Spline-based model specification and prediction for least squares and quantile regression

Dissertation zur Erlangung des Grades eines

Doktors der Wirtschaftswissenschaft

eingereicht an der Wirtschaftswissenschaftlichen

Fakultät der Universität Regensburg

vorgelegt von: Kathrin Kagerer

Berichterstatter:

Prof. Dr. Rolf Tschernig, Universität Regensburg

Prof. Dr. Harry Haupt, Universität Bielefeld

Tag der Disputation: 16. März 2012

Contents

List of Figures

List of Tables

1 Introduction and overview

1.1 Introduction and methodological background

> *Systema [...] maxime probabile valorum incognitarum [...] id erit, in quo quadrata differentiarum inter [...] valores observatos et computatos summam minimam efficiunt.*

> [...] that will be the most probable system of values of the unknown quantities [...] in which the sum of the squares of the differences between the observed and computed values [...] is a minimum.

This insight still is of crucial interest more than 200 years after Carl Friedrich Gauss stated it in 1809 (Gauss, 1809, p. 245, translation from Davis, 1857, p. 260). The method of least squares is historically used to describe the course of planets by Gauss (1809) and Legendre (1805) who independently suggested the same method. Today there are many more fields that apply the method of least squares in regression analysis. Among these are geography, biology/medicine and economics.

Years before Gauss' and Legendre's method of least squares, Laplace suggested to minimize the sum of absolute differences between observed and calculated observations (Laplace, 1789). Portnoy & Koenker (1997) provide a historical review and a detailed comparison of both methods. Minimizing the sum of absolute differences, as introduced by Laplace, is a special case of the method which is nowadays known as quantile regression.

1.1.1 Classical linear least squares regression analysis

Classical linear least squares regression can be applied to quantify the change of the expected outcome of the response y given x_1 and potential other factors $x_2, \ldots, x_q$ when x_1 varies by some amount while the other covariates $x_2, \ldots, x_q$ are held fixed. Accordingly, the expected value of y given the covariates $x_1, \ldots, x_q$, $\mathrm{E}(y|x_1, \ldots, x_q)$, is a function of the covariates, that is, it can be expressed as

$$\mathrm{E}(y|x_1, \ldots, x_q) = f(x_1, \ldots, x_q), \tag{1.1}$$

where f is a (unknown) function that describes the relationship between $\mathrm{E}(y|x_1, \ldots, x_q)$ and the covariates $x_1, \ldots, x_q$. Hence, the relationship between y and $f(x_1, \ldots, x_q)$ is given by

$$y = f(x_1, \ldots, x_q) + u, \tag{1.2}$$

where $\mathrm{E}(u|x_1, \ldots, x_q) = 0$.

An often applied choice when estimating the function f, is to assume a linear relationship between $\mathrm{E}(y|x_1, \ldots, x_q)$ and the covariates $x_1, \ldots, x_q$. That is, the functional form f is a linear combination of the covariates,

$$f(x_1, \ldots, x_q) = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q, \tag{1.3}$$

where $\beta_0, \ldots, \beta_q$ are unknown parameters that need to be estimated. For a given sample $i = 1, \ldots, n$ the parameters may be estimated by solving

$$\min_{\tilde{\beta}_0, \ldots, \tilde{\beta}_q \in \mathbb{R}} \sum_{i=1}^{n} \left(y_i - (\tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \ldots + \tilde{\beta}_q x_q)\right)^2.$$

Hence, the estimates for the parameters $\beta_0, \ldots, \beta_q$ are those that minimize the sum of squared differences between the observed values y_i and the computed values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_q x_{iq}$. That is, the parameters are estimated by applying Gauss' method of least squares.

To allow for other functional forms, $\mathrm{E}(y|x_1, \ldots, x_q)$ could, for example, be expressed as linear combination of higher order polynomials or other transformations of the covariates. This implies a possibly extensive specification search for f. To avoid this,

non- or semiparametric specifications for f can be applied. Spline regression is one of these nonparametric methods. An outline of splines and spline regression is given in Section 1.1.2.

For simplification, consider a bivariate relationship between one covariate x and the response y. Then, Equation (1.1) simplifies to

$$\mathrm{E}(y|x) = f(x). \tag{1.4}$$

If f is incorrectly specified for the estimation, several conclusions (e.g. interpretations of marginal effects and hypothesis tests) that build on a correctly specified model are invalid. Hence, a correct specification of f is crucial.

1.1.2 Splines

Basis functions

Truncated power basis and splines Piecewise polynomial functions constitute an easy approach to adopt a flexible functional form without being demanding to implement. These functions can be generated in different ways. An intuitive way to understand the main principle is to consider the truncated power basis (see for example Ruppert et al., 2003, ch. 3, Dierckx, 1993, ch. 1.1, de Boor, 2001, ch. VIII as general references). Consider a straight line on some interval $[\kappa_0, \kappa_{m+1}]$ (e.g. $[\kappa_0, \kappa_{m+1}] = [\min_i(x_i), \max_i(x_i)]$ for a sample $i = 1, \ldots, n$) with a kink at some position κ_1 where $\kappa_0 < \kappa_1 < \kappa_{m+1}$. It can be described by a weighted sum (i.e. a linear combination) of the basis functions 1, x and $(x - \kappa_1)_+$, where the truncation function

$$(x - \kappa_1)_+ = \begin{cases} x - \kappa_1 & \text{for } x \geq \kappa_1 \\ 0 & \text{else} \end{cases}$$

gives the positive part of $x - \kappa_1$. The reasoning for the basis functions 1, x and $(x - \kappa_1)_+$ is as follows: To obtain a straight line f that is folded at κ_1 but continuous there, this function f can be written as $\beta_0 + \beta_1 x$ for $x < \kappa_1$ and as $\beta_0' + (\beta_1 + \alpha_1) x$ for $x \geq \kappa_1$. That means, the slope is β_1 until $x = \kappa_1$ and from $x = \kappa_1$ on the

slope is changed by α_1. As f is constrained to be continuous at κ_1, this requires $\beta_0 + \beta_1 \kappa_1 = \beta_0' + (\beta_1 + \alpha_1) \kappa_1$ to hold or equivalently $\beta_0' = \beta_0 - \alpha_1 \kappa_1$. Overall, f then is given by

$$\begin{aligned} f(x) &= (\beta_0 + \beta_1 x) \cdot I_{\{x<\kappa_1\}} + \left(\beta_0' + (\beta_1 + \alpha_1) x\right) \cdot I_{\{x \geq \kappa_1\}} \\ &= (\beta_0 + \beta_1 x) \cdot I_{\{x<\kappa_1\}} + (\beta_0 + \beta_1 x + \alpha_1 (x - \kappa_1)) \cdot I_{\{x \geq \kappa_1\}} \\ &= \beta_0 + \beta_1 x + \alpha_1 (x - \kappa_1) I_{\{x \geq \kappa_1\}} \\ &= \beta_0 + \beta_1 x + \alpha_1 (x - \kappa_1)_+ \end{aligned}$$

where $I_{\{A\}}$ is the indicator function which is 1 if A holds and 0 else. That is, f can be written as a linear combination of the basis functions 1, x and $(x - \kappa_1)_+$.

Note that linear least squares regression with a constant and one covariate x has the two basis functions 1 and x and that for example an additional quadratic term leads to the additional basis function x^2.

Analogously to the line folded only at κ_1, a line folded m times at the positions $\kappa_1, \ldots, \kappa_m$ can be written as the weighted sum of the basis functions 1, x, $(x - \kappa_1)_+, \ldots, (x - \kappa_m)_+$, where the κ_j's are called knots. With a proper choice of the knots κ_j, the functions generated from this basis can approximate other functions rather well. However, it yields a curve with sharp kinks that is not differentiable at the knots. These sharp kinks as well as the lacking differentiability are often undesirable. Smoother curves can be obtained by using higher powers of the basis functions. The respective basis of degree $p \geq 0$ consists of the functions $1, x, \ldots, x^p$, $(x - \kappa_1)_+^p, \ldots, (x - \kappa_m)_+^p$, where $(x - \kappa_j)_+^p := \left((x - \kappa_j)_+\right)^p$ and $0^0 := 0$, and is called truncated power basis of degree p. The truncated power function $(x - \kappa_j)_+^p$ is $(p-1)$-times continuously differentiable at κ_j. Hence, also the linear combinations (called splines, Ruppert et al., 2003, p. 62) of the truncated power basis functions are $(p-1)$-times continuously differentiable at the knots κ_j, $j = 1, \ldots, m$.

Additionally including lower powers ($< p$) of $(x - \kappa_j)_+$ in the basis, changes the differentiability properties. That is, if the basis consists for example of the functions $1, x, \ldots, x^p$, $(x - \kappa_1)_+^p, \ldots, (x - \kappa_1)_+^{p-m_1+1}$, $(x - \kappa_2)_+^p, \ldots, (x - \kappa_m)_+^p$, the resulting linear combinations are $(p - m_1)$-times continuously differentiable at κ_1 and $(p-1)$-times continuously differentiable at $\kappa_2, \ldots, \kappa_m$, where m_1 can be regarded as multiplic-

ity of the knot κ_1 (e.g. Eubank, 1984, p. 447f.). For $m_j = p+1$, functions constructed as linear combination from the truncated power basis functions of degree p have a discontinuity/jump at κ_j. If $m_j > 1$, the respective knots and basis functions are denoted as $(x-\kappa_j)_+^p, \ldots, (x-\kappa_{j+m_j-1})_+^{p-m_1+1}$, where $\kappa_j = \ldots = \kappa_{j-m_j+1}$. That is, the knot κ_j has multiplicity m_j (and so have $\kappa_{j+1}, \ldots, \kappa_{j-m_j+1}$ where $m_j = \ldots = m_{j-m_j+1}$). This notation is consistent with the notation for the equivalent B-spline basis which is described later on. Further, the truncated power basis of degree p thus always consists of $p+1+m$ basis functions that are identified once the knot sequence is given.

The panels in the upper row of Figure 1.1 show the functions of the above described bases: the basis for linear functions $(1,\ x)$, the basis for cubic functions $(1,\ x,\ x^2,\ x^3)$, the truncated power basis of degree 1 with one knot at 0.4 $(1,\ x,\ (x-0.4)_+)$, the truncated power basis of degree 1 with four knots at 0.2, ..., 0.8 $(1,\ x,\ (x-0.2)_+$, ..., $(x-0.8)_+)$ and the truncated power basis of degree 3 with four knots at 0.2, ..., 0.8 $(1,\ x,\ x^2,\ x^3,\ (x-0.2)_+^3,\ \ldots,\ (x-0.8)_+^3)$. Additionally, the lower row shows an arbitrary example for a linear combination of the basis functions for each of the illustrated bases.

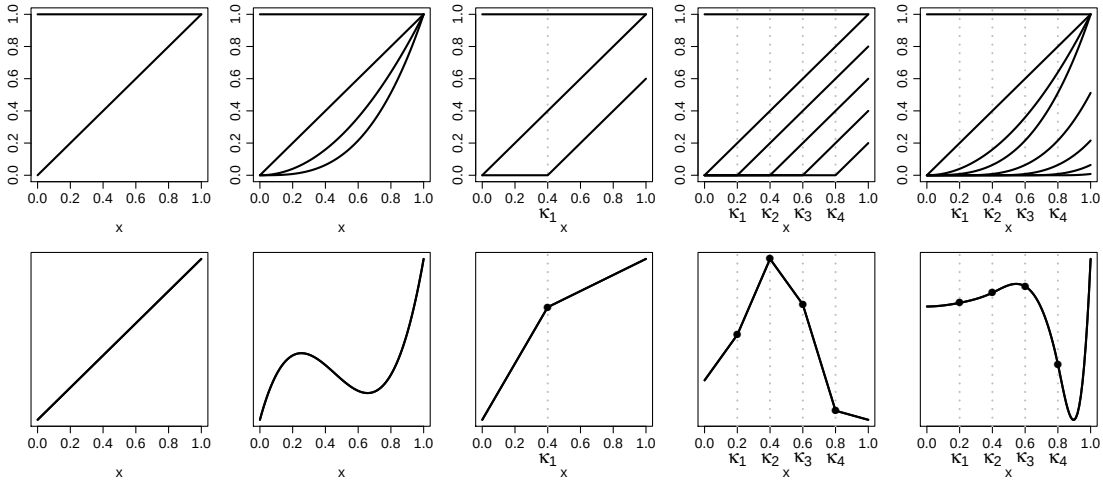

Figure 1.1: Top: different bases (from the left): basis functions for straight line, for cubic polynomial, for straight line with one kink, for straight line with four kinks, for cubic polynomial with four kinks. **Bottom**: arbitrary examples for linear combinations of the basis functions from the corresponding upper panel.

A disadvantage of the truncated power basis is that the basis functions are correlated and hence estimation results are often numerically instable (Ruppert et al., 2003, p. 70).

An equivalent basis that does not have this problem (Dierckx, 1993, p. 5, Ruppert et al., 2003, p. 70, de Boor, 2001, p. 85f.) and leads (apart from computational accuracy) to the same fit on $[\kappa_0, \kappa_{m+1}]$ (Ruppert et al., 2003, p. 70) is the B-spline basis of order $k = p + 1 \geq 1$ where p is the degree of the truncated power basis. In the following, B-splines are introduced.

B-spline basis and splines In a nutshell, the functions from the B-spline basis are piecewise polynomial functions of order k that are connected at the knots and have only small support. Then the spline of order k, which is a linear combination of the basis functions, is also a piecewise polynomial function of order k. It exhibits the same properties as the respective linear combination of the functions from the truncated power basis of degree $p = k - 1$ with the same knots $\kappa_1, \ldots, \kappa_m$. A short review concerning B-splines can be found e.g. in the work of Eilers & Marx (1996) who summarize the definition and properties of B-splines while de Boor (2001), Dierckx (1993) and Ruppert et al. (2003) provide a more extensive discussion of splines.

To derive the B-spline basis of order k, some definitions are necessary. Let $\boldsymbol{\kappa} = (\kappa_{-(k-1)}, \ldots, \kappa_{m+k})$ be a non-decreasing sequence of knots (i.e. $\kappa_{-(k-1)} \leq \ldots \leq \kappa_{m+k}$), where at most k adjacent knots coincide (i.e. $\kappa_j \neq \kappa_{j+k}$). The two boundary knots κ_0 and κ_{m+1} define the interval of interest and the m knots $\kappa_1, \ldots, \kappa_m$ are called inner knots. The remaining $2\,(k-1)$ exterior knots $\kappa_{-(k-1)}, \ldots, \kappa_{-1}$ and $\kappa_{m+2}, \ldots, \kappa_{m+k}$ are required to ensure regular behavior on the interval $[\kappa_0, \kappa_{m+1}]$. The B-spline basis functions (also called B-splines) are denoted as $B_j^{\boldsymbol{\kappa},k}$, $j = -(k-1), \ldots, m$.

B-splines can be motivated in different ways. One of them is to derive them by using divided differences (for a definition of divided differences and the corresponding representation of B-splines see for example de Boor, 2001, p. 3, 87). Another way that does not involve the definition of divided differences and can be shown to lead to the same result (cf. de Boor, 2001, p. 88) is to recursively calculate the B-splines of order $k > 1$ from the B-splines of lower order using the recurrence relation from de Boor (2001, p. 90)

$$B_j^{\boldsymbol{\kappa},k}(x) = \frac{x - \kappa_j}{\kappa_{j+k-1} - \kappa_j} B_j^{\boldsymbol{\kappa},k-1}(x) \; - \; \frac{x - \kappa_{j+k}}{\kappa_{j+k} - \kappa_{j+1}} B_{j+1}^{\boldsymbol{\kappa},k-1}(x),$$

where

$$B_j^{\boldsymbol{\kappa},1}(x) = (\kappa_{j+1} - x)_+^0 - (\kappa_j - x)_+^0 = \begin{cases} 1 & \text{for } \kappa_j \leq x < \kappa_{j+1} \\ 0 & \text{else} \end{cases}$$

is the B-spline of order 1 (de Boor, 2001, p. 89) and the index j runs from $-(k-1)$ to m. To obtain the properties of B-splines on the complete interval $[\kappa_0, \kappa_{m+1}]$ but not only on $[\kappa_0, \kappa_{m+1})$, the definition of $B_m^{\boldsymbol{\kappa},1}$ and $B_{m+1}^{\boldsymbol{\kappa},1}$ is modified such that $B_m^{\boldsymbol{\kappa},1}(x) = 1$ for $x = \kappa_{m+1}$ and $B_{m+1}^{\boldsymbol{\kappa},1}(x) = 0$ for $x = \kappa_{m+1}$ (cf. de Boor, 2001, p. 94).

For equidistant knots, that is $\Delta\kappa_j = \kappa_j - \kappa_{j-1} =: h$ for $j = -(k-1)+1, \ldots, m+k$, it can be shown (based on an extension of Problem 2 from de Boor, 2001, p. 106) that the function $B_j^{\boldsymbol{\kappa},k}$ can also be written as

$$B_j^{\boldsymbol{\kappa},k}(x) = \frac{1}{h^{k-1}(k-1)!}\,\Delta^k(\kappa_{j+k} - x)_+^{k-1}$$

where $\Delta^k := \Delta(\Delta^{k-1})$.

Figure 1.2 gives examples for B-spline bases of different orders k for an equidistant knot sequence $\boldsymbol{\kappa}$ with $m = 3$ inner knots. In the first row, the basis functions $B_j^{\boldsymbol{\kappa},k}$, $j = -(k-1), \ldots, m$, and their sum are shown. Further details of Figure 1.2 are given in the following when the respective features are explained.

B-splines have several useful properties. Firstly, they form a partition of unity on the interval $[\kappa_0, \kappa_{m+1}]$, that is, on $[\kappa_0, \kappa_{m+1}]$ it holds that

$$\sum_{j=-(k-1)}^{m} B_j^{\boldsymbol{\kappa},k}(x) = 1 \tag{1.5}$$

(de Boor, 2001, p. 96). This can be observed in the first row of Figure 1.2. Moreover, each of the basis functions has only small support. More precisely, the support of the function $B_j^{\boldsymbol{\kappa},k}$ is the interval (κ_j, κ_{j+k}) (de Boor, 2001, p. 91), hence $B_j^{\boldsymbol{\kappa},k}(x) \cdot B_{j+d}^{\boldsymbol{\kappa},k}(x)$ is zero for $|d| \geq k$, which is the reason for the numerical stability mentioned on page 6. Further, the B-splines are up to $(k - m_j - 1)$-times continuously differentiable at the knot κ_j and the $(k - m_j)$th derivative has a jump at κ_j where m_j is the multiplicity of the knot κ_j (e.g. $\kappa_j = \ldots, \kappa_{j+m_j-1}$) (Dierckx, 1993, p. 9, de Boor, 2001, p. 99). This property carries over to linear combinations of the basis functions (called spline,

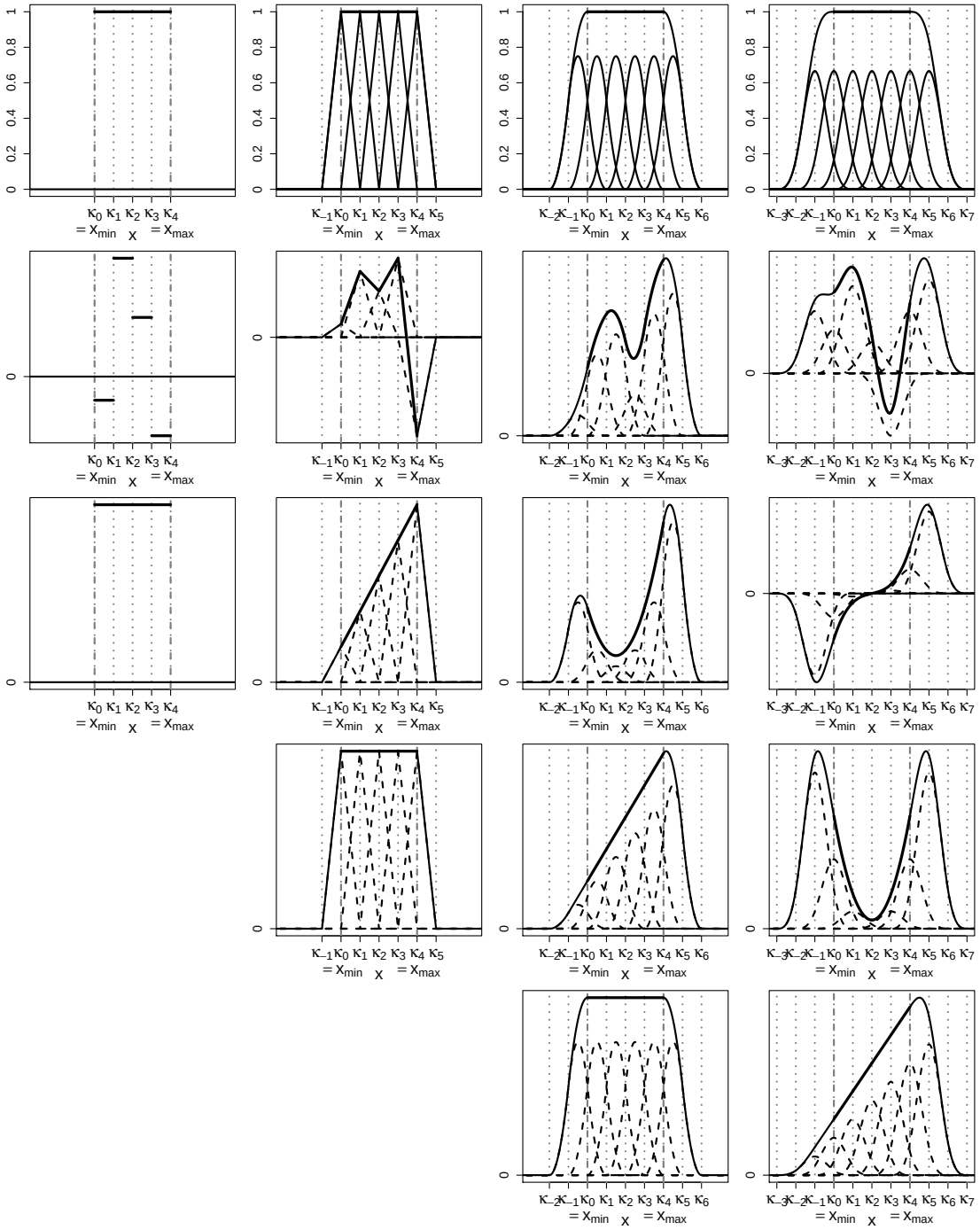

Figure 1.2: B-spline bases, sums and linear combinations for different orders k with equidistant knots and $m = 3$. **First row**: B-spline basis functions of orders 1, 2, 3 and 4, and their sum, which is 1 on $[\kappa_0, \kappa_{m+1}]$. **Second row**: arbitrarily weighted B-spline basis functions of orders 1, 2, 3 and 4, and their sum. **Third row**: B-spline basis functions of order 1, 2, 3 and 4, weighted such that their sum is a polynomial of degree $k-1$ on $[\kappa_0, \kappa_{m+1}]$. **Fourth row**: B-spline basis functions of orders 2, 3 and 4, weighted such that their sum is a polynomial of degree $k-2$ on $[\kappa_0, \kappa_{m+1}]$. **Fifth row**: B-spline basis functions of orders 2, 3 and 4, weighted such that their sum is a polynomial of degree $k-3$ on $[\kappa_0, \kappa_{m+1}]$.

de Boor, 2001, p. 93), that is, the functions

$$B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x) = \sum_{j=-(k-1)}^{m} \alpha_j B_j^{\boldsymbol{\kappa},k}(x), \tag{1.6}$$

where $\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{-(k-1)} & \dots & \alpha_m \end{pmatrix}'$, are also $(k - m_j - 1)$-times continuously differentiable at the knot κ_j. The second row of Figure 1.2 shows examples for linear combinations of the basis functions from the first row with arbitrarily chosen α_j.

The linear combinations of the B-spline basis functions of order k can generate all polynomial functions (in contrast to piecewise polynomial functions) of degree smaller than k on the interval $[\kappa_0, \kappa_{m+1}]$. Note that this is also a spline/polynomial of order k. This justifies the use of the notation order instead of degree since all polynomials of degree $< k$ are polynomials of order k. In the third (fourth, fifth) row of Figure 1.2, examples for polynomials of degree $k-1$ ($k-2$, $k-3$) are given for $k \geq 1$ ($k \geq 2$, $k \geq 3$).

The first two rows of Figure 1.3 show B-spline bases with $k = 2, 3, 4$ where two knots of the knot sequence $\boldsymbol{\kappa}$ coincide. In the first row, $\alpha_j = 1$ for all j, $j = -(k-1), \dots, m$, and in the second row, the α_j are chosen arbitrarily. For $k = 2$ the resulting spline now has a jump where the double knot is placed. For $k = 3$ the spline is still continuous, but its derivative is not, and for $k = 4$ the first derivative is also continuous, but the second derivative is not. For $k = 1$ no graphic is presented since twofold knots with $\kappa_j = \kappa_{j+1}$ do not make sense in this case, because $B_j^{\boldsymbol{\kappa},1}$ would be zero and could be excluded from the basis. The third row of Figure 1.3 shows weighted B-spline bases and the respective resulting splines for $k = 3$ and $k = 4$ where $\kappa_j = \kappa_{j+1} = \kappa_{j+2}$ for some j (i.e. $m_j = 3$). Then the spline of order 3 has a jump at κ_j and the spline of order 4 is continuous. For $k = 2$ one of the threefold knots is meaningless (as discussed for $k = 1$ and twofold knots). In the last line of Figure 1.3, a spline of order 4 is shown that has a jump at a fourfold knot position ($m_j = 4$ there).

To contrast equidistant knot sequences to those with non-equidistant knots, Figure 1.4 exemplarily shows $B_j^{\boldsymbol{\kappa},k}$, $j = -(k-1), \dots, m$, and $\sum_{j=-(k-1)}^{m} B_j^{\boldsymbol{\kappa},k}(x)$ for a non-equidistant knot sequence. This is analogous to the first row of Figure 1.2 with the only difference that the knots are not equidistant and hence the basis functions $B_j^{\boldsymbol{\kappa},k}$ look different. But still they sum up to unity on $[\kappa_0, \kappa_{m+1}]$.

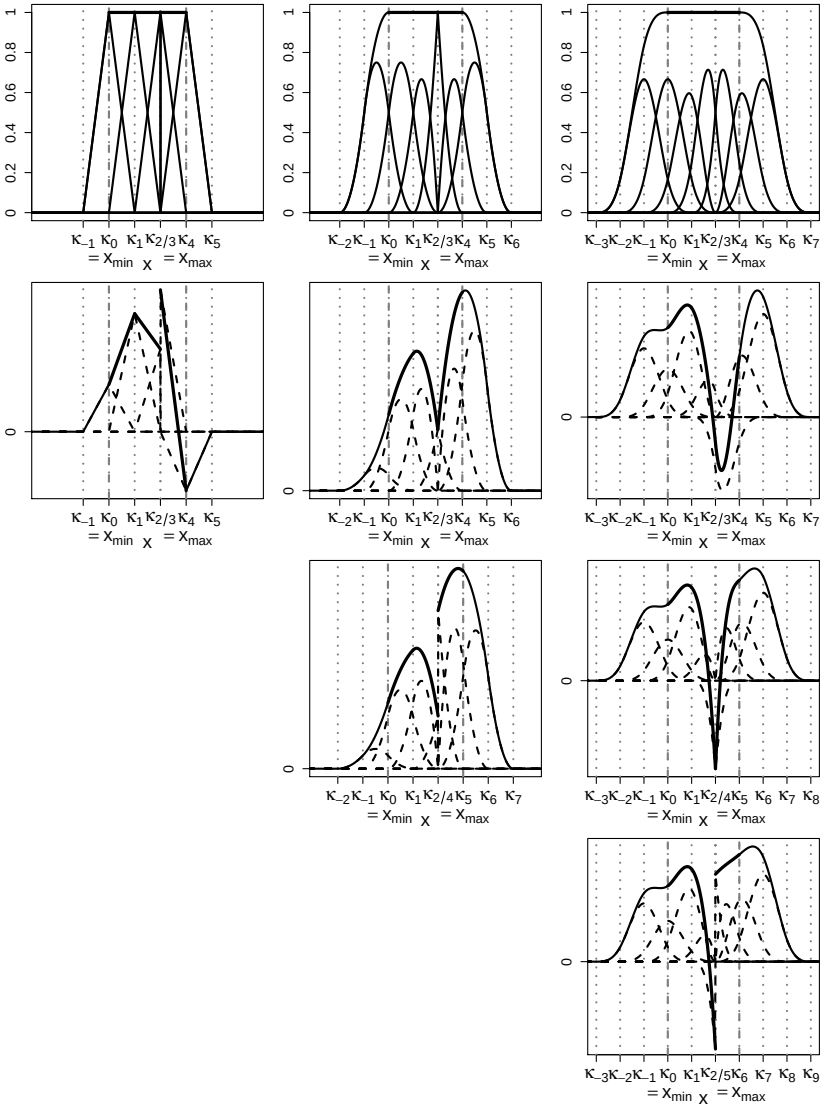

Figure 1.3: B-spline bases, sums and linear combinations for different orders k with equidistant knots where one knot position is multiple occupied. **First row**: B-spline basis functions of orders 2, 3 and 4, where $\kappa_2 = \kappa_3$, and their sum which is 1 on $[\kappa_0, \kappa_{m+1}]$. **Second row**: arbitrarily weighted B-spline basis functions of orders 2, 3 and 4, where $\kappa_2 = \kappa_3$, and their sum. **Third row**: arbitrarily weighted B-spline basis functions of orders 3 and 4, where $\kappa_2 = \kappa_3 = \kappa_4$, and their sum. **Fourth row**: arbitrarily weighted B-spline basis functions of order 4, where $\kappa_2 = \kappa_3 = \kappa_4 = \kappa_5$, and their sum.

Derivative and monotonicity The first derivative of the B-spline functions is

$$\frac{\partial B_j^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{k-1}{\kappa_{j+k-1} - \kappa_j} B_j^{\boldsymbol{\kappa},k-1}(x) - \frac{k-1}{\kappa_{j+k} - \kappa_{j+1}} B_{j+1}^{\boldsymbol{\kappa},k-1}(x) \tag{1.7}$$

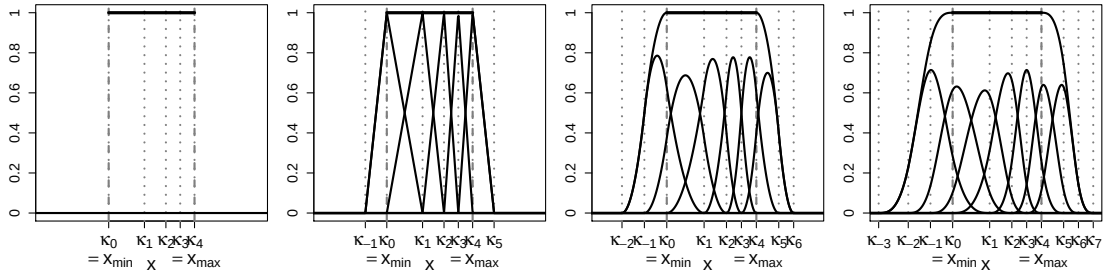

Figure 1.4: B-spline basis functions of orders 1, 2, 3 and 4 with non-equidistant knots ($m = 3$), and their sum which is 1 on $[\kappa_0, \kappa_{m+1}]$.

for $k > 1$ (de Boor, 2001, p. 115). For $k = 1$ it is defined to be 0 according to the argumentation in de Boor (2001, p. 117). Hence, the first derivative of a B-spline of order k is a spline of order $k-1$ since it is a linear combination of B-splines of order $k-1$. From Equation (1.7) it can be shown that the first derivative of a spline as linear combination of the B-spline basis functions is given by

$$\frac{\partial B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{\partial}{\partial x} \sum_{j=-(k-1)}^{m} \alpha_j B_j^{\boldsymbol{\kappa},k}(x) = (k-1) \sum_{j=-(k-1)}^{m+1} \frac{\alpha_j - \alpha_{j-1}}{\kappa_{j+k-1} - \kappa_j} B_j^{\boldsymbol{\kappa},k-1}(x) \quad (1.8)$$

where $\alpha_{-(k-1)-1} := 0 =: \alpha_{m+1}$ (de Boor, 2001, p. 116). On the interval $[\kappa_0, \kappa_{m+1}]$ it holds that $B_{-(k-1)}^{\boldsymbol{\kappa},k-1}(x) = B_{m+1}^{\boldsymbol{\kappa},k-1}(x) = 0$ and hence the summation reduces to

$$\frac{\partial B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x} = (k-1) \sum_{j=-(k-1)+1}^{m} \frac{\alpha_j - \alpha_{j-1}}{\kappa_{j+k-1} - \kappa_j} B_j^{\boldsymbol{\kappa},k-1}(x) \quad (1.9)$$

on $[\kappa_0, \kappa_{m+1}]$. For equidistant knot sequences, Equations (1.7) and (1.8)/(1.9) simplify to

$$\frac{\partial B_j^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{1}{h} B_j^{\boldsymbol{\kappa},k-1}(x) - \frac{1}{h} B_{j+1}^{\boldsymbol{\kappa},k-1}(x)$$

and

$$\frac{\partial B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x} = \frac{1}{h} \sum_{j=-(k-1)}^{m+1} (\alpha_j - \alpha_{j-1}) B_j^{\boldsymbol{\kappa},k-1}(x) = \frac{1}{h} \sum_{j=-(k-1)+1}^{m} (\alpha_j - \alpha_{j-1}) B_j^{\boldsymbol{\kappa},k-1}(x), \quad (1.10)$$

respectively, on $[\kappa_0, \kappa_{m+1}]$. Higher order derivatives can also be calculated from Equations (1.7) or (1.8).

Since all of the terms $k-1$, $\kappa_{j+k-1}-\kappa_j$ and $B_j^{\boldsymbol{\kappa},k-1}(x)$, $j=-(k-1)+1,\ldots,m$ are greater or equal to zero (de Boor, 2001, p. 91), the sign of $\frac{\partial B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x}$ only depends on the differences $\alpha_j-\alpha_{j-1}=:\delta_j$, $j=-(k-1)+1,\ldots,m$. It can be seen from Equation (1.8) that

$$\alpha_j \geq \alpha_{j-1} \quad \text{(i.e. } \delta_j \geq 0\text{)}, \quad j=-(k-1)+1,\ldots,m, \tag{1.11}$$

ensures a completely non-negative first derivative of $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$, and hence $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$ is monotonically increasing. Analogously, $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$ is monotonically decreasing if

$$\alpha_j \leq \alpha_{j-1} \quad \text{(i.e. } \delta_j \leq 0\text{)}, \quad j=-(k-1)+1,\ldots,m. \tag{1.12}$$

If $\alpha_j \geq \alpha_{j-1}$, $j=-(k-1)+1,\ldots,m$, holds (where $\alpha_j \neq 0$ for at least one j), it follows that $\alpha_{-(k-1)} \not\geq \alpha_{-(k-1)-1}$ or $\alpha_{m+1} \not\geq \alpha_m$ since the auxiliary parameters $\alpha_{-(k-1)-1}$ and α_{m+1} from Equation (1.8) are zero. This implies that the spline is only monotonically increasing on the interval $[\kappa_0,\kappa_{m+1}]$. Hence, the derivative (1.9) where the sum is indexed from $j=-(k-1)+1$ to $j=m$ and not (1.8) with $j=-(k+1),\ldots,m+1$ has to be regarded. For $\alpha_j \leq \alpha_{j-1}$ these considerations apply analogously.

Equations (1.11) and (1.12) can also be written in matrix notation as $\mathbf{C}\,\boldsymbol{\alpha} \geq \mathbf{0}$ with

$$\mathbf{C}=\begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & -1 & 1 \\ & & & \ddots \; \ddots \end{pmatrix} \qquad \text{and} \qquad \mathbf{C}=\begin{pmatrix} 1 & -1 & & \\ & 1 & -1 & \\ & & 1 & -1 \\ & & & \ddots \; \ddots \end{pmatrix},$$

respectively.

Trivially, for $k=1$ the conditions (1.11) and (1.12) are each necessary and sufficient conditions for monotonicity. For $k>1$ this issue is illustrated in Figures 1.5 and 1.6 for equidistant knot sequences but the same reasoning holds for non-equidistant knot sequences. The upper rows of both figures show splines $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$ and the underlying basis functions $B_j^{\boldsymbol{\kappa},k}$ weighted by α_j while the lower rows picture the respective derivatives $\frac{1}{h}\sum_{j=-(k-1)}^{m+1}(\alpha_j-\alpha_{j-1})B_j^{\boldsymbol{\kappa},k-1}(x)=\frac{1}{h}\sum_{j=-(k-1)}^{m+1}\delta_j B_j^{\boldsymbol{\kappa},k-1}(x)$ and the underlying basis functions $B_j^{\boldsymbol{\kappa},k-1}$ weighted by δ_j. Figure 1.5 gives an example where $\alpha_j \geq \alpha_{j-1}$ (i.e. $\delta_j \geq 0$) holds for all $j=-(k-1)+1,\ldots,m$. Hence all splines in the upper row are monotonically increasing on the interval $[\kappa_0,\kappa_{m+1}]$. In contrast, in Figure 1.6 the condition $\delta_j \geq 0$ is hurt for some j. For $k=2$ and $k=3$ this implies a derivative that is

negative within a certain interval and hence the spline is non-monotone on $[\kappa_0, \kappa_{m+1}]$. But for $k = 4$ (and also for $k \geq 5$ what is not illustrated here) some combinations of α_j exist where the respective spline $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$ is monotonically increasing on $[\kappa_0, \kappa_{m+1}]$ even if δ_j is negative for some j. That is, for $k \leq 3$ the conditions (1.11) and (1.12) are each necessary and sufficient for monotonicity and for $k \geq 4$ they are each sufficient but not necessary for monotonicity. A compendious discussion can also be found in Dierckx (1993, sec. 7.1).

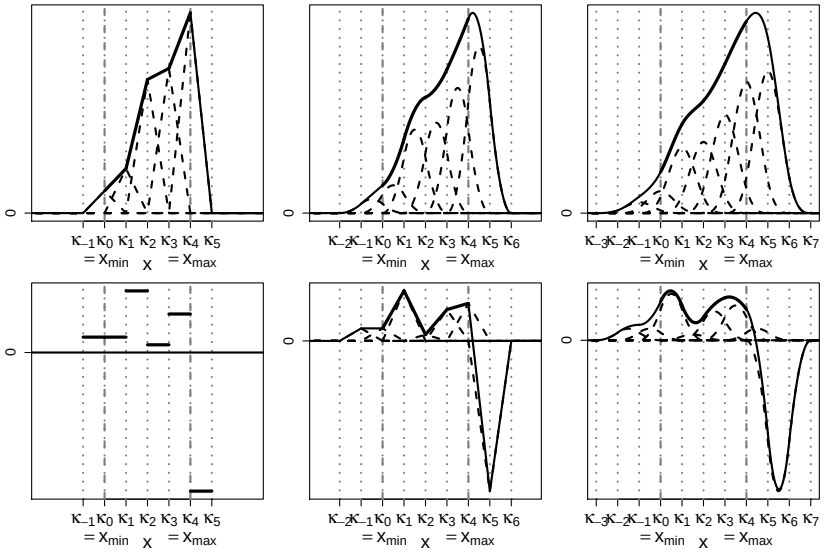

Figure 1.5: Top: monotonically weighted B-spline basis functions of orders 2, 3 and 4, and their sum. **Bottom**: first derivative of the spline from the top with underlying respectively weighted B-spline basis functions.

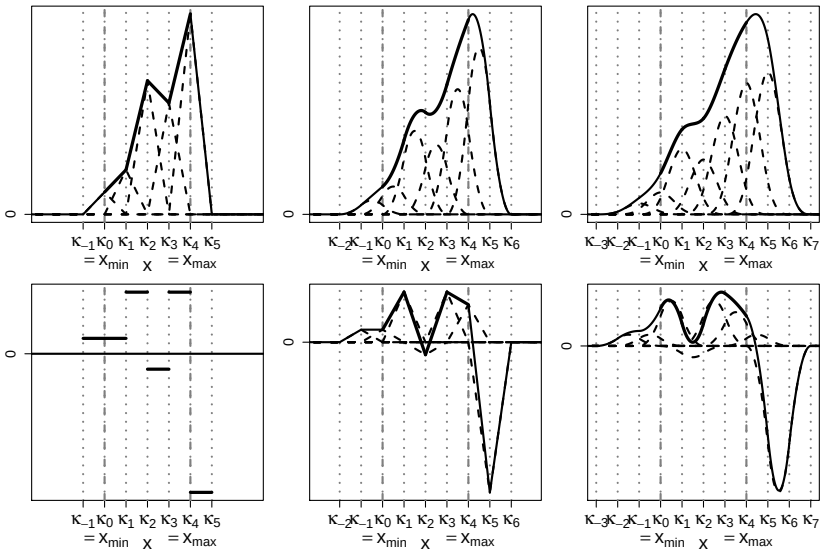

Figure 1.6: Top: apart from one time monotonically weighted B-spline basis functions of orders 2, 3 and 4, and their sum. **Bottom**: first derivative of the spline from the top with underlying respectively weighted B-spline basis functions.

Spline regression

Spline specification For a given order k and a knot sequence $\boldsymbol{\kappa}$, the regression function f for the estimation of $\mathrm{E}(y|x) = f(x)$ in (1.4) can be specified as

$$f(x) = \sum_{j=-(k-1)}^{m} \alpha_j B_j^{\boldsymbol{\kappa},k}(x), \tag{1.13}$$

where the parameters α_j, $j = -(k-1), \ldots, m$, have to be estimated for a given sample (y_i, x_i), $i = 1, \ldots, n$. Hence, the regression function is flexible since it can generate all piecewise polynomial functions of order k with differentiability depending on the knot sequence. Piecewise polynomial functions can well approximate quite arbitrary functions. This can be justified by Weierstrass' approximation theorem (cf. Mackenzie et al., 2005, p. 396) applied to pieces of f defined by two neighboring knots.

The task of the researcher when applying splines for regression purposes is to specify the order k of the spline and the knot sequence $\boldsymbol{\kappa}$ (i.e. the number and position of the knots) and if necessary impose some restrictions and/or penalties on the parameters to estimate. Several approaches concerning these issues as well as how to estimate the parameters α_j, $j = -(k-1), \ldots, m$, are presented in this section.

The different approaches of splines estimation are regression splines, penalized splines and smoothing splines (e.g. Cao et al., 2010, p. 892, Eilers & Marx, 1996, p. 89). The term regression splines denotes splines as in Equation (1.6) in the regression context. Penalized splines are regression splines with an additional penalty on the parameters. These penalties are detailed further down. Finally, smoothing splines can be shown to be penalized splines with knots at all distinct sample values of the covariate x.

Order of the spline For regression splines and penalized splines, the order and the knot sequence have to be specified in advance of the estimation. Though Ruppert et al. (2003, p. 124f.) state that the order of the spline basis nearly does not matter as long as enough knots are used, there might exist some requirements to the resulting fitted spline function. For example, to construct a spline with continuous first derivative, at least the order $k = 3$ has to be chosen. Ruppert's (2002, p. 742) opinion is that $k = 3$ is enough and his experience shows that the results for $k = 3$ and $k = 4$ are usually

similar. He & Shi (1998, p. 644) recommend to use $k = 3$ since then the monotonicity constraints (1.11) and (1.12) are "if and only if" constraints (cf. the explanation on page 12). Many studies use cubic splines which corresponds to $k = 4$. This is also the order that is recommended by Dierckx (1993, p. 45). He argues that splines of order $k = 4$ are computationally efficient and provide a good fit. Further, they allow the researcher to implement constraints on the parameters to guarantee monotonicity or convexity of the fitted regression curve (Dierckx, 1993, p. 119f.). According to Wegman & Wright (1983, p. 354), $k = 4$ is also the smallest order yielding visual smoothness. If more than two continuous derivatives are required, higher order splines with $k > 4$ are to be used (e.g. Dierckx, 1993, p. 45).

Knot sequence for regression splines In specifying the order of the spline basis, regression splines and penalized splines are treated alike. But this is different for specifying the knots. For regression splines the choice of the knots is very important. A trivial choice is to use equidistant knots or knots at equally spaced sample quantiles of x. Thus only the number of the knots or equivalently the number of the inner knots m has to be specified. For the choice of the latter, information criteria or cross-validation can be applied. For example Huang & Shen (2004), Huang et al. (2004) and Landajo et al. (2008) follow this approach. Alternatively, a rule of thumb can be applied. Asymptotic results for regression splines as for example given in the works of He & Shi (1998) or Huang et al. (2004) suggest

$$m \approx n^{1/5} - 1 \tag{1.14}$$

for cubic splines as a rule of thumb which is used for the applications in Sections 2, 3 and 4. Note that for knots at the equally spaced sample quantiles of x, the exterior knots usually are chosen to coincide with the boundary knots since no obvious other positions exist (as opposed to equidistant knots). On $[\kappa_0, \kappa_{m+1}]$ their positions do not matter for the resulting spline (though those basis functions, for which one of the boundary knots is included in the support, differ). For $\kappa_{-(k-1)} = \ldots = \kappa_0$ and $\kappa_{m+1} = \ldots = \kappa_{m+k}$, it holds that $B_j^{\boldsymbol{\kappa},k}(x) = B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x) = 0$ for $x \notin [\kappa_0, \kappa_{m+1}]$. This discussion can be applied to other non-equidistant knot sequences analogously.

For non-equidistant knots there exist many proposals to choose the knot sequence (i.e. the number and positions of the knots). On the one hand, knot selection algorithms based on information criteria can be applied. Then knots are stepwise deleted from or inserted to a given starting knot sequence which is usually equidistant or consists of sample quantiles of x. Examples for knot selection algorithms can be found in He & Shi (1998), Lee (2000) and Wand (2000). On the other hand, the knot positions can be estimated together with the other parameters. Dierckx (1993, sec. 4.2), Eubank (1984, sec. 4) and Huang et al. (2004) give an overview.

Penalized splines The elaborate and often computationally intensive search for the knot sequence can be circumvented if a penalty term is added in the optimization. Then a rather long knot sequence can be chosen since the penalty term avoids a too rough fit of the estimated regression curve. Usually the knot sequence is taken to be equidistant or the knots are placed at equally spaced sample quantiles of x where quite many knots are contained in the knot sequence $\boldsymbol{\kappa}$. However, it is not clear whether equidistant knots or sample quantiles are to be preferred (cf. e.g. the discussion between Eilers & Marx, 2010 and Crainiceanu et al., 2007). Lu et al. (2009, p. 1064) find nearly no differences in their study. As a rule of thumb for the number of knots, about $\min(n/4, 35)$ knots (Ruppert, 2002, p. 753) or 20-40 knots (Lang & Brezger, 2004, p. 186) can be employed.

Instead of a penalty term depending on the (second) derivative of the fitted function as in the work of O'Sullivan (1986, 1988), Eilers & Marx (1996) propose to penalize high second order differences (or differences of another order) of the estimated parameters and thus introduce P-splines as a computationally advantageous special case of penalized splines. The respective minimization problem is given by

$$\min_{\tilde{\boldsymbol{\alpha}} \in \mathbb{R}^{m+k}} \sum_{i=1}^{n} \left(y_i - \sum_{j=-(k-1)}^{m} \tilde{\alpha}_j B_j^{\boldsymbol{\kappa},k}(x_i) \right)^2 + \lambda \sum_{j=-(k-1)+2}^{m} (\Delta^2 \tilde{\alpha}_j)^2 \tag{1.15}$$

for an equidistant knot sequence. The penalty term in (1.15) can also be stated in

matrix notation as $\lambda\,\tilde{\boldsymbol{\alpha}}^T\,\mathbf{D}\,\tilde{\boldsymbol{\alpha}}$ where

$$\mathbf{D} = \begin{pmatrix} 1 & -2 & 1 & & & & & & \\ -2 & 5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & 1 & -4 & 6 & -4 & 1 & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 5 & -2 & \\ & & & & & 1 & -2 & 1 & \end{pmatrix}$$

(derivation using $\Delta^2\tilde{\alpha}_j = \begin{pmatrix} 0 & \dots & 0 & 1 & -2 & 1 & 0 & \dots & 0 \end{pmatrix}\tilde{\boldsymbol{\alpha}} =: \mathbf{d}_j\,\tilde{\boldsymbol{\alpha}}$, $(\Delta^2\tilde{\alpha}_j)^2 = \tilde{\boldsymbol{\alpha}}^T\,\mathbf{d}_j^T\,\mathbf{d}_j\,\tilde{\boldsymbol{\alpha}}$, $\sum_{j=-(k-1)+2}^{m}\mathbf{d}_j^T\,\mathbf{d}_j = \mathbf{D}$). While for $\lambda = 0$ the fit from (1.15) corresponds to that from the unpenalized model which is potentially overfitting, for $\lambda \to \infty$ the fit is a straight line (Eilers & Marx, 1996, p. 93) which may be oversmoothed. Hence, the selection of the smoothing/penalty parameter λ is an important and demanding task since the estimation results are sensitive with respect to λ. Eilers & Marx (1996) suggest to use the Akaike information criterion or (generalized) cross-validation, but many other criteria are possible, too (e.g. Imoto & Konishi, 2003).

The term $\sum_{j=-(k-1)+2}^{m}(\Delta^2\tilde{\alpha}_j)^2$ of the penalty in Equation (1.15) is motivated by the second derivative of $\sum_{j=-(k-1)}^{m}\tilde{\alpha}_j B_j^{\boldsymbol{\kappa},k}(x)$ (with equidistant knots) which can be derived using Equation (1.10) and is given by

$$\frac{\partial^2 B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x^2} = \frac{1}{h^2}\sum_{j=-(k-1)+2}^{m}\left(\Delta^2\alpha_j\right)B_j^{\boldsymbol{\kappa},k-2}(x).$$

Hence, the term $\sum_{j=-(k-1)+2}^{m}(\Delta^2\tilde{\alpha}_j)^2$ penalizes high second derivatives (which can also be interpreted as changes in the first derivative) of the fitted regression function. Analogously, the second derivative for non-equidistant knot sequences,

$$\frac{\partial^2 B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}(x)}{\partial x^2} = (k-1)(k-2)\sum_{j=-(k-1)+2}^{m}\frac{\frac{\alpha_j-\alpha_{j-1}}{\kappa_{j+k-1}-\kappa_j} - \frac{\alpha_{j-1}-\alpha_{j-2}}{\kappa_{j+k-2}-\kappa_{j-1}}}{\kappa_{j+k-2}-\kappa_j}\,B_j^{\boldsymbol{\kappa},k-2}(x)$$

can be used to formulate the respective term in the penalty as

$$\sum_{j=-(k-1)+2}^{m}\left(\frac{\frac{\Delta\alpha_j}{\kappa_{j+k-1}-\kappa_j} - \frac{\Delta\alpha_{j-1}}{\kappa_{j+k-2}-\kappa_{j-1}}}{\kappa_{j+k-2}-\kappa_j}\right)^2. \tag{1.16}$$

Because the constant factor $(k-1)^2\,(k-2)^2\,h_p^4$ with $h_p = \frac{\kappa_{m+1}-\kappa_0}{m+1}$ does not influence the optimization process, it can optionally be multiplied to (1.16) to guarantee that $\sum_{j=-(k-1)+2}^{m}(\Delta^2\alpha_j)^2$ results for equidistant knots.

Smooting splines Concerning the selection of the order k and the knot sequence, smoothing splines play a special role. Kimeldorf & Wahba (1970a,b) showed that $\hat{f}$ from the minimization of

$$\sum_{i=1}^{n} \left(y_i - \tilde{f}(x_i)\right)^2 + \lambda \int_{\min x_i}^{\max x_i} \left(\frac{\partial^\gamma \tilde{f}(x)}{\partial x^\gamma}\right)^2 dx \tag{1.17}$$

with respect to $\tilde{f}$ for some integer $\gamma > 0$ is a spline of order 2γ with possible knots at the distinct observations x_i (Wegman & Wright, 1983, p. 353). That is, the order (e.g. linear or cubic) of the smooting spline results from the choice of γ and only splines with even order can be obtained from the minimization of (1.17). Further, the knots do not have to be chosen since every (distinct) observation constitutes a knot. The smoothing parameter λ can be chosen as for penalized splines using several information criteria.

Hence, penalized splines are in-between regression splines and smoothing splines (e.g. Claeskens et al., 2009, p. 529). On the one hand, for $\lambda = 0$ penalized splines correspond to regression splines. If on the other hand the knots are chosen to be at the distinct values of x and the penalty is based on a derivative of the estimated spline, then penalized splines correspond to smoothing splines.

Monotonicity For many applications a monotone relationship between x and y is assumed a priori. Applying the monotonicity constraint (1.11) or (1.12) for the parameter estimation, ensures a monotone fit. As already explained on page 13, for $k \geq 4$ condition (1.11) or (1.12) is not necessary for $B_{\boldsymbol{\alpha}}^{\boldsymbol{\kappa},k}$ to be monotone on $[\kappa_0, \kappa_{m+1}]$. Hence, Wood (1994) provides necessary conditions for monotonicity but these are not as easy to implement as constraining the parameters by (1.11) or (1.12). Another approach for applications with assumed monotone relationship is to use a second penalty term which penalizes derivations from a monotone fit (Bollaerts et al., 2006, p. 193). In addition, imposing a monotonicity constraint has a smoothing effect on the estimated regression function (Dierckx, 1993, p. 119). This can also be observed in the applications of Sections 2 and 3.

B-spline basis and truncated power basis A basis which is equivalent to the B-spline basis of order k with knot sequence $\boldsymbol{\kappa}$ where $\kappa_{-(k-1)} < \ldots < \kappa_{m+k}$ and with

basis functions $B_j^{\boldsymbol{\kappa},k}$, $j = -(k-1), \ldots, m$, is given by the truncated power basis of degree $k-1$ with basis functions $1, x, \ldots, x^{k-1}, (x-\kappa_1)_+^{k-1}, \ldots, (x-\kappa_m)_+^{k-1}$ (de Boor, 2001, ch. IX). For knot sequences with multiple knots, the respective truncated power basis functions are as described on page 5. Both bases have the same number of basis functions and hence the same number of parameters to estimate. Note that regression splines, penalized splines as well as smoothing splines can be generated from either of the bases where both bases have clear advantages and drawbacks. Eilers & Marx (2010) compare the B-spline basis and the truncated power basis. They clearly favor the B-spline basis, especially due to its computational advantages (see also Eubank, 1984, p. 440 and the discussion on page 6). Apart from computational aspects, the truncated power basis can be explained more intuitively since the piecewise character of the resulting spline is immediately obvious. Further, testing whether a knot is active (i.e. whether it is necessary for the fitted spline and with it the respective estimated parameter) or testing for a polynomial regression function of order k, is much easier when the truncated power basis is applied. In this case, it is sufficient to test whether the respective parameter(s) are significant (e.g. Eubank, 1984, p. 443, Landajo et al., 2008, p. 236f.). But monotonicity constraints are not as easy to obtain compared to estimations using the B-spline basis. When applying P-splines based on the truncated power basis, the second order differences in the penalty term of Equation (1.15) have to be replaced by the squares of the parameters of the truncated power functions (e.g. Eilers & Marx, 2010, p. 638, Kauermann, 2005, p. 57).

Multiple regression The concept of using a B-spline basis to formulate a flexible regression model can be extended to the multiple regression framework. Customary, the multiple regression model is assumed to be additive. Since only models with additive spline components are applied in Sections 2 to 5 of this work, non-additive spline specifications are not discussed here. Some references concerning that issue are Dierckx (1993), Ruppert et al. (2003, ch. 13) and He & Shi (1996). For the additive models one or more covariates can be modeled using splines. Suppose that the covariates $x_1, \ldots, x_s$ have a nonlinear conditional effect on the response y which is approximated by splines and the remaining covariates $x_{s+1}, \ldots, x_q$ have linear conditional influence on y. That

is, a semiparametric model is specified where the covariates $x_1, \ldots, x_s$ constitute the nonparametric part of the model and the remaining covariates $x_{s+1}, \ldots, x_q$ form the parametric part. Since the B-spline basis represents a partition of unity (see Equation (1.5)), incorporating a separate basis for each of the covariates $x_1, \ldots, x_s$ in a model leads to multicollinearity. To avoid this, the bases for $x_1, \ldots, x_s$ have to be slightly modified. From Equation (1.5), each basis function $B_j^{\kappa,k}$ can be written as

$$B_j^{\kappa,k}(x) = 1 - \sum_{\substack{j'=-(k-1) \\ j' \neq j}}^{m} B_{j'}^{\kappa,k}(x).$$

Hence, a basis that is equivalent to $B_j^{\kappa,k}$, $j = -(k-1), \ldots, m$, is given by 1, $B_j^{\kappa,k}$, $j = -(k-1)+1, \ldots, m$, and (1.13) can be reformulated as

$$f(x) = \alpha_{-(k-1)} + \sum_{j=-(k-1)+1}^{m} (\alpha_j - \alpha_{-(k-1)}) B_j^{\kappa,k}(x), \tag{1.18}$$

for the regression case with only one covariate x. To ease notation, Equation (1.18) is written as

$$f(x) = \beta_0' + \sum_{j=-(k-1)+1}^{m} \alpha_j B_j^{\kappa,k}(x) \tag{1.19}$$

though the parameters α_j, $j = -(k-1)+1, \ldots, m$, are not the same as in Equation (1.13). Analogously, the conditional expectation $\mathsf{E}[y|x_1, \ldots, x_q] = f(x_1, \ldots, x_q)$ in the multivariate case of Equation (1.1) is assumed to be given by

$$\begin{aligned} f(x_1, \ldots, x_q) = \beta_0 &+ \sum_{j=-(k_1-1)+1}^{m_1} \alpha_{1j} B_{1j}^{\kappa_1,k_1}(x_1) \\ &+ \ldots + \sum_{j=-(k_s-1)+1}^{m_s} \alpha_{sj} B_{sj}^{\kappa_s,k_s}(x_s) + \sum_{j=s+1}^{q} \beta_j x_j \end{aligned} \tag{1.20}$$

where $\beta_0 = \beta_{10}' + \ldots + \beta_{s0}'$. Further, monotonicity constraints and penalties have to be adjusted. The monotonicity constraints (1.11) and (1.12) are rendered to

$$\alpha_j \geq \alpha_{j-1} \geq 0, \quad j = -(k-1)+2, \ldots, m, \qquad \text{with} \qquad \mathbf{C} = \begin{pmatrix} 0 & 1 & & \\ & -1 & 1 & \\ & & -1 & 1 \\ & & & \ddots \; \ddots \end{pmatrix},$$

and

$$\alpha_j \leq \alpha_{j-1} \leq 0, \quad j = -(k-1)+2, \ldots, m, \qquad \text{with} \qquad \mathbf{C} = \begin{pmatrix} 0 & -1 & & \\ & 1 & -1 & \\ & & 1 & -1 \\ & & & \ddots \; \ddots \end{pmatrix},$$

for the case with only one covariate (1.19) and analogous reasoning can be applied for the multivariate case (1.20). For the penalty term of (1.15) and (1.16) it can be shown that $\alpha_{-(k-1)}$ has to be replaced by 0 for the case with only one covariate. But note that the parameters α_j from (1.15) and (1.16) are not the same as those from (1.18) or (1.20). The matrix $\mathbf{D}$ modifies to

$$\mathbf{D} = \begin{pmatrix} 0 & 0 & 0 & & & & & \\ 0 & 5 & -4 & 1 & & & & \\ 0 & -4 & 6 & -4 & \cdot & & & \\ & 1 & -4 & 6 & \cdot & \cdot & & \\ & & 1 & -4 & \cdot & \cdot & 1 & \\ & & & 1 & \cdot & \cdot & -4 & 1 \\ & & & & \cdot & \cdot & 6 & -4 & 1 \\ & & & & & \cdot & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}.$$

This can be applied analogously to the multivariate case (1.20) with a separate penalty matrix $\mathbf{D}_1$, ..., $\mathbf{D}_s$ for each of the covariates x_1, ..., x_s.

1.1.3 Quantile regression

Basics and loss function Estimating the conditional expectation $\mathrm{E}(y|x_1,\ldots,x_q)$ gives insight into the central tendency of the conditional distribution of y given x_1, ..., x_q. But many aspects of this distribution (as for example skewness) are left unconsidered if only $\mathrm{E}(y|x_1,\ldots,x_q)$ is paid regard to. By analyzing (several) quantiles of the conditional distribution, these unconsidered aspects can be illuminated. Many of the research results for quantile regression are summarized in the monograph of Koenker (2005).

With $0 < \vartheta < 1$, the ϑ-quantile $\mathrm{Q}_\vartheta(y|x_1,\ldots,x_q)$ of the conditional distribution $F(y|x_1,\ldots,x_q)$ of y given $x_1,\ldots,x_q$ is defined as

$$\mathrm{Q}_\vartheta(y|x_1,\ldots,x_q) := \inf\left\{y : F(y|x_1,\ldots,x_q) \geq \vartheta\right\}.$$

Interpreting this conditional ϑ-quantile of y as a function of the covariates $x_1,\ldots,x_q$, that is,

$$\mathrm{Q}_\vartheta(y|x_1,\ldots,x_q) = f_\vartheta(x_1,\ldots,x_q), \tag{1.21}$$

it follows that $\mathrm{Q}_\vartheta(u_\vartheta|x_1,\ldots,x_q) = 0$ holds for the model

$$y = f_\vartheta(x_1,\ldots,x_q) + u_\vartheta.$$

For the specification of f_ϑ in (1.21), the same proceeding as for the specification of f in (1.1) is applicable. Hence, polynomials, transformations or nonparametric terms (e.g. splines as described in Section 1.1.2) of the covariates $x_1, \ldots, x_q$ can be applied to represent the corresponding conditional relationship. Further the functional form may differ across ϑ.

While in the case of the estimation of the conditional expectation $\mathrm{E}(y|x_1, \ldots, x_q)$, the sum of squared residuals $\hat{u}_i := y_i - \hat{y}_i$ has to be minimized, the sum of ϑ-weighted residuals is the objective when the conditional ϑ-quantile $\mathrm{Q}_\vartheta(y|x_1, \ldots, x_q)$ of the respective distribution is estimated (Koenker & Bassett, 1978). The corresponding optimization problem is

$$\min_{\tilde{f}_\vartheta} \sum_{i=1}^{n} \rho_\vartheta \left(y_i - \tilde{f}_\vartheta(x_1, \ldots, x_q) \right). \tag{1.22}$$

The quantile regression function f_ϑ is not specified more precisely here to allow for parametric as well as nonparametric functional forms. The ϑ-weighting function ρ_ϑ (also called check function, Koenker, 2005, p. 5) is defined to be

$$\rho_\vartheta(u) = \left(\vartheta - I_{\{u<0\}} \right) u,$$

where again $I_{\{A\}}$ is the indicator function. An equivalent definition of the ϑ-weighting function is given by

$$\rho_\vartheta(u) = \begin{cases} \vartheta\, u & \text{if } u \geq 0, \\ (1-\vartheta)\, |u| & \text{if } u < 0. \end{cases}$$

That is, in the optimization process the absolute value of positive residuals is weighted by ϑ and that of negative residuals is weighted by $1 - \vartheta$. Hence, for example if $\vartheta = 0.9$, positive residuals have far higher weights than negative residuals and hence the estimated function is pulled upwards compared to estimations with $\vartheta < 0.9$. The respective loss functions for the estimation of the conditional expectation as well as of the conditional median ($\vartheta = 0.5$) and the conditional lower quartile ($\vartheta = 0.25$) are contained in Figure 1.7.

Note that the optimization in (1.22) has to be implemented using numerical optimization algorithms since no closed form solution exists. An extensive overview on computational aspects of quantile regression is given by Koenker (2005, ch. 6).

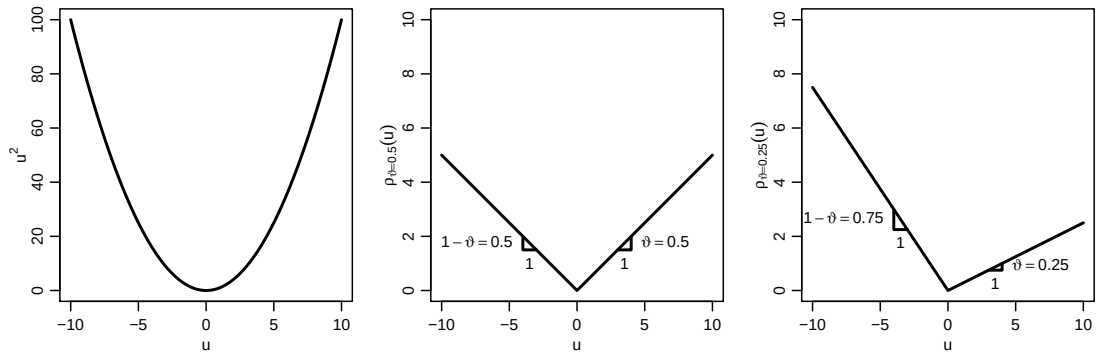

Figure 1.7: Loss functions for least squares estimation (left) and quantile regressions for $\vartheta = 0.5$ (middle) and $\vartheta = 0.25$ (right).

Illustrative examples In the parametric case with a single covariate x, the estimated regression curves for several values of ϑ can be plotted into an x-y-scatter plot. Figure 1.8 shows estimation results with $f_\vartheta(x) = \beta_0(\vartheta) + \beta_1(\vartheta)\, x$ for simulated data with homoskedastic symmetric (`hosy`), homoskedastic skewed (`hosk`), heteroskedastic symmetric (`hesy`) and heteroskedastic skewed (`hesk`) errors with the following specifications:

$$u_{\texttt{hosy}}|x \sim N(0,4), \qquad u_{\texttt{hesy}}|x \sim N(0,4)\cdot(2\,x+0.5),$$
$$u_{\texttt{hosk}}|x \sim \chi^2(2)-2, \qquad u_{\texttt{hesk}}|x \sim (\chi^2(2)-2)\cdot(2\,x+0.5).$$

The remaining parameters of the simulation are $x \sim U(0,1)$ and $y_\bullet = 1 + 8\,x + u_\bullet$ (with $n = 500$) for the respective error distribution $\bullet$.

The first column shows the example where the errors u are homoskedastic and not skewed. Hence, the resulting estimated regression lines (illustrated for $\vartheta = 0.1, \ldots, 0.9$) all have about the same slopes. This can also be observed in the lowest panel of the first column. The respective intercepts in the middle panel of the first column reflect the error distribution. For quantiles with ϑ near 0 or 1 the intercept varies faster with ϑ than for ϑ near 0.5. This is due to the underlying normal conditional error distribution which has more mass in the middle what leads to closer estimated intercepts and hence a slower increase with increasing ϑ for ϑ around 0.5.

In the second column the conditional error distribution is still homoskedastic but skewed. Due to the homoskedasticity, the estimated slopes are again very similar and

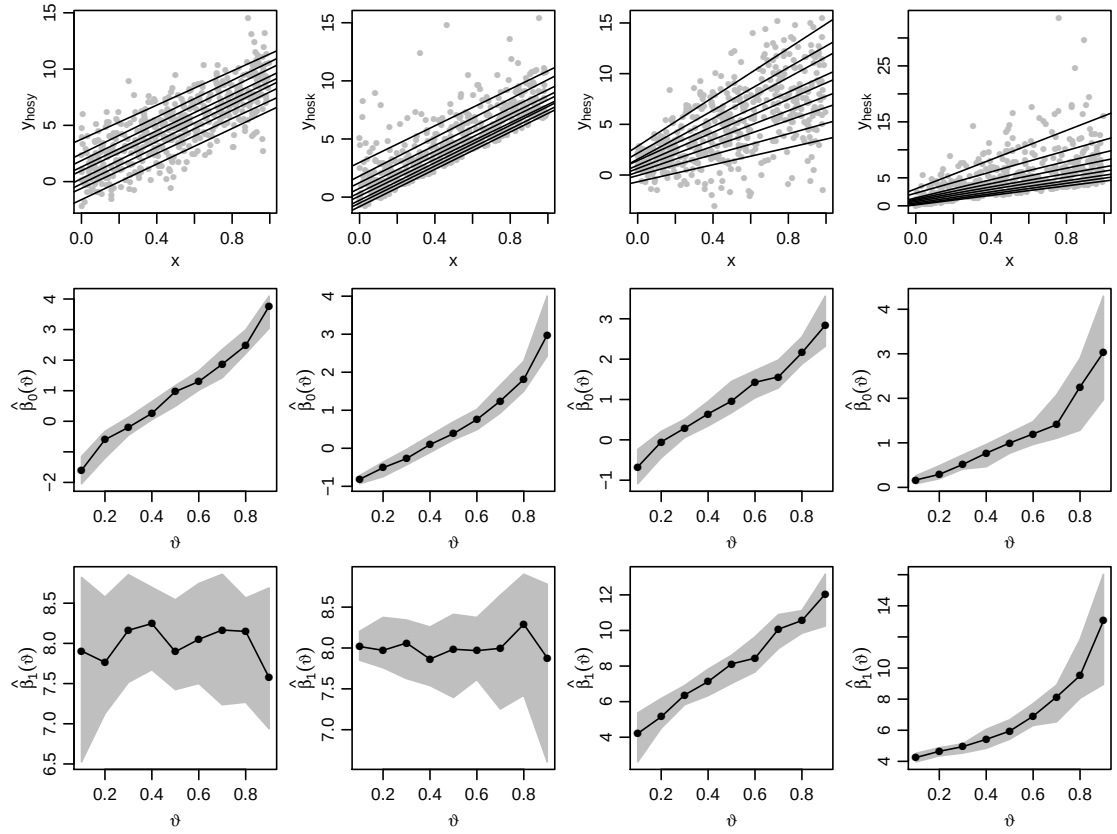

Figure 1.8: Results from quantile regressions with $\vartheta = 0.1, 0.2, \ldots, 0.9$ for simulated data. **From top to bottom**: x-y-scatter plot with estimated regression curves (top), estimated intercepts (middle) and slopes (bottom) of the respective regression with 90% confidence intervals. **From left to right**: data with homoskedastic symmetric, homoskedastic skewed, heteroskedastic symmetric and heteroskedastic skewed errors.

the intercepts reflect the underlying conditional error distribution. Since more mass of the right-skewed error distribution is on the left, the lower conditional quantiles lie closer to each other than the higher conditional quantiles. Hence, the increase with ϑ is slower for lower ϑ than for higher ϑ.

For the heteroskedastic examples in the last two columns, the estimated slopes vary with ϑ. In the given symmetric case, the slopes increase rather constantly with ϑ while they increase faster with higher ϑ for the right-skewed case. In a left-skewed example, the increase would be faster for lower ϑ and slower for higher ϑ, respectively. Furthermore, the estimated intercepts cannot be reasonably put in relation to the underlying conditional error distribution since their order depends on the data range of x. If for example x was replaced by $x + 3$, the conditional quantile lines would cross

at positive values of x and their order at $x = 0$ would be inversed and with it the order of the estimated intercepts. This complication can be avoided by transforming the covariate x such that its mean is 0. Then, the estimated intercepts correspond to the estimated conditional quantiles of the response for $x = \overline{x}$ and are increasing in ϑ (follows from Koenker, 2005, Theorem 2.5, p. 56). In that case, the estimated intercepts can be interpreted as the distribution of the response given $x = \overline{x}$. In the presented simulated example, only the scaling of the ordinate of the intercept panel would change if x was centered such that $\overline{x} = 0$ (but the error term kept the same).

The fact that in the homoskedastic case the estimates of the slope vary around a theoretical constant, can be used to construct a test on homoskedasticity. Such tests can be found in the seminal works of Koenker & Bassett (1982a,b).

Specification under heteroskedasticity For heteroskedastic cases, the functional form of f_ϑ may differ across quantiles. First, consider the heteroskedastic example where $y = 1 + 8\,x + u$ with $u|x \sim N(0,4) \cdot (4\,x^2 + 0.5)$ and $x \sim U(0,1)$ (with $n = 300$). A linear model is appropriate to estimate $\mathrm{Q}_\vartheta(y|x)$ for $\vartheta = 0.5$ but for quantiles closer to 0 or 1 a quadratic model has to be specified. Figure 1.9 shows the example with estimation results for $f_\vartheta(x) = \beta_0(\vartheta) + \beta_1(\vartheta)\,x + \beta_2(\vartheta)\,x^2$ and $\vartheta = 0.1, \ldots, 0.9$. As expected, the fitted curve is rather linear for $\vartheta = 0.5$ (i.e. $\hat{\beta}_2(0.5) \approx 0$) while for ϑ moving away from 0.5 the curvature becomes more and more pronounced (i.e. $|\hat{\beta}_2(\vartheta)|$ increases).

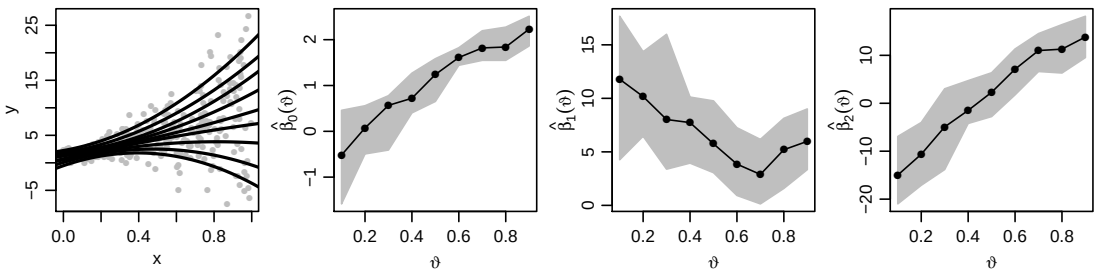

Figure 1.9: Scatter plot with estimated quadratic quantile regression curves and estimated parameters $\hat{\beta}_0(\vartheta)$, $\hat{\beta}_1(\vartheta)$, $\hat{\beta}_2(\vartheta)$ for $\vartheta = 0.1, \ldots, 0.9$ for simulated heteroskedastic symmetric data.

The reason for this behavior is as follows. The ϑ-quantile of a $N(\mu, \sigma^2)$-distributed

random variable u is $\sigma\,\Phi^{-1}(\vartheta)+\mu$, where Φ is the cumulative density function (cdf) of the standard normal distribution. Hence, for the conditionally $N(0,4)\cdot(4\,x^2+0.5)$-distributed random variable u, the conditional ϑ-quantile is $2\,(4\,x^2+0.5)\,\Phi^{-1}(\vartheta)$. This implies that the conditional ϑ-quantile of $y=1+8\,x+u$ is

$$\begin{aligned} Q_\vartheta(y|x) &= Q_\vartheta(1+8\,x+u|x) = 1+8\,x+Q_\vartheta(u|x) \\ &= 1+8\,x+2\,(4\,x^2+0.5)\,\Phi^{-1}(\vartheta) \\ &= (1+\Phi^{-1}(\vartheta))+8\,x+8\,\Phi^{-1}(\vartheta)\,x^2. \end{aligned}$$

Since $\Phi^{-1}(0.5)=0$, a linear model can be used for $\vartheta=0.5$ and for all other ϑ a quadratic model is correctly specified.

In a second example, assume that the error term given the covariate x is logistically distributed with expectation and median equal to 0 and variance depending on x, that is

$$F(u|x)=\left(1+\exp\left(-\frac{u}{b(x)}\right)\right)^{-1}, \qquad \mathrm{Var}(u|x)=\frac{\pi^2}{3}\,b(x)^2,$$

where $x\sim U(0,1)$ and again $y=1+8\,x+u$. Then, the conditional ϑ-quantile of u is

$$Q_\vartheta(u|x)=b(x)\,\log\left(\frac{\vartheta}{1-\vartheta}\right)$$

and

$$Q_\vartheta(y|x)=Q_\vartheta(1+8\,x+u|x)=1+8\,x+Q_\vartheta(u|x)=1+8\,x+b(x)\,\log\left(\frac{\vartheta}{1-\vartheta}\right).$$

As long as $b(x)$ is linear in x, linear regression functions $f_\vartheta(x)=\beta_0(\vartheta)+\beta_1(\vartheta)\,x$ are correctly specified for all ϑ. But if $b(x)$ is a nonlinear function in x (e.g. $b(x)=4\,x^2+1$), linear $f_\vartheta(x)$ are misspecified for $\vartheta\neq 0.5$ (note that $\log(\frac{\vartheta}{1-\vartheta})=0$ for $\vartheta=0.5$). Figure 1.10 shows the conditional ϑ-quantiles $Q_\vartheta(y|x)$ with $\vartheta=0.1,\ldots,0.9$ for two examples of $b(x)$. In the left example, a linear model is adequate for all quantiles since $b(x)=4\,x+1$ and thus $Q_\vartheta(y|x)=(1+\log(\frac{\vartheta}{1-\vartheta}))+(8+4\,\log(\frac{\vartheta}{1-\vartheta}))\,x$. In the right example, however, a linear specification is not correct since $b(x)=4\,x^2+1$ and thus $Q_\vartheta(y|x)=(1+\log(\frac{\vartheta}{1-\vartheta}))+8\,x+4\,\log(\frac{\vartheta}{1-\vartheta})\,x^2$. Hence for $\vartheta\neq 0.5$, a linear model would ignore the nonlinearity of the conditional quantiles.

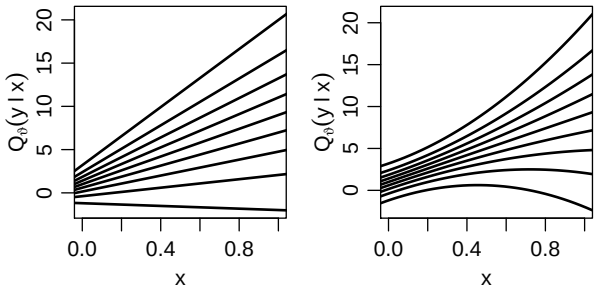

Figure 1.10: Conditional ϑ-quantiles $Q_\vartheta(y|x)$ for $\vartheta = 0.1, \ldots, 0.9$ with $y = 1 + 8\,x + u$ where u is logistically distributed with $b(x) = 4\,x + 1$ (left) and $b(x) = 4\,x^2 + 1$ (right).

Finally, assume that $u = \varepsilon \cdot s(x)$, where s is a strictly positive function and the random variable ε is independent of x and has cdf F. Then, the ϑ-quantile of ε is $F^{-1}(\vartheta)$ and hence, the conditional ϑ-quantile of u is

$$Q_\vartheta(u|x) = Q_\vartheta(\varepsilon \cdot s(x)|x) = s(x) \cdot Q_\vartheta(\varepsilon) = s(x) \cdot F^{-1}(\vartheta).$$

Consequently, the conditional ϑ-quantile of $y = f(x) + u$ is

$$Q_\vartheta(y|x) = Q_\vartheta(f(x) + u|x) = f(x) + Q_\vartheta(u|x) = f(x) + s(x) \cdot F^{-1}(\vartheta).$$

Hence, the correct specification of $Q_\vartheta(u|x)$ does not depend on the distribution of ε, but on f as well as on s. If, for example, $f(x)$ and $s(x)$ are linear in x, it is correct to specify $Q_\vartheta(y|x)$ linearly in x as was the case in the examples from Figure 1.8.

The above representation of $Q_\vartheta(y|x)$ can also be found in Koenker & Bassett (1982a, p. 45) and corresponds to the location-scale model in Koenker (2005).

Model evaluation The possible nonlinearity in parts of the conditional distribution caused by heteroskedasticity gives reason to estimate not only the central tendency of the conditional distribution but several quantiles and further to specify the regression curve $f_\vartheta(x_1, \ldots, x_q)$ more flexibly, for example using higher order polynomials, transformations of the covariates or splines as in Section 1.1.2. To specify the final model for the estimation, model selection criteria can be applied analogously to well-known criteria for least squares estimations. Due to the different loss functions in the respective minimization problems, the criteria have to be adapted suitably. Roughly

speaking, in the criteria the squares for the least squares estimation have to be replaced by ϑ-weighting for quantile regression.

$R^1(\vartheta)$ can be used as a goodness-of-fit measure to evaluate the estimated model for the ϑ-quantile analogously to R^2 in the least squares case. $R^1(\vartheta)$ is defined as

$$R^1(\vartheta) = 1 - \frac{\sum_{i=1}^{n} \rho_\vartheta \left(y_i - \hat{f}_\vartheta(x_{i1}, \ldots, x_{iq}) \right)}{\sum_{i=1}^{n} \rho_\vartheta (y_i - y_\vartheta)} \tag{1.23}$$

(based on Koenker & Machado, 1999, p. 1297), where y_ϑ is the sample ϑ-quantile of y which can also be obtained by regressing y on a constant only (e.g. Koenker & Bassett, 1978, p. 38). Further, the average ϑ-weighted error ($ATWE(\vartheta)$)

$$ATWE(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} \rho_\vartheta \left(y_i - \hat{f}_\vartheta(x_{i1}, \ldots, x_{iq}) \right)$$

can be evaluated analogously to the average squared error for least squares estimates. For the goodness-of-fit measure, it holds that $0 \leq R^1(\vartheta) \leq 1$ (based on Koenker & Machado, 1999, p. 1297). $R^1(\vartheta)$ cannot exceed 1 because both, numerator and denominator are non-negative and it cannot be negative since the nominator is at most as large as the denominator since otherwise the sample quantile y_ϑ would be a solution to (1.22) and not $\hat{f}_\vartheta$. Analog to R^2, $R^1(\vartheta)$ cannot decrease when additional covariates enter the estimated model (based on Koenker & Machado, 1999, p. 1297). Hence it is not suitable for model specification purposes.

Information criteria penalize the adding of further covariates. For quantile regression they are derived analogously to those for least squares regression by using the maximum likelihood estimator but are based on an asymmetric Laplace distribution. The Schwarz information criterion for some model is defined as $ll - \frac{1}{2} d \log n$ where ll is the logarithmized likelihood for the model and d is the dimension of the model (Schwarz, 1978, p. 461). It has to be maximized to find the best model. Analogously, defining the criterion as

$$SIC = -ll + \frac{1}{2} d \log n,$$

it has to be minimized.

For the asymmetric Laplace distribution $f_U(u_\vartheta) = \frac{\vartheta(1-\vartheta)}{\sigma} \exp(-\frac{1}{\sigma} \rho_\vartheta(u_\vartheta))$, where σ is a scale parameter (cf. Koenker & Machado, 1999, p. 1298) and the errors

$u_\vartheta = y - f_\vartheta(x_1, \ldots, x_q)$ are independent, the log-likelihood for the model $Q_\vartheta(y|x) = f_\vartheta(x_1, \ldots, x_q)$ is

$$ll(\vartheta) = n \log(\vartheta(1-\vartheta)) - n \log \sigma - \sum_{i=1}^{n} \frac{1}{\sigma} \rho_\vartheta(y - f_\vartheta(x_1, \ldots, x_q)). \tag{1.24}$$

Differentiating (1.24) with respect to the scale parameter σ and equating 0 yields

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^{n} \rho_\vartheta(y - f_\vartheta(x_1, \ldots, x_q)).$$

By inserting $\hat{\sigma}$ into (1.24), the estimated log-likelihood is

$$\hat{ll}(\vartheta) = n \log(\vartheta(1-\vartheta)) - n \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\vartheta(y - \hat{f}_\vartheta(x_1, \ldots, x_q)) \right) - n.$$

The terms $n \log(\vartheta(1-\vartheta))$ and $-n$ as well as the factor $\frac{1}{n}$ are irrelevant for the minimization of $SIC(\vartheta)$, hence it is

$$SIC(\vartheta) = \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\vartheta \left(y_i - \hat{f}_\vartheta(x_{i1}, \ldots, x_{iq}) \right) \right) + \frac{d \log n}{2n}. \tag{1.25}$$

The Akaike information criterion can be obtained analogously based on $-2\,ll + 2\,d$ (Akaike, 1974, p. 719) or equivalently $AIC = -ll + d$ and hence, using the asymmetric Laplace distribution, it is given by

$$AIC(\vartheta) = \log \left(\frac{1}{n} \sum_{i=1}^{n} \rho_\vartheta \left(y_i - \hat{f}_\vartheta(x_{i1}, \ldots, x_{iq}) \right) \right) + \frac{d}{n}. \tag{1.26}$$

$SIC(\vartheta)$ and $AIC(\vartheta)$ can also be found for example in Koenker (2005, p. 135 for the median case and respective Errata).

Properties of quantile regression The dimension d of the estimated model is $q+1$ in the linear case as in Equation (1.3). Equivalently it is obtained by counting the zero-residuals from the estimation (Koenker, 2005, p. 33) or in practice the number of residual with absolute value smaller than some tolerance value (Koenker & Mizera, 2004, p. 156). Besides this exact-fit property there are more useful properties of quantile regression. One of the most important features of quantile regression is the invariance under monotone transformations g,

$$Q_\vartheta(g(y)|x_1, \ldots, x_q) = g\left(Q_\vartheta(y|x_1, \ldots, x_q)\right)$$

(Koenker, 2005, p. 39). If for example the model to be estimated is $\log(y) = f(x_1, \ldots, x_q) + u$ (that is $g = \log$), this means that the conditional ϑ-quantile of y can be directly estimated by applying the exponential function to the estimated ϑ-quantile of $\log(y)$. This does not hold for the (conditional) expectation. There, corrections based on the conditional distribution of u have to be made (see e.g. Wooldridge, 2009, p. 210ff. for an example). Further, the quantile regression results are robust to outliers in the response y to a certain degree (Koenker, 2005, sec. 2.3). While for least squares regression, the estimated regression plane in general changes with every change in y, for quantile regression the fitted plane only changes when the sign of the respective residual changes. Hence, as long as the initial and the new y lie on the same side of the initially estimated regression plane, this estimated plane does not change (Koenker, 2005, p. 44).

Though (in general) no (strong) distributional assumption is made about the errors, it can be shown (Koenker & Bassett, 1978, Koenker, 2005, ch. 3, 4) that the estimated parameters are asymptotically normally distributed under very general conditions on the regressors and the error distribution (Koenker, 2005, p. 120f.). The estimation of the covariance matrix of the estimated parameters has been explored by many authors. A survey of existing methods can be found in Koenker (2005, ch. 3, 4, A.5) or Kocherginsky et al. (2005).

Prediction intervals can be constructed directly from estimated quantiles. Exact prediction intervals that are based on the estimated conditional expectation generally rely on the distributional assumptions that have been incorporated. Mostly the normal distribution is applied and the prediction intervals are constructed symmetrically around the least squares point prediction. But if the distributional assumption (including for example assumptions on a certain form of heteroskedasticity or skewness) does not hold, the intervals based on the least squares estimation are misspecified. Using (correctly specified) quantile regressions, prediction intervals can directly be estimated. For example the estimated conditional quantile curves for $\vartheta = 0.1$ and $\vartheta = 0.9$ can be combined to a 80% prediction interval for the response y. Consider again the example illustrated in the left panel of Figure 1.9. Here it is instantly obvious that the 80% prediction interval based on the conditional ϑ-quantiles for $\vartheta = 0.1$ and $\vartheta = 0.9$ clearly

differs from a simple 80% prediction interval based on the conditonal expectation and a normal distribution with homoskedastic variance. The latter interval would consist of two parallel rather straight lines while the former interval does not. More detailed issues concerning confidence and prediction intervals based on quantile regression are discussed by Zhou & Portnoy (1996) and Koenker (2011a).

Quantile crossing Though quantile regression is robust in many regards, it also has a weak point, which is illustrated in Figure 1.11 and has already been mentioned on page 24. It shows the estimated conditional quantile lines for $\vartheta = 0.3$ (solid) and $\vartheta = 0.4$ (dashed) for the example with $y = 1 + 8\,x + u$ where $x \sim U(0,1)$ and $u|x \sim N(0,4) \cdot (2\,x + 0.5)$ (with $n = 100$). It can be observed that for small values of x, the estimated line for $\vartheta = 0.3$ is situated above the estimated line for $\vartheta = 0.4$ and they cross at a x-value between 0 and 1.

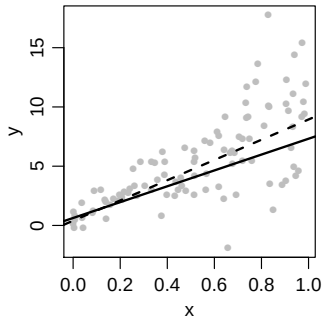

Figure 1.11: Scatter plot with crossing estimated quantile regression lines for $\vartheta = 0.3, 0.4$ (solid, dashed, respectively) for simulated heteroskedastic symmetric data.

This phenomenon is known as quantile crossing and has to be avoided if possible, since $Q_\vartheta(y|x_1,\ldots,x_q)$ interpreted as function of ϑ is naturally monotonically increasing in ϑ (Koenker, 2005, p. 56). Several approaches to avoid quantile crossings are proposed in the literature (e.g. He, 1997, Neocleous & Portnoy, 2008, Chernozhukov et al., 2010, Shim et al., 2009). Further, quantile crossings can be used to detect misspecification. If quantile crossings occur for many x, the estimated regression functions f_ϑ may be misspecified (Koenker, 2005, p. 57).

Quantile splines As already mentioned, the quantile regression function f_ϑ can be specified flexibly using splines where the shapes of the estimated conditional quantiles are allowed to differ across ϑ. Just as for the estimation of the conditional expectation, regression splines, penalized splines and smoothing splines can be applied, and, if necessary, constraints can be imposed on the parameters. Due to the optimization algorithm for quantile regression (Koenker & Bassett, 1978, Koenker, 2005, ch. 6), inequality constraints such as monotonicity conditions (1.11) or (1.12) can easily be adopted in the algorithm (He & Ng, 1999, Koenker & Ng, 2005, Ng & Maechler, 2007). Bollaerts et al. (2006) use P-splines analog to those from Eilers & Marx (1996) who estimate the conditional expectation. The respective minimization problem for the case with a single covariate is given by

$$\min_{\tilde{\boldsymbol{\alpha}}(\vartheta)\in\mathbb{R}^{m+k}} \sum_{i=1}^{n} \rho_\vartheta \left(y_i - \sum_{j=-(k-1)}^{m} \tilde{\alpha}(\vartheta)_j B_j^{\boldsymbol{\kappa},k}(x_i) \right) + \lambda \sum_{j=-(k-1)+d}^{m} |\Delta^d \tilde{\alpha}_j(\vartheta)|.$$

Bollaerts et al. (2006) also discuss the computational issues concerning this minimization problem (for a given penalty parameter). Further they impose monotonicity by adding a second penalty term that punishes deviations from a non-monotone fit. For the specification of the smoothing parameter λ, they apply cross-validation amongst others. Finally, smoothing splines (i.e. a knot at every distinct sample value of x and a penalty term avoiding a rough fit) can also be applied for quantile regression. They are obtained from minimizing

$$\sum_{i=1}^{n} \rho_\vartheta \left(y_i - \tilde{f}_\vartheta(x_i) \right) + \lambda \left(\int_{\min x_i}^{\max x_i} |\tilde{f}''_\vartheta(x)|^\gamma dx \right)^{1/\gamma}$$

with respect to $\tilde{f}_\vartheta$ (Koenker et al., 1994). The authors show that $\gamma = 1$ yields a linear spline for $\hat{f}_\vartheta$ and for $\gamma = \infty$ a quadratic spline results. The smoothing parameter λ can for example be chosen using $SIC(\vartheta)$ as defined in Equation (1.25) (Koenker et al., 1994, p. 677). Yuan (2006) also discusses this issue and suggests further criteria to choose the smoothing parameter. Monotonicity can easily be imposed as described in Koenker & Ng (2005).

1.1.4 Computational aspects

The estimations carried out in this thesis are performed using the open-source software R (R Development Core Team, 2011, www.r-project.org). In Sections 2, 3 and 5, version 2.13.0 was applied and version 2.11.1 in Section 4.

For all quantile regressions, functions from the package quantreg of Koenker (2011b) are applied. The parametric models as well as the semiparametric spline models without additional constraints or penalties are estimated by rq. Monotonicity constraints can be implemented by using the option method="fnc" in rq where the constraints on the parameters have to be supplied in the form $\mathtt{R}\,\mathtt{b} \geq \mathtt{r}$ (with $\mathtt{R} = \mathbf{C}$, $\mathtt{b} = \boldsymbol{\alpha}$ and $\mathtt{r} = \mathbf{0}$ from page 20). Quantile smoothing splines can be estimated using rqss. The functions summary.rq and summary.rqss are used for inference.

B-splines and their derivatives can be evaluated at a value x using the function splineDesign from the base package splines, where the knot sequence and the order have to be specified. The B-splines evaluated at the observed x are treated as regressors in the functions rq for quantile regressions or lm for least squares regressions. To implement penalties or inequality constraints, the function pcls from the package mgcv of Wood (2011) is applied for least squares estimations.

The kernel estimations in Sections 4 and 5 are carried out using the functions of the np package of Hayfield & Racine (2011): npreg and npregbw for the least squares estimations and npcdensbw and npqreg for quantile regressions.

1.2 Outline of the projects

The main part of this dissertation consists of four projects contained in Sections 2 to 5. This section provides an outline of each of the projects which have in common that they discuss model specification and/or prediction issues for spline regression. In the first application, several quantiles and the expectation of the conditional response distribution are estimated where one of the covariates enters the model using a spline component. The next project analyzes the same application but the focus is on prediction. Quantile regressions with spline components are compared to fully linear and fully nonparametric quantile regressions in the third project where estimation as well as prediction results are analyzed. Finally in the fourth project, prediction methods for observations with covariate value outside the interval $[\kappa_0, \kappa_{m+1}]$ are suggested and compared.

1.2.1 Beyond mean estimates of price and promotional effects in scanner-panel sales-response regression

The first project (see Section 2) contains a marketing application of monotone B-spline quantile regression. The aspects of quantile regression and (monotone) B-spline estimation which are relevant for the project are briefly discussed. Further a short review over recent research concerning the study of nonlinearities in the field of sales response is given.

For the application, a large marketing data set is processed. It contains scanner information on sales, prices and promotional activities of nine competing orange juice brands. Using this data set, a parametric benchmark model is specified which explains sales of a certain brand by its own price, prices of the competing brands, promotional activity as well as time and store information. The continuous variables in the model are log-transformed and hence the estimated parameters are interpreted as elasticities. The results from the estimation of the parametric model for the nine brands are in accordance with existing marketing theory and studies.

The parametric benchmark model is compared to the respective model where the

own-price is modeled using a spline component that is monotonicity constrained. An equidistant knot sequence is applied where the number of inner knots is chosen according to the rule of thumb in Equation (1.14). The results show that the own-price elasticities from the semiparametric spline model are clearly non-constant and hence the parametric benchmark model is misspecified.

Even though the covariates are log-transformed, the normality assumption on the errors of the model still is violated. Hence, median regression is implemented as an alternative to estimate the central tendency of the conditional sales distribution. Besides the conditional median, several other quantiles are estimated.

In summary, the project provides an example where classical parametric least squares estimation provides misleading conclusions due to the different mechanisms across the conditional sales distribution and the nonlinearities in the estimated conditional curves. Further, the spline specification process is clearly documented to serve as a guidance for users of (monotonicity constrained) spline regression that are new in this field.

1.2.2 Using quantile regression to predict brand sales from retail scanner data

The application of the project described in Section 1.2.1 is considered again in Section 3 but the data set is restricted to constitute a balanced panel covering 88 weeks for several stores. While in the previous project the main focus is on specification, estimation, interpretation and qualitative comparison of the estimates of the models for the different brands, the focus now is on sales prediction for one specific brand. Again, a parametric benchmark model is compared to several spline models using least squares and quantile regression. The semiparametric spline models for this application are an unconstrained B-spline model, a monotonicity constrained B-spline model and a monotonicity constrained smoothing spline model.

First, the in-sample fit of the four models is compared. In accordance with the discussion concerning the loss function for quantile regression and respective information criteria in Section 1.1.3, the models are compared using the Schwarz and Akaike

information criteria defined in Equations (1.25) and (1.26). Due to the possible overfitting when using too flexible models, the paper focuses on the evaluation of the out-of-sample predictive performance of the competing models. Therefore, the data set is split several times into an estimation and a prediction subsample. Unlike many applications, the sample is not split randomly but 52 consecutive weeks are used for estimation and sales one or four weeks ahead are predicted. To compare the predictive performance, the average squared error of prediction ($ASEP$) and the average ϑ-weighted error of prediction ($ATWEP$) are calculated for the predictions based on the estimation of the conditional expectation and the conditional ϑ-quantiles, respectively.

In a more detailed analysis, the $ASEP$ and the $ATWEP$ are examined in a more disaggregated fashion exploiting the panel structure of the data set. $ASEP$ and $ATWEP$ are calculated separately for every week, store and observation. Using graphical tools, these measures are compared across the four estimated models for the conditional expectation as well as for the conditional quantiles. In doing so, particularly one week can be detected for which the unconstrained spline model performs extremely bad. By regarding the estimated conditional regression curves for this week, it can be observed that the data to be predicted lie in a sparse region where the unconstrained curve fluctuates a lot.

Finally, using estimates with $\vartheta = 0.1$ and $\vartheta = 0.9$, 80% in-sample confidence intervals are constructed. Due to the violation of the assumptions required for the least squares-based intervals (such as symmetry of the conditional distribution), the quantile-based intervals and the least squares-based intervals are found to differ substantially where the empirical coverage of the quantile-based intervals is much closer to the theoretical coverage.

Overall, the two monotonicity constrained models (B-splines and smoothing splines) perform best with respect to in-sample fit as well as with respect to predictive accuracy. Further, it can be observed that the monotonicity constraint prevents the estimated conditional curve from being too rough.

1.2.3 Cross-validating fit and predictive accuracy of nonlinear quantile regressions

The paper in Section 4 focuses on a comparison of different model classes in the context of quantile regression. Again, a parametric benchmark model is compared to more flexible specifications which are a semiparametric B-spline specification and a fully nonparametric kernel specification. For the latter, the approach of Li & Racine (2008) is applied. It allows to estimate the quantiles of the conditional response distribution by completely estimating this conditional distribution using mixed kernels in a first step. This implies that the specification is restricted to be the same for all values of ϑ but it is flexible with respect to the shape of the conditional distribution including possible interactions among the covariates. However, the additional flexibility results in higher computational costs.

After a discussion describing the differences, strengths and shortcomings of the three model classes, criteria to suitably evaluate and compare their performance are presented. As the dimension of the fully nonparametric quantile model cannot be determined, information criteria like $AIC(\vartheta)$ and $SIC(\vartheta)$ as in Equations (1.26) and (1.25) cannot be applied. Hence the in-sample performance is evaluated using $R^1(\vartheta)$ as in Equation (1.23). To account for potential overfitting (what could be done by $AIC(\vartheta)$ and $SIC(\vartheta)$ by penalizing large model dimensions and hence too high flexibility), the average ϑ-weighted error ($ATWE(\vartheta)$) is also calculated for out-of-sample observations.

For the evaluation of $R^1(\vartheta)$ and $ATWE(\vartheta)$, the sample is split R times disjointly into an estimation and a prediction subsample. From each of the R estimation subsamples, $R^1(\vartheta)$ is calculated for each of the three models. $ATWE(\vartheta)$ is determined for the prediction subsamples from the predictions based on the estimates of the respective estimation subsample.

In a first step, this cross-validation approach is applied to the well-known Boston housing data set (e.g. Newman et al., 1998). Appropriate specifications are chosen for the parametric and the semiparametric model. This is avoided when using the fully nonparametric kernel model but the bandwidths have to be specified as detailed in

Section 4. The results from the evaluation of $R^1(\vartheta)$ are as to be expected: since the parametric model is nested in the semiparametric splines model, the latter naturally outperforms the former; further the nonparametric model performs superior compared to the semiparametric model due to its flexibility. The $ATWE(\vartheta)$ results still favor the non- and semiparametric models compared to the parametric specification though the differences are not as large as in the $R^1(\vartheta)$-case. The trade-off between in-sample fit and out-of-sample performance is illustrated using an $R^1(\vartheta)$-$ATWE(\vartheta)$-plot.

In a Monte Carlo simulation, an analog analysis is performed for data sets obtained from several data generating processes (DGPs). All DGPs contain two categorical covariates and one continuous covariate but they differ with respect to the distribution of the continuous covariate, the error distribution and the signal-to-noise ratio. In general, the results from the simulation correspond to those from the empirical application, although the differences in $R^1(\vartheta)$ and $ATWE(\vartheta)$ across the specifications vary with respect to the DGP.

Summarizing, the paper presents an approach that allows to compare arbitrary quantile regression specifications. Several methods for the comparison are suggested, including proper graphical tools.

1.2.4 Out-of-sample predictions for penalized splines

The two previous projects considered prediction from splines models. However, both did not discuss the case where the covariate value of the prediction observation does not lie within the interval $[\kappa_0, \kappa_{m+1}]$. This issue is considered in Section 5 for least squares P-spline regressions.

For estimations using P-splines, the smoothing parameter λ from Equation (1.15) can be chosen for example by applying information criteria like SIC or generalized cross-validation. Both depend on the dimension of the estimated model. Since a penalty and possibly a monotonicity constraint is imposed on the estimation, the dimension of the estimated model does not necessarily equal the (effective) number of estimated parameters. In least squares estimation, the effective number of estimated parameters can also be obtained by the trace of the respective hat matrix. Hence, a formula for

the hat matrix of monotonicity constrained penalized spline estimation is derived and implemented in the applications later on.

For splines estimations the estimated regression curve cannot be expediently continued straightforward outside the range of the given sample. Hence, different methods to overcome this problem for P-splines are proposed. On the one hand, the methods are extrapolating the regression curve at the boundary knot constantly or linearly. On the other hand the B-spline basis is enlarged by one additional B-spline and the respective coefficient is estimated from the coefficients of the original spline estimation by using autoregression (AR) techniques.

The presented prediction methods are compared in a Monte Carlo simulation for six very different DGPs. For each DGP, 1000 data sets are simulated. Each data set is split several times (by varying t) into an estimation and prediction subsample where the observations from the estimation sample have covariate values that lie within $[\kappa_0, \kappa_t]$ and those from the prediction sample have covariate values within the interval $(\kappa_t, \kappa_{t+1}]$, $t = s, \ldots, m$, where $s \approx m/2$. The prediction samples are evaluated with respect to the $ASEP$. It is analyzed in how many cases one of the presented prediction methods outperforms the other methods. The results are illustrated in compact graphical form and suggest in manifold ways that the simplest alternative, i.e. extrapolating the estimated regression curve constantly, is the most favorable method. The linear continuation of the regression curve and the AR-methods often provide similar results.

To evaluate the performance of the prediction methods for larger distances to the boundary knots, wider prediction horizons are considered. Again, the constant method promises the most reliable results since its $ASEP$ across the prediction horizons is rather stable. Further, the spline-based predictions are compared to kernel-based predictions where the kernel-based predictions are clearly outperformed since their boundary behavior is very volatile.

In a nutshell, the paper presents several methods that allow to predict based on spline estimation though the covariate value of the prediction observation does not lie within the range of the estimation sample.

2 Beyond mean estimates of price and promotional effects in scanner-panel sales-response regression

This essay is joint work with Harry Haupt[1].

It is published in the *Journal of Retailing and Consumer Services* (Haupt & Kagerer, 2012):
http://www.sciencedirect.com/science/article/pii/S0969698912000653.

[1]Centre for Statistics, Department of Economics and Business Administration, Bielefeld University, PO Box 100131, 33501 Bielefeld. Email: hhaupt@wiwi.uni-bielefeld.de

3 Smooth quantile based modeling of brand sales, price and promotional effects from retail scanner panels

This essay is joint work with Harry Haupt[1] and Winfried J. Steiner[2].

It will be published in the *Journal of Applied Econometrics* (Haupt et al. 2013): http://onlinelibrary.wiley.com/doi/10.1002/jae.2347/abstract.

[1]Centre for Statistics, Department of Economics and Business Administration, Bielefeld University, PO Box 100131, 33501 Bielefeld, Germany. Email: hhaupt@wiwi.uni-bielefeld.de

[2]Department of Marketing, Clausthal University of Technology, Julius-Albert-Straße 2, 38678 Clausthal-Zellerfeld, Germany. Email: winfried.steiner@tu-clausthal.de

4 Cross-validating fit and predictive accuracy of nonlinear quantile regressions

This essay is joint work with Harry Haupt[1] and Joachim Schnurbus[2].

It is published in the *Journal of Applied Statistics* (Haupt et al., 2011):
http://www.tandfonline.com/doi/abs/10.1080/02664763.2011.573542#.U0xfeaI3WuI.

[1]Centre for Statistics, Department of Economics and Business Administration, Bielefeld University, PO Box 100131, 33501 Bielefeld, Germany. Email: hhaupt@wiwi.uni-bielefeld.de

[2]Centre for Statistics, Department of Economics and Business Administration, Bielefeld University, PO Box 100131, 33501 Bielefeld, Germany. Email: jschnurbus@wiwi.uni-bielefeld.de

5 Out-of-Sample Prediction for Penalized Splines

This essay was under review at *Computational Statistics*.

Summary. Splines are an attractive nonparametric estimation technique as they are computationally fast and inexpensive. But they have the drawback that the estimated curves are only applicable within the range defined by the given sample. Several methods for predictions using spline estimates for such out-of-sample observations are proposed and compared by an extensive Monte Carlo study and two empirical examples using well-known data sets. Further, the paper gives a formula for computing the hat matrix of a penalized and inequality constrained splines estimator. Its trace estimates the dimension of the estimated model which is necessary for the calculation of several information criteria or the standard error of the regression.

Keywords: Out-of-sample, spline, prediction, monotonicity, hat matrix, Monte Carlo simulation.

5.1 Introduction

Non- and semiparametric regression has recently become widespread practice (e.g. Ruppert et al., 2009 who give an overview over semiparametric regression during the period 2003 to 2007) due to avoiding the often challenging and complex task of specifying a complete parametric functional form. Plenty of different options for non- and semiparametric estimation of the conditional mean or conditional quantiles exist. The one extreme is given by fully nonparametric estimation techniques (e.g. kernel regression, Härdle, 1990, Racine & Li, 2004), which result in computationally extensive settings and moreover face the curse of dimensionality. On the other hand, semiparametric and/or additive estimation techniques are available (e.g. Hastie & Tibshirani, 1990 or for spline regression Eilers & Marx, 1996, He & Shi, 1998). They do not share all drawbacks of the fully nonparametric methods and still allow for flexibility of the functional form, though less than the fully nonparametric methods.

While for fully parametric regressions it is practically possible to continue the regression line/hyperplane further to the right/left or across regions without data, this is often infeasible for non- and semiparametric regressions. The latter methods are data-driven and hence, in regions where no data is observed (as is the case outside the boundaries of the sample) the functional form cannot be determined. For example when using smoothing splines, there is no information about the spline basis outside the given data range.

This paper proposes and compares some methods for predicting the expected value of the response y for out-of-sample observations when the conditional relationship between y and a covariate x is estimated using splines. Thereby, the term "out-of-sample observations" denotes observations whose value of x does not lie within the sample range of x, i.e. does not lie within the interval $[\min_i(x_i), \max_i(x_i)]$. The proposed methods can also easily be expanded to additive semiparametric models with covariate vector $\mathbf{x}$.

The remainder of this paper is organized as follows: Section 5.2 summarizes the splines specifications used in this paper and gives the hat matrix / smoothing matrix for these. Different suggestions to predict y for out-of-sample observations from a

spline estimation are presented in Section 5.3. These are compared using a Monte Carlo study in Section 5.4 and using two empirical examples in Section 5.5. Further issues concerning prediction are discussed in Section 5.6. These include the analysis of large prediction horizons as well as a comparison to kernel-based predictions. Section 5.7 concludes.

5.2 Splines and their hat matrix

5.2.1 Splines with penalties and monotonicity constraints

A spline can be used to approximate other functions (e.g. Ruppert et al., 2003, de Boor, 2001, or Dierckx, 1993 as general references for splines). It consists of piecewise polynomial functions which are connected at knots and satisfy certain continuity conditions at these knots. The order of the piecewise polynomial functions is determined by the order k of the spline. The knot sequence $\boldsymbol{\kappa} = (\kappa_{-(k-1)}, \ldots, \kappa_{m+k})$ consists of $m + 2k$ non-decreasing knot positions κ_j, $j = -(k-1), \ldots, m+k$, where κ_0 and κ_{m+1} are the boundary knots that usually coincide with the bounds of the interval of interest, i.e. in the case of a scalar covariate x these are $\kappa_0 = \min_i(x_i)$ and $\kappa_{m+1} = \max_i(x_i)$. Then the spline is defined by the function

$$s(\cdot) = \sum_{j=-(k-1)}^{m} \alpha_j B_j^{\boldsymbol{\kappa},k}(\cdot), \tag{5.1}$$

which is a weighted sum of the B-spline basis functions $B_j^{\boldsymbol{\kappa},k}$. Each of the basis functions $B_j^{\boldsymbol{\kappa},k}$ is positive on the interval (κ_j, κ_{j+k}) and zero outside. Those basis functions and their unweighted (i.e. $\alpha_j = 1$ for all j) sum, which is 1 on $[\kappa_0, \kappa_{m+1}]$, are displayed as the solid lines in Figure 5.1.

For regression purposes, splines can be used to estimate the unknown regression curve. Consider the bivariate functional relationship

$$\mathrm{E}(y|x) = f(x), \tag{5.2}$$

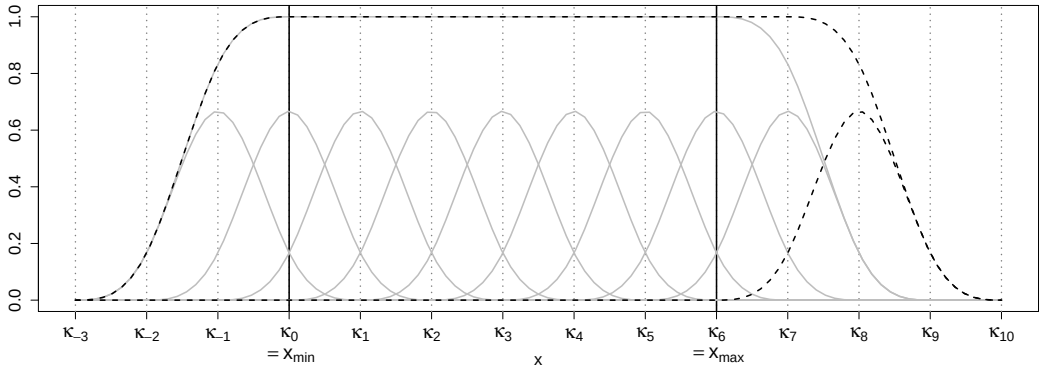

Figure 5.1: Grey, solid: Cubic ($k = 4$) B-spline basis functions for equidistant knot sequence with $m = 5$ inner knots and their unweighted sum. **Black, dashed**: One additional basis function on the right and resulting sum.

where f is to be estimated using spline regression, i.e. minimizing

$$\sum_{i=1}^{n} \left(y_i - \tilde{f}(x_i)\right)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=-(k-1)}^{m} \tilde{\alpha}_j B_j^{\boldsymbol{\kappa},k}(x_i) \right)^2 \tag{5.3}$$

with respect to the $m + k$ parameters $\tilde{\alpha}_j$ for a given sample $i = 1, \ldots, n$.

Restricting the estimated parameters $\hat{\alpha}_j$ such that they are in decreasing order,

$$\hat{\alpha}_j \geq \hat{\alpha}_{j+1}, \quad j = -(k-1), \ldots, m-1, \tag{5.4}$$

ensures a monotone decreasing estimated spline function and analogously, an increasing function results for increasing parameters (e.g. Dierckx, 1993, Section 7.1).

Cubic splines ($k = 4$) are commonly used in practice (e.g. Bollaerts et al., 2006, Eilers & Marx, 1996). They are easy to handle, exhibit a good fit and can be subject to several constraints as for example monotonicity or convexity (cf. Dierckx, 1993, Sections 3.2, 7.1). Hence, cubic splines are also used here, though the prediction methods presented in Section 5.3 can be applied to other spline orders, too.

Together with the knot sequence, the order of the spline fully determines the functions $B_j^{\boldsymbol{\kappa},k}$ of the B-spline basis. Eilers & Marx (1996) and Ruppert et al. (2003, Section 3.4) state some studies where the choice of the knot sequence (i.e. the number and location of the knots) is automated but computationally expensive. However, if the knot sequence is restricted to be equidistant, only the number of knots has to be chosen. Using many knots can result in a rough fit, while using only few knots may not

reflect the conditional relationship (5.2) well. Hence, Eilers & Marx (1996) propose the use of quite many equidistant knots while penalizing a rough fit. This is achieved for example by avoiding large second-order differences of the estimated parameters $\hat{\alpha}_j$, i.e. by penalizing large $\Delta^2\tilde{\alpha}_j = \tilde{\alpha}_j - 2\tilde{\alpha}_{j-1} + \tilde{\alpha}_{j-2}$. The objective function of the resulting minimization problem then is

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=-(k-1)}^{m}\tilde{\alpha}_j B_j^{\boldsymbol{\kappa},k}(x_i)\right)^2 + \lambda \sum_{j=-(k-1)+2}^{m}(\Delta^2\tilde{\alpha}_j)^2, \tag{5.5}$$

where λ is the smoothing parameter which controls the amount of smoothing and has to be chosen by the researcher (see below). Note that for $\lambda = 0$ the unpenalized fit as in Equation (5.3) results and for $\lambda \to \infty$ the fit is given by a straight line (cf. Eilers & Marx, 1996 with cubic splines and a penalty on the second-order differences of the estimated parameters). Still the number of knots has to be specified, though this is not that influential (see Ruppert et al., 2003, Sections 5.1, 5.5). In this work it is essential to use equidistant knots since most of the proposed methods for prediction in Section 5.3 cannot be applied for non-equidistant knot sequences. The number of inner knots m is chosen to be the rounded value of

$$\min(n/10 - 1,\, 35). \tag{5.6}$$

This is similar to Ruppert (2002) who proposes to use roughly $\min(n/4, 35)$ inner knots as a rule of thumb. The reason to use $n/10 - 1$ here instead of $n/4$ is to ensure having at least one observation in each of the $m + 1$ knot intervals with a quite high probability. For data sets with $n > 354$ the two choices will be the same anyway.

Now only the smoothing parameter λ is left to be specified. It can be chosen for example by the (generalized) cross validation criterion (CV, GCV) or the Akaike information criterion (AIC) (e.g. summarized in Ruppert et al., 2003, Section 5.3). All these criteria are based on the elements of the diagonal of the hat matrix of the estimation, e.g. the dimension of the estimated model which is given by the trace of the hat matrix. Section 5.2.2 explains how to obtain the hat matrix for the considered splines estimations.

5.2.2 A hat matrix for monotonicity constrained P-splines

The hat matrix $\mathbf{H}$ of the minimization problem $\min_{\tilde{f}} \sum_{i=1}^{n}(y_i - \tilde{f}(\mathbf{x}_i))^2$ with covariate vector $\mathbf{x}_i$ is defined to be the matrix for which $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$. In case of a linear regression, i.e. $\min_{\tilde{\boldsymbol{\alpha}}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i\tilde{\boldsymbol{\alpha}})^2$, the hat matrix is given by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, where $\mathbf{X} = \begin{pmatrix}\mathbf{x}_1^T & \cdots & \mathbf{x}_i^T & \cdots & \mathbf{x}_n^T\end{pmatrix}^T$.

For penalized estimations with a general penalty matrix $\mathbf{D}$ (for an example see Equation (5.9)) where $\sum_{i=1}^{n}(y_i - \mathbf{x}_i\tilde{\boldsymbol{\alpha}})^2 + \lambda\tilde{\boldsymbol{\alpha}}^T\mathbf{D}\tilde{\boldsymbol{\alpha}}$ is minimized with respect to $\tilde{\boldsymbol{\alpha}}$, the hat matrix can be determined as

$$\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^T$$

(e.g. Ruppert et al., 2003, Section 3.10).

For regression problems with general inequality constraints but without penalty, i.e. $\min_{\tilde{\boldsymbol{\alpha}}} \sum_{i=1}^{n}(y_i - \mathbf{x}_i\tilde{\boldsymbol{\alpha}})^2$ subject to $\mathbf{C}\tilde{\boldsymbol{\alpha}} \geq \mathbf{0}$ (for an example see Equation (5.10)), the hat matrix can be derived from the work of Paula (1999) and is given by

$$\mathbf{H}_{\text{constr}} = \mathbf{X}\left(\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}_R^T(\mathbf{C}_R(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}_R^T)^{-1}\mathbf{C}_R\right)(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \qquad (5.7)$$

where the matrix $\mathbf{C}_R$ contains the rows of $\mathbf{C}$ satisfying $\mathbf{C}\hat{\boldsymbol{\alpha}} = \mathbf{0}$ (cf. Paula, 1993, 1999).

Penalized estimations with inequality constraints are obtained by minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i\tilde{\boldsymbol{\alpha}})^2 + \lambda\tilde{\boldsymbol{\alpha}}^T\mathbf{D}\tilde{\boldsymbol{\alpha}}$ subject to $\mathbf{C}\tilde{\boldsymbol{\alpha}} \geq \mathbf{0}$ with respect to $\tilde{\boldsymbol{\alpha}}$. As the penalized estimation without constraints can be interpreted as ordinary least-squares problem with $\mathbf{X}^* = \begin{pmatrix}\mathbf{X}^T & \sqrt{\lambda}(\mathbf{D}^{1/2})^T\end{pmatrix}^T$ and $\mathbf{y}^* = \begin{pmatrix}\mathbf{y}^T & \mathbf{0}^T\end{pmatrix}^T$ (e.g. Eilers & Marx, 1996), these two hat matrices can be combined, resulting in the hat matrix for inequality constrained penalized estimations:

$$\begin{aligned}\mathbf{H}_{\lambda,\text{ constr}} = \mathbf{X}\Big(\mathbf{I} - (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{C}_R^T(\mathbf{C}_R(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{C}_R^T)^{-1}\mathbf{C}_R\Big) \\ \cdot (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D})^{-1}\mathbf{X}^T.\end{aligned} \qquad (5.8)$$

For penalized monotonicity constrained spline estimations ((5.5) with constraint (5.4)) with equidistant knots for a scalar covariate x, the (i,j)-th entry of the $n \times (m+k)$ matrix $\mathbf{X}$ is $B_{j-k}^{\kappa,k}(x_i)$. The penalty matrix $\mathbf{D}$ is the matrix for which $\sum_{j=-(k-1)+2}^{m}(\Delta^2\tilde{\alpha}_j)^2 =$

$\tilde{\boldsymbol{\alpha}}^T \mathbf{D} \tilde{\boldsymbol{\alpha}}$ holds and hence is given by

$$\mathbf{D} = \begin{pmatrix} 1 & -2 & 1 & & & & & \\ -2 & 5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & \cdot & & & \\ & 1 & -4 & 6 & \cdot & \cdot & & \\ & & 1 & -4 & \cdot & \cdot & 1 & \\ & & & 1 & \cdot & \cdot & -4 & 1 \\ & & & & \cdot & \cdot & 6 & -4 & 1 \\ & & & & & \cdot & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \tag{5.9}$$

The constraint matrix $\mathbf{C}$ required to obtain a monotonically decreasing fit is

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & 1 & -1 & \\ & & & \ddots & \ddots \end{pmatrix} \tag{5.10}$$

and for a monotone increasing fit it has to be multiplied by -1.

5.3 Predictions using splines

Let $\hat{y}_i = \hat{f}(x_i)$ be the fitted value of $\mathrm{E}(y_i|x_i) = f(x_i)$ for the observation (y_i, x_i), $i = 1, \ldots, n$. Given the data-driven nature of a nonparametric function, outside the sample range, i.e. outside the interval $[\kappa_0, \kappa_{m+1}]$, the estimated spline function $\hat{f}$ cannot be continued in analogy to the line resulting from a fully parametric regression. For observations from the given data set no problems arise and the fitted values can be calculated. Also for observations with covariate-values that lie within the sample range, predictions can be made in a straightforward manner. However, if the covariate-value of an observation whose expected y is to be predicted is outside this interval, no obvious prediction method exists.

In the following, several ways for predictions for out-of-sample observations based on penalized spline estimation are suggested. Note that without loss of generality only predictions for observations with values of $x > \kappa_{m+1}$ but not for $x < \kappa_0$ are considered in the following.

The easiest way to predict observations with $x > \kappa_{m+1}$ is to use the same fitted value as at the boundary knot κ_{m+1}. Hence, constant prediction means that the fitted value is defined to be $\hat{f}(x) = \hat{f}(\kappa_{m+1})$ for $x > \kappa_{m+1}$.

The prediction $\hat{f}(x)$ for x could also be continued linearly for $x > \kappa_{m+1}$. That is,

for $x > \kappa_{m+1}$ it holds that $\hat{f}'(x) = \hat{f}'(\kappa_{m+1})$, where the intercept of $\hat{f}$ for $x > \kappa_{m+1}$ is chosen such that $\hat{f}$ is continuous at κ_{m+1}.[1]

An alternative approach is to enlarge the knot sequence $\boldsymbol{\kappa} = (\kappa_{-(k-1)}, \ldots, \kappa_{m+k})$ by one additional knot κ_{m+1+k}, resulting in the new knot sequence $\boldsymbol{\kappa}'$. Along with this, the B-spline basis is enlarged by one additional basis function $B_{m+1}^{\boldsymbol{\kappa}',k}$ which is positive on $(\kappa_{m+1}, \kappa_{m+1+k})$. Now the resulting spline is valid on the interval $[\kappa_0, \kappa_{m+2}]$ in contrast to the estimation where it is valid only on $[\kappa_0, \kappa_{m+1}]$. The functions of the basis $\boldsymbol{\kappa}'$ and their sum are shown in Figure 5.1 as combination of the solid and dashed lines. If the weight $\hat{\alpha}_{m+1}$ of the new basis function $B_{m+1}^{\boldsymbol{\kappa}',k}$ is known, predictions for all x within the interval $(\kappa_{m+1}, \kappa_{m+2}]$ are straightforward. The objective of the next paragraph is to find the weight for the new basis function.

The weights $\hat{\alpha}_j$ together with the knots κ_j, $j = -(k-1), \ldots, m$, may be interpreted as a (time) series. Figure 5.2 illustrates this representation for the motorcycle example (see Section 5.5.1). First, the weight $\hat{\alpha}_{m+1}$ is predicted by an AR(1) process using the "sample" $\hat{\alpha}_{-(k-1)}, \ldots, \hat{\alpha}_m$. Further, for monotone functions an additional trend is added to the AR(1) process or the AR(1) process is estimated in first differences. Moreover, an AR(2) process is estimated. Higher order AR(p) models are not employed in this work. The basis enlarging methods require equidistant data and quite many estimated parameters to estimate the AR process. Hence, they need to be based on a spline estimation with many equidistant knots what is met when using P-splines.

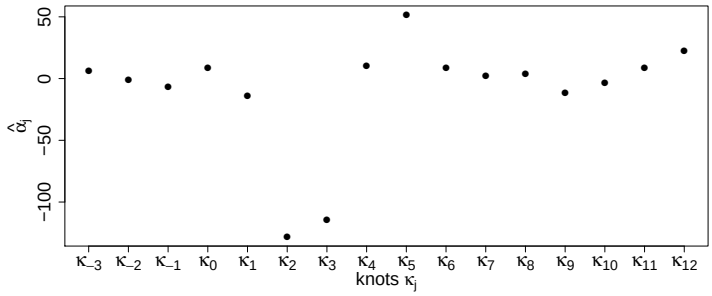

Figure 5.2: Plot of knot positions κ_j versus estimated spline parameters $\hat{\alpha}_j$ for the motorcycle data analyzed in Section 5.5.1.

[1] As cubic splines are used, additional to the constant and linear continuation of the fitted curve, quadratic pieces could be added. But this option is omitted here since the results are very unstable and not reliable as they are too sensitive with respect to the data. For cubic continuation the third derivative has to equal $\hat{f}^{(3)}(\kappa_{m+1})$ which is not defined for cubic splines as they are only differentiable up to order $k - 2 = 4 - 2 = 2$ at the knots.

Summarizing, the following prediction methods are presented:

class	method	description
boundary extrapolation	`con`	l-th order continuation of fitted curve at boundary knot with $l = 1$, i.e. constant continuation
	`lin`	$l = 2$, i.e. linear continuation
basis enlargement	`ar1`	estimate weight of new basis function of enlarged basis by an AR(1) process
	`art`	AR(1) process with trend (only monotone functions)
	`ard`	AR(1) process in first differences (only monotone functions)
	`ar2`	AR(2) process

Table 5.1: Classification and summary of prediction methods.

In the next sections the proposed prediction methods are compared. First in Section 5.4, an extensive Monte Carlo simulation is conducted. Thereafter in Section 5.5, the methods are applied to real data examples.

5.4 Monte Carlo simulation

5.4.1 Data generating processes

In this section only bivariate data generating processes (DGPs) are considered, i.e.

$$y = f(x) + u.$$

For all DGPs, the scalar covariate x and the errors u are assumed to be distributed as

$$x \sim U(0,1), \qquad u|x \sim N(0,\sigma^2).$$

To be able to use the same knot sequence for all replications, x is rescaled such that $\min_i(x_i) = 0$ and $\max_i(x_i) = 1$ for each sample. The nature of the results presented in Section 5.4.2 does not change for normally distributed x.

Six different regression functions f are studied:

$$
\begin{aligned}
f_1(x) &= 1 - x + 0.0625 \, \sin(5\pi \, x), \\
f_2(x) &= \frac{1}{1 + \exp(10(x - 0.5))}, \\
f_3(x) &= 0.5 + \sqrt{x(1-x)} \cdot \sin\left(\frac{2\pi \left(1 + 2^{-0.6}\right)}{x + 2^{-0.6}}\right), \\
f_4(x) &= 0.4 \left(x + 2 \exp\left(-(16(x - 0.5))^2\right)\right), \\
f_5(x) &= x + 0.2 \, \sin(6\pi \, x), \\
f_6(x) &= 1 - x.
\end{aligned}
$$

The functions f_1, f_2 and f_6 are monotonically decreasing, where f_1 is a shifted sine function with decreasing trend, f_2 is the mirrored CDF of the logistic distribution with parameters $a = 0.5$ and $b = 0.1$ and f_6 is a simple linear function. The function f_3 was also studied (up to the summand 0.5) for example in the works of Wand (2000) and Ruppert (2002). The so called bump function f_4 was e.g. used (up to the factor 0.4) by Ruppert (2002) and with variations by Hurvich et al. (1998) and Crainiceanu et al. (2007). Finally, f_5 is the sine function with higher periodicity and an increasing trend. All functions are chosen such that (approximately) $f(x) \in [0, 1]$ for $x \in [0, 1]$, hence the same error variance σ^2 is appropriate for all DGPs and is chosen to equal $\sigma^2 = 0.09$. The lower panels of Figures 5.3 to 5.8 show the respective functions f.

For all estimations the open source software `R` (`www.r-project.com`) is used. The (constrained) P-spline regressions are based on the base package `splines` and the `mgcv` package from Wood (2011).

5.4.2 Simulation results

For each replication $r = 1, \ldots, R$, $R = 1000$, of the Monte Carlo simulation, a sample of size $n = 300$ is drawn for x and u and the corresponding y for the regression functions f_1 to f_6 are calculated. According to Equation (5.6), the knot sequence for the spline basis $\boldsymbol{\kappa}$ contains $m = 29$ equidistant inner knots and is subject to the constraint (5.4) for f_1, f_2 and f_6. For each function the smoothing parameter λ has to

be chosen. In the simulations the true function f is known. Hence, λ can be selected for each of the six functions f from Section 5.4.1 by minimizing the mean integrated squared error ($MISE$). Then the smoothing parameter is chosen as

$$\lambda = \arg\min_{\tilde{\lambda}} \frac{1}{R} \sum_{r=1}^{R} \frac{1}{n} \sum_{i=1}^{n} \left(f(x_{i,r}) - \hat{f}_{\tilde{\lambda},r}(x_{i,r}) \right)^2,$$

where $x_{i,r}$ is the ith observation in the rth replication and $\hat{f}_{\tilde{\lambda},r}$ is the estimate of f for the rth replication and a given value $\tilde{\lambda}$ for the smoothing parameter. The results for λ chosen by minimizing the $MISE$ largely coincide with the smoothing parameter analogously obtained by the average GCV criterion. The latter is feasible for real data problems when f is unknown and hence is applied for the empirical examples in Section 5.5.

Then, for a given sample the observations are partitioned several times. Those observations with $x \in [\kappa_0, \kappa_t]$, $t = s, \ldots, m$, are used to estimate the respective part of f and the observations with $x \in (\kappa_t, \kappa_{t+1}]$ are used to evaluate the predictive performance of the different methods discussed in Section 5.3. The value for s is chosen such that at least approximately half of the data is available for the estimation, i.e. in the given example s is chosen to be 15. For each interval $I_t = (\kappa_t, \kappa_{t+1}]$ used for prediction, the average squared error of prediction ($ASEP$) is calculated by

$$ASEP_{t,r} = \frac{1}{n_{t,r}} \sum_{i,\, x_{i,r} \in I_t} (y_{i,r} - \hat{y}_{i,r})^2,$$

where $n_{t,r}$ is the number of observations $(y_{i,r}, x_{i,r})$ with $x_{i,r} \in I_t$ for the rth replication and $\hat{y}_{i,r}$ is the prediction for $\mathrm{E}(y_{i,r}|x_{i,r})$ for the rth replication.

The numbers in the gray boxes in the middle panel of Figures 5.3 to 5.8 give the percentage of replications for which the respective prediction method has the lowest $ASEP$. The darker the box, the higher the corresponding percentage. It can be observed for all DGPs that predicting the out-of-sample observations by constant (`con`) continuation is the best alternative (i.e. has the lowest $AESP$) in most cases. For the monotone DGPs, the AR(1) methods with (`art`) and without (`ar1`) trend also show quite low values of $ASEP$ for some intervals. For intervals where the curvature changes, prediction using the AR(2) method (`ar2`) is often the best. For intervals with

(approximately) linear f the linear method (`lin`) also performs well. This is the case throughout the interval $[\kappa_s, \kappa_{m+1}]$ for the linear function f_6, but still the method `con` has the lowest $ASEP$ in most cases.

Up to now, only the method with the lowest $ASEP$ was taken into account. The advantage over the other methods in terms of $ASEP$ may be considered as well, hence the average $ASEP$ across the replications is considered. Further approaches that may be used to evaluate the results may be found in the work of Haupt et al. (2011). The upper panels of Figures 5.3 to 5.8 show the average $ASEP$ across the replications in the respective interval, i.e.

$$\overline{ASEP}_t = \frac{1}{R} \sum_{r=1}^{R} ASEP_{t,r}.$$

It can be observed that all methods except the constant prediction exhibit very similar average $ASEP$s. Hence, the linear prediction method and those enlarging the B-spline basis by one knot cannot be clearly distinguished from each other in most cases in terms of average $ASEP$. Further, it can be observed that the constant prediction outperforms the other methods in terms of lowest average $ASEP$. Additionally, the highest average $ASEP$ from the constant prediction is lower than the highest average $ASEP$ from the other prediction methods for most DGPs and if that is not the case, the differences are mostly negligible. This suggests that the constant prediction method bears the smallest risk for a wide range of DGPs.

For practical proposes, the rightmost interval $I_m = (\kappa_m, \kappa_{m+1}]$ is of crucial interest. For each of the $R = 1000$ replications the prediction method which yields the lowest $ASEP$ in most of the intervals I_t for $t = s, \ldots, m+1$ is compared to the prediction method which has the lowest $ASEP$ in the rightmost prediction interval I_m in the respective replication. The first row of Table 5.2 shows the percentage of the R replications for which they coincide. This percentage corresponds approximately to the percentage of replications where the constant prediction method is preferred in terms of lowest $ASEP$ in interval I_m (see Figures 5.3 to 5.8). Hence, again the outstanding role of the constant prediction method becomes obvious. The second line of Table 5.2 gives an alternative way to evaluate the rightmost interval I_m. Therefore, first for each of the R replications the share of intervals I_t, $t = s+1, \ldots, m$, where

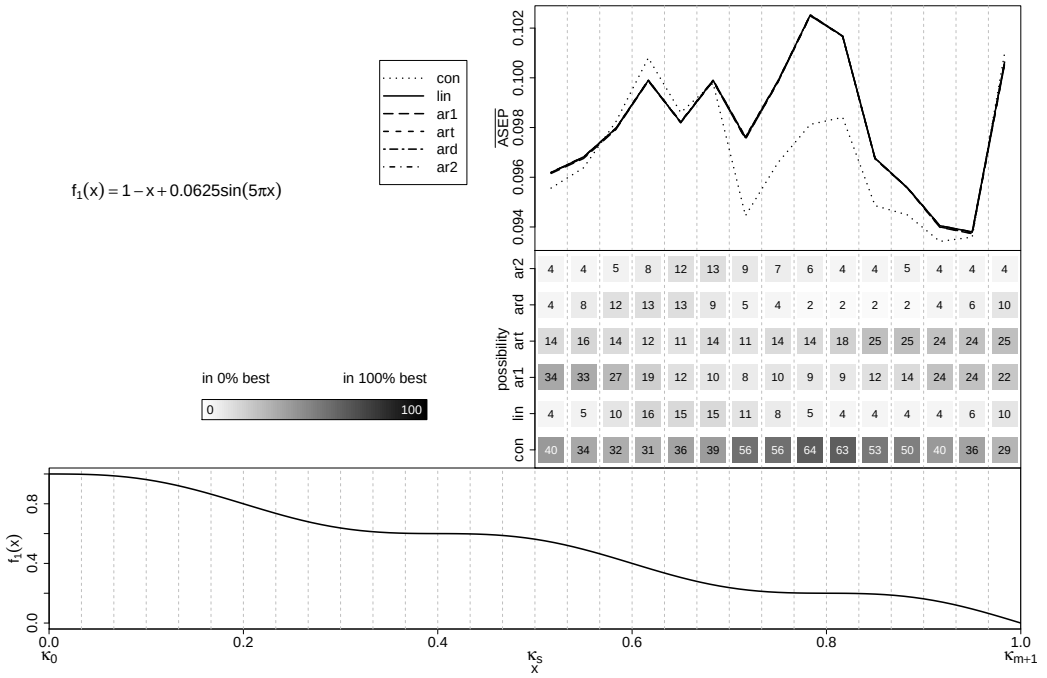

Figure 5.3: Lower panel: $f_1(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

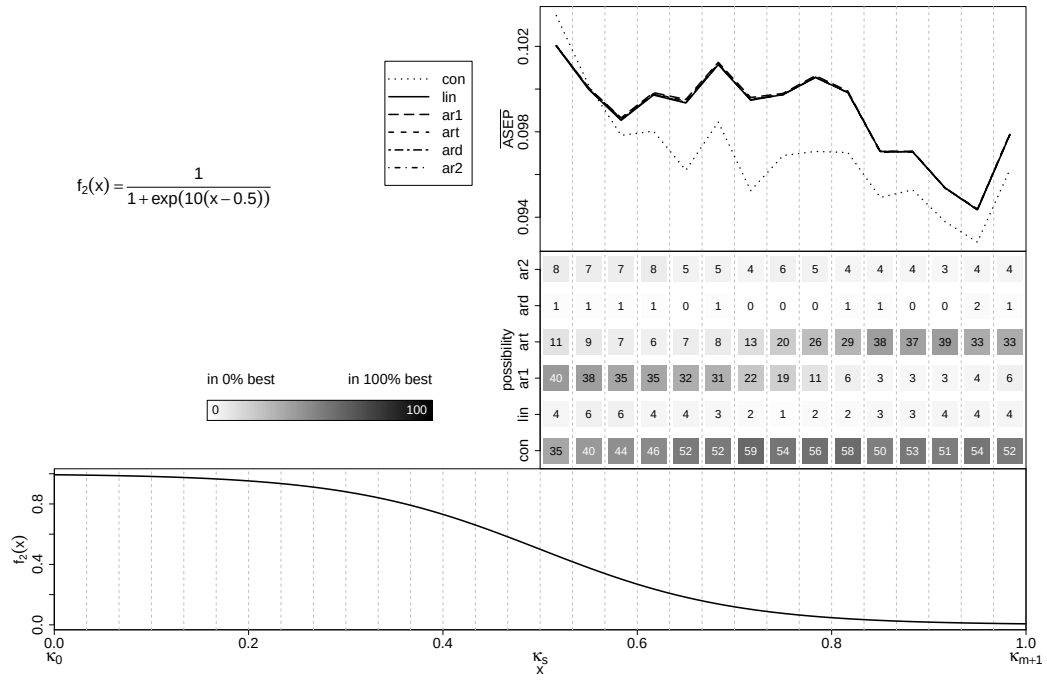

Figure 5.4: Lower panel: $f_2(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

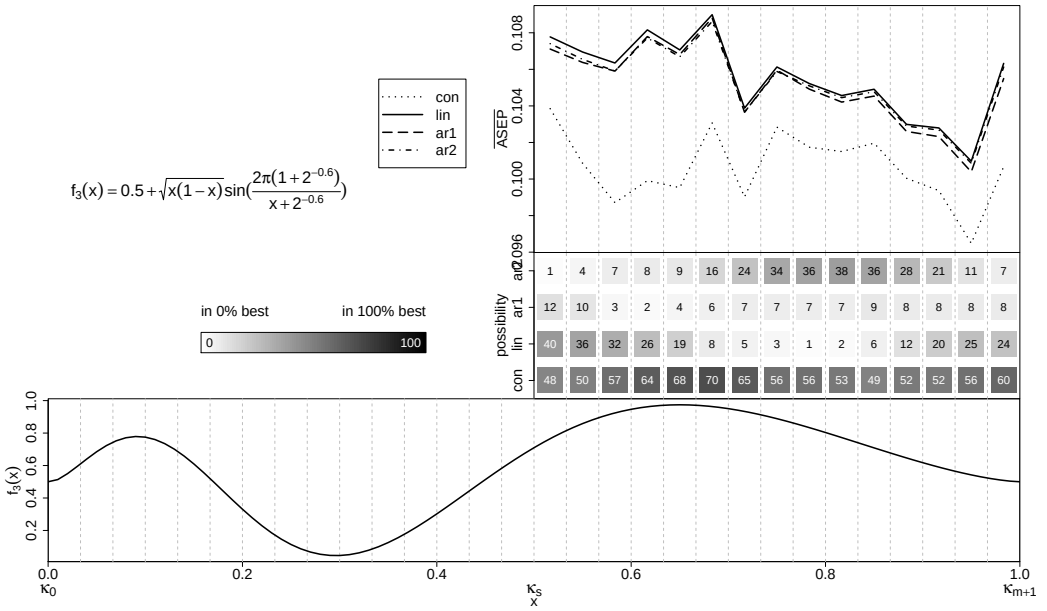

Figure 5.5: Lower panel: $f_3(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

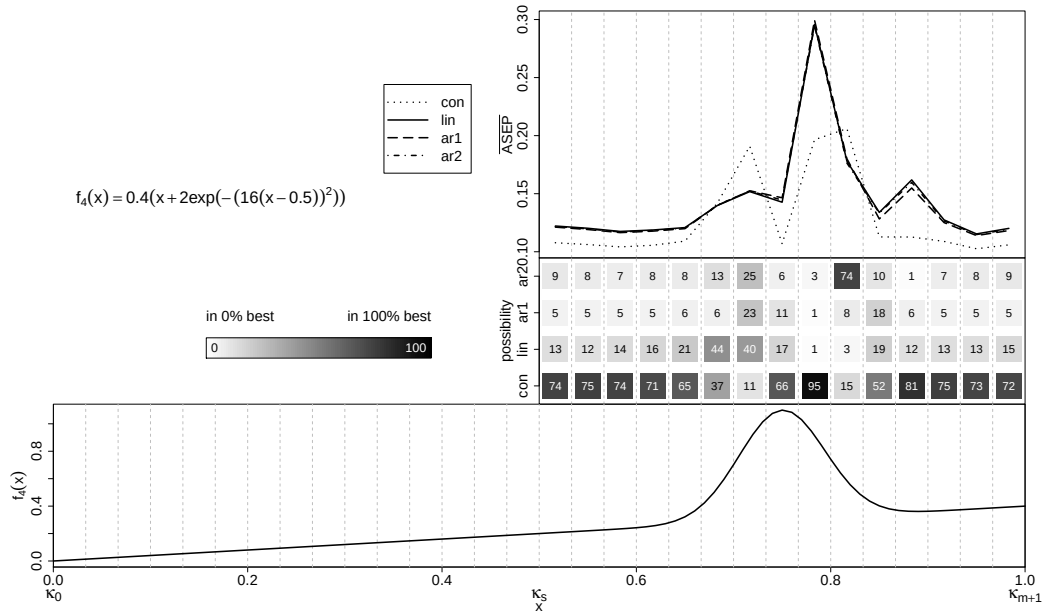

Figure 5.6: Lower panel: $f_4(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

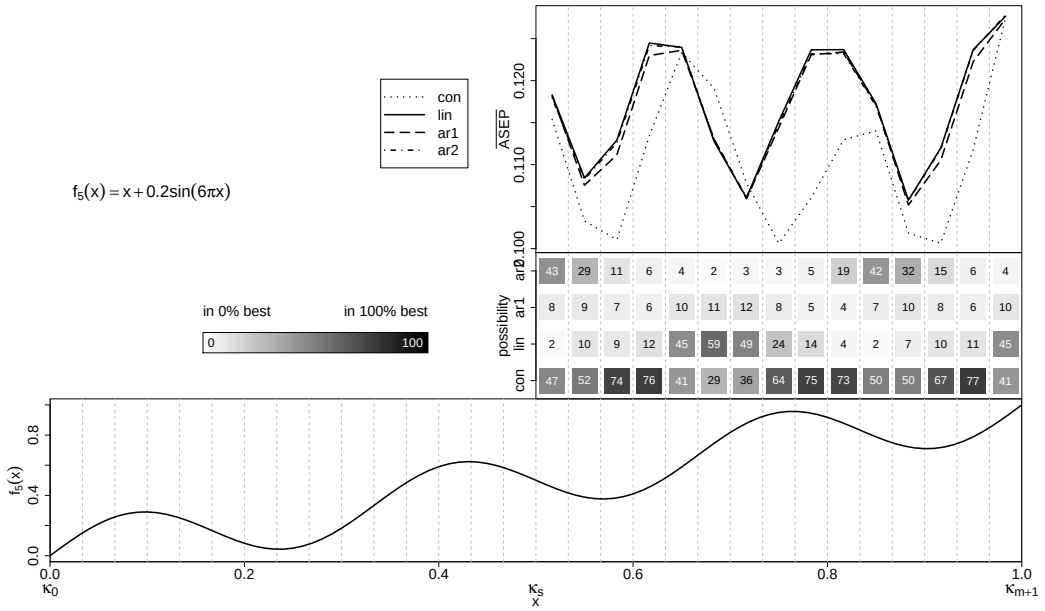

Figure 5.7: Lower panel: $f_5(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

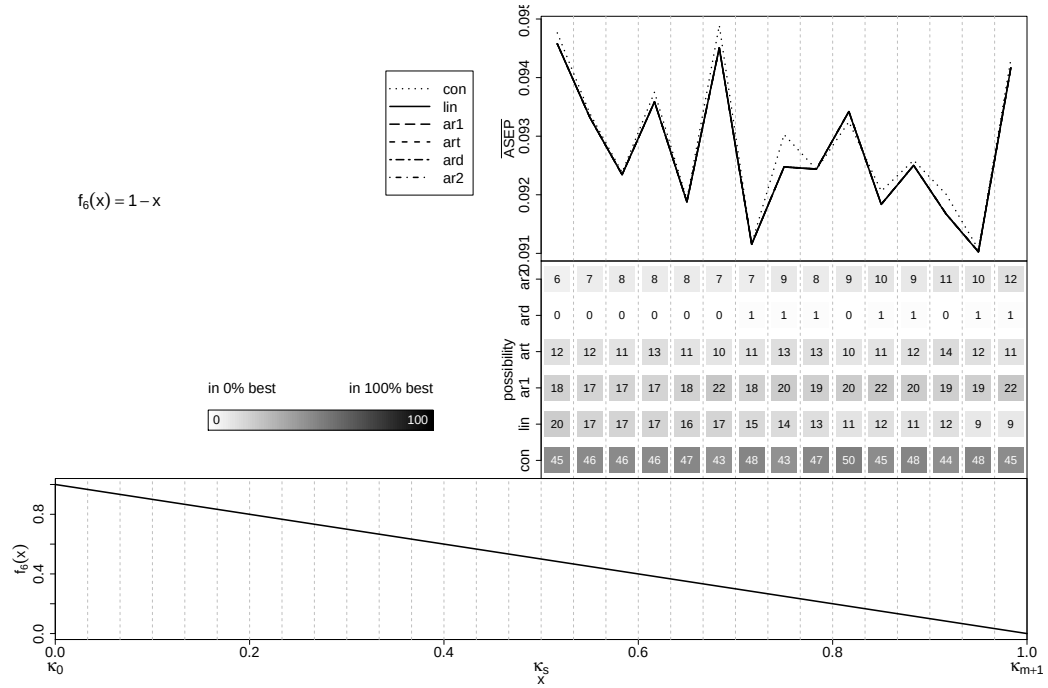

Figure 5.8: Lower panel: $f_6(x)$ with inner knots and boundary knots κ_0 and κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of cases where the respective prediction method has the lowest $ASEP$ in the respective knot interval. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval.

the prediction method with lowest $ASEP$ is the same in I_t and I_{t-1} is calculated and then this is averaged over the R replications. The latter share is about 29-34% for the monotone DPGs and about 38-41% for the non-monotone DGPs.

In summary the Monte Carlo results strongly favor the constant prediction method since it yields the lowest $ASEP$ in most replications as well as the lowest average $ASEP$ across the replications for many of the prediction intervals. Further, comparing the highest average $ASEP$s also justifies the constant prediction method. Hence, the results suggest to apply the constant prediction method in case of doubt.

functions	f_1	f_2	f_3	f_4	f_5	f_6
P(most times best before m equals best in m)	0.295	0.504	0.597	0.714	0.409	0.415
P(best in t equals best in $t-1$)	0.288	0.339	0.381	0.407	0.384	0.344

Table 5.2: Upper line: Share of R replications where the prediction method with lowest $ASEP$ for knot interval $I_m = (\kappa_m, \kappa_{m+1}]$ is the same as the method with lowest $ASEP$ for the majority of knot intervals $I_t = (\kappa_t, \kappa_{t+1}]$, $t = s, \ldots, m-1$. **Lower line**: Share of $R \times (m-s)$ cases where the prediction method with lowest $ASEP$ for knot interval $I_t = (\kappa_t, \kappa_{t+1}]$, $t = s+1, \ldots, m$, is the same as for knot interval $I_{t-1} = (\kappa_{t-1}, \kappa_t]$.

5.5 Empirical examples

In empirical practice usually only one data set is available. To conduct analyses in analogy to the Monte Carlo simulation, subsamples from the full sample can be drawn. For the presented examples the proportion of observations in the subsamples is 80 percent of all observations.

In the following, the two well-known datasets containing motorcycle acceleration and LIDAR data, respectively, are analyzed. In the latter case the relationship is assumed to be monotone.

For both examples, cubic splines are applied, m is chosen according to Equation (5.6) and λ with respect to $GCV(\lambda) = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(1-\frac{1}{n}\text{tr}(\mathbf{H}))^2}$ (e.g. Ruppert et al., 2003, Section 5.3) with $\mathbf{H}$ being the respective hat matrix as in Equation (5.7) or (5.8), respectively. For the monotone LIDAR example, constraint (5.4) is applied.

5.5.1 Motorcycle acceleration

First, the motorcycle data set which has been examined especially with nonparametric techniques (e.g. Silverman, 1985, Eilers & Marx, 1996, Yu et al., 2003) is analyzed. It is available for example in the R-package MASS from Venables & Ripley (2002). The data set includes information on $n = 133$ observations from a simulated motorcycle accident where the acceleration (y) is given for several time points after the impact (x).

The subsampling prediction results can be inspected in Figure 5.9. To demonstrate the analogy for predictions to the left, the results for the latter can be found for this example in Figure 5.10. The results are similar to those in the simulations: the linear and the basis enlarging methods show similar results among each other in most cases. Further, the constant prediction method is the mostly favored alternative. It yields the lowest $ASEP$ in most cases and the lowest average $ASEP$ for most intervals. In regions where the curvature changes the AR(2) method performs well, too.

For the empirical examples the full-sample results are also evaluated in addition to the results from the $R = 1000$ subsamples. The $ASEP$ from the full sample for each method is illustrated as symbols in the upper panels of Figures 5.9 and 5.10 and the lowest line in the middle panels states which method has the lowest $ASEP$ in the full sample. In most intervals again the constant prediction method is superior in terms of $ASEP$. Hence, the full-sample results confirm the subsampling results.

5.5.2 LIDAR

The LIDAR (light detection and ranging) data set with $n = 221$ is the second example. It also has been analyzed in several nonparametric studies (e.g. Ruppert et al., 2003, Ruppert & Carroll, 2000, Schnabel & Eilers, 2009) and can be found on the homepage of the book of Ruppert et al. (2003, `http://www.uow.edu.au/~mwand/webspr/data.html`). The dependent variable y is $logratio$, the logarithm of the ratio of received light from two laser sources, which is explained by the covariate $x = range$, the distance the light traveled before being reflected back to its source. For this

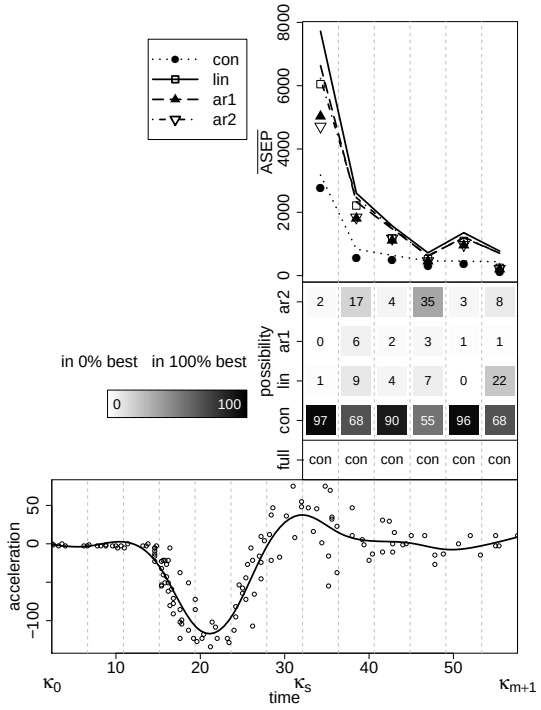

Figure 5.9: Predictions to the right. **Lower panel**: Motorcycle data and estimated regression function with knots κ_0 to κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of subsamples where the respective prediction method has the lowest $ASEP$ in the respective knot interval and additional full-sample results. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval and additional full-sample results.

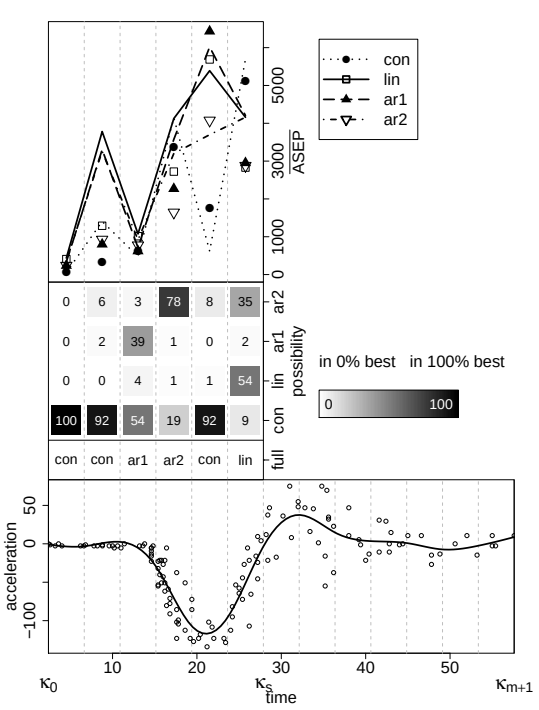

Figure 5.10: Predictions to the left. **Lower panel**: Motorcycle data and estimated regression function with knots κ_0 to κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of subsamples where the respective prediction method has the lowest $ASEP$ in the respective knot interval and additional full-sample results. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval and additional full-sample results.

example the fitted regression curve is restricted to be monotone decreasing.

The estimation results are summarized in Figure 5.11 and confirm the conclusions from the simulation study: again, the constant method performs best and like in the simulations with monotone DGPs, the AR(1) methods with and without trend also perform quite well. Further, the full-sample results lead to the same conclusions.

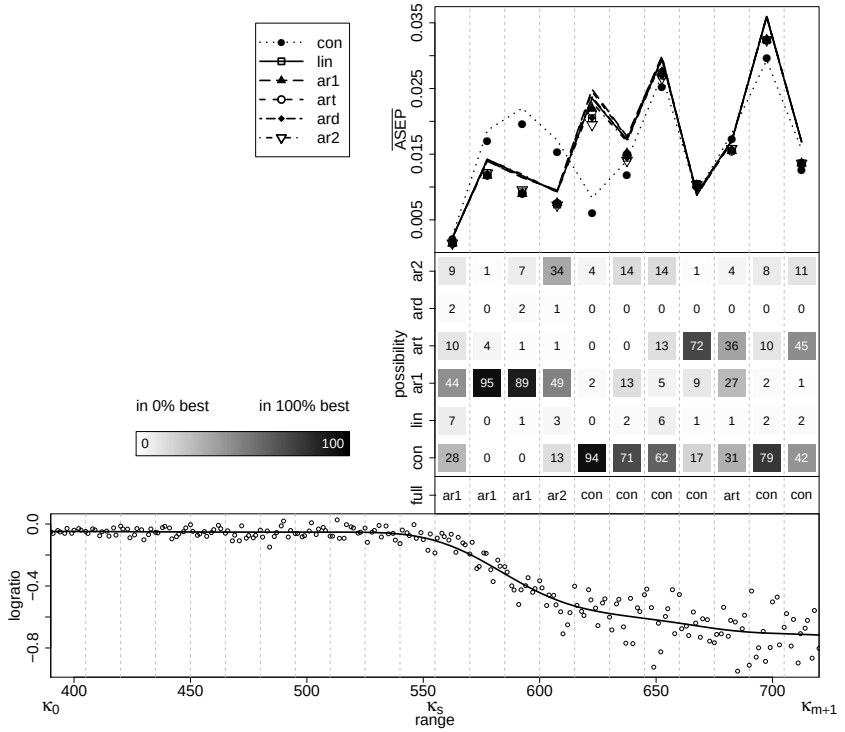

Figure 5.11: Lower panel: LIDAR data and estimated regression function with knots κ_0 to κ_{m+1} from knot sequence $\boldsymbol{\kappa}$. **Middle panel**: Percentage of subsamples where the respective prediction method has the lowest $ASEP$ in the respective knot interval and additional full-sample results. **Upper panel**: Average $ASEP$ for the respective prediction method in the respective knot interval and additional full-sample results.

5.6 Extensions

5.6.1 Larger prediction horizons

Up to now, prediction was considered only within one knot interval from the boundary knot, i.e. for observations whose covariate value lies within $I_{m+1} = (\kappa_{m+1}, \kappa_{m+2}]$. The methods proposed in this paper can easily be extended to predictions within two (i.e. predictions for $x \in I_{m+2} = (\kappa_{m+2}, \kappa_{m+3}]$) or three (i.e. predictions for $x \in I_{m+3} = (\kappa_{m+3}, \kappa_{m+4}]$) or even more knot intervals. In the following, predictions based on estimations with data from the interval $[\kappa_0, \kappa_{m+1}]$ are made for the interval I_{m+2} (I_{m+3}, ..., I_{m+6}, respectively). That is, the prediction horizon is $2h$ ($3h$, ..., $6h$, respectively), where h denotes the distance of two adjacent knots in the equidistant knot sequence.

Figures 5.12 to 5.17 show the average $ASEP$ results for the examples from the Monte Carlo simulation in Section 5.4 for the prediction horizons h to $6h$ denoted as $\overline{ASEP}(1)$ to $\overline{ASEP}(6)$. There is a striking pattern that is moving to the right with increasing prediction horizon. It reflects the dependence of the prediction (no matter what the prediction horizon is) on the data at the boundary of the estimation sample. For example for f_4, in Figure 5.15 can be observed that the prediction interval with the highest average $ASEP$ is always the one for which the respective estimation is based on the estimation sample that contains the bump of f_4 in the right boundary interval.

Still the non-constant prediction methods show a similar pattern in the average $ASEP$ among each other. As may be expected, however this feature diminishes with increasing prediction horizon especially for the non-monotone functions f_3 to f_5. Overall it remains unclear which one of the non-constant methods is to be preferred as this varies across DGPs and prediction horizons.

The constant prediction method behaves similarly across all prediction horizons and its average $ASEP$ does not increase as much as those of the other prediction methods. Although for the monotone functions the constant method is outperformed by the others in some intervals, it overall promises the most stable results.

The only exception is the linear DGP with regression function f_6. Here, the non-constant prediction methods perform better than the constant method with increasing prediction horizon. The reason is that the linear pattern of f_6 does not change outside the data range. Hence, continuing the regression curve linear is just appropriate and the AR methods can fit the next coefficient quite exactly, too. However, it can be seen that the magnitude of the average $ASEP$ for the constant prediction method still is reasonable when comparing the results with those for functions f_1 to f_5.

5.6.2 Predictions with minimal penalty

In Section 5.3 several methods for estimating α_{m+1} by AR techniques were discussed. Since splines with a second-order difference penalty are used for the regression, the weight $\hat{\alpha}_{m+1}$ can further be chosen such that the resulting additional term $(\Delta^2\hat{\alpha}_{m+1})^2$ in the penalty is minimal, i.e. is 0. Solving for $\hat{\alpha}_{m+1}$ results in the weight $\hat{\alpha}_{m+1} =$

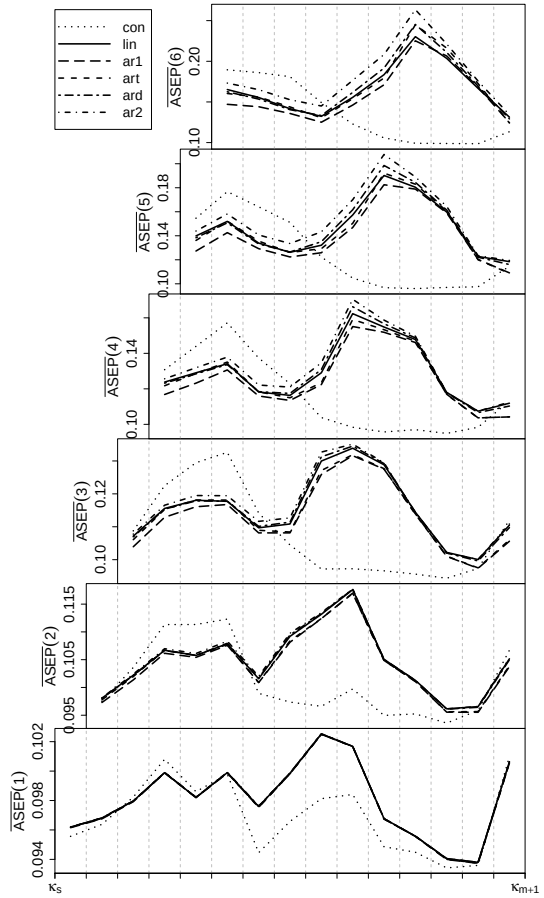

Figure 5.12: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_1(x)$ and a 1-, ..., 6-intervals horizon.

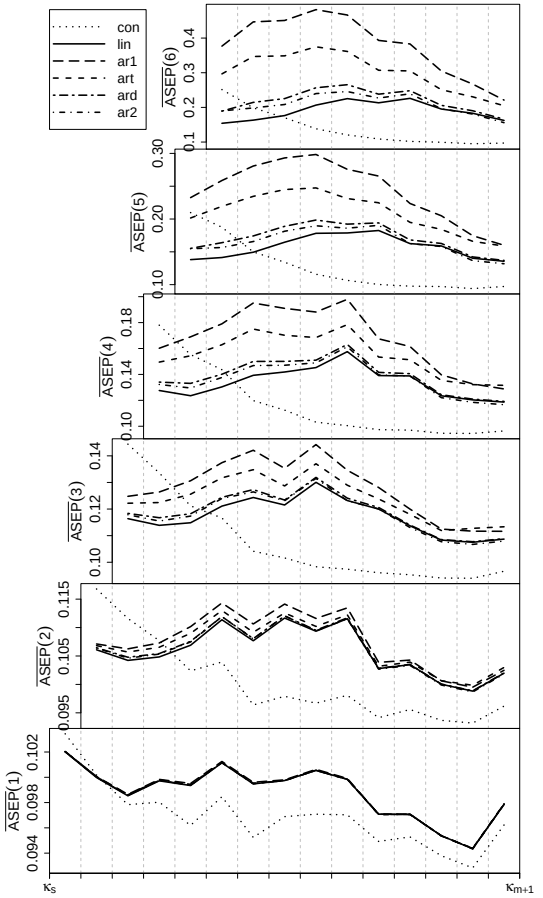

Figure 5.13: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_2(x)$ and a 1-, ..., 6-intervals horizon.

$2\hat{\alpha}_m - \hat{\alpha}_{m-1}$.

Note that if $\hat{\alpha}_{j-3}, \ldots, \alpha_j$ is linear in j, i.e. $\Delta^2\hat{\alpha}_j = \Delta^2\hat{\alpha}_{j-1} = 0$, the cubic spline $\sum_{j=-3}^{m} \alpha_j B_j^{\boldsymbol{\kappa},k}(x)$ is linear on the interval $[\kappa_j, \kappa_{j+1}]$ (cf. Eilers & Marx, 1996). Hence, the method of choosing $\hat{\alpha}_{m+1}$ to be $2\hat{\alpha}_m - \hat{\alpha}_{m-1}$ is very similar to the linear alternative. The results for continuing the spline outside the interval $[\kappa_0, \kappa_{m+1}]$ using this method are not presented since they largely coincide with those for the linear method as may be expected.

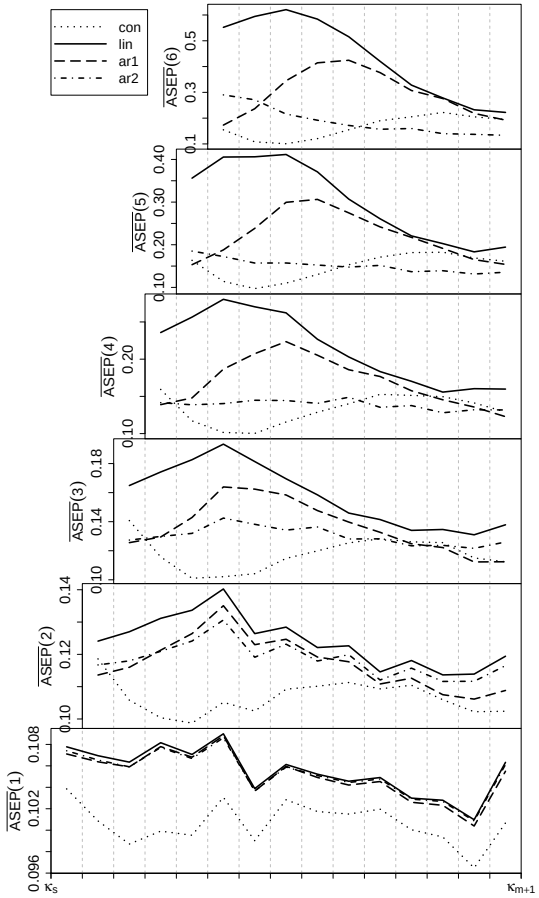

Figure 5.14: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_3(x)$ and a 1-, ..., 6-intervals horizon.

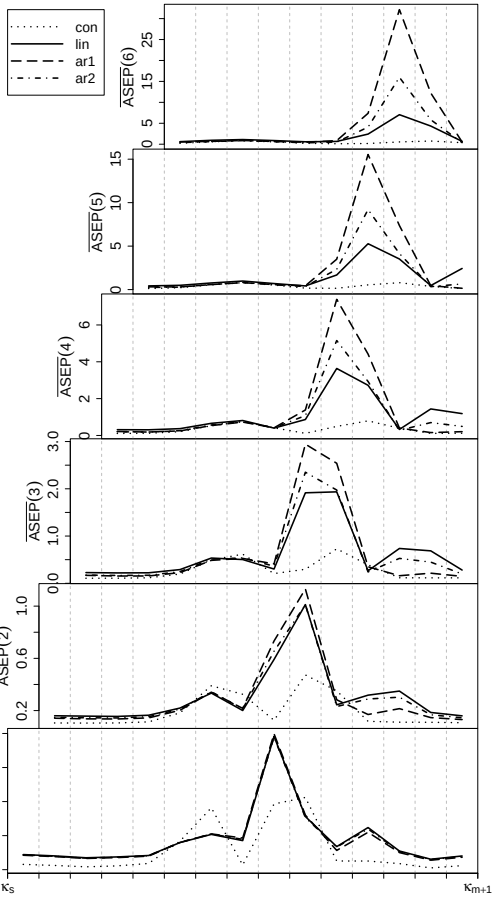

Figure 5.15: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_4(x)$ and a 1-, ..., 6-intervals horizon.

5.6.3 Re-locate boundary knot

In cases where only the prediction for out-of-sample observations is of interest but not the estimate of f on $[\kappa_0, \kappa_{m+1}]$, another method can be added to the analysis. If the boundary knot κ_{m+1} is re-located to the maximum value of x of the prediction observations, predictions can be made straightforward after the estimation. Re-locating κ_{m+1} hence means to give up the assumption of equidistant knots between the knots κ_m and κ_{m+1}. Further, the penalty matrix $\mathbf{D}$ is different from the matrix given in Section 5.2.2 (derivation using the derivative in de Boor, 2001, Chapter X). With the re-located knot κ_{m+1}, another knot sequence and with it another estimated regression curve on

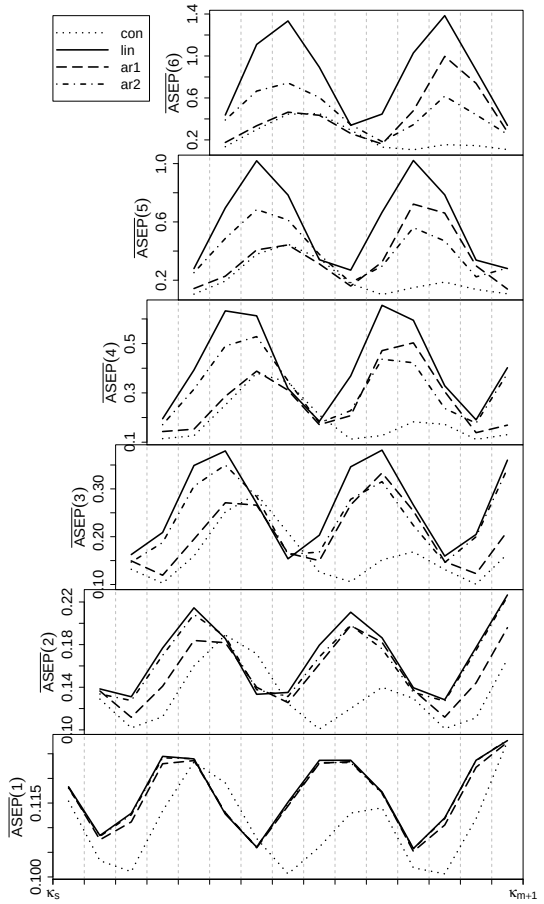

Figure 5.16: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_5(x)$ and a 1-, ..., 6-intervals horizon.

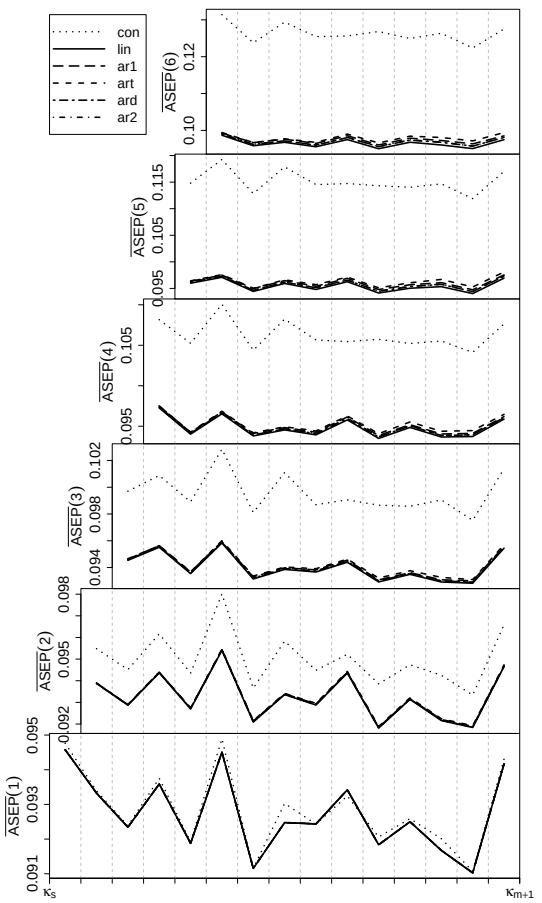

Figure 5.17: Average $ASEP$ for the respective prediction method in the respective knot interval for $f_6(x)$ and a 1-, ..., 6-intervals horizon.

$[\min_i(x_i), \max_i(x_i)]$ is obtained. The latter is the reason why this method was not regarded in the previous sections as the in-sample properties would change. The results for the newly included prediction method are quite diverse. While it performs well for some the monotone DGPs, the average $ASEP$ is up to 15 times and even 50 times the average $ASEP$ of the methods from Section 5.3 for f_5 and f_3, respectively. Since the observations from the prediction sample were not included in the optimization process, the results for the non-monotone DGPs are not satisfactory in intervals containing local extrema.

5.6.4 Kernel estimation

For an additional comparison, kernel estimates and predictions are calculated for the Monte Carlo study. Note that the estimation results and the fitted values within the interval $[\min_i(x_i), \max_i(x_i)]$ are different from those from the splines estimations. Two variants of kernel estimation (summarized e.g. in Ruppert et al., 2003, Section 3.15.1, Racine & Li, 2004) are considered: local constant and local linear estimation. For both a second order Gaussian kernel is applied and the bandwidth is chosen by least squares cross validation. Computations are carried out using the R-package `np` from Hayfield & Racine (2011).

Both variants show controversial results. In general they constitute the best variant in terms of $ASEP$ for many of the R replications. But in terms of average $ASEP$ they perform worse than the other presented prediction methods. This implies that their results are very volatile. In detail, with increasing prediction horizon, the difference of the average $ASEP$ of the local linear predictions and that of the remaining prediction methods becomes enormous. Hence, this method does not constitute a reasonable alternative. The average $ASEP$ lines of the local constant predictions run similar to those of the constant prediction method `con`, though the latter usually is minor. This analogy especially appears for the non-monotone functions. When comparing only `con` and the local constant predictions it can be observed that the kernel method outperforms the spline method slightly for intervals where the slope of the respective non-monotone function changes. This is due to the boundary bias (summarized e.g. in Ruppert et al., 2003, Section 3.15.1) of the local constant estimation, which is an advantage in this circumstance.

Summarizing, predictions using splines are in general more appropriate for out-of-sample predictions than kernel based predictions since they are more reliable with respect to their lower variability in terms of $ASEP$ even for larger prediction horizons and for monotone functions as well as for non-monotone functions.

5.7 Conclusion

For spline estimation it is not feasible to continue the estimated regression curve from a bivariate regression outside the boundary knots and predict the expected value of y given an x that lies outside the boundary knots. Hence, several methods to circumvent this problem are presented. These include continuing the estimated regression curve constantly or linearly at the boundary knots. Alternatively, the given knot sequence is enlarged by one knot and the parameter of the additional B-spline basis function is estimated using different AR techniques.

Extensive Monte Carlo simulations as well as two empirical examples lead to the conclusion that the constant continuation at the boundary knot is the most reliable alternative for predictions for out-of-sample observations. Further, it is observed that the methods that estimate the parameter for the additional basis function exhibit similar results to the linear continuation method. The constant continuation from the splines estimation moreover outperforms kernel-based local constant and local linear predictions.

The proposed approach can easily be adapted to additive regression settings with several other covariates that can be modeled either parametrically or using splines. Moreover, the suggested approach can be applied to quantile regression, when estimating conditional quantiles instead of the conditional mean of y.

Acknowledgments

Thanks to Harry Haupt, Joachim Schnurbus, Rolf Tschernig, Enzo Weber and Roland Weigand for many very helpful comments.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, 19(6), 716–723.

Bollaerts, K., Eilers, P. H. C., & Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L_1-norm. *Statistical Modelling*, 6(3), 189–207.

Cao, Y., Lin, H., Wu, T. Z., & Yu, Y. (2010). Penalized spline estimation for functional coefficient regression models. *Computational Statistics & Data Analysis*, 54(4), 891–905.

Chernozhukov, V., Fernández-Val, I., & Galichon, A. (2010). Quantile and probability curves without crossing. *Econometrica*, 78(3), 1093–1125.

Claeskens, G., Krivobokova, T., & Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3), 529–544.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., & Goodner, B. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational & Graphical Statistics*, 16(2), 265–288.

Davis, C. H. (1857). *Theory of the motion of the heavenly bodies moving about the sun in conic sections: a translation of Gauss's "Theoria Motus"*. Little, Brown and Company.

de Boor, C. (2001). *A Practical Guide to Splines*, volume 27 of *Applied Mathematical Sciences*. Berlin: Springer, revised edition.

Dierckx, P. (1993). *Curve and Surface Fitting with Splines*. Numerical Mathematics and Scientific Computation. Oxford: Oxford University Press.

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

Eilers, P. H. C. & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653.

Eubank, R. L. (1984). Approximate regression models and splines. *Communications in Statistics - Theory and Methods*, 13(4), 433–484.

Gauss, C. F. (1809). Theoria motus corporum coelestium. In *Carl Friedrich Gauss – Werke*. Königliche Gesellschaft der Wissenschaften zu Göttingen (1906).

Härdle, W. (1990). *Applied Nonparametric Regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, 1 edition.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Haupt, H. & Kagerer, K. (2012). Beyond mean estimates of price and promotional effects in scanner-panel sales-response regression. *Journal of Retailing and Consumer Services*, 19, 470–483.

Haupt, H., Kagerer, K., & Schnurbus, J. (2011). Cross-validating fit and predictive accuracy of nonlinear quantile regressions. *Journal of Applied Statistics*, 38(12), 2939–2954.

Haupt, H., Kagerer, K., & Steiner, W. J. (2013). Smooth quantile based modeling of brand sales, price and promotional effects from retail scanner panels. *Journal of Applied Econometrics*, (pp. n/a–n/a).

Hayfield, T. & Racine, J. S. (2011). *np: Nonparametric kernel smoothing methods for mixed data types*. R package version 0.40-7.

He, X. (1997). Quantile curves without crossing. *The American Statistician*, 51(2), 186–192.

He, X. & Ng, P. T. (1999). COBS: qualitatively constrained smoothing via linear programming. *Computational Statistics*, 14(3), 315–337.

He, X. & Shi, P. (1996). Bivariate tensor-product b-splines in a partly linear model. *Journal of Multivariate Analysis*, 58(2), 162–181.

He, X. & Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, 93(442), 643–650.

Huang, J. Z. & Shen, H. (2004). Functional coefficient regression models for non-linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics*, 31(4), 515–534.

Huang, J. Z., Wu, C. O., & Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14(3), 763–788.

Hurvich, C. M., Simonoff, J. S., & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society Series B - Statistical Methodology*, 60(Part 2), 271–293.

Imoto, S. & Konishi, S. (2003). Selection of smoothing parameters in B-spline non-parametric regression models using information criteria. *Annals of the Institute of Statistical Mathematics*, 55(4), 671–687.

Kauermann, G. (2005). A note on smoothing parameter selection for penalized spline smoothing. *Journal of Statistical Planning and Inference*, 127(1-2), 53–69.

Kimeldorf, G. S. & Wahba, G. (1970a). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41(2), 495–502.

Kimeldorf, G. S. & Wahba, G. (1970b). Spline functions and stochastic processes. *Sankhya: The Indian Journal of Statistics (A)*, 32, 173–180.

Kocherginsky, M., He, X., & Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational & Graphical Statistics*, 14(1), 41–55.

Koenker, R. (2005). *Quantile Regression*. Number 38 in Econometric Society Monographs. Cambridge: Cambridge University Press.

Koenker, R. (2011a). Additive models for quantile regression: model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3), 239–262.

Koenker, R. (2011b). *quantreg: Quantile Regression*. R package version 4.67.

Koenker, R. & Bassett, G. W. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50.

Koenker, R. & Bassett, G. W. (1982a). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1), 43–61.

Koenker, R. & Bassett, G. W. (1982b). Tests of linear hypotheses and l_1 estimation. *Econometrica*, 50(6), 1577–1583.

Koenker, R. & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296–1310.

Koenker, R. & Mizera, I. (2004). Penalized triograms: total variation regularization for bivariate smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1), 145–163.

Koenker, R. & Ng, P. T. (2005). Inequality constrained quantile regression. *Sankhya: The Indian Journal of Statistics*, 67(2), 418–440.

Koenker, R., Ng, P. T., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 84(4), 673–680.

Landajo, M., De Andrés, J., & Lorca, P. (2008). Measuring firm performance by using linear and non-parametric quantile regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(2), 227–250.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational & Graphical Statistics*, 13(1), 183–212.

Laplace, P.-S. (1789). *Sur quelques points du système du monde*. Mémoires de l'Académie des Sciences de Paris.

Lee, T. C. M. (2000). Regression spline smoothing using the minimum description length principle. *Statistics & Probability Letters*, 48(1), 71–82.

Legendre, A.-M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*.

Li, Q. & Racine, J. S. (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26(4), 423–434.

Lu, M., Zhang, Y., & Huang, J. (2009). Semiparametric estimation methods for panel count data using monotone B-splines. *Journal of the American Statistical Association*, 104(487), 1060–1070.

Mackenzie, M. L., Donovan, C., & McArdle, B. (2005). Regression spline mixed models: A forestry example. *Journal of Agricultural, Biological and Environmental Statistics*, 10(4), 394–410.

Neocleous, T. & Portnoy, S. (2008). On monotonicity of regression quantile functions. *Statistics & Probability Letters*, 78(10), 1226–1229.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases [`http://www.ics.uci.edu/~mlearn/MLRepository.html`]. Irvine, CA: University of california, department of information and computer science.

Ng, P. T. & Maechler, M. (2007). A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4), 315–328.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4), 502–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing*, 9(2), 363–379.

Paula, G. A. (1993). Assessing local influence in restricted regression-models. *Computational Statistics & Data Analysis*, 16(1), 63–79.

Paula, G. A. (1999). Leverage in inequality-constrained regression models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(4), 529–538.

Portnoy, S. & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, 12(4), 279–296.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Racine, J. & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1), 99–130.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4), 735–757.

Ruppert, D. & Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, 42(2), 205–223.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge: Cambridge University Press.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193–1256.

Schnabel, S. K. & Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53(12), 4168–4177.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.

Shim, J., Hwang, C., & Seok, K. H. (2009). Non-crossing quantile regression via doubly penalized kernel machine. *Computational Statistics*, 24(1), 83–94.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society Series B - Methodological*, 47(1), 1–52.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, fourth edition.

Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics*, 15(4), 443–462.

Wegman, E. J. & Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association*, 78(382), 351–365.

Wood, S. N. (1994). Monotonic smoothing splines fitted by cross-validation. *SIAM Journal on Scientific Computing*, 15(5), 1126–1133.

Wood, S. N. (2011). *mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL*. R package version 1.7-6.

Wooldridge, J. M. (2009). *Introductory Econometrics*. South-Western, 4. edition.

Yu, K., Lu, Z., & Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society Series D - The Statistican*, 52(Part 3), 331–350.

Yuan, M. (2006). GACV for quantile smoothing splines. *Computational Statistics & Data Analysis*, 50(3), 813–829.

Zhou, K. Q. & Portnoy, S. L. (1996). Direct use of regression quantiles to construct confidence sets in linear models. *Annals of Statistics*, 24(1), 287–306.