Fast Adaptive Penalized Splines

Tatyana Krivobokova* Ciprian M. Crainiceanu†
Katholieke Universiteit Leuven Johns Hopkins University

Göran Kauermann‡
Universität Bielefeld

17th April 2007

Abstract

This paper proposes a numerically simple method for locally adaptive smoothing. The heterogeneous regression function is modeled as a penalized spline with a varying smoothing parameter modeled as another penalized spline. This is formulated as a hierarchical mixed model, with spline coefficients following zero mean normal distribution with a smooth variance structure. The modeling framework is similar to Baladandayuthapani, Mallick & Carroll (2005) and Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006). In contrast to these papers the Laplace approximation of the marginal likelihood is used for estimation. This method is numerically simple and fast. The idea is extended to spatial and non-normal response smoothing.

Keywords: Function of locally varying complexity; Hierarchical mixed model; Laplace approximation

*ORSTAT, K.U. Leuven, Naamsestraat 69 - bus 3555, B-3000 Leuven, Belgium (email: Tatyana.Krivobokova@econ.kuleuven.be)

†Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N.Wolfe St. E3636 Baltimore, MD 21205 (email: ccrainic@jhsph.edu)

‡Department of Economics, University of Bielefeld, Postfach 100131, 33501 Bielefeld, Germany (email: gkauermann@wiwi.uni-bielefeld.de)

1 Introduction

Penalized spline smoothing has become increasingly popular in recent years. Originally introduced by O'Sullivan (1986) it was Eilers & Marx (1996) who gave it the name P-spline smoothing. The idea is quite simple. A smooth unknown regression function is estimated by assuming a functional parametric shape constructed via a high dimensional basis function. The dimension of the basis is chosen generously to achieve the desired flexibility, while the basis coefficients are penalized to ensure smoothness of the resulting functional estimates. The idea of penalized splines has led to a powerful and applicable smoothing technique which is demonstrated and motivated in the book by Ruppert, Wand & Carroll (2003). The actual dimension of the basis used has little influence on the fit as has been shown in Ruppert (2002) who concludes that "at most 35 to 40 knots (basis functions) could be recommended for all sample sizes and for all smooth functions without too many oscillations".
The penalized spline methodology is particularly appealing because it can be shown that the penalized spline fit is the Best Linear Unbiased Predictor (BLUP) in a particular mixed model Wand (2003), Ruppert, Wand & Carroll (2003), Kauermann (2004). Thus, standard mixed models software can be used for smoothing, see Ngo & Wand (2004), Crainiceanu, Ruppert & Wand (2005) or Lang & Brezger (2004).
Even though penalized spline smoothing is simple and practical, using a single penalty parameter may fail to correctly capture the features of functions that exhibit strong heterogeneity. This type of function occurs, for example, in applications where the signal changes rapidly in some regions while remaining relatively smooth in others. Regression with strongly heterogenous mean function has been addressed before in the literature and a number of solutions are available. For kernel based methods Fan & Gijbels (1995) and Herrmann (1997) may serve as references. For

spline smoothing Luo & Wahba (1997) suggested hybrid adaptive splines. The idea is to replace the n dimensional spline basis, where n is the sample size, by an adaptively selected subset of the basis functions. This idea has similarities to adaptive knot selection for regression splines as suggested in Friedman & Silverman (1989). An alternative approach is to allow the smoothing parameter to vary locally. Using a reproducing kernel Hilbert space formulation this has been suggested in Pintore, Speckman & Holmes (2005) using piecewise constant smoothing parameters. Similarly, using the penalized spline idea Ruppert & Carroll (2000) allow the penalty to act differently for each spline basis, where the smoothing parameters are then selected using a multivariate generalized cross validation. A similar approach is suggested in Wood, Jiang & Tanner (2002) working with mixtures of splines in a fully Bayesian framework.

In this paper, we focus on a framework similar to the one in Ruppert & Carroll (2000) and achieve spatial adaptivity by imposing a functional structure on the smoothing parameters. This is in line with Baladandayuthapani, Mallick & Carroll (2005) and Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006) who also estimate the error process variance function. However, these two papers require the use of Bayesian MCMC methods and are limited to the case of normal responses. Our paper shows how the MCMC techniques can be circumvented by simple Laplace approximation. While this might be viewed as a small step back in terms of methodology, it actually is a forward leap in terms of numerical simplicity. This allows us to develop fast R software and extend the methodology to non-normal responses.

The paper is organized in Sections that contain their own comparative simulation studies. Section 2 introduces our spatially adaptive modeling methodology. Section 3 extends the methods to spatial smoothing. Section 4 generalizes the approach to

the non-normal responses and provides an example of adaptive bivariate smoothing for binary data. Section 5 contains our conclusions.

2 Smoothly varying local penalties for P-spline regression

2.1 Hierarchical penalty model

We start with the following model

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

where $m(x)$ is a function modeled as a truncated polynomial spline

$$m(x) = \beta_0 + \beta_1 x + ... + \beta_q x^q + \sum_{s=1}^{K_b} b_s \{x - \tau_s^{(b)}\}_+^q, \tag{2}$$

$\tau_1^{(b)}, ..., \tau_{K_b}^{(b)}$ are knots covering the range of xs, and $\{x-\tau_s^{(b)}\}_+^q$ is equal to $\{x-\tau_s^{(b)}\}^q$ if $\{x-\tau_s^{(b)}\} > 0$ and zero otherwise. The dimension K_b of the basis is chosen generously and the knots $\tau_s^{(b)}$ are placed over the range of xs, e.g. at the empirical quantiles of xs. In practice we follow the suggestion of Ruppert (2002) and set $K_b \geq \min(n/4, 40)$. The penalized spline approach is to impose a penalty on the coefficients b_s in (2). A standard approach is to minimize the sum of squares plus a quadratic penalty $\lambda b^T D b$, where λ is the penalty parameter and D is the penalty design matrix. For truncated polynomials the matrix D is the identity matrix and the penalty is $\lambda b^T b$. For B-splines the penalty is constructed from differences between neighboring spline coefficients (see Eilers & Marx, 1996). The methodology presented in this paper is developed for any type of penalized splines and different basis functions will be used for illustration. Notational simplicity is our only reason for using truncated polynomial bases.

An important feature of penalized spline smoothing is its link to linear mixed models.

It can be shown that the penalized spline fit is the BLUP in a particular mixed model. More precisely, in this model $b \sim N(0, \sigma_b^2 D^-)$ where b is the vector of spline coefficients, $\sigma_b^2 = \sigma_\epsilon^2/\lambda$, and D^- is a (generalized) inverse of D. For truncated polynomials $D = I$ and $b \sim N(0, \sigma_b^2 I)$. One limitation of this approach is that a single parameter, σ_b^2, is used to shrink all the spline coefficients. Because of the global nature of the smoothing parameter, σ_b^2, information is borrowed from regions exhibiting high oscillations to regions without and viceversa. The Doppler curve in Figure (1) is a typical example of such a complex function. To avoid this pitfall we allow the coefficients $b_1, ..., b_{K_b}$ to have different prior variances

$$b_s \sim N[0, \sigma_b^2\{\tau_s^{(b)}\}], \quad s = 1, ..., K_b,$$

and assume that the shrinkage variance process $\sigma_b^2\{\tau_s^{(b)}\}$ is a smooth function modeled as a log penalized spline

$$\sigma_b^2\{\tau^{(b)}\} = \exp[\gamma_0 + \gamma_1\tau^{(b)} + ... + \gamma_p\tau^{(b)^p} + \sum_{t=1}^{K_c} c_t\{\tau^{(b)} - \tau_t^{(c)}\}_+^p], \tag{3}$$

where $\tau_1^{(c)}, ..., \tau_{K_c}^{(c)}$ is a second layer of knots covering the range of $\tau_1^{(b)}, \ldots, \tau_{K_b}^{(b)}$. Theoretically we could have $K_c = K_b$ but in practice we choose K_c much smaller than K_b. Our hierarchical penalized smoothing model is completed by the shrinkage assumption $c_t \sim N(0, \sigma_c^2)$, $t = 1, \ldots, K_c$, where the variance σ_c^2 is constant. We now rewrite the model in matrix form. Let

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X_b = \begin{pmatrix} 1 & \ldots & x_1^q \\ \vdots & & \vdots \\ 1 & \ldots & x_n^q \end{pmatrix}, \quad Z_b = \begin{pmatrix} \{x_1 - \tau_1^{(b)}\}_+^q & \ldots & \{x_1 - \tau_{K_b}^{(b)}\}_+^q \\ \vdots & & \vdots \\ \{x_n - \tau_1^{(b)}\}_+^q & \ldots & \{x_n - \tau_{K_b}^{(b)}\}_+^q \end{pmatrix},$$

$\beta = (\beta_0, \ldots, \beta_q)^T$, and $b = (b_1, \ldots, b_{K_b})^T$. We also define

$$X_c = \begin{pmatrix} 1 & \tau_1^{(b)} & \ldots & \tau_1^{(b)^p} \\ \vdots & & & \vdots \\ 1 & \tau_{K_b}^{(b)} & \ldots & \tau_{K_b}^{(b)^p} \end{pmatrix}, \quad Z_c = \begin{pmatrix} \{\tau_1^{(b)} - \tau_1^{(c)}\}_+^p & \ldots & \{\tau_1^{(b)} - \tau_{K_c}^{(c)}\}_+^p \\ \vdots & & \vdots \\ \{\tau_{K_b}^{(b)} - \tau_1^{(c)}\}_+^p & \ldots & \{\tau_{K_b}^{(b)} - \tau_{K_c}^{(c)}\}_+^p \end{pmatrix}.$$

Thus, our hierarchical smoothing model can be written as

$$\begin{aligned} Y|b,c &= X_b\beta + Z_b b + \epsilon,\ \epsilon \sim N(0,\sigma_\epsilon^2 I_n), \\ b|c &\sim N(0,\Sigma_b),\ \Sigma_b = \mathrm{diag}\{\exp(X_c\gamma + Z_c c)\}, \\ c &\sim N(0,\sigma_c^2 I_{K_c}). \end{aligned} \tag{4}$$

The marginal likelihood of model (4) is

$$\begin{aligned} L(\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) &= f(Y;\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) \\ &= (2\pi)^{-\frac{(n+K_c)}{2}}\sigma_\epsilon^{-n}\sigma_c^{-K_c}\int_{R^{K_c}}\exp\{-g(c)\}dc, \end{aligned} \tag{5}$$

where

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

and $V_\epsilon = I_n + Z_b\Sigma_b Z_b^T/\sigma_\epsilon^2$. Note that Σ_b and V_ϵ depend on c and γ, but this additional notational burden will be omitted throughout the paper. The integral in (5) is not available analytically, which explains why other authors chose to use Bayesian MCMC techniques. In contrast, we use the Laplace approximation, which is justifiable because K_c and K_b are bounded while sample size n is growing, i.e. $K_c < K_b \ll n$. Thus, the Laplace approximation has an error of order $O(n^{-1})$ (see Severini, 2000), which makes it an attractive alternative to simulation based techniques. The Laplace log-likelihood approximation is, up to a constant,

$$\begin{aligned} -2l(\beta,\gamma,\sigma_\epsilon^2,\sigma_c^2) &\approx n\log\sigma_\epsilon^2 + K_c\log\sigma_c^2 + \log|V_\epsilon(\hat{c})| + \log|I_{cc}(\hat{c})| \\ &+ \hat{c}^T\hat{c}/\sigma_c^2 + (Y-X_b\beta)^T V_\epsilon^{-1}(\hat{c})(Y-X_b\beta)/\sigma_\epsilon^2, \end{aligned} \tag{6}$$

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

$$\frac{\partial g(\hat{c})}{\partial c_i} = \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}\right) + \frac{c_i}{\sigma_c^2} - \frac{1}{2\sigma_\epsilon^2}(Y-X_b\beta)^T V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}(Y-X_b\beta) = 0, \tag{7}$$

$$\{I_{cc}(c)\}_{ij} = E\left(\frac{\partial^2 g(c)}{\partial c_i \partial c_j}\middle| c\right) = \frac{\delta_{ij}}{\sigma_c^2} + \frac{1}{2}\mathrm{tr}\left(V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_i}V_\epsilon^{-1}\frac{\partial V_\epsilon}{\partial c_j}\right), \tag{8}$$

and δ_{ij} is Kronecker's delta. It is easy to show that the derivative appearing in the above equations results in $\partial V_\epsilon/\partial c_i = Z_b \mathrm{diag}(Z_{c,i})\Sigma_b Z_b^T/\sigma_\epsilon^2$, where $Z_{c,i}$ stands for the ith column of the matrix Z_c. The prediction of b is defined by $Z_b^T V_\epsilon^{-1}(y - X_b\beta) = \sigma_\epsilon^2\Sigma_b^{-1}\hat{b}$ and $\mathrm{tr}\,(V_\epsilon^{-1}\partial V_\epsilon/\partial c) = Z_c^T w_{df}$, where w_{df} is the K_b dimensional vector of diagonal elements of $A := Z_b^T Z_b(\sigma_\epsilon^2\Sigma_b^{-1} + Z_b^T Z_b)^{-1}$. Thus, (7) and (8) become

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

and

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

respectively. Here v_{df} is the K_b dimensional vector of diagonal elements of A^2. Note that $df_b = \sum_{s=1}^{K_b} w_{df} = 1_{K_b}^T w_{df}$ is the number of degrees of freedom used for fitting b. Note that $\partial v_{df}/\partial\gamma_i = 2\mathrm{diag}\{(A^2 - A^3)\mathrm{diag}(X_{c,i})\}$ with $X_{c,i}$ as the ith column of the matrix X_c indicating that the weights v_{df} vary very slowly as a function of γ. Thus, using this observation we can estimate γ and c simultaneously using the following iterated weighted least squares (IWLS) algorithm. If $\theta = (\gamma^T, c^T)^T$ the algorithm starts by setting

$$\hat{\theta} = \frac{1}{2}\left(\frac{1}{2}W_c^T \mathrm{diag}\,(v_{df})W_c + \frac{D_c}{\sigma_c^2}\right)^{-1} W_c^T \mathrm{diag}\,(v_{df})\,u, \tag{9}$$

where $W_c = (X_c, Z_c)$, $D_c = \mathrm{diag}\{0_{(p+1)\times(p+1)}, I_{K_c}\}$, and $u = W_c\theta + \mathrm{diag}(v_{df}^{-1})(\Sigma_b^{-1}\hat{b}^2 - w_{df})$. Minimizing (6) for fixed $\hat{\theta}$ one obtains

$$\begin{aligned}
\hat{\sigma}_c^2 &= \hat{c}^T\hat{c}/w_{df}^c \\
\hat{\beta} &= (X_b^T V_\epsilon^{-1}(\hat{\theta})X_b)^{-1}X_b^T V_\epsilon^{-1}(\hat{\theta})y, \\
\hat{\sigma}_\epsilon^2 &= (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/n,
\end{aligned} \tag{10}$$

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

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

The latter steps are standard in linear mixed model methodology. Estimation can now be carried out using the EM type algorithm (see e.g. Searle, Casella, & McCulloch, 1992 or Breslow & Clayton, 1993) by iterating between (9) and (10) until convergence. It should be noted that the estimation consists of two simple steps and is, therefore, numerically fast. In fact, for $n = 1000$ the fit is obtained in seconds on standard computers. In contrast Bayesian MCMC methods (Crainiceanu, Ruppert, Carroll, Adarsh & Goodner, 2006 or Baladandayuthapani, Mallick & Carroll, 2005) require more than 10 minutes on similar computers. As we will show in subsequent sections our proposed method can be extended to non-normal responses and is robust to small changes of the model. This allowed us to develop the R package "AdaptFit" which implements the methodology described in this paper (please see the Appendix).

2.2 Restricted maximum likelihood

The above results are presented for maximum likelihood estimates. The use of restricted maximum likelihood (REML) is, however, more common in mixed models (see Harville, 1977). The restricted maximum log-likelihood for the model (4) is

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

with $l(\beta, \gamma, \sigma_\epsilon^2, \sigma_c^2)$ as given in (6). The estimation method is similar to the one in Section 2.1, with the matrix A in w_{df} and v_{df} replaced by

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

and the variance estimate replaced by $\hat{\sigma}_\epsilon^2 = (y - X_b\hat{\beta})^T V_\epsilon^{-1}(\hat{\theta})(y - X_b\hat{\beta})/(n-q-1)$. In our performance study the ML and REML methods provided similar results.

2.3 Variance estimation

We denote by $\tilde{m}(x)|c = X_b\tilde{\beta} + Z_b\tilde{b}|c$ the BLUP of the function $m(x)|c = X_b\beta + Z_b b|c$, where $\tilde{\beta} = (X_b^T V_\epsilon^{-1} X_b)^{-1} X_b^T V_\epsilon^{-1} y$ and $\tilde{b}|c = \Sigma_b Z_b^T V_\epsilon^{-1}(y - X_b\tilde{\beta})/\sigma_\epsilon^2$. In the mixed model framework the function $m(x)|c$ is random because b is random and $\tilde{m}(x)|c$ is unbiased for $m(x)|c$. Thus, confidence intervals for $m(x)|c$ can be obtained from

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

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

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

Let $\hat{m}(x)|c = X_b\hat{\beta} + Z_b\hat{b}|c$ denote the EBLUP obtained from $\tilde{m}(x)|c$ by plugging in the estimates of variance parameters. This can be used to obtain a plug in estimate $\widehat{\mathrm{Var}}\{\hat{m}(x) - m(x)\} \approx \hat{\sigma}_\epsilon^2 S(\hat{\theta})$.

The variance estimate can also be justified in the Bayesian framework. Assuming that the parameters $\Sigma_b = \mathrm{diag}\{\exp(W_c\theta)\}$ and σ_ϵ^2 are known, the posterior distribution of $m(x)$ is $N\{\hat{m}(x,\theta), \sigma_\epsilon^2 S(\theta)\}$, where $\hat{m}(x,\theta) = S(\theta)y$. An empirical Bayes approach would replace the unknown values Σ_b and σ_ϵ^2 in the prior by estimates and then treat these parameters as if they were known. Thus, the approximate posterior distribution of $m(x)$ is $N\{\hat{m}(x,\hat{\theta}), \hat{\sigma}_\epsilon^2 S(\hat{\theta})\}$, yielding the same confidence intervals as the frequentist mixed model approach.

The variance formula is simple but does not account for the estimation variability

of θ, the parameters of the shrinkage process. This is the price to pay when using Laplace's method instead of a full Bayesian approach. For further discussion we refer to Morris (1983), Laird & Louis (1987), Kass & Steffey (1989) or Ruppert & Carroll (2000). To correct for this additional variability, we use the delta-method correction proposed by Kass & Steffey (1989) and obtain

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

As estimate of $\mathrm{Var}(\hat{\theta})$ one can use the inverse of the Fisher information matrix $I_{\theta\theta}(\hat{\theta})$ obtained from the last iteration of the estimation algorithm. The derivative in the last term, ignoring the dependence of $\hat{\sigma}^2_\epsilon$ on θ, is

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

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

2.4 Numerical implementation

Our proposed algorithm is simple and can be presented as a sequence of mixed model fits using standard mixed effects software:

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

2. Get next estimates for $\hat{\theta}$ and $\hat{\sigma}^2_c$ from (9) and (10);

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

4. Iterate between 2 and 3 until convergence.

We implemented this algorithm in the package "AdaptFit" described below. We compared B-splines of different degrees and penalty orders, quadratic and cubic truncated polynomials as well as cubic thin plate splines. Although all spline bases produced almost indistinguishable results, the cubic thin plate splines were slightly more numerically stable and were preferred in the subsequent simulation studies.

2.5 Simulations and comparisons with other univariate smoothers

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

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

with $j = 6$. We performed 500 simulations with $K_b = 80$ and $K_c = 20$, choosing the number of knots to be consistent with Ruppert & Carroll (2000). The top plot of Figure 1 shows the fit and 95% confidence intervals for one simulated data set. The corresponding estimated variance of random effects is shown in the middle plot. The average MSE over all x's obtained as $AMSE = n^{-1}\sum_{i=1}^{n}\{\hat{m}(x_i) - m(x_i)\}^2$ equals 0.0034, which is comparable with 0.0027 reported in Baladandayuthapani, Mallick & Carroll (2005) and 0.0026 of Ruppert & Carroll (2000). We also computed the pointwise coverage probabilities of the 95% confidence intervals over all 500 simulated datasets. The smoothed pointwise coverage probabilities can be seen in the bottom plot of Figure 1. For small values of $x \leq 0.1$, i.e. in the region with low signal-to-noise ratio, there is clear under-coverage but beyond 0.1 the coverage prob-

ability exceeds 95% being slightly conservative. The average coverage probability is 94.95%.

Next, we considered the heterogeneous regression function

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

with $n = 1000$, the x values being equally spaced on $[0, 1]$ and $\epsilon_i \sim N(0, 0.5^2)$. We applied our approach to 500 simulated datasets, using $K_b = 40$ and $K_c = 4$, following the choice of Ruppert & Carroll (2000) and Baladandayuthapani, Mallick & Carroll (2005). The top and middle plots of Figure 2 represent one of the simulated fits and estimated variance of random effects, respectively. The resulting AMSE is equal 0.0048, which is smaller than 0.0061 and 0.0065, obtained by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000), respectively. The smoothed pointwise coverage probabilities can be seen in the bottom plot of Figure 2. The average coverage probability for this function equals 95.94%, which is comparable with 95.22% and 96.28% reported by Baladandayuthapani, Mallick & Carroll (2005) and Ruppert & Carroll (2000), respectively. For the same setting Baladandayuthapani, Mallick & Carroll (2005) reported the simulation results for the BARS approach of DiMatteo, Genovese & Kass (2001). BARS employs free-knots splines with the random number and location of knots, using reversible jump MCMC for estimation. The AMSE based on this approach is 0.0043, while the average coverage probability is 94.72%, both comparable to our approach.

To show the robustness of our approach to the choice of number of subknots K_c, we performed simulations both for $m_1(x)$ and $m_2(x)$ with different values of K_c. AMSE based on 500 simulations for the function $m_1(x)$ using 10, 20 and 30 subknots, respectively, resulted in 0.00344025, 0.00344029 and 0.00344023. AMSE based on 500 simulations for the function $m_2(x)$ based on K_c equal to 4, 10 and 15, respectively,

results in 0.0048405, 0.0048330 and 0.0048313. In general there should be enough subknots to capture the structure of the variance of random effects and further increase of K_c has little effect on the fit.

Our approach provides good results even when adaptive smoothing is not necessary. In such cases, the variance of the random effects is estimated to be nearly constant and has little effect on the resulting fit. To show this we run 150 simulations based on the function $\sin(2\pi x)$. We used $n = 400$ points and $\epsilon_i \sim N(0, 0.3^2)$. The adaptive and non-adaptive estimates approaches provided nearly indistinguishable results, with the AMSE values being 0.0017038 and 0.0017351, respectively.

Overall, our method provides comparable results to other approaches, but with significantly less numerical effort.

3 Spatial smoothing

3.1 Hierarchical modeling

We now generalize the ideas from the previous section to spatial smoothing. We consider the model $y_i \sim N\{m(\boldsymbol{x}_i), \sigma_\epsilon^2\}$ for $i = 1, \ldots, n$, where $\boldsymbol{x}_i \in R^2$ and $m(.)$ is a smooth function of two covariates. Following Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006) we use radial basis functions (for details see Ruppert, Wand & Carroll, 2003) and choose K_b knots $\boldsymbol{\tau}_1^{(b)}, ..., \boldsymbol{\tau}_{K_b}^{(b)} \in R^2$. In this case, the fixed effects matrix X_b is the matrix with i-th row equal to $[1, \boldsymbol{x}_i^T]_{1 \le i \le n}$ and the random effects matrix is equal to $Z_b = Z_{K_b} \Omega_{K_b}^{-1/2}$, where $Z_{K_b} = [\|\boldsymbol{x}_i - \boldsymbol{\tau}_s^{(b)}\|^2 \log \|\boldsymbol{x}_i - \boldsymbol{\tau}_s^{(b)}\|]_{1 \le s \le K_b, 1 \le i \le n}$ and $\Omega_{K_b} = [\|\boldsymbol{\tau}_t^{(b)} - \boldsymbol{\tau}_s^{(b)}\|^2 \log \|\boldsymbol{\tau}_t^{(b)} - \boldsymbol{\tau}_s^{(b)}\|]_{1 \le s,t \le K_b}$ with $\|.\|$ denoting the Euclidean norm in R^2. The penalized thin plate spline spatial smoother is equivalent to the following linear mixed model

$$Y|b = X_b\beta + Z_b b + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2 I_n), \quad b \sim N(0, \Sigma_b). \tag{11}$$

As in the case of univariate smoothers, local adaptiveness is achieved by allowing the coefficients b to have spatially variable smoothing parameters. As in Section 2, we set spatial subknots $\boldsymbol{\tau}_1^{(c)}, ..., \boldsymbol{\tau}_{K_c}^{(c)} \in R^2$, $K_c < K_b$ and define matrices X_c and Z_c similarly to the corresponding definition of matrices X_b and Z_b. More precisely, $X_c^s = [1, (\boldsymbol{\tau}_s^{(b)})^T]_{1 \le s \le K_b}$, $Z_c = Z_{K_c}\Omega_{K_c}^{-1/2}$ with $Z_{K_c} = [\|\boldsymbol{\tau}_s^{(b)} - \boldsymbol{\tau}_t^{(c)}\|^2 \log \|\boldsymbol{\tau}_s^{(b)} - \boldsymbol{\tau}_t^{(c)}\|]_{1 \le s \le K_b, 1 \le t \le K_c}$ and $\Omega_{K_c} = [\|\boldsymbol{\tau}_t^{(c)} - \boldsymbol{\tau}_s^{(c)}\|^2 \log \|\boldsymbol{\tau}_t^{(c)} - \boldsymbol{\tau}_s^{(c)}\|]_{1 \le s,t \le K_c}$, where the $\boldsymbol{x}$ covariates are replaced by knots $\boldsymbol{\tau}^{(b)}$ and the knots are replaced with subknots $\boldsymbol{\tau}^{(c)}$. The model is completed by adding to (11) the hierarchical structure

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

Estimation can now be carried out as in the case of univariate smoothers. There are many ways of choosing the knots. We used the *clara* algorithm described in Kaufman & Rousseeuw (1990) and implemented in the R package "cluster".

3.2 Simulations and comparisons with other surface fitting methods

For comparison with Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006) and Lang & Brezger (2004) we consider the following regression function with moderate spatial variability

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

with x_1 and x_2 independent and uniformly distributed on $[0, 1]$. We used $n = 300$, $\sigma = 1/4\text{range}(m_3)$ and equally-spaced 12×12 and 5×5 knot grids for $\tau_i^{(b)}$ and $\tau_j^{(c)}$, respectively, as suggested by Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006). Figure 3 displays the true function (top left plot) and the resulting fit for one simulation using our approach (top right plot) together with the estimated variance of random effects (left bottom plot). We simulated 500 datasets to compare

log(MSE) of our estimator with values reported in Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006) and Lang & Brezger (2004). Our simulations provide a median of log(MSE) of -3.79 with an interquartile range $[-4.17, -3.80]$ and a range $[-4.96, -2.27]$. This outperforms the results in Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006) (median -3.67, interquartile range $[-3.80, -3.53]$ and a range $[-4.21, -3.13]$) which, in turn, outperforms the findings of Lang & Brezger (2004). The average coverage probability of the 95% confidence intervals is 94.31%. The smoothed coverage probabilities are displayed in the right bottom plot of Figure 3. Similarly to the Crainiceanu, Ruppert, Carroll, Adarsh & Goodner (2006), the coverage probability is lowest for $x_1 \in [0.2, 0.5]$ due to the low signal-to-noise ratio in this region.

4 Non-normal response models

4.1 Hierarchical modeling

In this section we extend the methods to the case of non-normal response models. Consider the following generalized linear hierarchical mixed model

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

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

$$\exp\{ql(\beta, \gamma, \sigma_c^2)\} = (2\pi)^{-\frac{(K_b+K_c)}{2}} \sigma_c^{-K_c} \int_{R^{K_b}} \int_{R^{K_c}} \exp\{-k_1(b,c)\} db\, dc, \quad (12)$$

where

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

and $q_i(y,\mu) = -2\int_y^\mu (y-t)/v(t)dt$ is the deviance function. Assuming that conditionally on b and c the observations are drawn from the exponential family $Y|b,c \sim \exp([y\vartheta(x) - b\{\vartheta(x)\}]/\phi + c(y,\phi))$, the quasi-likelihood (12) is the likelihood of the data. Using Laplace's method for approximation of the integral over b one obtains

$$\exp\{ql(\beta,\gamma,\sigma_c^2)\} \approx (2\pi)^{-\frac{K_c}{2}}\sigma_c^{-K_c}\int_{R^{K_c}}\exp\{-k_2(c)\}dc, \tag{13}$$

where

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

$\hat{b}$ is the solution to

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

Here W denotes the $n\times n$ diagonal matrix of the GLM iterated weights with diagonal elements $w_i = [\phi v(\mu_i^{b,c})\{\tilde{g}'(\mu_i^{b,c})\}^2]^{-1}$, using the simplifying assumption that the iterative weights w_i vary slowly as a function of the mean.

Substituting the current estimate $\hat{b}$ into (13) and replacing the deviance $\sum q_i(y_i,\mu_i^{b,c})$ in $k_2(.)$ by the Pearson chi-squared statistic $\sum(y_i-\mu_i^{b,c})^2/v_i(\mu_i^{b,c})$ provides the following approximation

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

where $k_3(c) = \frac{1}{2}\log|V| + \frac{c^Tc}{2\sigma_c^2} + (U-X_b\beta)^TV^{-1}(U-X_b\beta)$, $V = W^{-1} + Z_b\Sigma_bZ_b^T$, and $U = X_b\beta + Z_b\hat{b} + \mathrm{diag}\{\tilde{g}'(\mu^{b,c})\}(Y-\mu^{b,c})$. Applying again Laplace's method we end up with the following quasi-log-likelihood for the remaining parameters

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

where $k_3^{cc} = \partial^2 k_3(c)/\partial c \partial c^T$. As in Section 2 the estimation of parameter $\theta = (\gamma^T, c^T)^T$ can be carried out from the score equation

$$\frac{\partial k_3(\hat{\theta})}{\partial \theta} = -\frac{1}{2} W_c^T \Sigma_b^{-1} \left(\hat{b}^2 - w_{df} \sigma_b^2 \right) + D_c \theta / \sigma_c^2 = 0. \tag{14}$$

As in the normal case, the algorithm iterates between estimation of $\hat{\theta}$ and generalized linear mixed models fitting implemented in standard software.

4.2 Simulations for non-normal response models

We consider the following model for Bernoulli data $Y_i \sim B(1, \pi_i)$ with canonical link $\text{logit}(\pi_i) = m_2(x_i)$ where $m_2(.)$ is the function introduced in Section 2. The top plot of Figure 4 represents the adaptive fit (bold line) for a data set of size $n = 5000$. For comparison the non-adaptive fit (dashed line) is also shown indicating the superiority of the adaptive fit. While the benefits of adaptive fit with binary data are more clearly visible for large sample sizes, the adaptive procedure does not require very large sample sizes to provide functional estimates. Depending on the sample size and signal-to-noise ratio the adaptive fit would produce fits that are closer or further away from its non-adaptive counterpart.

Since the method in this paper is the first one that can provide adaptive smoothing for binary data, we cannot compare our method to others. However, the BARS procedure of DiMatteo, Genovese & Kass (2001) allows fitting Poisson responses. We performed a number of simulations to compare the performance of our routine with the BARS implemented for this setting. We simulated $n = 800$ Poisson variates with means $\exp\{m_1(x)\}$, where $m_1(.)$ was introduced in Section 2.5, and use $j = 4$ which corresponds to moderate heterogeneity. We estimated the data with our approach using $K_b = 60$ and $K_c = 10$ and with the BARS procedure. We ran 10000 MCMC iterations with a burn-in sample of 2000. The bottom plot of Figure

4 displays estimates based on our adaptive approach (bold) and BARS (dashed). The AMSE of the five fits based on our approach is 0.010672, while the AMSE for BARS based fits is 0.020547. We did not perform a more extensive simulation study, since a single BARS fit required more than 4 hours estimation time on an up-to-date computer. For comparison, our function `asp` required only one minute. We experimented with other mean functions and sample sizes and obtained very similar results.

4.3 Example

As an illustration we apply spatially adaptive smoothing methods to a dataset on the absenteeism of workers at a company in Germany. Parts of the data have been analyzed before in Kauermann & Ortlieb (2004) with a different focus. We consider absenteeism spells and model the probability of returning to work after a sick leave. Denoting the duration of such a leave by d, we model the discrete hazard rate as $P(d = t|d \geq t) = h(t)$, where $t \geq 1$. The duration is measured in days and the event of interest is "recovery", which allows workers to return to work. If the worker reported sick on one day but returns to work on a consecutive working day thereafter, we count this as an event and the duration is the number of working days the worker had been absent. If the last day of absenteeism and the first day of returning to work are not consecutive working days the duration is viewed as censored with d being the number of days of absenteeism. To make this more explicit, assume that a worker reports sick on Friday but returns to work the Monday after. It is unclear when the worker actually recovered, either Friday, Saturday or Sunday. It is however known, that the worker was at least sick on one day and the observation is therefore $d = 1$ with censoring being indicated. Let δ denote the censoring indicator, which

is either zero, for censoring, or 1, otherwise. For each absence spell we transform d to the binary variables $y_1, ..., y_d$ with $y_l = 0$ for $l < d$ and $y_d = \delta$. The hazard function is then the probability $P(y_t = 1|y_l = 0, l < t)$. We concentrate on short term absenteeism spells truncated at $d = 10$ and define longer spells as censored observations. Our final model depends both on time since leave, t, and on actual calendar time, c,

$$\text{logit}P(d = t|d \geq t, c) = m(t, c). \tag{15}$$

Data were collected in a company in southern Germany and we analyze the data of about 370 employees. Not all of these employees were employed during the same time period with the observation period ranging from 1981 to 1998. On average, about 75% of the employees reported sick at least once per calender year. For illustration purposes we assume that the durations of different sick leaves of the same worker are independent. One might, of course, argue whether this is an appropriate assumption, but for the sake of simplicity we do not address this problem here. Figure 5 shows the fit of the model (15) using non-adaptive (top) and adaptive (bottom) smoothing. Both fits were obtained using 196 knots (14 for each dimension) and low-rank thin spline basis as defined in Section 3.1. The variance structure for the adaptive fit was modelled with 100 knots (10 for each dimension). The differences in the plots are quite obvious and interesting. Both fits show a higher hazard for years 1992 and 1993 around day 3. The adaptive fit has a sharper peak in this region. In other regions, particularly for longer absenteeism time, the non-adaptive fit is quite wiggly while the adaptive approach estimates a much smoother surface. The latter fit looks more reasonable and easier to interpret and shows the benefits of spatial adaptivity in a real context.

The peak at year 1993 and duration time at day 3 allows for an interesting economic

interpretation. In 1992/93 the company went through a major downsizing process with more than 50% of the workers being dismissed. While this economic situation has hardly any effect on the hazard function for days $d \geq 5$, it does affect the hazard rate for short absenteeism times, particularly for $d = 3$. According to the German law, workers who report sick for more than 3 consecutive working days have to provide a medical certificate at the latest at the third day of their sick leave. For shorter periods no special attestation is required. Apparently, during the downsizing period the duration of sick leaves is shorter with more employees returning after 3 days. This provides indication that economically critical conditions of a company have a direct influence on the absenteeism of employees.

5 Conclusion

We showed that local adaptive smoothing can be easily carried out by formulating penalties on the spline coefficient as a hierarchical mixed model. Our major contribution was to show that the Laplace approximation of the marginal likelihood allows fast fitting of adaptive smoothing models with application to univariate, bivariate smoothing as well as to non-normal responses. For reasonably sized data sets and normal responses our algorithm requires seconds while Bayesian MCMC simulations require tens of minutes. For non-normal responses our algorithm requires less than a minute while Bayesian MCMC requires more than several hours. In addition, small changes to the model, such as adding a covariate, standard random effects, or other smooth components, can be handled very easily by our methods, whereas it may take hours, days, or even weeks for developing new MCMC software. An important reason for having a fast and accurate procedure is that in many situations the smoothing procedure needs to be applied repeatedly. One trivial example is when

doing simulations (note, for example how long it would take to perform the BARS simulations in Section 4.2).

A R Package "AdaptFit"

To implement our approach we developed an R package. We took advantage of the R package "SemiPar", written by M.P. Wand to accompany the book Ruppert, Wand & Carroll (2003). The function `spm` of this package performs scatterplot, spatial and generalized (binomial and poisson) smoothing using the (generalized) mixed models representation of penalized splines. This function handles additive models as well. To perform adaptive smoothing we had to integrate the Fisher scoring procedure (9) for θ with updates of the remaining parameters by subsequent calls of function `spm`. The current version of our package "AdaptFit" with the function `asp` is available at http://cran.r-project.org. The function `asp` is similar to that of function `spm` but may use, in addition, adaptive smoothing. For example, estimation of the function $m_1(x)$ described in Section 2.5 can be performed by

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

Switching between maximum likelihood and restricted maximum likelihood estimation can be done by specifying `spar.method="ML"`. In the additive model case some

components of the model can be fitted non-adaptively. Other examples are provided within the package.

Acknowledgements

The work was conducted while the first author was affiliated to the University Bielefeld. She wishes to thank for a productive working environment. The fist and third author are also indebted to the Deutsche Forschungsgemeinschaft (DFG) for partial support.

References

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

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

Crainiceanu, C., Ruppert, D., Carroll, R., Adarsh, J., and Goodner, B. (2006). Spatially adaptive Bayesian P-splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, (to appear).

Crainiceanu, C., Ruppert, D., and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *Journal of statistical software* **14(14)**.

DiMatteo, I., Genovese, R., and Kass, R. (2001). Bayesian curve-fitting with free-knots splines. *Biometrika* **88**, 1055–1071.

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

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B* **57**, 371–394.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modelling. *Technometrics* **31**, 3–39.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association.* **72**, 320–338.

Herrmann, E. (1997). Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics* **6**, 35–54.

Kass, R. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayesian models). *Journal of the American Statistical Association.* **84**, 717–726.

Kauermann, G. (2004). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planing and Inference* **127**, 53–69.

Kauermann, G. and Ortlieb, R. (2004). Temporal pattern in the number of staff on sick leave: the effect of downsizing. *Journal of the Royal Statistical Society, Series C* **53**, 353–367.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Laird, N. and Louis, T. (1987). Empirical Bayes confidence intervals based on bootstrap samples (with discussion). *Journal of the American Statistical Association.* **82**, 739–757.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13**, 183–212.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association* **92**, 107–116.

Morris, C. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). *Journal of the American Statistical Association.* **78**, 47–65.

Ngo, L. and Wand, M. (2004). Smoothing with mixed model software. *Journal of statistical software* **9(1)**.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (c/r: P519-527). *Statistical Science* **1**, 502–518.

Pintore, A., Speckman, P., and Holmes, C. C. (2005). Spatially adaptive smoothing splines. *Biometrika*, (to appear).

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **42**, 205–224.

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

Searle, S., Casella, G., , and McCulloch, C. (1992). *Variance Components*. Wiley.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics* **18**, 223–249.

Wood, S., Jiang, W., and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika* **89**, 513–528.

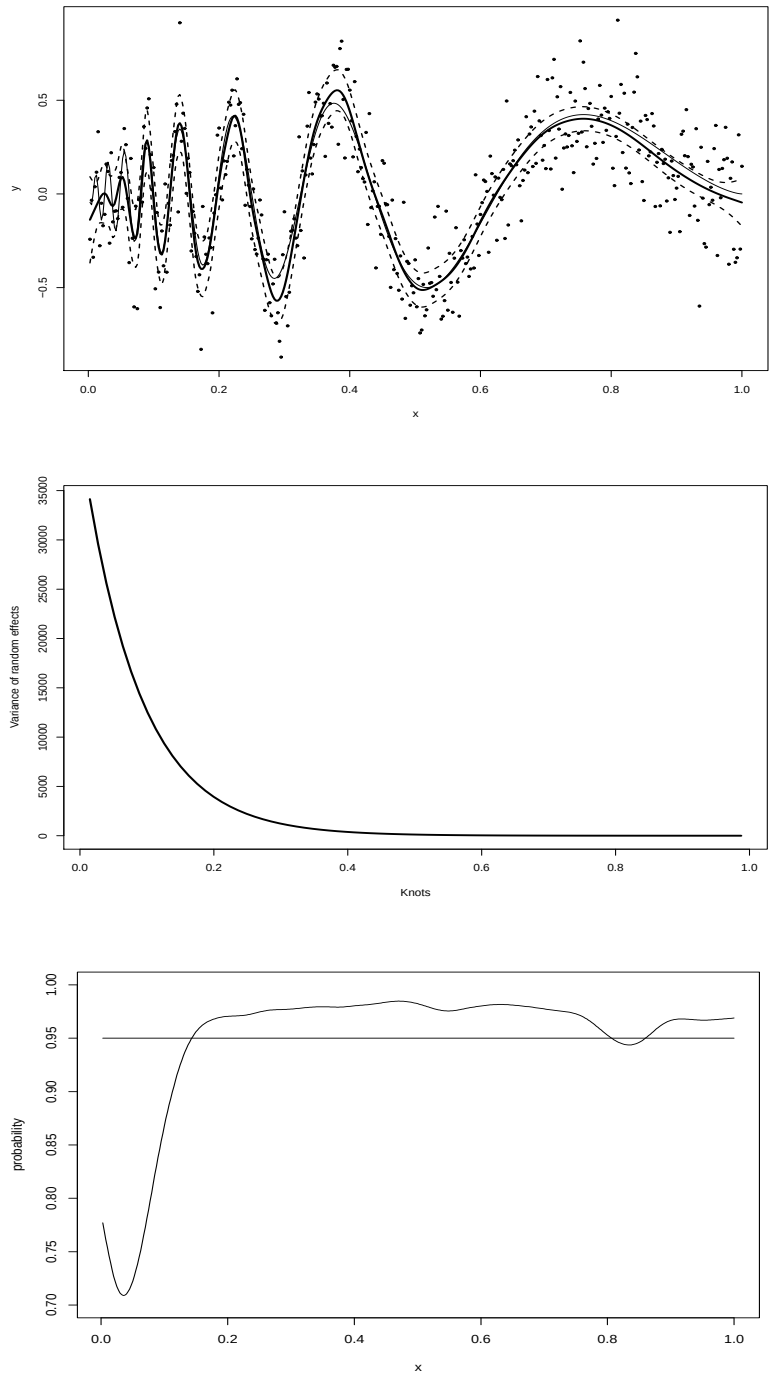

Figure 1: Top: Estimated regression function $m_1(x)$ (bold) with confidence intervals (dashed) and true function. Middle: Estimated variance of random effects. Bottom: Smoothed pointwise coverage probabilities of 95% confidence intervals for 500 simulated datasets.

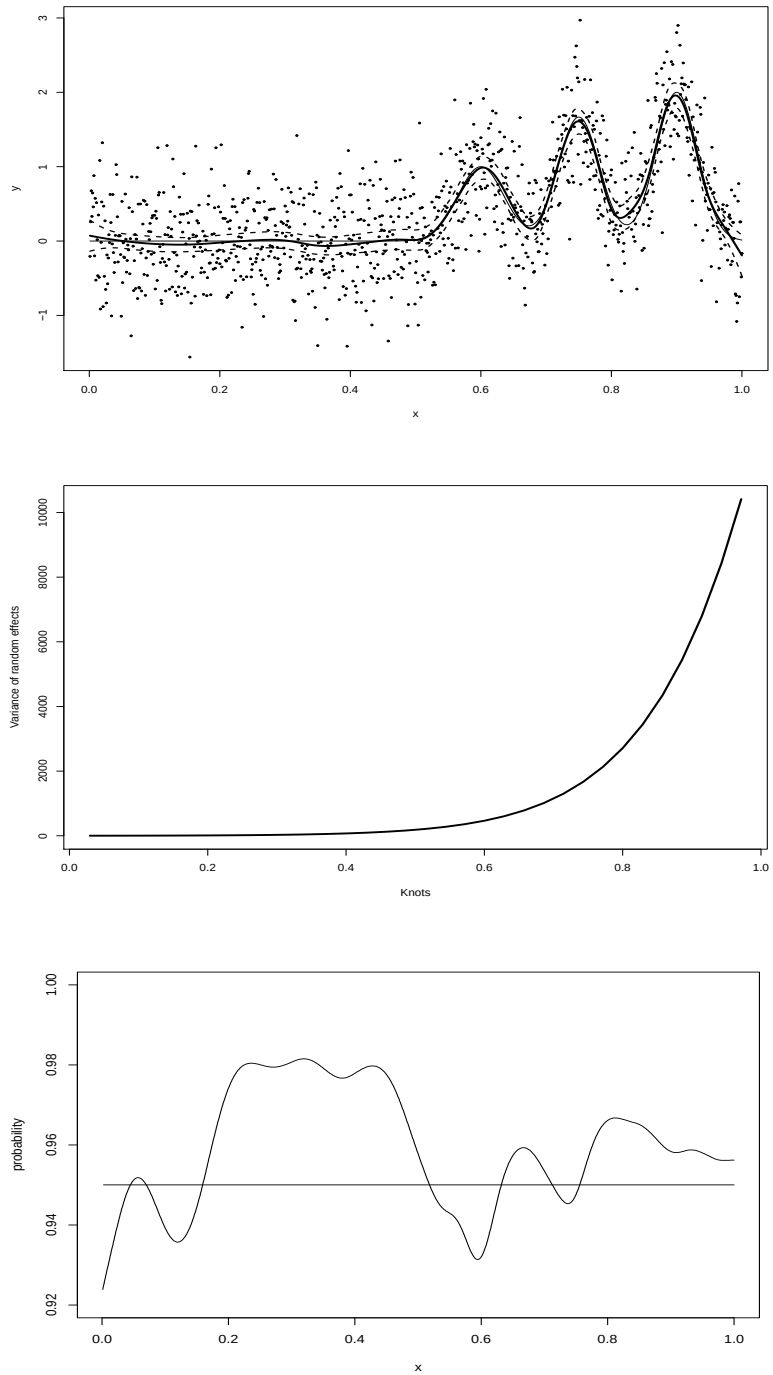

Figure 2: Top: Estimated regression function $m_2(x)$ (bold) with confidence intervals (dashed) and true function. Middle: Estimated variance of random effects. Bottom: Smoothed pointwise coverage probabilities of 95% confidence intervals for 500 simulated datasets.

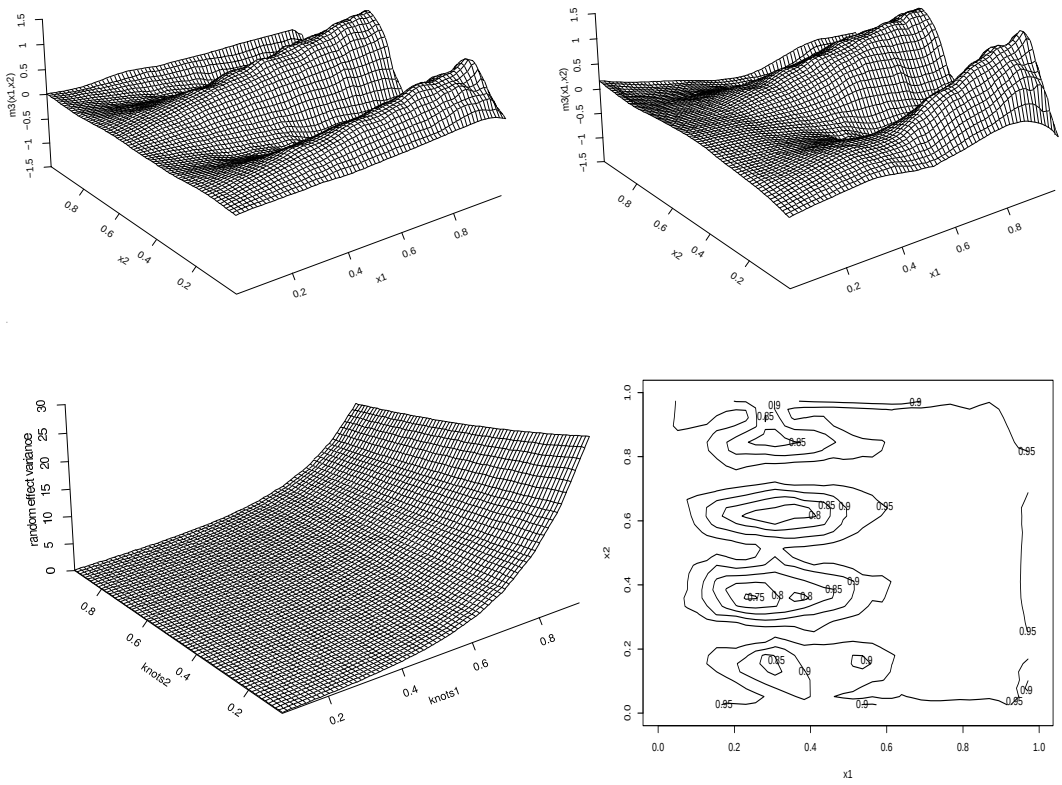

Figure 3: Top row: True regression function $m_3(x_1, x_2)$ (left) and its estimate with adaptive smoothing parameter (right). Bottom row: Estimated variance of random effects (left) and smoothed coverage probability of 95% confidence intervals for 500 simulated datasets (right).

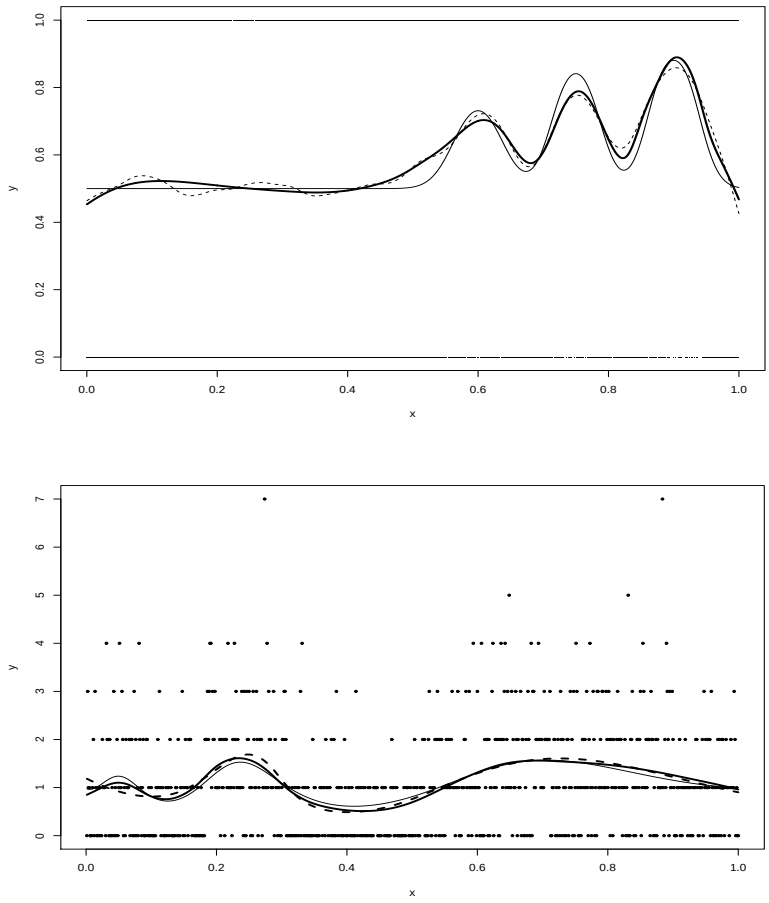

Figure 4: Top: Estimated regression function $\pi = \text{logit}^{-1}[m_2(x)]$ with adaptive penalty (bold), with global smoothing parameter (dashed) and true function for 5000 binary data (middle). Bottom: Estimated regression function $\exp[m_1(x)]$ based on our adaptive approach (bold) and BARS (dashed).

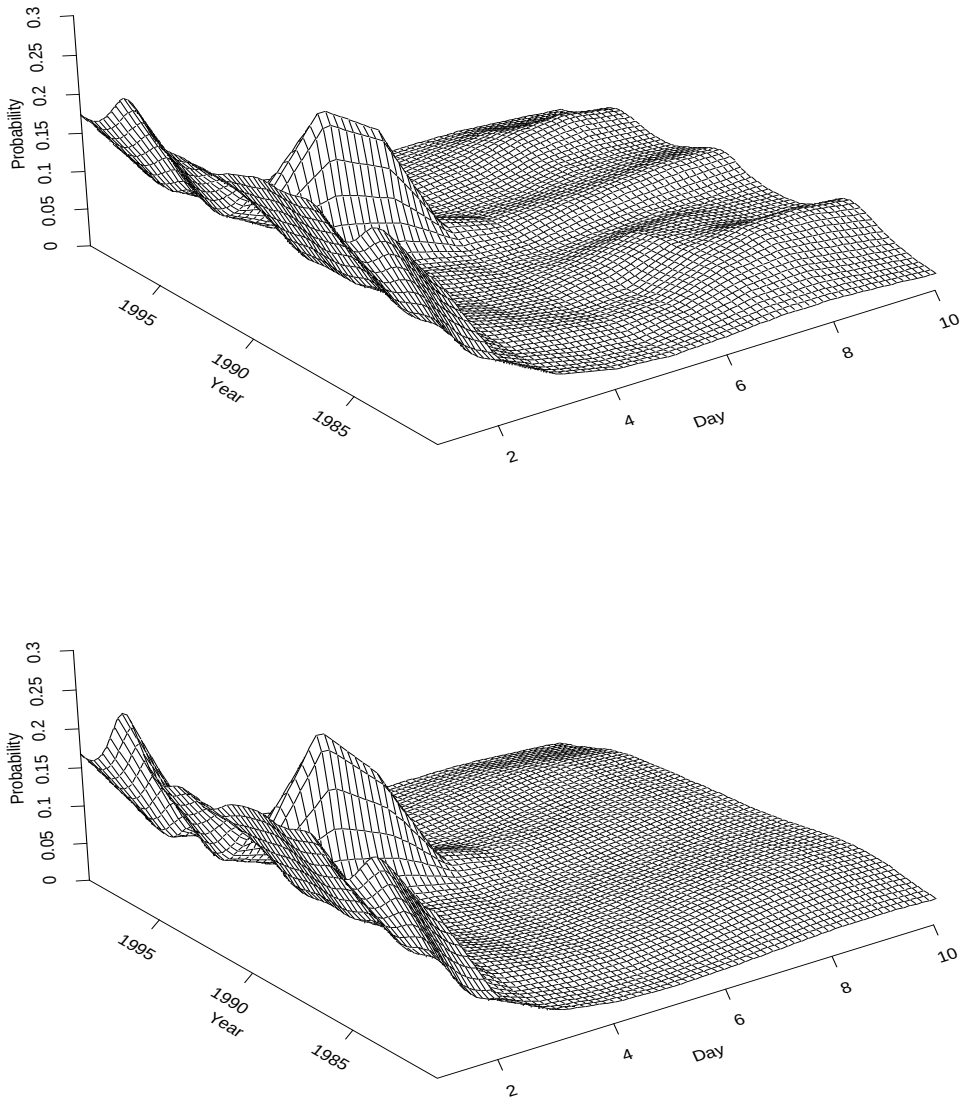

Figure 5: Estimated regression function $P(d = t|d \geq t, c) = \text{logit}^{-1}[m(t, c)]$ with global (top) and adaptive (bottom) smoothing parameter.