Some Asymptotic Results on Generalized Penalized Spline Smoothing

Göran Kauermann
University of Bielefeld

Tatyana Krivobokova
Katholieke Universiteit Leuven

Ludwig Fahrmeir
Ludwig-Maximilians-Universität München

20th March 2008

Abstract

The paper discusses asymptotic properties of penalized spline smoothing if the spline basis increases with the sample size. The proof is provided in a generalized smoothing model allowing for non-normal responses. The results are extended in two ways. First, assuming the spline coefficients to be a priori normally distributed links the smoothing framework to generalized linear mixed models (GLMM). We consider the asymptotic rates such that Laplace approximation is justified and the resulting fits in the mixed model correspond to penalized spline estimates. Secondly, we make use of a fully Bayesian viewpoint by imposing a priori distribution on all parameters and coefficients. We argue that with the postulated rates at which the spline basis dimension increases with the sample size the posterior distribution of the spline coefficients is approximately normal. The validity of this result is investigated in finite samples by comparing Markov Chain Monte Carlo (MCMC) results with their asymptotic approximation in a simulation study.

1 Introduction

Recent years have seen an increasing use of penalized spline estimation as smoothing technique. Originally suggested by O'Sullivan (1986), the approach has achieved general attention with the paper by Eilers & Marx (1996) who phrased the routine as P-spline smoothing. A general introduction and a description of the flexibility of penalized spline smoothing is found in Ruppert, Wand & Carroll (2003). Even though penalized splines are practically convincing, theoretical investigations of their performance and properties are less explored. A recent investigation is found in Opsomer & Hall (2005) who reformulate the approach as white noise representation. Some first results were provided in Wand (1999) and Aerts, Claeskens & Wand (2002) who use simplifying assumption that the dimension of the spline basis is fixed. Though this is a stringent assumption in theoretical terms, it has little practical impact if the dimension of the spline basis is chosen in lush and generous manner, see Ruppert (2002). The theoretical advantage of fixing the number of spline functions in advance is that asymptotically one achieves a parametric model and penalization looses its influence.

In this paper, we start from a penalized spline approach, but allow the number of spline basis functions to depend on the sample size. Recently, Claeskens, Krivobokova & Opsomer (2008) showed that depending on the assumption formulated for the number of knots the asymptotic properties of penalized splines are either similar to those of regression splines (for a "small" number of knots) or to those of smoothing splines (for a "large" number of knots), with a clear breakpoint between two asymptotic scenarios. Cardot (2002) considered penalized splines with adaptive penalties and presented some results in the first asymptotic scenario with a "small" number of knots. Recently, Li & Ruppert (2008) provided first theoretical results in the second asymptotic scenario with a "large" number of knots, deriving equivalence between kernel smoothing and penalized splines. All these results are based on a normal response model. We go a step towards generalized response models of the

form

$$\mu(x) = \mathrm{E}\{y(x)\} = h\{\eta(x)\}, \tag{1.1}$$

with x as a continuous covariate and y as response, assumed to be distributed according to an exponential family distribution. Function $h(\cdot)$ is a known invertible (inverse) link function while function $\eta(x)$ is supposed to be smooth and will be estimated via penalized spline smoothing. In this paper we pursue the asymptotic scenario with a "small" number of knots growing with the sample size.

Penalized spline smoothing has an interesting link to mixed models, by comprehending the penalty imposed on the spline coefficients as a Gaussian prior, see Wand (2003). In this case, the smoothing parameter steering the amount of penalization becomes the ratio of the dispersion parameter over the a priori variance of the random spline effect. This has the practical impact that smoothing parameter selection can now be carried out by Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) estimation (see e.g. McCulloch & Searle, 2001 or Kauermann, 2005). If a generalized smoothing model like (1.1) is assumed, penalized spline fitting can be linked to generalized linear mixed models (GLMM), again by writing the penalty as a priori normal distribution on the spline coefficients. Integrating out the random spline coefficients using a Laplace approximation is then equivalent to a penalized spline fit (see also Wang, 1998 or Lin & Zhang, 1999 for the connection in case of classical spline smoothing). In general, the equivalence of penalized spline fitting and generalized linear mixed models is asymptotically justified only if the Laplace approximation holds. It has been shown in Breslow & Lin (1995) or more generally in Shun & McCullagh (1995) that Laplace approximation can fail for clustered data in generalized linear mixed models. The asymptotic scenario for penalized spline smoothing is, however, conceptionally different to clustered data. Here splines play the role of clus-

ters and the number of spline bases functions is small compared to the sample size n, while the number of observations for each spline is increasing with the sample size. This, however, is exactly the condition in which Laplace approximation works (see Severini, 2000). In this paper we investigate how the number of spline coefficients may increase without disturbing the accuracy of the Laplace approximation.

The mixed model approach to penalized splines smoothing can also be interpreted from an empirical Bayes viewpoint. This can be extended by taking completely the Bayesian perspective, that is assuming all parameters to have a prior distribution. Based on Fahrmeir, Kneib & Lang (2004) we consider this fully Bayes approach. Exact finite sample size posterior inference can be carried out using MCMC simulation. Numerically this is available for instance with the software package BayesX (Brezger, Kneib & Lang, 2005) which is available from `www.stat.uni-muenchen.de/~bayesx/`. The MCMC calculation provides posterior distributions without relying on the Laplace or other asymptotic approximations. Our investigation focuses the question whether approximative numerically less demanding methods based on Laplace, and exact results, based on MCMC, are comparable with respect to their accuracy. We investigate the difference between Laplace approximation and MCMC theoretically as well as in simulations. It is a standard result in Bayesian analysis that the posterior converges with increasing sample size to a normal distribution with mean as Maximum Likelihood estimate (the mode) and the variance as inverse Fisher matrix, as long as the dimension of the parameter space is fixed, see for instance Bernardo & Smith (2005, chapter 5.3). Our results go in this direction, but the dimension of the parameter space increases with the sample size, since the spline basis dimension is allowed to grow. We explore the derived results also empirically through a simulation study. Our results confirm that the approximation is accurate.

The paper is organized as follows. Section 2 discusses the generalized smoothing model while Section 3 investigates the mixed model formulation. Section

4 looks at the Bayesian perspective before a short discussion concludes the paper. Some technical details are found in the Appendix.

2 Generalized P-Spline Smoothing

We consider the generalized smoothing model (1.1) where y for given x is assumed to follow an exponential family distribution with notation

$$y|x \sim \exp\left\{\frac{y\vartheta(x) - b\{\vartheta(x)\}}{\phi} + c(y,\phi)\right\}, \tag{2.2}$$

with $\vartheta(x) = \vartheta\{\eta(x)\}$ as the natural parameter of the underlying exponential family and ϕ as dispersion parameter. Functions $b(\cdot)$ and $c(\cdot)$ are determined by the distribution. For simplicity we ignore the role of the dispersion parameters in (2.2) and set $\phi \equiv 1$. Functions $\vartheta(x)$ and $\mu(x)$ stand in the unique relationship $b'(\vartheta) = \mu$, so that $\vartheta\{\eta(x)\} = b'^{-1}[h\{\eta(x)\}]$. Choosing the link function $h(\cdot) = b'(\cdot)$ provides the natural link. We observe the independent observations $(x_i, y_i), i = 1, \ldots, n$. Function $\eta(x)$ is assumed to be smooth in x and for fitting we decompose $\eta(x)$ to

$$\eta(x) = X(x)\beta + Z(x)u + \delta(x), \tag{2.3}$$

where $\delta(x) = \eta(x) - \{X(x)\beta + Z(x)u\}$ will be called *approximation bias* subsequently. The vector $X(x)$ is thereby a low dimensional polynomial basis, i.e. $X(x) = (1, x, x^2/2, \ldots, x^q/q!)$, while $Z(x)$ is high-dimensional, built from *truncated polynomials*, i.e.

$$Z(x) = \left\{\frac{(x-\tau_1)_+^q}{q!}, \ldots, \frac{(x-\tau_{k-1})_+^q}{q!}\right\},$$

where $(x)_+^q = x^q$ for $x > 0$ and zero otherwise and $0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = 1$. To distinguish between vectors and matrices we follow the following notation. Lower case letters with indices refer to scalars, lower case

letters without indices refer to column vectors, while capital letters without indices denote row vectors. Finally, matrices will be denoted with bold capital letters. Following this convention and ignoring the approximation bias $\delta(x)$ we obtain the log likelihood

$$l(\theta) = \sum_{i=1}^{n} y_i \vartheta(P_i\theta) - b\left\{\vartheta(P_i\theta)\right\}, \tag{2.4}$$

where $P_i = P(x_i) = \{X(x_i), Z(x_i)\} = (X_i, Z_i)$ and $\theta = (\beta^T, u^T)^T$. Maximizing $l(\theta)$ will lead to a wiggled estimate if the spline dimension $k + q$ is large. Therefore, a penalty is imposed on θ. For truncated polynomials one can employ a simple shrinkage, that is we consider the penalized likelihood

$$l_p(\theta, \lambda) = l(\theta) - \frac{\lambda}{2} u^T u, \tag{2.5}$$

where λ is the smoothing or penalty parameter. Both, $l_p(.)$ as well as $l(.)$ depend on the sample size n which is suppressed in our notation for simplicity of presentation. The penalty in (2.5) can also be written as $\theta^T \mathbf{D}_k \theta$ where $\mathbf{D}_k$ is a block diagonal with zero entries in the upper left $(q+1) \times (q+1)$ block and identity matrix $\mathbf{I}_{k-1}$ in the bottom right block. Increasing λ to infinity leads to a purely parametric fit with a q-th order polynomial.

Our model is formulated with truncated polynomials in order to utilize the straightforward connection of such representation to the mixed and Bayesian models. However, the use of *B-splines* (de Boor, 2001) is more advisable numerically and also allows for simple handling of theoretical developments. In fact, both approaches are equivalent in the following sense (see also Hämmerlin & Hoffmann, 1992). We define with $\mathbf{P}_{q,k}$ the n by $(q+k)$ dimensional truncated spline basis with rows

$$P_i = P_{q,k}(x_i) = \left(1, x_i, \frac{x_i^2}{2!}, \ldots, \frac{x_i^q}{q!}, \frac{(x_i - \tau_1)_+^q}{q!}, \ldots, \frac{(x_i - \tau_{k-1})_+^q}{q!}\right),$$

$i = 1, \ldots, n$. From $\mathbf{P}_{q,k}$ we can construct the normed B-spline basis via $\mathbf{B}_{q,k} = k^q \mathbf{P}_{q,k} \mathbf{L}_{q,k}$ where $\mathbf{L}_{q,k}$ is a $(q+k) \times (q+k)$ dimensional invertible matrix constructed from the $(q+1)$ order difference matrix (see Fahrmeir, Kneib & Lang, 2004 or Claeskens, Krivobokova & Opsomer, 2008 for more details). The spline representation can now be written as $\mathbf{P}_{q,k}\theta = \mathbf{B}_{q,k}\omega$ with $\omega = k^{-q}\mathbf{L}_{q,k}^{-1}\theta$ as coefficient vector for the B-spline basis. Note that the coefficient vectors θ and ω both depend on k which is suppressed for the ease of notation. We can now formulate the penalized likelihood (2.5) in terms of parameter vector ω leading to

$$l_p(\omega, \lambda) = l(\omega) - \frac{\lambda k^{2q}}{2} \omega^T \tilde{\mathbf{D}}_k \omega, \tag{2.6}$$

where $\tilde{\mathbf{D}}_k = \mathbf{L}_{q,k}^T \mathbf{D}_k \mathbf{L}_{q,k}$. Note that $\tilde{\mathbf{D}}_k$ does not have full rank.
We will now investigate how k may grow with increasing sample size, that is we allow k to depend on n. To do so we will make the following assumptions

(A1) We assume that design points x_i are distributed according to a design density with compact support on $[0, 1]$. This implies that the distance between two adjacent values x_i and x_j, say, converges to zero with order $O(n^{-1})$.

(A2) The knots for the spline basis are equidistantly distributed (for simplicity) so that $0 = \tau_0 < \tau_1 < \ldots < \tau_{k-1} < \tau_k = 1$ with $\tau_j - \tau_{j-1} = k^{-1}$ for $j = 1, \ldots, k$.

(A3) The penalty parameter λ is assumed to grow with the sample size with order

$$\lambda = O\left(n^{\gamma}\right), \ \gamma \leq \frac{2}{2q+3} \tag{2.7}$$

(A4) We assume that the dimension of the spline basis grows with the sample

size with order

$$k \sim C n^{\frac{1}{2q+3}}, \tag{2.8}$$

for some constant $C > 0$.

(A5) Function $\eta(x)$ is assumed to be $(q+1)$ times differentiable and except of a finite number of isolated points in $[0,1]$ it is continuously differentiable. Finally, $\eta(x)$ is bounded so that $\mu(x) = h\{\eta(x)\}$ is in the interior of the mean parameter space for all x. This guarantees that the likelihood contributions are all of the same asymptotic order $O_p(1)$.

Let $\mathbf{P}_{q,k}\theta_0 = \mathbf{B}_{q,k}\omega_0$ be the best spline approximation of the unknown function $\eta(x)$ based on a Kullback Leibler measure, that is $\theta_0 = \text{argmax}\, \mathrm{E}\{l(\theta)|\eta\}$ or equivalently $\omega_0 = k^{-q}L_{q,k}^{-1}\theta_0 = \text{argmax}\, \mathrm{E}\{l(\omega)|\eta\}$, where the expectation is calculated with the unknown predictor $\eta(x)$. Accordingly, we define with $\delta_0(x) = \eta(x) - P_{q,k}(x)\theta_0 = \eta(x) - B_{q,k}(x)\omega_0$ the smallest approximation bias with $B_{q,k}(x)$ as B-spline basis evaluated at x. Note that θ_0 and ω_0, respectively, depend on k and therewith on n, which is suppressed notationally. We can now decompose the Mean Squared Error to

$$\mathrm{MSE}\{\hat{\eta}(x)\} \;=\; \mathrm{E}\left[\{\hat{\eta}(x) - B_{q,k}(x)\omega_0\}^2\right] + \delta_0^2(x) - 2\delta_0(x)\mathrm{E}\{\hat{\eta}(x) - B_{q,k}(x)\omega_0\}.$$

The first component mirrors a conventional Mean Squared Error in penalized parametric regression, while the remaining two components include the approximation bias. The central result of this section can now be stated as follows.

Theorem 1 With assumptions (A1) to (A5) we find that the penalized estimate $\hat{\eta}(x) = B_{q,k}(x)\hat{\omega}$ obtained from (2.6) is consistent with the Mean Squared Error of order

$$\mathrm{MSE}\{\hat{\eta}(x)\} = O\left(n^{-\frac{2q+2}{2q+3}}\right).$$

In particular we can expand the estimate $\hat{\eta}(x)$ as

$$\hat{\eta}(x) - \eta(x) = \left[B_{q,k}(x)\mathbf{F}(\lambda)^{-1} \left\{ \frac{\partial l(\omega)}{\partial \omega} - \lambda k^{2q} \tilde{\mathbf{D}}_k \omega \right\} - \delta(x) \right] \{1 + o_p(1)\} \quad (2.9)$$

with $\mathbf{F}(\lambda) = \mathrm{E}\left(-\partial^2 l(\omega)/\partial\omega\partial\omega^T + \lambda k^{2q}\tilde{\mathbf{D}}_k \right)$. The leading stochastic component in (2.9) has the asymptotic order $O_p\left(n^{-\frac{1}{2}\frac{2q+2}{2q+3}}\right)$.

The proof of the theorem is provided in the Appendix.

Remarks

1. Based on the expansions we can use (2.9) to derive an approximate distribution for the estimate. Using the central limit theorem we get

$$\hat{\eta}(x) - \eta(x) \overset{a}{\sim} \mathcal{N}\left[\mathrm{bias}\left\{ \hat{\eta}(x) \right\}, \mathrm{Var}\left\{ \hat{\eta}(x) \right\} \right], \quad (2.10)$$

with $\mathrm{bias}\{\hat{\eta}(x)\} = -B_{q,k}(x)\mathbf{F}(\lambda)^{-1}\lambda k^{2q}\tilde{\mathbf{D}}_k\omega_0 - \delta(x)$ and

$$\mathrm{Var}\{\hat{\eta}(x)\} = B_{q,k}(x)\mathbf{F}^{-1}(\lambda)\mathbf{F}(\lambda = 0)\mathbf{F}^{-1}(\lambda)B_{q,k}^T(x). \quad (2.11)$$

2. The variance of $\hat{\eta}(x)$ is build in a sandwich form from Fisher type matrices. Due to the fact that the dimension k of ω grows with the sample size, the dimension of the Fisher matrix grows as well. It is shown in the Appendix that the sandwich type variance in (2.11) is decreasing to zero with order $O(k/n)$.

3 P-Spline Smoothing and Mixed Models

3.1 Laplace Approximation

Penalized spline smoothing can be linked to mixed models by comprehending the penalty as *a priori* normal distribution on the spline coefficients. We show

now that the penalized estimate is asymptotically equivalent to the posterior Bayes estimate resulting in the mixed model. This equivalence holds exactly in the normal response model with identity link and normal distribution imposed on the spline coefficients. The smoothing parameter λ plays the role of the ratio of the residual variance and the prior variance of the spline coefficients. Consequently, based on the mixed model, the smoothing parameter can be estimated by maximizing the marginal likelihood, or an adjusted version of it yielding a Restricted Maximum Likelihood estimate (REML). For generalized response models, however, integration over the spline coefficients is not available analytically and alternative methods have to be used. The link to penalized spline estimation results by pursuing a Laplace approximation. The latter is justified asymptotically only, if the remaining correction terms converge to zero with growing sample size. In the following section we show that the Laplace approximation is justified if we assume the spline dimension to grow with the previously proposed order $k \sim Cn^{\frac{1}{2q+3}}$.

We now model spline coefficient vector u as a priori normally distributed. Moreover, we assume in this section for the sake of simplicity that link function $h(.)$ is the canonical link. This leads to the generalized linear mixed model (GLMM)

$$E(y|u) = h(\mathbf{X}\beta + \mathbf{Z}u), \quad u \sim N\left(0, \sigma_u^2 \mathbf{I}_{k-1}\right), \tag{3.12}$$

with $y = (y_1, \ldots, y_n)^T$ and $\mathbf{X}$ and $\mathbf{Z}$ as matrices with rows X_i and Z_i, respectively. Integrating out the random spline effects leads to the marginal likelihood (up to a constant)

$$L(\beta, \sigma_u^2) = \sigma_u^{-(k-1)} \int_{R^{k-1}} \exp[-g(u)]du, \tag{3.13}$$

with $g(u) = -y^T(\mathbf{X}\beta + \mathbf{Z}u) + 1_n^T b(\mathbf{X}\beta + \mathbf{Z}u) + u^T u/(2\sigma_u^2)$, with $1_n = (1, \ldots, 1)^T$. The integral in (3.13) does not generally have an analytic so-

lution. We therefore make use of a Laplace approximation to obtain the marginal likelihood (3.13). Note that $g(u) = g(\beta, u, \sigma_u^2)$, that is $g(u)$ depends on other quantities as well which are omitted in (3.13). It is not difficult to see that $\partial g(\hat{\beta}, \hat{u}, \sigma_u^2)/\partial(\beta, u) = 0$ defines the penalized estimating equation $\partial l_p(\theta, \lambda)/\partial\theta = 0$ with $l_p(\theta, \lambda)$ as defined in (2.5) and $\lambda = \sigma_u^{-2}$ playing the role of the penalization parameter. Instead of deriving a Laplace approximation for the integral (3.13) directly, we use a B-spline formulation for technical reasons. Let therefore the difference matrix $\mathbf{L}_{q,k}$ from above be decomposed as

$$\mathbf{L}_{q,k} = \begin{pmatrix} \mathbf{L}_{11} & \mathbf{L}_{12} \\ \mathbf{L}_{21} & \mathbf{L}_{22} \end{pmatrix},$$

according to the dimension of β and u, i.e. $\mathbf{L}_{11} \in \mathbb{R}^{(q+1)\times(q+1)}$. Since the elements of $\mathbf{L}_{12}$ are all equal to zero it is easy to see that $\mathbf{P}_{k,q}\theta = \mathbf{B}_{q,k}\omega$ can be represented as

$$\mathbf{X}\beta + \mathbf{Z}u = k^q(\mathbf{X}\mathbf{L}_{11} + \mathbf{Z}\mathbf{L}_{21})\omega_1 + k^q\mathbf{Z}\mathbf{L}_{22}\omega_2 =: \mathbf{B}_{q,k,1}\omega_1 + \mathbf{B}_{q,k,2}\omega_2, \quad (3.14)$$

with $\omega_1 := k^{-q}\mathbf{L}_{11}^{-1}\beta$ and $\omega_2 := k^{-q}\mathbf{L}_{22}^{-1}(u - \mathbf{L}_{21}\mathbf{L}_{11}^{-1}\beta)$. In this notation the integral (3.13) takes the form

$$L(\beta, \sigma_u^2) = \sigma_u^{-(k-1)}k^{(k-1)q}|\mathbf{L}_{22}| \int_{R^{k-1}} \exp\{-\tilde{g}(\omega_2)\}d\omega_2, \quad (3.15)$$

where $\tilde{g}(\omega_2) := \tilde{g}(\omega_1, \omega_2) = g\{\theta(\omega)\} = g(\beta, u)$. The integral in (3.15) is approximated using a Laplace approximation by

$$\int_{R^{k-1}} \exp\{-\tilde{g}(\omega_2)\}d\omega_2 = |\tilde{\mathbf{G}}|^{-1/2}(2\pi)^{\frac{(k-1)}{2}} \exp\{-\tilde{g}(\hat{\omega}_2)\} \{1 + O(\varepsilon_0)\}, \quad (3.16)$$

where $\tilde{\mathbf{G}} = \tilde{\mathbf{G}}(\hat{\omega}_2)$ denotes the second order derivative $\partial^2\tilde{g}(\hat{\omega}_2)/\partial\omega_2\partial\omega_2^T$, evaluated at $\hat{\omega}_2$ that minimizes $\tilde{g}(.)$. The objective is now to evaluate the asymp-

totic order of the correction term ε_0. Let $\tilde{g}_{jl}$ denote the (j,l)-th element of $\tilde{\mathbf{G}}$. Accordingly, third and forth order derivatives of $\tilde{g}(\cdot)$ are denoted by $\tilde{g}_{jlr}$ and $\tilde{g}_{jlrs}$, respectively. Moreover with $\tilde{g}^{jl}$ we refer to the (i,j)-th element of the inverse of $\tilde{\mathbf{G}}$. Following the results provided in Barndorff-Nielsen & Cox (1989) or Shun & McCullagh (1995) and using Einstein's summation convention we can write the correction term in (3.16) as

$$\varepsilon_0 \;=\; -\tilde{g}_{jlrs}\tilde{g}^{jl}\tilde{g}^{rs}[3]/24 + \tilde{g}_{jlr}\tilde{g}_{stv}\left(\tilde{g}^{jl}\tilde{g}^{rs}\tilde{g}^{tv}[9] + \tilde{g}^{js}\tilde{g}^{lt}\tilde{g}^{rv}[6]\right)/72.$$

In (3.17) equal super and subscript imply a summation over the corresponding indices and the bracketed terms refer to the (number of) possible permutations over the indices, e.g. the first component in (3.17) is a short form for $1/24\tilde{g}_{jlrs}\left(\tilde{g}^{jl}\tilde{g}^{rs}+\;\tilde{g}^{jr}\tilde{g}^{ls}+\tilde{g}^{js}\tilde{g}^{rl}\right)$. The objective is now to show that the term ε_0 vanishes asymptotically with the sample size n increasing. Let $\mathbf{W}$ be the diagonal matrix of the conditional variances $\mathrm{Var}(y_i|u)$, that is $\mathbf{W} = \mathrm{diag}\left\{b''(\mathbf{B}_{q,k}\omega)\right\}$ and we denote with $\hat{\mathbf{W}}$ matrix $\mathbf{W}$ with ω_2 replaced by $\hat{\omega}_2$. Note that

$$\tilde{g}_{jl} = B^T_{q,k,j}\hat{\mathbf{W}}B_{q,k,l} + \frac{k^{2q}}{\sigma_u^2}(\mathbf{L}_{22}^T\mathbf{L}_{22})_{jl}, \tag{3.17}$$

with $B_{q,k,l}$ denoting the l-th column of $\mathbf{B}_{q,k,2}$ defined in (3.14) and $(\mathbf{L}_{22}^T\mathbf{L}_{22})_{jl}$ as (j,l)-th element of $\mathbf{L}_{22}^T\mathbf{L}_{22}$. With assumptions (A1) and (A2) we obtain that the number of non zero elements for each column of the spline basis $\mathbf{B}_{q,k}$ is of order $O(n/k)$. Consequently the first element in (3.17) equals 0 if $|j-l| > q$ and is of order $O_p(n/k)$ otherwise. Considering the definition of $\mathbf{L}_{22}^T\mathbf{L}_{22}$ we find that the second component in (3.17) has similar structure and takes values 0 if $|j-l| > q+1$ and has order $O_p(k^{2q}/\sigma_u^2)$ otherwise. Hence

$$\tilde{g}_{jl} = \begin{cases} O_p\left(n/k + k^{2q}/\sigma_u^2\right), & |j-l| \le q+1 \\ 0, & \text{otherwise} \end{cases}$$

For higher order derivatives we get with the same arguments

$$\tilde{g}_{jlr} = \begin{cases} O_p\left(n/k\right), & |j-l| \leq q \text{ and } |j-r| \leq q \text{ and } |l-r| \leq q \\ 0, & \text{otherwise} \end{cases}$$

and according results for $\tilde{g}_{jlrs}$. Considering the inverse matrix $\tilde{g}^{jl}$ we can make use of results derived in Demko (1977). In the line of the arguments of Remark 10 in the Appendix we get

$$\tilde{g}^{jl} = \rho^{|j-l|} O_p \left\{ \left(\frac{n}{k} + k^{2q} \sigma_u^{-2} \right)^{-1} \right\}$$

for some $0 < \rho < 1$. The proof is in line with the arguments used to derive (A.32) in the Appendix and therefore not explicitly listed here again.
These orders imply that ε_0 has the order

$$\varepsilon_0 = O_p \left\{ n \left(\frac{n}{k} + \frac{k^{2q}}{\sigma_u^2} \right)^{-2} \right\} + O_p \left\{ n^2 \left(\frac{n}{k} + \frac{k^{2q}}{\sigma_u^2} \right)^{-3} \right\}. \tag{3.18}$$

The second component in (3.18) is asymptotically dominating, and letting now k grow with order $n^{\frac{1}{2q+3}}$ allows to rewrite the asymptotic order of (3.18) to

$$\varepsilon_0 = O_p \left\{ n^{-\frac{2q}{2q+3}} \left(1 + n^{-\frac{2}{2q+3}} \sigma_u^{-2} \right)^{-3} \right\}.$$

If we set $\sigma_u^{-2} = O(n^\gamma)$ with $\gamma \leq 2/(2q+3)$ we get $\varepsilon_0 = O_p\left(n^{-\frac{2q}{2q+3}}\right)$ so that the Laplace approximation is asymptotically justified for $q > 0$. The condition imposed on σ_u^2 resembles assumption (A3) in the previous Section. We therefore reformulate (A3) to

(A3′) The *a priori* variance σ_u^2 is assumed to have the asymptotic order

$$\sigma_u^{-2} = O\left(n^\gamma\right), \; \gamma \leq \frac{2}{2q+3} \tag{3.19}$$

It remains to demonstrate that this assumption is sound and justified which will be discussed in Section 3.3 below. Due to the equivalence of B-splines and truncated polynomials the result transfers directly to the original formulation (3.12) with truncated polynomials. The latter is formulated in the following theorem.

Theorem 2 With assumptions (A1), (A2), (A3′) and (A4) we find that the marginal likelihood function of the generalized linear mixed model (3.12) can be approximated using Laplace approximation, that is

$$L(\beta, \sigma_u^2) = \left[\sigma_u^{-(k-1)}|\mathbf{G}|^{-1/2} \exp\left\{-g(\beta, \hat{u}, \sigma_u^2)\right\}\right] \{1 + o_p(1)\}, \quad (3.20)$$

with $g(\hat{u}, \beta, \sigma_u^2) = -y^T(\mathbf{X}\beta + \mathbf{Z}\hat{u}) + 1_n^T b(\mathbf{X}\beta + \mathbf{Z}\hat{u}) + \hat{u}^T\hat{u}/2\sigma_u^2$ where $y = (y_1, \ldots, y_n)^T$ and $\hat{u}$ as minimizer of $g(\beta, u, \sigma_u^2)$. Matrix $\mathbf{G}$ is defined through $\partial^2 g(\beta, \hat{u}, \sigma_u^2)/\partial u \partial u^T = \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + \mathbf{I}_{k-1}/\sigma_u^2$.

Remarks

3. Shun & McCullagh (1995) showed that the Laplace approximation of a likelihood for some k dimensional parameter based on n data points (from an exponential family) is reliable provided that $k = o(n^{1/3})$. This is satisfied for our choice $k \sim Cn^{\frac{1}{2q+3}}$, given $q > 0$, but for $q = 0$ the Laplace approximation fails.

3.2 Posterior Cumulants

For further estimation it is common to ignore the dependence of $\hat{\mathbf{W}}$ on β (and $\hat{u}$). This is motivated in Breslow & Clayton (1993) and can be justified since the dependence is mirrored in higher order asymptotic terms only. We will therefore subsequently treat matrix $\hat{\mathbf{W}}$ as deterministic and look now more generally at posterior cumulants of the spline coefficients based on the generalized linear mixed model (3.12). Based on (3.13), the corresponding

moment generating function is defined through

$$M_{\omega_2|y}(t) = \frac{\int_{R^{k-1}} \exp(t^T\omega_2)\exp\{-\tilde{g}(\omega_2)\}d\omega_2}{\int_{R^{k-1}} \exp\{-\tilde{g}(\omega_2)\}d\omega_2}. \tag{3.21}$$

Following the results from above, the denominator in (3.21) can be approximated by (3.16). Applying Laplace approximation in the same style to the nominator of (3.21) (see Barndorff-Nielsen & Cox, 1989) we obtain for the numerator in (3.21)

$$(2\pi)^{(k-1)/2}|\tilde{\mathbf{G}}|^{-1/2}\exp\{-\tilde{g}(\hat{\omega}_2)\}\left[M_z(t) + \exp(t^T\hat{\omega}_2)O_p\left\{\varepsilon_0 + \varepsilon_1(t)\right\}\right],$$

where $M_z(t)$ denotes the moment generating function of the normally distributed random variable $z \sim N(\hat{\omega}_2, \tilde{\mathbf{V}})$ with $\tilde{\mathbf{V}} = \tilde{\mathbf{G}}^{-1}$. The correction term ε_0 is defined in (3.17) and $\varepsilon_1(t)$ results to $\varepsilon_1(t) = \hat{\tilde{g}}_{jlr}t_s\hat{\tilde{g}}^{jl}\hat{\tilde{g}}^{rs}[3]/6$, where t_s is the s-th elements in t. Note that ε_0 and $\varepsilon_1(t)$ are of the same asymptotic order for any fixed value of $t > 0$. Applying Laplace approximation to denominator and nominator of (3.21) gives

$$M_{\omega_2|y}(t) = M_z(t)\frac{1 + \exp(-t^T\tilde{\mathbf{V}}t/2)O_p\left\{\varepsilon_0 + \varepsilon_1(t)\right\}}{1 + O_p(\varepsilon_0)}.$$

and the corresponding cumulant generating function can be written as

$$K_{\omega_2|y}(t) = K_z(t) + \tilde{H}(t) + O_p(\varepsilon_0), \tag{3.22}$$

with $\tilde{H}(t) = \exp(-t^T\tilde{\mathbf{V}}t/2)O_p\left\{\varepsilon_0 + \varepsilon_1(t)\right\}$. Hence, the pth derivative of $\tilde{H}(t)$ with respect to $t_{j_1}, ..., t_{j_p}$ evaluated at $t = 0$ defines the difference between the pth order posterior cumulant of the ω_2 given y and the approximate cumulant of the $N(\hat{\omega}_2, \tilde{\mathbf{V}})$. We will now show that the derivatives of $\tilde{H}(t)$ are asymptotically negligible. Using the subscript notation from above, that is $\tilde{H}_{j_1,...,j_p}(t) = \partial^p\tilde{H}(t)/\partial t_{j_1}...\partial t_{j_p}$ and bearing in mind that the derivatives of

$O_p\{\varepsilon_0 + \varepsilon_1(t)\}$ with respect to t are equal to zero for $p > 2$, we obtain

$$\begin{aligned}\tilde{H}_{j_1,\ldots,j_p}(t) &= [\exp(-t^T\tilde{\mathbf{V}}t/2)]_{j_1,\ldots,j_p} O_p\{\varepsilon_0 + \varepsilon_1(t)\} \\ &+ p[\exp(-t^T\tilde{\mathbf{V}}t/2)]_{j_1,\ldots,j_{p-1}} O_p(\varepsilon_{j_p}),\end{aligned}$$

with $\varepsilon_{j_p} = \tilde{g}_{rst}\tilde{g}^{rs}\tilde{g}^{tj_p}[3]/6$. From standard results on the multivariate normal distribution we find $[\exp(-t^T\tilde{\mathbf{V}}t/2)]_{j_1,\ldots,j_p} = \tilde{h}_{j_1,\ldots,j_p}\exp(-t^T\tilde{\mathbf{V}}t/2)$, where $\tilde{h}_{j_1,\ldots,j_p}$ are the Hermite tensors (see McCullagh, 1987, pages 149-151). We are now interested in $\tilde{H}_{j_1,\ldots,j_p}(t=0)$ and since $\tilde{h}_j(t=0) = 0$ we immediately obtain

$$\begin{aligned}\tilde{H}_j(0) &= O(\varepsilon_j) \quad , \quad \tilde{H}_{jr}(0) = O(-\tilde{v}_{jr}\varepsilon_0) \\ \tilde{H}_{jrs}(0) &= O(-3\tilde{v}_{jr}\varepsilon_s) \quad , \quad \tilde{H}_{jrst}(0) = O(\tilde{v}_{jr}\tilde{v}_{st}[3]\varepsilon_0)\end{aligned}$$

and so on, where $\tilde{v}_{jr}$ denotes the (j,r)-th element of matrix $\tilde{\mathbf{V}} = \tilde{\mathbf{G}}^{-1}$. Since $\tilde{v}_{jr} = \tilde{g}^{jr} = \rho^{|j-r|}O(k/n)$ for some $\rho \in (0,1)$, see also remark 10 in the Appendix, and $\varepsilon_j = O(k/n)$, it results that $\tilde{H}(t)$ is asymptotically negligible in (3.22) and posterior cumulants can be approximated by the cumulant generating function of a normal distribution. In particular we have $E(\omega_2^j|y) \approx \omega_2^j$ and $\mathrm{Cov}(\omega_2^j, \omega_2^r|y) \approx \tilde{g}^{jr}$. Using the connection between ω and $\theta = (\beta^T, u^T)^T$ we find that $\hat{u}$ approximates the posterior mean and $\mathbf{V} = \mathbf{G}^{-1}$ approximates the posterior variance of u given y (and β and σ_u^2). Higher oder cumulants tend to zero.

3.3 Maximum Likelihood Estimation

We have assumed above that the inverse *a priori* variance σ_u^{-2} increases with order $O(n^\gamma)$, $\gamma \leq 2/(2q+3)$. We will now demonstrate that this rate of convergence is sound by looking at the Maximum Likelihood estimate. Based on (3.20) the leading term in the Laplace approximated log likelihood is

written as

$$l(\beta, \sigma_u^2) \approx -\frac{1}{2}\log|\mathbf{G}| - g(\hat{u}) - \frac{k-1}{2}\log\sigma_u^2, \tag{3.23}$$

with $\mathbf{G} = \mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z} + \mathbf{I}_{k-1}/\sigma_u^2$. Inserting the estimate for β and differentiating (3.23) with respect to σ_u^2 yields

$$\begin{aligned} \frac{\partial l(\hat{\beta}, \sigma_u^2)}{\partial \sigma_u^2} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \sigma_u^2}\right) + \frac{\hat{u}^T\hat{u}}{2\sigma_u^4} - \frac{k-1}{2\sigma_u^2} && (3.24) \\ &= \frac{1}{2\sigma_u^2}\left\{\frac{\hat{u}^T\hat{u}}{\sigma_u^2} - df(\sigma_u^2)\right\}, && (3.25) \end{aligned}$$

where $df(\sigma_u^2) = \mathrm{tr}\left\{\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\right\}$. We thereby ignored the dependence of $\mathbf{W}$ on β (and σ_u^2), as this leads to correction terms of negligible asymptotic order. We get from (3.24) to (3.25) by reflecting the definition of $\mathbf{G}$ and using the fact that $\mathrm{tr}(\mathbf{G}^{-1}\mathbf{G}) = k-1$. The estimate is now defined through

$$\hat{\sigma}_u^2 = \frac{\hat{u}^T\hat{u}}{df(\sigma_u^2)} = \frac{\hat{\theta}^T\mathbf{D}_k\hat{\theta}}{df(\sigma_u^2)} = k^{2q}\frac{\hat{\omega}^T\tilde{\mathbf{D}}_k\hat{\omega}}{df(\sigma_u^2)}. \tag{3.26}$$

It should be remarked that (3.26) does not provide an analytic estimate, since the right hand side of the equation contains the unknown parameter as well. For our investigation we can however make use of (3.26) be treating σ_u^2 on the right hand side as true *a priori* variance. It is not difficult to see that $\mathrm{E}(\hat{\sigma}_u^2) = \sigma_u^2$, so that we investigate the variance, expressed here as Fisher matrix. Tedious calculations yield thereby

$$\begin{aligned} \frac{\partial^2 l(\beta, \sigma_u^2)}{\partial \sigma_u^2 \partial \sigma_u^2} &= \left[\frac{1}{2\sigma_u^4} + \frac{1}{2\sigma_u^2}\right]\frac{\partial l(\beta, \sigma_u^2)}{\partial \sigma_u^2} && (3.27) \\ &+ \frac{1}{2\sigma_u^4}\left[\mathrm{tr}(\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}) - \frac{2}{\sigma_u^2}\hat{u}^T\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\hat{u}\right] \end{aligned}$$

Our intention is to show that variational coefficient $\mathrm{Var}(\hat{\sigma}_u^2)/\sigma_u^4$ (and thus the $\mathrm{Var}(\hat{\sigma}_u^2)$) is decreasing to zero if σ_u^{-2} has the above assumed order. To do so, we look at the Fisher information. Note that the first component in (3.27) has zero expectation. We also derived in the previous section that $\hat{u} \approx \mathrm{E}_u(u|y)$ so that with $\mathrm{E}_u(u) = 0$ we get $\mathrm{E}_y(\hat{u}\hat{u}^T) \approx \mathrm{Var}_y\{\mathrm{E}_u(u|y)\}$. Moreover we have shown that $\mathbf{G}^{-1} \approx \mathrm{Var}_u(u|y)$ which does not depend on y. This yields $\mathrm{E}_y(\hat{u}\hat{u}^T) = \sigma_u^2\mathbf{I}_{k-1} - \mathbf{G}^{-1}$ and in turn $E(\hat{u}^T\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\hat{u}/\sigma_u^2) = \mathrm{tr}(\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z})$. Using the relationship $k^q\mathbf{Z}\mathbf{L}_{22} = \mathbf{B}_{q,k,2}$ as defined in (3.14) we get the equality $\mathrm{tr}(\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}\mathbf{G}^{-1}\mathbf{Z}^T\hat{\mathbf{W}}\mathbf{Z}) = \mathrm{tr}(\tilde{\mathbf{G}}^{-1}\mathbf{B}_{q,k,2}^T\hat{\mathbf{W}}\mathbf{B}_{q,k,2}\tilde{\mathbf{G}}^{-1}\mathbf{B}_{q,k,2}^T\hat{\mathbf{W}}\mathbf{B}_{q,k,2})$, which allows now to apply similar arguments as used before to calculate the asymptotic order of the Fisher information. To be specific, we get

$$\begin{aligned}
\mathrm{E}\left\{-\frac{\partial^2 l(\beta,\sigma_u^2)}{\partial\sigma_u^2\partial\sigma_u^2}\right\} &= \frac{1}{2\sigma_u^4}\mathrm{tr}(\tilde{\mathbf{G}}^{-1}\mathbf{B}_{q,k,2}^T\hat{\mathbf{W}}\mathbf{B}_{q,k,2}\tilde{\mathbf{G}}^{-1}\mathbf{B}_{q,k,2}^T\hat{\mathbf{W}}\mathbf{B}_{q,k,2}) \\
&= O\left\{\sigma_u^{-4}n^{\frac{1}{2q+3}}\left(1+\sigma_u^{-2}n^{-\frac{2}{2q+3}}\right)^{-2}\right\}.
\end{aligned}$$

For the variational coefficient this leads to

$$\frac{\mathrm{Var}(\hat{\sigma}_u^2)}{\sigma_u^4} = O\left\{n^{-\frac{1}{2q+3}}\left(1+\sigma_u^{-2}n^{-\frac{2}{2q+3}}\right)^{2}\right\},$$

which tends to zero if σ_u^{-2} is of order $O(n^\gamma)$ with $\gamma \leq \frac{2}{2q+3}$ and the ML estimate for σ_u^2 is consistent.

4 Bayesian P-Spline Smoothing and MCMC Inference

4.1 Asymptotic Bayesian Framework

The mixed model approach of Section 3 can be interpreted as *empirical Bayes* inference for penalized splines smoothing. In the truncated polynomial representation $\mathbf{X}\beta + \mathbf{Z}u$ coefficient u is taken as random with a Gaussian prior while β is t considered as a fixed, unknown parameter vector which is, together with σ_u^2 estimated through maximum likelihood based on the Laplace approximation. We will now go a step ahead by taking a *fully Bayesian* perspective and consider both, β and σ_u^2 to be random with appropriate priors. Our interest is thereby to investigate the posterior distributions of β, u and σ_u^2. We will see, that an approximate posterior normal distribution results, even though the spline dimension is growing with the sample size. Let us start by formulating a prior for coefficient β. It is standard to specify either a noninformative flat prior $p(\beta) \propto 1$ or weakly informative Gaussian. For spline coefficients u we make use of the normal prior as used in the mixed model above, that is

$$p(u|\sigma_u^2) \propto \exp\left(-\frac{1}{2\sigma_u^2}u^T u\right). \tag{4.28}$$

Finally for the remaining parameter σ_u^2 we make use of a weakly informative inverse Gamma prior $IG(a_u, b_u)$, i.e.

$$p(\sigma_u^2) \propto \frac{1}{(\sigma_u^2)^{a_u+1}} \exp\left(-\frac{b_u}{\sigma_u^2}\right), \tag{4.29}$$

with small hyperparameters $a_u = b_u$ (e.g. with values 0.001), see e.g. Lang & Brezger (2004). Note that the asymptotic scenario implies that the number of components in u increase with growing sample size. This is, apparently,

a non standard setting in the fully Bayesian model as the parameter space changes with the sample size. To relate the fully Bayesian setting to the results derived above we need a coherent formulation of assumption (A3′) which is given as follows.

(A3″) We assume that σ_u^2 has a prior distribution with the constraint

$$\sigma_u^{-2} = O_p\left(n^{\gamma}\right),\ \gamma\ \leq \frac{2}{2q+3}.$$

It is not difficult to show that (A3″) is fulfilled for the prior (4.29) if $a_u > 0$ and $b_u = o\left(n^{-\gamma}\right)$. Finally, in the normal response regression model we will also assign inverse Gamma or flat priors for the unknown variance σ_ϵ^2 of the errors. Fully Bayesian inference is now based on the posterior $p(\theta, \sigma_u^2|y)$, where the additional residual variance σ_ϵ^2 occurs in the normal response model. As remarked before, for computational reasons, a B-spline basis representation can be advantageous. In this case we assume for $\omega = k^{-q}\mathbf{L}_{q,k}\theta$ the Gaussian prior $p(\omega) \propto \exp\left(-k^{2q}\omega^T\tilde{\mathbf{D}}\omega/2\sigma_u^2\right)$ which is partially improper, with rank deficiency equal to the dimension $q+1$ of β. Fahrmeir & Kneib (2006) provide theorems guaranteeing propriety of the posterior under fairly mild regularity conditions, which are fulfilled here.

Inference is carried out via MCMC simulation, drawing iteratively from the full conditionals $p(\theta|\sigma_u^2, \sigma_\epsilon^2; y)$, $p(\sigma_u^2|\theta, \sigma_\epsilon^2; y)$ and $p(\sigma_\epsilon^2|\theta, \sigma_u^2; y)$. In the Gaussian case, the full conditional for θ is Gaussian again, and the full conditionals for σ_u^2 and σ_ϵ^2 are inverse Gamma, so that a Gibbs sampler results. In the non-Gaussian case, σ_ϵ^2 is fixed, whereas the full conditional for θ has no analytic form and sampling can be done through Metropolis-Hastings steps, see Brezger & Lang (2006) for details. After a burn in phase, the sample $\{\theta^{(t)}, t = 1, ..., T\}$ can be used to (approximately) compute the marginal posterior density $p(\theta|y)$ through its empirical density. Our intention is now to compare the fully Bayesian approach to its approximate version which can be derived from the results from above. In principle we could take the prior

for β and (4.28) and apply a Laplace approximation as done in the previous section. The only difference is that integration in (3.13) is not only carried out over u but also over β. Since β is low dimensional it is a classical result in Bayesian statistics that the posterior is approximately normal with $\hat{\beta}$ as posterior mean and the Fisher matrix as posterior variance. In the same line we can generalize the results derived in (3.21) to derive the posterior distribution for $\theta = (\beta^T, u^T)^T$ or $\omega = (\omega_1^T, \omega_2^T)^T$, respectively if integration takes place over θ or ω. In fact, with assumptions (A1), (A2), (A3′) and (A4) we find

$$\theta|y, \sigma_u^2 \overset{a}{\sim} N\left(\hat{\theta}, \hat{\mathbf{V}}\right) \quad \text{with } \hat{\mathbf{V}} = \left\{\mathbf{P}_{q,k}^T \hat{\mathbf{W}} \mathbf{P}_{q,k} + \text{diag}(0_{q+1}, 1_{k-1}\sigma_u^{-2})\right\}^{-1} \tag{4.30}$$

This result shows that the conditional posterior (or full conditional) $p(\theta|y, \sigma_u^2)$ is approximately normal for given σ_u^2, with σ_u^2 obeying assumption (A3′). Fully Bayesian inference is based, however, on the marginal posterior $p(\theta|y) = \int p(\theta|y, \sigma_u^2)p(\sigma_u^2|y)\mathrm{d}\sigma_u^2$ after integrating out σ_u^2. Even in a classical Bayesian linear model setting, the marginal posterior does not necessarily follow a standard distribution, see e.g. O'Hagan & Foster (2004). Avoiding numerical integration, samples from this marginal posterior can be obtained through MCMC with Metropolis-Hastings steps, drawing from the full conditionals (i) $p(\theta|y, \sigma_u^2)$ and (ii) $p(\sigma_u^2|y, \theta)$. For σ_u^2 fulfilling (A3″) we may use the normal approximation (4.30) for $p(\theta|y, \sigma_u^2)$. This leads to an approximate but simple MCMC scheme with Gibbs steps instead of Metropolis-Hastings steps used for (i). It remains to check that the sample drawn from (ii) also fulfills the order (A3″). In this case we may use (4.30) iteratively in the MCMC steps. Note that $p(\sigma_u^2|y, \theta) \propto p(y|\theta)p(\theta|\sigma_u^2)p(\sigma_u^2)$. Using the normal distribution (4.28) for $p(\theta|\sigma_u^2)$ it follow directly that with σ_u^2 fulfilling (A3″) we have $\sigma_u^{-2}|y, \theta = O_p(n^\gamma)$, $\gamma \leq \frac{2}{2q+3}$.

A Bayesian approach avoiding MCMC at all has been recently suggested by Rue & Martino (2005) by combining a Laplace approximation for the posterior $p(\sigma_u^2|y)$ with numerical integration, see also Rue & Held (2005). A

simple approximation is to replace σ_u^2 in (4.30) by its posterior mode estimate, which corresponds to a (restricted) ML estimate. By doing so one is back in a frequentist penalized likelihood setting. We compare the true marginal distribution based on the MCMC sample with the approximate estimates derived in the upper way which use Laplace approximations and setting σ_u^2 to its ML estimate. The corresponding simulation results are discussed in the next section.

4.2 Simulation Study

To explore the theoretical findings empirically we run a small simulation study. We simulate $n = 500$ data points from the binomial model

$$\text{logit}\left\{P(y_i = 1|x_i)\right\} = \sin(2\pi x_i)$$

with x_i as equidistantly distributed on $[0, 1]$, $i = 1, \ldots, n$. For fitting we make use of a truncated linear basis (i.e. $q = 1$) with $K = 30$ knots. The model is fitted in two ways, first following the mixed model framework of section 3 we apply a Laplace approximation and estimate the remaining parameters, that is β and σ_u^2, by maximum likelihood. Secondly, we follow the fully Bayesian framework as described above by using an MCMC approach (with burn in sample size 2000 and a Markov chain of length 52.000, storing every 50th simulation as draw from the posterior distribution providing 1000 replicants of the posterior distribution). The resulting fits based on the posterior mean are shown in Figure 1. Our focus is on the posterior distribution of $\eta(x)|y = P_{q,k}(x)\theta$ with θ as random given $y = (y_1, \ldots, y_n)$. The theoretical arguments derived above state that $\eta(x)|y$ is approximative normal with mean $\hat{\eta}(x) = P_{q,k}(x)\hat{\theta}$ obtained from the Laplace approximation and variance $P_{q,k}(x)\hat{\mathbf{V}}P_{q,k}^T(x)$. To check this approximation we now compare the true posterior obtained from the MCMC output with the approximating distribution. We repeat the simulation 50 times, with a MCMC sample of the above

stated size in each simulation. For each repeated simulation we calculate at locations $x = 0.25$, $x = 0.5$, $x = 0.75$ and $x = 1$ the Kolmogorov-Smirnov statistics between the (simulated) posterior distribution based on the MCMC sample, standardized with the moments obtained from the Laplace approximation and standard normal distribution. The corresponding values of the statistics are shown in Figure 2, boxplot a). We now modify the simulation setting by using $K = 80$ knots (boxplot b) and increasing the sample size to $n = 1000$, with $K = 30$ and $K = 80$ knots, respectively (boxplots c and d). As reference we draw 1000 random variables from the standard normal and calculate the resulting Kolmogorov-Smirnov statistics (boxplot e). Overall we see accordance to the theoretical findings that the Laplace approximation provides a usable alternative to the full MCMC approach. This has been found empirically in numerous other examples, described for instance in the PhD thesis by Kneib (2006).

5 Conclusion

The paper shows that the Mixed Model framework and its usage for penalized spline smoothing is asymptotically justified even if the dimension of the spline basis is allowed to increase with the sample size with the rate $n^{1/(2q+3)}$, provided that $q \geq 1$. This implies that critiques published concerning the Penalized Quasi Likelihood (PQL) approach (see e.g. Breslow & Clayton, 1993, and Breslow & Lin, 1995) are not applicable in this framework of penalized spline smoothing. We derive asymptotic rates which balances bias and variance and which in the same way guarantee the equivalence between penalized spline smoothing and PQL estimation. Therewith, the use of mixed model software using PQL and Laplace approximation for smoothing is justified for non-normal response models. Moreover, a fully Bayesian formulation of the model yields approximately the same results as a Laplace approximation, again even for growing dimensions of the spline basis.

A Technical Details

A.1 Proof of Theorem 1

Our asymptotic scenario is built on the assumptions (A1) to (A5).
Before we get deeper in the proof we want to give the following remarks.

Remarks

4. An essential component in the subsequent proof is the order of the penalized Fisher matrix defined through

$$\mathbf{F}(\lambda) = \mathbf{B}_{q,k}^T \hat{\mathbf{W}} \mathbf{B}_{q,k} + \lambda k^{2q} \tilde{\mathbf{D}}_k, \tag{A.31}$$

where $\hat{\mathbf{W}}$ is the n dimensional diagonal weight matrix resulting from the variance function. It should be noted that $\mathbf{F}(\lambda)$ is band diagonal with $2q+1$ diagonal bands having elements of order $O(n/k + \lambda k^{2q})$ and the outer $2q+2$ band with elements of order $O(\lambda k^{2q})$. Inserting the order of λ and k, respectively, i.e. using (2.7) and (2.8), we find $\mathbf{F}(\lambda)$ as band diagonal matrix with elements of order $O(n^{(2q+2)/(2q+3)})$. Normalizing $\bar{\mathbf{F}}_k(\lambda) = \mathbf{F}(\lambda) n^{-(2q+1)/(2q+3)}$ we obtain from the structure of B-splines and with the penalty matrix $\tilde{\mathbf{D}}_k$ that the (j,l)-th element of the matrix $\bar{\mathbf{F}}(\lambda)$ denoted with $\bar{f}_{jl}$ is decreasing in $|j-l|$ and maximal is on the diagonal. That is to say that $\bar{\mathbf{F}}(\lambda)$ is a strictly diagonal dominant matrix. Making use of results derived in Demko (1977) we can therefore bound the elements of the inverse matrix $\bar{f}^{jl}(\lambda) \leq \text{const}\rho^{|j-l|}$ with $0 < \rho < 1$, or equivalently

$$f^{jl}(\lambda) = \rho^{|j-l|} O\left\{ \left(\frac{n}{k} + \lambda k^{2q} \right)^{-1} \right\} \tag{A.32}$$

where $f^{jl}(\lambda)$ is the (j,l)-th elements of $\mathbf{F}(\lambda)$.

For the proof of Theorem 1 we use the following notation. Let $l(\vartheta) = \sum_{i=1}^{n} y_i^T \vartheta(x_i) - b\{\vartheta(x_i)\}$ define the log-likelihood function and denote the derivative with respect to the vector $\vartheta = (\vartheta(x_1), \ldots, \vartheta(x_n))$ as $l_\vartheta(\vartheta) := \partial l(\vartheta)/\partial\vartheta = \left[y_i - \mu\{\vartheta(x_i)\}\right]_{i=1,\ldots,n}$. Accordingly we write $l_\eta(\eta)$ for the n dimensional column vector

$$l_\eta(\eta) := \frac{\partial \vartheta^T}{\partial \eta} \cdot l_\vartheta(\vartheta) = \left(\frac{\partial \vartheta(x_i)}{\partial \eta(x_i)} \left[y_i - \mu \left\{ \vartheta(x_i) \right\} \right] \right)_{i=1,\ldots,n} \tag{A.33}$$

Let $\eta_0 = \mathbf{B}_{q,k}\omega_0$, where ω_0 is the best coefficient in the sense that ω_0 minimizes the Kullback-Leibler distance, that is $E\left[\mathbf{B}_{q,k}^T \, l_\eta \left\{\vartheta(\mathbf{B}_{q,k}\omega_0)\right\}\right] = 0$, where the expectation is carried out with respect to the true function $\eta(x)$. Coefficient ω_0 defines the optimal approximation bias $\delta(x)$ in (2.3) through

$$\delta_0(x) = \eta(x) - B_{q,k}(x)\omega_0. \tag{A.34}$$

The proof of the theorem follows now by decomposing

$$\begin{aligned} \mathrm{E}\left[\{\hat{\eta}(x) - \eta(x)\}^2\right] &= \underbrace{\mathrm{E}\left[\{\hat{\eta}(x) - \eta_0(x)\}^2\right]}_{1} \\ &\quad + \underbrace{\delta_0^2(x)}_{2} - \underbrace{2\mathrm{E}\left\{\hat{\eta}(x) - \eta_0(x)\right\}\delta_0(x)}_{3}. \end{aligned} \tag{A.35}$$

We consider the separate components in (A.35). We show first convergence of $\hat{\omega}$ to ω_0. Note that the penalized estimating equation for $\hat{\omega}$ results to

$$0 = \mathbf{B}_{q,k}^T \, l_\eta \left\{\vartheta(\mathbf{B}_{q,k}\hat{\omega})\right\} - \lambda \, k^{2q} \, \tilde{\mathbf{D}}_k \hat{\omega}. \tag{A.36}$$

The subsequent proof will make use of Einstein's summation convention (see McCullagh, 1987 or Barndorff-Nielsen & Cox, 1989). To apply the technique we need some additional notation. Let the j-th component of vector ω be subsequently denoted with a superscript instead of a subscript, that

is $\omega = (\omega^1, \omega^2, \ldots, \omega^{k+q})$. With $0 = l_{p,j}(\hat{\omega})$ we denote the j-th component of equation (A.36), that is $l_{p,j}(\hat{\omega}) = \partial l_p(\omega)/\partial\omega^j|_{\omega=\hat{\omega}}$ with $l_p(\cdot)$ as defined in (2.6). The objective is now to expand $l_{p,j}(\hat{\omega})$ around $l_{p,j}(\omega_0)$. We use the convention that we omit the explicit listing of parameters if the best coefficient ω_0 is used, that is $l_{p,j} = l_{p,j}(\omega_0)$. Moreover, the hat notation $\hat{l}_{p,j}$ is used for $l_{p,j}(\hat{\omega})$. Finally, higher order derivatives are notated by multiple subscripts, e.g. $l_{p,jl} = \partial^2 l_p(\omega_0)/\partial\omega^j\partial\omega^l$. We are now able to expand $\hat{l}_{p,j}$ around $l_{p,j}$. Using the Einstein summation convention implies that equal sub and superscripts are being summed over. This allows to write the expansion as

$$0 = \hat{l}_{p,j} = l_{p,j} + l_{p,jl}(\hat{\omega}^l - \omega_0^l) + \frac{1}{2} l_{p,jlr}(\hat{\omega}^l - \omega_0^l)(\hat{\omega}^r - \omega_0^r) + \ldots \quad \text{(A.37)}$$

Solving (A.37) for $\hat{\omega}^l - \omega_0^l$ can be done with series inversion (see Barndorff-Nielsen & Cox, 1989) and we get

$$\hat{\omega}^j - \omega_0^j = -l_p^{jl} l_{p,l} - \frac{1}{2} l_p^{jlr} l_{p,l} l_{p,r} + \ldots \quad \text{(A.38)}$$

with l_p^{jl} as (j,l)-th element of the matrix inverse of $l_{p,jl}, j, l = 1, \ldots, q+k$, and $l_p^{jlr} = l_p^{js} l_p^{lt} l_p^{ru} l_{p,stu}$. The remaining components not explicitly listed in (A.37) and (A.38) are of lower asymptotic order and are therefore omitted. In the style of classical Maximum Likelihood theory (see McCullagh, 1987) we simplify (A.38) using the following arguments. First, we decompose $l_{p,jl} = f_{jl}(\lambda) + s_{jl}$, with $f_{jl}(\lambda) = f_{jl}(0) + \lambda k^{2q} \tilde{d}_{jl}$, where $f_{jl}(0)$ is the weight or Fisher matrix contribution $-\mathrm{E}(\partial^2 l(\omega_0)/\partial\omega^j\partial\omega^l)$, $\tilde{d}_{jl}$ is the jl element of $\tilde{\mathbf{D}}$ and s_{jl} is the stochastic component of the second order derivative without the penalty, i.e. $s_{jl} = l_{jl} - f_{jl}(0)$. Using assumption (A3) we find from (A.32) that $f^{jl}(\lambda) = \rho^{|j-l|} O(k/n)$. Similarly we get that matrix s_{jl} is block diagonal, with elements of order $O_p\left\{(n/k)^{1/2}\right\}$. The first component in (A.38) can then be simplified using

$$l_p^{jl} \;=\; f^{jl}(\lambda) - f^{jr}(\lambda) f^{ls}(\lambda) s_{rs} + \ldots = f^{jl}(\lambda)\left[1 + O_p\left\{\left(\frac{n}{k}\right)^{-1}\left(\frac{n}{k}\right)^{1/2}\right\}\right].$$

With the same arguments we see that $l_{p,stu}$ is of diagonal structure, meaning that $l_{p,stu}$ is zero if $\max\{|s-t|, |s-u|, |t-u|\} > q+1$, otherwise the element has order $O(n/k)$. This allows to quantify the remaining components in (A.38) and we get with tedious but simple calculations

$$\hat{\omega}^j - \omega_0^j = f^{jl}(\lambda) l_{p,l} + o_p\left(\frac{k}{n}\right) \tag{A.39}$$

With (A.39) we can now also rewrite the leading component in (A.39) in matrix notation

$$\hat{\omega} - \omega_0 = \mathbf{F}^{-1}(\lambda)\left(\mathbf{B}_{q,k}^T l_\eta - \lambda\, k^{2q}\tilde{\mathbf{D}}_k\omega_0\right) + \ldots \tag{A.40}$$

with $l_\eta = l_\eta\{\vartheta(\mathbf{B}_{q,k}\omega_0)\}$. Note that $k^{2q}\tilde{\mathbf{D}}_k\omega_0 = k^{q-1}\mathbf{L}_{q,k}^T\mathbf{D}_k k^{q+1}\mathbf{L}_{q,k}\omega_0 = k^{q-1}\mathbf{L}_{q,k}^T\mathbf{D}_k k\theta_0$. Defining the qth order difference vector as $\eta_0^{(q)} = k\{\eta_0^{(q-1)}(\tau_q) - \eta_0^{q-1}(\tau_{q-1}), \eta_0^{(q-1)}(\tau_{q+2}) - \eta_0^{(q-1)}(\tau_{q+1}), \ldots\}$, $q > 1$ we obtain $\eta_0^{(q)}$ as a discretized version of the qth order derivative of $\eta_0(x)$. Since $k\mathbf{D}_k\theta_0 = (0_q, \eta_0^{(q+1)})$, with 0_q as q-dimensional zero vector, we obtain with (A5) and the definition of θ_0 that the elements of $k\mathbf{D}_k\theta_0$ have order $O(1)$ to achieve differentiability in the limit. Based on the structure of $\mathbf{L}_{q,k}$ this implies that $\|k^{2q}\tilde{\mathbf{D}}_k\omega_0\|_\infty = O(k^{q-1})$.
Consequently, with (A3) the Mean Squared Error for $\hat{\omega}$ has the leading terms

$$\begin{aligned} E\left[\hat{\omega} - \omega_0\right] &= -\mathbf{F}^{-1}(\lambda)\lambda k^{2q}\tilde{\mathbf{D}}_k\omega_0\left\{1 + o(1)\right\} \\ &= O\left\{\left(\frac{n}{k}\right)^{-1}\lambda k^{q-1}\right\} = O(n^{-1}k^{q+2}), \end{aligned} \tag{A.41}$$

$$\begin{aligned} \mathrm{Var}(\hat{\omega}) &= \mathbf{F}^{-1}(\lambda)\mathbf{F}(\lambda=0)\mathbf{F}^{-1}(\lambda)\left\{1 + o(1)\right\} \\ &= O\left\{\left(\frac{n}{k}\right)^{-2}\frac{n}{k}\right\} = O\left(\frac{k}{n}\right) \end{aligned} \tag{A.42}$$

The Mean Squared Error taking ω_0 as true coefficient has the order

$$\mathrm{MSE}(\hat{\omega}|\omega_0) \;=\; \mathrm{E}(\hat{\omega}-\omega_0)^2 + \mathrm{Var}(\hat{\omega}) = O(n^{-2}k^{2q+4}) + O\left(\frac{k}{n}\right)$$

which is minimized for k postulated in (A4). The Mean Squared Error for $\hat{\omega}$ is then of order $O\left(n^{-\frac{2q+2}{2q+3}}\right)$. The asymptotic orders for $\hat{\eta}(x)$ and $\hat{\omega}$ is the same.

In the second part of the proof we focus the approximation bias $\delta_0(x)$ given in (A.34). Since $\eta(x)$ is approximated in each interval $[\tau_j,\, \tau_{j+1}]$ by a polynomial of order q we find for $\eta(.)$ being $(q+1)$ times differentiable by Taylor series an approximation bias $\delta_0(x)$ of order $O(k^{-(q+1)})$. Observing the order of k given in (2.8) we find the squared approximation bias to be of order $O\left(k^{-2(q+1)}\right) = O\left(n^{-\frac{2q+2}{2q+3}}\right)$. Hence, the second and first components in (A.35) carry the same asymptotic order. The explicit formula for approximation bias as appropriate scaled Bernoulli polynomials can be found in Barrow & Smith (1978).

Finally, the third component in (A.35) results by multiplication of the bias (A.41) and the approximation bias. Keeping the above results in mind we find with the same arguments as used before, that this component is also of order $O\left(n^{-\frac{2q+2}{2q+3}}\right)$ so that (A.35) is a decomposition with elements having all the same asymptotic order.

Combining the results we get the final expansion

$$\begin{aligned}
\hat{\eta}(x) - \eta(x) &= B_{q,k}(x)\hat{\boldsymbol{\omega}} - \eta(x) && \text{(A.43)}\\
&= B_{q,k}(x)\left\{\mathbf{B}_{q,k}^T\hat{\mathbf{W}}\mathbf{B}_{q,k} + \lambda\, n^{\frac{2q}{2q+3}}\tilde{\mathbf{D}}_k\right\}^{-1} \\
&\quad \times \left\{\mathbf{B}_{q,k}^T\mathbf{l}_\eta - \lambda\, n^{\frac{2q}{2q+3}}\tilde{\mathbf{D}}_k\omega\right\} + O_p\left(n^{-\frac{q+1}{2q+3}}\right), && \text{(A.44)}
\end{aligned}$$

where $\hat{\mathbf{W}} = \mathrm{diag}\{b''(\mathbf{B}_{q,k}\hat{\omega})\}$. Finally, with (A.40) we see that the dominant stochastic part of $\hat{\omega}-\omega_0$ is $\mathbf{F}^{-1}(\lambda)\mathbf{B}_{q,l}^T l_\eta$. Since l_η is a vector of independent random variables the central limit theorem applies so that with (A.41) and (A.42) we get (2.10).

References

Aerts, M., Claeskens, G., and Wand, M. (2002). Some theory for penalized spline additive models. *Journal of Statistical Planing and Inference* **103**, 455–470.

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic Techniques for use in Statistics*. London: Chapman and Hall.

Barrow, D. L. and Smith, P. W. (1978). Asymptotic properties of best $L_2[0, 1]$ approximation by splines with variable knots. *Quarterly of Applied Mathematics 36*(3), 293–304.

Bernardo, J. and Smith, A. F. M. (2005). *Bayesian Theory*. New York: Wiley.

de Boor, C. (2001). *A Practical Guide to Splines (Revised Edition)*. Berlin: Springer.

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

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

Brezger, A., Kneib, T., and Lang, S. (2005). BayesX: Analysing Bayesian structured additive regression models. *Journal of Statistical Software* **14**, 11.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis* **50**, 967–991.

Cardot, H. (2002). Local roughness penalties for regression splines. *Computational Statistics* **17**, 89–102.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2008). Asymptotic properties of penalized spline estimators. Technical Report, Katholieke Universiteit Leuven, Faculty of Business and Economics.

Demko, S. (1977). Inverses of band matrices and local convergence of spline projections. *SIAM J. Numer. Anal.* **14(4)**, 616–619.

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

Fahrmeir, L. and Kneib, T. (2006). Propriety of posteriors in structured additive regression models: Theory and empirical evidence. *Discussion paper 510, SFB 386*.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica* **14**, 715–745.

Hämmerlin, G. and Hoffmann, K.-H. (1992). *Numerische Mathematik*. Berlin: Springer.

Kauermann, G. (2005). Penalised spline fitting in multivariable survival models with varying coefficients. *Computational Statistics and Data Analysis* **49**, 169–186.

Kneib, T. (2006). *Mixed model based inference in structured additive regression*. Ph. D. thesis, Ludwig-Maximilians-Universität München, Dr. Hut-Verlag.

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

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized slines. to appear in Biometrika.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Series B 55*(2), 381–400.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McCulloch, C. and Searle, S. (2001). *Generalized, Linear and Mixed Models*. New York: Wiley.

O'Hagan, A. and Foster, J. (2004). *Kendall's Advanced Theory of Statistics: Volume 2B: Bayesian Inference* (second ed.). Arnold Publication.

Opsomer, J. D. and Hall, P. (2005). Theory for penalised spline regression. *Biometrika* **92**, 105–118.

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

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields; Theory and Applications*. Chapman & Hall/CRC.

Rue, H. and Martino, S. (2005). Approximate inference for hierarchical Gaussian Markov random fields models. Technical Report 7/2005, Department of Mathematical Sciences, Norwegian University of Science and Technology.

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

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

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

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B* **57**, 749–760.

Wand, M. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika* **86**, 936–940.

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

Wang, Y. (1998). Mixed effects smoothing splinr analysis of variance. *Journal of the Royal Statistical Society, Series B* **60**, 159–174.

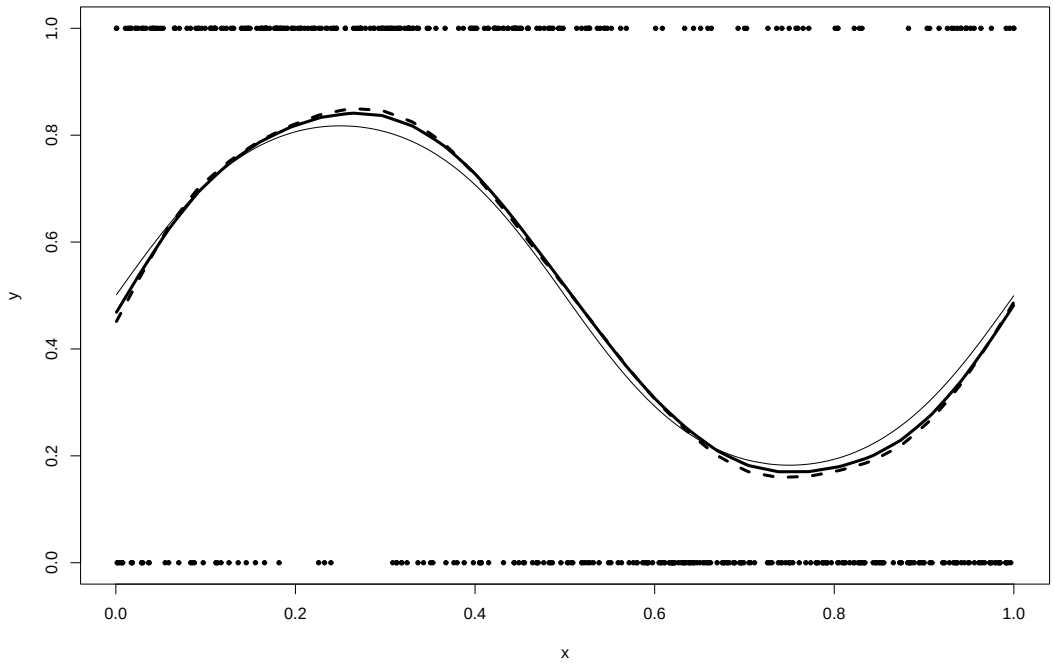

Figure 1: Simulated data and corresponding estimates based on a Laplace approximation (bold) and as mean from the MCMC posterior (dashed), respectively. The true curve is shown as thin line.

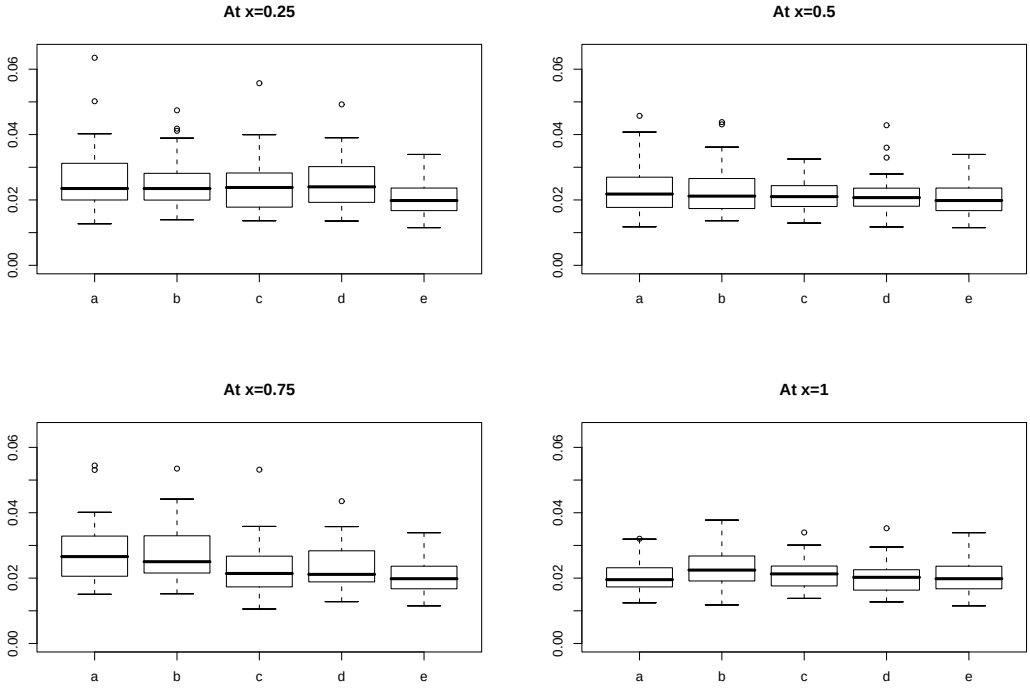

Figure 2: Kolmogorov-Smirnov statistics for comparing the posterior distribution based on MCMC standardized with the moments obtained from the Laplace approximation and the standard normal distribution a) $n = 500$ and $K = 30$, b) $n = 500$ and $K = 80$, c) $n = 1000$ and $K = 30$ and d) $n = 1000$ and $K = 80$. Plot e) shows as reference the distribution of the Kolmogorov-Smirnov statistics if random variables are drawn from a normal distribution.