University of Innsbruck

Working Papers in Economics and Statistics

Simultaneous probability statements for Bayesian P-splines

Andreas Brezger and Stefan Lang

2007-08

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by
- Department of Economics (Institut für Wirtschaftstheorie, Wirtschaftspolitik und Wirtschaftsgeschichte)
- Department of Public Finance (Institut für Finanzwissenschaft)
- Department of Statistics (Institut für Statistik)

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: +43 512 507 7151
Fax: +43 512 507 2970
e-mail: finanzwissenschaft@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://www.uibk.ac.at/fakultaeten/volkswirtschaft_und_statistik/forschung/wopec

For a list of recent papers see the backpages of this paper.

Simultaneous probability statements for Bayesian P-splines

Andreas Brezger
Hypovereinsbank München
email: andreas.brezger@hvb.de

Stefan Lang
University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria
email: stefan.lang@uibk.ac.at

Abstract

P-splines are a popular approach for fitting nonlinear effects of continuous covariates in semiparametric regression models. Recently, a Bayesian version for P-splines has been developed on the basis of Markov chain Monte Carlo simulation techniques for inference. In this work we adopt and generalize the concept of Bayesian contour probabilities to additive models with Gaussian or multicategorical responses. More specifically, we aim at computing the maximum credible level (sometimes called Bayesian p-value) for which a particular parameter vector of interest lies within the corresponding highest posterior density (HPD) region. We are particularly interested in parameter vectors that correspond to a constant, linear or more generally a polynomial fit. As an alternative to HPD regions simultaneous credible intervals could be used to define pseudo contour probabilities. Efficient algorithms for computing contour and pseudo contour probabilities are developed. The performance of the approach is assessed through simulation studies. Two applications on the determinants of undernutrition in developing countries and the health status of trees show how contour probabilities may be used in practice to assist the analyst in the model building process.

1 Introduction

The additive model

$$y_i = \eta_i + \varepsilon_i = f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

models the mean of a continuous response variable y_i as the sum of nonlinear but sufficiently smooth functions $f_1, \ldots, f_p$ of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$. It is a popular and widely used tool for refined exploratory data analysis. Part of the success is due to modern software for fitting the models, see particularly the R packages mgcv (Wood (2006), Wood (2006b)) and GAMLSS (Stasinopoulos et al. (2005)) and BayesX (Brezger et al. (2005a), Brezger et al. (2005b)). In many applications, however, the resulting estimates suggest simple functional forms for at least some of the nonlinear effects. Sometimes it is even questionable whether a particular covariate should be included in the model. Two questions arise frequently in consulting cases and in cooperations:

- The functional form of the effect of x_j is close to a linear, quadratic or cubic fit. Is it justified to replace the unspecified nonlinear function by a polynomial of low

degree in order to ease interpretation? In this context interpretation means that the *parameters* are interpretable (not only the estimated effect).

- The contribution of x_j to the fit seems to be quite small. Is a more parsimonious model with covariate x_j omitted sufficient?

Representative for a number of cooperations we discuss two applications in more detail. The first application analyzes determinants of childhood undernutrition for two African countries. The nutritional status of a child is measured by an anthropometric score (variable *zscore*) which compares the height of a child with the median height of children in a reference population of the same age. One of the most important determinants of the nutritional status is the age of the child. Figures 4 and 5 show the estimated effect of age for Zambia and Tanzania. The estimated functions are based on Bayesian P-splines as developed by Lang and Brezger (2004). For both countries a clear nonlinear effect can be observed with decreasing nutritional status until the age of 20 months and an almost constant effect thereafter. However, a more parsimonious modeling by a low order polynomial of degree two or three seems to be a reasonable alternative.

The plan of this paper is to provide methodology that assists the applied statistician to answer the questions stated above. The basis for the methodological development is a Bayesian approach to generalized additive models (and extensions) based on P(enalized)-splines as the main building block. P-splines have been originally suggested by O'Sullivan (1986) and made popular by Eilers and Marx (1996), see also Marx and Eilers (1998) and Eilers and Marx (2004). The Bayesian version is due to Lang and Brezger (2004) and Brezger and Lang (2006). Bayesian inference is carried out using modern Markov Chain Monte Carlo (MCMC) simulation techniques that allow to draw random numbers from the posterior of the model. Based on a random sample posterior quantities of interest like the posterior mean or standard deviation are easily estimated via their empirical counterparts. Pointwise credible intervals, also called Bayesian confidence intervals, are obtained by computing respective quantiles of the sampled parameters. For example a 95% credible interval is defined by the 2.5% and 97.5% empirical quantiles of the random sample. However, the decision whether a nonparametrically estimated effect could be replaced by a more parsimonious polynomial fit, requires *simultaneous probability statements* about the parameters. The primary goal of this paper is to develop techniques for obtaining such simultaneous probability statements in the context of P-splines smoothing. Our approach adapts ideas recently proposed by Held (2004) for estimating and computing contour probabilities or Bayesian p-values. The definition of contour probabilities is based on highest posterior density (HPD) regions which are constructed such that the posterior density within the region is higher than outside. Bayesian p-values are defined as the *maximum credible level* for which a particular parameter vector of interest lies within the corresponding HPD region. We are particularly interested in parameter vectors that correspond to a constant, linear or more generally a polynomial fit. The final goal is to assist the analyst in the model building process towards more parsimonious models. For instance, if the contour probability for a linear fit is small but relatively high for a quadratic fit, a more parsimonious model with a parametric linear fit could be used. As an alternative to HPD regions, simultaneous credible intervals as proposed by Besag, Green, Higdon and Mengersen (1995) could be used to define pseudo contour probabilities.

Currently methodology is available for Gaussian responses and (multi)categorical logit and probit models. As an example for multicategorical data we consider an application on the health status of trees measured in three ordered categories.

Additionally to Bayesian p-values model choice could be based on a global goodness of fit measure that allows to compare competing models. The classical instrument for comparing models in a Bayesian framework is the Bayes factor (e.g. Kass and Raftery (1995), DiCiccio et al. (1997)). Suppose we are given two competing models M_1 and M_2 with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. If the prior probabilities $p(M_j)$ are equal, i.e. $p(M_j) = 1/2$, the two models can be compared via the posterior odds, or in other words the Bayes factor,

$$BF = \frac{p(M_1|\mathbf{y})}{p(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)},$$

with

$$p(\mathbf{y}|M_j) = \int p(\mathbf{y}|\boldsymbol{\theta}_j, M_j)\, p(\boldsymbol{\theta}_j|M_j)\, d\boldsymbol{\theta}_j \quad j = 1, 2. \tag{2}$$

In many practical cases the computation of Bayes factors is difficult. An approximation is given by the Bayesian information criterion (BIC) which is defined as

$$BIC = -2l(\hat{\boldsymbol{\theta}}) + \log(n)\, q$$

where $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood evaluated at the posterior mode, n the sample size and q the number of parameters in the model. A nice justification of the BIC can be found in Hastie et al. (2003). However, the applicability of the Bayes factor and with it the BIC is restricted to models with proper prior distributions for the parameters. Since the prior distribution used for Bayesian P-splines is partially improper these tools are not available for model choice.

Another widely used tool for Bayesian model selection is the Deviance information criterion (DIC) developed by Spiegelhalter et al. (2002). Similar to the AIC the derivation of the DIC is based on information theoretic arguments. It is designed to compare complex hierarchical models where the number of parameters overestimates the complexity of the model. The DIC is based on a measure p_D for the effective number of parameters in a model as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. Specifically, $p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ where $\overline{D(\boldsymbol{\theta})}$ is the posterior mean deviance and $D(\bar{\boldsymbol{\theta}})$ is the deviance of the posterior mean of $\boldsymbol{\theta}$. Adding p_D to the posterior mean deviance gives the DIC

$$\text{DIC} = \overline{D} + p_D.$$

The computation of the DIC is well suited to simulation based inference and is obtained more or less as a by product of a MCMC sampler. The straightforward computation of the DIC is certainly one of the reasons for its widespread use. However, from a practical point of view the DIC comes with two limitations. First, the DIC is subject to sampling error. Hence, a clear cut decision between two models with similar DIC is possible (if at all) only if repeated MCMC samples are available allowing to assess the sampling error involved. Second, for every model under consideration new and time consuming MCMC samples are required to compute the DIC. This limits its applicability to the comparison of only a few competing models. It will be the Bayesian p-values suggested in this paper that will we helpful in assisting the analyst in detecting a small number of promising models that could additionally be compared via the DIC.

The reminder of the paper is organized as follows:

The first subsection of the next section reviews additive models based on P-splines and illustrates how the Bayesian version is constructed. Consecutive subsections are devoted

to computing contour or pseudo contour probabilities for P-splines. Technical details and proofs are deferred to two appendices. In section 3 the performance of the different approaches is assessed through extensive simulation studies. In section 4 results for the two applications on determinants of undernutrition and the health status of trees show how the proposed methodology could be used in practice. The final section gives a brief summary of the paper.

2 Bayesian P-Splines and model selection via contour probabilities

2.1 Bayesian P-splines

Our approach for modeling the nonlinear functions f_j in (1) is based on P(enalized)-splines. The approach assumes that the unknown functions f_j can be approximated by a polynomial spline of degree l with equally spaced knots $x_{j,min} = \kappa_{j0} < \kappa_{j1} < \cdots < \kappa_{j,r-1} < \kappa_{jr} = x_{j,max}$ within the domain of x_j. The spline can be written in terms of a linear combination of $m = r + l$ B-spline basis functions B_{jk}, i.e.

$$f_j(x_j) = \sum_{k=1}^{m} \beta_{jk} B_{jk}(x_j). \tag{3}$$

By defining the design matrices $\boldsymbol{X}_j$, where the element in row i and column k is given by $\boldsymbol{X}_j(i,k) = B_{jk}(x_{ij})$, we can rewrite the predictor in (1) in matrix notation as

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \ldots + \boldsymbol{X}_p\boldsymbol{\beta}_p.$$

The Penalized version assumes a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on squared differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. This leads to a penalized least squares approach where

$$PLS(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p) = (\mathbf{y} - \boldsymbol{\eta})'(\mathbf{y} - \boldsymbol{\eta}) + \lambda_1 \sum_{k=d+1}^{m} \Delta^d \beta_{1k} + \ldots + \lambda_p \sum_{k=d+1}^{m} \Delta^d \beta_{pk} \tag{4}$$

is minimized with respect to $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. The index d indicates the order of differences. Usually $d = 1$ or $d = 2$ is used leading to first differences $\beta_{jk} - \beta_{j,k-1}$ or second differences $\beta_{j,k} - 2\beta_{j,k-1} + \beta_{j,k-2}$, respectively. The trade off between fidelity to the data (governed by the least squares term) and smoothness (governed by the p penalty terms) is controlled by the *smoothing parameters* λ_j. The larger the smoothing parameters the smoother the resulting fit.

Recently, Lang and Brezger (2004) and Brezger and Lang (2006) developed a Bayesian version of P-splines. The Bayesian point of view has several advantages. First, we gain new insight in how the P-spline approach works. Second, it allows for a unified approach for simultaneously obtaining point and interval estimates of the regression coefficients *and* the smoothing parameters. The main difference of the Bayesian approach is that the parameters in the model are treated as random variables for which a prior distribution is specified. Note, that the assumption of a prior distribution for the parameters does not necessarily imply that the parameters are random. The prior distribution rather reflects

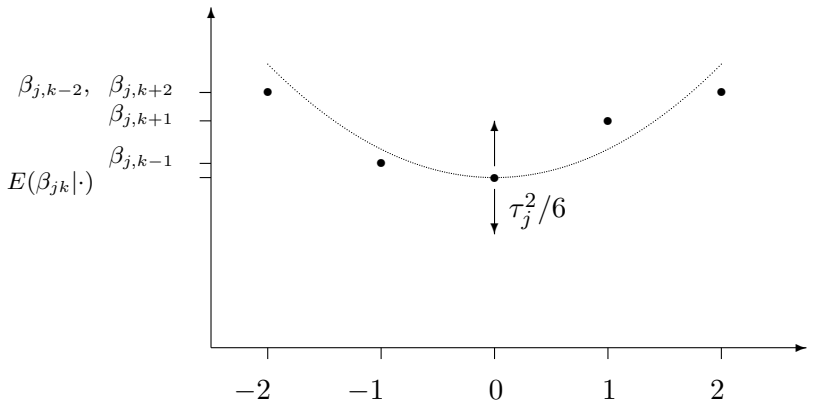

Figure 1: Illustration of the conditional mean of a parameter β_{jk} given the left and right neighbors $\beta_{j,k-2}$, $\beta_{j,k-1}$, $\beta_{j,k+1}$,$\beta_{j,k+2}$ for a second order random walk prior. The conditional distribution is Gaussian with mean $-1/6\beta_{j,k-2} + 2/3\beta_{j,k-1} + 2/3\beta_{j,k+1} - 1/6\beta_{j,k+2}$ and variance $\tau_j^2/6$.

the degree of uncertainty about the parameters. In many modern applications the prior distribution models structural assumptions, e.g. smoothness, the degree of continuity etc. The Bayesian version of P-splines is based on stochastic analogues of difference penalties as smoothness priors for the regression coefficients. More specifically, first or second order random walks are used as smoothness prior, i.e.

$$\beta_{jk} = \beta_{j,k-1} + u_{jk}, \quad \text{or} \quad \beta_{jk} = 2\beta_{j,k-1} - \beta_{j,k-2} + u_{jk} \tag{5}$$

with Gaussian errors $u_{jk} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The random walk priors could have been equivalently defined via the conditional distributions of β_{jk} given the left *and* right neighboring parameters, i.e. $\beta_{j,k-1}$, $\beta_{j,k+1}$ in case of a first order random walk and $\beta_{j,k-2}$, $\beta_{j,k-1}$,$\beta_{j,k+1}$, $\beta_{j,k+2}$ in case of a second order random walk. From (5) it follows that these conditional distributions must be Gaussian. To shed more light on the nature of the prior we study the conditional means of the regression parameters given the neighboring parameters. For both (equivalent) specifications of the prior the respective conditional means can be nicely interpreted. We restrict ourselves to second order random walks which are most frequently used in practise. The conditional mean of β_{jk} given the two left neighbors $\beta_{j,k-1}$, $\beta_{j,k-2}$ is the extrapolation of the linear trend formed by the values $\beta_{j,k-1}$, $\beta_{j,k-2}$. The conditional mean of β_{jk} given the left and right neighbors is obtained by fitting a quadratic polynomial to the four points $(\beta_{j,k-2}, -2)$, $(\beta_{j,k-1}, -1)$, $(\beta_{j,k+1}, 1)$ and $(\beta_{j,k+2}, 2)$. The conditional mean $E(\beta_{jk} \,|\, \cdot)$ is then the point on the quadratic polynomial at value 0, see figure 1.

Applying the law of total probability the priors (5) can be equivalently written in the form of a global smoothness priors

$$\boldsymbol{\beta}_j | \tau_j^2 \propto \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j' \boldsymbol{K}_j \boldsymbol{\beta}_j\right) \tag{6}$$

with penalty matrix $\boldsymbol{K}_j = \mathbf{D}'\mathbf{D}$ where $\mathbf{D}$ is a difference matrix of order one or two. The penalty matrix $\mathbf{K}_j$ is rank deficient with $rk(\mathbf{K}_j) = d - 1$ for first order differences respectively a first order random walk and $rk(\mathbf{K}_j) = d - 2$ for a second order random

walk. Hence (6) has the form of a singular normal distribution and the prior for the regression coefficients is improper. However, it can be shown that the posterior is proper, see Fahrmeir and Kneib (2006).

The variances τ_j^2 and the smoothing parameters λ_j are related by $\lambda_j = \tau_j^2/\sigma^2$ where σ^2 is the error variance. Hence, τ_j^2 may be seen as an inverse smoothing parameter. Large variances correspond to small smoothing parameters and result in more wiggled estimates. Smaller variances result in smoother estimates. To be able to estimate the amount of smoothness simultaneously with the regression parameters additional invers Gamma distributed hyperpriors $p(\tau_j^2) \sim IG(a_j, b_j)$ independent from the priors on $\boldsymbol{\beta}_j$ are assigned to the variances τ_j^2 (and the overall variance parameter σ^2). We assume $a_j = b_j = 0.01$ which corresponds to noninformative priors (on the log scale).

Bayesian inference is entirely based on the posterior distribution which is proportional to the product of the likelihood and the prior. For given variance parameters the posterior mode estimate for the regression coefficients is obtained by minimizing the PLS criterion (4). Fully Bayesian inference for all parameters involved can be based on MCMC simulation. For Gaussian responses a Gibbs sampler can be used to successively update the parameters $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p, \tau_1^2, \ldots, \tau_p^2$, see Lang and Brezger (2004) for details. For most models with categorical responses a representation of the models in terms of underlying latent continuous variables may be utilized for MCMC simulation. In this case the Gibbs sampler for Gaussian responses can be adapted to models with categorical responses. Details can be found in Albert and Chib (1993) for binary and ordinal probit models, Fahrmeir and Lang (2001) for multivariate probit models and Holmes and Held (2006) for binary and multicategorical logit models. For other responses such as Poisson regression the sampling schemes are more complicated, see Brezger and Lang (2006) for details.

Based on a sample of simulated parameters from the posterior most quantities of interest are easily estimated. The posterior mean as a point estimator is estimated by the empirical means of the simulated parameters. Pointwise credible intervals are based on the respective quantiles of sampled parameters. As has been pointed out by various authors Bayesian confidence intervals are preferable to their frequentist counterparts in terms of coverage probabilities, see Wood (2006c) for a recent account. However, a much more challenging task are *simultaneous probability statements* about the parameters. This will be the topic of the next subsection.

2.2 Contour probabilities

In order to keep the notation as simple as possible the development in this section is presented for a particular covariate x with regression parameters $\boldsymbol{\beta}$. Hence the index j in (3) and everywhere else is suppressed.

Suppose we are interested in simultaneous posterior probability statements for a particular parameter vector $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. The posterior *contour probability* $P(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})$ of $\boldsymbol{\beta}^*$ is defined as 1 minus the content of the HPD region of $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ which just covers $\boldsymbol{\beta}^*$, i.e.

$$P(\boldsymbol{\beta}^* \,|\, \boldsymbol{y}) = P\{p(\boldsymbol{\beta} \,|\, \boldsymbol{y}) \leq p(\boldsymbol{\beta}^* \,|\, \boldsymbol{y}) \,|\, \boldsymbol{y}\}, \tag{7}$$

see Box and Tiao (1973) and Held (2004). Note that $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ is treated here as a random variable. In the following we briefly review concepts for estimating the probability (7) from posterior samples $\boldsymbol{\beta}^{(t)}$, $t = 1, \ldots, T$ obtained via MCMC simulation.

Held (2004) proposes to estimate (7) by

$$\widehat{P(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{p(\boldsymbol{\beta}^{(t)} \,|\, \boldsymbol{y}) \leq p(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})\}, \tag{8}$$

i.e. the proportion of MCMC samples for which the posterior density is smaller than the density of the point of interest $\boldsymbol{\beta}^*$.

Unfortunately the functional form of the marginal density $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ is unknown (otherwise MCMC would not be necessary) and we have to employ some method of density estimation to obtain estimates $\widehat{p(\boldsymbol{\beta}^{(t)} \,|\, \boldsymbol{y})}, t = 1, \ldots, T$, and $\widehat{p(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})}$. For Gaussian responses and multicategorical data with latent Gaussian responses the full conditionals $p(\boldsymbol{\beta} \,|\, \cdot)$, i.e. the conditional densities of $\boldsymbol{\beta}$ given the data and the remaining parameters, are available and an approach based on Rao-Blackwellization seems natural (Held 2004). The Rao-Blackwell estimate is more efficient than any other density estimate based on $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}$ and there is no smoothing parameter involved. Using the Rao-Blackwell theorem estimates for the marginal density $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ evaluated at an arbitrary parameter vector $\boldsymbol{\beta}$ are obtained by

$$\widehat{p(\boldsymbol{\beta} \,|\, \boldsymbol{y})} = \frac{1}{T}\sum_{v=1}^{T} p(\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}), \tag{9}$$

where $\boldsymbol{\alpha}_{-}^{(v)}$ comprises all model parameters excluding $\boldsymbol{\beta}$ and hence $p(\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y})$ denotes the full conditional density of $\boldsymbol{\beta}$. The density estimator (9) is a simple average of the full conditional evaluated at $\boldsymbol{\beta}$. Averaging is done by conditioning subsequently on all sampled parameters $\boldsymbol{\alpha}_{-}^{(v)}$, $v = 1, \ldots, T$.

As an alternative to the mean in (9) Held (2004) suggests to use the median, i.e.

$$\widehat{p(\boldsymbol{\beta} \,|\, \boldsymbol{y})} = med_{1 \leq v \leq T} \left\{ p(\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}) \right\}. \tag{10}$$

As an advantage, the estimated contour probabilities are invariant to monotonic transformations of $p(\boldsymbol{\beta} \,|\, \boldsymbol{y})$ in (7). For instance, one could replace $p(\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y})$ in (10) by the log density, i.e.

$$\log(\widehat{p(\boldsymbol{\beta} \,|\, \boldsymbol{y})}) = med_{1 \leq v \leq T} \left\{ \log(p(\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y})) \right\}. \tag{11}$$

Usually, this is computationally more favorable than using the density directly (see appendix B) and also more robust against extreme samples.

In order to estimate (8) the marginal densities $p(\boldsymbol{\beta}^{(t)} \,|\, \mathbf{y})$ evaluated at the sampled random numbers $\boldsymbol{\beta}^{(t)}$, $t = 1, \ldots, T$, have to be estimated. Using for instance (10) estimates are now easily obtained by

$$\widehat{p(\boldsymbol{\beta}^{(t)} \,|\, \boldsymbol{y})} = med_{1 \leq v \leq T} \left\{ p(\boldsymbol{\beta}^{(t)} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}) \right\}.$$

Summarizing, the contour probability (7) is estimated by replacing the marginal densities with (9), (10), or (11) if log densities are used. Using (10) we obtain

$$\widehat{P(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\left\{ med_{1 \leq v \leq T} \left\{ p(\boldsymbol{\beta}^{(t)} \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}) \right\} \leq med_{1 \leq v \leq T} \left\{ p(\boldsymbol{\beta}^* \,|\, \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}) \right\} \right\} \tag{12}$$

Pseudo contour probabilities based on credible intervals

As an alternative to the definition of contour probabilities via HPD regions, we could base the definition on simultaneous credible intervals for the parameter $\boldsymbol{\beta}^*$ of interest. Besag et al. (1995) propose to define a simultaneous credible interval as the hyperrectangular defined by

$$[\beta_k^{[T+1-t^*]}, \beta_k^{[t^*]}] \quad k = 1, \ldots, r+l, \tag{13}$$

where $\beta_k^{[t]}$, $t = 1, \ldots, T$ denotes the ordered samples of the parameter β_k. The index t^* is the smallest integer such that the hyperrectangular (13) contains at least 100α percent of the samples $\boldsymbol{\beta}^{(1)}, \ldots, \boldsymbol{\beta}^{(T)}$ if α is the desired level of the credible interval.

The *(pseudo) contour probability* $P(\boldsymbol{\beta}^* \,|\, \boldsymbol{y})$ for $\boldsymbol{\beta}^*$ can now be defined as 1 minus the smallest credible level, for which $\boldsymbol{\beta}^*$ is contained in the corresponding credible interval.

The advantage of pseudo contour probabilities is that they are much less computationally demanding. However, as is shown through simulations they are less reliable than contour probabilities.

2.3 Contour probabilities for P-Splines

In the context of P-splines, we are particularly interested in parameters $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ that lead to a constant, linear or in general a polynomial fit. Since P-splines are centered around zero a constant fit corresponds to $\boldsymbol{\beta}^* = \mathbf{0}$, i.e. the corresponding covariate is excluded from the predictor. In this section we determine conditions on the regression parameters that lead to a polynomial fit rather than a piecewise polynomial as is generally the case.

It can be shown that a spline $f(x)$ expressed in terms of (3) reduces to a polynomial of degree $s \leq l$ if the $(s+1)$-th differences of the regression parameters are zero, i.e.

$$\Delta^{s+1}\beta_k = 0, \quad k = s+2, \ldots, r+l, \tag{14}$$

or in matrix notation

$$\boldsymbol{D}_{s+1}\boldsymbol{\beta} = \mathbf{0},$$

where $\boldsymbol{D}_{s+1}$ is a difference matrix of order $s+1$. A proof can be found in appendix A.

In order to compute (pseudo) contour probabilities the full conditional of $\boldsymbol{D}_s\boldsymbol{\beta}$ must be computed. The full conditional of $\boldsymbol{\beta}$ is multivariate Gaussian

$$\boldsymbol{\beta} \,|\, \boldsymbol{\alpha}_-, \boldsymbol{y} \sim N(\boldsymbol{m}, \boldsymbol{P}^{-1}) \tag{15}$$

with

$$\boldsymbol{P} = \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} + \frac{1}{\tau_j}\boldsymbol{K}, \quad \boldsymbol{m} = \boldsymbol{P}^{-1}\frac{1}{\sigma^2}\boldsymbol{X}'(\boldsymbol{y} - \tilde{\boldsymbol{\eta}}).$$

Here, $\tilde{\boldsymbol{\eta}}$ is the part of the predictor associated with all remaining effects in the model. Thus $\boldsymbol{D}_s\boldsymbol{\beta} =: \tilde{\boldsymbol{\beta}}$ is also multivariate Gaussian

$$\tilde{\boldsymbol{\beta}} \,|\, \boldsymbol{\alpha}_-, \boldsymbol{y} \sim N(\tilde{\boldsymbol{m}}, \tilde{\boldsymbol{P}}^{-1}), \tag{16}$$

with mean $\tilde{\boldsymbol{m}} = \boldsymbol{D}_s\boldsymbol{m}$ and precision matrix $\tilde{\boldsymbol{P}} = \boldsymbol{D}_s\boldsymbol{P}^{-1}\boldsymbol{D}_s'$. Note that for the special case $s = 0$, i.e. $\boldsymbol{D}_s = \boldsymbol{I}$, we recover (15) as full conditional for $\boldsymbol{D}_s\boldsymbol{\beta}$.

Computational aspects of the estimator are discussed in appendix B.

2.4 Software

Bayesian P-splines and contour probabilities are implemented in the software package BayesX which is available free of charge at `www.stat.uni-muenchen.de/~bayesx/`. A detailed description of the usage is given in the accompanying manuals. As an example we take our application on undernutrition in Zambia and Tanzania. The following code estimates for Zambia nonlinear effects of the age of the child and the mother's body mass index:

```
delimiter=;

dataset d;
d.infile using c:\zambia.raw;

bayesreg b;
b.regress zscore = agc(psplinerw2,countourprob=4) +
bmi(psplinerw2,countourprob=4), approx family=gaussian
iterations=12000 burnin=2000 step=10 predict using d;
```

Option `contouprob=4` specifies that contour probabilities for difference orders zero to four are computed. By defining the global option `approx` the computation of contour probabilities is based on stochastic approximations for quantiles as described in Tierney (1983), see appendix B for details.

Note that BayesX allows an arbitrary combination of P-splines with other components for nonparametric modeling. Examples are random intercepts and slopes, spatial effects, varying coefficients, two dimensional surface smoothers etc. Details are available from the BayesX manuals.

3 Simulations

We realized an extensive simulation study in order to asses the performance of contour probabilities and to compare them to pseudo contour probabilities. In particular we were interested in the following questions: First, how successful is our tool in detecting a parametric (polynomial), or nonparametric effect, given different signal to noise ratios (SNR). Second, are there substantial differences regarding the type of definition for the contour probabilities, and third, what is the extent of improvement that is achieved by using contour probabilities rather than pseudo contour probabilities. Of course, given a credible level $1 - \alpha$, say, the HPD region is by definition the region with the smallest area, and therefore PCPs will always underestimate the p-value. However, it is unclear in advance how dramatic the loss of efficiency is in practice, specifically in the context of P-spline modeling.

We investigated the functions

$$y_i = 1 + k \cdot \sin(2\pi x_i) + \epsilon_i, \tag{17}$$

and

$$y_i = 1 + x_i + k \cdot \sin(2\pi x_i) + \epsilon_i, \tag{18}$$

with different values for k. For x we chose 100 equidistant design points in the interval $[0, 1]$ and generated data sets with 250 replications of each of the models (17) and (18)

with $\epsilon_i \sim N(0, 0.5)$. This corresponds to a signal to noise ratio (SNR) of $0.0, 1.0$ and 2.25 for $k = 0.0, 1.0, 1.5$ and model (17), and a SNR of approximately $0.17, 0.53$ and 1.47 for $k = 0.0, 1.0, 1.5$ and model (18), respectively. We used an IG(0.001,0.001) prior for the scale parameter σ^2 and the variance parameter τ^2. Alternatively, we assigned an IG(1.0,0.005) and a uniform prior. However, the results proved to be insensitive regarding the choice of prior.

We compare the results in terms of the 'p-values' obtained from contour probabilities based on the median and the mean of the log-density, and from pseudo contour probabilities. Figure 2 shows boxplots of p-values for three selected SNRs. Note that $\Delta^s\beta = 0$ corresponds to a constant fit (i.e. no effect) for both, $s = 0$ and $s = 1$. Figure 3 shows the outcome of an alternative model selection according to the DIC. The results of both models can be summarized as follows:

- *No effect (SNR=0.0)*
  As we could have expected, for a signal to noise ratio of 0.0 the contour probabilities are close to one for all difference orders considered, i.e. p-values give no evidence of any influence of the covariate at all. Pseudo contour probabilities do not suggest the existence of an influence of the covariate either, though they are considerably lower than the contour probabilities.

  It is striking that pseudo contour probabilities show a noticeable difference between difference orders $s = 0$ and $s = 1$, though both correspond to the probability for no effect of the covariate. Held (2004) reports severe underestimation for $s = 0$ and conjures that this comes from strong correlations between successive parameters. Since the correlation decreases when considering first differences of the parameters instead of the parameters directly the problem becomes less distinctive. This may explain the big differences between $s = 0$ and $s = 1$.

  The DIC exhibits features (similar to AIC) that have been observed in many cases. In approximately 25 percent of the cases a linear or even more complex effect is found where there is essentially none.

- *Very low to low signal to noise ratio (SNR=0.17, 0.53, 1.0)*
  For the very low and low signal to noise ratios (0.17, 0.53 in model (18), 1.0 in model (17)) the p-values clearly decrease for all difference orders smaller than 4, i.e. the posterior probabilities for a (at least) cubic effect increase, as the SNR increases from 0.17 to 1.0. For the model with $SNR = 1.0$ contour probabilities actually speak against the hypothesis of the covariate having no effect. However, neither contour probabilities nor pseudo contour probabilities give clear cut results and hence further investigation is advisable. An exception are p-values from pseudo contour probabilities based on 0-th order differences. Here, pseudo contour probabilities exhibit mainly very low p-values. However, this may be due to the underestimation mentioned by Held (2004).

  The DIC clearly is in favor of more complex models. Cubic modeling is selected in approximately 70 percent of the cases and nonparametric modeling in 30 percent.

- *Medium signal to noise ratio (SNR=1.47, 2.25)*
  For medium signal to noise ratios (1.47 in model (18), 2.25 in model (17)) the contour probabilities for parametric fits with polynomials of degree smaller than three (i.e. difference order smaller than 4) are very small, suggesting that a more

flexible modeling is needed. However, the need of a polynomial of degree higher than 3 is rather unlikely a posteriori. This is in perfect agreement with the data, since a sine curve can be approximated by a polynomial of degree 3 without major deviations. Pseudo contour probabilities, on the other hand, perform very poorly for difference orders higher than 1.

Similar to low signal to noise ratios the DIC is able to detect the nonlinear effects. In roughly 50 percent of cases a cubic fit is preferred and for the rest of the cases even nonparametric modeling.

- *Contour probabilities versus pseudo contour probabilities*
  It turns out that p-values based on pseudo contour probabilities are apparently smaller than that obtained from the contour probabilities for very low signal to noise ratios. This is in accordance with findings of Held (2004) who reported severe underestimation of p-values especially in the case of difference order $s = 0$, but also - to a smaller degree - when considering first differences.

  In contrast, pseudo contour probabilities behave rather conservative regarding higher differences compared to contour probabilities. For a SNR of 2.25 p-values in favor of a parametrization by polynomials of a degree higher than quadratic are still reasonably close to one.

- *Contour probabilities versus the DIC*

  The main difference is that contour probabilities are more conservative than the DIC. The DIC detects even nonlinear effects with low signal to noise ratio whereas for contour probabilities clear decisions are possible only for moderate (or larger) SNR's. However, there is nothing like a free lunch. In a considerable number of cases the DIC detects an effect where there is none.

- *Contour probabilities based on the median/mean of log-density*
  Estimated p-values may differ slightly regarding on which definition they are based. In our simulation study we compared p-values based on the median or on the mean of the log-density, respectively. We found p-values based on the mean of the log-density to be noticeably higher than the ones based on the median.

Based on the findings we draw the following conclusions:

- Pseudo contour probabilities underestimate the p-values regarding the decision whether a covariate has an effect on the response or not, whereas for the decision of modeling an effect linearly (or by a polynomial of higher degree) they seem to behave too conservative and are therefore not recommended.

- Contour probabilities seem to give highly reasonable results for medium or higher SNRs . In situations where the SNR is small, it is difficult to base model selection solely on contour probabilities in cases when the obtained p-values lie in a medium range (i.e. between 0.1 and 0.4, approximately).

- The DIC detects nonlinear effects even for small SNR's. On the other hand in a considerable number of cases too complex models are selected.

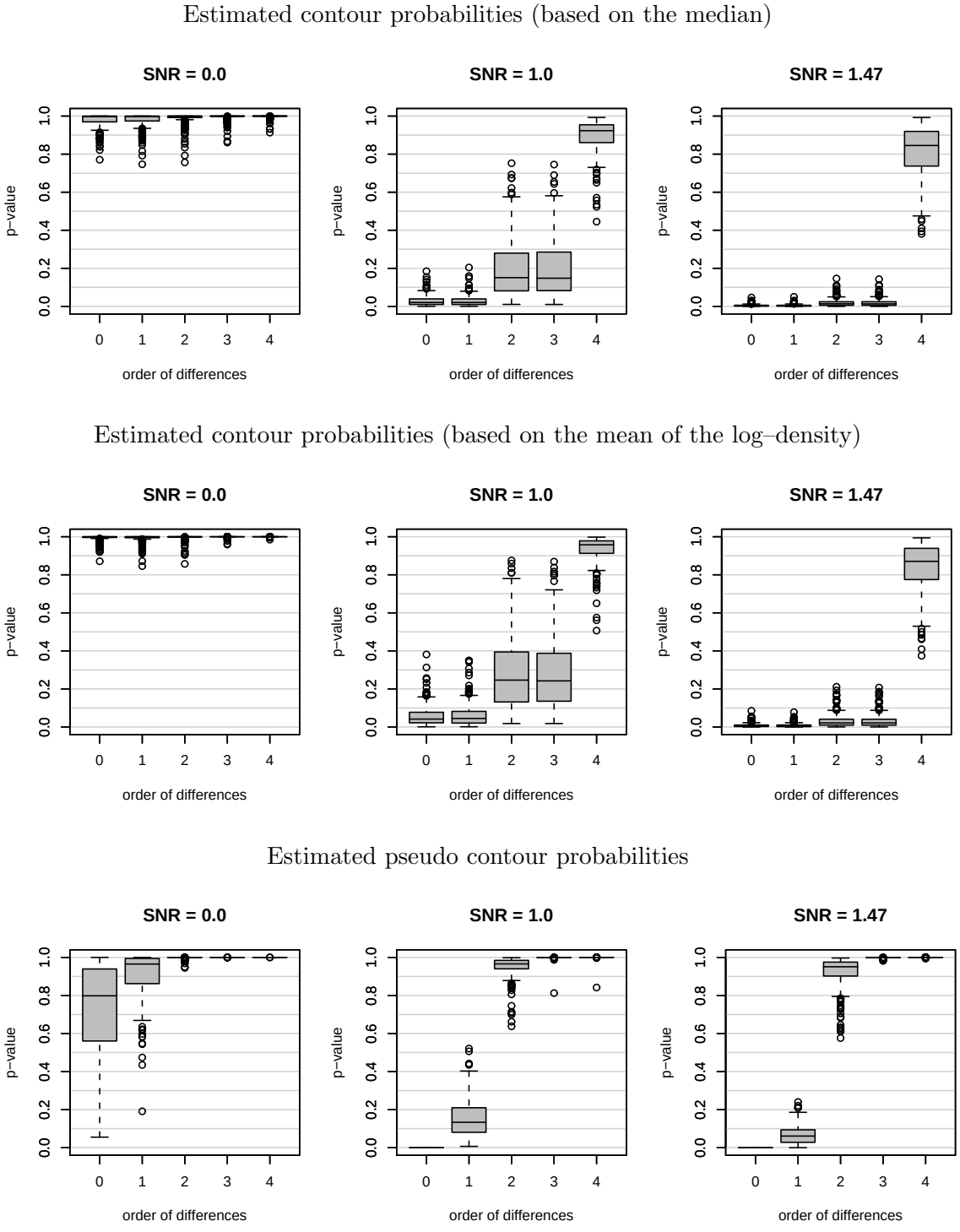

Figure 2: Boxplots of p-values obtained from contour probabilities based on the median (top), contour probabilities based on the mean of the log-density (middle), and pseudo contour probabilities (bottom) for different SNRs and difference orders. Difference order $s = 0$ *and* $s = 1$ *corresponds to no effect,* $s = 2$ $(3, 4)$ *corresponds to a linear (quadratic, cubic) effect.*

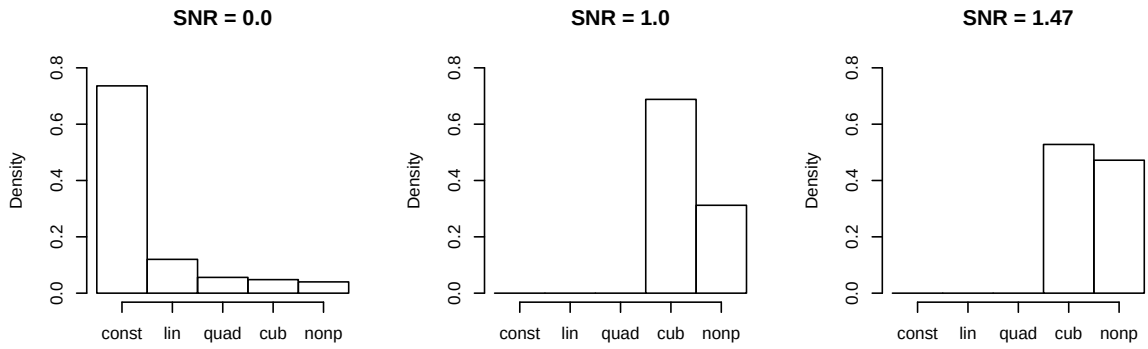

Figure 3: Percentage of cases in which the DIC selected the constant, linear, quadratic, cubic or nonparametric model for $SNR = 0.0, 1.0, 1.47$.

4 Applications

In this section we illustrate the application of the previously described model selection tools by applications that result from cooperations. Our first example investigates undernutrition of children in Zambia and Tanzania and is based on data already analyzed by Kandala et al. (2001). The second analysis investigates the impact of covariates on the degree of defoliation of trees measured in three categories. This is an application of contour probabilities to non-Gaussian responses. Based on the results of the previous section we restrict the discussion to contour probabilities and do not report corresponding pseudo contour probabilities. We propose the following modeling strategy in applications:

- We start with a complex model where effects of continuous covariates are estimated nonparametrically by P-splines. The model building process at this stage is primarily guided by knowledge from previous studies and the scientific background.
- We compute contour probabilities in order to identify a small number of reasonable and potentially less complex models.
- We additionally compute the DIC as a global goodness of fit measure for the models under consideration.
- Sensitivity analysis regarding the hyperpriors, outliers and influential points is performed for the most promising models. In the remainder details to this step are omitted because this is not the topic of this paper.
- Based on the results we extract the undoubted core findings with least degree of uncertainty. Usually for some aspects of the analysis considerable uncertainty remains because of unstable results.

4.1 Undernutrition in Zambia and Tanzania

The Demographic and Health Surveys (DHS) of Tanzania and Zambia, both conducted in 1992, draw a representative sample of women in reproductive age in the two countries. Thereafter they administer a questionnaire and an anthropometric assessment of themselves and their children that where born within the previous five years. The data contains 6299 cases in Zambia and 8138 cases in Tanzania. Kandala et al. (2001) use this data to

explore determinants of undernutrition measured through stunting, which is insufficient height for age, indicating chronic undernutrition. Stunting for a child i is determined by a Z-score

$$Z_i = \frac{AI_i - MAI}{\sigma_R},$$

where AI refers to the childs height at a certain age, MAI refers to the median of a reference population, and σ_R denotes the standard deviation of the reference population. Kandala et al. (2001) estimate separate additive models for each country with a predictor

$$\eta = \gamma_0 + f_1(bmi) + f_2(agc) + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x},$$

where the mothers body mass index bmi and the age of the child agc are modeled nonparametrically with Bayesian P-splines. The expression $f_{spat}(d)$ denotes a spatial effect associated with the district d the child lives in, and is modeled as the sum of i.i.d and spatially correlated random effects for Zambia. For Tanzania the i.i.d random effects are excluded from the model. The fixed effects $\boldsymbol{\gamma}$ include categorical variables concerning the education and employment situation of the mother, the gender of the child and the characteristic of the area (urban or rural), where the child resides. For more details on the analysis we refer the reader to Kandala et al. (2001).

Here, our aim is to investigate whether the nonparametric modeling of bmi and agc is necessary by employing contour probabilities. As a starting point, we use the model developed by Kandala et al. (2001) and model the effects of both continuous covariates, bmi and agc, nonparametrically by P-splines. The resulting fits together with 80 and 95 percent pointwise credible intervals are given in the first rows of figures 4 and 5. The effects of the body mass index are both slightly U-shaped and close to linearity. The U-shape is more pronounced for Tanzania. Overall this is in line with the literature on undernutrition. Mothers who exhibit a very low BMI, indicating their poor nourishment, are likely to have poorly nourished children. At the same time, parents with a very high BMI might also have poorly nourished children as the obesity associated with their high BMI indicates poor quality of nutrition and might therefore indicate poor quality of nutrition for their children. For the latter presumption support from the data is weak as both effects are close to linearity.

The age effect indicates a continuous worsening of the nutritional status up until about 20 months of age. This deterioration sets in right after birth and continues, more or less linearly, until 20 months. After 20 months, stunting stabilizes at a low level. Again the effect is in line with the literature. The main difference between the two countries is an additional local maximum in the interval $[25, 30]$ for Tanzania. This is picking up the effect of a change in the data set that makes up the reference standard. Until 24 months, the currently used international reference standard is based on white children in the US of high socioeconomic status, while after 24 months, it is based on a representative sample of all US children. Since the latter sample exhibits worse nutritional status, comparing the Tanzanian children to that sample may lead to a sudden improvement of their nutritional status at 24 months.

For both effects it is reasonable to think of a more parsimonious modeling. The effect of bmi seems close to linearity. The effect of agc may possibly be modeled by a quadratic or cubic polynomial. To shed more light on this issue we take a look at the contour probabilities displayed in table 1. They suggest a quadratic, cubic or (for Tanzania) even nonparametric effect for the age of the child. The p-values based on contour probabilities

in favor of 'no effect' of *bmi* are in a medium region and allow no clear decision. In either case, a linear effect for *bmi* seems to be sufficient. Based on these observations we investigate a number of additional model specifications summarized as follows:

$$
\begin{array}{lllll}
\boldsymbol{\eta}_1 = \gamma_0 + f_1(bmi) & + f_2(agc) & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_2 = \gamma_0 + & + f_2(agc) & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_3 = \gamma_0 + & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_4 = \gamma_0 + & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_5 = \gamma_0 + \alpha_1 bmi & + f_2(agc) & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_6 = \gamma_0 + \alpha_1 bmi & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x} \\
\boldsymbol{\eta}_7 = \gamma_0 + \alpha_1 bmi & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & & + f_{spat}(d) + \boldsymbol{\gamma}'\boldsymbol{x}
\end{array}
$$

In table 2 values for the DIC of the models are summarized. Models are ordered according to the DIC.

For Zambia the best fit in terms of the DIC is achieved by model 5, featuring a nonparametric fit for *agc* and a linear fit for *bmi*. However, models 7 (*bmi* linear, *agc* cubic) and 1 (both nonparametrically) perform roughly equally well. Model 7 is the model which is best in line with the observed contour probabilities. The second row of figure 4 compares the three different fits for *bmi* and *agc* obtained by models 1, 5 and 7. Since the two linear fits for *bmi* and the two nonparametric fits for *agc* are almost identical the respective second fit is omitted. We conclude that a linear effect for *bmi* and a cubic effect for *agc* is sufficient to describe the variability of the data.

Slightly different results are obtained for Tanzania. The best fit in terms of the DIC is achieved by model 1 with nonparametric modeling of *bmi* and *agc*. Model 5 which is best in line with the observed contour probabilities shows the second best fit. The DIC for this model is, however, considerably larger than for model 1. The bottom line of figure 5 compares the different fits. Obviously, the cubic fit totally misses the local maximum exhibited by the nonparametric estimate of the effect of *agc*. Regarding the effect of *bmi* contour probabilities and the DIC exhibit the largest differences. While contour probabilities are in favor of (at maximum) a linear fit the DIC clearly supports the U-shaped effect obtained by the nonparametric modeling. However, the U-shaped effect is mostly driven by a few observations with very large body mass index. A sensitivity analysis reveals that exclusion of the 6 observations with $bmi > 39$ yields that both models perform equally well in terms of the DIC criterion. We conclude that the effect of *agc* should be modeled nonparametrically because of the local maximum of the effect in the interval $[25, 30]$. The effect of the *bmi* is clearly linear for mothers with $bmi < 32$. For $bmi > 32$ there is considerable uncertainty about the effect. There is some support for the U-shaped effect predicted in the literature but more information is needed for a clear decision.

4.2 Forest health study

In this longitudinal study on the health status of trees, we demonstrate the usefulness of our approach for non Gaussian data. We analyze the influence of calendar time t, age of trees A (in years), canopy density CP (in percent) and location L of the stand on the defoliation degree of beeches. Data have been collected in yearly forest damage inventories carried out in the forest district of Rothenbuch in northern Bavaria from 1983 to 2001. The state of a tree is assessed by the degree of defoliation measured in three ordered categories, with $y_{it} = 1$ for "bad" state of tree i in year t, $y_{it} = 2$ for "medium"

Table 1: Undernutrition in Zambia and Tanzania: Contour probabilities for the effects of bmi and agc. Displayed are the results for model 1.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
Zambia					
bmi (based on median)	0.29	0.38	1.0	1.0	1.0
bmi (based on mean of log-density)	0.30	0.42	1.0	1.0	1.0
agc (based on median)	0.0	0.0	0.0	0.09	0.84
agc (based on mean of log-density)	0.0	0.0	0.0	0.12	0.87
Tanzania					
bmi (based on median)	0.33	0.14	0.79	0.93	0.93
bmi (based on mean of log-density)	0.42	0.25	0.86	0.97	0.97
agc (based on median)	0.0	0.0	0.0	0.0	0.09
agc (based on mean of log-density)	0.0	0.0	0.0	0.01	0.20

Table 2: Undernutrition in Zambia and Tanzania: DIC for models 1-5. The models are ordered according their DIC.

Zambia	DIC	Tanzania	DIC
M5	12733.9	M1	15555.8
M7	12736.9	M5	15563.1
M1	12737.9	M7	15592.8
M2	12756.2	M2	15606.5
M4	12761.7	M4	15638.4
M6	12779.5	M6	15672.6
M3	12804.2	M3	15723.6

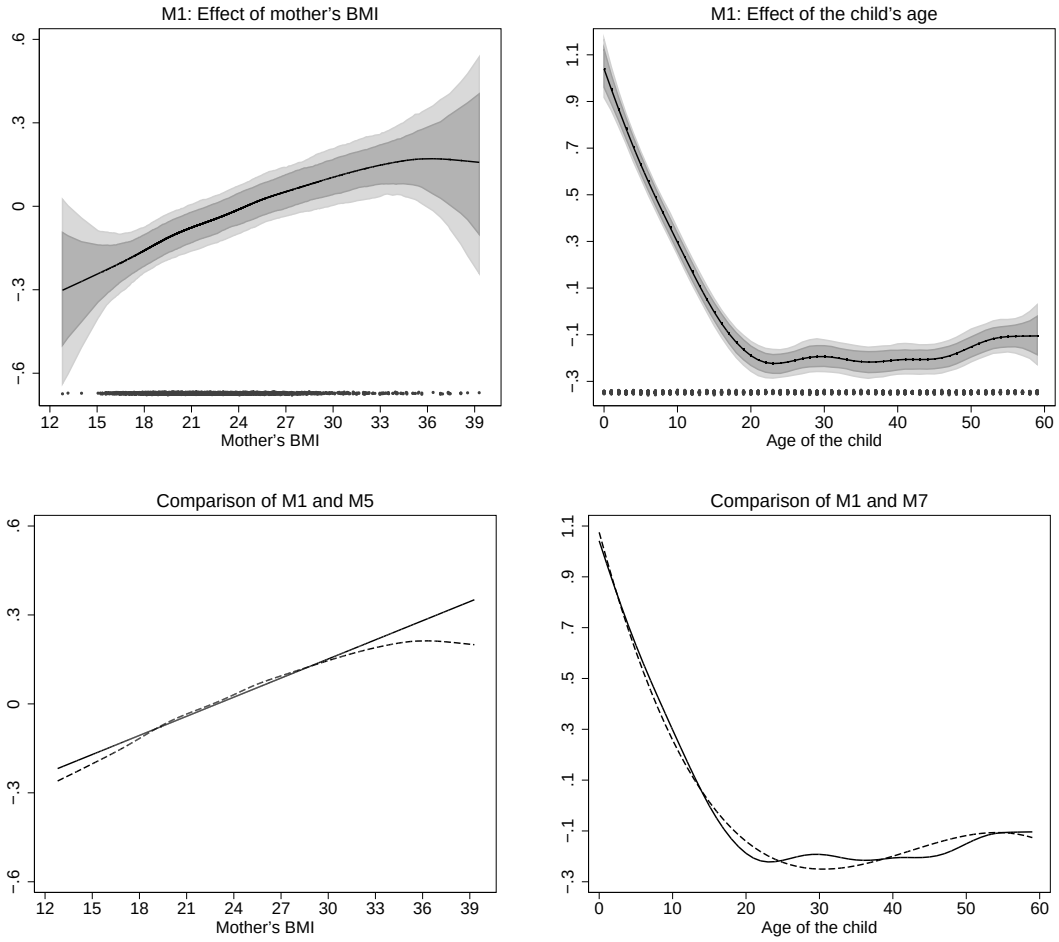

Figure 4: Zambia: Effects of bmi (left panel) and agc (right panel) for the nonparametric model 1. The second row compares the nonparametric fits with competing parametric fits obtained from models 5 and 7. The dots at the bottom of the graphs indicate the distribution of the data.

and $y_{it} = 3$ for "good". A detailed description of the data can be found in Göttlein and Pruscha (1996).

We use a three-categorical ordered probit model with predictor

$$\eta_{it} = \gamma_0 + f_1(t) + f_2(A_{it}) + f_3(CP_{it}) + f_{spat}(L_i). \tag{19}$$

The functions f_1 to f_3 are modeled nonparametrically by cubic P-splines. The expression f_{spat} denotes the spatial effect of the location L_i of the tree, modeled by a spatially correlated random effect. The data have already been analyzed in Fahrmeir and Lang (2001) (for the years 1983-1997 only), where nonlinear functions have been modeled solely by random walk priors.

Again, we are interested in the appropriateness of nonparametric modeling of the function f_1 to f_3 compared to parametric alternatives. Table 3 summarizes the obtained p-values and figure 6 displays the estimated effects. Looking at contour probabilities, nonparametric modeling is clearly suggested for the effects of time and age, whereas modeling of

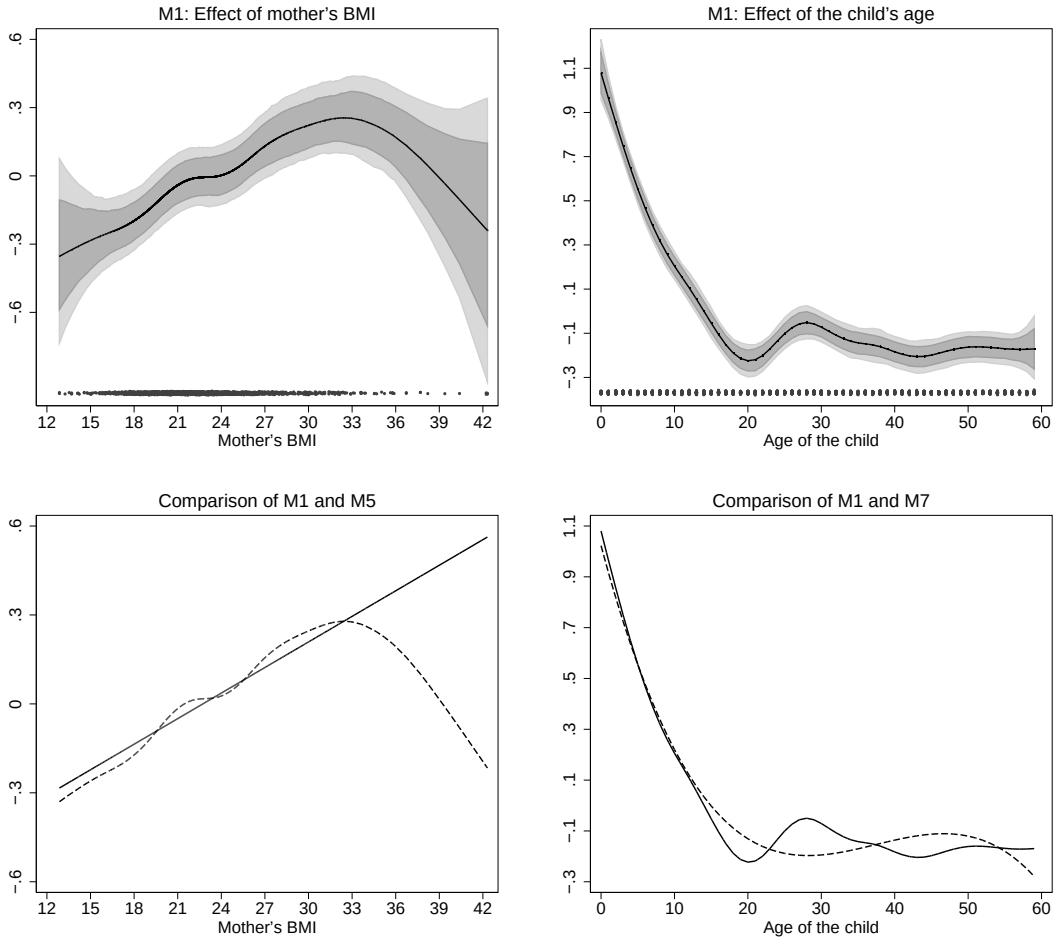

Figure 5: Tanzania: Effects of bmi (left panel) and agc (right panel) for the nonparametric model M1. The second row compares the nonparametric fits with competing parametric fits obtained from models 5 and 7. The dots at the bottom of the graphs indicate the distribution of the data.

canopy density is suggested as at most linearly. Based on the relatively clear results we additionally estimated only one competing model with linear effect for CP. The DIC for this model is 1350.7 and is slightly less than for the nonparametric model. In this application the observed contour probabilities and the DIC are fully consistent in the sense that they lead to the same conclusion: The effects of calendar time and the age of the trees can not be modeled by polynomials whereas the effect of canopy density may be modeled linearly.

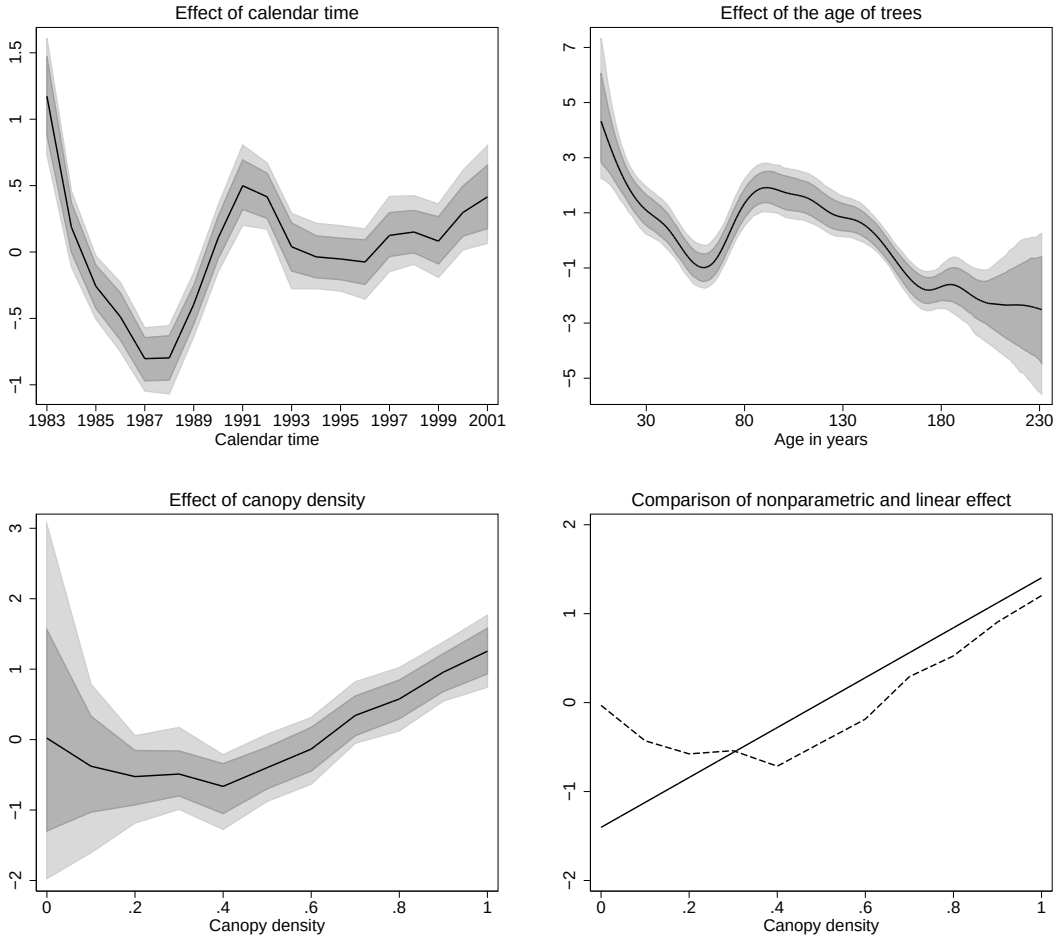

Figure 6: Nonparametric effects of t, A and CP together with 80% and 95% pointwise credible intervals. Additionally the nonparametric and the linear fit for canopy density are compared.

Table 3: Contour probabilities for the effects of `t`, `A` *and* `CP`. *Displayed are the results for model 1.*

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`t` (based on median)	0.0	0.0	0.0	0.0	0.01
`t` (based on mean of log-density)	0.01	0.0	0.0	0.01	0.02
`A` (based on median)	0.0	0.0	0.0	0.0	0.0
`A` (based on mean of log-density)	0.0	0.0	0.0	0.0	0.0
`CP` (based on median)	0.04	0.09	0.51	0.87	0.98
`CP` (based on mean of log-density)	0.06	0.12	0.44	0.85	0.99

5 Conclusion

The paper developed contour probabilities and pseudo contour probabilities for Bayesian P-splines in order to decide whether nonparametric modeling of continuous covariates is necessary or if parametric modeling by polynomials of small degree is sufficient. Currently the methodology is available for Gaussian responses and for multicategorical logit or probit models with latent Gaussian responses. Estimating Bayesian p-values for general distributions from an exponential family is computationally much more expensive since the marginal distributions are no longer available by Rao-Blackwellization. A possible approach could be based on a paper by Chib and Jeliazkov (2001) who present methodology for computing posterior ordinates of densities based on MCMC. However, the approach is quite computer intensive and currently not available for routine use in applications.

The simulation study shows that contour probabilities are a valuable tool for model choice at least for moderate or larger signal to noise ratios. They also proved to be superior compared to pseudo contour probabilities. However, we do not recommend to base model selection solely on contour probabilities. We rather propose to use a combination of simultaneous probability statements, goodness of fit measures and sensitivity analysis to perform this complex task. Examples of the proposed model selection approach are given by the two applications on undernutrition in developing countries and the health status of trees.

Acknowledgement:
This research has been financially supported by grants from the German Science Foundation (DFG), Sonderforschungsbereich 386 "Statistical Analysis of Discrete Structures". We are grateful to two referees and the editors for many valuable suggestions that helped to improve the first version of the paper.

A Proof of equation (14)

For a proof of (14) we exploit the fact that the B-spline basis functions in (3) for representing the spline can be computed as differences of truncated power functions (Eilers and Marx (2004)), i.e.

$$B_k(x) = -1^{l+1}\Delta^{l+1}t(x,k)/(h^l l!), \quad k = 1,\ldots,r+l \tag{20}$$

where h is the distance between two neighboring knots and $t(x,k) := (x-(\kappa_0+kh))_+^l$ is the truncated power function that corresponds to the knot $\kappa_k = \kappa_0 + kh$.

Assume first that $s = 0$, which corresponds to a constant fit. Then we get

$$\frac{(h^l l!)}{-1^{l+1}} f(x) = \frac{(h^l l!)}{-1^{l+1}} \sum_{k=1}^{r+l} B_k(x)\beta_k = \sum_{k=1}^{r+l} \Delta\Delta^l t(x,k)\beta_k = \sum_{k=1}^{r+l} \Delta^l t(x,k)\beta_k - \sum_{k=1}^{r+l} \Delta^l t(x,k-1)\beta_k$$

Rearranging the two sums by combining the respective k-th summand of the first sum and the $(k+1)$-th summand of the second sum yields

$$\frac{(h^l l!)}{-1^{l+1}} f(x) = -\sum_{k=1}^{r+l-1} \Delta^l t(x,k)\Delta\beta_{k+1} + \Delta^l t(x,r+l)\beta_{r+l} - \Delta^l t(x,0)\beta_1. \tag{21}$$

Provided that $\Delta\beta_k = 0$, the summands in the first term are all zero. The second term in (21) is zero within the range $[x_{min}, x_{max}]$ of x because the polynomial part of $t(x, r+l)$ starts at x_{max}. In the third term the truncated power function $t(x, 0)$ is a polynomial of degree l within the range of x. Since the l-th difference of a polynomial of degree l is a constant (compare, e.g. Schlittgen and Streitberg, p. 39f), the spline $f(x)$ reduces to a constant as claimed in (14).

For an arbitrary degree $s \leq l$ the proof is based on analogous arguments. Using again relationship (20) we get

$$\begin{aligned} \frac{(h^l l!)}{-1^{l+1}} f(x) &= \sum_{k=1}^{r+l} \Delta^{s+1}\Delta^{l-s} t(x,k)\beta_k \\ &= a_1 \sum_{k=1}^{r+l} \Delta^{l-s} t(x,k)\beta_k + \cdots + a_{s+2} \sum_{k=1}^{r+l} \Delta^{l-s} t(x, k-(s+1))\beta_k \end{aligned} \tag{22}$$

with constants $a_1, \ldots, a_{s+2}$ given by

$$a_j = (-1)^{s+j} \binom{s+1}{j-1}, \quad j = 1, \ldots, s+2.$$

Combining the k-th summand of the first sum, $(k+1)$-th summand of the second sum, to the $(k+s+1)$-th summand of the $(s+2)$-th sum, $k = 1, \ldots, r+l-s-1$, we obtain

$$\frac{(h^l l!)}{-1^{l+1}} f(x) = (-1)^{s+1} \sum_{k=1}^{r+l-s-1} \Delta^{l-s} t(x,k) \Delta^{s+1} \beta_{k+s+1} + R_1 + R_2 \tag{23}$$

with

$$\begin{aligned} R_1 &= a_1 \left(\Delta^{l-s} t(x, r+l-s)\beta_{r+l-s} + \cdots + \Delta^{l-s} t(x, r+l)\beta_{r+l} \right) \\ &\quad + \cdots + a_{s+1} \Delta^{l-s} t(x, r+l)\beta_{r+l-s} \end{aligned}$$

and

$$R_2 = a_2 \Delta^{l-s} t(x,0)\beta_1 + \cdots + a_{s+2} \left(\Delta^{l-s} t(x,-s)\beta_1 + \cdots + \Delta^{l-s} t(x,0)\beta_{s+1} \right).$$

Provided that $\Delta^{s+1}\beta_k = 0$, the sum in (23) is zero. The expression R_1 is zero within the range $[x_{min}, x_{max}]$ of x. Since the $(l-s)$-th difference of a polynomial of degree l is a polynomial of degree s (compare Schlittgen and Streitberg, p. 39f) all differences of the truncated power functions appearing in R_2 are polynomials of degree $l-s$ within the range of x. Hence R_2, and therefore the spline $f(x)$, is a polynomial of degree s.

B Computational aspects

This section is concerned with computational aspects of the estimator (12). We will distinguish the two cases $s = 0$ and $s > 0$.

In the case $s = 0$ we have to evaluate

$$log(p(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y})) = \frac{1}{2} \log(|\boldsymbol{P}^{(v)}|) - \frac{1}{2} (\boldsymbol{\beta}^{(t)} - \boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} (\boldsymbol{\beta}^{(t)} - \boldsymbol{m}^{(v)}) \tag{24}$$

for $t, v = 1, \ldots, T$ in order to estimate (12). Here, $\boldsymbol{P}^{(v)}$ is the posterior precision matrix evaluated at the v-th sample of τ^2 and σ^2 and $\boldsymbol{m}^{(v)}$ is the posterior mean evaluated at the v-th sample of $\boldsymbol{P}$, σ^2 and $\tilde{\boldsymbol{\eta}}$. It is useful to decompose the quadratic form in (24) by

$$(\boldsymbol{\beta}^{(t)} - \boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} (\boldsymbol{\beta}^{(t)} - \boldsymbol{m}^{(v)}) =$$
$$\frac{1}{(\sigma^2)^{(v)}} (\boldsymbol{\beta}^{(t)})' \boldsymbol{X}' \boldsymbol{X} \boldsymbol{\beta}^{(t)} + \frac{1}{(\tau^2)^{(v)}} (\boldsymbol{\beta}^{(t)})' \boldsymbol{K} \boldsymbol{\beta}^{(t)} + (\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{m}^{(v)} - 2(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{\beta}^{(t)},$$

This shows that (12) can be evaluated by computing and storing the samples $\log(|\boldsymbol{P}^{(t)}|)$, $(\boldsymbol{\beta}^{(t)})' \boldsymbol{X}' \boldsymbol{X} \boldsymbol{\beta}^{(t)}$, $(\boldsymbol{\beta}^{(t)})' \boldsymbol{K} \boldsymbol{\beta}^{(t)}$, $(\boldsymbol{m}^{(t)})' \boldsymbol{P}^{(t)} \boldsymbol{m}^{(t)}$ and $(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{\beta}^{(t)}$. Except $(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{\beta}^{(t)}$ these quantities are obtained as a by product of the MCMC simulation run. For $t \leq v$, $t, v = 1, \ldots, T$ it is also possible to store $(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{\beta}^{(t)}$. For $t > v$ the quantity $(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)} \boldsymbol{\beta}^{(t)}$ must be computed *after* the MCMC simulation. This is facilitated by storing $(\boldsymbol{m}^{(v)})' \boldsymbol{P}^{(v)}$ after every iteration of the MCMC sampler.
The case $s > 0$ is computationally more demanding. In this case the log densities

$$log(p(\tilde{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\alpha}_-^{(v)}, \boldsymbol{y})) = \frac{1}{2} \log(|\tilde{\boldsymbol{P}}^{(v)}|) - \frac{1}{2} (\tilde{\boldsymbol{\beta}}^{(t)} - \tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} (\tilde{\boldsymbol{\beta}}^{(t)} - \tilde{\boldsymbol{m}}^{(v)})$$

must be computed. Evaluation of the quadratic form yields

$$(\tilde{\boldsymbol{\beta}}^{(t)} - \tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} (\tilde{\boldsymbol{\beta}}^{(t)} - \tilde{\boldsymbol{m}}^{(v)}) = (\tilde{\boldsymbol{\beta}}^{(t)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)} + (\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{m}}^{(v)} - 2(\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}.$$

Hence the quantities $\log(|\tilde{\boldsymbol{P}}^{(v)}|)$ and $(\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{m}}^{(v)}$ can be computed as a by product of the MCMC sampler and stored in every iteration. However, the quantities $(\tilde{\boldsymbol{\beta}}^{(t)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}$ and $(\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}$ can only be stored for $t \leq v$. For $t > v$ both quantities must be computed *after* the MCMC run.
Now we can compute $med_v \left\{ \log(p(\tilde{\boldsymbol{\beta}}^{(t)} | \boldsymbol{\alpha}_-^{(v)}, \boldsymbol{y})) \right\}$ for all t in two ways which differ in the order of evaluations:

Algorithm 1:

For $t = 1, \ldots, T$:

1. For $v = 1, \ldots, T$:
   (a) If $t > v$:
   Compute $\tilde{\boldsymbol{P}}^{(v)}$ and with it the quantities $(\tilde{\boldsymbol{\beta}}^{(t)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}$ and $(\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}$.
   (b) Compute $\log(p(\tilde{\boldsymbol{\beta}}^{(t)} | \boldsymbol{\alpha}_-^{(v)}, \boldsymbol{y}))$.
2. Compute $med_v \left\{ \log(p(\tilde{\boldsymbol{\beta}}^{(t)} | \boldsymbol{\alpha}_-^{(v)}, \boldsymbol{y})) \right\}$.

This algorithm is very time consuming, because $\tilde{\boldsymbol{P}}^{(v)}$ has to be computed $T(T-1)/2$ times.

The second algorithm is:

Algorithm 2:

1. For $v = 1, \ldots, T$:
   (a) Compute $\tilde{\boldsymbol{P}}^{(v)}$.
   (b) For $t = 1, \ldots, T$:
   If $t \leq v$: Compute $\log(p(\tilde{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}))$ based on the stored quantities.
   If $t > v$:
   Compute first $(\boldsymbol{\beta}^{(t)})' \tilde{\boldsymbol{P}}^{(v)} \boldsymbol{\beta}^{(t)}$ and $(\tilde{\boldsymbol{m}}^{(v)})' \tilde{\boldsymbol{P}}^{(v)} \tilde{\boldsymbol{\beta}}^{(t)}$ and then $\log(p(\tilde{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}))$.
2. For $v = 1, \ldots, T$: Compute $med_v \left\{ \log(p(\tilde{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y})) \right\}$.

The drawback of this algorithm is that it takes an enormous amount of memory space, because we have to create a $T \times T$ matrix to store all values $\log(p(\tilde{\boldsymbol{\beta}}^{(t)} \mid \boldsymbol{\alpha}_{-}^{(v)}, \boldsymbol{y}))$, $t, v = 1, \ldots, T$, before computing the median. A remedy is to take only every k-th sample to estimate $\widehat{p(\tilde{\boldsymbol{\beta}}^{(t)} | \boldsymbol{y})}$, but the memory requirement is still quite high.
As an alternative to the direct computation of the medians, we propose to use the method of stochastic approximation as described in Tierney (1983). The advantage is, that the quantiles can be estimated by a very space-efficient recursive procedure. Throughout this work we use Algorithm 2 together with stochastic approximation of quantiles to avoid extensive use of memory space.

References

Albert, J. and Chib, S., 1993: Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88, 669-679.

Besag, J.E., Green, J.P., Higdon, D.M. and Mengersen, K.L., 1995: Bayesian Computation and Stochastic Systems (with discussion). *Statistical Science*, 10, 3-66.

Box, G. E. P. and Tiao, G. C., 1973: *Bayesian Inference in Statistical Analysis.* Reading, MA: Addision-Wiley. Reprint by Wiley in 1992 in the Wiley Classics Library Edition.

Brezger, A. and Kneib, T. and Lang, S., 2005a: BayesX: Analyzing Bayesian Structured Additive Regression Models. *Journal of Statistical Software*, 14, 1–22.

Brezger, A. and Kneib, T. and Lang, S., 2005b: BayesX Manuals, Available at http://www.stat.uni-muenchen.de/~bayesx

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

Chib, S., and Jeliazkov, I., 2001: Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96, 270-281.

DiCiccio, T.J., Kass, A.E., Raftery, A.E. and Wasserman, L., 1997: Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, 92, 903-915.

Eilers, P.H.C. and Marx, B.D. (1996), Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder). *Statistical Science*, 11, 89–121.

Eilers, P.H.C. and Marx, B.D. (2004), Splines, Knots and Penalties. Technical report. Available at `http://www.stat.lsu.edu/bmarx/`.

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

Göttlein, A. and Pruscha, H. (1996), Der Einfluß von Bestandskenngrößen, Topographie, Standord und Witterung auf die Entwicklung des Kronenzustandes im Bereich des Forstamtes Rothenbuch, *Forstwissenschaftliches Centralblatt*, 114, 146–162.

Hastie, T. J. and Tibshirani, R. J. and Friedman, J., 2003: *The Elements of Statistical Learning*. Springer, New York.

Held, L., 2004: Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, 13, 20-35.

Holmes, C., and Held, L., 2006: Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, 1, 145-168.

Kandala, N. B., Lang, S., Klasen, S. and Fahrmeir, L. (2001), Semiparametric Analysis of the Socio–Demographic and Spatial Determinants of Undernutrition in Two African Countries, *Research in Official Statistics*, 1, 81–100.

Kass, R.E. and Raftery, A.E., 1995: Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

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

Marx, B.D. and Eilers, P.H.C. (1998), Direct generalized additive modeling with penalized likelihood, *Computational Statistics and Data Analysis*, 28, 193-209.

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

Schlittgen R., Streitberg, B.H. (1995), *Zeitreihenanalyse*, Oldenbourg, 1995.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A., 2002: *Bayesian measures of model complexity and fit*, *Journal of the Royal Statistical Society B*, 65, 583–639.

Stasinopoulos, M. and Rigby, B. and Akanziliotou, P., 2005: Instructions on how to use the GAMLSS package in R. Availabe at http://studweb.north.londonmet.ac.uk/ stasinom/gamlss.html.

Tierney, L., 1983: A space-efficient recursive Procedure for estimating a Quantile of an unknown Distribution. *SIAM J. Sci. Stat. Comput.*, 4, 706-711.

Wood, S. N., 2006: *Generalized Additive Models: An Introduction with R.* Chapman & Hall / CRC, Boca Raton, FL.

Wood, S. N., 2006b: *R - Manual: The mgcv package, version 1.3 - 22.*

Wood, S. N., 2006c: On Confidence Intervals for Generalized Additive Models Based On Penalized Regression Splines. *Australian and New Zealand Journal of Statistics*, 48,1-20.

University of Innsbruck – Working Papers in Economics and Statistics
Recent papers

2007-08 **Andreas Brezger and Stefan Lang:** Simultaneous probability statements for Bayesian P-splines

2007-07 **Georg Meran and Reimund Schwarze:** Can minimum prices assure the quality of professional services?

2007-06 **Michal Brzoza-Brzezina and Jesus Crespo Cuaresma:** Mr. Wicksell and the global economy: What drives real interest rates?.

2007-05 **Paul Raschky:** Estimating the effects of risk transfer mechanisms against floods in Europe and U.S.A.: A dynamic panel approach.

2007-04 **Paul Raschky and Hannelore Weck-Hannemann:** Charity hazard - A real hazard to natural disaster insurance.

2007-03 **Paul Raschky:** The overprotective parent - Bureaucratic agencies and natural hazard management.

2007-02 **Martin Kocher, Todd Cherry, Stephan Kroll, Robert J. Netzer and Matthias Sutter:** Conditional cooperation on three continents.

2007-01 **Martin Kocher, Matthias Sutter and Florian Wakolbinger:** The impact of naïve advice and observational learning in beauty-contest games.

University of Innsbruck

Working Papers in Economics and Statistics

2007-08

Andreas Brezger and Stefan Lang

Simultaneous probability statements for Bayesian P-splines

Abstract

P-splines are a popular approach for fitting nonlinear effects of continuous covariates in semiparametric regression models. Recently, a Bayesian version for P-splines has been developed on the basis of Markov chain Monte Carlo simulation techniques for inference. In this work we adopt and generalize the concept of Bayesian contour probabilities to additive models with Gaussian or multicategorical responses. More specifically, we aim at computing the maximum credible level (sometimes called Bayesian p-value) for which a particular parameter vector of interest lies within the corresponding highest posterior density (HPD) region. We are particularly interested in parameter vectors that correspond to a constant, linear or more generally a polynomial fit. As an alternative to HPD regions simultaneous credible intervals could be used to define pseudo contour probabilities. Efficient algorithms for computing contour and pseudo contour probabilities are developed. The performance of the approach is assessed through simulation studies. Two applications on the determinants of undernutrition in developing countries and the health status of trees show how contour probabilities may be used in practice to assist the analyst in the model building process.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)