ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Simultaneous selection of variables and smoothing parameters in structured additive regression models

Christiane Belitz [a], Stefan Lang [b,*]

[a] Department of Statistics, University of Munich, Ludwigstr. 33, D-80539 Munich, Germany
[b] Department of Statistics, University of Innsbruck, Universitätsstr. 15, A-6020 Innsbruck, Austria

ARTICLE INFO

Article history:
Received 21 August 2007
Received in revised form 30 May 2008
Accepted 31 May 2008
Available online xxxx

ABSTRACT

In recent years, considerable research has been devoted to developing complex regression models that can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different types. Much less effort, however, has been devoted to model and variable selection. The paper develops a methodology for the simultaneous selection of variables and the degree of smoothness in regression models with a structured additive predictor. These models are quite general, containing additive (mixed) models, geoadditive models and varying coefficient models as special cases. This approach allows one to decide whether a particular covariate enters the model linearly or nonlinearly or is removed from the model. Moreover, it is possible to decide whether a spatial or cluster specific effect should be incorporated into the model to cope with spatial or cluster specific heterogeneity. Particular emphasis is also placed on selecting complex interactions between covariates and effects of different types. A new penalty for two-dimensional smoothing is proposed, that allows for ANOVA-type decompositions into main effects and an interaction effect without explicitly specifying the main effects. The penalty is an additive combination of other penalties. Fast algorithms and software are developed that allow one to even handle situations with many covariate effects and observations. The algorithms are related to backfitting and Markov chain Monte Carlo techniques, which divide the problem in a divide and conquer strategy into smaller pieces. Confidence intervals taking model uncertainty into account are based on the bootstrap in combination with MCMC techniques.

1. Introduction

In recent years, substantial progress has been made towards realistic, complex regression models. Their capabilities for data analysis go far beyond the traditional linear or generalized linear model. Prominent examples are additive and generalized additive models (Hastie and Tibshirani, 1990; Rigby and Stasinopoulos, 2005; Wood, 2006b), geoadditiv models (Fahrmeir and Lang, 2001; Kamman and Wand, 2003), generalized additive mixed models (Lin and Zhang, 1999; Ruppert et al., 2003), varying coefficient models (Hastie and Tibshirani, 1993), geographically weighted or space varying regression (Fotheringham et al., 2002; Gamerman et al., 2003) and structured additive regression (Fahrmeir et al., 2004; Brezger and Lang, 2006). The latter is the most general model class that can deal simultaneously with nonlinear covariate effects and time trends, unit- or cluster specific heterogeneity, spatial heterogeneity and complex interactions between covariates of different type. Moreover, the models can routinely be estimated using user-friendly statistical

* Corresponding author. Tel.: +43 512 507 7110; fax: +43 512 507 2851.
E-mail addresses: christiane.belitz@stat.uni-muenchen.de (C. Belitz), stefan.lang@uibk.ac.at (S. Lang).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.032

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

Table 1
Determinants of undernutrition: Overview of covariates

Variable	Description
ageC	Child's age in months
bfmC	Months child was breastfed
agebirM	Mother's age at child's birth
bmiM	Mother's body mass index
heightM	Mother's height in cm
sex	Gender of the child
district	District in India the mother and her child lives

software packages, particularly the R packages mgcv (Wood, 2004, 2006b) and GAMLSS (Stasinopoulos et al., 2005) and BayesX (Brezger et al., 2005a,b).

Much less effort, however, has been devoted to model and variable selection for these complex models. For linear models, a number of highly efficient model and variable selection routines is available, particularly stepwise selection procedures (see Miller (2002) for an overview), LASSO (Tibshirani, 1996), least angle regression (Efron et al., 2004), boosting approaches (Bühlmann, 2006) and Bayesian variable selection (Casella and Moreno, 2006). To our knowledge, for semiparametric regression models, the only available procedures are the stepwise algorithm implemented in S-plus for additive models (Chambers and Hastie, 1991), and a selection approach implemented in the mgcv package of R (Wood, 2006a). There is also a Bayesian approach proposed by Shively and Wood (1999) and Yau et al. (2003), but (user-friendly) software is not available.

In this paper we propose algorithms for simultaneous selection of variables and the degree of smoothness in regression models with structured additive predictor. Our algorithms are able to

- decide whether a particular covariate is included in the model,
- decide whether a continuous covariate enters the model linearly or nonlinearly,
- decide whether a spatial effect enters the model,
- decide whether a unit- or cluster specific heterogeneity effect is included in the model,
- select complex interaction effects (two-dimensional surfaces, varying coefficient terms),
- select the degree of smoothness of nonlinear covariate, spatial or cluster specific heterogeneity effects.

Particular emphasis is devoted to modeling and selecting interaction terms. As a side aspect, we propose a new penalty for two-dimensional P-splines. The penalty is an additive combination of two other penalties. It allows an ANOVA type decomposition into main effects and an interaction effect without explicitly specifying the main effects. A particular special case is a main effects model with nonlinear effects modeled by one dimensional P-splines. This allows one to discriminate between a main effects model and a more complex model with additional interaction term within the model selection process.

The basis for inference is a penalized least squares view developed in the next section. A Bayesian approach (without model selection) for models with structured additive predictor has been developed in Lang and Brezger (2004) and Brezger and Lang (2006). We will utilize the close connection to the Bayesian approach for computing (pointwise) confidence intervals for nonlinear effects. Another basis for inference not employed in the paper is a mixed model representation, see Fahrmeir et al. (2004) and Ruppert et al. (2003).

The development of methodology is motivated by various cooperations and consulting cases. In a considerable number of applications, we are confronted with large datasets, many potential covariates of different types and a lack of theory guiding the analyst when specifying promising models. Moreover, the existence or non-existence of complex interactions between covariates is often one of the central scientific questions. The estimation of a particular model including the choice of smoothing parameters poses no difficulty. However, the number of competing models is usually too large for model selection by 'hand'. Hence, flexible, automated and fast procedures are necessary to perform this difficult task.

In this paper we discuss an application from development economics. We analyze data from the second National Family Health Survey (NFHS-2) from India, which was conducted in the years 1998 and 1999. The aim is to study the determinants of child undernutrition measured through an anthropometric indicator. Potential covariates used in this paper are given in Table 1. An important issue is also spatial heterogeneity induced by latent or unobserved covariate effects. It is also well known that there is a gender bias in undernutrition and mortality among new born children in India. Hence, one of the most important scientific questions are possible gender related differences in undernutrition. This leads to the model

$$\begin{aligned} zscore &= f_1(ageC, bfmC) + g_1(ageC, bfmC) \cdot sex + f_2(agebirM) \\ &\quad + g_2(agebirM) \cdot sex + f_3(bmiM) + g_3(bmiM) \cdot sex + f_4(heightM) \\ &\quad + g_4(heightM) \cdot sex + f_5(district) + f_5(district) \cdot sex + \gamma_0 + \gamma_1 sex + \varepsilon, \end{aligned}$$

where $f_1 - f_4$ are possibly nonlinear functions of the continuous covariates, f_5 is a district specific spatial effect and $g_1 - g_5$ are possible interactions with sex. The effect of *ageC* and *bfmC* is modeled by a two-dimensional (nonlinear) effect because an interaction between both variables is very likely a priori. For a child of age three months a duration of three months of breastfeeding is quite normal whereas for a 24 months old child it is much less than recommended by health organizations.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

The literature on undernutrition expects that the age effect depends on the amount of breastfeeding. For children that have not been breastfed enough (according to the health organizations) the age effect is supposed to be below the effect for children with adequate duration of breastfeeding. It is clear that model choice by hand is very tedious for this complex model and an automated approach is required for simultaneous selection of relevant terms and the degree of smoothness of nonlinear functions.

The plan of the paper is as follows: The next section describes models with structured additive predictors mostly from a penalized least squares point of view. Section 3 presents algorithms for simultaneous selection of variables and smoothing parameters while Section 4 briefly discusses interval estimates both conditional and unconditional on the selected model. Section 5 presents results of extensive simulations to study the performance of the approach and to compare it to other methodology. Section 6 is devoted to the application on undernutrition in india. The final section summarizes the paper.

2. Models with structured additive predictor: A penalized least squares view

Suppose that observations $(y_i, \mathbf{z}_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, are given, where y_i is a continuous response variable, and $\mathbf{z}_i = (z_{i1}, \ldots, z_{iq})'$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ are vectors of covariates. For the variables in $\mathbf{z}$ possibly nonlinear effects are assumed, whereas the variables in $\mathbf{x}$ are modeled in the usual linear way. The components of $\mathbf{z}$ are not necessarily continuous covariates. A component may also indicate a time scale, a spatial index denoting the region or district a certain observations pertains to, or a unit- or cluster index denoting the unit (e.g. community) a certain observation pertains to. Moreover, the components of $\mathbf{z}$ may be two- or even three dimensional in order to model interactions between covariates. Summarizing, the vector $\mathbf{z}$ contains covariates of different type and dimension with possibly nonlinear effects. We assume an additive decomposition of the effects of z_{ij} (and x_{ij}) and obtain the model

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{i1}) + \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_i. \tag{1}$$

Here, $f_1 - f_q$ are nonlinear functions of the covariates z_{ij} and $\mathbf{x}_i'\boldsymbol{\gamma}$ is the usual linear part of the model. The errors ε_i are assumed to be mutually independent Gaussian with mean 0 and variance σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

Throughout the paper the nonlinear functions f_j are modeled by a basis functions approach, i.e. a particular nonlinear function f is approximated by a linear combination of basis functions

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z).$$

The B_k are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is a vector of unknown regression coefficients to be estimated. To ensure enough flexibility, typically a large number of basis functions is defined. To avoid overfitting a roughness penalty on the regression coefficients is additionally specified. We use quadratic penalties of the form $\boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta}$ where $\mathbf{K}(\boldsymbol{\lambda})$ is a penalty matrix. The penalty depends on one or multiple smoothing parameters $\boldsymbol{\lambda}$ that govern the amount of smoothness imposed on the function f. In the case of a single smoothing parameter, i.e. $\boldsymbol{\lambda}$ is a scalar rather than a vector, we will also write $\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta}$. From a Bayesian point of view the quadratic penalty $\boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta}$ corresponds to a Gaussian (improper) prior for the regressions coefficients $\boldsymbol{\beta}$, i.e.

$$p(\boldsymbol{\beta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta}\right).$$

The choice of basis functions $B_1, \ldots, B_K$ and penalty matrix $\mathbf{K}(\boldsymbol{\lambda})$ depends on our prior assumptions about the smoothness of f as well as the type and dimension of z. We will give specific examples below. Defining the $n \times K$ design matrix $\mathbf{Z}$ with elements $\mathbf{Z}[i, k] = B_k(z_i)$ the vector $\mathbf{f} = (f(z_1), \ldots, f(z_n))'$ of function evaluations can be written in matrix notation as $\mathbf{f} = \mathbf{Z}\boldsymbol{\beta}$. Accordingly, for model (1) we obtain

$$\mathbf{y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \cdots + \mathbf{Z}_q\boldsymbol{\beta}_q + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the design matrix for linear effects, $\boldsymbol{\gamma}$ is the vector of regression coefficients for linear effects, and $\mathbf{y}$ and $\boldsymbol{\varepsilon}$ are the vectors of observations and errors. In the next subsections we will give specific examples for modeling the unknown functions f_j or in other words for the choice of basis functions and penalty matrices. We start with modeling the effect of continuous covariates using splines.

2.1. Continuous covariates

2.1.1. P(enalized)-splines

Suppose first that a particular component z of $\mathbf{z}$ is univariate and continuous. There is a considerable amount of literature on basis functions approaches in combination with a (quadratic) roughness penalty for continuous covariates. Prominent examples are smoothing splines, P(enalized)-splines, Thin-plate splines and radial basis functions, see the monographs Hastie and Tibshirani (1990), Hastie et al. (2003), Fahrmeir and Tutz (2001), Wood (2006b) and Fahrmeir et al. (2007) and

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

the references cited therein. Here, we apply the P-splines approach introduced by Eilers and Marx (1996). The approach assumes that the unknown functions can be approximated by a polynomial spline of degree l and with equally spaced knots

$$z_{\min} = \zeta_0 < \zeta_1 < \cdots < \zeta_{m-1} < \zeta_m = z_{\max}$$

over the domain of z. The spline can be written in terms of a linear combination of $K = m + l$ B-spline basis functions (De Boor, 2001). The columns of the design matrix $\mathbf{Z}$ are given by the B-spline basis functions evaluated at the observations z_i. To overcome the well known difficulties involved with regression splines, Eilers and Marx (1996) suggest a relatively large number of knots (usually between 20 and 40) to ensure enough flexibility, and to introduce a roughness penalty on adjacent regression coefficients based on squared r-th order differences, i.e.

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \lambda \sum_{k=r+1}^{K} (\Delta^r \beta_k)^2.$$

The penalty matrix is given by $\mathbf{K}(\lambda) = \lambda \mathbf{D}_r' \mathbf{D}_r$ where $\mathbf{D}_r$ is a r-th order difference matrix. Typically, second or third order differences are used. The limiting behavior $\lambda \to \infty$ depends both on the order of the spline and the order of the penalty. If the order of the spline is equal to or higher than the order of the penalty, which is typically the case, then a polynomial fit of degree $r - 1$ is obtained in the limit.

The approach can be extended to impose monotonicity or more general shape constraints. We will not need P-splines with shape constraints in the applications of this paper. We therefore omit the details and refer to Bollaerts et al. (2006). However, our software is fully capable of fitting P-Splines with shape constraints.

2.1.2. Tensor product P-splines

Assume now that z is two-dimensional, i.e. $z = \left(z^{(1)}, z^{(2)}\right)'$ with continuous components $z^{(1)}$ and $z^{(2)}$. The aim is to extend the univariate P-spline from the preceding section to two dimensions. A common approach is to approximate the unknown surface $f(z)$ by the tensor product of one dimensional B-splines, i.e.

$$f\left(z^{(1)}, z^{(2)}\right) = \sum_{k=1}^{K_1} \sum_{s=1}^{K_2} \beta_{ks} B_{1,k}(z^{(1)}) B_{2,s}(z^{(2)}), \tag{2}$$

where $B_{11}, \ldots, B_{1K_1}$ are the basis functions in $z^{(1)}$ direction and $B_{21}, \ldots, B_{2K_2}$ in $z^{(2)}$ direction. See Gu (2002), Eilers and Marx (2003), Wood (2003) and Lang and Brezger (2004) for some recent references to surface fitting based on tensor products. The $n \times K = n \times K_1 K_2$ design matrix $\mathbf{Z}$ now consists of products of basis functions.

Several alternatives are available for the penalty matrix $\mathbf{K}(\boldsymbol{\lambda})$:

(a) *Penalty based on first differences:* The two-dimensional generalization of a penalty based on first differences is given by combining row- and column wise quadratic differences

$$\sum_{k=2}^{K_1} \sum_{s=1}^{K_2} (\beta_{ks} - \beta_{k-1,s})^2 = \boldsymbol{\beta}'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)\boldsymbol{\beta}$$

$$\sum_{k=1}^{K_1} \sum_{s=2}^{K_2} (\beta_{ks} - \beta_{k,s-1})^2 = \boldsymbol{\beta}'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})\boldsymbol{\beta}$$

to the penalty

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda \left[(\mathbf{I}_{K_2} \otimes \mathbf{D}_1)'(\mathbf{I}_{K_2} \otimes \mathbf{D}_1) + (\mathbf{D}_2 \otimes \mathbf{I}_{K_1})'(\mathbf{D}_2 \otimes \mathbf{I}_{K_1})\right] \boldsymbol{\beta}.$$

Another way of expressing the penalty is given by

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda \left[\mathbf{I}_{K_2} \otimes \mathbf{K}_1 + \mathbf{K}_2 \otimes \mathbf{I}_{K_1}\right] \boldsymbol{\beta}, \tag{3}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are the respective one dimensional penalty matrices. In the limit $\lambda \to \infty$ a constant fit is obtained.

(b) *Penalty based on second differences:* In a similar way two-dimensional penalties based on higher order differences are constructed. A second order difference penalty is obtained if $\mathbf{K}_1$ and $\mathbf{K}_2$ in (3) correspond to penalty matrices based on second rather than first differences. Similar to one dimensional P-splines the limit $\lambda \to \infty$ results in a linear fit, i.e.

$$f\left(z^{(1)}, z^{(2)}\right) = c_0 + c_1 z^{(1)} + c_2 z^{(2)} + c_3 z^{(1)} z^{(2)}.$$

(c) *Anisotropic penalty:* The two penalties considered so far are not capable of different penalization in $z^{(1)}$ and $z^{(2)}$ direction, respectively. Following Eilers and Marx (2003) anisotropic penalties are obtained by assuming separate smoothing parameters $\lambda^{(1)}$ and $\lambda^{(2)}$ in $z^{(1)}$ and $z^{(2)}$ direction. The penalty is then given by

$$\boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta} = \boldsymbol{\beta}' \left[\lambda^{(1)} \mathbf{I}_{K_2} \otimes \mathbf{K}_1 + \lambda^{(2)} \mathbf{K}_2 \otimes \mathbf{I}_{K_1}\right] \boldsymbol{\beta}. \tag{4}$$

The resulting fit in the limit $\lambda^{(1)} \to \infty$ and $\lambda^{(2)} \to \infty$ depends on the penalty used to construct $\mathbf{K}_1$ and $\mathbf{K}_2$. If $\mathbf{K}_1$ and $\mathbf{K}_2$ correspond to a first order difference penalty a constant fit is obtained in the limit. Second order difference penalties result in a linear fit for $f\left(z^{(1)}, z^{(2)}\right)$.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

(d) *Penalties with main effects in the limit:* Sometimes, it is desirable to decompose the effect of the two covariates $z^{(1)}$ and $z^{(2)}$ into two main effects modeled by one dimensional functions and a two-dimensional interaction effect, i.e.

$$f\left(z^{(1)}, z^{(2)}\right) = f_1\left(z^{(1)}\right) + f_2\left(z^{(2)}\right) + f_{1|2}\left(z^{(1)}, z^{(2)}\right). \tag{5}$$

See Gu (2002) for extensive discussion of such ANOVA-type decompositions. Usually a two-dimensional surface smoother, together with two additional one dimensional P-splines (or other smoothers) are estimated. However, the approach has some drawbacks. First, additional identifiability constraints have to be imposed on the three functions. Second, in cases where the overall hat matrix is not available (as in the estimation approach of this paper), the sum of the degrees of freedom of the three smoothers is not usable as an approximation to the degrees of freedom of the overall effect, see the next section. We therefore follow a different approach. We specify a two-dimensional surface based on tensor product P-splines and compute the decomposition of the resulting surface into main effects and the interaction effect *after* estimation. Moreover, we specify a penalty that allows for a main effects only model as a special case. This allows one to discriminate between a simple main effects model and a more complicated two way interactions model. A penalty that guarantees a main effects model in the limit is defined by the Kronecker product of two penalty matrices for one dimensional P-splines based on first order differences, i.e.

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \boldsymbol{\beta}'\lambda\mathbf{K}_1 \otimes \mathbf{K}_2\boldsymbol{\beta}. \tag{6}$$

The drawback of this penalty is that the limit $\lambda \to \infty$ yields *unpenalized* main effects, i.e. quite complex and wiggly functions. We therefore propose to use a modified penalty which is effectively a combination of the two penalties (4) and (6). More specifically we define

$$\boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta} = \boldsymbol{\beta}'\left[\frac{\lambda^{(1)}}{K_1}\mathbf{I}_{K_2} \otimes \mathbf{K}_1 + \frac{\lambda^{(2)}}{K_2}\mathbf{K}_2 \otimes \mathbf{I}_{K_1} + \lambda^{(3)}\tilde{\mathbf{K}}_1 \otimes \tilde{\mathbf{K}}_2\right]\boldsymbol{\beta}, \tag{7}$$

where $\mathbf{K}_1$ and $\mathbf{K}_2$ are penalty matrices corresponding to one dimensional P-splines based on first or second order differences. The matrices $\tilde{\mathbf{K}}_1$ and $\tilde{\mathbf{K}}_2$ are penalty matrices of P-splines based on first order differences. This penalty has the following nice properties (see Appendix A of Belitz (2007) for a derivation):

- The limit $\lambda^{(3)} \to \infty$ results in a mere main effects model. The main effects are one dimensional P-splines with smoothing parameters $\lambda^{(1)}$ and $\lambda^{(2)}$.
- The limit $\lambda^{(3)} \to 0$ yields the anisotropic penalty (4) as a special case.
- The limit $\lambda^{(1)} \to 0$ and $\lambda^{(2)} \to 0$ yields the Kronecker product penalty (6) as a special case.
- The limit $\lambda^{(1)} \to \infty$, $\lambda^{(2)} \to \infty$ and $\lambda^{(3)} \to \infty$ results in a main effects model with linear or constant main effects depending on the difference order used to construct $\mathbf{K}_1$ and $\mathbf{K}_2$.

It is important that the penalty matrices $\tilde{\mathbf{K}}_1$ and $\tilde{\mathbf{K}}_2$ used in the third Kronecker product in (7) are based on first order differences. For higher order differences, for instance second order differences, the limit for $\lambda^{(3)} \to \infty$ results in a model with main effects *and* additional complex interactions (varying coefficient terms), i.e. the nice property of a main effects model in the limit is lost.

After estimation, the main effects $f_1\left(z^{(1)}\right)$, $f_2\left(z^{(2)}\right)$ and the interaction effect $f_{1|2}\left(z^{(1)}, z^{(2)}\right)$ are computed from $f\left(z^{(1)}, z^{(2)}\right)$ as follows: We first compute the interaction surface by

$$f_{1|2}\left(z^{(1)}, z^{(2)}\right) = f\left(z^{(1)}, z^{(2)}\right) - \bar{f}_1\left(z^{(2)}\right) - \bar{f}_2\left(z^{(1)}\right) + \bar{f},$$

where

$$\bar{f}_1\left(z^{(2)}\right) = \frac{1}{\text{range}(z^{(1)})}\int f\left(z^{(1)}, z^{(2)}\right) \mathrm{d}z^{(1)}$$

$$\bar{f}_2\left(z^{(1)}\right) = \frac{1}{\text{range}(z^{(2)})}\int f\left(z^{(1)}, z^{(2)}\right) \mathrm{d}z^{(2)}$$

$$\bar{f} = \frac{1}{\text{range}(z^{(1)}) \cdot \text{range}(z^{(2)})}\iint f\left(z^{(1)}, z^{(2)}\right) \mathrm{d}z^{(1)}\mathrm{d}z^{(2)}$$

are the row wise, column wise and overall means of f, respectively. The next step is to extract the two main effects as

$$f_1\left(z^{(1)}\right) = \bar{f}_2\left(z^{(1)}\right) - \bar{f} \qquad f_2\left(z^{(2)}\right) = \bar{f}_1\left(z^{(2)}\right) - \bar{f}.$$

Finally, the intercept γ_0 is corrected by adding the overall mean $\bar{f}$ of f. The approach guarantees that the row wise, column wise and overall means of the interaction $f_{1|2}$ as well as the means of the two main effects are zero, see also Chen (1993) and Stone et al. (1997). Moreover, by inserting the tensor product representation of f into $\bar{f}_1\left(z^{(2)}\right)$, respectively $\bar{f}_2\left(z^{(1)}\right)$, it is easily shown that the two main effects are P-splines.

2.2. Spatial heterogeneity

In this subsection we assume that z represents the location a particular observation pertains to. The location is typically given in two ways. If exact locations are available, $z = (z^{(1)}, z^{(2)})'$ is two-dimensional, and the components $z^{(1)}$ and $z^{(2)}$

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

correspond to the coordinates of the location. In this case the spatial effect $f(z^{(1)}, z^{(2)})$, could be modeled by two-dimensional surface estimators as described in the preceding section.

In many applications, however, exact locations are not available. Typically, a geographical map is available and $z \in \{1, \ldots, K\}$ is an index that denotes the region (e.g. district) an observation pertains to. For instance, the data on undernutrition in India contains information in which of the approximately 400 districts the mother and her child lives. A common approach is to assume $f(z) = \beta_z$, i.e. separate parameters $\beta_1, \ldots, \beta_K$ for each region are estimated. The $n \times K$ design matrix $\mathbf{Z}$ is an incidence matrix whose entry in the i-th row and k-th column is equal to one if observation i has been observed at location k and zero otherwise. To prevent overfitting, a penalty based on squared differences is defined that guarantees that parameters of neighboring regions are similar. Typically two regions are assumed to be neighbors if they share a common boundary although other neighborhood definitions are possible. The penalty is defined as

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \lambda \sum_{k=2}^{K} \sum_{s \in N(k), s<k} (\beta_k - \beta_s)^2, \tag{8}$$

where $N(k)$ denotes all sites that are neighbors of site k. The elements of the penalty matrix are given by

$$\mathbf{K}[s, r] = \lambda \begin{cases} -1 & k \neq s, k \sim s, \\ 0 & k \neq s, k \nsim s, \\ |N(k)| & k = s. \end{cases} \tag{9}$$

If we adopt a Bayesian point of view, the prior resulting from penalty matrix (9) is an example of an intrinsic Gaussian autoregression or a Gaussian Markov random field prior.

Depending on the prior belief on smoothness of the spatial effect, several alternatives to penalty (9) are available. If a very smooth effect is assumed, the two-dimensional smoothers discussed in the preceding section could be used as an alternative. Since exact locations are not available, the centroids of the regions could be used instead. In some situations a smooth spatial effect is not justified, because of local spatial heterogeneity. In this case, the assumption of spatial dependence of neighboring parameters is not meaningful. Instead, a simple ridge type penalty

$$\boldsymbol{\beta}'\mathbf{K}(\lambda)\boldsymbol{\beta} = \lambda\boldsymbol{\beta}'\boldsymbol{\beta} = \lambda \sum_{k=1}^{K} \beta_k^2$$

with penalty matrix $\mathbf{K}(\lambda) = \lambda\mathbf{I}$ may be defined. This penalty does not assume any spatial dependence, but prevents highly variable estimates induced by small samples, for some regions or sites.

2.3. Unit- or cluster specific heterogeneity

There is a vast literature on modeling unit- or cluster specific heterogeneity, see e.g. Verbeke and Molenberghs (2000). An important special case arises for longitudinal data, where individuals are repeatedly observed over time (Diggle et al., 2002). Typically, unit- or cluster specific random effects are introduced, to account for heterogeneity. In its simplest form, a random intercept β_z with $\beta_z \sim N(0, \tau^2)$ is introduced. Here, $z \in \{1, \ldots, K\}$ is an index variable that denotes the cluster a particular observation pertains to. This is equivalent to a penalized least squares approach with function $f(z) = \beta_z$, penalty matrix $\mathbf{I}$ and smoothing parameter $\lambda = \sigma^2/\tau^2$. The $n \times K$ design matrix $\mathbf{Z}$ is a 0/1 incidence matrix whose entry in the i-th row and k-th column is equal to one, if observation i belongs to the k-th cluster and zero otherwise. Random slopes could be treated in the same way, see the next subsection.

Note that more than one random intercept with respect to different cluster variables are possible. In many cases, there exists a hierarchical ordering of clusters. Models with such hierarchical clusters are also called multilevel models, see for instance Skrondal and Rabe-Hesketh (2004).

2.4. Varying coefficients

In our application on undernutrition in India, an important scientific question is concerned with gender related differences. Hence a modeling framework is required that allows for sex specific covariate effects. More generally, suppose that the effect of a continuous covariate $z^{(2)}$ is assumed to vary with respect to a categorical covariate $z^{(1)}$. For notational convenience, we restrict the discussion to binary covariates $z^{(1)}$. The generalization to (multi)categorical covariates is straightforward. The interaction between $z^{(2)}$ and $z^{(1)}$ can be modeled by a predictor of the form

$$\eta = \cdots + f_1(z^{(2)}) + g(z^{(2)})z^{(1)} + \cdots,$$

where f_1 and g are smooth functions (modeled by P-splines). The interpretation of the two functions f_1 and g depends on the coding of the binary variable $z^{(1)}$. If $z^{(1)}$ is in dummy-coding, the function f_1 corresponds to the effect of $z^{(2)}$ subject to $z^{(1)} = 0$, and g is the difference effect for observations with $z^{(1)} = 1$. If $z^{(1)}$ is in effect-coding, the function f_1 can be interpreted as an average effect of $z^{(2)}$, and g respectively $-g$, is now the deviation from f_1 for $z^{(1)} = 1$ and $z^{(1)} = -1$. It turns out that the coding of $z^{(2)}$ is not only important for interpretation, but also crucial for inference. Since the estimation

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

described in the next section is based on an iterative backfitting-type procedure, dependence between f_1 and g should be minimized, to avoid convergence problems. Hence, effect coding for $z^{(2)}$ is an effective yet simple device, to avoid convergence problems.

Models with interaction effects of the form $g(z^{(2)})z^{(1)}$ are known as varying coefficient models (Hastie and Tibshirani, 1993), because the effect of $z^{(1)}$ varies smoothly with respect to the continuous covariate $z^{(2)}$. Covariate $z^{(2)}$ is called the effect modifier of $z^{(1)}$. The approach can be easily extended to a two-dimensional effect modifier with components $z^{(2)}$ and $z^{(3)}$. The interaction effect is then given by $g(z^{(2)}, z^{(3)})z^{(1)}$ where $g(z^{(2)}, z^{(3)})$ is a two-dimensional surface, which is modeled by the tensor product P-splines discussed in Section 2.1.2. For instance, for the India data we model the sex specific effect of the age of the child (*ageC*) and the duration of breastfeeding (*bfmC*) by the interaction effect $g(ageC, bfmC)sex$. Another modification arises if the effect modifier is the location, either given as the coordinates or as a spatial index. In this case, we have a space varying effect of $z^{(1)}$. Models of this kind are also known as geographically weighted regression, see Fotheringham et al. (2002). A final modification is obtained for a unit- or cluster index as effect modifier. The effect of $z^{(1)}$ is now assumed to be unit- or cluster specific and typically referred to as a random slope.

Independent of the specific type of the effect modifier, the interaction term $g\left(z^{(2)}\right)z^{(1)}$ (or $g\left(z^{(2)}, z^{(3)}\right)z^{(1)}$) can be cast into our general framework by defining

$$f\left(z^{(1)}, z^{(2)}\right) = g\left(z^{(2)}\right)z^{(1)} \quad \text{or} \quad f\left(z^{(1)}, z^{(2)}, z^{(3)}\right) = g\left(z^{(2)}, z^{(3)}\right)z^{(1)}. \tag{10}$$

The overall design matrix $\mathbf{Z}$ is given by $\mathrm{diag}(z_1^{(1)}, \ldots, z_n^{(1)})\mathbf{Z}^{(1)}$ where $\mathbf{Z}^{(1)}$ is the usual design matrix for P-Splines, tensor product P-splines, spatial- or cluster specific effects.

3. Simultaneous selection of variables and smoothing parameters

A main building block of our algorithms are smoothers of the form

$$S(\mathbf{y}, \boldsymbol{\lambda}) = \mathbf{Z}\hat{\boldsymbol{\beta}} \qquad \hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z} + \mathbf{K}(\boldsymbol{\lambda}))^{-1}\mathbf{Z}'\mathbf{y},$$

where $\mathbf{Z}$ and $\mathbf{K}(\boldsymbol{\lambda})$ are design and penalty matrices, corresponding to the smooth covariate effects discussed in the preceding section. The cross product matrices $\mathbf{Z}'\mathbf{Z} + \mathbf{K}(\boldsymbol{\lambda})$ are band matrices, or can be transformed into a matrix with band structure. For the spatial penalty (8) the cross product matrix is not a priori a band matrix, but sparse. It can be transformed into a band matrix (with differing band size in every row), by reordering the regions using the reverse Cuthill Mc-Kee algorithm (see George and Liu (1981) p. 58 ff). Hence, numerical evaluation of the smoothers is done by using matrix operations for sparse matrices, in particular Cholesky decompositions. In our implementation, we use the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981).

For fixed smoothing parameter(s), $\hat{\boldsymbol{\beta}}$ is the minimizer of the penalized least squares criterion

$$PLS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{Z}\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{K}(\boldsymbol{\lambda})\boldsymbol{\beta}.$$

Consecutively applying smoothers S_j corresponding to the j-th function f_j in (1) to the current partial residual, reveals the well known backfitting algorithm to minimize the overall PLS-criterion

$$PLS = \left(\mathbf{y} - \sum_{j=1}^{q}\mathbf{Z}_j\boldsymbol{\beta}_j - \mathbf{X}\boldsymbol{\gamma}\right)'\left(\mathbf{y} - \sum_{j=1}^{q}\mathbf{Z}_j\boldsymbol{\beta}_j - \mathbf{X}\boldsymbol{\gamma}\right) + \sum_{j=1}^{q}\boldsymbol{\beta}_j'\mathbf{K}_j(\boldsymbol{\lambda}_j)\boldsymbol{\beta}_j.$$

The complexity of the fit may be determined by the equivalent degrees of freedom df, as a measure of the effective number of parameters. In concordance with linear models, the degrees of freedom of the fit are defined as

$$df = \mathrm{trace}(\mathbf{H}),$$

where $\mathbf{H}$ is the prediction matrix that projects the observations $\mathbf{y}$ on their fitted values $\hat{\mathbf{y}}$, i.e. $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In complex models with many effects, the trace of $\mathbf{H}$ is difficult, and computationally intensive to compute. Therefore df is typically approximated by the sum of the degrees of freedom of individual smoothers, i.e.

$$df = \sum_{j=1}^{q} df_j + p,$$

where p is the number of linear effects in the model and df_j is in most cases computed as

$$df_j = \mathrm{trace}(\mathbf{Z}_j(\mathbf{Z}_j'\mathbf{Z}_j + \mathbf{K}_j(\boldsymbol{\lambda}_j))^{-1}\mathbf{Z}_j') - 1. \tag{11}$$

The substraction of one from the trace is necessary, because terms are usually centered around zero to guarantee identifiability, and as a consequence, one degree of freedom is lost. In some situations, varying coefficient terms are identifiable, and substraction of one is not required. To give an example, the varying coefficient terms in $\eta = \cdots + g_1(z_1)z + g_2(z_2)z + \gamma_0 + \gamma_1 z + \cdots$ are not identified, whereas in $\eta = \cdots + g_1(z_1)z + \gamma_0 + \cdots$, the varying coefficient term is identifiable. In the first case, centering and substraction of one is necessary, in the second case, it is not.

The approximation provides good results, provided that correlations among individual terms, and correlations between nonlinear terms and linear effects are moderate. In addition, the number of observations should be considerably larger than the estimated number of parameters. Two remarks are in order:

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

- *Two-dimensional P-splines:* As already mentioned, a two-dimensional effect of $z^{(1)}$ and $z^{(2)}$ is frequently specified, together with corresponding main effects, i.e. the predictor is of the form (5). In this case, the approximation of the degrees of freedom by the sum of individual degrees of freedom, is clearly not valid. Instead, the overall projection matrix of the three smoothers must be used to compute the degrees of freedom. In this paper, we therefore follow the alternative approach already proposed in Section 2.1.2, i.e. the penalty matrix is constructed such that a main effects model based on P-splines is a special case, see penalty (7). Then the approximation (11) is applicable. The approach is also much less computationally intensive, and gives similar results in terms of goodness of fit, than the more conventional approach, see also the results of the simulation study in Section 5.
- *Unit- or cluster specific effect:* Consider for simplicity the predictor $\eta = \gamma_0 + f(z)$, where $f(z)$ is a cluster specific effect with respect to cluster variable z. Particular for small smoothing parameters, there is increasing linear dependence with the intercept. We therefore base computation of the degrees of freedom on an augmented design matrix, that contains an additional column of ones, i.e. $\mathbf{Z}_j$ in (11) is replaced by $(\mathbf{Z}_j\mathbf{1})$.

Our variable selection procedure described below aims at minimizing a goodness of fit criterion. The following options are available:

- *Test- and validation sample*
  Provided that enough data are available, the best strategy is to divide the data into a test- and validation sample. The test data set is used to estimate the parameters of the models. The fit of different models is assessed via the validation data set. In the case of a continuous response, typically the mean squared prediction error is minimized.
- *AIC*, AIC_c, *BIC*
  In such cases where data are sparse, an estimate for the prediction error is based on goodness of fit criteria that penalize the complexity of models. A general form for a wide range of criteria is given by $C = n\log(\hat{\sigma}^2) + penalty$, where the penalty depends on the degrees of freedom (df) of the model. The most widely used criteria are the *AIC* ($penalty = 2df$) and the *BIC* ($penalty = \log(n)df$). A bias corrected version AIC_c of *AIC* is widely used with regression models, see Hurvich et al. (1998). Here, the penalty is given by $penalty = 2df + \frac{2\cdot df(df+1)}{n-df-1}$. Experience from extensive simulations suggests using AIC_c for the models considered in this article.
- *Cross validation*
  Cross validation mimics the division into test- and validation samples. The data are split into K parts (typically $K = 5$ or $K = 10$), of roughly equal size. One part is used to estimate prediction errors, and the other $K - 1$ parts are used to fit the models. This is done for every part, and the estimated prediction errors are added. K-fold cross validation is in most cases slightly more time consuming than using *AIC* or *BIC*, but an approximation of the degrees of freedom is not required. Hence, cross validation is recommended if the approximation to the degrees of freedom is likely to be erroneous e.g. because of strong correlations among covariates.

We are now ready to describe our approach for simultaneous selection of variables and smoothing parameters. We first exclude smoothers with multiple smoothing parameters, as is particularly the case for the two-dimensional surface estimator with penalty (7). The basic algorithm works as follows:

Algorithm 1. (1) **Initialization**
Define for every possible nonlinear term $f_j, j = 1, \ldots, q$, a discrete number M_j of decreasing smoothing parameters $\lambda_{j1} > \cdots > \lambda_{jM_j}$. The smoothing parameters are chosen such that they correspond to certain equivalent degrees of freedom. For example, for one dimensional P-splines with 20 inner knots (corresponding to 22 basis functions) and second order difference penalty we define $\lambda_{j1} = \infty$ corresponding to $df_{j1} = 1$, i.e. a linear fit, and the other smoothing parameters λ_{js}, $s = 2, \ldots, 22$, are defined such that $df_{js} = s$. For model terms with a large number of basis functions typically 30–40, different smoothing parameters are defined.

(2) **Start model**
Choose a start model with current predictor

$$\hat{\boldsymbol{\eta}} = \hat{\mathbf{f}}_1 + \cdots + \hat{\mathbf{f}}_q.$$

where $\hat{\mathbf{f}}_j$ is the vector of function evaluations at the observations. For example, set $\hat{\mathbf{f}}_j \equiv \mathbf{0}, j = 1, \ldots, q$ which corresponds to the empty model. Choose a goodness of fit criteria C.

(3) **Iteration**
(a) For $j = 1, \ldots, q$:
For $m = 0, \ldots, M_j$:
Compute the fits

$$\hat{\mathbf{f}}_{jm} := \begin{cases} \mathbf{0} & m = 0 \\ \mathbf{S}_j(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[j]}, \lambda_{jm}) & m = 1, \ldots, M_j \end{cases}$$
$$= \begin{cases} \mathbf{0} & m = 0 \\ (\mathbf{Z}_j'\mathbf{Z}_j + \mathbf{K}(\lambda_{jm}))^{-1}\mathbf{Z}_j'(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[j]}) & m = 1, \ldots, M_j \end{cases}$$

and the corresponding predictors $\hat{\boldsymbol{\eta}}_{jm} := \hat{\boldsymbol{\eta}}_{[j]} + \hat{\mathbf{f}}_{jm}$. Here, $\hat{\boldsymbol{\eta}}_{[j]}$ is the current predictor with the j-th fit $\hat{\mathbf{f}}_j$ removed.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

Compute the updated estimate

$$\hat{\mathbf{f}}_j = \operatorname{argmin} C(\hat{\mathbf{f}}_{jm}),$$

i.e. among the fits $\hat{\mathbf{f}}_{jm}$ for the j-th component, choose the one that minimizes the goodness of fit criteria C.

(b) The linear effects part $\mathbf{x}'\boldsymbol{\gamma}$ typically consists of the intercept γ_0 and dummy variables for the categorical covariates. For the moment, suppose that $\mathbf{x}$ contains dummies representing only one categorical variable. Then we compare the fits $\hat{\gamma}_0 = \overline{y - \eta_{[lin]}}$, $\gamma_1 = 0, \ldots, \gamma_q = 0$ (covariate removed from the model) and $\hat{\boldsymbol{\gamma}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\eta}}_{[lin]})$ where $\hat{\boldsymbol{\eta}}_{[lin]}$ is the current predictor with the linear effects removed and $\overline{y - \eta_{[lin]}}$ is the mean of the elements of the partial residual vector $\mathbf{y} - \hat{\boldsymbol{\eta}}_{[lin]}$. If more than one categorical covariate is available the procedure is repeated for every variable.

(4) **Termination**

The iteration cycle in 3. is repeated until the model, regression and smoothing parameters do not change anymore.

If a two-dimensional surface with penalty (7) is specified, the basic Algorithm 1 must be adapted. The following algorithm describes modifications for a two-dimensional surface $f\left(z^{(1)}, z^{(2)}\right)$, which may be decomposed into two main effects f_1 and f_2, and an interaction effect $f_{1|2}$ as in (5):

Algorithm 2. (1) **Initialization**

Define for the three components of $\boldsymbol{\lambda} = (\lambda^{(1)}, \lambda^{(2)}, \lambda^{(3)})'$ ordered lists of smoothing parameters

$$\lambda_1^{(j)} > \cdots > \lambda_{M_j}^{(j)} \quad j = 1, 2, 3.$$

We first choose the list of smoothing parameters for the first and second component $\lambda^{(1)}$ and $\lambda^{(2)}$. This is done by temporarily assuming $\lambda^{(3)} = \infty$, i.e. a main effects model. Then the smoothing parameters $\lambda_1^{(j)}, \ldots, \lambda_{M_j}^{(j)}, j = 1, 2$ are defined in the same way as for univariate P-Splines, i.e. they are chosen in accordance to certain equivalent degrees of freedom. The next step is to define a list of smoothing parameters for the third component $\lambda^{(3)}$. For that purpose we temporarily assume $\lambda^{(1)} = \lambda^{(2)} = 0$, which corresponds to penalty (6). We then assign again the smoothing parameters in accordance to certain degrees of freedom. The largest smoothing parameter in the list is set to infinity, i.e. $\lambda_{M_3}^{(3)} = \infty$, to guarantee that a main effects model is included as a special case.

(2) **Start model**

We usually start with a main effects model, i.e. for $\lambda^{(3)}$ the largest smoothing parameter $\lambda_{M_1}^{(3)} = \infty$ in the list is set.

(3) **Iteration**

We distinguish the two cases (a) $\lambda^{(3)} = \infty$ and (b) $\lambda^{(3)} \neq \infty$ for the current value of the third component of $\boldsymbol{\lambda}$. Case (a) corresponds to a main effects model, i.e. $f_{1|2} \equiv 0$ and f_1 and f_2 are one dimensional P-splines in (5).

(a) *Current value of* $\lambda^{(3)}$ *is infinity (main effects model)*

· *Update* $\lambda^{(1)}$: Compute the P-spline fits

$$\hat{\mathbf{f}}_{1m} := \begin{cases} \mathbf{0} & m = 0 \\ \left(\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{K}_1\left(\lambda_m^{(1)}\right)\right)^{-1}\mathbf{Z}_1'(\mathbf{y} - \hat{\boldsymbol{\eta}}) & m = 1, \ldots, M_1 \end{cases}$$

for the main effect f_1 in (5). The matrices $\mathbf{Z}_1$ and $\mathbf{K}_1$ correspond to the design matrix and penalty matrix of a one dimensional P-spline. The corresponding predictors to $\hat{\mathbf{f}}_{1m}$ are computed as $\hat{\boldsymbol{\eta}}_m := \hat{\boldsymbol{\eta}} + \hat{\mathbf{f}}_{1m}$, where $\hat{\boldsymbol{\eta}}$ is the current predictor with the first main effect removed. Finally, the updated estimate is computed as

$$\hat{\mathbf{f}}_1 = \operatorname{argmin} C(\hat{\mathbf{f}}_{1m}).$$

· *Update* $\lambda^{(2)}$: This is done in complete analogy to the previous step.

· *Update* $\lambda^{(3)}$: The third component is updated only if the main effects are not completely removed, i.e. if $\hat{\mathbf{f}}_1 \neq \mathbf{0}$ or $\hat{\mathbf{f}}_2 \neq \mathbf{0}$. In this case, we fix the first and second component of $\boldsymbol{\lambda}$ at its current values and compute the fits

$$\hat{\mathbf{f}}_m := \left(\mathbf{Z}'\mathbf{Z} + \mathbf{K}\left(\lambda^{(1)}, \lambda^{(2)}, \lambda_m^{(3)}\right)\right)^{-1}\mathbf{Z}'(\mathbf{y} - \hat{\boldsymbol{\eta}}) \quad m = 1, \ldots, M_3$$

for the surface f. The corresponding predictors are given by $\hat{\boldsymbol{\eta}}_m := \hat{\boldsymbol{\eta}} + \hat{\mathbf{f}}_m$, with $\hat{\boldsymbol{\eta}}$ being the current predictor with the surface removed. Finally compute the updated estimate

$$\hat{\mathbf{f}} = \operatorname{argmin} C(\hat{\mathbf{f}}_m).$$

(b) $\lambda^{(3)} \neq \infty$ (*model with interaction*)

The main difference to the previous case is that the main effects can not be removed, because we do not allow for a surface f with interaction, but without main effects. For $j = 1, 2, 3$ the j-th component $\lambda^{(j)}$ of $\boldsymbol{\lambda}$ is updated by fixing the two remaining components, and computing the fits

$$\hat{\mathbf{f}}_m := \left(\mathbf{Z}'\mathbf{Z} + \mathbf{K}\left(\ldots, \lambda_m^{(j)}, \ldots\right)\right)^{-1}(\mathbf{y} - \hat{\boldsymbol{\eta}}) \quad m = 1, \ldots, M_j$$

for the surface f. The updated estimates are again given by

$$\hat{\mathbf{f}} = \operatorname{argmin} C(\hat{\mathbf{f}}_m).$$

We conclude this section with a few remarks on the two algorithms:

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

(1) *Avoid backfitting*

When updating the function estimates $\hat{\mathbf{f}}_j$, the other terms in the model are *not* re-estimated as in a backfitting procedure. However, the algorithm automatically collapses to backfitting, as soon as the variables and smoothing parameters included in the model no longer change. Avoiding backfitting in Step (2) dramatically reduces computing time, without substantial loss of estimation accuracy. In the majority of cases, the selected models with backfitting included in step 2 are identical to the models selected with our algorithm. In a few cases, the fit of our algorithm is slightly worse, but estimated functions are still visually indistinguishable.

(2) *Increasing values for the goodness of fit criterion*

The price for avoiding backfitting, is that sometimes the value of the goodness of fit criteria slightly increases after one of the cycles in step 3a of Algorithm 1. This might happen because meanwhile other functions have been adjusted. If backfitting could be included after every cycle in step 3a, an increase in the goodness of fit criteria would be impossible. It is therefore necessary to stop the algorithm only if the model selection has finished, and the algorithm has already collapsed into backfitting.

(3) *Connection to boosting*

Taking two modifications our (basis) algorithm turns into a boosting approach. For a description of boosting see e.g. Tutz and Binder (2007), who discuss boosting for ridge regression, which is conceptually similar to the methods in this paper. Boosting is obtained if we compute in every iteration only one fit for every nonlinear term. The fits must be based on comparably large λ_j's in order to guarantee that the resulting smoothers are "weak learners". As a second modification, the smoothers are applied to the current residuals $\mathbf{y} - \hat{\boldsymbol{\eta}}$ rather than to $\mathbf{y} - \hat{\boldsymbol{\eta}}_{[j]}$.

(4) *Large linear models*

A referee suggested to express our models as large linear models as successfully proposed in Marx and Eilers (1998) for GAM's and Eilers and Marx (2002) for models with additive smooth structures. Then standard methodology for variable selection in linear models as e.g. described in Miller (2002), could be used (with modifications). In principle, this approach is possible with two advantages. First, possible convergence problems due to highly correlated terms can be avoided. Second, an approximation to the degrees of freedom of the fit required for computing some of the goodness of fit criteria is not necessary, as a trace of the smoother matrix is available. However, the approach is not feasible, or at least very time consuming for the complex models this paper is about. The difference to the situation in Marx and Eilers (1998) and Eilers and Marx (2002) is that our models may contain many more parameters, because spatial effects and/or cluster specific random effects as described in Sections 2.2 and 2.3 are typically involved. For instance, the full model for the data on undernutrition in India (see Section 6), with penalty (8) for the spatial effect, has roughly 1000 parameters. In some models with spatially varying effects, or many cluster specific effects, several thousand parameters are involved. The advantage of our adaptiv backfitting-type algorithm is that the problem can be divided into smaller pieces. Another important advantage is also that the sparsity of the cross product matrices $\mathbf{X}'\mathbf{X} + \mathbf{K}(\lambda)$ can be utilized to substantially decrease computing time.

(5) *Derivative free optimization*

One might ask why we have not attempted a derivative based approach for minimizing the goodness of fit criteria, as has been successfully proposed in Wood (2000), and also in Kauermann and Opsomer (2004). The reason is that our models may contain a large number of parameters, particularly because spatial and/or cluster specific effects with as many parameters as sites or clusters are frequently included. The derivative based algorithm of Wood (2000) is of cubic order in the number of parameters. Hence, computationally intense, if not infeasible, algorithms would result. However, we have experimented with other derivative free optimization as described e.g. in Brent (2003). While conceptually more demanding, substantial improvements in terms of computing time, or model fit have not been achieved. Therefore details are omitted.

(6) *Practical application*

Our algorithms aim at simultaneously estimating and selecting a model which fits well, in terms of a preselected goodness of fit measure. It is not guaranteed that the global minimum is found, but the solution of the algorithms are at least close. However, the selection process is subject to sampling error, and it is of course not guaranteed that the selected model is correct (if one believes that a correct model exists). Moreover, we do not believe in the one single model that best represents and summarizes the available data. We think of the selected best model more as a good starting point for further investigation.

(7) *Software*

The selection algorithms, as well as the methodology for obtaining confidence intervals described in the next section, are included in the public domain software package BayesX (version 2.0). Our software avoids most of the drawbacks inherent in many of the selection routines for linear models. For instance, it is possible to force a particular term into the model, or to include and remove all dummies of categorical variables simultaneously. For all covariates, the user has full control over the prior assumptions of the corresponding model term, by manually specifying the list of permitted smoothing parameters. Hence, mindless data dredging can be avoided.

4. Confidence intervals

We first describe the computation of pointwise confidence intervals conditional on the selected model. Frequentist confidence intervals are based on the assumption that the distribution of the estimated regression coefficients $\hat{\boldsymbol{\beta}}_j$ is

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

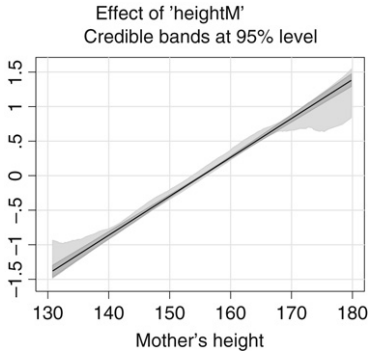

Fig. 1. estimated effect of the mother's height on the Z-score together with 95% conditional (dark-grey shaded area) and unconditional confidence intervals (light-grey shaded area).

approximately Gaussian. As has been noticed by various authors, frequentist confidence intervals are unsatisfactory, because of penalty induced bias; see Wood (2006c) for a recent discussion. Alternatively, we can resort to a Bayesian point of view, which implicitly considers the prior belief about smoothness induced through the penalty terms. Bayesian inference for models with structured additive predictors is extensively discussed in Lang and Brezger (2004) and Brezger and Lang (2006); see also Jullion and Lambert (2007). Inference can be based on MCMC methods, and pointwise (Bayesian) confidence intervals are simply obtained by computing the respective quantiles of simulated function evaluations. If credible intervals are conditional on the selected model, the smoothing parameters are fixed, and MCMC simulation is very fast, because simulation requires only to draw random numbers from multivariate normal distributions with fixed precision and covariance matrix. As will be shown through simulations, the coverage rates are surprisingly close to the nominal level, although model uncertainty is not taken into account.

Much more difficult and computer intensive are unconditional confidence intervals that explicitly take into account model uncertainty. We adapt an approach proposed by Wood (2006c) which works as follows:

(1) For $k = 1, \ldots, B$ use parametric bootstrapping to obtain B bootstrap response vectors $\mathbf{y}^{(k)}$.

(2) Apply the model selection algorithms of Section 3 on the bootstrap responses $\mathbf{y}^{(k)}$ resulting in estimates $\hat{\boldsymbol{\lambda}}_1^{(k)}, \ldots, \hat{\boldsymbol{\lambda}}_q^{(k)}$ for the smoothing parameters.

(3) Simulate for each of the B smoothing parameter combinations $\hat{\boldsymbol{\lambda}}_1^{(k)}, \ldots, \hat{\boldsymbol{\lambda}}_q^{(k)}$, and the smoothing parameters of the originally selected model random numbers from the posterior. Simulation is based conditional on the bootstrapped smoothing parameters and the observed data vector $\mathbf{y}$. Typically we set $B = 99$, resulting in 100 different smoothing parameter combinations. For every smoothing parameter combination, 10 random numbers are drawn from the posterior.

(4) Pointwise confidence intervals for the regression parameters $\boldsymbol{\gamma}$, and the nonlinear functions f_j are based on the whole sample of (typically 1000) simulated random numbers, and are obtained via the respective empirical quantiles.

Note that bootstrapping is only used for getting replicated samples of the smoothing parameters, whereas samples for regression coefficients are based on the original data. This avoids the typical bias of bootstrap confidence intervals in semiparametric regression models, see e.g. Kauermann et al. (2006). We have experimented with several proposals for bias reduction, see e.g. Politits and Romano (1994). These approaches are, however, less straightforward and proved to be not competitive to the conceptionally simple approach by Wood (2006c). Another advantage of the Wood approach is that a relatively small number of bootstrap replicates are sufficient, because bootstrapping is limited to the sampling of the smoothing parameters. This is particularly important because every bootstrap replicate requires a rerun of the relatively costly model selection algorithms of Section 3. More details and justifications for the bootstrap approach are given in Wood (2006c).

In the majority of applications, conditional and unconditional confidence intervals, are almost undistinguishable. In some exceptional cases, however, significant differences can be observed. As an example take Fig. 1, which plots the estimated effect of the mother's height on the Z-score in our data example on undernutrition in India, see also Section 6. The selected best model contains a linear effect of the mother's height. Since the unconditional confidence intervals are based on 100 (bootstrapped) smoothing parameter combinations in a significant number of cases, a sightly nonlinear effect for mother's height is selected. This increased uncertainty is reflected in the unconditional confidence intervals, which suggests that a slightly nonlinear effect is also reasonable. In this particular case, the unconditional confidence intervals are favorable to conditional confidence intervals, as they provide further valuable insight.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

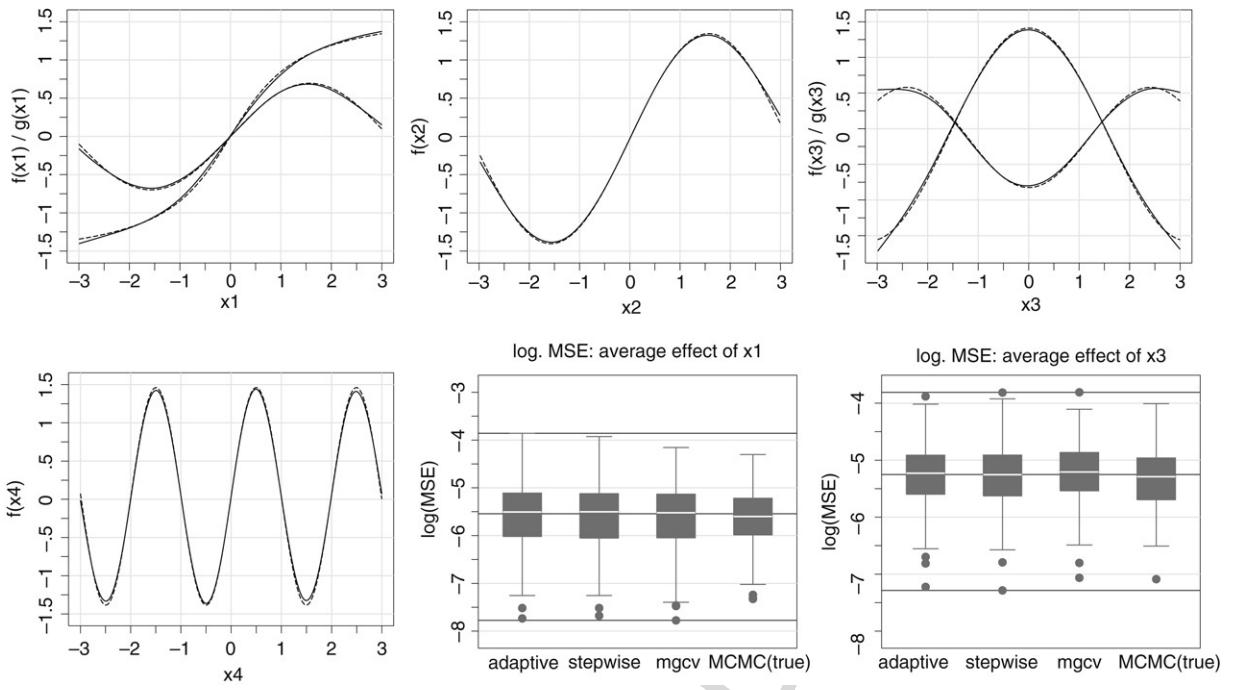

Fig. 2. Simulation study 1. Average estimated functions (dashed lines) based on adaptive search (top left to bottom left). For comparison the true functions are superimposed (solid lines). Boxplots of the empirical log MSEs for the various estimators and selected model terms.

5. Simulation results

5.1. Additive model with varying coefficient terms

A number of simulation experiments have been conducted to analyze the performance of our approach. As a representative example, we report results for a complex model that includes smooth additive effects, interactions between continuous and categorical covariates, spatial heterogeneity, and a spatially varying effect. The model imitates the kind of models estimated for the Indian data on undernutrition.

The simulation design is as follows: six continuous covariates $z_1, \ldots, z_6$ and a binary variable x have been simulated with values uniformly distributed in $[-3, 3]$ and $\{-1, 1\}$ respectively. The results given below are based on uncorrelated covariates. Experiments with moderately correlated covariates give virtually the same results, and are therefore omitted. The most general predictor with possibly nonlinear effects of the continuous covariates, possibly nonlinear interaction effects between the continuous covariates and the binary variable x, a spatial effect of a region indicator r and a spatially varying effect of x is given by

$$\eta = f_1(z_1) + g_1(z_1) \cdot x + \cdots + f_6(z_6) + g_6(z_6) \cdot x + f_7(r) + g_7(r) \cdot x + \gamma_0 + \gamma_1 \cdot x. \tag{12}$$

The simulated model contains nonlinear main effects $f_1 - f_4$ of the covariates $z_1 - z_4$, two interaction effects g_1 and g_3 between z_1 and x and z_3 and x (see Fig. 2), a spatial effect $f_7(r)$ and a spatially varying effect $g_7(r)$. The true predictor is therefore given by

$$\eta = f_1(z_1) + g_1(z_1) \cdot x + f_2(z_2) + f_3(z_3) + g_3(z_3) \cdot x + f_4(z_4) + f_7(r) + g_7(r) \cdot x + \gamma_0 + \gamma_1 x$$

and the remaining functions f_5, f_6 and $g_2, g_4 - g_6$ are identical to zero. All functions $f_j, j = 1, \ldots, 4, 7$, were chosen such that $\sigma_{f_j} = 1$, whereas the effect of the interaction functions $g_j, j = 1, 3, 7$, is weaker with $\sigma_{g_j} = 0.5$. We assumed a Gaussian model with a standard deviation of $\sigma_\varepsilon = 0.82$, leading to a moderate signal to noise ratio of $\sigma_\eta / \sigma_\varepsilon = 3$. We simulated 3 observations for each of the 309 regions, resulting in $n = 927$ observations for a single replication of the model. Overall, 250 replications have been simulated.

We compared the following estimators: our adaptive search based on Algorithm 1, the model selection tool of the mgcv package (version 1.3-22) in R, the stepwise algorithm described in Chambers and Hastie (1991) (using a fast implementation in BayesX, version 2.0) and a fully Bayesian approach (without model selection), as described in Lang and Brezger (2004). The Bayesian approach may be regarded as a reference, because estimation is based on the correct model, i.e. model selection is not included. Selected results of the simulation study are summarized in Tables 2 and 3 and Figs. 2 and 3. We draw the following conclusions:

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

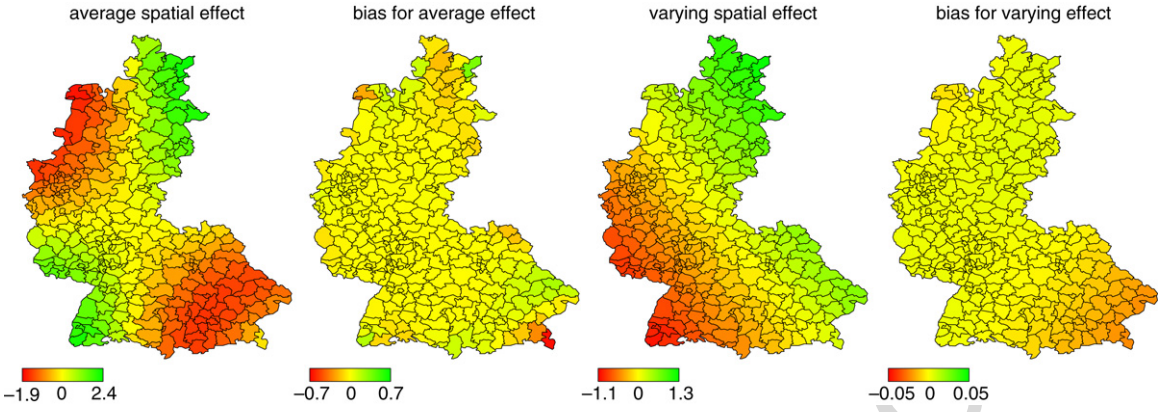

Fig. 3. Simulation study 1. Average estimates and empirical bias of the spatial main and interaction effect f_7 and g_7 for adaptive search.

Table 2
Simulation study 1

Approach	Algorithm 1	Stepwise	mgcv	MCMC (true)
Runtime	0:07	0:42	3:05	1:31

Computing times in hours for the first 25 replications.

- *Computing time:* The computing times displayed in Table 2 show that model selection is considerably faster using our adaptive Algorithm 1 compared with other approaches.
- *Bias and MSE:* All nonlinear terms are estimated with almost negligible bias, see Figs. 2 and 3 for our adaptive search algorithm. Results for the competitors are similar, and therefore omitted. The empirical mean squared errors (MSEs) are usually very close for all approaches, i.e. they perform more or less equally well (see Fig. 2 for two typical examples). Even the MSEs for the fully Bayesian approach are only slightly lower than for the model selection approaches. The empirical MSEs of the unimportant functions (not shown) are always below 0.02, indicating that individual estimated functions are close to zero.
- *Irrelevant terms:* On average, our search algorithm included 1.32 terms with no effect, whereas the relevant terms were selected in 100% of the replications. Hence, all selection errors are due the additional inclusion of irrelevant terms. However, each irrelevant function is removed from the model in 72%–80% of the replications.
- *Confidence intervals:* Table 3 shows average coverage rates for the different model terms. In order to assess the dependence of results on the sample size, we additionally included coverage rates based on a doubled sample size of $n = 1854$. The results are as follows:
  - For conditional confidence intervals, coverage rates tend to be below the nominal level, whereas the coverage rates of Bayesian confidence intervals based on MCMC simulation, are considerably above the nominal level. Average coverage rates of unconditional confidence intervals typically match or are slightly above the nominal level. For all approaches, average coverage rates are closer to the nominal level if the sample size is increased.
  - For irrelevant terms and spatial effects, average coverage rates are considerably above the nominal level, particularly for unconditional confidence intervals.

5.2. Two-dimensional surface

This section presents the results of a second simulation study that aims at examining the performance of the ANOVA type decomposition of a two-dimensional surface into main effects, and an interaction as described in Section 2.1.2. The true model is given by

$$y = \gamma_0 + f_1\left(z^{(1)}\right) + f_2\left(z^{(2)}\right) + f_{1|2}\left(z^{(1)}, z^{(2)}\right) + \varepsilon \tag{13}$$

with functions

$$f_1\left(z^{(1)}\right) = 12 \cdot \left(z^{(1)} - 0.5\right)^2 - 1.13,$$
$$f_2\left(z^{(2)}\right) = 1.5 \cdot \sin\left(3 \cdot \pi \cdot z^{(2)}\right) - 0.28,$$
$$f_{1|2}\left(z^{(1)}, z^{(2)}\right) = 3 \cdot \sin\left(2 \cdot \pi \cdot z^{(1)}\right) \cdot \left(2z^{(2)} - 1\right).$$

The true functions visualized in Figs. 4 and 5 are chosen such that the row and column wise means of the interaction, as well as the overall means of the three functions are zero. The errors ε are assumed to be Gaussian with variance $\sigma^2 = 1.16$

Table 3
Simulation study 1

		Conditional	Uncond.	MCMC	mgcv	Conditional	Uncond.	MCMC	mgcv
		f_1				g_1			
n	95%	0.898	0.933	0.969	0.906	0.952	0.963	0.973	0.889
$2n$	95%	0.932	0.959	0.970	0.933	0.939	0.958	0.972	0.941
n	80%	0.738	0.769	0.834	0.735	0.799	0.828	0.856	0.757
$2n$	80%	0.781	0.816	0.843	0.769	0.781	0.814	0.844	0.789
		f_2				g_2			
n	95%	0.951	0.962	0.971	0.952	0.947	0.983	–	0.992
$2n$	95%	0.943	0.958	0.969	0.955	0.956	0.991	–	0.995
n	80%	0.807	0.830	0.857	0.992	0.812	0.924	–	0.995
$2n$	80%	0.794	0.827	0.848	0.818	0.852	0.930	–	0.946
		f_3				g_3			
n	95%	0.921	0.939	0.966	0.861	0.923	0.939	0.961	0.903
$2n$	95%	0.940	0.953	0.970	0.935	0.948	0.962	0.970	0.950
n	80%	0.748	0.767	0.837	0.670	0.763	0.781	0.819	0.741
$2n$	80%	0.780	0.798	0.838	0.763	0.794	0.819	0.844	0.794
		f_4				g_4			
n	95%	0.940	0.949	0.964	0.950	0.938	0.979	–	0.992
$2n$	95%	0.944	0.951	0.958	0.950	0.945	0.983	–	0.989
n	80%	0.782	0.793	0.822	0.796	0.851	0.917	–	0.928
$2n$	80%	0.789	0.802	0.817	0.799	0.846	0.920	–	0.936
		f_5				g_5			
n	95%	0.930	0.967	–	0.982	0.938	0.978	–	0.992
$2n$	95%	0.950	0.981	–	0.992	0.947	0.982	–	0.980
n	80%	0.803	0.885	–	0.906	0.857	0.920	–	0.922
$2n$	80%	0.864	0.935	–	0.948	0.867	0.930	–	0.911
		f_6				g_6			
n	95%	0.951	0.982	–	0.993	0.948	0.983	–	0.986
$2n$	95%	0.942	0.983	–	0.992	0.955	0.986	–	0.982
n	80%	0.877	0.938	–	0.936	0.867	0.935	–	0.926
$2n$	80%	0.844	0.918	–	0.945	0.868	0.933	–	0.899
		f_{spat}				g_{spat}			
n	95%	0.988	0.984	0.990	0.917	0.984	0.985	0.987	0.960
$2n$	95%	0.994	0.991	0.995	0.926	0.980	0.983	0.987	0.940
n	80%	0.945	0.926	0.951	0.766	0.912	0.917	0.927	0.829
$2n$	80%	0.966	0.952	0.969	0.779	0.904	0.912	0.925	0.799

Average coverage probabilities for the individual functions based on nominal levels of 95% and 80%. Values that are more than 2.5% below (above) the nominal level are indicated in dark gray (light gray).

corresponding to a signal to noise ratio of 3. The sample size is $n = 300$. Overall, 250 replications of the model have been simulated.

We computed and compared the following estimators:

(a) Tensor product cubic P-splines on a 12 by 12 grid of knots with penalty (7) (henceforth anova). The penalty matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ in (7) correspond to second order difference penalties. Estimation has been carried out using Algorithm 2.
(b) Tensor product cubic P-splines on a 12 by 12 grid of knots using penalty (3) with penalty matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ based on second order differences (henceforth surface). Estimation has been carried out using Algorithm 1.
(c) Bayesian tensor product cubic P-splines as described in Lang and Brezger (2004) (henceforth mcmc). In addition, two main effects are explicitly specified. Estimation is carried out using MCMC simulation techniques. Note that model selection is not part of the estimation, i.e. in contrary to estimator (a) a correctly specified model is always estimated.

Figs. 4–6 summarize the results of the simulation study. We can draw the following conclusions:

- *Mean squared error (MSE):* In terms of the empirical MSE estimators (a) and (c) give comparable results whereas estimator (b) performs considerably worse (Fig. 4). Note that estimator (a) is competitive although it contains less parameters and model selection is part of the estimation process.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

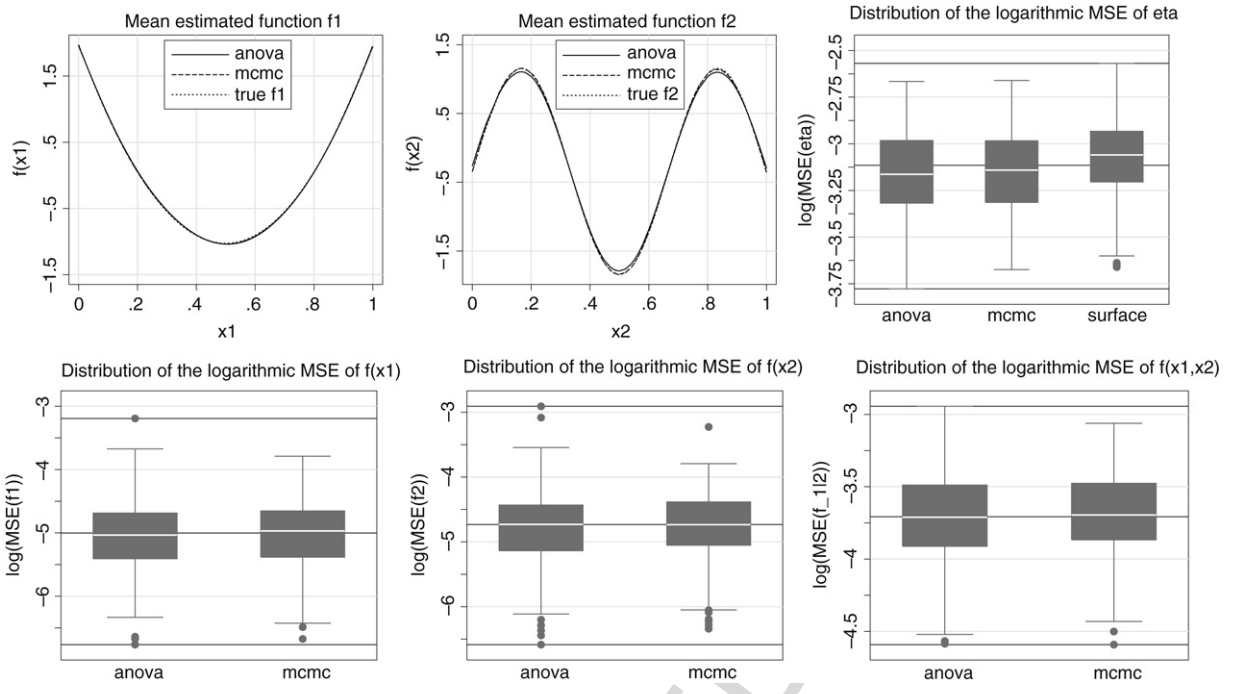

Fig. 4. Average estimated main effects together with the true underlying functions f_1 and f_2 (first row, first two graphs). Boxplots of $\log(MSE)$ for the predictor η and the nonlinear functions f_1, f_2 and $f_{1|2}$.

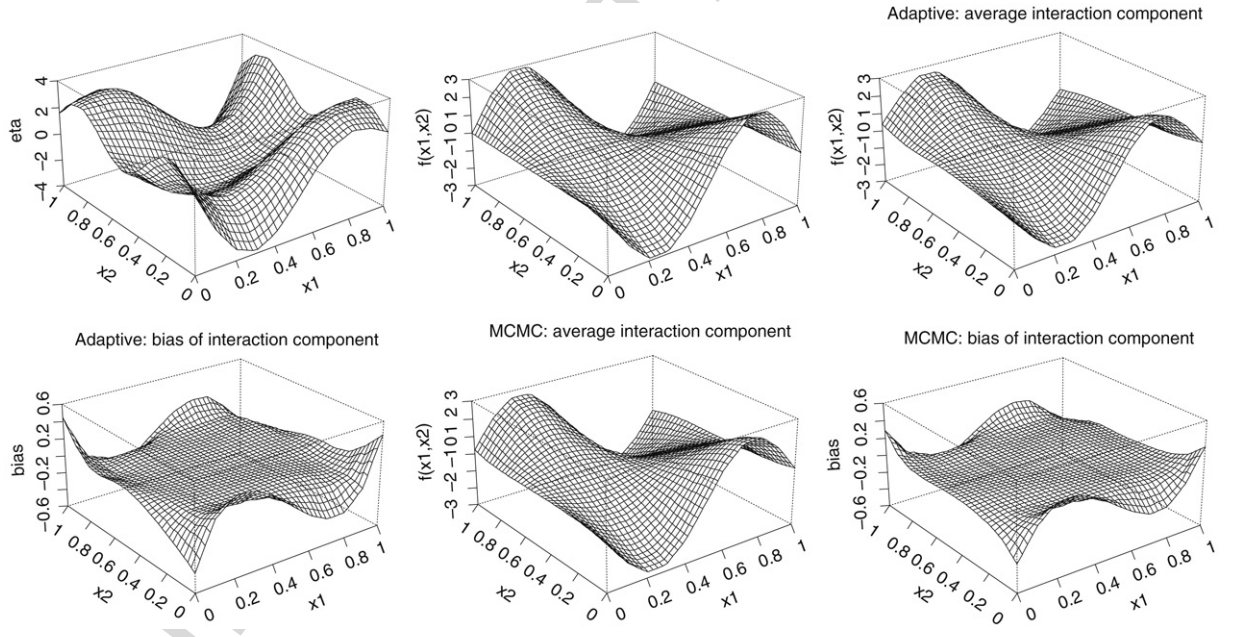

Fig. 5. First row: true predictor, true interaction component, average of estimated interaction component. Second row: bias of interaction component, average of estimated interaction component based on MCMC, bias of interaction component based on MCMC.

- *Bias:* The average of the estimated functions and the predictors in Figs. 4–6 reveal practically unbiased estimates for the two main effects f_1 and f_2. For the interaction $f_{1|2}$, and the predictor η, estimators (a) and (c) show a small bias at the corners and edges. For estimator (b), a considerably larger bias is observed for η. We conclude that estimators (a) and (c) perform equally well, whereas estimator (b) performs substantially worse.
- *Selection error rate:* The overall selection error rate for estimator (a) is zero, i.e. in 250 out of 250 replications; both main effects and the interaction effect are included in the selected model.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

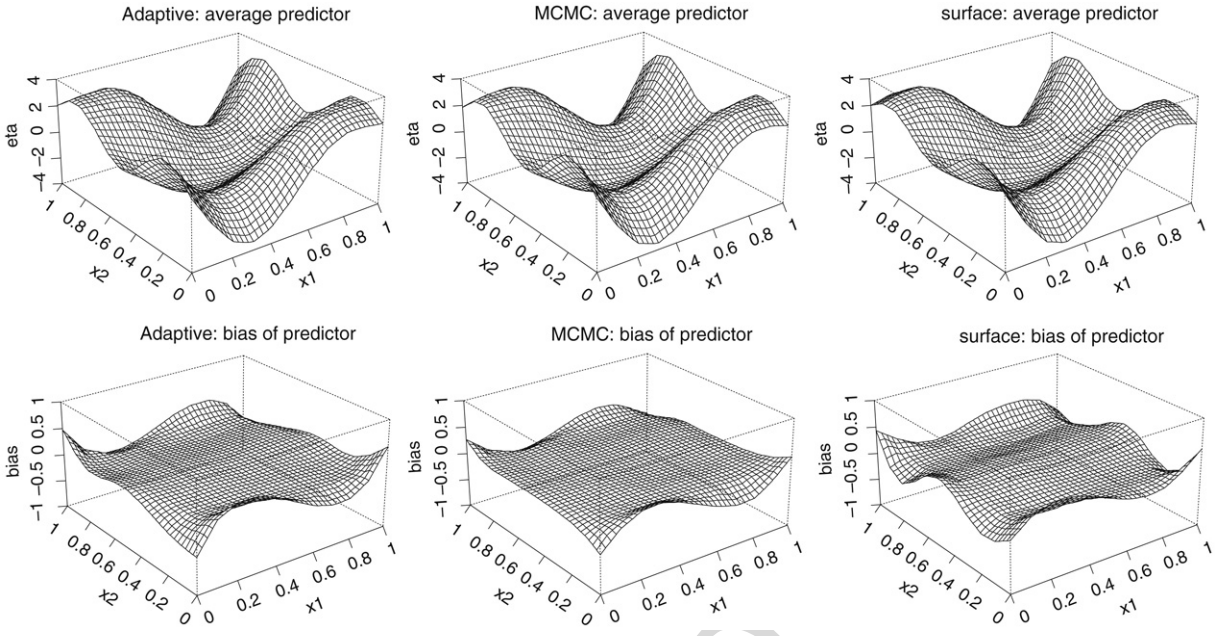

Fig. 6. Average estimated predictors (first row) and bias (second row).

In order to investigate the ability of our estimators to select the true model, a further simulation experiment has been conducted. The true model still contains the two main effects, but no interaction, i.e. $f_{1|2} \equiv 0$. The variance of the errors has been decreased to $\sigma^2 = 20.63$, in order to pertain the signal to noise ratio of 3. Our surface estimator (a) selected the correct model in roughly 74% of the cases, i.e. an interaction effect is not included in the final model. However, in 24% of the cases, an interaction is included, although not necessary. For these cases we investigated the AIC_c differences between the AIC_c best model, that misleadingly includes the interaction, and the main effects model. In the majority of cases, the AIC_c differences are well below 3. Hence in most cases, the main effects model is within the set of models that are not substantially worse than the AIC_c-best model.

6. Case study: Determinants of undernutrition

Very high prevalence of childhood undernutrition as well as very large gender bias, are two of the most severe development problems in India. In this case study, we analyze the determinants of undernutrition, and possible sex related differences. We thereby focus mostly on the statistical issues of the problem.

The analysis is based on micro data from the second National Family Health Survey (NFHS-2) from India which was conducted in the years 1998 and 1999. Among others, the survey collected detailed health, nutrition and anthropometric information on children born in the 3 years preceding the survey. There are approximately 13 000 observations of male and 12 000 observations of female children.

Undernutrition among children is usually measured by determining the anthropometric status of the child, relative to a reference population of children known to have grown well. Researchers distinguish different types of undernutrition. In this paper, we focus on stunting or insufficient height for age, indicating chronic undernutrition. For a child i, stunting is determined using a Z-score which is defined as

$$Z_i = \frac{AI_i - MAI}{\sigma}, \tag{14}$$

where AI refers to the height of the child, MAI and σ refers to the median height, and the standard deviation of children in the reference population at the same age. The aim of the analysis is

- to select and analyze important socio-demographic, environmental and health specific determinants of undernutrition
- to determine the functional form of the effects
- to investigate possible sex-specific differences of undernutrition and
- to account for spatial (and other sources of) heterogeneity.

The covariates used in this study are described in Table 1. This is a selection of the most interesting covariates (from a statisticians point of view). We have not included the full list of available covariates, because this would be beyond the scope of the paper.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

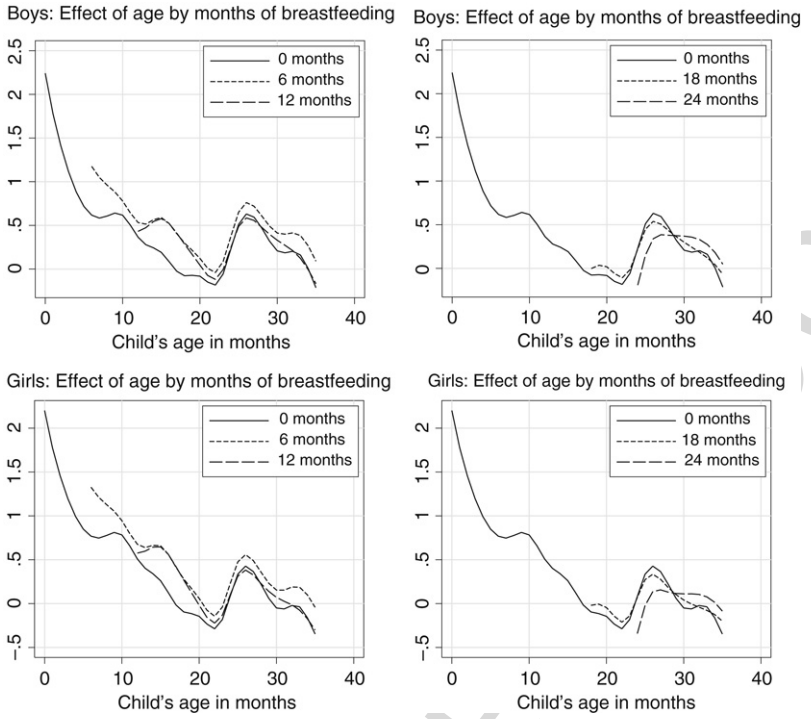

Fig. 7. Undernutrition in India: Estimated effects of child's age and duration of breastfeeding for boys and girls. Shown are slices through the two-dimensional surfaces for 0, 6, 12, 18 and 24 months of breastfeeding.

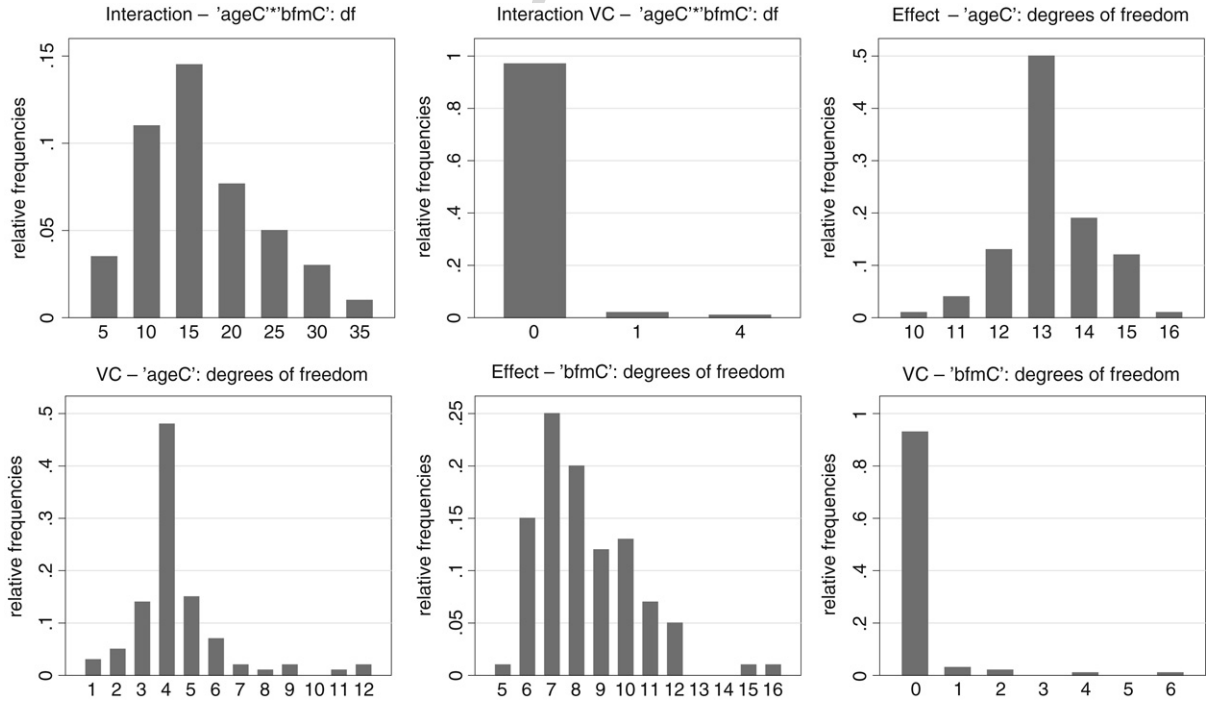

Fig. 8. Undernutrition in India: Sampling distribution of the degrees of freedom for the two-dimensional main, and interaction effect of child's age and duration of breastfeeding. VC is a shortcut for 'varying coefficient'. The top row shows from left to right: df's for the interaction $f_{1,1|2}$, the interaction $g_{1,1|2}$ of the VC term and the main effect $f_{1,1}$ of *ageC*. The bottom row shows the VC main effect $g_{1,1}$ of *ageC*, the main effect $f_{1,2}$ of *bfmC* and the VC main effect $g_{1,2}$ of *bfmC*.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

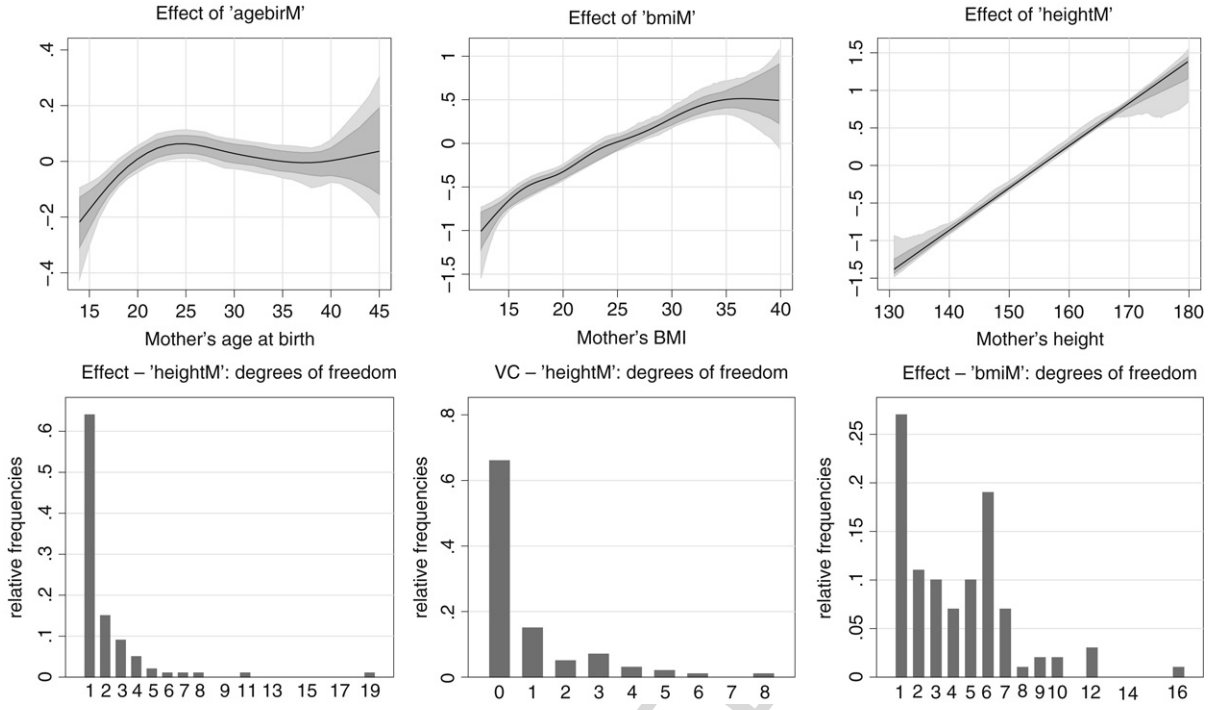

Fig. 9. Undernutrition in India: Estimated effects including 80% and 95% pointwise unconditional confidence intervals of mother's age at birth, body mass index and height.

The full predictor containing all effects, including interactions with sex is given by

$$\eta = f_1(ageC, bfmC) + g_1(ageC, bfmC) \cdot sex + f_2(agebirM) + g_2(agebirM) \cdot sex + f_3(bmiM) + g_3(bmiM) \cdot sex + f_4(heightM) + g_4(heightM) \cdot sex + f_5(district) + f_5(district) \cdot sex + \gamma_0 + \gamma_1\, sex,$$

where f_1 is a two-dimensional smooth surface of the child's age, and the duration of breastfeeding modeled by a tensor product P-spline with penalty (7). The effect of *ageC* and *bfmC* is a priori modeled by a two-dimensional effect, because an interaction between both variables is expected. Because of the penalty matrix used f_1 may be decomposed into two main effects, and an interaction effect, i.e.

$$f_1(ageC, bfmC) = f_{1,1}(ageC) + f_{1,2}(bfmC) + f_{1,1|2}(ageC, bfmC).$$

Algorithm 2 automatically selects which of the components are relevant. The penalty particularly allows one to select a main effects model as a special case with $\lambda^{(3)} = \infty$. f_2, f_3 and f_4 are one dimensional P-splines based on second order difference penalties, and f_5 is a district specific spatial effect. The functions $g_1, \ldots, g_5$ are possible interactions with sex. Again g_1 may be decomposed into main effects $g_{1,1}, g_{1,2}$ and an interaction $g_{1,1|2}$.

Using AIC_c the selected model is

$$\eta = f_1(ageC, bfmC) + g_1(ageC) \cdot sex + f_2(agebirM) + f_3(bmiM) + \gamma_2(heightM) + f_5(district) + f_5(district) \cdot sex + \gamma_0 + \gamma_1\, sex.$$

The model is much more parsimonious than the full model. In particular, most of the possible interactions with sex are not selected. Exceptions are the sex specific effects of child's age and district. Moreover, the selected model still contains the interaction between child's age and duration of breastfeeding. The effect of mother's height is linear in the best model, although there is at least evidence for a slight decrease of the slope for very tall women, see below.

The estimated nonlinear effects are given in Figs. 7–10. Partly included are unconditional (pointwise) 80% and 95% confidence intervals, and the sampling distribution of the degrees of freedom of the model. The general trend of the child's age effect is typical for studies on undernutrition, see e.g. Kandala et al. (2001). We observe a continuous worsening of the nutritional status, up until about 20 months of age. This deterioration sets in right after birth. After 20 months, stunting stabilizes at a low level. We also see a sudden improvement of the Z-score around 24 months of age. 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 socio-economic 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 Indian children to that sample, leads to a sudden improvement of their nutritional status at 24 months. The different curves for

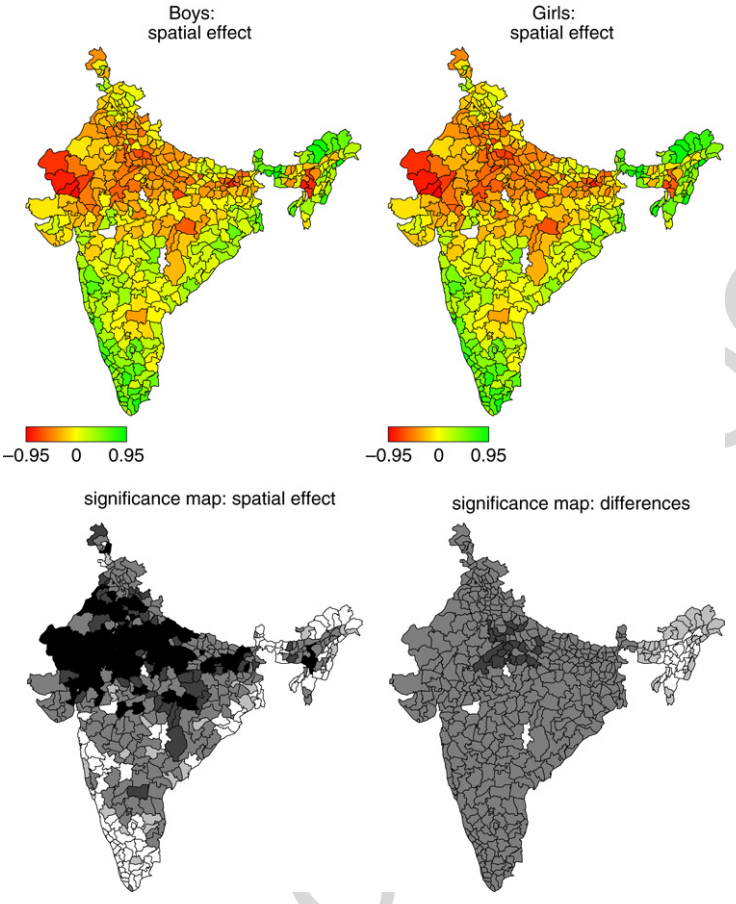

Fig. 10. Undernutrition in India: Estimated spatial effects for boys and girls (top panel) and significance maps (bottom panel) based on 80% and 95% confidence intervals.

children with different breast-feeding durations give evidence that both sexes are sensitive to breast-feeding. Boys and girls that are breastfed for 6 or twelve months have a better nutritional status throughout. Long breast-feeding durations (18 or 24 months) carry no benefits, however, and are probably an indicator of poor availability of alternative nutrition. Note, that the effect of breastfeeding is relatively weak. Therefore definite answers are not possible and further investigation, (using e.g. data of other countries) is necessary. The sampling distributions of the degrees of freedom are very helpful for assessing the uncertainty of the selected effect. It is relatively clear that a complex two-dimensional interaction effect with sex is not required. Instead a (nonlinear) interaction between the child's age and sex is sufficient.

The effect of mother's age at birth is clearly nonlinear, with improving nutritional status for older mothers. The effects of the mother's body mass index, and her height are very close to linearity. The effect of BMI is close to an inverse U form as suggested by the literature. However, obesity of the mother (possibly due to a poor quality diet) is likely to pose less of a risk for the nutritional status of the child, as very low BMIs which suggest acute undernutrition of the mother. Moreover, the sampling distribution of the degrees of freedom, as well as the unconditional confidence band show that a linear effect is at least reasonable. The Z-score is highest (and thus stunting lowest), at a BMI of around 30–35. The effect of the mother's height is linear in the selected best model. The sampling distribution of the degrees of freedom, and particularly the unconditional confidence intervals give at least evidence for a slight decrease of the slope for very tall women.

The spatial effect is relatively strong, showing a clear North–South pattern, with better nourished children in the South of India. The spatial interaction effect with sex is much weaker, but there seem to be some sex specific differences.

7. Conclusion

This paper proposes simultaneous selection of relevant terms, and smoothing parameters in models with structured additive predictors. A particular focus is devoted to developing fast algorithms for model choice. As a side aspect, a new penalty for two-dimensional surface smoothing is proposed. The proposal allows an additive decomposition into main effects, and an interaction effect without explicitly specifying and estimating complicated ANOVA type models. The paper also investigates conditional and unconditional (pointwise) confidence intervals for nonlinear terms. In the majority of

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

cases, conditional and unconditional confidence intervals are almost undistinguishable. In some important cases, however, both types of confidence interval are considerably different and unconditional confidence intervals provide further valuable insight.

We see several directions for future research:

- First of all, the approach can be generalized to models with non-Gaussian responses. This could be done by combining the usual iteratively weighted least squares (IWLS) algorithm for fitting generalized linear models (Fahrmeir and Tutz, 2001) with our backfitting-type algorithm. The combination of IWLS and backfitting is also known as local scoring (Hastie and Tibshirani, 1990).
- Another promising direction is a fully Bayesian approach, which is, however, very challenging both from a methodological and a computational point of view.
- Finally, we plan to investigate the alternative algorithmic approach discussed in remark 4 (see Algorithm 2). Presumably, the approach will not work for the most complex models, but it may be a promising alternative for models of moderate complexity.

Acknowledgments

This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen". We thank two anonymous referees for valuable comments that helped to improve the first version of the paper.

References

Belitz, C., 2007. Model selection in generalised structured additive regression models. Ph.D. Thesis, Dr.Hut–Verlag. Available online at: http://edoc.ub.uni-muenchen.de/.
Bollaerts, K., Eilers, P., Van Mechelen, I., 2006. Simple and multiple p-spline regression with shape constraints. British Journal of Mathematical and Statistical Psychology 59, 451–469.
Brent, R., 2003. Algorithms for Minimization Without Derivatives. Dover Publications.
Brezger, A., Kneib, T., Lang, S., 2005a. Bayesx: Analyzing Bayesian structured additive regression models. Journal of Statistical Software 14, 1–22.
Brezger, A., Kneib, T., Lang, S., 2005b. Bayesx manuals. Technical Report, Department of Statistics, University of Munich. Available at: http://www.stat.uni-muenchen.de/~bayesx.
Brezger, A., Lang, S., 2006. Generalized additive regression based on Bayesian P-splines. Computational Statistics and Data Analysis 50, 967–991.
Bühlmann, P., 2006. Boosting for high-dimensional linear models. Annals of Statistics 34, 559–583.
Casella, G., Moreno, E., 2006. Objective bayesian variable selection. Journal of the American Statistical Association 101, 157–167.
Chambers, J.M., Hastie, T., 1991. Statistical Models in S. Chapman and Hall.
Chen, Z., 1993. Fitting multivariate regression functions by interaction spline models. Journal of the Royal Statistical Society B 55, 473–491.
De Boor, C., 2001. A Practical Guide to Splines. Springer, New York.
Diggle, P.J., Heagerty, P., Liang, K.-L., Zeger, S.L., 2002. Analysis of Longitudinal Data (2. Auflage). Oxford University Press, Oxford.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Annals of Statistics 32, 407–451.
Eilers, P., Marx, B., 2002. Generalized linear additive smooth structures. Journal of Computational and Graphical Statistics 11, 758–783.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing using b-splines and penalized likelihood. Statistical Science 11, 89–121.
Eilers, P.H.C., Marx, B.D., 2003. Multidimensional calibration with temperature interaction using two-dimensional penalized signal regression. Chemometrics and Intelligent Laboratory Systems 66, 159–174.
Fahrmeir, L., Kneib, T., Lang, S., 2004. Penalized structured additive regression for space–time data: A Bayesian perspective. Statistica Sinica 14, 731–761.
Fahrmeir, L., Kneib, T., Lang, S., 2007. Regression. Modelle, Methoden und Anwendungen. Springer Verlag, Berlin.
Fahrmeir, L., Lang, S., 2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. Applied Statistics 50, 201–220.
Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models (2. Auflage). Springer, New York.
Fotheringham, A., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.
Gamerman, D., Moreira, A., Rue, H., 2003. Space-varying regression models: Specifications and simulation. Computational Statistics and Data Analysis 42, 513–533.
George, A., Liu, J.W., 1981. Computer Solution of Large Sparse Positive Definite Systems. Prentice-Hall, Englewood Cliffs, NJ.
Gu, C., 2002. Smoothing Spline ANOVA Models. Springer, New York.
Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. Journal of the Royal Statistical Society B 55, 757–796.
Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman & Hall / CRC, London.
Hastie, T.J., Tibshirani, R.J., Friedman, J., 2003. The Elements of Statistical Learning. Springer, New York.
Hurvich, C., Simonoff, J., Tsai, C., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. Journal of the Royal Statistical Society B 60, 271–293.
Jullion, A., Lambert, P., 2007. Robust specification of the roughness penalty prior distribution in spatially adaptive bayesian p-splines models. Computational Statistics and Data Analysis 51, 2542–2558.
Kamman, E.E., Wand, M.P., 2003. Geoadditive models. Applied Statistics 52, 1–18.
Kandala, N.B., Lang, S., Klasen, S., 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.
Kauermann, G., Claeskens, G., Opsomer, J.D., 2006. Bootstrapping for penalized spline rregression. Research Report KBI_0609, Faculty of Economics and Applied Economics, Katholieke Universiteit Leuven.
Kauermann, G., Opsomer, J., 2004. Generalized cross-validation for bandwidth selection of backfitting estimates in generalized additive models. Journal of Computational and Graphical Statistics 13, 66–89.
Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics 13, 183–212.
Lin, X., Zhang, D., 1999. Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical Society B 61, 381–400.
Marx, B., Eilers, P., 1998. Direct generalized additive modeling with penalized likelihood. Computational Statistics and Data Analysis 28, 193–209.
Miller, A., 2002. Subset Selection in Regression. Chapman & Hall / CRC, Boca Raton, FL.
Politits, D., Romano, J., 1994. The stationary bootstrap. Journal of the American Statistical Association 89, 1303–1313.
Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape. Applied Statistics 54, 507–554.
Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032

Shively, T., R, K., Wood, S., 1999. Variable selection and function estimation in additive nonparametric regression models using a data-based prior (with discussion). Journal of the American Statistical Association 94, 777–806.
Skrondal, A., Rabe-Hesketh, S., 2004. Generalized Latent Variable Modelling. Chapman & Hall / CRC, Boca Raton, FL.
Stasinopoulos, M., Rigby, B., Akanziliotou, P., 2005. Instructions on how to use the gamlss package in r. Technical Report. Availabe at: http://studweb.north.londonmet.ac.uk/stasinom/gamlss.html.
Stone, C.J., Hansen, M.H., Kooperberg, C., Truong, Y.K., 1997. Polynomial splines and their tensor products in extended linear modeling. Annals of Statistics 25, 1371–1470.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58, 267–288.
Tutz, G., Binder, H., 2007. Boosting ridge regression. Computational Statistics and Data Analysis 51, 6044–6059.
Verbeke, G., Molenberghs, G., 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.
Wood, S., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society B 62, 413–428.
Wood, S., 2004. On confidence intervals for gams based on penalized regression splines. Technical Report, University of Glasgow, Department of Statistics.
Wood, S., 2006a. R-manual: The mgcv package, version 1.3 - 22. Technical Report.
Wood, S.N., 2003. Thin-plate regression splines. Journal of the Royal Statistical Society B 65 (1), 95–114.
Wood, S.N., 2006b. Generalized Additive Models: An Introduction with R. Chapman & Hall / CRC, Boca Raton, FL.
Wood, S.N., 2006c. On confidence intervals for GAMs based on penalized regression splines. Australian and New Zealand Journal of Statistics 48 (4), 445–464.
Yau, P., Kohn, R., Wood, S., 2003. Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. Journal of Computational and Graphical Statistics 12, 23–54.

Please cite this article in press as: Belitz, C., Lang, S., Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis (2008), doi:10.1016/j.csda.2008.05.032