%. However, the homoscedastic method does not take into account the increased variability and produces credible intervals that are too short.

For this example both methods performed roughly similar on simulated data sets that did not require variance estimation. However, when the error variance is not constant the method using log–P–splines to estimate the error variance considerably outperforms the method that does not.

These simulations required approximately $1,000$ hours of were very computationally intensive, with the heteroscedastic method requiring roughly four times as much time.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

7 Low rank multivariate smoothing

In this section we generalize the ideas in Section 2 to multivariate smoothing while preserving the appealing geometric interpretation of the smoother. Consider the following regression $y_i = m(\boldsymbol{x}_i) + \epsilon_i$ where $m(\cdot)$ is a smooth function of L covariates. We will use radial basis functions which have the advantage of being independent of rotations. Suppose that $\boldsymbol{x}_i \in \mathbb{R}^L$, $1 \leq i \leq n$ are n vectors of covariates and $\kappa_k \in \mathbb{R}^L$, $1 \leq k \leq K_m$ are K_m knots. Consider the following distance function

$$C(\boldsymbol{r}) = \begin{cases} ||\boldsymbol{r}||^{2M-L} & \text{for } L \text{ odd} \\ ||\boldsymbol{r}||^{2M-L} \log ||\boldsymbol{r}|| & \text{for } L \text{ even} \end{cases},$$

where $||\cdot||$ denotes the Euclidean norm in $\mathbb{R}^L$, the integer M controls the smoothness of $C(\cdot)$, $\boldsymbol{X}$ the matrix with ith row $\boldsymbol{X}_i = [1 \ \boldsymbol{x}_i^T]$, $\boldsymbol{Z}_{K_m} = \{C(||\boldsymbol{x}_i - \boldsymbol{\kappa}_k||)\}_{1 \leq i \leq n, 1 \leq k \leq K_m}$, $\boldsymbol{\Omega}_{K_m} = \{C(||\boldsymbol{\kappa}_k - \boldsymbol{\kappa}_{k'}||)\}_{1 \leq k \leq K, 1 \leq k' \leq K}$ and define $\boldsymbol{Z} = \boldsymbol{Z}_K \boldsymbol{\Omega}_K^{-1/2}$, where $\boldsymbol{\Omega}_K^{-1/2}$ is the principal square root of $\boldsymbol{\Omega}_K$. With these notations the low rank approximation of thin plate spline regression can be obtained as the BLUP in the LMM (Kamman and Wand, 2003; Ruppert, Wand and Carroll, 2003)

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \ \mathrm{E}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \ \mathrm{Cov}\begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{\epsilon} \end{pmatrix} = \begin{pmatrix} \sigma_b^2 \boldsymbol{I}_{K_m} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\epsilon^2 \boldsymbol{I}_n \end{pmatrix}. \quad (8)$$

This model contains only one variance component, σ_b^2, for controlling the shrinkage of $\boldsymbol{b}$, which is equivalent to one global smoothing parameter $\lambda = \sigma_b^2/\sigma_\epsilon^2$ and implicitly assumes homoscedastic errors. To relax these assumptions, we consider a new set of knots $\{\boldsymbol{\kappa}_1^*, \ldots, \boldsymbol{\kappa}_{K_b}^*\}$ and define $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$ similarly with the corresponding definition of matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ where the $\boldsymbol{x}$-covariates are replaced by the knots $\boldsymbol{\kappa}_k$ and the knots are replaced by the subknots $\boldsymbol{\kappa}_k^*$. Consider the following model for the response

$$\begin{cases} y_i &= \beta_0 + \sum_{j=1}^L \beta_j x_{i,j} + \sum_{k=1}^{K^m} b_k z_{i,k} + \epsilon_i \\ b_k &\sim N\{0, \sigma_b^2(\boldsymbol{\kappa}_k)\}, \quad k = 1, \ldots, K_m \\ \epsilon_i &\sim N\{0, \sigma_\epsilon^2(\boldsymbol{x}_i)\}, \quad i = 1, \ldots, n \end{cases}, \quad (9)$$

where the error variance, $\log\{\sigma_\epsilon^2(\boldsymbol{x}_i)\}$, is modeled as a low rank log-thin plate spline

$$\begin{cases} \log\{\sigma_\epsilon^2(\boldsymbol{x}_i)\} &= \gamma_0 + \sum_{j=1}^L \gamma_j x_{i,j} + \sum_{k=1}^{K_\epsilon} c_k z_{i,k} \\ c_k &\sim N(0, \sigma_c^2), \quad k = 1, \ldots, K_\epsilon \end{cases}, \quad (10)$$

and the random coefficient variance, $\log\{\sigma_b^2(\boldsymbol{\kappa}_k)\}$, is modeled as another low rank log-thin plate spline

$$\begin{cases} \log\{\sigma_b^2(\boldsymbol{\kappa}_k)\} &= \delta_0 + \sum_{j=1}^L \delta_j x_{k,j}^* + \sum_{j=1}^{K_b} d_j z_{k,j}^* \\ d_j &\sim N(0, \sigma_d^2), \quad j = 1, \ldots, K_b \end{cases}, \quad (11)$$

where $x_{k,j}^*$ and $z_{k,j}^*$ are the entries of $\boldsymbol{X}^*$ and $\boldsymbol{Z}^*$ matrices respectively and c_k and d_s are assumed mutually independent.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

In the case of low rank smoothers the set of knots for the covariates and subknots for modeling the shrinkage process have to be chosen. Our approach is to use a fixed set of knots using the *space filling* design (Nychka and Saltzman, 1998), which is based on the maximal separation principle. This design avoids wasting knots and is likely to lead to better approximations in sparse regions of the data. The `FUNFITS` module (Nychka et al. 1998) provides software for space filling knot selection. The program can be slow for large n and K_m and a remedy is to apply the algorithm to a random sample of the $\boldsymbol{x}$'s. The set of subknots can be obtained by applying the space filling algorithm with the covariates replaced by the knots from the previous stage. Given the relatively small number of subknots the algorithm for choosing the subknots does not present the same computational challenges.

8 The Noshiro example

This example comes from Ruppert (1997). Noshiro, Japan was the site of a major earthquake. Much of the damage from the quake was due to soil movement. Professor Thomas O'Rourke, a geotechnical engineer, was investigating the factors that might help predict soil movement during a future quake. One factor thought to be of importance was the slope of the land. To estimate the slope, Ruppert (1997) used a data set with 799 observations where $\boldsymbol{x}$ was longitude and latitude and y was elevation at locations in Noshiro. The object of primary interest is the gradient, at least its magnitude and possibly its direction.

We used the thin-plate spline models (9) for the mean of $\boldsymbol{y}$ with $K_m = 100$ knots equally spaced on a rectangular grid in the $[0,1]\times[0,1]$ square. The log-thin plate spline described in (10) with the same $K_\epsilon = 100$ knots was used to model the variance of the error process $\sigma^2_\epsilon(\boldsymbol{x}_i)$. The log-thin plate spline described in (11) with $K_b = 16$ equally spaced knots was used to model the variance $\sigma^2_b(\boldsymbol{\kappa}_k)$ which controls the amount of shrinkage of the b parameters.

Figure 10 displays the posterior mean of the mean response function. There is a rather sharp peak around $(0.42, 0.4)$ and the function displays a slow decay in the general south direction with a much sharper decay in all other directions. The function seems much smoother towards the boundary. This is exactly the type of function for which adaptive spatial smoothing can substantially improve the fit.

Figure 11 shows that in large areas of the map $\sigma_\epsilon(\cdot)$ is smooth and has very small values. However, in a neighborhood of the peak of the mean function, $\sigma_\epsilon(\cdot)$ displays two relatively sharp maxima located N-N-W and S-S-E of the peak. Severe heteroscedasticity can also be noted near the eastern boundary of the map, especially in the N-E and S-E. Another area displaying heteroscedasticity is the S-W corner of the map. To check whether these characteristics of the function are present in the data we also performed a two-stage frequentist analysis. In the first stage we used a low rank thin plate spline for the

CORE
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

mean function and the results from this regression were used to obtain residuals. We then fitted another low rank thin plate spline to the absolute values of these residuals. While this map was not identical to the one in Figure 11, it did exhibit the same patterns of heteroscedasticity.

The process $\log\{\sigma_b^2(\boldsymbol{\kappa}_k)\}$ controlling the shrinkage of the b_k is displayed in Figure 12. Smaller values of this function correspond to less shrinkage of b_k towards zero and more local behavior of the smoother. Figure 12 indicates that in a neighborhood of the peak of the mean function the shrinkage is smaller to allow the mean function to change rapidly. Away from the peak, the shrinkage is larger corresponding to a smoother mean. If a fixed smoothing parameter were used, then the representation of its posterior mean in Figure 12 would be the hyperplane $Z = -4.71$.

To provide a better understanding of the results we posted 4 movies on the website `www.people.cornell.edu/pages/cmc59/moviefiles/`. The first 3 movies show the posterior mean of the mean, standard deviation and shrinkage functions. The 4th movie displays various realizations from the posterior distribution of the mean function.

9 Comparisons with other adaptive surface fitting methods

Lang and Brezger (2004) compared their adaptive surface smoothing method with several methods used in a simulation study by Smith and Kohn, 1997: MARS (Friedman, 1991), "locfit" (Cleveland and Grosse, 1991), "tps" (bivariate cubic thin plate splines with a single smoothing parameter), tensor product cubic smoothing splines with five smoothing parameters, and a parametric linear interaction model. We will use the following regression functions, also used by Lang and Brezger

- $f_1(x_1, x_2) = x_1 \sin(4\pi x_2)$ where x_1 and x_2 are distributed independently uniform on $[0, 1]$.
- $f_2(x_1, x_2) = 1/5 \exp(-8x_1^2) + 3/5 \exp(-8x_2^2)$ where x_1 and x_2 are distributed independently normal with mean 0.5 and variance 0.1.

Function $f_1(\cdot)$ correspond to models with interactions and $f_2(\cdot)$ has only main effects. Function $f_1(\cdot)$ has moderate spatial variability, while function $f_2(\cdot)$ is much smoother. We used a sample size $n = 300$, $\sigma = 1/4\text{range}(f)$ and 250 simulations from each model.

Our model is more general than the models considered by Lang and Brezger (2004) because it incorporates simultaneous nonparametric estimation of the error variance. To assess the performance of the spatially adaptive component of our model we modeled the mean function as a low rank thin–plate spline with an equally–spaced 12×12 knot grid. We

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

modeled the smoothing parameter using a low rank log-thin plate spline with an equally–spaced 5×5 knot grid. We considered two estimators, one with constant error variance (CEV) and one with estimated error variance using a log–thin plate spline (EEV) with the same 12×12 knot grid as for the mean function. and a constant error variance. The CEV estimator is the one that can directly be compared with the estimators considered by Lang and Brezger (2004). One expects that the CEV outperforms the EEV estimator when the true mean error process is homoscedastic and is outperformed by EEV when it is heteroscedastic. We investigate this in our simulations and we compare the performance of our estimators with the performance of estimators considered by Lang and Brezger (2004).

The performance of all estimators was measured by the empirical mean squared error given by $\text{MSE}(\hat{f}) = 1/n \sum_{i=1}^{n} \{f(x_i) - \hat{f}(x_i)\}^2$ and we compared log(MSE) for our method with the values reported by Lang and Brezger (2004).

For the CEV estimator and function $f_1(\cdot)$ corresponding to moderate spatial heterogeneity we obtained a median log(MSE) of -3.67 with an interquartile range $[-3.80, -3.53]$ and a range $[-4.21, -3.13]$. These values are better than the ones reported by Lang and Brezger's methods in their Figure 5–b) and are also better than all the other methods considered in the comparative study of Lang and Brezger. Remarkably, the EEV estimator performed almost as well as the CEV estimator in this case and outperformed all the other methods considered. The median log(MSE) for EEV was of -3.59 with an interquartile range $[-3.74, -3.43]$ and a range $[-4.14, -3.02]$. These results are consistent with univariate results reported by Ruppert and Carroll (2000), Baladandayuthapani, Mallick and Carroll (2004) and Lang and Brezger (2004) who found that allowing the smoothing parameter to vary smoothly outperforms other adaptive techniques when the function requires adaptive smoothing.

For function $f_1(\cdot)$ figure 13 displays the coverage probabilities for the 95% credible intervals calculated over a 20×20 equally spaced grid in $[0, 1]^2$. For one grid point we calculated the frequency with which the 95% pointwise credible interval covers the true value of the function at the grid point. As expected, these coverage probabilities show strong spatial correlation with lower coverage probabilities along the ridges of the sinus function. Coverage probability is lowest when x_1 is in the $[0.2, 0.5]$ range. The signal–to–noise in this region is about half the signal–to–noise ratio in the region corresponding to x_1 close to 1. This explains why the coverage probability is smaller in this region. Another region with high coverage probability is $x_1 < 0.15$ which corresponds to high degree of attenuation of the sinus function. Interestingly, the zeros of the true function appearing at $x_2 = 0.25,\ 0.50,\ 0.75$ are covered with at least 95% probability. Another interesting feature is the lower coverage probabilities near the north, south and east boundaries. These features of the pointwise coverage probability map are not consistent with the coverage probabilities reported by Lang and Brezger for their Bayesian P–spline method.

UCLA
Biostatistics
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

For our CEV estimator and function $f_2(\cdot)$ corresponding to very low spatial heterogeneity we obtained a median log(MSE) of -5.95 with an interquartile range $[-6.12, -5.66]$ and a range $[-6.51, -5.41]$. We compare this results with the results reported by Lang and Brezger in their Figure 5–a). In this case our method performs roughly similar to the Bayesian P–spline method of Lang and Brezger and to the cubic thin plate spline (tps), being outperformed only by Lang and Brezger's adaptive P–spline method with two main effects. We could also add two main effects to our bivariate adaptive smoother to improve MSE for this example, but this is not our concern in this paper. Again, the EEV estimator performed very similarly to our CEV estimator.

The last simulation study was done using the function $f_1(\cdot, \cdot)$ where x_1, x_2 are independent uniformly distributed in $[0, 1]$ with an error standard deviation function

$$\sigma_\epsilon(x_1, x_2) = \frac{r}{32} + \frac{3r}{32} x_1^2 \, ,$$

where $r = \text{range}(f_1)$. We compared our CEV and EEV estimators described at the beginning of this section and, as expected, the EEV estimator outperformed the CEV both in term on log(MSE) and coverage probabilities. More precisely for log(MSE) we obtained a median of -5.26 for CEV and -5.35 for EEV, an interquartile range $[-5.37, -5.15]$ for CEV and $[-5.48, -5.24]$ for EEV, and a range $[-4.65, -5.72]$ for CEV and $[-4.84, -5.84]$ for EEV. For both estimators the general patterns for coverage probabilities were the ones displayed in Figure 13. EEV outperformed CEV in terms of coverage probabilities. For example, the regions where the nominal level is exceeded for both estimators have a roughly similar shape and location, but the EEV coverage probabilities tend to be closer to their nominal level.

The simulation studies reported here required more than $2,000$ hours of computation time.

10 Implementation using MCMC

We will now provide some details for MCMC simulations of model (1), where the variances $\sigma_\epsilon^2(x_i)$ and $\sigma_b^2(\kappa_k)$ are modeled by equations (3) and (4). The implementation for other models, e.g., for the multivariate smoothing in Section 7, is similar. Consider independent normal priors for the coefficients of the monomials: $\beta_i \sim N(0, \sigma_{0,\beta}^2)$, $\gamma_i \sim N(0, \sigma_{0,\gamma}^2)$, $i = 0, \ldots, p$, $\delta_i \sim N(0, \sigma_{0,\delta}^2)$, $i = 0, \ldots, q$, and independent inverse Gamma priors for the variance components: $\sigma_c^2 \sim \text{IGamma}(a_c, b_c)$ and $\sigma_d^2 \sim \text{IGamma}(a_d, b_d)$. Using these priors many full conditionals of the posterior distribution are easy to derive, while a few have complex multivariate forms. Our implementation of the MCMC using multivariate Metropolis-Hastings steps proved to be unstable with poor mixing properties. A simple and reliable solution was to change the model by adding error terms to the log-spline models, that is

$$\begin{cases} \log\{\sigma_\epsilon^2(x_i)\} &= \gamma_0 + \ldots + \gamma_p x_i^p + \sum_{s=1}^{K_\epsilon} c_s (x_i - \kappa_s^\epsilon)_+^p + u_i \\ \log\{\sigma_b^2(\kappa_k^m)\} &= \delta_0 + \ldots + \delta_q (\kappa_k^m)^q + \sum_{s=1}^{K_b} d_s (\kappa_k^m - \kappa_s^b)_+^q + v_k \end{cases} , \tag{12}$$

http://biostats.bepress.com/jhubiostat/paper61

where $u_i \sim N(0, \sigma_u^2)$ and $v_k \sim N(0, \sigma_v^2)$. This idea was also used for σ_b^2 by Baladandayuthapani, Mallick, and Carroll (2004). We fixed the values of $\sigma_u^2 = \sigma_v^2 = 0.01$, as these variances appear not identifiable or only weakly identifiable and a standard deviation of 0.1 is small on a log-scale. This device reduces the computational costs because one can now use univariate MH steps to simulate from complex full conditionals.

Define as before by $\boldsymbol{\Sigma}_\epsilon$, $\boldsymbol{\Sigma}_b$, $\boldsymbol{\theta}_X$, and define $\boldsymbol{\theta}_\epsilon = (\boldsymbol{\gamma}^T, \boldsymbol{c}^T)^T$, $\boldsymbol{\theta}_b = (\boldsymbol{\delta}^T, \boldsymbol{d}^T)^T$, $\boldsymbol{C}_X = (\boldsymbol{X}\ \boldsymbol{Z})$, $\boldsymbol{C}_\epsilon = (\boldsymbol{X}_\epsilon\ \boldsymbol{Z}_\epsilon)$, $\boldsymbol{C}_b = (\boldsymbol{X}_b\ \boldsymbol{Z}_b)$, where $\boldsymbol{X}$, $\boldsymbol{X}_\epsilon$, $\boldsymbol{X}_b$ contain the monomials and $\boldsymbol{Z}$, $\boldsymbol{Z}_\epsilon$, $\boldsymbol{Z}_b$ contain the truncated polynomials of the spline models for m, $\log\{\sigma_\epsilon^2\}$, and $\log\{\sigma_b^2\}$. respectively. Also, denote by

$$\boldsymbol{\Sigma}_{0X} = \begin{bmatrix} \sigma_{0,\beta}^2 \boldsymbol{I}_{p+1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_b \end{bmatrix} \boldsymbol{\Sigma}_{0\epsilon} = \begin{bmatrix} \sigma_{0,\gamma}^2 \boldsymbol{I}_{p+1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_c^2 \boldsymbol{I}_{K_\epsilon} \end{bmatrix} \boldsymbol{\Sigma}_{0b} = \begin{bmatrix} \sigma_{0,\delta}^2 \boldsymbol{I}_{q+1} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_d^2 \boldsymbol{I}_{K_m} \end{bmatrix}.$$

The full conditionals of the posterior are detailed below

1. $[\boldsymbol{\theta}_X] \sim N\left(\boldsymbol{M}_X \boldsymbol{C}_X^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{Y}, \boldsymbol{M}_X\right)$, where $\boldsymbol{M}_X = \left(\boldsymbol{C}_X^T \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{C}_X + \boldsymbol{\Sigma}_{0X}^{-1}\right)^{-1}$.

2. $[\boldsymbol{\theta}_\epsilon] \sim N\left(\boldsymbol{M}_\epsilon \boldsymbol{C}_\epsilon^T \boldsymbol{Y}_\epsilon / \sigma_u^2, \boldsymbol{M}_\epsilon\right)$, where $\boldsymbol{Y}_\epsilon = [\log(\sigma_{\epsilon,1}^2), \ldots, \log(\sigma_{\epsilon,n}^2)]^T$ and $\boldsymbol{M}_\epsilon = (\boldsymbol{C}_\epsilon^T \boldsymbol{C}_\epsilon / \sigma_u^2 + \boldsymbol{\Sigma}_{0\epsilon}^{-1})^{-1}$.

3. $[\boldsymbol{\theta}_b] \sim N\left(\boldsymbol{M}_b \boldsymbol{C}_b^T \boldsymbol{Y}_b / \sigma_v^2, \boldsymbol{M}_b\right)$ where $\boldsymbol{Y}_b = \left[\log(\sigma_1^2), \ldots, \log(\sigma_{K_m}^2)\right]^T$ and $\boldsymbol{M}_b = (\boldsymbol{C}_b^T \boldsymbol{C}_b / \sigma_v^2 + \boldsymbol{\Sigma}_{0b}^{-1})^{-1}$.

4. $[\sigma_c^2] \sim \text{IGamma}\left(a_c + K_\epsilon/2, b_c + ||\boldsymbol{c}||^2/2\right)$.

5. $[\sigma_d^2] \sim \text{IGamma}\left(a_d + K_b/2, b_d + ||\boldsymbol{d}||^2/2\right)$.

6. $[\sigma_{\epsilon,i}^2] \propto \sigma_{\epsilon,i}^{-3} \exp\left\{-(y_i - \mu_i)^2/(2\sigma_{\epsilon,i}^2) - \left[\log(\sigma_{\epsilon,i}^2) - \eta_i\right]^2/(2\sigma_u^2)\right\}$, where μ_i and η_i are the ith components of $\boldsymbol{C}_X \boldsymbol{\theta}_X$ and $\boldsymbol{C}_\epsilon \boldsymbol{\theta}_\epsilon$ respectively.

7. $[\sigma_k^2] \propto \sigma_k^{-3} \exp\left\{-b_k^2/(2\sigma_k^2) - \left[\log(\sigma_k^2) - \zeta_k\right]^2/(\sigma_v^2)\right\}$, where ζ_k is the kth component of $\boldsymbol{C}_b \boldsymbol{\theta}_b$.

All the above conditionals have an explicit form with the exception of the $n + K_m$ one-dimensional conditionals from 6. and 7. For these distributions we use the Metropolis-Hastings algorithm with a normal proposal distribution centered at the current value and small variance.

An appealing feature of our methodology is that it can be implemented in high-level Bayesian software such as WinBUGS. Simulations implemented in WinBUGS and MATLAB gave similar results. However, in the simulation study in Section 6 WinBUGS produced credible intervals with lower coverage probability. This is probably due to its sampling inefficiency when parameters are very highly correlated. Programs for our two examples are available on the website `http://www.people.cornell.edu/pages/cmc59/adaptiveprograms/`. For

CORRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

the full model for both our univariate examples, we obtained more than 100 simulations per second from the target distribution (3.4GHz CPU, 3.4Gb RAM PC). We discarded the first $20,000$ burn-in simulations and used $100,000$ additional simulations from the target distribution for our inferences. For one model and one data set this took approximately 20 minutes of computation time. For the Noshiro example with a full model we obtained approximately 3 simulations per second. we discarded the first $4,000$ simulations and used $10,000$ additional simulations from the target distribution for our inference. This took approximately 3 hours of computation time.

In complex models the amount of simulation needed for accurate estimation of the posterior depends on the parameters monitored. In the LIDAR case, accurate inference for the parameters modeling $\log\{\sigma_b^2(\cdot)\}$ requires tens of millions of simulations whereas the mean function only requires several thousand simulations. This seems due to some parameters being highly correlated or only very weakly identified from the data. Another issue is possible multimodality of the posterior. In a few very long runs we have noted that parameters in $\log(\sigma_b^2)$ tend to shift between 2–3 mean levels, suggesting the need for longer runs if these are more than nuisance parameters. Fortunately, the estimates and credible intervals for $m(x)$ do not change much with these changes in $\log(\sigma_b^2)$.

Acknowledgements

Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

References

Baladandayuthapani, V., Mallick, B.K., and Carroll, R.J. (2004). Spatially Adaptive Bayesian Regression Splines, *J. of Comp. and Graph. Statist.*, to appear.

Brumback, B., Ruppert, D. and Wand, M.P., (1999). Comment on Variable selection and function estimation in additive nonparametric regression using data-based prior by Shively, Kohn, and Wood. *J. Am. Statist. Assoc.*, 94, 794–797.

Carroll, R.J. (2003). Variances are not always nuisance parameters: The 2002 R. A. Fisher Lecture. *Biometrics*, 59, 211–220.

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J.R. Statist. Soc. – Series B*, 66(1), 165–185.

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

Crainiceanu, C.M., Ruppert, D., and Wand, M.P. (2004). Bayesian analysis for penalized spline regression using WinBUGS, *submitted*.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B–plines and penalties (with discussion). *Statist. Sci.*, 11, 89–121.

Eilers, P.H.C. and Marx, B.D. (2004). Splines, Knots, and Penalties. Manuscript.

Friedman, J.H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, 19, 1–141.

Gelman, A. (2004). Prior distributions for variance parameters in hierarchical models, manuscript.

Gray, R.J. (1994). Spline-based tests in survival analysis. *Biometrics*, 50, 640–652.

Härdle, W. (1990), *Applied Nonparametric Regression*. Cambridge University Press, New York.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. and Edner, H. (1996). Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of LIDAR measurements, *Environmetrics*, 7, 401–416.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Appl. Statist.*, 52, 1–18.

Kauermann, G. (2002). A note on bandwidth selection for penalised spline smoothing. *Technical Report 02-13*, Department of Statistics, University of Glasgow

Lang, S., and Bretzger, A. (2004). Bayesian P-splines. *J. of Comp. and Graph. Statis.*, 13, 183–212.

Lang, S., Fronk, E.M. and Fahrmeir, L. (2004). Function estimation with locally adaptive dynamic models, *to appear*.

Natarajan, R., and Kass, R.E. (2000), Reference Bayesian methods for generalized linear mixed models, *J. of the Amer. Statist. Assoc.,* 95, 227-237.

Ngo, L. and Wand, M.P., (2004). Smoothing with mixed model software, *J. Statist. Software*, 9.

Nychka, D., Haaland, P., O'Connell, M., and Ellner, S. (1998). FUNFITS, data analysis and statistical tools for estimating functions. In D. Nychka, W.W. Piegorsch and L.H. Cox (Eds.) *Case studies in Environmental Statistics* (Lecture Notes in Statistics, vol. 132), pp. 159–179. New York: Springer Verlag.
`http://www.cgd.ucar.edu/stats/Software/Funfits/`

Nychka, D., and Saltzman, N. (1998). Design of air quality monitoring networks. In D. Nychka, W.W. Piegorsch and L.H. Cox (Eds.) *Case studies in Environmental Statistics* (Lecture Notes in Statistics, vol. 132), pp. 51–76. New York: Springer Verlag.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

Ruppert, D. (1997). Local polynomial regression and its applications in environmental statistics, In *Statistics for the Environment*, Volume 3, Barnett, V., and Turkman, F., ed., pp. 155–173, Chicester: John Wiley.

Ruppert, D., (2002). Selecting the number of knots for penalized splines. *J. of Comp. and Graph. Statis.*, 11, 735–757.

Ruppert, D. and Carroll, R.J., (2000). Spatially-adaptive penalties for spline fitting. *Aust. and New Zeal. J. of Statistics*, 42(2), 205–223

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003) *Semiparametric Regression*. Cambridge University Press, Cambridge.

Sigrist, M. (ed.), (1994). Air Monitoring by Spectroscopic Techniques (Chemical Analysis Series, Vol. 127), Wiley, New York.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non–parametric regression curve fitting, *J.R. Statist. Soc. – Series B*, 47(1), 1–52.

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

Wand, M.P. (2000). A comparison of regression spline smoothing procedures. *Comp. Statist.*, 15, 443–462.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

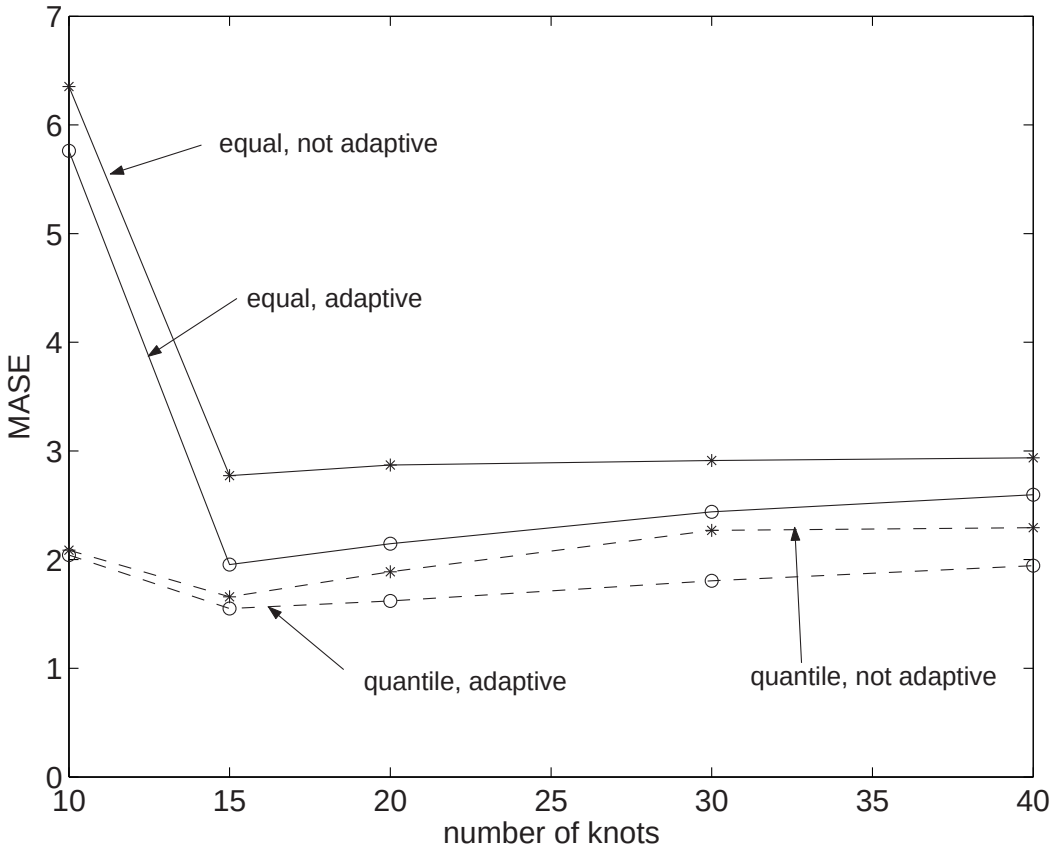

Figure 1: MASE for non-Bayesian estimators with non-adaptive penalties and equal (EM) or quantile (QS) knot spacings. MASE is based on 1000 simulations from the model $Y_i = 10\sin\{60x_i/(1+3x_i)\} + \epsilon_i$ with $\epsilon_i \sim N(0,9)$ where the covariate values are $x_i = (i - 1/99)^4$, $i = 1, \ldots, 100$. The number of knots was varied and equalled 10, 15, 20, 30, and 40.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

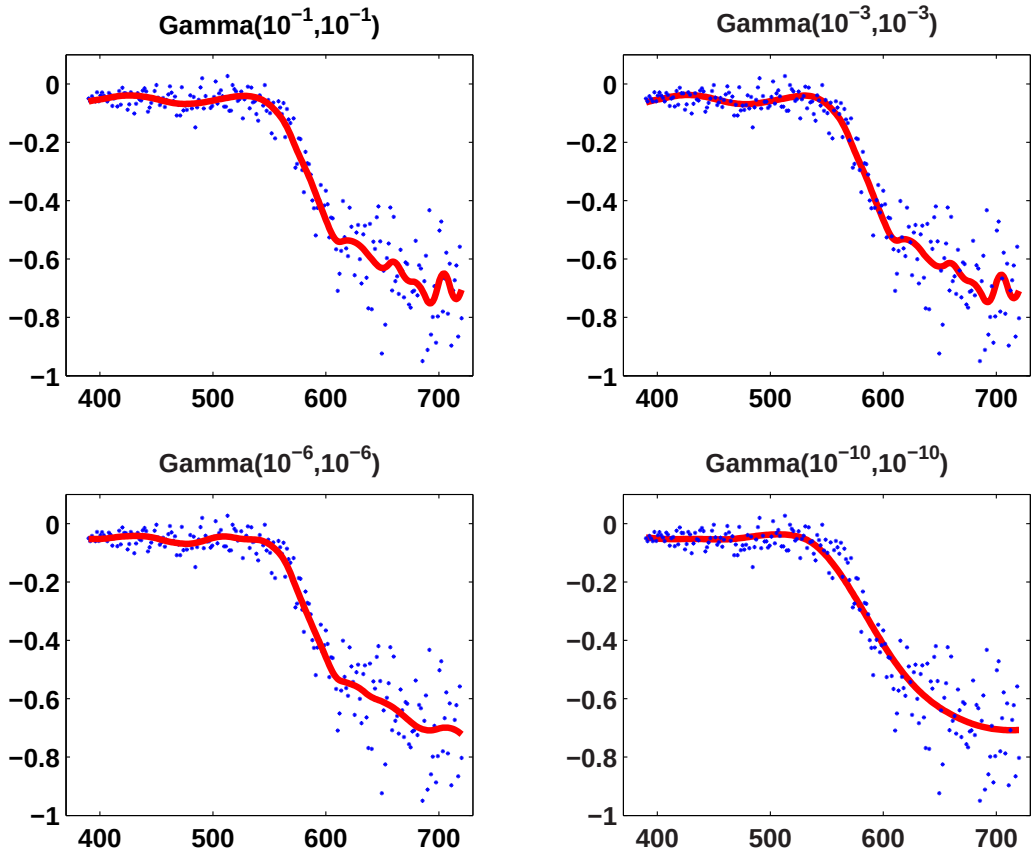

Figure 2: LIDAR data with unstandardized covariate: effect of four mean–one Gamma priors for the precision of the truncated polynomial parameters. The variances of these priors are 10, 10^3, 10^6, and 10^{10} respectively.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

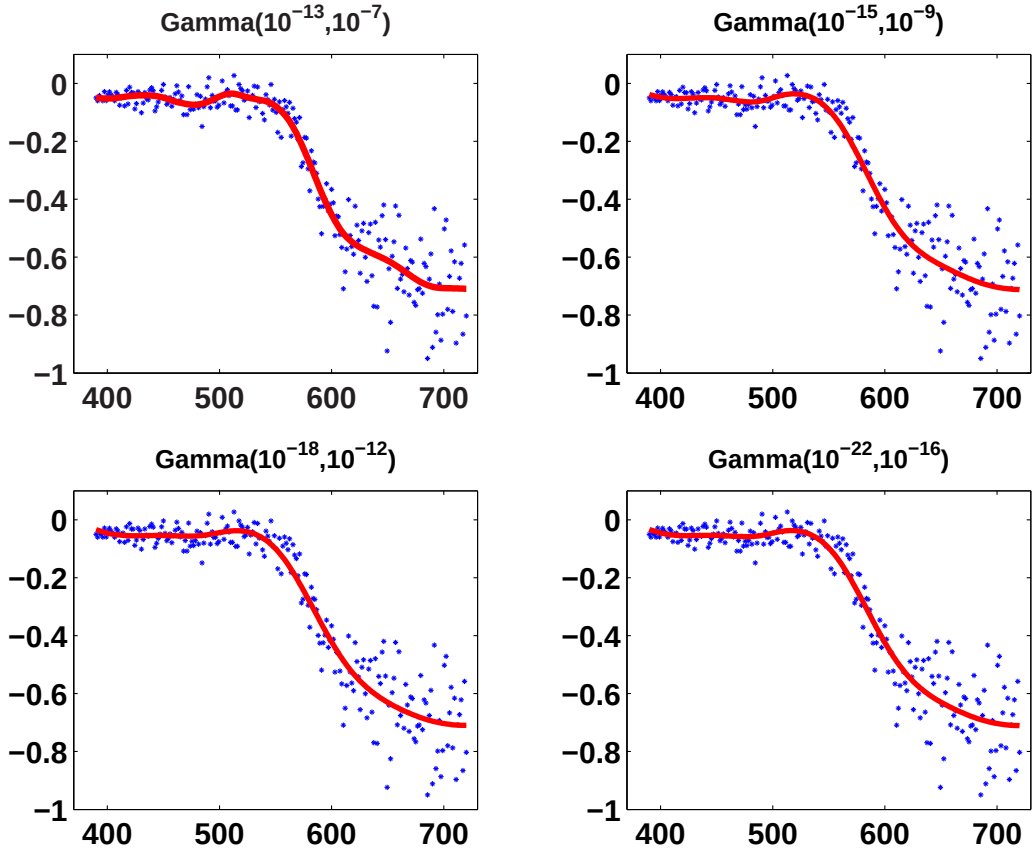

Figure 3: LIDAR data with unstandardized covariate: effect of four Gamma priors with mean 10^{-6} for the precision of the truncated polynomial parameters. The variances of these priors are 10, 10^3, 10^6, and 10^{10} respectively.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

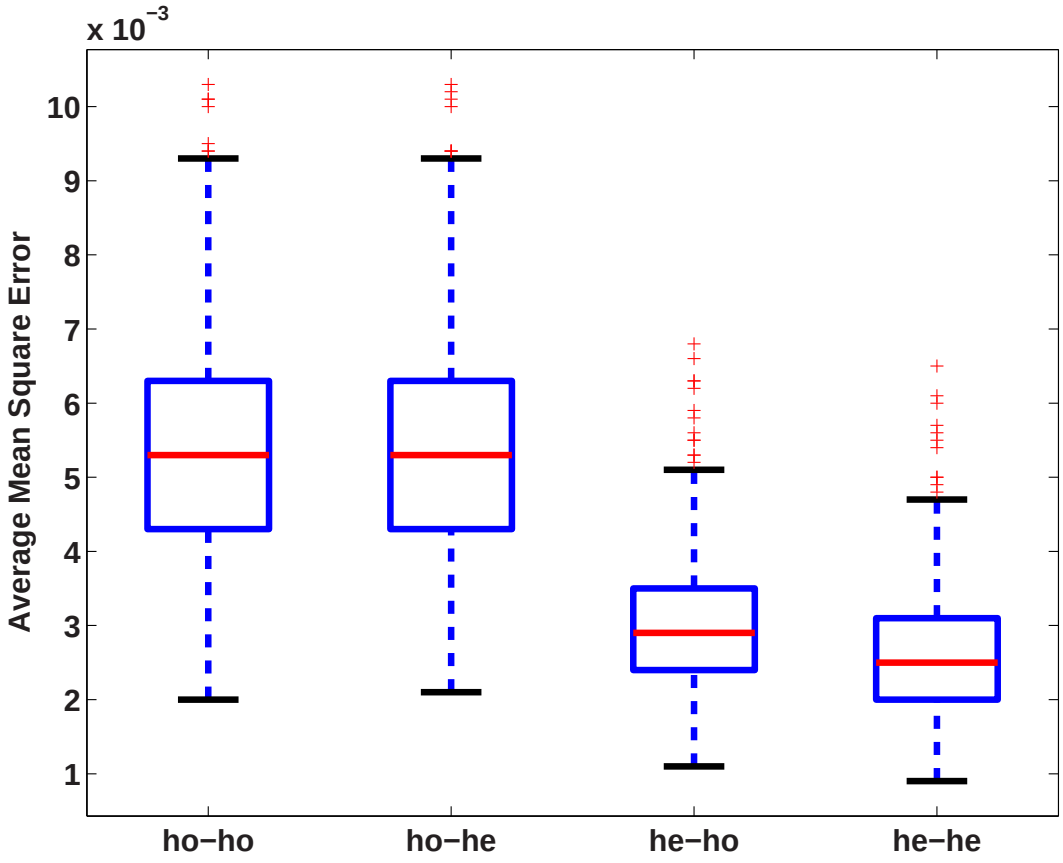

Figure 4: Mean Square Error based on 500 simulations from the models described in Section 6. The mean function was the same for each simulation study. The labels describe the combination of methods used for simulation and inference. For example, "ho–he" corresponds to homoscedastic errors for the simulation model and heteroscedastic errors used for the estimation model. Only the first and the last two boxplots are comparable because they correspond to the same model used for simulations.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

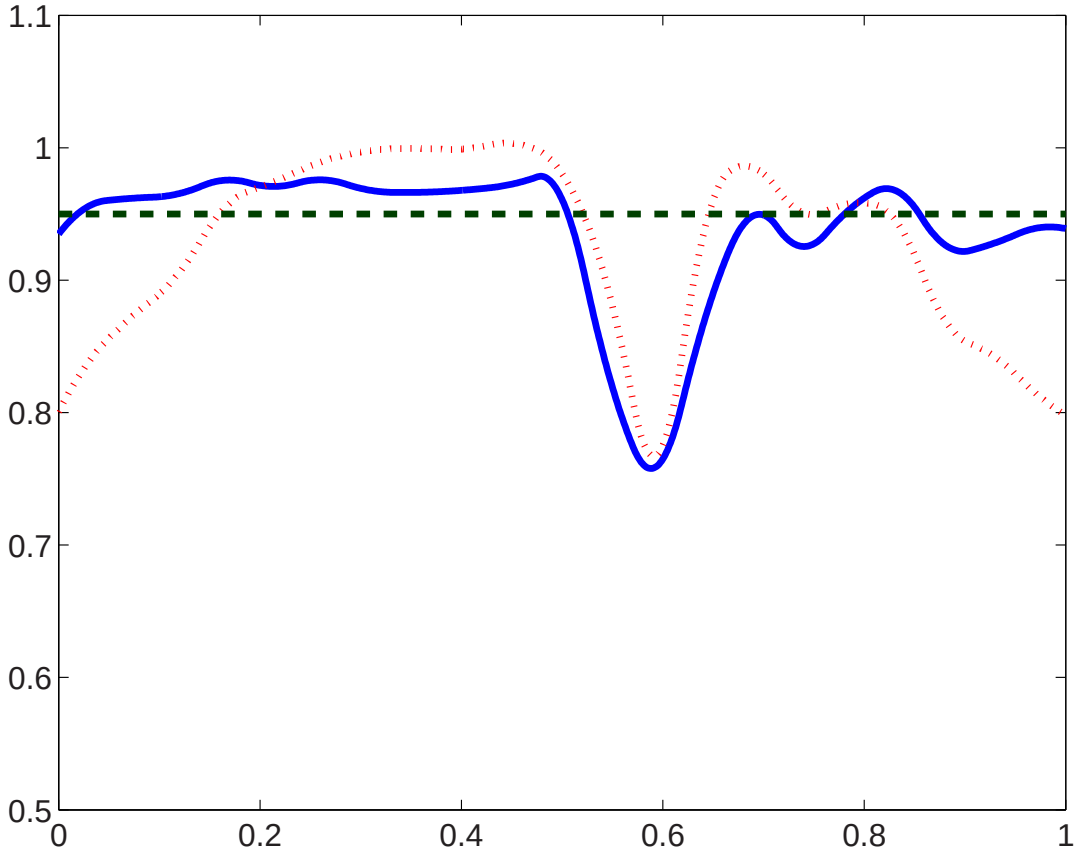

Figure 5: Comparison of pointwise coverage probabilities of the 95% credible intervals in 500 simulations. Data was simulated from the model described in Section 6 with heteroscedastic error variance. Estimation was done using two methods: Bayesian adaptive P–splines with homoscedastic errors (dotted), and Bayesian adaptive P–splines with heteroscedastic errors (solid). The coverage probabilities have been smoothed to remove Monte Carlo variability.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

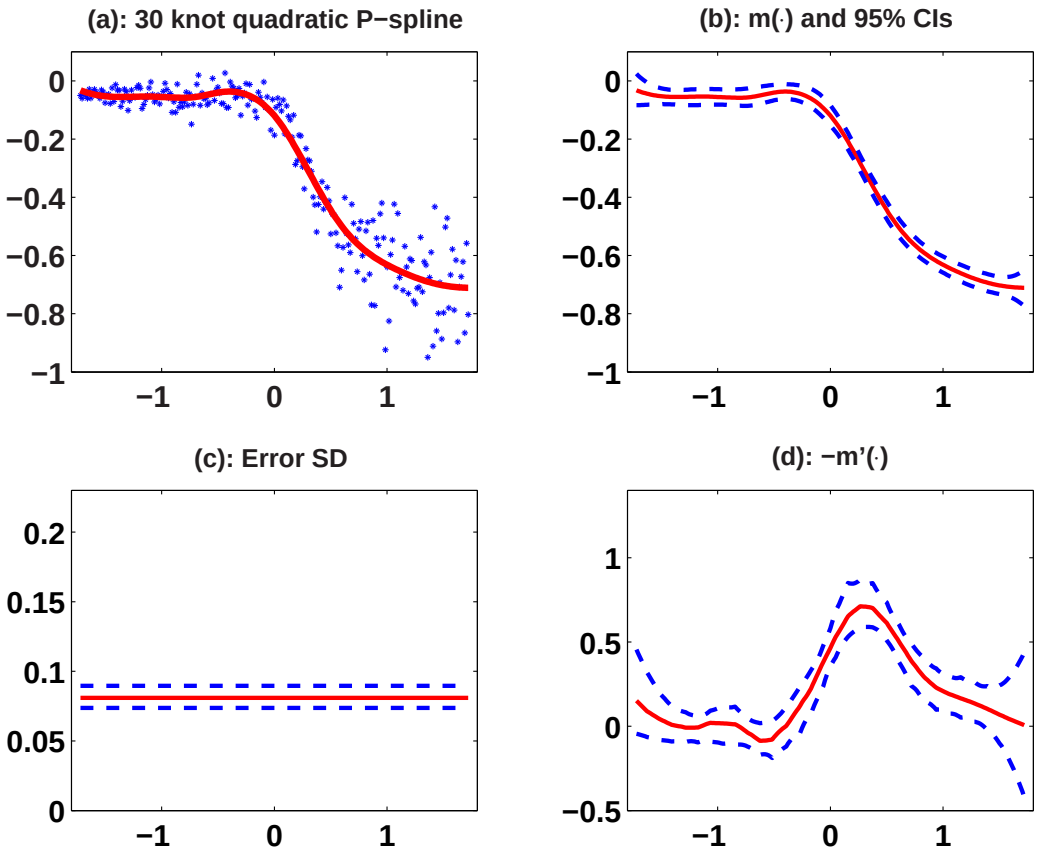

Figure 6: LIDAR data with standardized covariate: inference using Model I which has a global smoothing parameter and homoscedastic errors. The mean function was modeled by a quadratic spline function with 30 knots. (a) – data and posterior mean of the mean function $m(\cdot)$, (b) – posterior mean and simultaneous 95% credible intervals for the mean function $m(\cdot)$, (c) – posterior mean and simultaneous 95% credible interval for the standard error function, (d) – posterior mean and simultaneous 95% credible intervals for $-m'(\cdot)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

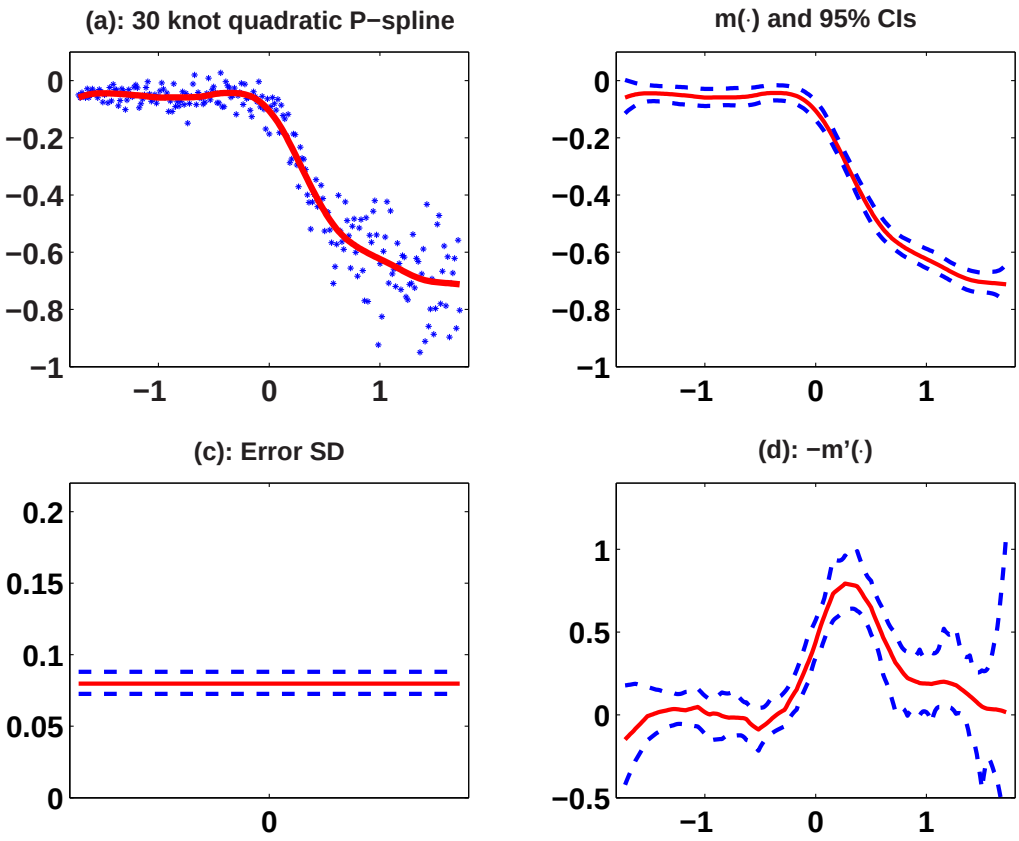

Figure 7: LIDAR data with standardized covariate: inference using Model II with a spatially adaptive smoothing parameter and homoscedastic errors. The mean function was modeled by a quadratic P–spline function with 30 knots. The smoothing parameter function was modeled as a linear log–P–spline function with 4 knots. (a) – data and posterior mean of the mean function $m(\cdot)$, (b) – posterior mean and simultaneous 95% credible intervals for the mean function $m(\cdot)$, (c) – posterior mean and simultaneous 95% credible interval for the standard error function, (d) – posterior mean and simultaneous 95% credible intervals for $-m'(\cdot)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

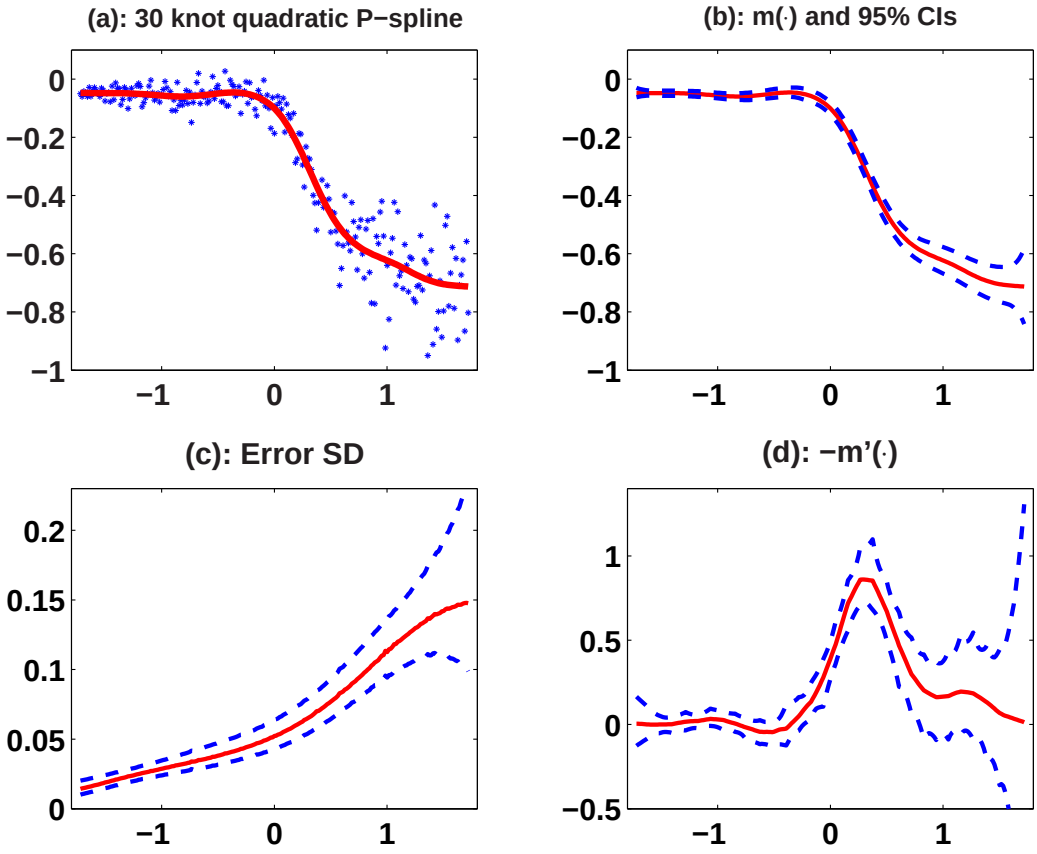

Figure 8: LIDAR data with standardized covariate: inference using Model III which has a spatially adaptive smoothing parameter and a nonparametric model for the heteroscedastic error process. The mean function was modeled by a quadratic P–spline function with 30 knots. The smoothing parameter function was modeled as a linear log–P–spline function with 4 knots. The error variance function was modeled as a quadratic log–P–spline function with 30 knots. (a) – data and posterior mean of the mean function $m(\cdot)$, (b) – posterior mean and simultaneous 95% credible intervals for the mean function $m(\cdot)$, (c) – posterior mean and simultaneous 95% credible interval for the standard error function, (d) – posterior mean and simultaneous 95% credible intervals for $-m'(\cdot)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

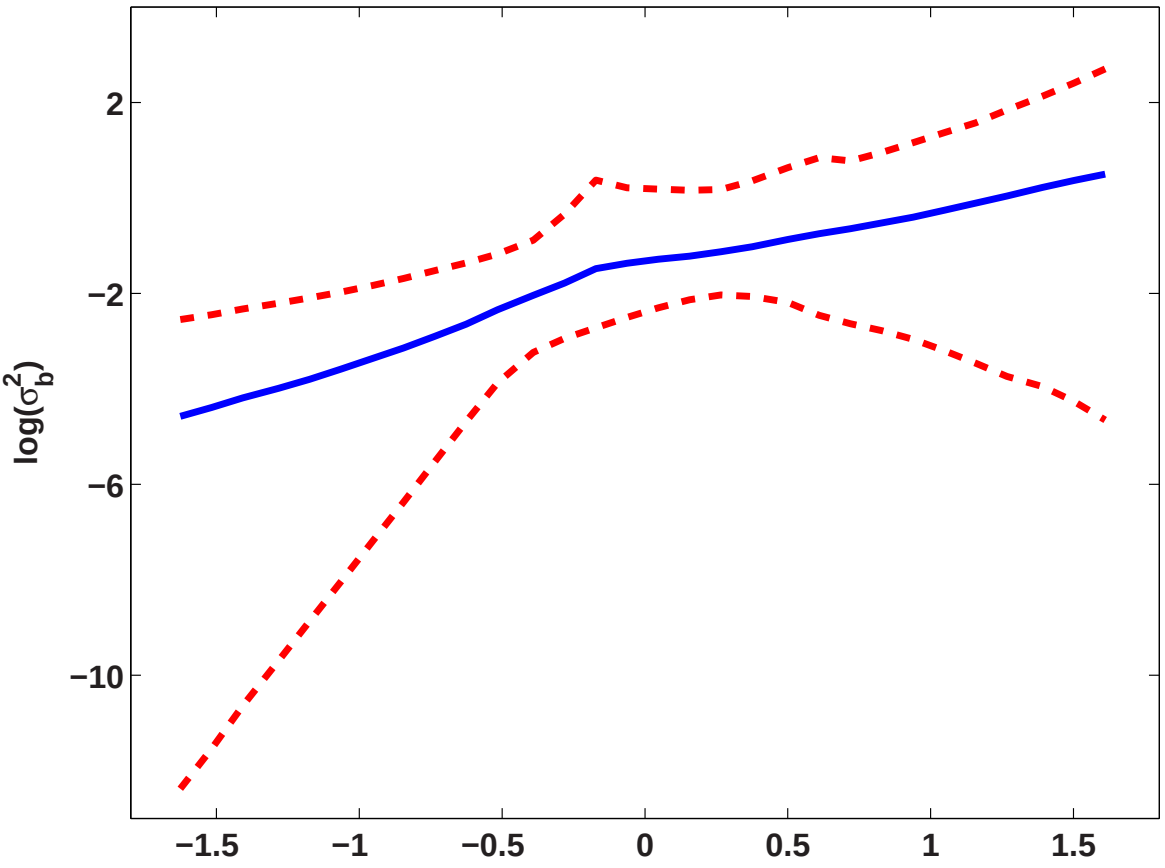

Figure 9: LIDAR data: Posterior mean and 95% simultaneous credible intervals for $\log(\sigma_b^2)$ using Model III. The $\sigma_b^2(\cdot)$ function controls the adaptive shrinkage of the truncated polynomial basis.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

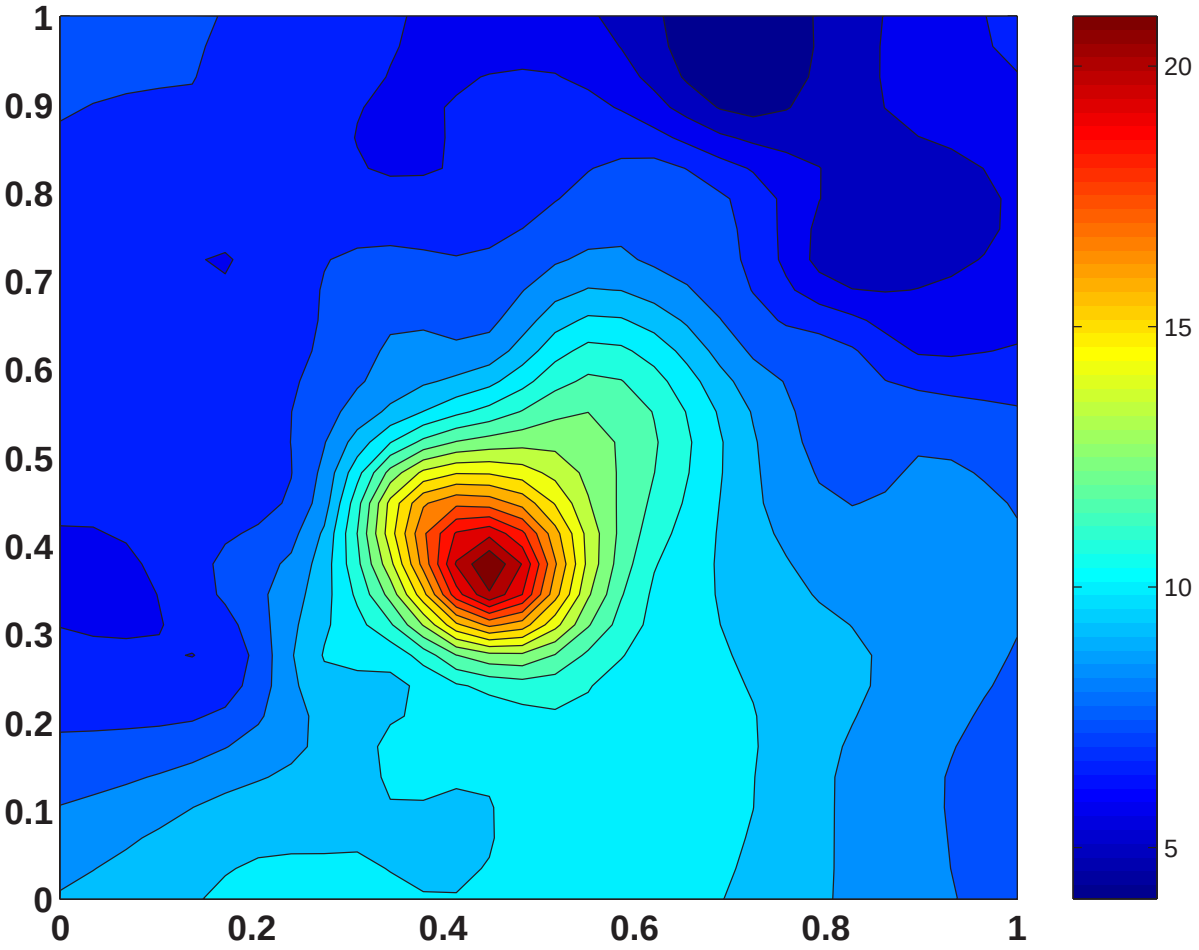

Figure 10: Posterior mean of the mean regression function for the Noshiro example. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin–plate spline basis was $m = 2$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

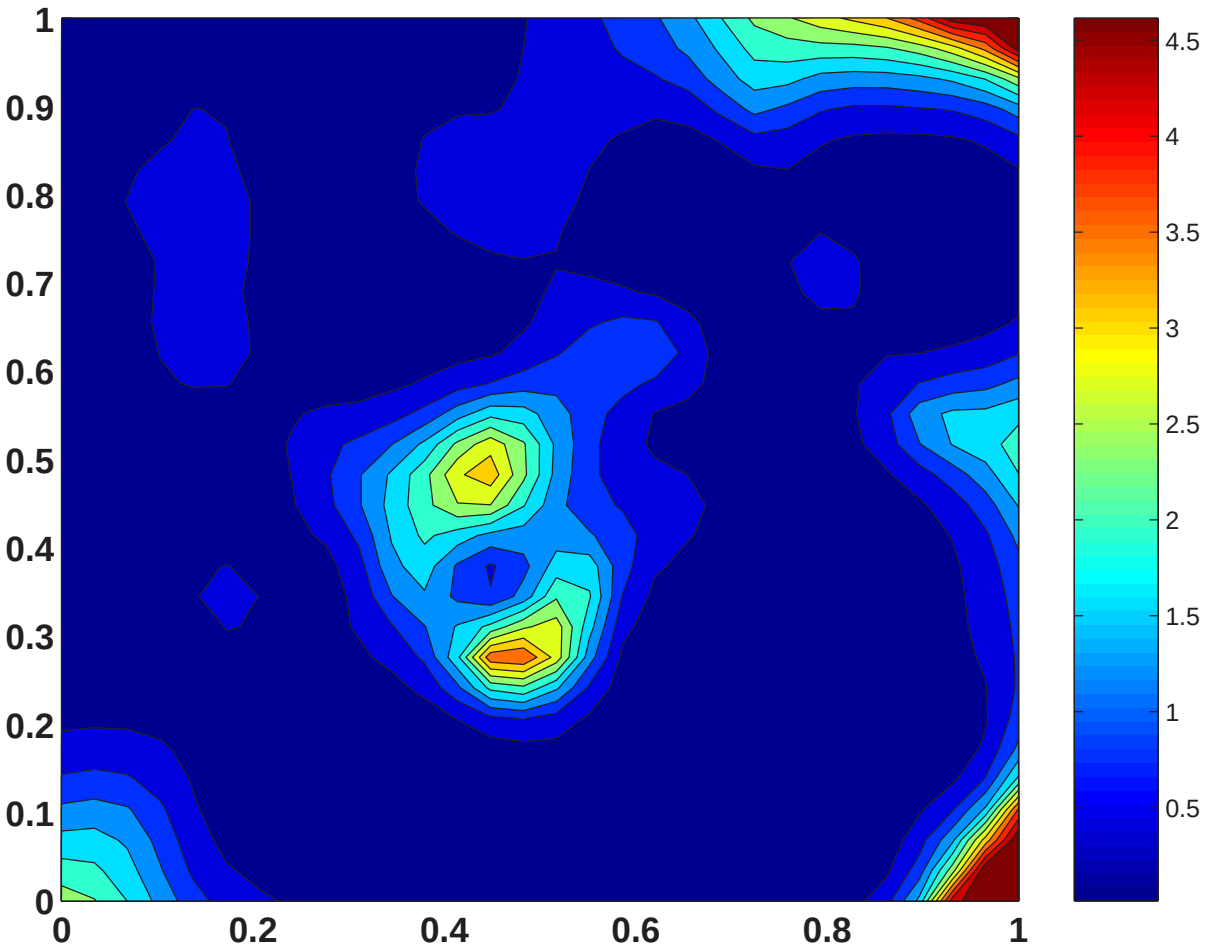

Figure 11: Posterior mean of the standard deviation, $\sigma_\epsilon(\boldsymbol{x}_i)$, of the error process function for the Noshiro example. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin–plate spline basis was $m = 2$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

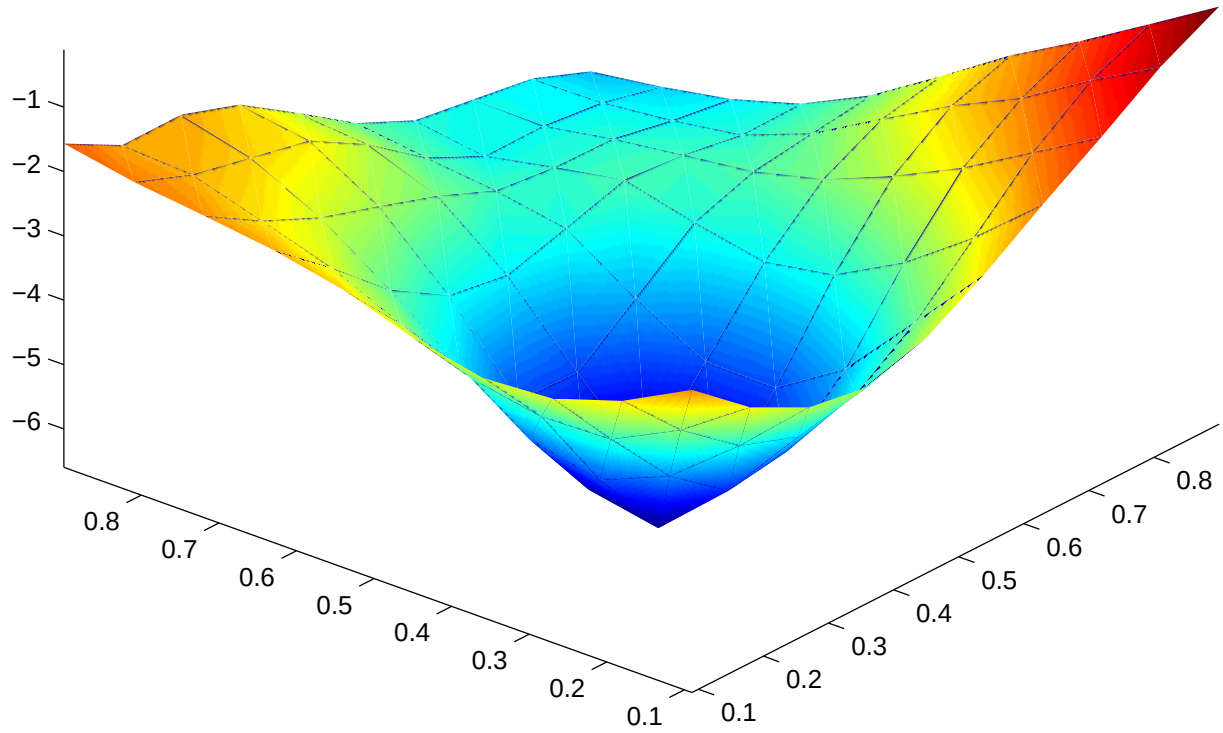

Figure 12: Posterior mean of the shrinkage process $-\log\{\sigma_b^2(\boldsymbol{\kappa}_k)\}$. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin–plate spline basis was $m = 2$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/jhubiostat/paper61

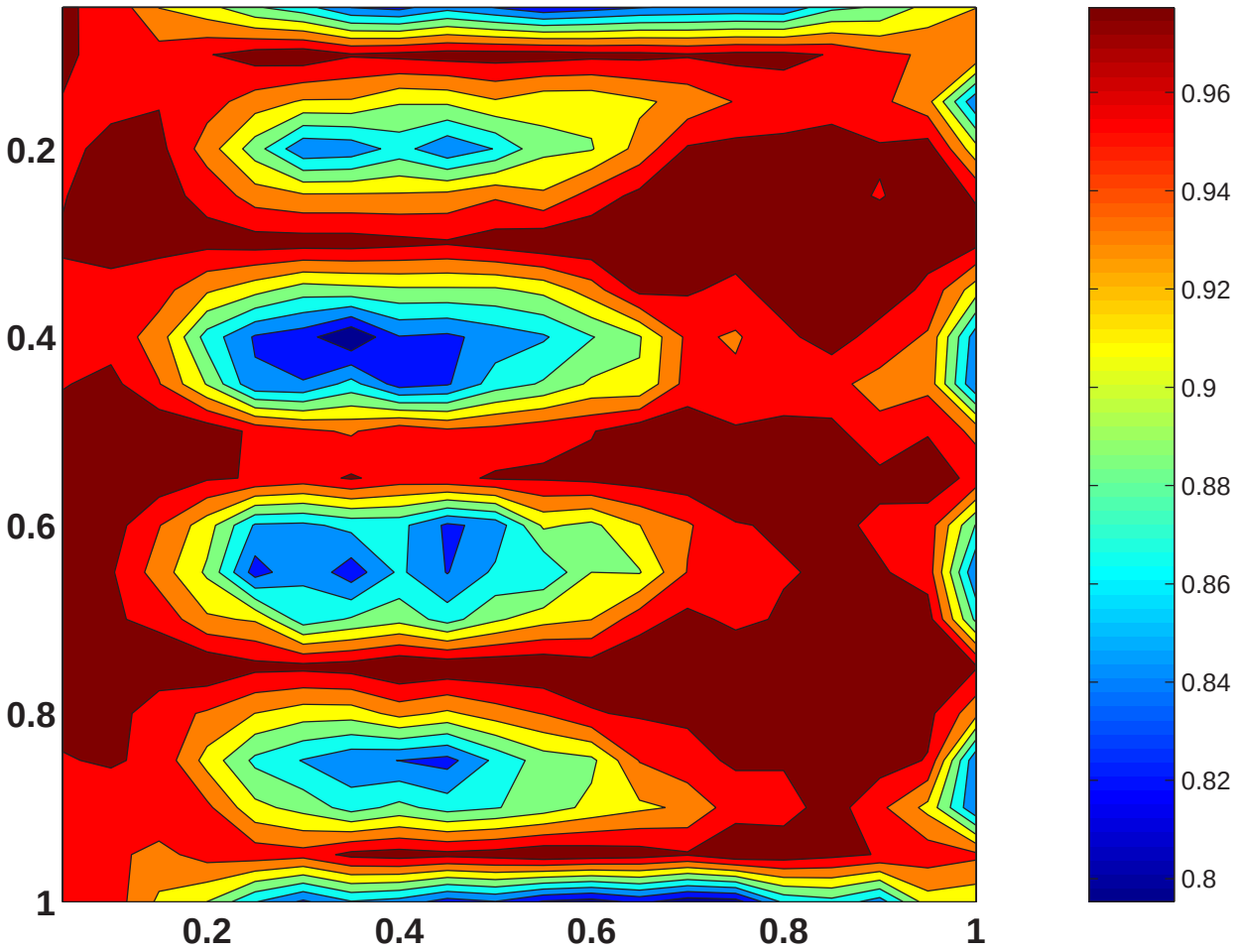

Figure 13: Coverage probability of the pointwise 95% credible intervals for the mean function for function $f_1(x_1, x_2) = x_1 \sin(4\pi x_2)$ with a constant error standard deviation $\sigma = range(f)/4$. Probabilities are computed on a 20×20 equally spaced grid of points in $[0, 1]^2$. The model used was a thin plate spline with $K = 100$ knots for the mean function, a log-thin plate spline with $K = 100$ knots for the variance function, and a log-thin plate spline with $K^* = 16$ knots for the spatially adaptive shrinkage parameter. The parameter controlling the degree of smoothness of the thin–plate spline basis was $m = 2$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press