Leitenstorfer, Tutz:

Knot selection by boosting techniques

Sonderforschungsbereich 386, Paper 481 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Knot Selection by Boosting Techniques

Florian Leitenstorfer *, Gerhard Tutz

Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München, Germany

Abstract

A novel concept for estimating smooth functions by selection techniques based on boosting is developed. It is suggested to put radial basis functions with different spreads at each knot and to do selection and estimation simultaneously by a componentwise boosting algorithm. The methodology of various other smoothing and knot selection procedures (e.g. stepwise selection) is summarized. They are compared to the proposed approach by extensive simulations for various unidimensional settings, including varying spatial variation and heteroskedasticity, as well as on a real world data example. Finally, an extension of the proposed method to surface fitting is evaluated numerically on both, simulation and real data. The proposed knot selection technique is shown to be a strong competitor to existing methods for knot selection.

Key words: Nonparametric regression, Knot selection, Radial basis functions, Boosting, Surface fitting.

* Corresponding author. Tel.: ++4989 2180 2925; fax.: ++4989 2180 5308.
Email addresses: leiten@stat.uni-muenchen.de (Florian Leitenstorfer), tutz@stat.uni-muenchen.de (Gerhard Tutz).

1 Introduction

In the last decades, a tremendous amount of methods has been developed for the estimation of smooth functions f in an uni-dimensional regression setting $y = f(x) + \epsilon$. Besides localized approaches (see e.g. Fan and Gijbels, 1996), one distinguishes between methods based on smoothing splines or regression splines. The former (see Eubank (1988) or Wahba (1990)) uses many knots (up to the sample size n) which are placed in the design space, and the roughness of the estimate is controlled by a specific penalty term. The latter, on which we will focus in the present paper, is based on an expansion of f into basis functions, $f = \sum \alpha_j B_j$. In this setting, the number of actually chosen knots is much less than n. To avoid overfitting, one uses penalization strategies (P-splines, see Eilers and Marx, 1996) as well as knot selection strategies which are based on well known variable selection techniques. Stone et al. (1997) and He and Ng (1999) use stepwise selection, whereas Osborne et al. (1998) propose knot selection by Lasso (see Tibshirani, 1996). Knot selection from a Bayesian perspective has been treated by Smith and Kohn (1996), Denison et al. (1998) and Lang and Brezger (2004).

In the present paper, we aim at knot selection by employing recent developments in variable selection based on boosting techniques. Bühlmann and Yu (2003) propose a boosting algorithm constructed from the L_2-loss, which is suitable for high dimensional predictors in an additive model context. Bühlmann (2006) extends L_2Boost to the special issue of fitting high-dimensional linear models, where the number of covariates may exceed the sample size. This approach can straightforwardly be adapted to a regression spline context. It is possible to work with a very high number of basis functions, which are selected componentwise in a stepwise fashion. In order to obtain high flexibility in the resulting fits, we recommend to use radial basis functions (e.g. Ripley, 1996) with spreads chosen data-adaptively by componentwise boosting. As simulations will show, this leads to superior performance for the estimation of functions with high spatial variation as well as to robustness against violations of model assumptions.

The outline of the paper is as follows: in Section 2, we give an outline of the boosting algorithm. Section 3 contains a brief review over some alternative smoothing methods. The procedures are compared by a simulations study in the style of Wand (2000). In Section 4, the approach is extended to surface fitting.

2 A Smoothing Procedure Based on Componentwise L_2-Boosting

We consider the problem of conventional uni-dimensional nonparametric regression. For a dependent variable y_i and a covariate x_i, $i = 1, \ldots, n$, the model

$$y_i = f(x_i) + \sigma(x_i)\epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \tag{1}$$

is assumed, where $f(.)$ is a smooth function and $\sigma(.)$ is a positive function. A very popular approach to this problem is the expansion of f into basis functions, i.e.

$$f(x_i) = \alpha_0 + \sum_{j=1}^{m} \alpha_j B_j(x_i), \tag{2}$$

where the α_j are unknown coefficients, m is the number of knots and B_j denote the basis functions. Basis functions that have often been used in the literature are e.g. the truncated power series basis (see Ruppert and Carroll (2000) or Wand (2000)), the B-spline basis (Eilers and Marx, 1996) and the natural spline basis (Green and Silverman, 1994). Alternative basis functions that are suggested in the neural network community are the so-called radial basis functions (e.g. Ripley, 1996). An example for the latter are localized Gaussian densities, given by

$$B_j(x) = \exp\left(-\frac{|x - \tau_j|^2}{2h^2}\right), \tag{3}$$

where τ_j is the center of the basis function, and h determines the spread.

In the following, we will focus on radial basis functions, since they have some properties which are useful for the proposed procedure based on componentwise boosting. First, if we assume a sequence of knots $\{\tau_j\}_{j=1}^m$, the $B_j(x)$ as given in (3) are only linked to one knot τ_j. Furthermore, radial basis functions provide support on the entire real line. In contrast, B-splines—which are widely-used due to their numerical stability—are determined by $q + 2$ knots if they are of degree q. They have local support, i.e. they take values greater than 0 only on $q + 2$ consecutive knots. This implies for knot selection that the whole B-spline basis has to be recomputed, if a certain knot is added or deleted. It entails further a re-estimation of all coefficients.

The use of basis functions as given in (3) raises the question how to choose the spread h appropriately. A simple concept would be to take the same h at each knot, and to determine it e.g. by a data driven choice. However, this strategy is doubtful if one aims at constructing a flexible smoothing procedure, which should also be able to handle the estimation of functions with high spatial variation (see for example Ruppert and Carroll, 2000). We suggest to put at each knot τ_j, $j = 1, \ldots, m$, several radial basis functions with different spread h. Suppose we have a sequence of r distinct spread variables $h_1 < \cdots < h_r$,

then the expansion into basis functions from (2) is given as

$$f(x_i) = \alpha_0 + \sum_{j=1}^{m} \sum_{k=1}^{r} \alpha_{jk} B_{jk}(x_i), \tag{4}$$

where $B_{jk}(x_i) = \exp(-|x_i - \tau_j|^2/2h_k^2)$. From (4), it is seen that the number of parameters is large, if the number of knots m is high enough and the grid of h_k is subtle enough to get satisfying flexibility. Thus, a procedure is needed which is able to estimate high dimensional problems and avoids overfitting. In our simulation studies we found that the use of more than one spread improves the performance distinctively. As we will demonstrate, it is useful to center the basis functions by their means, i.e. we suggest to use $\tilde{B}_{jk}(x_i) = B_{jk}(x_i) - \frac{1}{n}\sum_{i=1}^{n} B_{jk}(x_i)$.

In the following, we propose a boosting method that selects the basis functions which are important for the data set at hand. Related approaches based on statistical variable selection techniques are given in Stone et al. (1997). We will refer to it in detail later. Boosting has originally been developed in the machine learning community to improve classification procedures (e.g. Schapire, 1990). With Friedman's (2001) gradient boosting machine it has been extended to regression modelling (see Bühlmann and Yu (2003) and Bühlmann (2006)). The basic concept in boosting is to obtain a fitted function iteratively by fitting in each iteration a "weak" learner to the current residual. Componentwise boosting in the sense of Bühlmann and Yu (2003) means that in one iteration, only the contribution of one variable is updated. However, in the problem considered here, componentwise does not refer to variables but to basis functions. Thus in each iteration only the contribution of one basis function is updated. The procedure automatically selects a subset of basis functions which produce a proper fit. Bühlmann (2006) developed an algorithm for estimation and variable selection in high-dimensional linear models, which can be brought forward to the smoothing problem given in (4). The weak learner that we used is ridge regression as proposed by Hoerl and Kennard (1970); for boosted variants of ridge regression, see also Tutz and Binder (2005). Before outlining the algorithm, the data are given in matrix notation: $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{x} = (x_1, \ldots, x_n)'$. Then, the expansion into radial basis functions yields the data set $(\mathbf{y}, \tilde{\mathbf{B}})$, where

$$\tilde{\mathbf{B}} = (\tilde{B}_{11}(\mathbf{x}), \ldots, \tilde{B}_{m1}(\mathbf{x}), \ldots, \tilde{B}_{1r}(\mathbf{x}), \ldots, \tilde{B}_{mr}(\mathbf{x}))$$

denotes a $(n \times mr)$-matrix with columns $\tilde{B}_{jk}(\mathbf{x}) = (\tilde{B}_{jk}(x_1), \ldots, \tilde{B}_{jk}(x_n))'$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$.

One might consider several ways to handle the intercept term in boosting algorithms. Since we propose to use centered basis functions, a computationally efficient way is to set the intercept term fixed to $\hat{\alpha}_0 = \bar{y}$. It may be shown that this suggestion yields the same results as estimation with updating of the intercept in each step, but using basis functions which are not centered.

L2KnotSmooth

Step 1 (Initialization)

Standardize $\mathbf{y}$ to zero mean, i.e. set $\hat{\alpha}_0 = \bar{y}$, $\hat{\boldsymbol{\alpha}}^{(0)} = (\bar{y}, 0, \ldots, 0)'$ and $\hat{\boldsymbol{\mu}}^{(0)} = (\bar{y}, \ldots, \bar{y})'$.

Step 2 (Iteration)

For $l = 1, 2, \ldots$, compute the current residuals $\mathbf{u}^{(l)} = \mathbf{y} - \hat{\boldsymbol{\mu}}^{(l-1)}$.

(1) *Fitting step*
For $j = 1, \ldots, m$, $k = 1, \ldots, r$, compute the ridge regression estimator with tuning parameter λ for the linear regression model

$$\mathbf{u}^{(l)} = \alpha_{jk}\tilde{B}_{jk}(\mathbf{x}) + \boldsymbol{\epsilon}.$$

The resulting ridge estimate is given by $\hat{\alpha}_{jk} = \tilde{B}_{jk}(\mathbf{x})'\mathbf{u}^{(l)}/[\tilde{B}_{jk}(\mathbf{x})'\tilde{B}_{jk}(\mathbf{x}) + \lambda]$.

(2) *Selection step*
Choose from the pairs $(j, k) \in \{1, \ldots, m\} \times \{1, \ldots, r\}$ the pair (j_l, k_l) such that $||\mathbf{u}^{(l)} - \hat{\alpha}_{jk}\tilde{B}_{jk}(\mathbf{x})||^2$ is minimized.

(3) *Update*
Set

$$\hat{\alpha}_{jk}^{(l)} = \begin{cases} \hat{\alpha}_{jk}^{(l-1)} + \hat{\alpha}_{j_l k_l}, \text{ if } \quad (j,k) = (j_l, k_l), \\ \hat{\alpha}_{jk}^{(l-1)}, \text{ otherwise,} \end{cases}$$

and

$$\hat{\boldsymbol{\mu}}^{(l)} = \hat{\boldsymbol{\mu}}^{(l-1)} + \hat{\alpha}_{j_l k_l}\tilde{B}_{j_l k_l}(\mathbf{x}).$$

In order to prevent overfitting, it is necessary to include a stopping criterion. The often used cross-validation criterion is not recommended because it implies heavy computational effort. A much more appropriate criterion is the AIC criterion which balances goodness-of-fit with the degrees of freedom (for AIC in smoothing, see Hastie and Tibshirani, 1990). In order to use the AIC criterion, the hat matrix of the smoother has to be given. For the present procedure, it can be obtained in a similar way as for componentwise L2Boost in linear models. With $\mathbf{S}_l = \tilde{B}_{j_l k_l}(\mathbf{x})\tilde{B}_{j_l k_l}(\mathbf{x})'/[\tilde{B}_{j_l k_l}(\mathbf{x})'\tilde{B}_{j_l k_l}(\mathbf{x}) + \lambda]$, $l = 1, 2, \ldots$ and $\mathbf{S}_0 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$, $\mathbf{1}_n = (1, \ldots, 1)'$, one has in the lth iteration

$$\hat{\boldsymbol{\mu}}^{(l)} = \hat{\boldsymbol{\mu}}^{(l-1)} + \mathbf{S}_l\mathbf{u}^{(l)} = \hat{\boldsymbol{\mu}}^{(l-1)} - \mathbf{S}_l(\hat{\boldsymbol{\mu}}^{(l-1)} - \mathbf{y}),$$

and therefore

$$\hat{\boldsymbol{\mu}}^{(l)} = \mathbf{H}_l \mathbf{y},$$

where

$$\mathbf{H}_l = \mathbf{I} - (\mathbf{I} - \mathbf{S}_l)(\mathbf{I} - \mathbf{S}_{l-1}) \cdots (\mathbf{I} - \mathbf{S}_0) = \sum_{j=0}^{l} \mathbf{S}_j \prod_{i=0}^{j-1} (\mathbf{I} - \mathbf{S}_{j-i-1}). \tag{5}$$

Since $\mathbf{H}_l$ corresponds to the hat-matrix after the lth iteration, $\mathrm{tr}(\mathbf{H}_l)$ may be considered as degrees of freedom of the estimate. A possible stopping rule for boosting iterations is based on the corrected AIC criterion proposed by Hurvich et al. (1998), given by

$$\mathrm{AIC}_c(l) = \log(\hat{\sigma}^2) + \frac{1 + \mathrm{tr}(\mathbf{H}_l)/n}{1 - (\mathrm{tr}(\mathbf{H}_l) + 2)/n}, \tag{6}$$

where $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})'(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})$. An alternative stopping criterion that has been recently used by Bühlmann and Yu (2006) in a boosting context is the g-prior minimum description length (gMDL),

$$\mathrm{gMDL}(l) = \log[n\hat{\sigma}^2/\{n - \mathrm{tr}(\mathbf{H}_l)\}] + \frac{\mathrm{tr}(\mathbf{H}_l)}{n} \log\left[\frac{\sum_{i=1}^{n} y_i^2 - n\hat{\sigma}^2}{\mathrm{tr}(\mathbf{H}_l)n\hat{\sigma}^2/\{n - \mathrm{tr}(\mathbf{H}_l)\}}\right]. \tag{7}$$

It is a hybrid between AIC and BIC (see Schwarz (1978) and Hansen and Yu (2001)). Thus, the complexity of the fit is penalized stronger, and models using less basis functions are expected. The optimal number of boosting iterations, which in our framework plays the role of a smoothing parameter, is estimated by $l_{\mathrm{opt}}^{\mathrm{AIC}_c} = \arg\min_l \mathrm{AIC}_c(l)$ or $l_{\mathrm{opt}}^{\mathrm{gMDL}} = \arg\min_l \mathrm{gMDL}(l)$.

It is noteworthy to emphasize that L2KnotSmooth may choose more than one basis function with different spreads h_k at a certain knot τ_j. Furthermore, due to its componentwise fitting strategy, L2KnotSmooth is able to handle problems where mr is fairly large, even if it exceed the sample size n.

The hat matrix may also be used as a starting point for the derivation of standard deviations of function estimates. Assuming that $E(\epsilon_i^2) = \sigma^2$, one obtains

$$\mathrm{cov}(\hat{\boldsymbol{\mu}}^{(l)}) = \mathbf{H}_l \mathrm{cov}(\mathbf{y}) \mathbf{H}_l' = \sigma^2 \mathbf{H}_l \mathbf{H}_l',$$

Using $\hat{\sigma}_\epsilon^2 = \frac{1}{n - \mathrm{tr}(\mathbf{H}_l)}(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})'(\mathbf{y} - \hat{\boldsymbol{\mu}}^{(l)})$ as an estimate for σ^2, one obtains

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\mu}}^{(l)}) = \hat{\sigma}_\epsilon^2 \mathbf{H}_l \mathbf{H}_l', \tag{8}$$

from which confidence intervals for $\hat{\boldsymbol{\mu}}^{(l)}$ can be derived. Note that the resulting confidence intervals should be regarded as a rough approximation. As simulations show (not given), the estimated standard deviations obtained from (8) tend to underestimate the true one, especially when the latter is high. This is

presumably due to the fact that the hat matrix is implicitly a function of the response $\mathbf{y}$.

A referee suggested to use place a knot at each observation and to use one unique spread h which should be rather large. We tried this in several examples given in Section 3. The simulations indicated that this strategy in most cases yields inferior results compared to using less knots but basis functions with different spreads. Thus, we do not pursue this approach in the rest of the present paper.

3 Numerical Comparisons for Univariate Settings

3.1 Simulation Settings

In the following, we give an outline of a simulation study which aims on the comparison of several smoothing methods, including the boosting approach presented in Section 2. The simulations were conducted similar to the settings used by Wand (2000). In all investigated scenarios, we considered the model given in (1), where the x_i, $i = 1, \ldots, n$, were drawn from a $U[0, 1]$-distribution. In each of the three investigated settings, a particular factor has been modified:

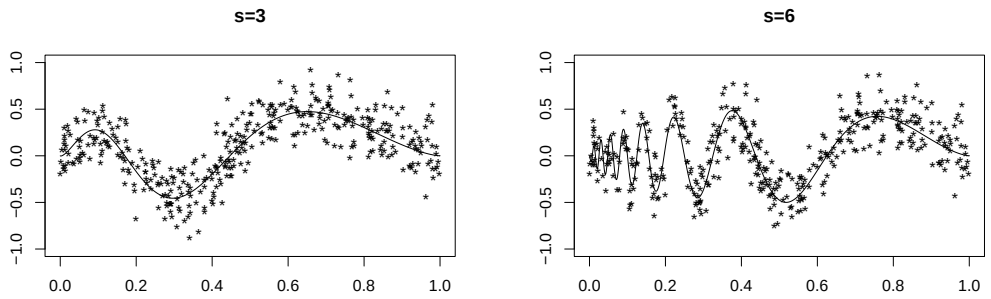

Figure 1. Setting I (spatial variation), typical data set for $s = 3$ (left panel) and $s = 6$ (right panel).

Setting I: Spatial variation Since the adaption to high spatial variability is one of the main issues of the proposed method, we consider the so called "Doppler" function which has also been used by Donoho and Johnstone (1994),

$$f_1(x) = \sqrt{x(1-x)}\,\sin\left(\frac{2\pi(1+2^{(9-4s)/5})}{x+2^{(9-4s)/5}}\right).$$

The oscillation can be controlled by the parameter $s = 1, \ldots, 6$, where $s = 1$ yields low spatial variation and $s = 6$ yields high spatial variation, respectively.

The standard deviation $\sigma(x_i)$ was set constant to $\sigma = 0.2$ in this setting. The sample size was $n = 400$. See Figure 1 for typical data sets.

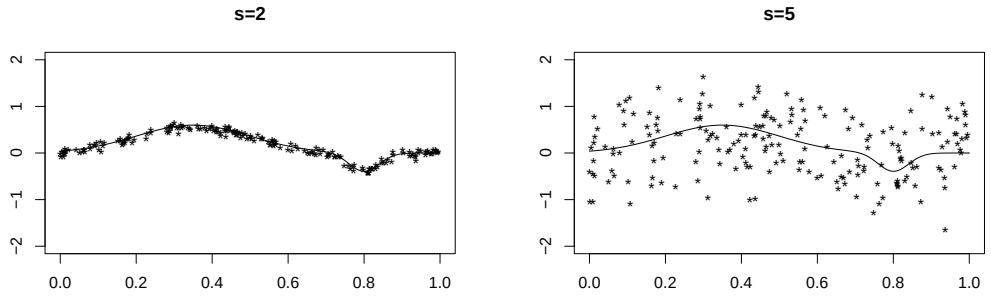

Figure 2. Setting II (varying noise level), typical data set for $s = 2$ (left panel) and $s = 5$ (right panel).

Setting II: Varying noise level This part of the study examines the influence of changes in the noise level on the resulting fits. The investigated function was

$$f_2(x) = 1.5\phi\left(\frac{x - 0.35}{0.15}\right) - \phi\left(\frac{x - 0.8}{0.04}\right), \tag{9}$$

where $\phi(.)$ denotes the standard normal density function. The standard deviation was set to $\sigma(x_i) \equiv \sigma_s$, where $\sigma_s = 0.02 + 0.04(s - 1)^2$ varied from $s = 1, \ldots, 6$. We used a sample size of $n = 200$ in this setting. See Figure 2 for typical data sets.

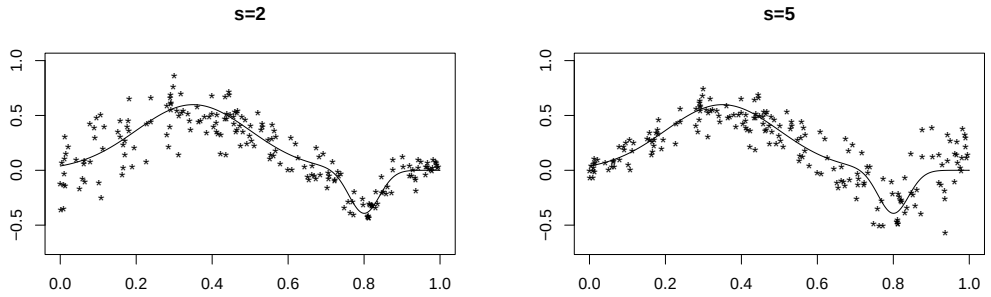

Figure 3. Setting III (heteroskedasticity), typical data set for $s = 2$ (left panel) and $s = 5$ (right panel).

Setting III: Heteroskedasticity Even though smoothing methods are mainly based on the homoskedasticity assumption of the error, it is interesting to compare the performance of smoothers if this assumption is dropped. Robustness against heteroskedastic errors may be an important property in

practice. Therefore, we considered the function from (9), and chose a standard deviation that depends on the response x,

$$\sigma_s(x) = 0.15\{1 + 0.4(2s - 7)(x - 0.5)\},$$

where s varied again from 1 to 6. Small values of s yield a higher variance for values of x near 0, whereas high ss result in higher variance for x-values near 1. Again a sample size of $n = 200$ was used. See Figure 3 for typical data sets.

The various fitting methods were compared by their out-of-sample performance. Therefore, 1000 new observations $x_i^{(\text{new})}, i = 1, \ldots, 1000$, were drawn from a $U[x_{\min}, x_{\max}]$-distribution where $x_{\min} = \min\{x_i\}$ and $x_{\max} = \max\{x_i\}$, and the averaged squared error (generalization error),

$$\text{ASE} = \frac{1}{1000} \sum_{i=1}^{1000} [\hat{f}(x_i^{(\text{new})}) - f(x_i^{(\text{new})})]^2,$$

was computed using the new observations.

The settings for the L2KnotSmooth procedure were as follows: $m = m' + 2$ knots τ_j, $j = 1, \ldots, m$, were chosen, where according to Wand (2000), m' interior knots were placed at $(x_{(dj')} + x_{(dj'+1)})/2$, $j' = 1, \ldots, m'$, with $m' = \lfloor n/d - 1 \rfloor$ and $d = \max\{4, \lfloor n/35 \rfloor\}$, ensuring that there are at least d observations between each knot. To keep dimensions comparable, we placed τ_1 at $\min\{x_i\}$ and τ_m at $\max\{x_i\}$. At each knot, $r = 13$ basis functions were allocated, with spreads h_k ranging from 0.02 to 0.2 (from 0.02 to 0.1 with step size 0.01, from 0.1 to 0.2 with step size 0.025). We tried several other ranges for h, especially with higher values for h, and found out that the results changed only marginally. For the ridge parameter, a fairly large value of $\lambda = 50$ was chosen. We believe that the chosen setting is a sensible compromise between accuracy and computational feasibility. The numerical analyses were carried out using the programming package R (R Development Core Team, 2006).

3.2 Alternative Approaches

In this section, we briefly review some alternative approaches for smoothing problems, which are included in the simulation study. Thereby, we take a closer look on methods which aim on knot selection.

The `mgcv` package in R The `mgcv` package (see Wood (2006)) aims mainly on the problem of multiple parameter selection for fitting generalized additive models with multiple smooth components. In the present case with an

unidimensional predictor and metric response, the fit is performed by penalized thin plate regression splines. This method bypasses the problem of knot placement by constructing an optimal approximating basis to thin plate splines (for details, see Wood, 2003). The roughness of the resulting fit is guided by a smoothing parameter, which is chosen by the GCV criterion. Hence, the comparison of this alternative to knot selection techniques presented in this paper is of special interest.

Multiple adaptive regression splines (MARS) The procedure suggested by Friedman (1991) is based on an expansion in basis function as in (2). This is done in a stepwise fashion, where basis functions are constructed successively from products of linear splines of the form $(x_i - \tau_j)_+$. The knots τ_j are chosen data adaptively by partition techniques. In each step, a new linear spline is included until a large model is constructed, which has to be pruned by backward deletion techniques (for details, see Friedman, 1991). For the present simulation study, we used the implementation `mars()` given in the R package `mda`, written by T. Hastie and R. Tibshirani. It should be noted that a MARS fit is not smooth, since it results from a linear combination of linear splines. It is more appropriate for multivariate design with nonlinear interactions.

Stepwise selection Stepwise selection of knots in regression splines goes back to Smith (1982). In the simulations, an algorithm is used which can be found more generally in Stone et al. (1997) and might be considered as a generalization of MARS. It has been formulated for uni-dimensional nonparametric regression in Wand (2000) and utilizes truncated power series of degree q, i.e. one has the matrix of basis functions

$$\mathbf{B} = (\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^q, (\mathbf{x} - \tau_1)_+^q, \ldots, (\mathbf{x} - \tau_m)_+^q), \tag{10}$$

where all operations are considered componentwise and $\{\tau_j\}_{j=1}^m$ denotes the sequence of knots. Let $B_j^{(t)}(\mathbf{x}) = (\mathbf{x} - \tau_m)_+^q$, $j = 1, \ldots, m$, then the algorithm is given as follows:

Step 1 (Initialization)

Define the minimal basis, $\mathbf{B}_{min} = (\mathbf{1}, \mathbf{x}, \ldots, \mathbf{x}^q)$ and set the current basis to $\mathbf{B}_c = \mathbf{B}_{\min}$.

Step 2 (Stepwise addition)

Repeat until $\mathbf{B}_c = \mathbf{B}$:

For each basis function that is not in $\mathbf{B}_c$, compute the Rao statistic,

$$R_j = \frac{B_j^{(t)}(\mathbf{x})'(\mathbf{I} - \mathbf{H}_c)\mathbf{y}}{\sqrt{B_j^{(t)}(\mathbf{x})'(\mathbf{I} - \mathbf{H}_c)B_j^{(t)}(\mathbf{x})}},$$

where $\mathbf{H}_c = \mathbf{B}_c(\mathbf{B}_c'\mathbf{B}_c)^{-1}\mathbf{B}_c'$, and include $B_j^{(t)}(\mathbf{x})$ in $\mathbf{B}_c$ which maximizes $|R_j|$. Fit the model with $\mathbf{B}_{c,\text{new}} = (\mathbf{B}_c, B_j^{(t)}(\mathbf{x}))$ by least squares and compute the corresponding GCV.

Step 3 (Stepwise deletion)

Repeat until $\mathbf{B}_c = \mathbf{B}_{\min}$:

For each basis $B_j^t(\mathbf{x})$ that is in $\mathbf{B}_c$, compute the Wald statistic,

$$W_j = \frac{[(\mathbf{B}_c'\mathbf{B}_c)^{-1}\mathbf{B}_c'\mathbf{y}]_j}{\sqrt{[(\mathbf{B}_c'\mathbf{B}_c)^{-1}]_{jj}}},$$

where $[.]_j$ denotes the jth component of a vector and $[.]_{jj}$ denotes the jth diagonal element of a matrix. Delete the basis function $B_j^{(t)}(\mathbf{x})$ that minimizes $|W_j|$. Fit the model with the reduced basis $\mathbf{B}_{c,\text{new}}$ by least squares and compute the corresponding GCV criterion.

We used the GCV criterion proposed by Stone et al. (1997), which is defined as

$$\text{GCV} = \frac{\hat{\sigma}^2}{(1 - a(J-1)/n)^2},$$

where $\hat{\sigma}^2$ is the residual sum of squares scaled by n, J denotes the number of parameters in the current model and a is an additional parameter that is set to 2.5. For the simulations conducted here, we followed Ruppert and Carroll (2000) by setting $q = 2$ and choosing the same initial set of knots as described for L2KnotSmooth (without the first knot at $\min\{x_i\}$ and the last knot at $\max\{x_i\}$, respectively).

Constrained B-spline smoothing A somewhat different approach to knot selection for regression splines has been proposed by He and Ng (1999). It is based on quantile regression techniques for smoothing problems (see e.g. Koenker et al., 1994). The idea is as follows: Consider bivariate random variables (x, y), then for the pth conditional quantile function of y given x, $f_p(x)$,

$$P(y \leq f_p(x)|x) = p$$

holds. When p is taken to 0.5, one obtains the conditional median function, which can be seen as a measure of central tendency. Consequently, it may be exploited in order to describe the relationship between x and y.

He and Ng (1999) distinguish between smoothing and regression spline approaches. In the present paper, we will focus on the latter (also called "median regression B-splines, see He and Shi (1994)). Let $p = 0.5$, then one is interested in estimating $f \equiv f_{0.5}$. The function f is decomposed into a B-spline basis (e.g. De Boor (1978) or Eilers and Marx (1996)), i.e. we have $f(x) = \sum_{j=1}^{m+q+1} \alpha_j B_j(x; q)$, where m is the number of internal knots and q denotes the degree of the B-splines. The initial sequence of knots is given by $\{\tau_j\}_{j=1}^{m+2(q+1)}$, and the selection of knots is carried out again in a stepwise fashion:

Step 1 (Initialization)

Compute median regression B-splines for $t = 0, \ldots, m$ interior knots of the initial sequence. They are obtained by minimizing the L_1-norm,

$$\min_{\alpha_t} |y_i - \sum_t \alpha_t B_t(x; q)|. \tag{11}$$

Select that t with the smallest AIC-criterion,

$$\mathrm{AIC}(T_t) = \log(\frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{f}_{T_t}(x_i)|) + 2(t + q)/n,$$

where T_t denotes the chosen knot sequence and $\hat{f}_{T_t}$ denotes the fit given by (11). Set $T_t = T_c$

Step 2 (Stepwise deletion)

Delete each of the t internal knots chosen in step 1 and compute the corresponding AIC. Choose the knot that results in the largest reduction of AIC and delete it in the updated current knot sequence T_c. Repeat this step until no further reduction in AIC occurs. Denote the remaining knot sequence as $T_{\min}$ and take the corresponding fit resulting from (11) as the final fit.

Note that for each change of the knot sequence, the B-spline basis has to be recomputed. Since the estimation is based on the L_1-norm, the procedure is expected to show robustness to outlying response variables. It is implemented in the library `cobs` of R. In the simulation studies, we used B-splines of degree $q = 2$. In order to attain comparability, the same set of initial knots as for L2KnotSmooth was used.

Knot selection using the Lasso This knot selection procedure aims on knot selection by the least absolute shrinkage and regression operator proposed by Tibshirani (1996) originally for linear regression. The idea is to solve a least squares problem under a constraint on the L_1-norm of the vector of coefficients. Osborne et al. (1998) transferred the approach to a regression spline framework. Assume an initial sequence of knots $\{\tau_j\}_{j=1}^m$ and a truncated power series basis of degree q, the design matrix of basis functions $\mathbf{B}$ is given by (10). The proceeding is as follows:

Let t be a shrinkage parameter that determines an upper bound of the L_1-norm of the parameters under constraint. For a grid of values t_u, minimize

$$(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})'(\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \quad \text{subject to} \quad \sum_{j=q+2}^{m+q+1} |\alpha_j| \leq t_u, \tag{12}$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{m+q+1})'$. Since Lasso is able to do variable selection, choose the sequence of knots $T_u = \{\tau_j | \alpha_{j+q+1} \neq 0, j = 1, \ldots, m\}$. Define the design matrix $\mathbf{B}_u$ that contains—besides the first $q+1$ bases—only the columns that belong to the knots in T_u and compute an ordinary least squares fit by using $\mathbf{B}_u$, yielding $\hat{\boldsymbol{\alpha}}_u$. Choose the optimal shrinkage parameter t_{opt} and the corresponding optimal fit by minimizing AIC, where the residual sum of the unconstrained fit is penalized by the number of nonzero parameters in $\hat{\boldsymbol{\alpha}}_u$.

One may wonder why an unconstrained fit is used instead of the coefficients obtained by Lasso. Osborne et al. (1998) found that the method performs much better with this strategy, and our experiments support their observations. In contrast to Osborne et al. (1998), in (12) we put only the coefficients under the constraint that belong to the truncated bases. It seems to be more sensible than a constraint on all parameters, since we aim on knot selection, and the first $q+1$ coefficients are not linked to a certain knot. The Lasso fit was performed by the function `l1ce()` of the library `lasso2` in R (for details, see Lokhorst (1999)), where the basis functions were not standardized. Furthermore, the adjustments and the initial knot mesh of the stepwise selection procedure were used. In the following, we refer to this method as OPT.

3.3 Results

The results of the spatial variation setting are given in Table 1, where for the various knot selection techniques the median of the relative change (in %) in ASE compared to MGCV, i.e.

$$\frac{\text{ASE} - \text{ASE(MGCV)}}{\text{ASE(MGCV)}} \times 100,$$

		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
MARS	$q_{.5}$ (%)	**79.47**	99.57	806.57	398.13	38.91	10.28
	$[q_{.25}, q_{.75}]$	[25.93,163.18]	[55.59,164.06]	[592.48,1222.81]	[348.32,460.40]	[27.83, 49.09]	[6.05, 14.68]
L2KS, AIC_c	$q_{.5}$ (%)	129.63	39.83	35.98	**-74.23**	**-83.14**	**-76.33**
	$[q_{.25}, q_{.75}]$	[60.03,230.56]	[14.30, 76.04]	[17.14, 66.73]	[-79.59,-69.38]	[-86.81,-79.85]	[-79.20,-72.86]
L2KS, gMDL	$q_{.5}$ (%)	83.82	**11.50**	**32.53**	-73.30	-81.44	-74.96
	$[q_{.25}, q_{.75}]$	[30.14,174.16]	[-4.09, 32.28]	[8.08, 58.55]	[-79.01,-67.82]	[-84.98,-76.74]	[-77.80,-70.91]
Stepwise	$q_{.5}$ (%)	80.71	52.52	73.68	-65.04	-81.37	-74.03
	$[q_{.25}, q_{.75}]$	[25.78,160.02]	[9.89,136.30]	[28.24, 144.28]	[-70.94,-58.34]	[-84.82,-76.24]	[-78.62,-68.92]
COBS	$q_{.5}$ (%)	123.93	53.64	60.70	-59.52	-70.40	-65.83
	$[q_{.25}, q_{.75}]$	[54.67,242.97]	[14.45,110.06]	[18.34, 129.07]	[-69.54,-50.04]	[-76.47,-64.39]	[-71.78,-59.00]
OPT	$q_{.5}$ (%)	152.39	24.30	38.79	**-74.19**	**-81.95**	**-75.44**
	$[q_{.25}, q_{.75}]$	[70.99,331.89]	[-13.19,130.30]	[0.65, 113.62]	[-79.44,-62.61]	[-85.04,-77.65]	[-78.80,-70.66]

Table 1
Setting I, median percentage change of ASE relative to MGCV over $S = 200$ simulated data sets, along with the corresponding 25th and 75th percentiles.

is reported over $S = 200$ simulated data sets (the best two performers are given in bold faces—if MGCV is among these, only one number is in bold face). It is seen that for lower spatial variation ($s = 1, \ldots, 3$), MGCV performs very well, compared to knot selection techniques, whereas for higher spatial variation, the procedure is less adequate than knot selection approaches. L2KnotSmooth shows good results and outperforms COBS in most settings. Stepwise selection and OPT are very competitive for high spatial variation, but the former does clearly worse than gMDL-stopped L2KnotSmooth for $s = 2$ and 3. Not surprisingly, AIC_c-stopped boosting yields slightly better results for high spatial variation, while gMDL works distinctively better in the case of lower spatial variation. The reason is that the complexity of the fit is penalized stronger by gMDL, which is an advantage for low spatial variation cases, where a smaller number of chosen basis functions is sufficient for a proper fit. High spatial variation implies a more complex model, which is provided by boosting stopped by the AIC_c-criterion, but differences are only marginal. Note that in the case of $s = 6$, there occurred 16 data sets where a distinct minimum of the AIC_c criterion was not found within the maximum number of 1000 boosting iterations, compared to only two where gMDL was used. When taking a look at the quartiles, it is seen that L2KnotSmooth tends to show comparable, sometimes even less variation than most of the competitors.

Table 2 shows the results of the varying noise setting. It is seen that L2KnotSmooth dominates the other knot selection methods in most experiments. The gMDL criterion does better for high noise, while AIC_c has marginal advantages in low noise cases. For a low signal-to-noise ratio ($s = 6$), MGCV performs better than the knot selection methods, but shows serious drawbacks for higher signal-to-noise ratio. This observation, together with the results of setting I, indicates that MGCV tends to oversmooth the data. Due to its piecewise linear structure, MARS cannot compete with L2KnotSmooth or stepwise selection in the case of low noise, but shows reasonable performance for high noise cases. OPT performs distinctively worse than the other knot selection approaches when noise is high.

		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
MARS	$q_{.5}$ (%)	-50.75	-41.08	-9.74	**-7.05**	23.09	49.28
	$[q_{.25}, q_{.75}]$	[-66.37,-32.72]	[-57.62,-21.46]	[-26.08, 5.37]	[-24.34, 14.71]	[-5.75, 60.89]	[13.88, 88.32]
L2KS, AIC_c	$q_{.5}$ (%)	**-96.37**	**-78.94**	**-27.96**	6.00	39.15	61.16
	$[q_{.25}, q_{.75}]$	[-97.17,-94.73]	[-83.35,-72.50]	[-40.06, -9.08]	[-12.34, 36.31]	[5.21, 82.73]	[11.37,123.89]
L2KS, gMDL	$q_{.5}$ (%)	**-95.64**	**-76.80**	**-38.66**	**-15.49**	**16.18**	44.48
	$[q_{.25}, q_{.75}]$	[-96.80,-94.13]	[-82.57,-70.91]	[-47.35,-21.05]	[-31.89, 2.21]	[-12.94, 49.91]	[1.95,106.17]
Stepwise	$q_{.5}$ (%)	-93.86	-65.79	9.15	38.53	33.89	**39.64**
	$[q_{.25}, q_{.75}]$	[-95.07,-92.04]	[-75.34,-54.94]	[-18.37, 49.01]	[11.87, 75.30]	[-3.10, 79.91]	[3.41,100.61]
COBS	$q_{.5}$ (%)	-93.31	-60.64	45.25	69.14	62.98	61.27
	$[q_{.25}, q_{.75}]$	[-95.11,-90.69]	[-70.76,-49.48]	[12.85, 89.67]	[28.79,117.28]	[18.89,132.80]	[9.79,124.72]
OPT	$q_{.5}$ (%)	-95.18	-69.67	15.99	67.20	137.64	175.03
	$[q_{.25}, q_{.75}]$	[-96.30,-92.87]	[-77.12,-55.41]	[-21.10, 75.96]	[18.91,194.97]	[59.46,279.17]	[83.21,418.56]

Table 2
Setting II, median percentage change of ASE relative to MGCV over $S = 200$ simulated data sets, along with the corresponding 25th and 75th percentiles.

		$s=1$	$s=2$	$s=3$	$s=4$	$s=5$	$s=6$
MARS	$q_{.5}$ (%)	-25.67	-22.33	-24.63	-14.96	-5.51	2.95
	$[q_{.25}, q_{.75}]$	[-42.48, -8.31]	[-38.73, -5.73]	[-37.14, -6.59]	[-30.79, 8.85]	[-24.52, 17.68]	[-16.91, 24.55]
L2KS, AIC_c	$q_{.5}$ (%)	**-25.88**	**-32.94**	**-36.85**	**-35.17**	**-23.86**	**-13.73**
	$[q_{.25}, q_{.75}]$	[-37.82, -2.28]	[-42.85,-17.12]	[-46.54,-20.87]	[-47.13,-17.49]	[-41.85, -0.71]	[-37.13, 17.01]
L2KS, gMDL	$q_{.5}$ (%)	**-33.14**	**-39.62**	**-42.23**	**-41.15**	**-36.25**	**-31.59**
	$[q_{.25}, q_{.75}]$	[-46.89,-19.21]	[-51.37,-26.02]	[-54.84,-29.08]	[-52.91,-27.23]	[-49.95,-22.09]	[-45.89,-12.35]
Stepwise	$q_{.5}$ (%)	17.18	-4.86	-12.28	-2.81	19.88	47.76
	$[q_{.25}, q_{.75}]$	[-14.38, 77.88]	[-27.92, 37.87]	[-31.88, 23.21]	[-27.17, 30.43]	[-12.60, 67.50]	[6.12,105.25]
COBS	$q_{.5}$ (%)	20.01	14.67	20.02	25.42	37.53	47.33
	$[q_{.25}, q_{.75}]$	[-9.86, 75.45]	[-9.49, 48.40]	[-4.96, 46.99]	[-0.44, 65.94]	[8.46, 83.60]	[14.21, 96.31]
OPT	$q_{.5}$ (%)	33.32	12.38	0.54	2.30	20.93	53.11
	$[q_{.25}, q_{.75}]$	[-10.92,120.96]	[-23.75, 86.20]	[-31.34, 40.34]	[-29.75, 60.44]	[-20.42,106.74]	[-1.33,152.42]

Table 3
Setting III, median percentage change of ASE relative to MGCV over $S = 200$ simulated data sets, along with the corresponding 25th and 75th percentiles.

Table 3 summarizes the results of the heteroskedasticity setting. The boosting methods outperform all other predictors whereas gMDL dominates the AIC_c criterion in all cases. This shows the robustness of L2KnotSmooth to the violation of common model assumptions. MARS does surprisingly well for small values of $s = 1, 2$ (i.e. for small values of x, the corresponding error variance is high) and outperforms the other knot selection techniques. Interestingly, the MGCV procedure yields fairly moderate results throughout all investigated settings.

In order to learn more about the functionality of L2KnotSmooth, it is interesting to look at the development of chosen basis functions and parameter estimates in dependence of the number of iterations. Figure 4 shows the number of chosen basis functions (left panel) and the L_1-norm of parameters, $\sum_{j,k} |\alpha_{j,k}|$ (right panel), for 300 boosting iterations on typical data sets of setting II with $s = 2$ (lower noise) and $s = 5$ (higher noise). The optimal number of boosting iterations, estimated by AIC_c and gMDL for the two data sets, is represented by vertical lines. From the left panel, it is seen that the number of chosen basis functions is similar up to iteration 60. For higher number of

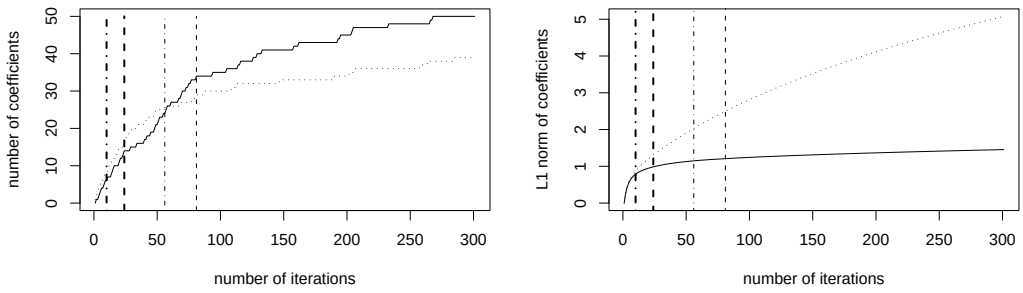

Figure 4. Number of selected basis functions (left panel) and L_1-norm of coefficients (right panel) vs. the number of boosting iterations for setting II, $s = 2$ (low noise, solid) and $s = 5$ (high noise, dotted). The thin vertical lines give the optimal number of iteration for $s = 2$ (AIC_c: dashed; gMDL: dash-dotted), the bold lines for $s = 5$, respectively.

iterations the procedure tends to choose more basis functions in the low noise case than for high noise. On the other hand, as the right panel shows, the curve of the L1-norm shows some saturation after 50 iterations for $s = 2$, whereas it grows considerably faster for $s = 5$. This behavior might be interpreted as sign for overfitting with increasing number of iterations, since a high L1-norm for the parameters indicates a wiggly curve. The optimal numbers of iterations suggest that the two proposed criteria show a reasonable resistance against overfitting.

3.4 Sensitivity to outliers

Often it is instructive to explore the behavior of a smoothing method when outliers are present in the data. Some approaches to robustifying the choice of the smoothing parameter have been suggested in the literature (see e.g. Härdle (1984)). In the following we do not develop robustified methods but investigate the behavior of L2KnotSmooth and the methods described in Subsection 3.2 when response outliers are present in a small simulation study similar to Leung (2005). To this end, regression function (9) was used, where we considered the case of no contamination of the response using $\sigma(x_i) \equiv 0.1$ and two types of contamination: for type 1, let $\sigma(x_i) = 0.1b_{1,i} + 0.9(1 - b_{2,i})$, where $b_{1,i} \overset{iid}{\sim} \mathcal{B}(1, 0.1)$ and for type 2, let $\sigma(x_i) = 0.1b_{2,i}+1.5(1-b_{2,i})$, where $b_{2,i} \overset{iid}{\sim} \mathcal{B}(1, 0.2)$. All other configurations were the same as in setting II. Typical data sets for the three scenarios are given in Figure 5.

In Table 4, the median ASE evaluated over $S = 200$ simulated data sets is given, as well as its median relative change compared to the results of the corresponding fitting method without contamination. For the latter, the fits by L2KnotSmooth show the best performance. When outliers are present, COBS

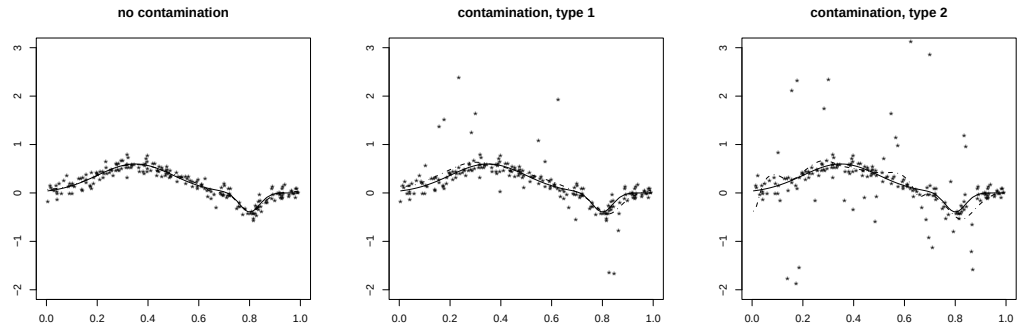

Figure 5. Exemplary data sets along with the true function (solid) and a L2KnotSmooth fit (dash-dotted; gMDL-stopped). Left panel: no contamination; mid panel: contamination, type 1; right panel: conatmination, type 2.

	med(no cont.)	med(cont. 1)	rel. change [%]	med(cont. 2)	rel. change [%]
MGCV	0.0017	0.0052	197.51	0.0141	764.47
MARS	0.0012	0.0047	300.69	0.0190	1396.95
L2KS, AIC_c	0.0007	0.0061	683.26	0.0237	3183.71
L2KS, gMDL	0.0007	0.0046	533.26	0.0197	2755.58
Stepwise	0.0011	0.0081	654.47	0.0243	2088.96
COBS	0.0014	0.0020	30.09	0.0053	268.15
OPT	0.0012	0.0100	726.02	0.0384	3428.91

Table 4
Median ASE for different degrees of contamination over S=200 simulated data sets, along with percentage of increase in ASE compared to the uncontaminated case

which is based on minimizing the L_1-norm is—not surprisingly—by far the best solution. The penalized regression spline fit obtained by `mgcv` seems to be less sensitive to outliers than the remaining knot selection techniques. It is seen that stopping L2KnotSmooth by gMDL yields more robust results than using AIC_c. One should emphasize that in absolute terms of ASE, gMDL-stopped L2KnotSmooth outperforms most of the knot selection approaches (except COBS) also in the contaminated cases.

3.5 Example: LIDAR Data

LIDAR (Light Detection And Ranging) is a method for detection of chemical compounds in the atmosphere; it uses the reflection of laser-emitted light. We consider a typical LIDAR data set which has been previously analyzed by Holst et al. (1996) and Ruppert and Carroll (2000). The dependent variable (range) measures the distance traveled before the light is reflected back to its source. The response (log-ratio) is the logarithm of received signals at frequencies on and off the resonance frequency of the chemical species of interest (mercury). The data set is of special interest to the proposed smoothing method, since

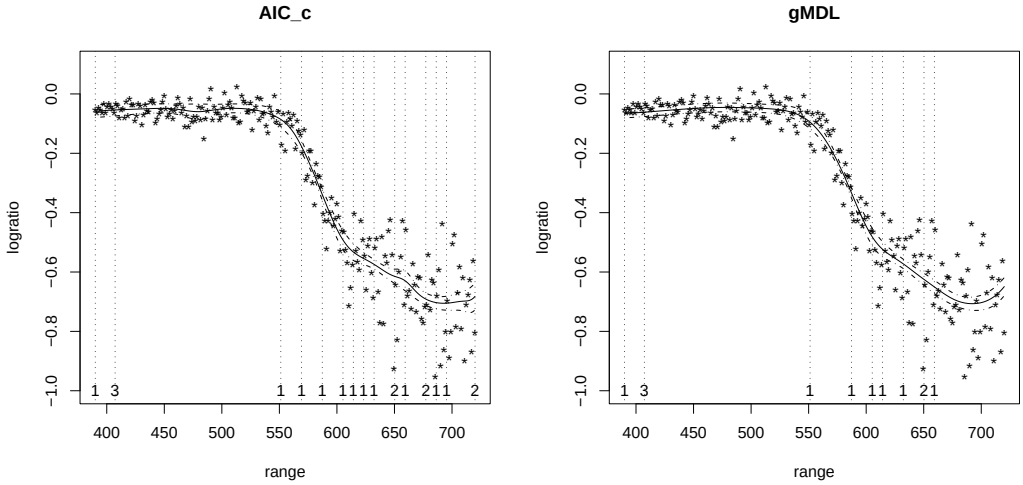

Figure 6. LIDAR data: Fits (and approximate pointwise 0.95 confidence intervals) for AIC_c-stopped (left panel) and gMDL-stopped (right panel) L2KnotSmooth, along with the location of chosen knots and the number of chosen basis functions at each knot.

there is heteroskedasticity in the data (see Figure 6) and the simulations above show good performance of L2KnotSmooth in that case.

In Figure 6, the L2KnotSmooth fits are shown, where boosting was stopped with AIC_c (left panel) and gMDL (right panel). All settings were the same as in the simulation study. Additionally, the location of the actually chosen knots are given, along with the number of contributing basis functions at each knot. It is seen that AIC_c-stopped boosting ($l_{\text{opt}} = 50$) chose more basis functions on the right end of range, yielding a somewhat more wiggly fit. In contrast, gMDL-stopped boosting ($l_{\text{opt}} = 24$) chose less knots especially for a range > 640. For smaller values of range the location and number of basis functions is the same for both criteria, and the fits are very similar in this region. We compared L2KnotSmooth with the alternative approaches presented above in terms of the corrected AIC. AIC_c-stopped L2KnotSmooth performs best ($\text{AIC}_c = -4.059$); the stepwise approach and MGCV comes closest (-4.033 and -4.019, respectively). We also give approximate confidence intervals according to (8). Note that this confidence bands should be regarded only as a rough guide, since heteroskedasticity of errors is not taken into account.

4 Extension to Surface Fitting

In the following, we consider the problem of estimating a smooth function $g : \mathbb{R}^p \mapsto \mathbb{R}$ of p covariates $x_1, \ldots, x_p$, i.e. we assume the model

$$y_i = g(x_{i1}, \ldots, x_{ip}) + \sigma(x_{i1}, \ldots, x_{ip})\epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1). \tag{13}$$

L2KnotSmooth can straightforwardly be extended to model (13) by expansion of g into p-dimensional radial basis functions. For simplicity, let $p = 2$. Consider a grid of knots $\{\boldsymbol{\tau}_{j,l}\}_{j,l=1}^{m}$ where $\boldsymbol{\tau}_{j,l} = (\tau_j^{(1)}, \tau_l^{(2)})' \in \mathbb{R}^2$, and two-dimensional radial basis functions which are given by

$$B_{jlk}(x_1, x_2) = \exp\left(-\frac{((x_1, x_2)' - \boldsymbol{\tau}_{j,l})'((x_1, x_2)' - \boldsymbol{\tau}_{j,l})}{2h_k^2}\right).$$

Thus, the basis expansion results in the model

$$g(x_{i1}, x_{i2}) = \alpha_0 + \sum_{j=1}^{m}\sum_{l=1}^{m}\sum_{k=1}^{r} \alpha_{jlk} B_{jlk}(x_{i1}, x_{i2}), \tag{14}$$

where at each knot $\boldsymbol{\tau}_{j,l}$, r radial basis functions with different spreads are allocated. To fit (14), the L2KnotSmooth algorithm can be applied directly, along with the corresponding stopping criteria. Note that we again suggest to center the basis functions by their mean, i.e. we use $\tilde{B}_{jlk}(x_{i1}, x_{i2}) = B_{jlk}(x_{i1}, x_{i2}) - \bar{B}_{jlk}$, where $\bar{B}_{jlk} = \frac{1}{n}\sum_{i=1}^{n} B_{jlk}(x_{i1}, x_{i2})$.

We analyze the performance of L2KnotSmooth for surface fitting in a small simulation study. The following settings were investigated:

- In **setting IV**, we used a function which has previously been considered by Smith and Kohn (1997),

  $$g_1(x_1, x_2) = x_1 \sin(4\pi x_2), \tag{15}$$

  see Figure 7 (left panel). It represents a nonlinear interaction that shows very different partial derivatives in the direction of x_1 and x_2. The covariates $\mathbf{x}_i = (x_{i1}, x_{i2})'$ were drawn independently from a uniform distribution on $[0, 1] \times [0, 1]$, which was also the case for settings V and VII below. In accordance with Smith and Kohn (1997), $n = 300$ observations were drawn and $\sigma(x_1, x_2)$ was set constant to 0.5.
- Since heteroskedasticity of the error was of special interest in the unidimensional case, in **setting V** it was also considered for surface fitting. Therefore, we followed Crainiceanu et al. (2004) and investigated the function given in (15), with the same sample size as in setting IV and a standard deviation function given by

  $$\sigma(x_1, x_2) = \frac{1}{16} + \frac{3}{16}x_1^2.$$
- **Setting VI** aims on assessing the performance of bivariate smoothers when the covariates are correlated. The regression function (15) was applied, and the $\mathbf{x}_i$ were generated from a bivariate normal distribution with mean $(0.5, 0.5)'$, variance 0.1 and a correlation of 0.5. We used the same error distribution as in setting IV.
- To examine the ability of smoothers to detect surfaces with complex struc-

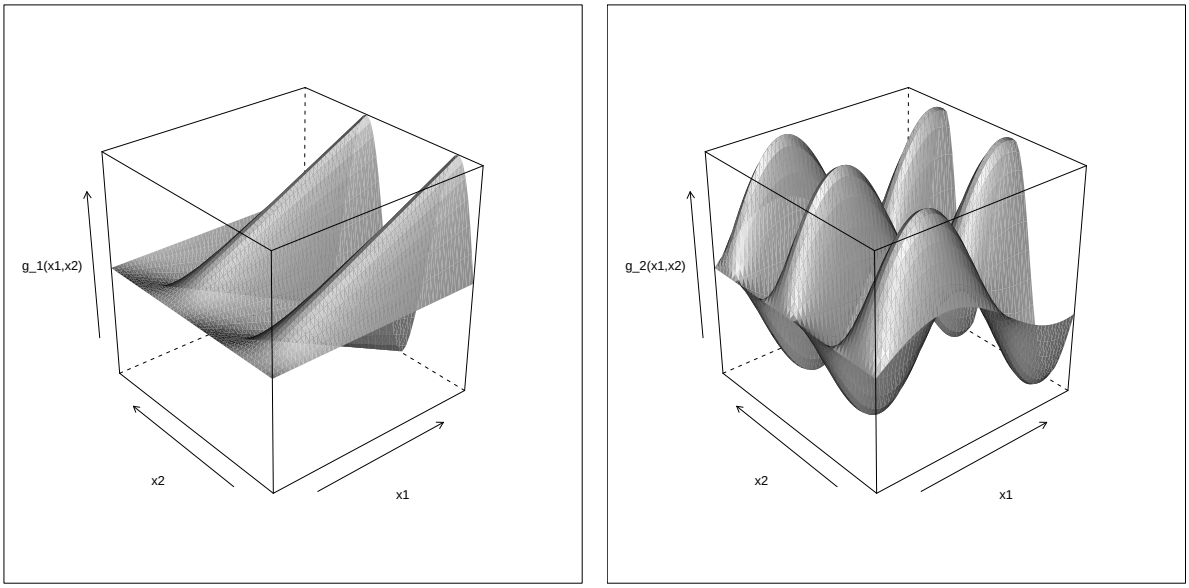

Figure 7. Plots of the surface functions used in the simulation study. Left panel: $g_1(.,.)$ (setting IV and V), right panel: $g_2(.,.)$ (setting VI).

tures, in **setting VII** the function

$$g_2(x_1, x_2) = \sin(5x_1)\cos(5(x_2+1)^2),$$

displayed in the right panel of Figure 7, was investigated. The standard deviation was set constant to $\sigma(x_1, x_2) \equiv 0.2$, and $n = 400$ observations were drawn.

We used an initial number of $m = \max\{20, \min\{\lfloor n/4, 150\rfloor\}\}$ basis functions for L2KnotSmooth, as suggested by Ruppert et al. (2003). The placement of the knots is not as simple as in the univariate case. One might consider to use marginal sample quantiles, but this might yield undesirable results when the design is not uniform across the unit square. Instead, we followed Crainiceanu et al. (2006), who use the function `cover.design()` from the R package `fields` (Nychka, 2005). This approach finds the initial knot mesh by minimizing a geometric space-filling criterion (for details, see also Johnson et al. (1990)). At each knot, $r = 13$ radial basis functions were allocated. The spreads h_k ranged from 0.02 to 0.4 (from 0.02 to 0.1 with step size 0.02, from 0.1 to 0.4 with step size 0.04). Experiments with other ranges (e.g. higher values for h) changed the results only marginally. For the ridge parameter we

chose $\lambda = 10$, since when higher values of λ were chosen, especially the AIC_c-criterion did not reach a distinct minimum after 1000 boosting iterations for many data sets.

We compared our approach with the two-dimensional version of penalized thin plate regression splines implemented in the library `mgcv` and with MARS, which is well suited for detecting interactions in the data, see Section 3.2. Furthermore, the local fitting method called "locfit" (Loader (1999), see also Cleveland and Grosse (1991)) was applied, which has also been considered e.g. by Smith and Kohn (1997) in the context of surface smoothing. It performs a local quadratic fit, where the nearest neighbor smoothing parameter has been chosen by GCV. An implementation can be found in the R library `locfit` (Loader, 2006).

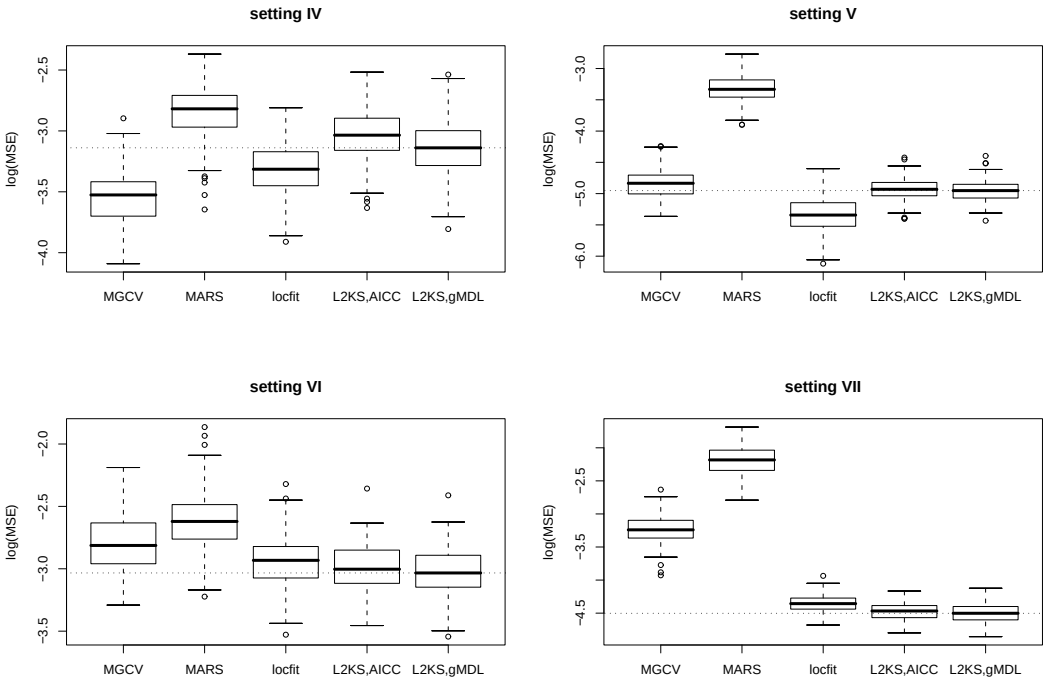

Figure 8. Surface fitting, boxplots of log(MSE) for various fitting methods, setting IV (upper left panel), setting V (upper right panel), setting VI (lower left panel) and setting VII (lower right panel). The dotted lines represent the median of gMDL-stopped L2KnotSmooth.

The results of the simulation study are presented in the boxplots of Figure 8, where log(MSE), with $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_i - \mu_i)^2$, is given. From the upper left panel it is seen that in setting IV MGCV is the strongest competitor and outperforms the other approaches. L2KnotSmooth does distinctively better than MARS, however. Interestingly, gMDL dominates the AIC_c criterion clearly in this case, but has inferior performance compared to locfit. In the heteroskedastic setting (setting V, upper right panel), locfit yields the best results. The two boosting alternatives show similar behavior in terms of the median and

outperform MGCV. MARS yields noticeably worse results. It should be noted that for 33 data sets, the AIC_c criterion did not reach a distinct minimum within the maximum number of 1000 iterations. In contrast, this phenomenon never occurred for gMDL-stopped L2KnotSmooth. In the case of a correlated design (setting VI), the proposed boosting approach works well. MGCV and MARS are distinctively outperformed, whereas locfit is more competitive. The results for setting VII are given in the lower right panel of Figure 8. It is seen that L2KnotSmooth performs best in fitting the complex surface. Note that also here AIC_c-stopped boosting exceeded the maximum number of iterations for 2 data sets, whereas gMDL never did.

The Noshiro example is taken from Ruppert (1997). In Noshiro (Japan) a major earthquake took place, where much of the damage was caused by soil movement activated by the quake. Since the slope of the land is supposed to be an important factor for soil movement, it is of interest to estimate it from land survey data. The data set used here, which has also been investigated by Crainiceanu et al. (2004), consists of $n = 799$ observations, where the independent variables are longitude and latitude, and the response is elevation, measured at several locations in Noshiro.

In Figure 9, the fitted surface and the corresponding contour plot of a gMDL-stopped L2KnotSmooth fit to the Noshiro data is shown. The same settings as in the simulation study were used, and boosting stopped after $l_{\text{opt}} = 81$ iterations. We concentrate on gMDL-stopped boosting, since the AIC_c criterion did not achieve a distinct minimum after 1000 iterations. The fitted surface shows a rather sharp peak at a longitude of 0.4 and a latitude 0.42 (after scaling the covariates to the unit interval), and seems much smoother at the boundary, which indicates the presence of spatial variation. Additionally, Crainiceanu et al. (2004) reported severe heteroskedasticity in the data. Thus, L2KnotSmooth is supposed to show good performance for this type of data. To emphasize this prospect, we compared our approach with MGCV and MARS in terms of gMDL. The gMDL-stopped boosting algorithm shows the best performance (gMDL $= -0.019$), clearly outperforming MGCV (0.362) and MARS (0.732).

5 Conclusion

A knot selection technique is introduced that is based on componentwise L_2-boosting. It shows high flexibility, especially in the case of spatial variation, and robustness against heteroskedastic errors. Simulations and examples show that it is a strong competitor to other knot selection approaches. The method can be easily extended to the fitting of surfaces of higher dimensions. An extension to additive models may be done straightforwardly.

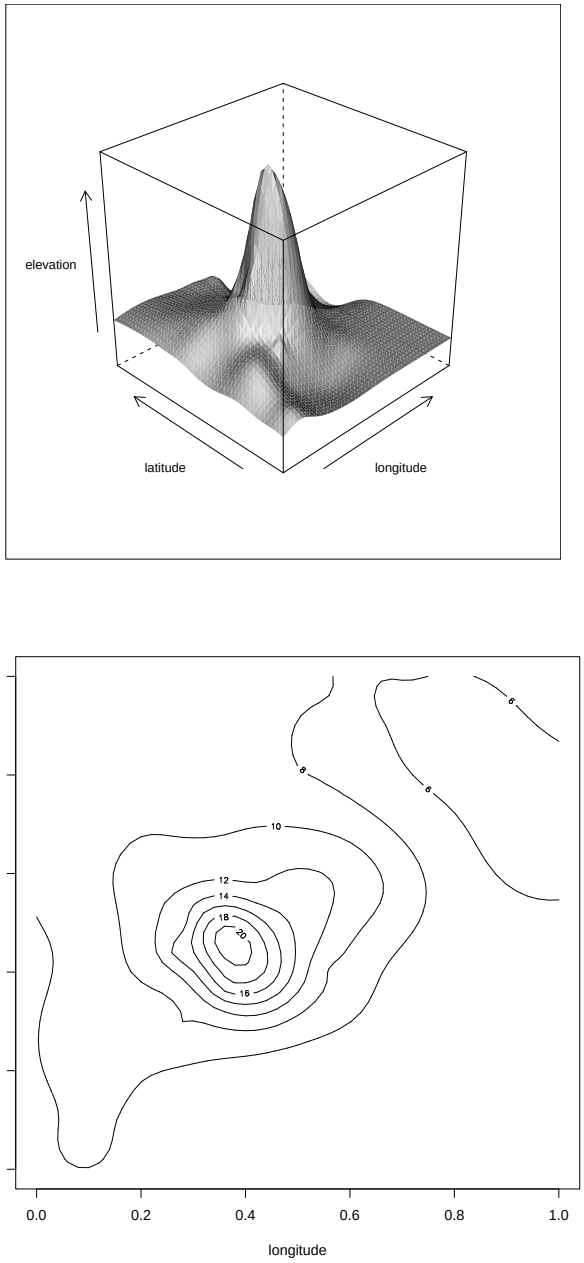

Figure 9. Noshiro data: Fit with gMDL-stopped L2KnotSmooth, surface (upper panel) and contour plot (lower panel)

Since boosting can be seen as a very general optimization technique in function space (e.g Friedman, 2001), our approach might also be extended to models with non-Gaussian response. The basic concept is to consider models of the type $E(y_i|x_i) = h(f(x_i))$ where h is a known link function and $f(x_i)$ is the unknown predictor which is parameterized in basis functions. More general one assumes that $y_i|x_i$ follows a simple exponential family, including binary and Poisson responses. The L_2 loss function used in L2KnotSmooth algorithm has to be replaced by the corresponding log-likelihood. Such approaches to likelihood based boosting can be found e.g. in Ridgeway (1999) or Tutz and

Binder (2006).

Acknowledgements

We gratefully acknowledge support from the Deutsche Forschungsgemeinschaft (SFB 386, "Statistical Analysis of Discrete Structures"). We thank David Ruppert for letting us use the LIDAR and Noshiro data and Berwin A. Turlach for helpful remarks.

References

Bühlmann, P., 2006. Boosting for high-dimensional linear models. Annals of Statistics 34, 559 –583.

Bühlmann, P., Yu, B., 2003. Boosting with the L_2-loss: regression and classification. Journal of the American Statistical Association 98, 324–339.

Bühlmann, P., Yu, B., 2006. Sparse boosting. Journal of Machine Learning Research 7, 1001–1024.

Cleveland, W. S., Grosse, E., 1991. Computational methods for local regression. Statist. Comput. 1, 47–62.

Crainiceanu, C. M., Diggle, P. J., Rowlingson, B., 2006. Bivariate binomial spatial modelling Loa Loa prevalence in tropical Africa. Working Paper 103, John Hopkins University, Dept. of Biostatistics.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., 2004. Spatially adaptive bayesian P-splines with heteroskedastic errors. Working Paper 61, John Hopkins University, Dept. of Biostatistics.

De Boor, C., 1978. A Practical Guide to Splines. Springer-Verlag, New York.

Denison, D., Mallick, B., Smith, A., 1998. Automatic Bayesian curve fitting. Journal of the Royal Statistical Society B60, 333–350.

Donoho, D., Johnstone, I., 1994. Ideal spatial adaptation via wavelet shrinkage. Biometrika 81, 425–455.

Eilers, P. H. C., Marx, B. D., 1996. Flexible smoothing with B-splines and Penalties. Statistical Science 11, 89–121.

Eubank, R. L., 1988. Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and Its Applications. Chapman & Hall, London.

Friedman, J., 1991. Multivariate adaptive regression splines (with discussion). Ann. Statist. 19, 1–141.

Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics 29, 1189–1232.

Green, D. J., Silverman, B. W., 1994. Nonparametric Regression and Gener-

alized Linear Models: A Roughness Penalty Approach. Chapman & Hall, London.

Hansen, M., Yu, B., 2001. Model selection and minimum description length principle. Journal of the American Statistical Association 96, 746–774.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

He, X., Ng, P., 1999. COBS: Qualitatively constrained smoothing via linear programming. Computational Statistics 14, 315–337.

He, X., Shi, P., 1994. Convergence rate of b-splines estimators of nonparametric conditional quantile functions. Journal of Nonparametric Statistics 3, 299–308.

Hoerl, A. E., Kennard, R. W., 1970. Ridge regression: Bias estimation for nonorthogonal problems. Technometrics 12, 55–67.

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., Edner, H., 1996. Locally weighted least squares kernel regression and statistical evaluation of lidar measurements. Environmetrics 7, 401–416.

Härdle, W., 1984. How to determine the bandwidth of some nonlinear smoothers in practice. In: Robust and Nonlinear Time Series Analysis. Vol. 26 of Lecture Notes in Statistics. Wiley, New York, pp. 163–184.

Hurvich, C. M., Simonoff, J. S., 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.

Johnson, M. E., Moore, L. M., Ylvisaker, D., 1990. Minimax and maximin distance designs. Journal of Statistical Planning and Inference 26, 131–148.

Koenker, R., Ng, P., Portnoy, S., 1994. Quantile smoothing splines. Biometrika 81, 673–680.

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

Leung, D. H.-Y., 2005. Cross-validation in nonparametric regression with outliers. Annals of Statistics 33, 2291–2310.

Loader, C., 1999. Local Regression and Likelihood. Springer-Verlag, New York.

Loader, C., 2006. locfit: Local Regression, Likelihood and Density Estimation. R package version 1.5-3.

Lokhorst, J., 1999. The lasso and generalised linear models. Honors project, Department of Statistics, University of Adelaide.

Nychka, D., 2005. fields: Tools for spatial data. R package version 2.3.

Osborne, M. R., Presnell, B., Turlach, B. A., 1998. Knot selection for regression splines via the lasso. In: Weisberg, S. (Ed.), Dimension Reduction, Computational Complexity, and Information. Vol. 30 of Computing Science and Statistics. Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460, pp. 44–49.

R Development Core Team, 2006. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ridgeway, G., 1999. The state of boosting. In: Berk, K., Pourahmadi, M.

(Eds.), Models, predictions and computing. Vol. 31 of Computing Science and Statistics. Interface Foundation of North America, Inc., Fairfax Station, VA 22039-7460, pp. 172–181.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. University Press, Cambridge.

Ruppert, D., 1997. Local polynomial regression and its applications in environmental statistics. In: Barnett, V., Turkman, F. (Eds.), Statistics for the Environment. Vol. 3. John Wiley, Chicester, pp. 155–173.

Ruppert, D., Carroll, R. J., 2000. Spatially-adaptive penalties for spline fitting. Australian & New Zealand Journal of Statistics 42, 205–223.

Ruppert, D., Wand, M. P., Carroll, R. J., 2003. Semiparametric Regression. Cambridge University Press, Cambridge.

Schapire, R. E., 1990. The strength of weak learnability. Machine Learning 5, 197–227.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Smith, M., Kohn, R., 1996. Nonparametric regression using Bayesian variable selection. Journal of Econometrics 75, 317–343.

Smith, M., Kohn, R., 1997. A Bayesian approach to nonparametric bivariate regression. Journal of the American Statistical Association 92, 1522–1535.

Smith, P. L., 1982. Curve fitting and modeling with splines using statistical variable selection techniques. Report NASA 166034, NASA, Langley Research center, Hampton.

Stone, C., Hansen, M., Kooperberg, C., Truong, Y., 1997. Polynomial splines and their tensor products in extended linear modeling. The 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., 2005. Boosting ridge regression. SFB Discussion Paper 418, LMU München.

Tutz, G., Binder, H., 2006. Generalized additive modelling with implicit variable selection by likelihood based boosting. Biometrics (to appear).

Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia.

Wand, M. P., 2000. A comparison of regression spline smoothing procedures. Computational Statistics 15, 443–462.

Wood, S. N., 2003. Thin plate regression splines. Journal of the Royal Statistical Society B 65, 95–114.

Wood, S. N., 2006. Generalized Additive Models: An Introduction with R. Chapman&Hall/CRC, London.