Brezger, Lang:

Simultaneous probability statements for Bayesian P-splines

Projektpartner

Simultaneous probability statements for Bayesian P-splines

Andreas Brezger and Stefan Lang
Department of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

April 25, 2005

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 Bayesian P-splines within a generalized additive models framework. 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 and applications to data for the Munich rental guide and on undernutrition in Zambia and Tanzania.

1 Introduction

Consider the additive model

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

where the mean of a continuous response variable y_i is the sum of nonlinear but sufficiently smooth functions $f_1, \ldots, f_p$ of the covariates $x_i = (x_{i1}, \ldots, x_{ip})'$.

Currently one of the most popular approaches for estimating the functions f_j is based on P(enalized)-splines as proposed by Eilers and Marx (1996), see also Marx and Eilers (1998) and Eilers and Marx (2004). The approach assumes that the unknown functions f_j can be approximated by a spline of degree l with equally spaced knots $x_{j,min} = \zeta_{j0} < \zeta_{j1} < \cdots < \zeta_{j,r-1} < \zeta_{jr} = x_{j,max}$ within the domain of x_j. The spline can be written in terms of a linear combination of $r + l$ B-spline basis functions $B_{j\rho}$, i.e.

$$f_j(x_j) = \sum_{\rho=1}^{r+l} \beta_{j\rho} B_{j\rho}(x_j). \tag{2}$$

By defining the design matrices X_j, where the element in row i and column ρ is given by $X_j(i, \rho) = B_{j\rho}(x_{ij})$, we can rewrite the predictor in (1) in matrix notation as

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p.$$

Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on squared differences of adjacent B-spline coefficients to guarantee sufficient smoothness of the fitted curves. Recently, Lang and Brezger (2004) developed a Bayesian version of P-splines which is based on stochastic analogues of difference penalties as priors for the regression coefficients. More specifically, first or second order random walks are used as smoothness prior, i.e.

$$\beta_{j\rho} = \beta_{j,\rho-1} + u_{j\rho}, \quad \text{or} \quad \beta_{j\rho} = 2\beta_{j,\rho-1} - \beta_{j,\rho-2} + u_{j\rho} \tag{3}$$

with Gaussian errors $u_{j\rho} \sim N(0, \tau_j^2)$ and diffuse priors $\beta_{j1} \propto const$, or β_{j1} and $\beta_{j2} \propto const$, for initial values, respectively. The priors (3) can be equivalently written in the form of a global smoothness priors

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

with appropriate penalty matrix K_j. In a further stage of the hierarchy, inverse Gamma hyperpriors $p(\tau_j^2) \sim IG(a_j, b_j)$ are assigned to the variances τ_j^2 (and the overall variance parameter σ^2). Bayesian inference for the regression and variance parameters can be based on MCMC simulation. For Gaussian responses, as primarily considered in this paper, a Gibbs sampler can be used to successively update the parameters $\beta_1, \ldots, \beta_p, \tau_1^2, \ldots, \tau_p^2$, see Lang and Brezger (2004) for details.

Currently, interval estimates are limited to *pointwise credible intervals* for the regression parameters β_j and the functions f_j evaluated at the observations. The primary goal of this paper is to develop techniques for obtaining *simultaneous probability statements* about the regression parameters and as a result about the unknown functions. 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. Since the functions f_j are centered around zero, a constant fit corresponds to $\beta_j = 0$, i.e. the particular covariate has no effect on the conditional mean of the response variable. 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.

The plan of the paper is as follows:

- In Section 2.1 we review ideas recently proposed by Held (2004) for estimating and computing contour probabilities or Bayesian p-values. 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.

- We derive in Section 2.2 conditions on the regression parameters that lead to a constant, linear or in general a polynomial fit and develop efficient algorithms for computing the corresponding (pseudo) contour probabilities. So far, algorithms

and software are available for models with Gaussian responses and models where latent Gaussian responses can be obtained through data augmentation. The latter is possible for most categorical regression models, see Albert and Chib (1993) for probit models and Holmes and Held (2005) for logit models.

- The performance of the different approaches is assessed through simulation studies (Section 3). We finally present in Section 4 applications to data for the Munich rental guide and on undernutrition in Zambia and Tanzania.

2 Contour probabilities for P-Splines

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 β. Hence the index j in (2) and everywhere else is suppressed.

2.1 Contour probabilities

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

$$P(\beta^* \mid y) = P\{p(\beta \mid y) \leq p(\beta^* \mid y) \mid y\}, \tag{4}$$

see Box and Tiao (1973) and Held (2004). Note that $p(\beta \mid y)$ is treated here as a random variable. In the following we briefly review concepts for estimating the probability (4) from posterior samples $\beta^{(t)}$, $t = 1, \ldots, T$ obtained via MCMC simulation.

Held (2004) proposes to estimate (4) by

$$\widehat{P(\beta^* \mid y)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{p(\beta^{(t)} \mid y) \leq p(\beta^* \mid y)\}, \tag{5}$$

i.e. the proportion of the MCMC samples for which the posterior density is smaller than the density of the point of interest β^*.

Unfortunately the functional form of the marginal density $p(\beta \mid y)$ is unknown (otherwise MCMC would not be necessary) and we have to employ some method of density estimation to obtain estimates $\widehat{p(\beta^{(t)} \mid y)}, t = 1, \ldots, T$, and $\widehat{p(\beta^* \mid y)}$. For (latent) Gaussian responses the full conditionals $p(\beta \mid \cdot)$, i.e. the conditional densities of β 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 $\beta^{(1)}, \ldots, \beta^{(T)}$ and no smoothing parameter is involved. Estimates for the marginal density $p(\beta \mid y)$ can be obtained using the Rao-Blackwell theorem

$$\widehat{p(\beta \mid y)} = \frac{1}{T} \sum_{v=1}^{T} p(\beta \mid \alpha_-^{(v)}, y), \tag{6}$$

where $\alpha_-^{(v)}$ comprises all model parameters excluding β and hence $p(\beta \mid \alpha_-^{(v)}, y)$ denotes the full conditional density of β. As an alternative to the mean in (6) Held (2004) suggests to use the median, i.e.

$$\widehat{p(\beta \mid y)} = med_{1 \leq v \leq T} \left\{ p(\beta \mid \alpha_-^{(v)}, y) \right\}. \tag{7}$$

As an advantage, the estimated contour probabilities are invariant to monotonic transformations of $p(\beta \,|\, y)$ in (4). For instance, one could replace $p(\beta \,|\, \alpha_-^{(v)}, y)$ in (7) by the log density, i.e.

$$\log(\widehat{p(\beta \,|\, y)}) = med_{1 \leq v \leq T}\left\{\log(p(\beta \,|\, \alpha_-^{(v)}, y))\right\}. \tag{8}$$

Usually, this is computationally more favorable than using the density directly (see Subsection 2.3) and also more robust against extreme samples.

Summarizing, the contour probability (4) is estimated by replacing the marginal densities with (6), (7), or (8) if log densities are used. Using (7), for instance, we obtain

$$\widehat{P(\beta^* \,|\, y)} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\left\{med_{1 \leq v \leq T}\left\{p(\beta^{(t)} \,|\, \alpha_-^{(v)}, y)\right\} \leq med_{1 \leq v \leq T}\left\{p(\beta^* \,|\, \alpha_-^{(v)}, y)\right\}\right\} \tag{9}$$

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 β^* of interest. For instance, Besag et al. (1995) propose to define a simultaneous credible interval as the hyperrectangular defined by

$$[\beta_\rho^{[T+1-t^*]}, \beta_\rho^{[t^*]}] \quad \rho = 1, \ldots, r + l, \tag{10}$$

where $\beta_\rho^{[t]}$, $t = 1, \ldots, T$ denotes the ordered samples of the parameter β_ρ. The index t^* is the smallest integer such that the hyperrectangular (10) contains at least 100α percent of the samples $\beta^{(1)}, \ldots, \beta^{(T)}$ if α is the desired level of the credible interval.

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

2.2 Contour probabilities for P-Splines

In the context of P-splines, we are particularly interested in parameters $\beta = \beta^*$ that lead to a constant, linear or in general a polynomial fit. Since P-splines are centered around zero a constant fit corresponds to $\beta^* = 0$, 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 $f(x)$ reduces to a polynomial of degree $s \leq l$ if the $(s+1)$-th differences of the regression parameters are zero, i.e.

$$\Delta^{s+1}\beta_\rho = 0, \quad \rho = s + 2, \ldots, r + l, \tag{11}$$

or in matrix notation

$$D_{s+1}\beta = 0,$$

where D_{s+1} is a difference matrix of order $s + 1$. A proof can be found in the Appendix. In order to compute (pseudo) contour probabilities the full conditional of $D_s\beta$ must be computed. The full conditional of β is multivariate Gaussian

$$\beta \,|\, \alpha_-, y \sim N(m, P^{-1}) \tag{12}$$

with

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

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

$$\tilde{\beta} \mid \alpha_-, y \sim N(\tilde{m}, \tilde{P}^{-1}), \tag{13}$$

with mean $\tilde{m} = D_s m$ and precision matrix $\tilde{P} = D_s P^{-1} D_s'$. Note that for the special case $s = 0$, i.e. $D_s = I$, we recover (12) as full conditional for $D_s\beta$.

2.3 Computational aspects

This section is concerned with computational aspects of the estimator (9). We will distinguish the two cases $s = 0$ and $s > 0$.

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

$$log(p(\beta^{(t)} \mid \alpha_-^{(v)}, y)) = \frac{1}{2}\log(|P^{(v)}|) - \frac{1}{2}(\beta^{(t)} - m^{(v)})'P^{(v)}(\beta^{(t)} - m^{(v)}) \tag{14}$$

for $t, v = 1, \ldots, T$ in order to estimate (9). Here, $P^{(v)}$ is the posterior precision matrix evaluated at the v-th sample of τ^2 and σ^2 and $m^{(v)}$ is the posterior mean evaluated at the v-th sample of P, σ^2 and $\tilde{\eta}$. It is useful to decompose the quadratic form in (14) by

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

This shows that (9) can be evaluated by computing and storing the samples $\log(|P^{(t)}|)$, $(\beta^{(t)})'X'X\beta^{(t)}$, $(\beta^{(t)})'K\beta^{(t)}$, $(m^{(t)})'P^{(t)}m^{(t)}$ and $(m^{(v)})'P^{(v)}\beta^{(t)}$. Except $(m^{(v)})'P^{(v)}\beta^{(t)}$ 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 $(m^{(v)})'P^{(v)}\beta^{(t)}$. For $t > v$ the quantity $(m^{(v)})'P^{(v)}\beta^{(t)}$ must be computed *after* the MCMC simulation. This is facilitated by storing $(m^{(v)})'P^{(v)}$ after every iteration of the MCMC sampler.

The case $s > 0$ is computationally more demanding. In this case the log densities

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

must be computed. Evaluation of the quadratic form yields

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

Hence the quantities $\log(|\tilde{P}^{(v)}|)$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{m}^{(v)}$ can be computed as a by product of the MCMC sampler and stored in every iteration. However, the quantities $(\tilde{\beta}^{(t)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ and $(\tilde{m}^{(v)})'\tilde{P}^{(v)}\tilde{\beta}^{(t)}$ can only be stored for $t \leq v$. For $t > v$ both quantities must be computed *after* the MCMC run.

Now we can compute $med_v \left\{ \log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y)) \right\}$ 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 $\tilde{P}^{(v)}$ and with it the quantities $(\tilde{\beta}^{(t)})' \tilde{P}^{(v)} \tilde{\beta}^{(t)}$ and $(\tilde{m}^{(v)})' \tilde{P}^{(v)} \tilde{\beta}^{(t)}$.

    (b) Compute $\log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y))$.

2. Compute $med_v \left\{ \log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y)) \right\}$.

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

Algorithm 2:

1. For $v = 1, \ldots, T$:

    (a) Compute $\tilde{P}^{(v)}$.

    (b) For $t = 1, \ldots, T$:
    If $t \leq v$: Compute $\log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y))$ based on the stored quantities.
    If $t > v$:
    Compute first $(\beta^{(t)})' \tilde{P}^{(v)} \beta^{(t)}$ and $(\tilde{m}^{(v)})' \tilde{P}^{(v)} \tilde{\beta}^{(t)}$ and then $\log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y))$.

2. For $v = 1, \ldots, T$: Compute $med_v \left\{ \log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y)) \right\}$.

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 $\log(p(\tilde{\beta}^{(t)} | \alpha_-^{(v)}, y))$, $t, v = 1, \ldots, T$, before computing the median. A remedy is to take only every k-th sample to estimate $\widehat{p(\tilde{\beta}^{(t)} | y)}$, 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.

3 Simulations

We realized an extensive simulation study in order to compare the performance of contour and pseudo contour probabilities. We investigated the functions

$$y_i = 1 + k \cdot \sin(2\pi x_i) + \epsilon_i, \quad k = 0.0, 1.0, 1.5, \tag{15}$$

and

$$y_i = 1 + x_i + k \cdot \sin(2\pi x_i) + \epsilon_i, \quad k = 0.0, 1.0, 1.5. \tag{16}$$

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 (15) and (16) 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 $k = 0, 1.0, 1.5$ and model (15) and a SNR of approximately $0.17, 0.53$ and 1.47 for $k = 0, 1.0, 1.5$ and model (16), respectively. We used an IG(0.001,0.001) prior for the scale parameter σ^2 and the variance parameter τ^2. Figures 1 and 2 display the simulated functions as well as a typical replication from the generated response for $k = 0, 1.0, 1.5$.

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 3 shows boxplots of p-values for model (15). Note that $\Delta^s\beta = 0$ corresponds to a constant fit (i.e. no effect) for both, $s = 0$ and $s = 1$. The results for model (16) exhibit mainly the same behavior and are displayed in Figure 4. 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 $s = 1$.

- *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 16, 1.0 in model 15) 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).

- *Medium signal to noise ratio (SNR=1.47, 2.25)*
  For medium signal to noise ratios (1.47 in model 16, 2.25 in model 15) 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

(compare Figure 5). Pseudo contour probabilities, on the other hand, perform very poorly for difference orders higher than 1.

- *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 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.

We conclude that pseudo contour probabilities seem to underestimate the p-values regarding the decision whether a covariate has an effect on the response or not, whereas for the decision for modeling an effect linearly (or by a polynomial of higher degree) they seem to behave too conservative. Contour probabilities seem to give the most reasonable results. However, 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).

4   Applications

In this section we illustrate the performance of the previously described model selection tools by applications to complex data sets. First, we reanalyze data from the official Munich rental guide of the year 2003 using the model developed by Fahrmeir, Biller, Brezger, Gieger, Hennerfeind, Jerak and Schmid (2003). A brief description of the data is given in Subsection 4.1. For a more detailed description we refer the reader to Fahrmeir et al. (2003).

Our second example investigates undernutrition of children in Zambia and Tanzania and is based on data already analyzed by Kandala, Lang, Klasen and Fahrmeir (2001). We give a brief description of the data in Subsection 4.2.

4.1   Rental guide

According to the German rental law, owners of apartments or flats can base an increase in the amount that they charge for rent on "average rents" for flats comparable in type, size, equipment, quality and location in a community. To provide information about these "average rents", most larger cities publish "rental guides", which can be based on regression analysis with rent as the dependent variable. We use data from the City of Munich, collected in 2002 by Infratest Sozialforschung for a random sample of more than 3000 flats. As response variable we choose

R monthly net rent per square meter in German Marks, that is the monthly rent minus calculated or estimated utility costs.

Covariates characterizing the flat were constructed from almost 200 variables out of a questionnaire answered by tenants of flats. In our reanalysis we use the same covariates as in the official Munich '03 rental guide. These are the continuous covariates "floor space" (F) and "year of construction" (Y) and a vector v of 23 binary covariates characterizing the quality of the flat, e.g. the kitchen and bath equipment, the quality of the heating or the quality of the warm water system. Another important covariate is the location L of the flat in Munich. For the official Munich '03 rental guide, location in the city was assessed in three categories (average, good, top) by experts.

Thus, our model has a semiparametric predictor of the form

$$\eta = \gamma_0 + f_1(F) + f_2(Y) + \gamma'x,$$

where F denotes the floor space in square meters, Y denotes the year of construction and γ contains all covariates to be modeled parametrically, e.g. the quality of the kitchen and bath equipment, location of the flat, etc. The main effects f_1 and f_2 of floor space and year of construction are modeled by cubic P-splines with 20 knots and a second order random walk penalty. Since we aim at deciding whether the nonparametric modeling of the continuous covariates F and Y could possibly be replaced by a more parsimonious polynomial fit, we do not compare the results for the fixed effects to the results of Fahrmeir et al. (2003).

By visual inspection, the nonparametric effects as depicted in Figure 6 seem justified and one presumes that they can not be adequately modeled by low degree polynomials. Table 1 displays the p-values obtained from contour probabilities and pseudo contour probabilities. Obviously, there is strong evidence for the need of nonparametric modeling of the effects for F and Y regarding contour probabilities, as all p-values are either exactly or at least near zero. However, pseudo contour probabilities only suggest a linear effect of both covariates. For comparison we additionally estimate the model with only fixed effects included, i.e. we model F and Y linearly. Comparison of the DIC (Table 2) shows that the semiparametric model clearly outperforms the parametric model suggested by pseudo contour probabilities.

Figure 7 depicts the linear and nonparametric estimates for F and Y together with the partial residuals

$$r_F = y - \hat{\eta}^{-F} \quad \text{and} \quad r_Y = y - \hat{\eta}^{-Y},$$

where $\hat{\eta}^{-F}$, $\hat{\eta}^{-Y}$ is the estimated predictor with the estimated effect of F and Y, respectively, excluded. Figure 8 displays the mean and the standard deviation of the partial residuals. Note, that for F the data is rounded to integer values to compute means and standard deviations. A comparison between the figures for the linear and the nonparametric model gives further evidence for the superiority of the semiparametric model, since the nonparametric estimates show a clearly better adaptation to the partial residuals.

We conclude that nonparametric modeling of the continuous covariates is reasonable according to the results obtained from both, the estimated contour probabilities and the DIC. Pseudo contour probabilities, however, are somewhat misleading and gave no satisfactory results, similar to the simulation results in Section 3

4.2 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

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

where the mothers body mass index bmi and the age of the child agc are modeled non-parametrically with Bayesian P-splines. The expression $f_{spat}(d)$ denotes the spatial effect associated with the district d the child lives in, and is modeled as sum of a structured and an unstructured random effect for Zambia. For Tanzania they exclude the unstructured effect from the model. The fixed effects γ 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 and pseudo contour probabilities. Moreover, we compare different models in terms of the DIC. In a first attempt, we use the model developed by Kandala et al. (2001) and model both continuous covariates, bmi and agc, nonparametrically by P-splines. Based on the contour probabilities for the nonparametric effects obtained from this model, we investigate a number of different model specifications, where bmi and agc are modeled nonparametrically, or parametrically with polynomials of different degrees. Following Kandala et al. (2001), spatial heterogeneity is captured by adding an unstructured and a structured random effect for Zambia and a structured random effect for Tanzania and the remaining covariates are modeled parametrically throughout our analysis. The models under consideration are:

$$
\begin{aligned}
\eta_1 &= \gamma_0 + f_1(bmi) & + f_2(agc) & & + f_{spat}(d) + \gamma'x \\
\eta_2 &= \gamma_0 + f_1(bmi) & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \gamma'x \\
\eta_3 &= \gamma_0 + f_1(bmi) & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & + f_{spat}(d) + \gamma'x \\
\eta_4 &= \gamma_0 + & + f_2(agc) & & + f_{spat}(d) + \gamma'x \\
\eta_5 &= \gamma_0 + & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \gamma'x \\
\eta_6 &= \gamma_0 + & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & + f_{spat}(d) + \gamma'x \\
\eta_7 &= \gamma_0 + \alpha_1 bmi & + f_2(agc) & & + f_{spat}(d) + \gamma'x \\
\eta_8 &= \gamma_0 + \alpha_1 bmi & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \gamma'x \\
\eta_9 &= \gamma_0 + \alpha_1 bmi & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & + f_{spat}(d) + \gamma'x \\
\eta_{10} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 & + f_2(agc) & & + f_{spat}(d) + \gamma'x \\
\eta_{11} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 & + \beta_1 agc + \beta_2 agc^2 & & + f_{spat}(d) + \gamma'x \\
\eta_{12} &= \gamma_0 + \alpha_1 bmi + \alpha_2 bmi^2 & + \beta_1 agc + \beta_2 agc^2 + \beta_3 agc^3 & + f_{spat}(d) + \gamma'x
\end{aligned}
$$

Table 3 shows the resulting contour probabilities and pseudo contour probabilities, respectively. In Table 4 values for the deviance, the effective degrees of freedom (pD) and the DIC are displayed. Models are ordered according to the DIC.

For the effect of agc the p-values obtained from both, contour probabilities based on the median and on the log-density, are zero for difference order 2 and lower, and still very low for difference order 3. This suggests cubic modeling for the effect of agc, since contour probabilities for difference order 4 are quite high. In contrast, pseudo contour probabilities suggest a linear effect. The p-values based on contour probabilities in favor of 'no effect' of bmi are in a medium region and allow no clear decision. Pseudo contour probabilities allow no clear decision either since the p-value for difference order $s = 0$ pleads for an effect, whereas the p-value for $s = 1$ tends towards 'no effect'.

The best fit in terms of the DIC is achieved by model 7, featuring a nonparametric fit for agc and a linear fit for bmi. However, the models 1, 9, 10 and 12 perform almost equally well. Summarizing table 4, we may derive that the DIC slightly favors nonparametric over cubic modeling of agc and clearly votes for a linear effect of bmi. Even a quadratic effect of bmi is considered.

In Figure 9 we compare the nonparametric, the linear and the quadratic fit for agc, and the nonparametric, the quadratic and the cubic effect for bmi. The depicted effects correspond to the best model in terms of DIC that includes the corresponding type of modeling of an effect. Here we see that the functional form of the two effects can indeed satisfactorily be modeled by a linear and a cubic term, respectively. The subjective reduction of the bias of the quadratic fit of bmi compared to the linear is mainly due to few observations at the borders.

The results for Tanzania are reported in Tables 5 and 6. A comparison of parametric and nonparametric estimations is displayed in Figure 10. The main differences compared to the results for Zambia are a more curved estimation of the effect of bmi and an additional local maximum of the effect of agc at in the interval $[25, 30]$.

This is confirmed by the obtained contour probabilities (based on the median and the mean of the log-density), which are clearly smaller for 4th differences for agc and still somewhat smaller for difference order $s = 1$ for bmi. Figure 10 shows that a cubic fit totally misses the local maximum exhibited by the nonparametric estimate of the effect of agc. Pseudo contour probabilities are smaller than for Zambia, too, but still prefer a linear effect of agc and exclude bmi from the model. Comparing the models in terms of the DIC, we observe a distinct improvement when using a nonparametric fit for agc (models 1, 4, 7, 10) instead of a cubic fit (models 3, 6, 9, 12). A quadratic fit for bmi (models 10, 11, 12) only slightly improves the model compared to a linear fit (models 7, 8, 9),

Summarizing, we observe that the contour probabilities are somewhat conservative compared to the DIC, which tends to chose more complex models. This is in agreement with the typical experience, that the DIC – like the AIC – is likely to prefer (too) complex models.

5   Conclusion

We applied contour probabilities and pseudo contour probabilities in order to decide whether nonparametric modeling of continuous covariates is necessary or if parametric modeling by polynomials of small degree is sufficient. In a simulation study we found contour probabilities to perform clearly superior compared to pseudo contour probabilities. In two applications we highlight that contour probabilities qualify as a helpful instrument for model selection. We conclude that contour probabilities give useful hints for a careful model selection. They seem to behave somewhat more conservative regarding the possible model fit compared to the DIC. Therefore, we recommend not to rely solely on them, but to take into account other model selection tools as for example the DIC, especially when the resulting contour probabilities are not close to 0 or 1.

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. Instead an approach of Chib and Jeliazkov (2001) could be used. This might be a challenge for future research.

A Proof of equation (11)

For a proof of (11) we exploit the fact that the B-spline basis functions in (2) for representing the spline can be computed as differences of truncated power functions (e.g. Eilers and Marx 2004), i.e.

$$B_\rho(x) = -1^{l+1}\Delta^{l+1}t(x,\rho)/(h^l l!), \quad \rho = 1,\dots, r+l \tag{17}$$

where h is the distance between two neighboring knots and $t(x,\rho) := (x - (\zeta_0 + \rho h))_+^l$ is the truncated power function that corresponds to the knot $\zeta_\rho = \zeta_0 + \rho h$.

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

$$\frac{(h^l l!)}{-1^{l+1}}f(x) = \frac{(h^l l!)}{-1^{l+1}}\sum_{\rho=1}^{r+l} B_\rho(x)\beta_\rho = \sum_{\rho=1}^{r+l}\Delta\Delta^l t(x,\rho)\beta_\rho = \sum_{\rho=1}^{r+l}\Delta^l t(x,\rho)\beta_\rho - \sum_{\rho=1}^{r+l}\Delta^l t(x,\rho-1)\beta_\rho$$

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

$$\frac{(h^l l!)}{-1^{l+1}}f(x) = -\sum_{\rho=1}^{r+l-1}\Delta^l t(x,\rho)\Delta\beta_{\rho+1} + \Delta^l t(x,r+l)\beta_{r+l} - \Delta^l t(x,0)\beta_1. \tag{18}$$

Provided that $\Delta\beta_\rho = 0$, the summands in the first term are all zero. The second term in (18) 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 (11).

For an arbitrary degree $s \leq l$ the proof is based on analogous arguments. Using again relationship (17) we get

$$
\begin{aligned}
\frac{(h^l l!)}{-1^{l+1}}f(x) &= \sum_{\rho=1}^{r+l}\Delta^{s+1}\Delta^{l-s}t(x,\rho)\beta_\rho \\
&= a_1\sum_{\rho=1}^{r+l}\Delta^{l-s}t(x,\rho)\beta_\rho + \cdots + a_{s+2}\sum_{\rho=1}^{r+l}\Delta^{l-s}t(x,\rho-(s+1))\beta_\rho
\end{aligned}
\tag{19}
$$

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

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

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

$$\frac{(h^l l!)}{-1^{l+1}}f(x) = (-1)^{s+1}\sum_{\rho=1}^{r+l-s-1}\Delta^{l-s}t(x,\rho)\Delta^{s+1}\beta_{\rho+s+1} + R_1 + R_2 \tag{20}$$

with

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

13

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 $\Delta^{s+1}\beta_\rho = 0$, the sum in (20) 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.

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.

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

Eilers, P.H.C. and Marx, B.D. (1996), Flexible smoothing using B-splines and penalized likelihood (with comments and rejoinder), *Statistical Science*, 11 (2), 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., Biller, C., Brezger, A., Gieger, C., Hennerfeind, A., Jerak, A. and Schmid, V. (2003), *Teil 2: Statistische Analyse der Nettomieten*, Gutachten zur Erstellung des Mietspiegels für München© 2003, Landeshauptstadt München, Sozialreferat – Amt für Wohnungswesen, *in German.*

Held, L., 2004: Simultaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, Vol.13, No.1, 20-35.

Holmes, C., and Held, L., 2005: On the simulation of Bayesian binary and polychotomous regression models using auxiliary variables. *Bayesian Analysis*, to appear.

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.

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.

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

Table 1: Contour probabilities based on median, mean of log-density and mean of density and pseudo contour probabilities for the effects of F and Y with IG(0.001,0.001) priors on τ_F^2, τ_Y^2 and σ^2.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
F (based on median)	0.0	0.0	0.0	0.01	0.03
F (based on mean of log-density)	0.0	0.0	0.0	0.01	0.04
F (pseudo contour probabilities)	0.0	0.0	0.46	0.94	1.0
Y (based on median)	0.0	0.02	0.01	0.11	0.12
Y (based on mean of log-density)	0.01	0.03	0.03	0.17	0.19
Y (pseudo contour probabilities)	0.0	0.0	0.70	0.99	0.99

Table 2: Deviance, effective degrees of freedom (pD) and DIC for the parametric and the semiparametric model using IG(0.001,0.001) priors on τ_F^2, τ_Y^2 (in the semiparametric model) and σ^2.

	Deviance $D(\bar{\theta})$	pD	DIC
parametric model	12629.7	26.0	12681.7
semiparametric model	12400.0	41.8	12483.6

Table 3: Contour probabilities for the effects of `bmi` and `agc` in Zambia. Displayed are the results for model 1.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`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
`bmi` (pseudo contour probabilities)	0.0	0.45	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
`agc` (pseudo contour probabilities)	0.0	0.0	0.83	1.0	1.0

Table 4: Deviance, effective degrees of freedom (pD) and DIC for models 1-12 for Zambia using IG(0.001,0.001) priors on τ^2_{bmi}, τ^2_{agc} and σ^2.

	Deviance $D(\bar{\theta})$	pD	DIC
Model 7	12640.5	46.7	12733.9
Model 10	12640	47.5	12735
Model 9	12651.5	42.7	12736.9
Model 1	12639.1	49.4	12737.9
Model 12	12650.9	43.9	12738.7
Model 3	12650.9	45.3	12741.5
Model 4	12663	46.6	12756.2
Model 6	12676.5	42.6	12761.7
Model 8	12696.3	41.6	12779.5
Model 11	12695.2	42.5	12780.2
Model 2	12694.3	44.0	12782.3
Model 5	12722	41.1	12804.2

Table 5: Contour probabilities for the effects of `bmi` and `agc` in Tanzania. Displayed are the results for model 1.

difference order	0	1	2	3	4
degree of polynomial	const	const	linear	quadratic	cubic
`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
`bmi` (pseudo contour probabilities)	0.02	0.04	0.89	0.84	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
`agc` (pseudo contour probabilities)	0.0	0.0	0.60	0.77	0.99

Table 6: Deviance, effective degrees of freedom (pD) and DIC for models 1-12 for Tanzania using IG(0.001,0.001) priors on τ^2_{bmi}, τ^2_{agc} and σ^2.

	Deviance $D(\bar{\theta})$	pD	DIC
Model 1	15477.8	39.0	15555.8
Model 10	15489.3	35.0	15559.3
Model 7	15494.7	34.2	15563.1
Model 3	15521.2	32.8	15586.8
Model 12	15531.4	28.9	15589.6
Model 9	15537	27.9	15592.8
Model 4	15539.9	33.3	15606.5
Model 6	15583.4	27.0	15638.4
Model 2	15601	32.0	15665
Model 11	15612.5	27.8	15668.1
Model 8	15618.8	26.9	15672.6
Model 5	15671.8	25.9	15723.6
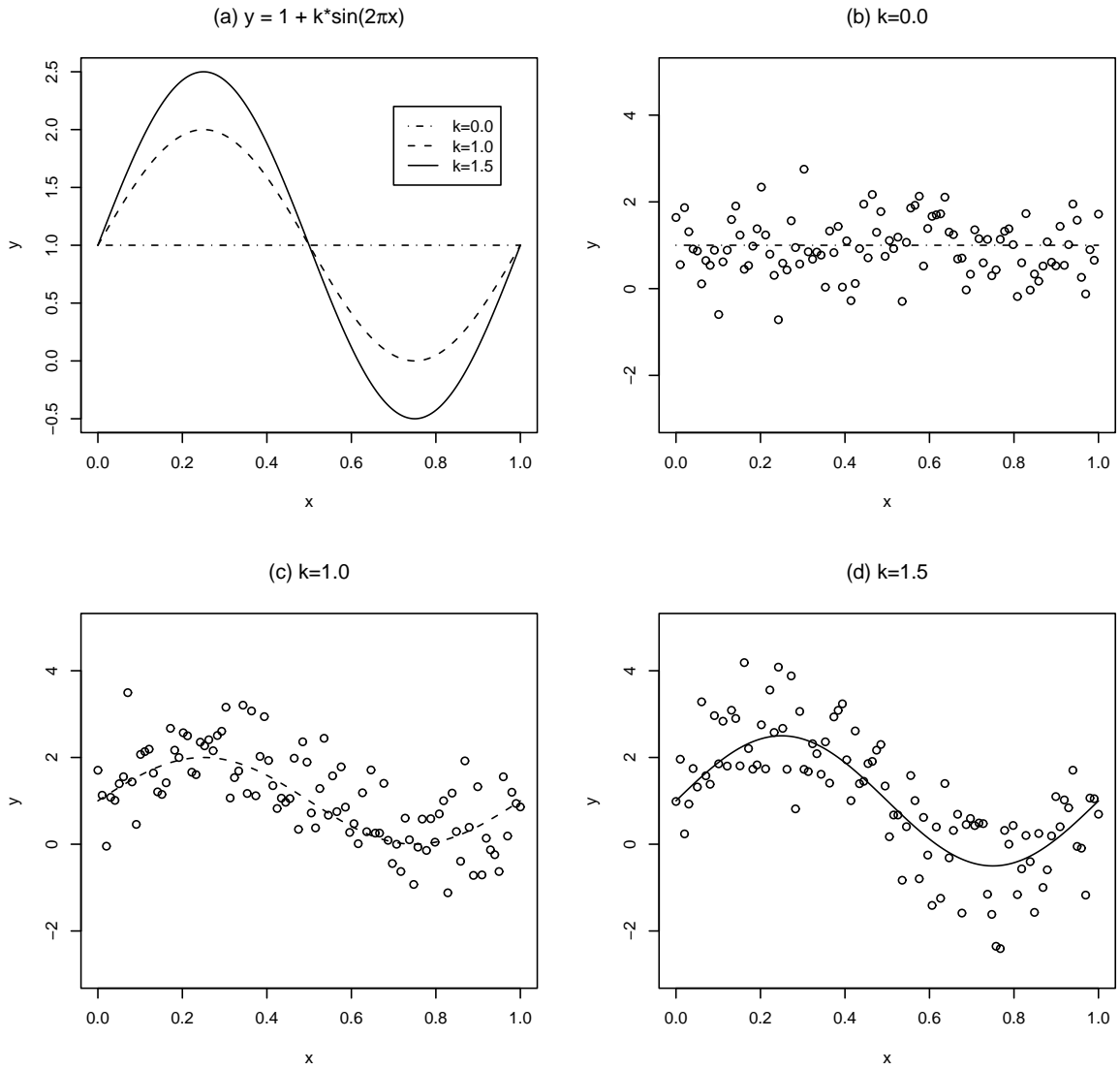
Figure 1: True functions for model (15) (a) and a typical replication of y for $k = 0, 1.0, 1.5$ (b)-(d).
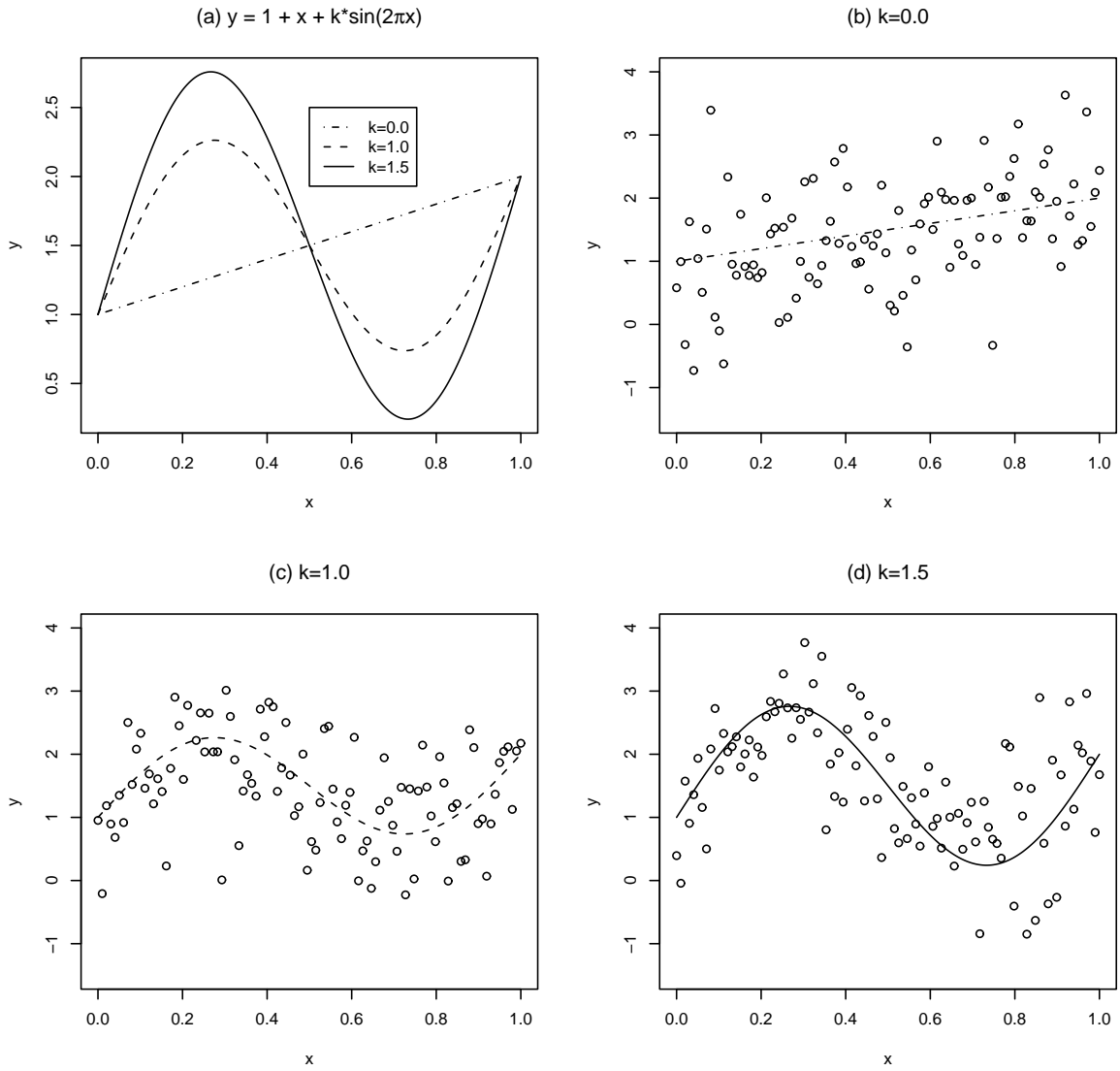
Figure 2: True functions for model (16) (a) and a typical replication of y for $k = 0, 1.0, 1.5$ (b)-(d).
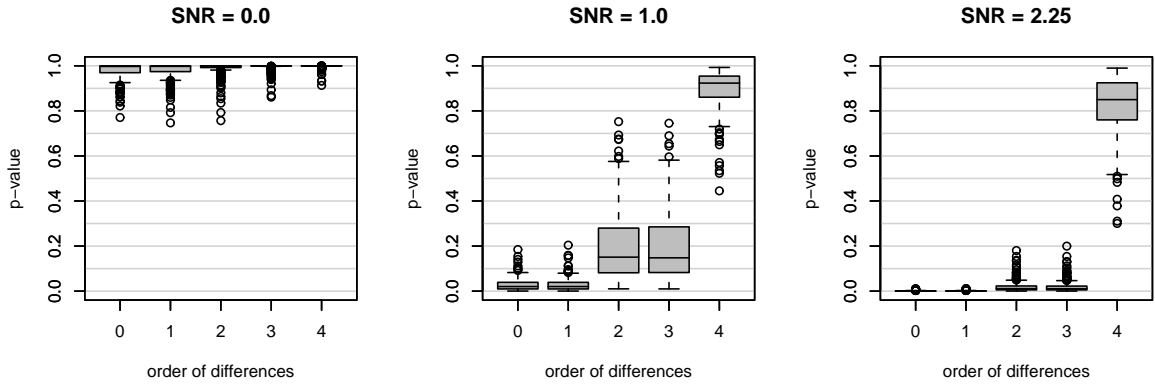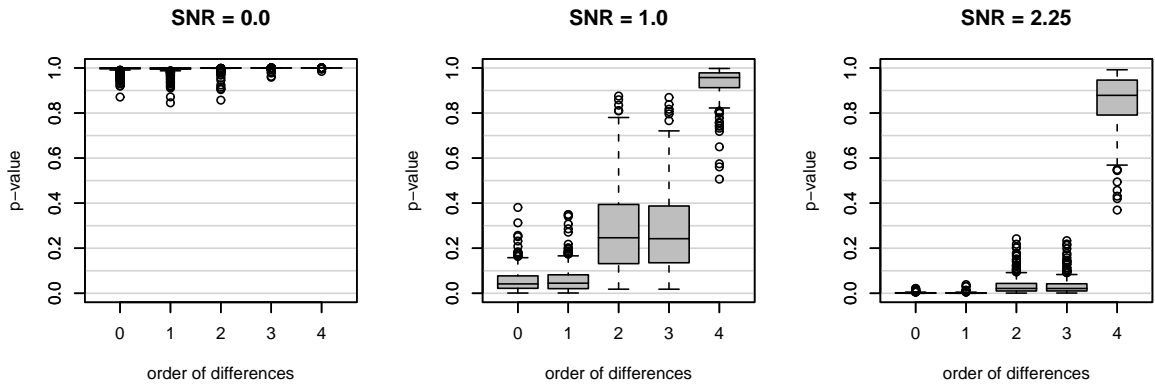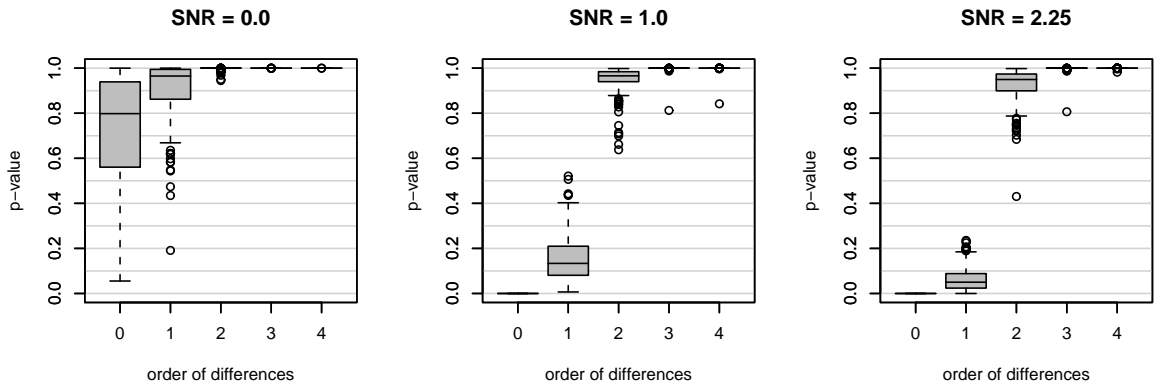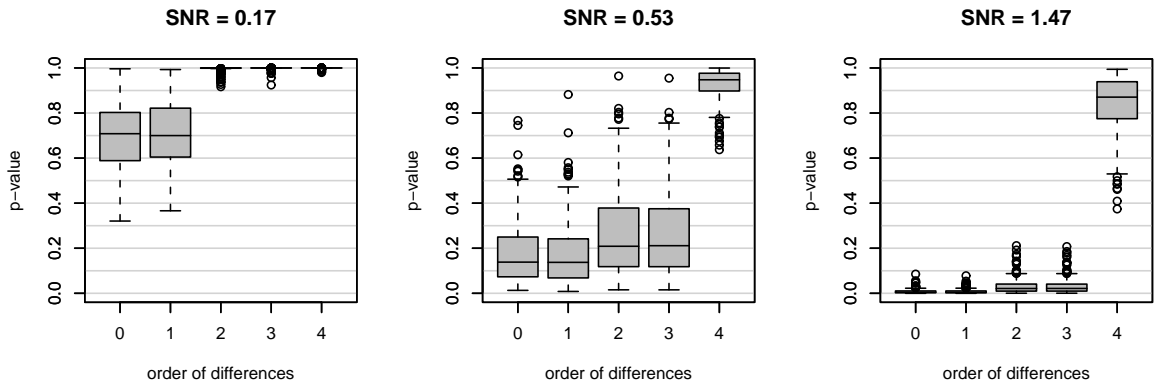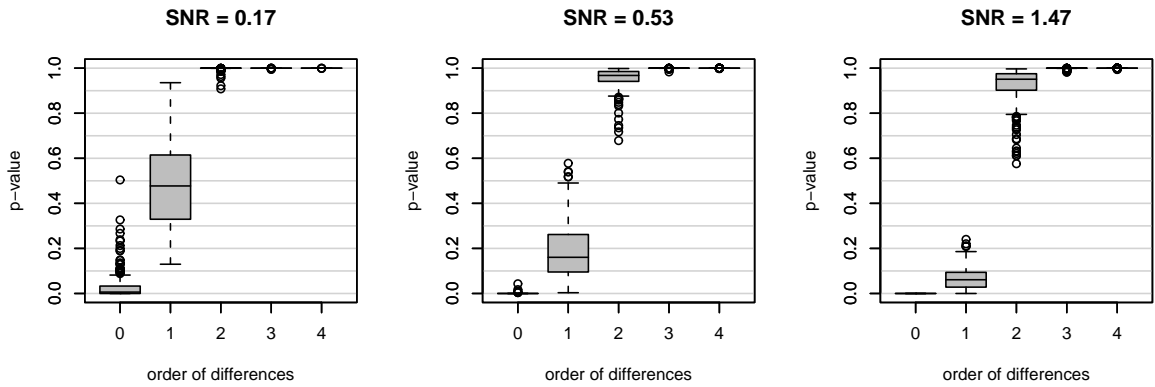
Figure 3: Boxplots of p-values obtained from contour probabilities based on the median (top), contour probabilities based on the mean of the log-density (middle), and pseudo contour probabilities (bottom) for different SNRs and difference orders (model 15). Difference order $s = 0$ and $s = 1$ corresponds to no effect, $s = 2$ $(3, 4)$ corresponds to a linear (quadratic, cubic) effect.
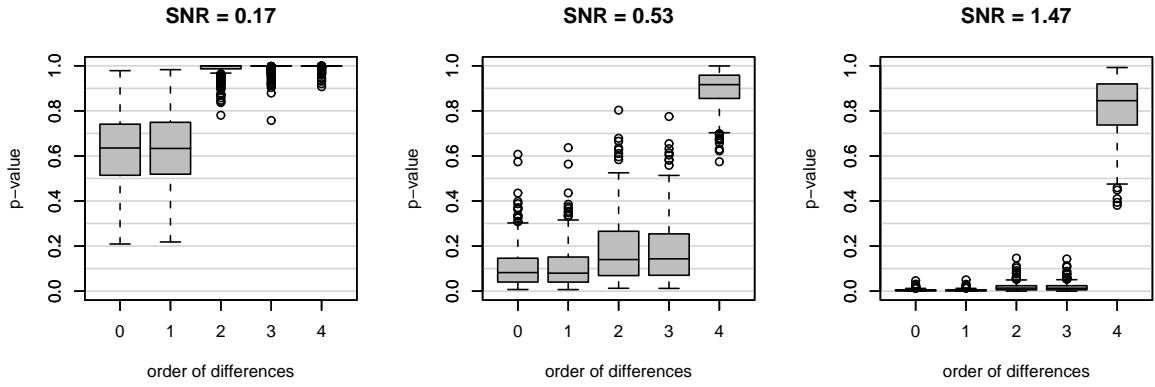
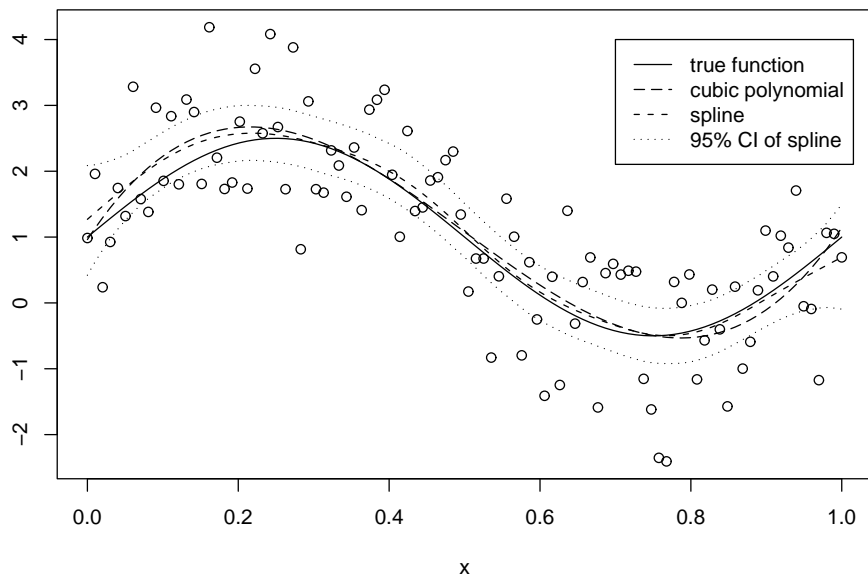Figure 4: Boxplots of p-values obtained from contour probabilities based on the median (top), contour probabilities based on the mean of the log-density (middle), and pseudo contour probabilities (bottom) for different SNRs and difference orders (model 16). Difference order $s = 0$ and $s = 1$ corresponds to no effect, $s = 2$ $(3, 4)$ corresponds to a linear (quadratic, cubic) effect.
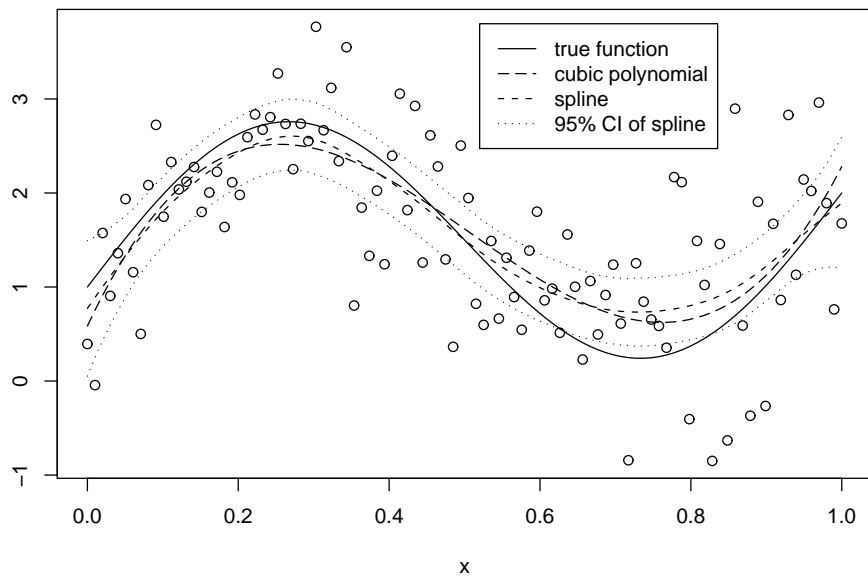
Figure 5: True function, cubic fit, spline fit and 95% pointwise credible levels for a selected replication of model (15) (top) and model (16) (bottom) for $k = 1.5$. Circles visualize the observation points.

Effect of F

floor space in square meters

Effect of Y

year of construction
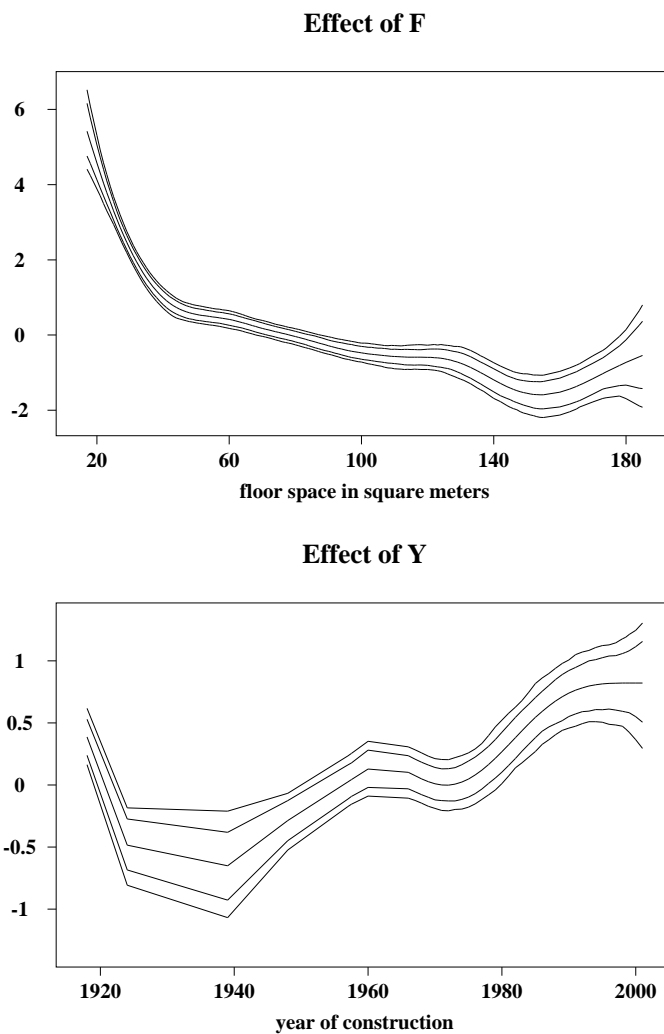
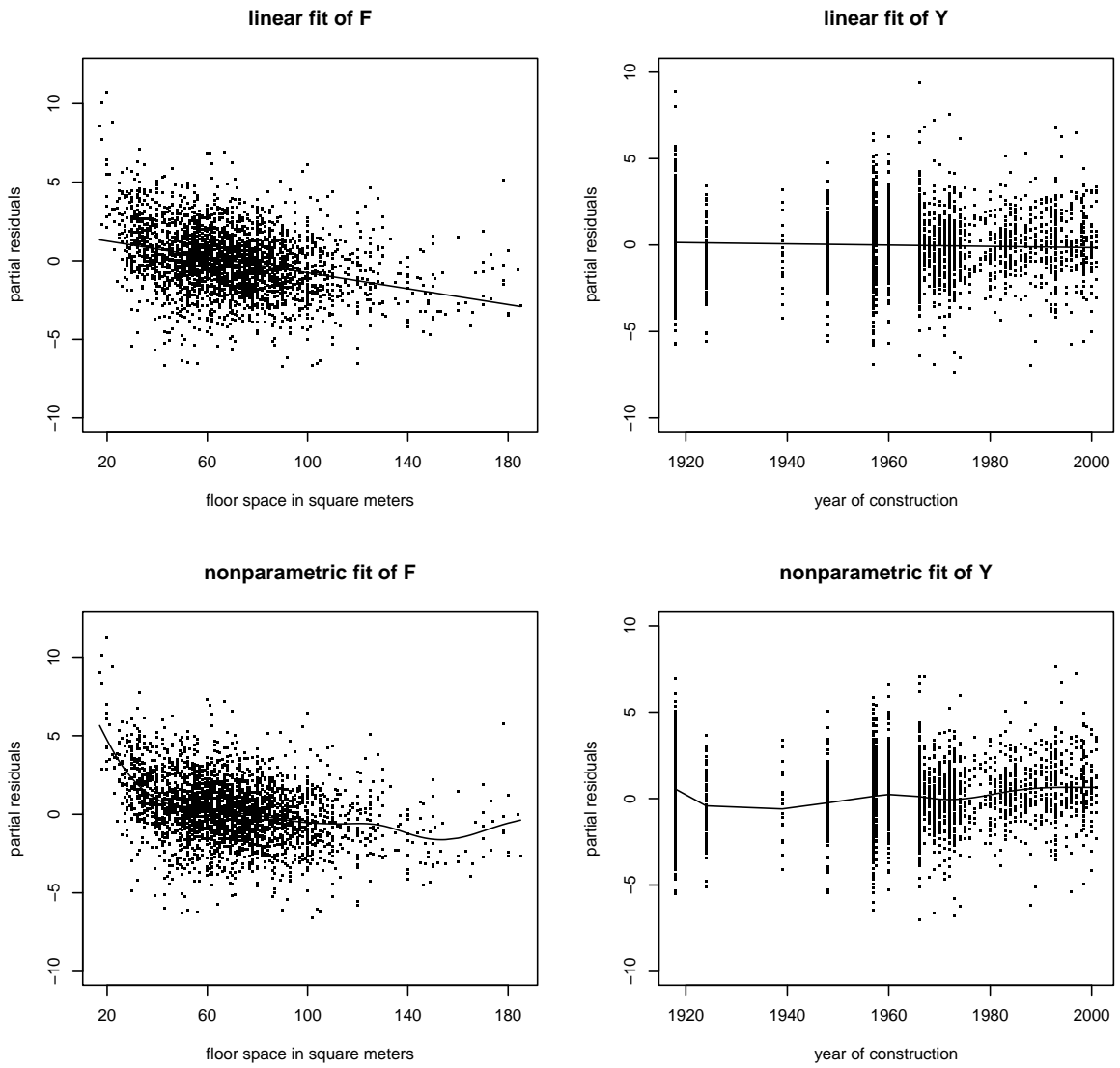Figure 6: Nonparametric effects of F and Y for IG(0.001,0.001) priors on τ_F^2, τ_Y^2 and σ^2.

Figure 7: Estimated function (solid line) and partial residuals (dots) for F (left panel) and Y (right panel). Shown are the results for the parametric (top) and the nonparametric fit (bottom).
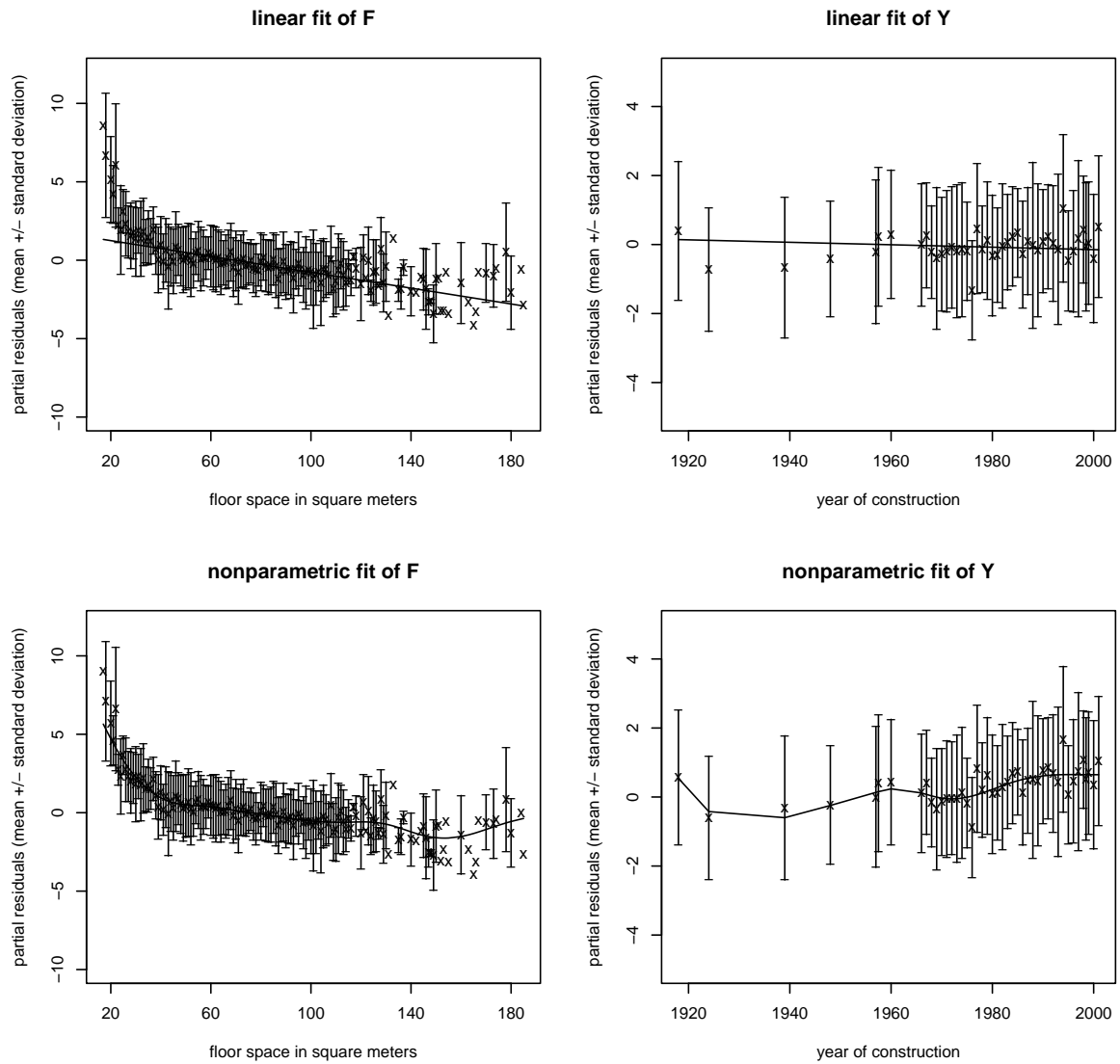
Figure 8: Estimated function (solid line), mean (symbol 'x') and mean ± standard deviation of partial residuals (symbol '-') for F (left panel) and Y (right panel). Shown are the results for the parametric (top) and the nonparametric fit (bottom).
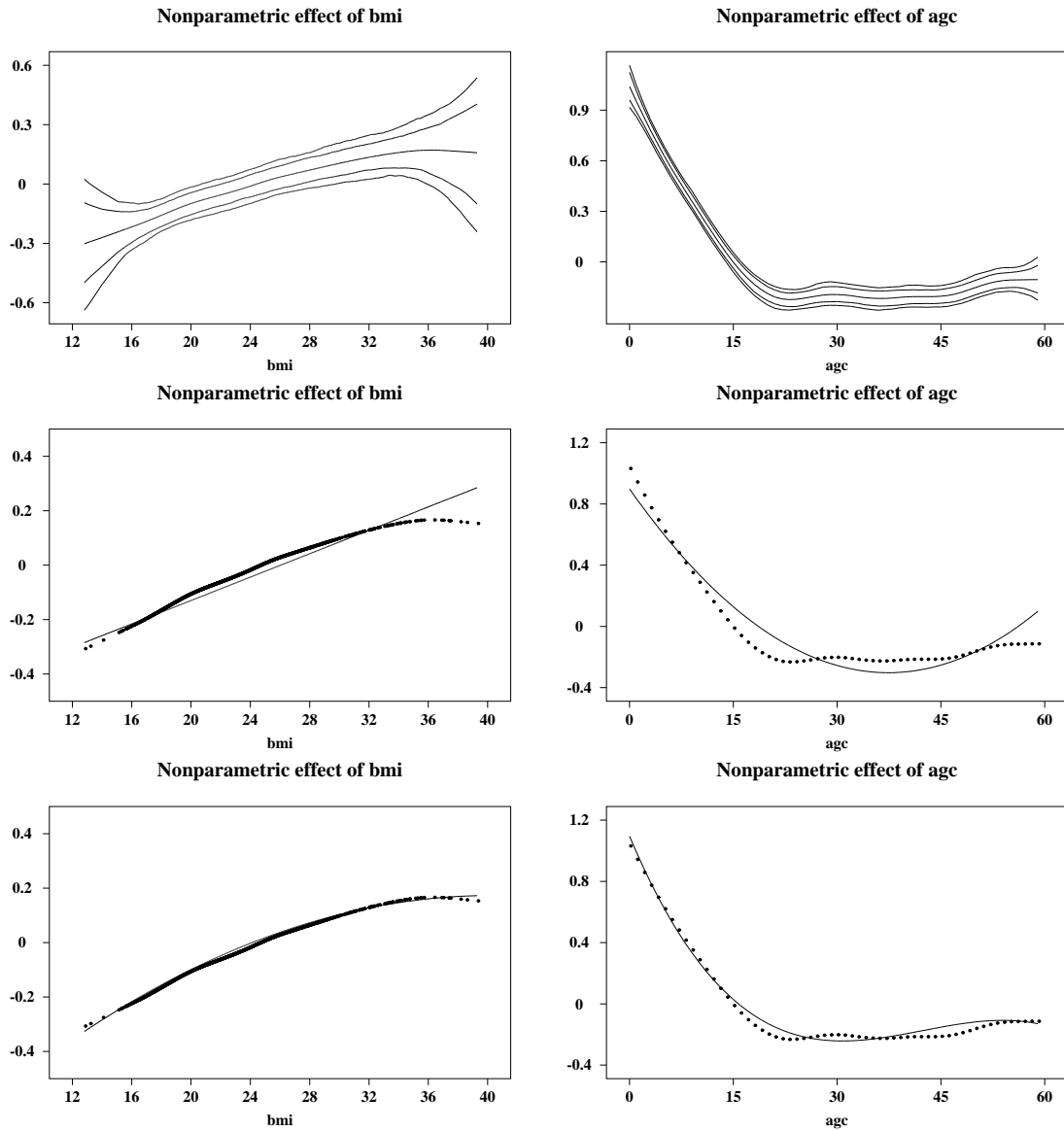
Figure 9: Effects of `bmi` (left panel) and `agc` (right panel) in Zambia for different model specifications. In the two lower panels the solid line corresponds to the parametric fit, the dotted curve displays the spline estimate.
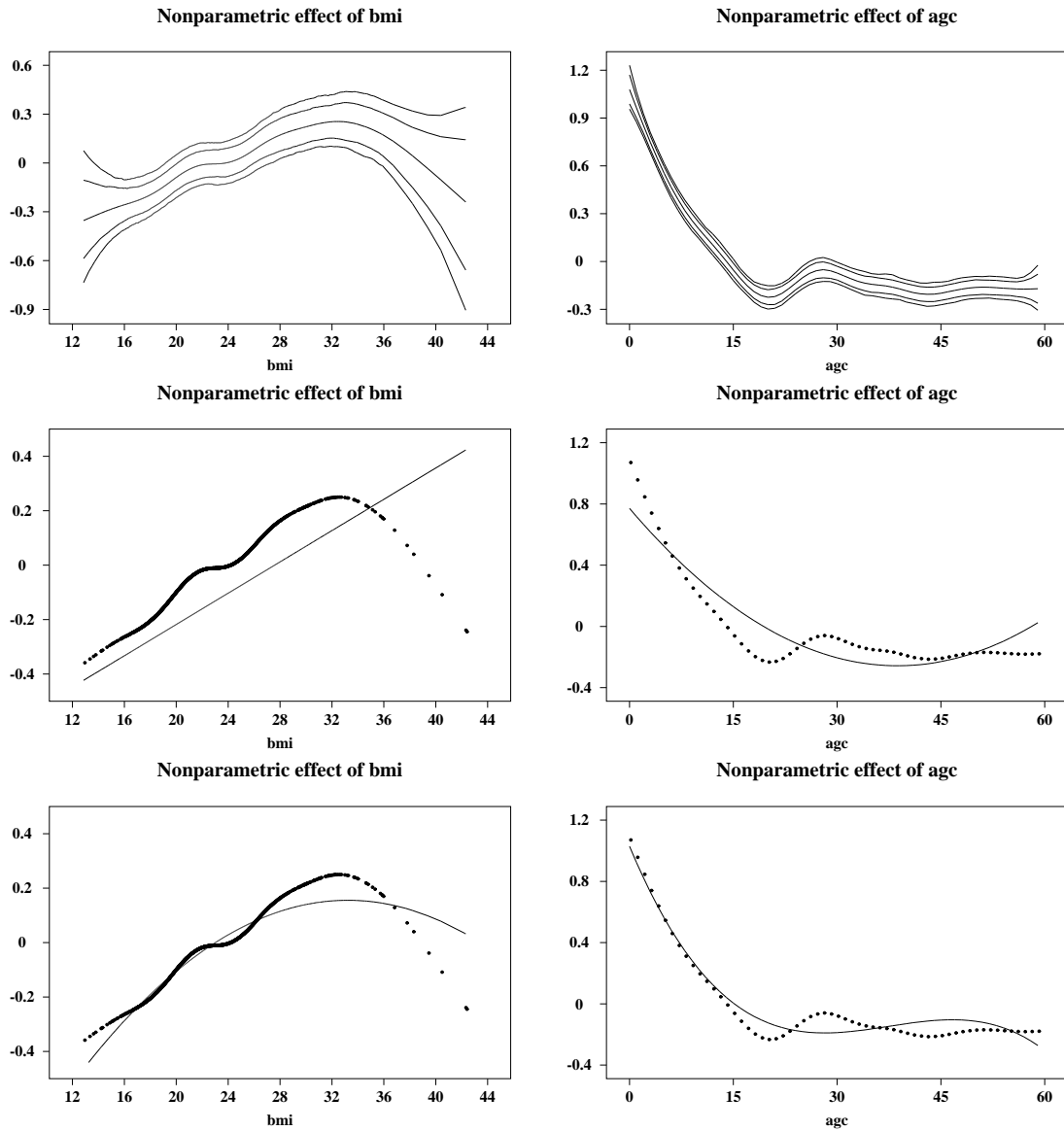
Figure 10: Effects of `bmi` (left panel) and `agc` (right panel) in Tanzania for different model specifications. In the two lower panels the solid line corresponds to the parametric fit, the dotted curve displays the spline estimate.