Semiparametric Mode Regression

Margret-Ruth Oelker[*][†], Fabian Sobotka[‡], Nadja Klein[‡] & Thomas Kneib[‡]

April, 2015

Abstract

Regression models for the conditional mode of a response distribution can provide useful additional information compared to mean and median analyses. Unfortunately, the conditional mode is intrinsically difficult to determine for continuous data. As a consequence, most of the previous approaches for mode regression resort to kernel density estimation from which the mode is then determined in a second step. We propose direct inference in semiparametric mode regression based on an iteratively re-weighed least squares (IRLS) optimization of a continuous, quadratic approximation of the conditional mode loss function. Adaptive tuning parameters within the algorithm provide stable estimates and avoid additional manual tuning. For linear predictors, a close link to kernel-based approaches allows to derive consistency and asymptotic normality of the estimator. The quadratic approximation of the loss function can also easily be combined with quadratic penalties in semiparametric extensions of mode regression comprising for example penalized spline or spatial components. We evaluate our conditional mode specification in several simulation studies with different error structures and illustrate its relevance along two real data sets.

Keywords: Mode regression; kernel regression; semiparametric regression; iteratively re-weighted least squares; penalized splines

[*] Corresponding author: margret.oelker@stat.uni-muenchen.de

[†] Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany

[‡] Faculty of Economic Sciences, Georg-August-Universität Göttingen, Germany

1 Introduction

Recent years have seen a tremendous increase in interest related to regression beyond the mean of the conditional distribution of a response given covariates. The most prominent examples are quantile regression and the special case of median regression. They are particularly attractive alternatives to mean regression due to two reasons: Due to their inherent robustness with respect to outliers, and due to the general information they provide concerning distributional features such as heteroscedasticity or skewness. Surprisingly, regression models for the conditional mode of the response distribution given covariates have received far less attention. This may partially be explained by the inherent difficulty to determine an estimate for the mode based on samples from a continuous distribution, where in theory each sampled value should appear only once almost surely, and therefore, there will be multiple "empirical modes". Still, estimating conditional modes is of high interest as

- the mode is by far the visually most prominent feature of a density as compared to the mean and the median,
- the mode is extremely robust with respect to outliers,
- the mode provides a location measure that is easily communicated to practitioners such that mode regression will be of high interest in applied regression situations,
- there may be situations where the dependence of the mode on covariates may be quite different from the dependence of the median and/or the mean,
- mode regression allows to deal with truncated dependent variables. It can still be estimated and interpreted as long as the modal part of the distribution is not truncated. This can, for example, be relevant in applications on income where quite often the upper part of the response distribution is truncated due to non-participation of the high income part of a society.

Consider the regression specification

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \varepsilon, \tag{1}$$

where y is the response variable of interest, $\boldsymbol{x} \in \mathbb{R}^q$ is a vector of covariates supplemented with regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^q$ and ε is the error term. Unlike in mean regression, we do not assume $\mathbb{E}(\varepsilon) = 0$ which leads to regression effects on the mean of the response variable, but

$$\arg\max_{\xi} f_{\varepsilon|\boldsymbol{x}}(\xi|\boldsymbol{x}) = 0. \tag{2}$$

That is, the conditional density of the error terms $f_\varepsilon(\cdot|\boldsymbol{x})$ is assumed to have a global mode at zero. In turn, this implies that the predictor $\boldsymbol{x}^T\boldsymbol{\beta}$ is the conditional mode of the response distribution $f_y(\cdot|\boldsymbol{x})$. The mode regression coefficient is obtained as

$$\boldsymbol{\beta} = \arg\max_{\boldsymbol{b}} f_\varepsilon(y - \boldsymbol{x}^T\boldsymbol{b}|\boldsymbol{x}). \tag{3}$$

An equivalent approach is based on the step loss function $\mathcal{L}_\epsilon(\xi) = 1 - \mathbb{1}(-\epsilon \leq \xi \leq \epsilon)$, where ϵ is a positive constant that defines a local environment around zero. With this loss function, we obtain

$$\boldsymbol{\beta}_\epsilon = \arg\min_{\boldsymbol{b}} \mathbb{E}\left[\mathcal{L}_\epsilon(y - \boldsymbol{x}^T\boldsymbol{b})|\boldsymbol{x}\right]. \tag{4}$$

In the limiting case $\epsilon \to 0$, $\mathcal{L}_\epsilon(\xi)$ approaches

$$L(\xi) = \mathbb{1}(\xi \neq 0), \tag{5}$$

and $\boldsymbol{\beta}_\epsilon$ approaches $\boldsymbol{\beta}$ from equation (3) (Manski, 1991). Held (2008, page 158) proves the equivalence of the two approaches in more detail. However, based on n independent observations $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$, from model (1) subject to the condition (2), an estimate for the mode regression coefficient can not be determined by an empirical analogue to (3) unless specific assumptions are made for the error density $f_\varepsilon(\cdot|\boldsymbol{x})$. In contrast, (4) is empirically minimized by

$$\hat{\boldsymbol{\beta}}_\epsilon = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\mathcal{L}_\epsilon(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})\right],$$

which does not require any other further assumptions than independence of the observations. However, for the limiting case $\epsilon \to 0$, this criterion is not useful for modal regression based on data with continuous error distributions since in general, there will be no unique solution – even if the density of the errors ε_i has a global mode. As a consequence, earlier attempts to mode regression usually either rely on nonparametric kernel methods from which the mode is then derived in a second step, or on different types of approximations (5).

Collomb et al. (1987) follow the first of these two routes and show the uniform convergence of the mode determined from a kernel density estimate to the conditional mode function for a certain class of processes. Lee (1989) approaches the estimation of conditional modes by an empirical approximation to the theoretical loss function defining the mode based on a rectangular kernel. Lee (1989) also shows identification and strong consistency of the resulting estimate, but this requires quite strong assumptions on the error distribution, which has either to be symmetric around the mode (in which case median or mean regression would be obvious alternatives to determine the mode) or – if assumed to be asymmetric – all error distributions have to be identical leading to an i.i.d. model. Lee (1993) extends his approach from 1989 by replacing the rectangular kernel with a quadratic kernel. This allows to construct a more efficient estimate, but it also requires stronger assumptions on the error term such as local symmetry around the mode. Yu and Aristodemou (2012) introduce Bayesian mode regression relying on a working likelihood corresponding to either a uniform or a triangular density.

Einbeck and Tutz (2006) again rely on a kernel regression estimate to implicitly derive the mode in a regression model, but they extend the linear regression specification to a semiparametric predictor. This allows for the nonlinear dependence of the conditional mode on the covariate of interest, but the approach is limited to one single predictive variable. A multivariate extension based on a product kernel for the multivariate covariate vector is outlined in Taylor and Einbeck (2011). However, the resulting estimate is hard to interpret beyond two-dimensional covariates since no additivity assumption can be placed on the predictor. Gannoun et al. (2010) follow a different approach by noting that for many distributions there exists a simple parametric relationship between mode, median and mean. As a consequence, once estimates for the mean and the median are available, the conditional mode can be derived based on this parametric relationship. Their approach is motivated by a forecasting problem in financial time series such that no interpretability for the regression effects on the mode is required, which would be difficult to achieve when combining mean and median estimates.

Kemp and Santos Silva (2012) return to the idea of Lee (1989, 1993). They use a modified kernel to approximate the limiting case (5) and to derive a consistent, asymptotically normal estimator for linear mode regression models. In this paper, we build upon Kemp and Santos Silva (2012) and

- provide a differentiable approximation of the limiting case (5) that is based on nested intervals such that an iteratively re-weighted least squares (IRLS) algorithm can be used to estimate the mode regression coefficients (Section 2),
- show the consistency and the asymptotic normality of the obtained estimator,
- investigate the practical performance of the approach in a simulation study,

- extend the purely linear mode regression model to additive models by combining nonparametric effects of several covariates in one penalized IRLS framework (Section 3),
- provide an extended analysis of the evolution of the BMI in England that has been already studied in Kemp and Santos Silva (2010) and where the polynomial specification of the effect of the age is replaced by a nonparametric specification (Section 4).
- perform a geoadditive analysis of the rents in the city of Munich combining penalized spline smoothing with spatial effects (Section 5).

The main advantage of this Nested Interval Least Squares (NILS) framework is that it allows to easily include extended regression functionality from (generalized) additive models which also rely on IRLS estimation. In fact, we can further exploit this connection by determining the smoothing parameters within the IRLS framework such that the proposed semiparametric mode regression is fully data-driven.

2 The Nested Interval Least Squares Approach

As seen in the introduction of this paper, there are two equivalent approaches to mode regression: maximizing the conditional density $f_\varepsilon(\cdot|\boldsymbol{x})$ and minimizing the expectation of the step loss function $\mathcal{L}_\epsilon(\xi)$ for the limiting case $\epsilon \to 0$. The reasoning behind the latter can be illustrated based on a set of simulated standard normal data: Iteratively reducing the environment $[-\epsilon, \epsilon]$ allows to determine the mode via nested intervals that contain the largest fraction of observations. Stacking these intervals upon each other allows to graphically indicate how reducing the width of the intervals captures the mode of the distribution. For comparison, in Figure 1, a kernel density estimate is added.

2.1 Construction of the Estimator

Our approach to mode regression follows a similar reasoning: The limiting case $L(\xi)$ is approximated such that it is zero not only for $\xi = 0$ but in a surrounding of $\xi = 0$. The approximation – denoted by $\mathcal{L}(\xi)$ – will replicate the nested interval approach, that is, $\mathcal{L}(\xi)$ will have a very broad minimum in the early iterations and it will be very close to $L(\xi)$ for the final iteration of the proposed algorithm. However, $L(\xi)$ is approximated by a continuously differentiable function. This has two important advantages: (i) The approximation $\mathcal{L}(\xi)$ can be linked to

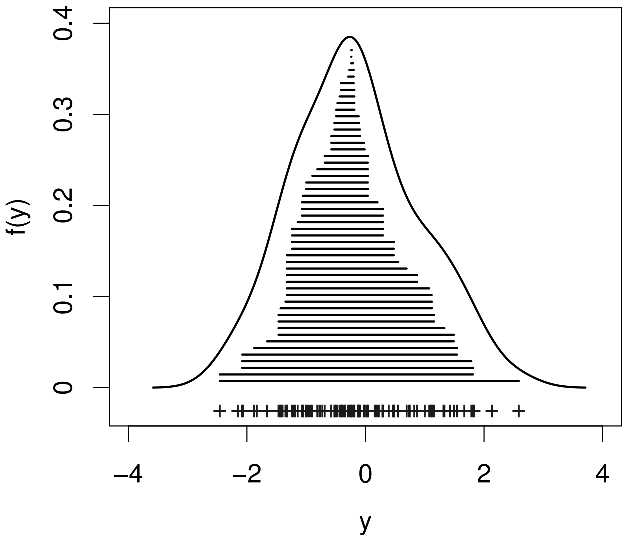

Figure 1: Determining the mode based on nested intervals: Based on 100 realizations from a standard normal distribution ("+"), nested intervals are constructed such that the interval covers the largest possible fraction of data points given a fixed width. Stacking these intervals upon each other allows to graphically indicate how reducing the width of the intervals captures the mode of the distribution. For comparison, a kernel density estimate is added.

iteratively re-weighted least squares estimation, and (ii) the smooth approximation allows to determine asymptotic properties such as consistency and asymptotic normality.

In detail, we employ the function

$$\mathcal{L}(\xi) = 1 - \exp(c^{\frac{1}{2g}} - ((k\xi)^{2g} + c)^{\frac{1}{2g}}), \tag{6}$$

depending on the set of tuning parameters $\mathcal{T} = \{g, k, c\}$ with $\lim_{\mathcal{T}\to T} \mathcal{L}(\xi) = L(\xi)$ for some set of limiting values T. The approximation $\mathcal{L}(\xi)$ is constructed as the scaled composition of the two functions $f(\xi)$ and $h(\xi)$. The former is given by $f(\xi) = 1 - \exp(-\xi)$. Let k be a positive number, then $f(k \cdot \xi)$ actually approximates the indicator $L(\xi)$ with the approximation being closer to $L(\xi)$ the larger k is. The latter function is defined as $h(\xi) = (\xi^{2g} + c)^{\frac{1}{2g}}$, where g is as a positive integer and c is a small, positive constant. As illustrated in Figure 2, $h(\xi)$ accounts for the broad minimum needed to imitate the nested interval approach. For the limiting value $g = 1$, $h(\xi)$ simply approximates the absolute value function. Due to the constant c, it is a continuously differentiable approximation of the absolute value. Scaling the composition $f(h(k \cdot \xi))$ gives function (6). As $\mathcal{L}(\xi)$ is continuously differentiable, an iteratively re-weighted

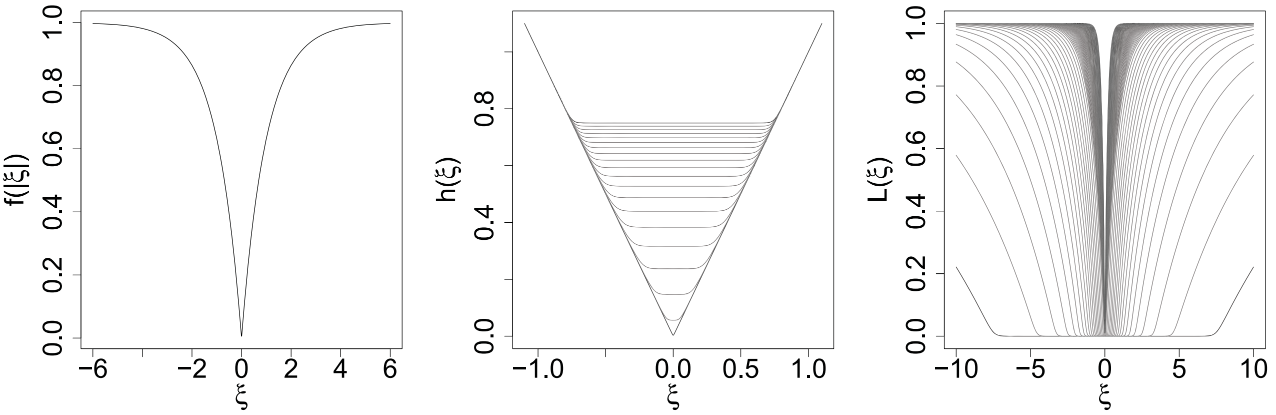

Figure 2: Illustration of the employed loss function. The left panel shows function $f(|\xi|)$. The panel in the middle depicts function $h(\xi)$, where the tuning parameter $c = 10^{-5}$ is fixed. Parameters k and g vary as follows: $g = 20, \ldots, 1$, $k = 0.1, \ldots, 6$ in 99 steps. The right panel shows the scaled composition $\mathcal{L}(\xi)$ for the same tuning parameters.

least squares algorithm is derived. The approximated objective

$$\mathcal{M}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathcal{L}(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})$$

is minimized by iterating

$$\hat{\boldsymbol{\beta}}_{(l+1)} \quad = \quad (1-\nu)\hat{\boldsymbol{\beta}}_{(l)} + \nu \boldsymbol{A}_{(l)}^{-1} \boldsymbol{a}_{(l)} \tag{7}$$

until convergence. Thereby

$$\begin{aligned} \boldsymbol{a}_{(l)} &= \boldsymbol{X}^T \mathrm{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)}} \right) \boldsymbol{y}, \\ \boldsymbol{A}_{(l)} &= \boldsymbol{X}^T \mathrm{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{(l)}} \right) \boldsymbol{X}, \end{aligned} \tag{8}$$

and $\mathcal{D}(\xi) = \frac{\partial \mathcal{L}(\xi)}{\partial \xi}$ denotes the derivative of the employed loss. The design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T \in \mathbb{R}^{n \times q}$ comprises the covariate vectors $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i,q-1})^T$, $i = 1, \ldots, n$, and the step length $\nu > 0$ controls the speed of convergence. The algorithm is terminated when

$$\frac{\sum_{j=0}^{q} |\boldsymbol{\beta}_{(l+1)} - \boldsymbol{\beta}_{(l)}|}{\sum_{j=0}^{q} |\boldsymbol{\beta}_{(l)}|} \leq \tau,$$

where τ is a small, positive constant. For a detailed derivation of (7), see Appendix A.

To imitate the idea of nested intervals, the tuning parameters have to be chosen such that g is relatively large in the early iterations of the IRLS algorithm while it should equal one for the final iteration. In contrast, k is relatively small in the beginning of the algorithm and as large as possible for the final iteration. The constant c is as small as possible. To allow for a smooth transition $\mathcal{T} \to T$ and thus reliable results, the algorithm will have a small step length ν (for example, $\nu = 0.25$) and thus relatively many iterations until convergence. In Section 2.3, the (data-driven) choice of the tuning parameters is discussed in more detail.

2.2 Asymptotic Properties

For the final iteration of the IRLS algorithm, it holds that $g = 1$ and that k is relatively large. Hence, to show asymptotic properties, we assume $g = 1$ and consider the properties of

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \mathcal{M}(\boldsymbol{\beta}) \tag{9}$$

for $k_n \to \infty$ and $n \to \infty$ at appropriate rates. The index n emphasizes the dependence on the sample size n. With $g = 1$, minimizing $\mathcal{M}(\boldsymbol{\beta})$ is equivalent to the minimization of $1 - K(u)$ where

$$K : \mathbb{R} \to \mathbb{R}, \quad u \mapsto K(u) = \frac{1}{2} \exp\left\{-\sqrt{u^2 + c}\right\}, \; 0 < c \leq 1, \tag{10}$$

and where $u = k_n \cdot (y - \boldsymbol{x}^T \boldsymbol{\beta})$. The kernel $K(u)$ in turn is an approximation of $\frac{1}{2}\exp(-|u|)$ which is the density of a Laplace distributed random variable U with mean $\mathbb{E}(U) = 0$ and variance $\mathbb{V}(U) = 2$. That is, for the final iteration, the proposed approximation can be interpreted as one minus a rounded (and thus, differentiable) Laplace kernel. As discussed in Section 1, approaching mode regression with kernel methods is well established and investigated. Kemp and Santos Silva (2012) derive asymptotic properties for mode regression for a general kernel $K(u)$, where $u = (y - \boldsymbol{x}^T\boldsymbol{\beta})/\delta_n$ with positive bandwidth δ_n depending on the sample size n and with the objective function $1 - \mathcal{M}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} (\delta_n^{-1}(y - \boldsymbol{x}^T\boldsymbol{\beta}))$. One can easily see that function (10) structurally fits in this framework as the tuning parameter k_n relates inversely to the bandwidth δ_n. Moreover, we show in Lemma 1 and Lemma 2 that a scaled version of function (10) meets all requirements needed to prove asymptotically consistent and normal estimates.

Consistency For proving consistency, we make the following extended assumptions following Kemp and Santos Silva (2012):

A1 $\{(\varepsilon_i, \boldsymbol{x}_i)\}_{i=1}^{\infty}$ is an independent and identically distributed (i.i.d.) sequence, where ε_i takes values in $\mathbb{R}$ and $\boldsymbol{x}_i$ takes values in $\mathbb{R}^q$ for some finite q.

A2 The parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^q$ and contains the true value $\boldsymbol{\beta}^*$.

A3 The distribution of $\boldsymbol{x}$ is such that:

(i) $\mathbb{E}(\|\boldsymbol{x}_i\|) < \infty$, where $\|\boldsymbol{a}\|$ denotes the Euclidean norm of $\boldsymbol{a}$ for any scalar or finite-dimensional vector $\boldsymbol{a}$,

(ii) $\mathbb{P}(\boldsymbol{x}_i^T \boldsymbol{c} = 0) < 1$ for all fixed $\boldsymbol{c} \neq \boldsymbol{0}$.

A4 There exists a version of the conditional density of ε given $\boldsymbol{x}$, denoted $f_{\varepsilon|\boldsymbol{x}}(\cdot|\cdot)$: $\mathbb{R} \times \mathbb{R}^q \to \mathbb{R}$, such that:

(i) $\sup_{\varepsilon \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^q} f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) \leq \infty$,

(ii) $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})$ is continuous for all ε and $\boldsymbol{x}$. In addition, there exists a set $A \subseteq \mathbb{R}^q$ such that $\mathbb{P}(\boldsymbol{x}_i \in A) = 1$ and $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) \leq f_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x})$ for all $\varepsilon \neq 0$ and $\boldsymbol{x} \in A$.

A5 $\{k_n\}_{n=1}^{\infty}$ is a strictly positive sequence such that:

(i) $k_n \to \infty$,

(ii) $n \left(k_n \ln(n)\right)^{-1} \to \infty$.

Assumptions A1 and A3 are standard assumptions. Together with A4 and Lemma 1, they are needed to prove that the objective function $\mathcal{M}$ has a global minimum at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ which is unique (compare Lemma 1, Kemp and Santos Silva, 2012). Note that A1 does imply an i.i.d. assumption for $\varepsilon_i|\boldsymbol{x}_i$, but not for ε_i. Furthermore, A1 could be relaxed even further, but then stronger assumptions on the distribution of $\boldsymbol{x}_i$ would be required. Assumptions A2 and A5 are required to prove that the objective function satisfies a uniform law of large numbers (Lemma 2, Kemp and Santos Silva, 2012). Assumption A4 (ii) imposes that the conditional density has a global mode at zero. Assumption A5 ensures that k_n is increasing with a moderate rate. This is the crucial factor in the algorithm since no symmetry assumptions are made on the conditional density. We come back to the choice of k_n in Section 2.3. The kernel needs to be a bounded density that can be normalized having a bounded derivative:

Lemma 1. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (10) *is differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} K(u) \mathrm{d}u = 1$,

(ii) $\sup_{u \in \mathbb{R}} |K(u)| = c_0 < \infty$,

(iii) $\sup_{u\in\mathbb{R}} |K'(u)| = c_1 < \infty$, *where* $K'(u) = \mathrm{d}K(u)/\mathrm{d}u$.

The proof of Lemma 1 can be found in Appendix B.2.

With these assumptions, we obtain the consistency of the mode regression estimate:

Theorem 1. *If assumptions A1–A5 hold, the IRLS-based mode regression estimate is consistent, that is*

$$\hat{\boldsymbol{\beta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\beta}^*.$$

Theorem 1 is a direct consequence of Kemp and Santos Silva (Theorem 1, 2012) and Lemma 1.

Asymptotic Normality To obtain asymptotic normality, we need the following additional assumptions:

B1 $\mathbb{E}(|\boldsymbol{x}_i|^{5+\xi}) < \infty$ for some $\xi > 0$.

B2 $\boldsymbol{\beta}^*$ belongs to the interior of $\boldsymbol{\mathcal{B}}$.

B3 $f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})$ is three times differentiable with respect to ε for all $\boldsymbol{x}$ such that:

(i) $f^{(j)}_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x}) = \partial^j f_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})/\partial\varepsilon^j$ is uniformly bounded for $j = 1, 2, 3$,

(ii) $\mathbb{E}\left[f^{(2)}_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x})\boldsymbol{x}\boldsymbol{x}^T\right]$ is negative definite.

B4 The sequence $\{k_n\}_{n=1}^{\infty}$ is such that:

(i) $n/k_n^7 = o(1)$,

(ii) $n\left(k_n^5 \ln(n)\right)^{-1} \to \infty$.

As expected, each of these assumptions is a stronger version of the assumptions A1–A5. In particular, further moments of the distribution of $\boldsymbol{x}_i$ are required to be finite (B1) and the true parameter has to be in the interior of the parameter space $\boldsymbol{\mathcal{B}}$ (B2). The latter assumption is standard in maximum-likelihood estimation. Assumption B3 guarantees the existence of a Taylor expansion of the first derivative $f^{(1)}_{\varepsilon|\boldsymbol{x}}(u/k_n|\boldsymbol{x})$ around $u = 0$. Note that no smoothness in $\boldsymbol{x}_i$ is required such that the theory also holds for categorical covariates. Finally, assumptions B4(i) and B4(ii) imply more constrained rates on k_n compared to assumption A5. We will see in Theorem 2 that B4(ii) implies that the speed of convergence of the estimate is at most $n^{2/7}$. For the kernel function, the following stronger assumptions about its smoothness are needed:

Lemma 2. *The kernel function* $K : \mathbb{R} \to \mathbb{R}$ *defined in* (10) *is three times differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$,

(ii) $\lim_{u\to\pm\infty} K(u) = 0$,

(iii) $\int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u = M_0 < \infty$,

(iv) $\int_{-\infty}^{\infty} |K'(u)|^2\mathrm{d}u = M_1 < \infty$,

(v) $\sup_{u\in\mathbb{R}} |K''(u)| = M_2 < \infty$,

(vi) $\sup_{u\in\mathbb{R}} |K'''(u)| = M_3 < \infty$,

(vii) $\int_{-\infty}^{\infty} |K''(u)|^2\mathrm{d}u = M_4 < \infty$.

With Lemma 2 and Theorems 2 and 3 from Kemp and Santos Silva (2012) asymptotic normality follows:

Theorem 2. *Under Assumptions A1–A5 and B1–B4, the IRLS-based mode regression estimate is asymptotically normal, that is*

$$\left(\frac{n}{k_n^3}\right)^{1/2}\left[\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}^*\right] \xrightarrow{d} N(0, \boldsymbol{\Omega}^*),$$

where the asymptotic covariance matrix is given by

$$\begin{aligned}
\boldsymbol{\Omega}^* &= \boldsymbol{C}^{*-1}\boldsymbol{B}^*\boldsymbol{C}^{*-1},\\
\boldsymbol{B}^* &= \lim_{n\to\infty} \mathbb{V}\left(\left(\frac{n}{k_n^3}\right)^{1/2}\left(-\frac{\partial\mathcal{M}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}\bigg|_{\boldsymbol{\beta}^*}\right)\right) = M_1\mathbb{E}\left(f_{\varepsilon|\boldsymbol{x}}(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right),\\
M_1 &= \int_{-\infty}^{\infty}|K'(u)|^2\mathrm{d}u < \frac{1}{4},\\
K'(u) &= \mathrm{d}K(u)/\mathrm{d}u,\\
\boldsymbol{C}^* &= \lim_{n\to\infty}\mathbb{E}\left(-\frac{\partial^2\mathcal{M}(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\bigg|_{\boldsymbol{\beta}^*}\right) = \mathbb{E}\left(f^{(2)}_{\varepsilon|\boldsymbol{x}}(\varepsilon|\boldsymbol{x})(0|\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T\right).
\end{aligned}$$

A consistent estimate for the asymptotic covariance matrix is obtained by

$$\hat{\boldsymbol{\Omega}}_n = \hat{\boldsymbol{C}}_n^{-1}\hat{\boldsymbol{B}}_n\hat{\boldsymbol{C}}_n^{-1} \xrightarrow{\mathbb{P}} \boldsymbol{\Omega}^*,$$

where

$$\begin{aligned}
\hat{\boldsymbol{B}}_n &= n^{-1}\sum_{i=1}^{n} k_n\left[K'\left(k_n\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right)\right)\right]^2(\boldsymbol{x}_i\boldsymbol{x}_i^T);\\
\hat{\boldsymbol{C}}_n &= n^{-1}\sum_{i=1}^{n} k_n^3K''\left(k_n\left(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_n\right)\right)(\boldsymbol{x}_i\boldsymbol{x}_i^T).
\end{aligned}$$

Remark The close connection to the approach of Kemp and Santos Silva (2012) provides not only the asymptotic theory for the NILS approach. Kemp and Santos Silva (2012) argue that their approach has two limiting cases: As they employ the Gaussian kernel, mean regression for $\delta_n \longrightarrow \infty$, and mode regression for $\delta_n \longrightarrow 0$. Considering $\mathcal{L}(\xi)$ as one minus a rounded Laplace kernel yields a similar interpretation for the NILS approach: The loss function $\mathcal{L}_\epsilon$ corresponds to the loss function of median regression for $k_n = g = 1$. For $g = 1$ and $k_n \longrightarrow \infty$, $\mathcal{L}(\xi) \longrightarrow L(\xi)$. Hence, depending on the choice of k_n, the NILS approach is closer to mode or to median regression. As $\bar{x} > \tilde{x}_{median} > \tilde{x}_{mode}$ for positively skewed distributions and $\bar{x} < \tilde{x}_{median} < \tilde{x}_{mode}$ for negatively skewed distributions, the NILS approach seems to be a natural choice to approximate mode regression.

2.3 Adaptive Tuning

As indicated in Section 2.1, the NILS approach requires tuning. The constant $c > 0$ guarantees that the loss function $\mathcal{L}(\xi)$ is differentiable. As long as it is sufficiently small, it has a minor impact on the performance and in our experience, $c = 10^{-5}$ works well. The integer g governs how broad the minimum of $\mathcal{L}(\xi)$ is. It should be large enough to guarantee $\mathcal{D}(\xi) \neq 0$ for the initial iteration and decreases towards 1 within the natural numbers while iterating. As the value of k affects the width of the minimum of $\mathcal{L}(\xi)$ for $g > 1$, it is possible to choose a fixed sequence for g (we propose to choose the fixed sequence from 10 to 1 for g) and to address all issues of tuning by a properly chosen sequence k_n of k. Since k determines how close $\mathcal{L}(\xi)$ and $L(\xi)$ are, it has to be chosen carefully and its impact on the asymptotic variance of the estimates has to be controlled. Thus, we propose to choose the sequence of values for k_n driven by the data and by the asymptotic theory. The initial value for k_n is determined as $k_{initial} = (n/const)^{1/7}$ where $const$ is chosen such that n fulfills both assumptions B4(i) and B4(iii): $const = n^{5/12} \cdot \log(n)^{7/12}/(n^{7/12})$. Then, $k_{initial}$ is increased up to $k_{final} = k_{initial} n^{1/7}/sd$ in 10 steps while iterating and where sd is the standard deviation of the residuals of a fitted median regression. After that, the final value for k_n is kept until convergence. In order to obtain smooth transitions and to reach a value of k_n that is sufficiently large, the step length is set to $\nu = 0.25$.

2.4 Numerical Experiments

To evaluate the performance of the NILS approach in finite samples, we consider the estimation accuracy and the applicability of the asymptotic results in a linear model. Concretely, we generate $n_{rep} = 100$ replications of the model

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \\ &= 1 + 0.2x_1 - 2x_2 + 3x_3 + \varepsilon, \end{aligned} \tag{11}$$

where x_1, x_2, x_3 are drawn from the continuous uniform distribution on $[0, 2]$. Thereby, different *model features* are systematically varied:

- The distribution of the errors ε is either Gaussian $\varepsilon \sim N(0, 1)$, log-normal $\varepsilon \sim LN(0, 1)$ or gamma $\varepsilon \sim Ga(s = 2, r = 2)$, where s and r denote the shape and the inverse scale parameter. That is, we consider a symmetric scenario where mean, median and mode coincide and two skew scenarios with differently shaped error distributions. As the mode of the skew distributions is unequal zero, they are shifted accordingly.
- Different sample sizes are considered: $n \in \{100, 500, 1000, 10000\}$.

For each replication of model (11), four different *methods* are compared:

- the NILS approach with adaptive tuning as proposed in Section 2.3,
- the approach of Kemp and Santos Silva (2012),
- mean regression and
- median regression.

According to Kemp and Santos Silva (2012), the bandwidth δ_n of their approach is chosen based on the median of the absolute deviation from the median least squares residual of a preceding mean regression: $\text{MAD} = \text{med}_i \left\{ \left| (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}) - \text{med}_j (y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}) \right| \right\}$ and $\delta_n = 1.2 \cdot \text{MAD} \cdot n^{1/7}$. The mean regression estimates are employed as starting values.

The results of the median regression are obtained by an IRLS algorithm that approximates the absolute loss function $|\xi|$ by $\sqrt{\xi + c}$, where c denotes a small positive constant, for example, $c = 10^{-5}$. This is advantageous as it allows for exactly the same computational structure for both, median and mode regression.

As the speed of convergence of the asymptotic theory in Section 2.2 is rather slow, we expect that the coverage rates of the confidence intervals (CI) based on the asymptotic covariance

Error Distribution:		$N(0,1)$				$LN(0,1)$				$Ga(2,2)$			
	n:	100	500	1000	10000	100	500	1000	10000	100	500	1000	10000
Mode	β_0	0.04	0.10	0.09	0.29	0.26	0.28	0.36	0.45	0.05	0.06	0.09	0.11
regression	β_1	0.04	0.06	0.13	0.32	0.32	0.48	0.53	0.79	0.08	0.15	0.12	0.26
NILS	β_2	0.09	0.10	0.07	0.29	0.32	0.50	0.57	0.78	0.09	0.18	0.12	0.21
	β_3	0.08	0.08	0.11	0.28	0.35	0.50	0.53	0.65	0.04	0.12	0.16	0.20
Mode	β_0	0.76	0.87	0.76	0.90	0.63	0.54	0.49	0.37	0.70	0.71	0.67	0.40
regression	β_1	0.83	0.79	0.80	0.93	0.80	0.85	0.76	0.83	0.79	0.75	0.84	0.85
NILS BS	β_2	0.73	0.79	0.84	0.86	0.81	0.83	0.83	0.79	0.73	0.72	0.80	0.83
	β_3	0.84	0.81	0.83	0.90	0.78	0.82	0.80	0.75	0.84	0.73	0.79	0.89
Mode	β_0	0.50	0.77	0.72	0.81	0.58	0.35	0.13	0.01	0.33	0.46	0.53	0.60
regression	β_1	0.59	0.77	0.69	0.86	0.83	0.87	0.80	0.88	0.39	0.59	0.46	0.64
Kemp	β_2	0.55	0.70	0.76	0.81	0.77	0.88	0.88	0.81	0.39	0.51	0.59	0.60
	β_3	0.57	0.74	0.76	0.86	0.71	0.84	0.81	0.83	0.39	0.44	0.55	0.57
Mode	β_0	0.75	0.83	0.74	0.85	0.67	0.46	0.21	0.03	0.64	0.77	0.78	0.78
regression	β_1	0.75	0.78	0.69	0.80	0.89	0.91	0.83	0.91	0.83	0.81	0.80	0.91
Kemp BS	β_2	0.78	0.72	0.74	0.77	0.88	0.91	0.90	0.83	0.74	0.80	0.79	0.85
	β_3	0.74	0.70	0.77	0.90	0.89	0.89	0.89	0.85	0.83	0.82	0.83	0.80

Table 1: Coverage rates of the confidence intervals estimated for different sample sizes and different error distributions; BS denotes that the results rely on $B = 1000$ bootstrap samples.

matrix $\hat{\boldsymbol{\Omega}}_n$ are reliable only for a rather large number of observations n. Hence, beside the CIs derived from asymptotic normality, we evaluate $(1-\alpha)$ CIs based on bootstrap samples of the residuals $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_n$. For each sample $(\boldsymbol{y}_b^*, \boldsymbol{X})_{b=1,\dots,B}$ with $\boldsymbol{y}_b^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_n + \hat{\varepsilon}_b^*$, the according model is estimated and we obtain the bootstrap estimates $\hat{\boldsymbol{\beta}}^*_{b=1,\dots,B}$. The pointwise $(1-\alpha)$ CI for the estimated coefficient $\hat{\boldsymbol{\beta}}_n$ is then defined by the $\alpha/2$ and the $1-\alpha/2$ quantile of the empirical distribution of the bootstrap estimates $\hat{\boldsymbol{\beta}}^*_{b=1,\dots,B}$. This approach assumes that the functional form of the regression model is correctly specified and that the errors are identically distributed (Fox, 2008, page 598). While this may seem rather restrictive, nonparametric bootstrap samples are not a good choice for mode regression as the samples $(\boldsymbol{y}, \boldsymbol{X})^*_{b=1,\dots,B}$ contain duplicated observations. Duplicated or even multiplied observations imply a mode of $\varepsilon|\boldsymbol{X}$ and can therefore render the estimation procedure unstable. Following Efron and Tibshirani (1993), we choose $B = 1000$ to determine the bootstrap CIs.

Results To judge the results, the estimation accuracy and the coverage rates of 95% CIs are considered. The left panel of Figure 3 shows boxplots of the resulting coefficients for $n = 100$ observations, $\varepsilon \sim N(0,1)$ (top) and $\varepsilon \sim LN(0,1)$ (bottom). The results can be summarized as follows:

- In the most simple scenario with standard normal errors, the estimation accuracy of the

NILS approach is not as precise as the results of mean and median regression which was to be expected since the error distribution is symmetric. Due to some outliers, the variations of the approach of Kemp and Santos Silva (2012) are slightly larger.

- In the scenario with log-normal errors, mean, median and mode of the error distribution differ by a location shift. The lower left plot of Figure 3 illustrates that this shift is captured by the estimates of the intercept β_0. The results of mean and median regression are clearly scattered around a value different from the true value which is indicated by a horizontal line. Again, this was to be expected as the structure of the errors is additive. Both, the NILS approach and the proposal of Kemp and Santos Silva (2012) are biased slightly regarding the intercept while the remaining coefficients are estimated equally well by all methods.
- The middle panel of Figure 3 shows the widths of the confidence intervals for each coefficient obtained with the asymptotic theory of Section 2.2. One sees that the interval widths for the NILS and the Kemp approach differ substantially as they depend on the choice of the tuning parameters k_n and δ_n. For the NILS approach, k_n is increased steadily while iterating whereas Kemp and Santos Silva (2012) choose a fixed bandwidth. Table 1 gives the corresponding coverage rates for normally, log-normally and gamma distributed errors, $n \in \{100, 500, 1000, 10000\}$. Both approaches perform differently well for different error distributions. Beside the different tunings employed, another reason for the partly insufficient results is the slow rate of convergence of at most $n^{2/7}$. In practice, we advise to apply bootstrap methods to assess the estimate's variance. The coverage rates relying on $B = 1000$ bootstrap samples in Table 1 seem to confirm this recommendation for both approaches.

3 Semiparametric Mode Regression

3.1 Semiparametric Modeling

So far, for mode regression, predictors have been restricted to parametric effects – either due to methodical reasons or to ensure numerical stability. In contrast, the NILS approach allows

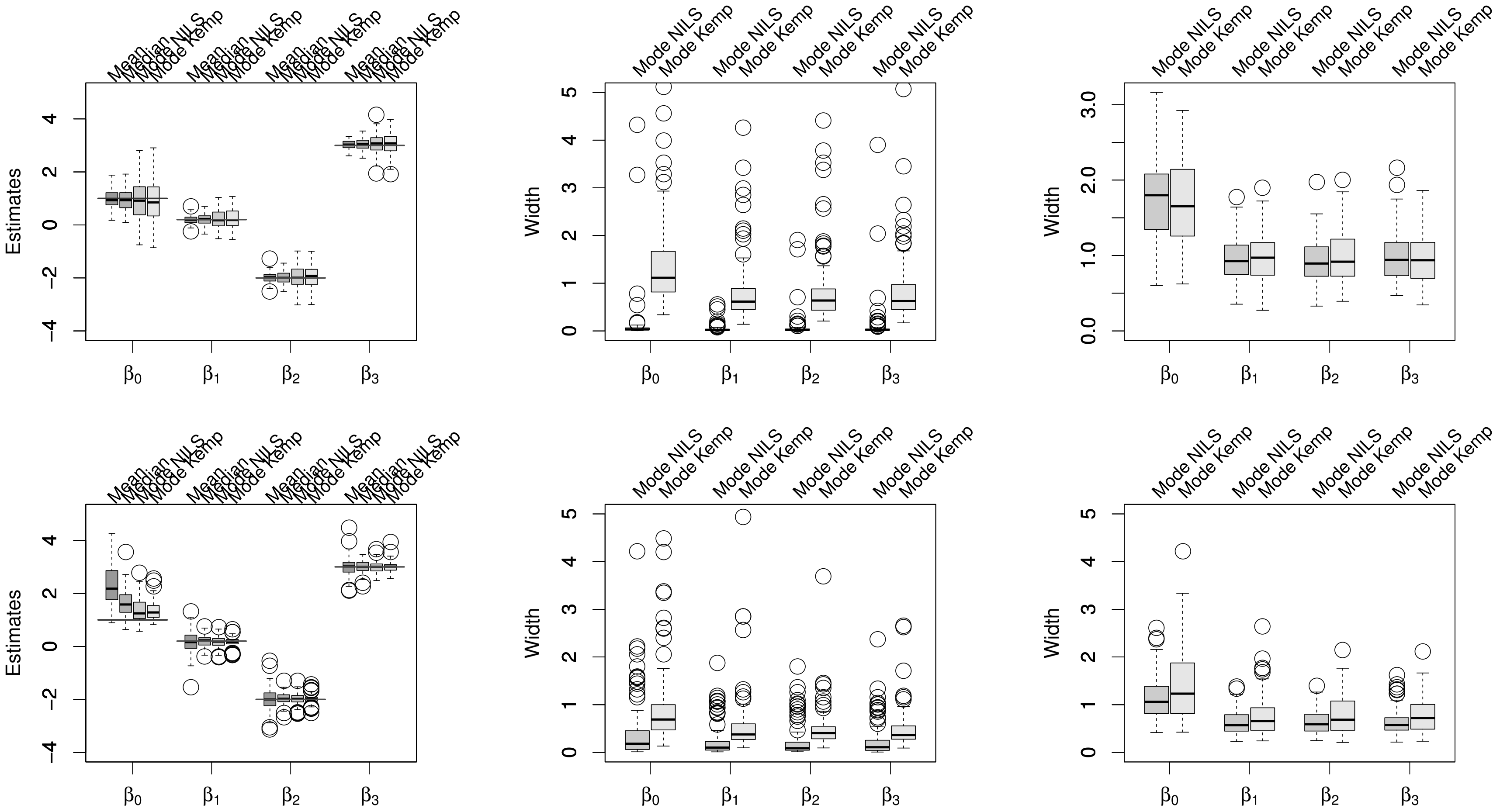

Figure 3: Estimated regression coefficients (left), estimated widths of the asymptotic confidence intervals (middle) and of the bootstrap confidence intervals (right); $n = 100$ observations; $\varepsilon \sim N(0,1)$ (top) and $\varepsilon \sim LN(0,1)$ (bottom). The true coefficients are indicated by horizontal lines.

to easily augment the linear predictor in model (1) to semiparametric predictors of the form

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \sum_{j=1}^{r} f_j(z_j) + f_{geo}(s) + \varepsilon,$$

where as before, $\boldsymbol{x}^T\boldsymbol{\beta}$ represent the linear effects. The functions f_j represent nonlinear smooth effects of continuous covariates z_j, $j = 1, \ldots, r$, modeled by penalized B-splines (Eilers and Marx, 1996) of degree 3 and with 20 outer knots as a default option. The effect f_{geo} allows to include spatial information which will be relevant in our application on the Munich rent index in Section 5.

Semiparametric models – which are also known as generalized additive (mixed) models (Hastie and Tibshirani, 1990; Wood, 2006) – are an established tool in many fields of regression modeling. And in fact, the predictor above is not the most general form. For mean regression, Fahrmeir et al. (2013) give an extensive overview of generic predictor representations where further effect types such as interactions between two continuous covariates or random effects can be included into the predictor. The general assumption is that each function f (independent of the type of the covariate x) can be written as a linear combination of appropriate basis functions, that is, $f(x) = \sum_{k=1}^{d} B_k(x)\beta_k$ which allows to write a vector of n function evaluations in matrix notation as $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{\beta}$. To achieve specific properties such as the smoothness of a function f, estimation is regularized by additional penalty terms. Specifically, we assume quadratic penalties of form $P_\lambda(\boldsymbol{\beta}) = \lambda\boldsymbol{\beta}^T\boldsymbol{K}\boldsymbol{\beta}$, where $\boldsymbol{K} \in \mathbb{R}^{q\times q}$ is an appropriate penalty matrix and $\lambda \geq 0$ is a penalty parameter that determines the strength of the regularization. Estimation in mode regression is then enabled by augmenting matrix $\boldsymbol{A}$ defined in equation (8):

$$\boldsymbol{A} \;=\; \boldsymbol{X}^T \mathrm{diag}\left(\frac{\mathcal{D}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(l)})}{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(l)}}\right)\boldsymbol{X} + \lambda\boldsymbol{K}.$$

Like a modular system and with none but the usual restrictions, mode regression can be combined with any quadratic penalty and/or smooth component. As we do work with an IRLS algorithm, the proposed approximation can be combined with the `R` package `mgcv` (Wood, 2011) such that a wide range of smooth components and several options to choose the penalty parameter λ are available. Note that the asymptotic theory of Section 2.2 does not include penalized estimates. While for fixed smoothing parameters one might argue that the asymptotic theory may carry over to penalized estimation, a much more careful investigation would be required for data-driven smoothing parameter estimates. Anyway, to achieve reliable results, a moderate

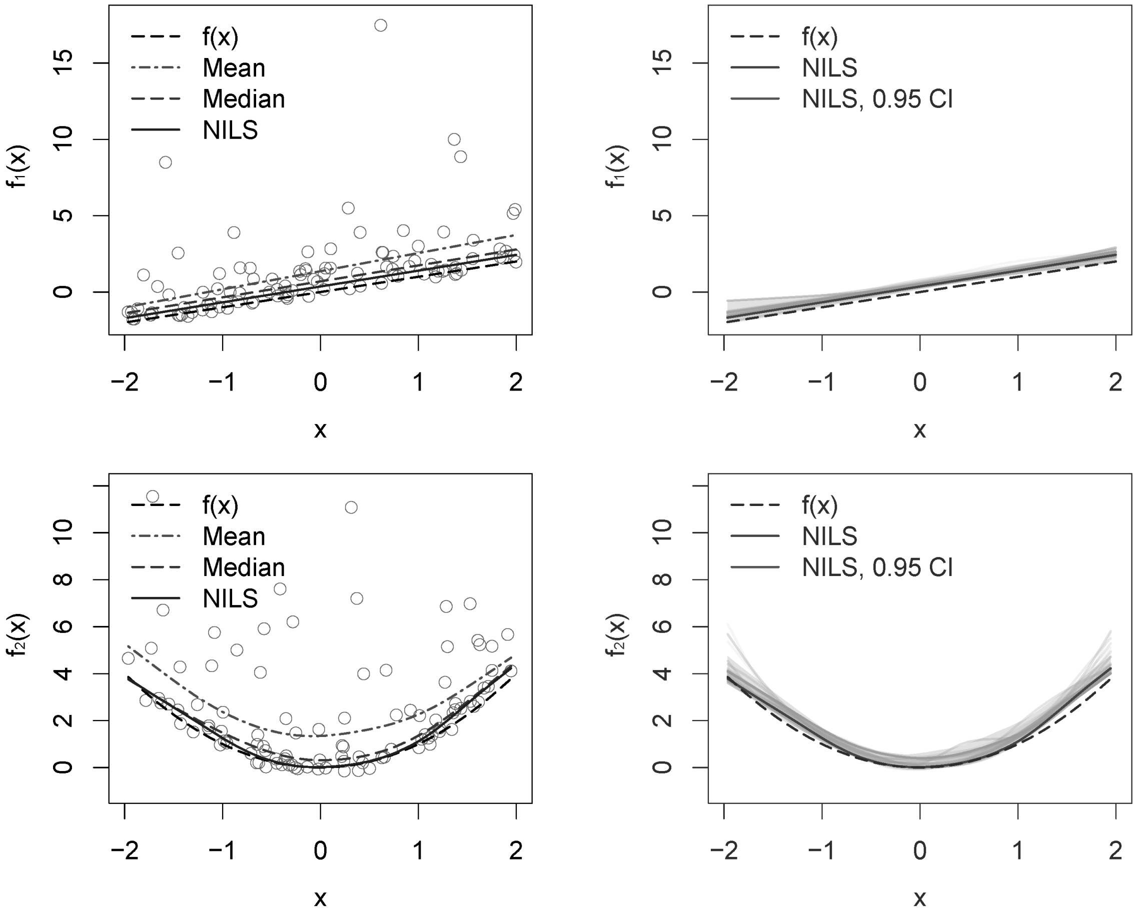

Figure 4: Examples for data fits of functions $f_1(x)$ and $f_2(x)$ (by rows) for $\varepsilon \sim LN(0, 1)$, $n = 100$. On the left, the results of mean, median and mode regression are compared. On the right, the bootstrap confidence intervals for the NILS approach are illustrated. The penalty parameter λ is chosen by the REML criterion.

number of observations relative to the model complexity was already required in a parametric setting, compare Section 2.4. Therefore, bootstrap methods turned out to be an attractive alternative. We will likewise employ bootstrap methods to asses how stable the estimated effects are in semiparametric mode regression. More specifically, we consider the pointwise $\alpha/2$ and $1 - \alpha/2$ quantiles of the functions fitted on the bootstrap samples in order to judge the variability of a fitted function.

3.2 Numerical Experiments

We investigate the performance of the proposed methods empirically. In contrast to the previous settings, penalized smooth components require to choose the penalty parameter(s) λ

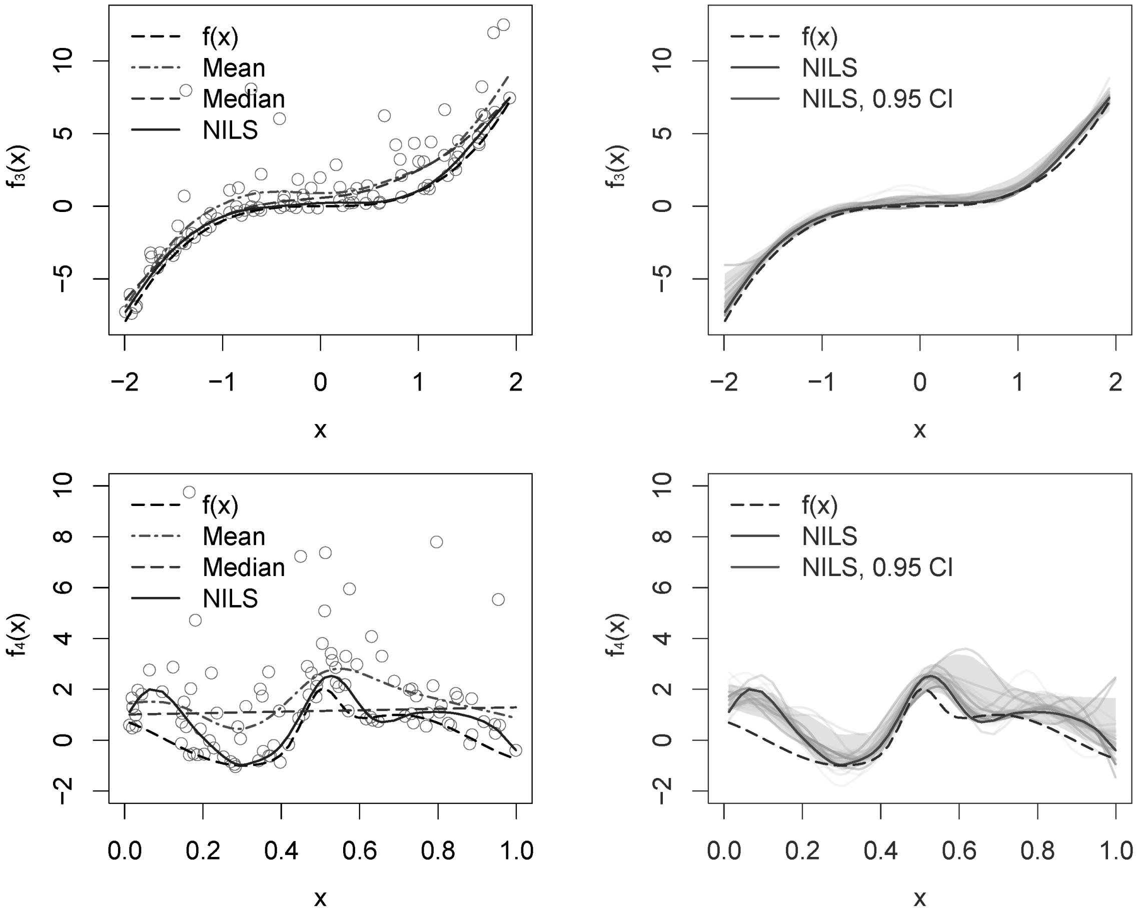

Figure 5: Examples for data fits of functions $f_3(x)$ and $f_4(x)$ (by rows) for $\varepsilon \sim LN(0,1)$, $n = 100$. On the left, the results of mean, median and mode regression are compared. On the right, the bootstrap confidence intervals for the NILS approach are illustrated. The penalty parameter λ is chosen by the REML criterion.

adequately. For mean regression, different strategies such as k-fold cross-validation with a specific loss criterion or the generalized cross-validation criterion of O'Sullivan et al. (1986) are available. Often, these criteria are based on rank estimation, that is, on the estimated hat matrix, or estimated degrees of freedom. It is not clear whether this makes sense for the employed loss function and if so, how data-sensitive the proposed method is. Moreover, combining the estimation with the R package mgcv implies that the estimation of the coefficient vector $\boldsymbol{\beta}$ and of the penalty parameter λ are interlaced. Hence, we consider not only the performance of semiparametric mode regression but compare the performance of different strategies for the choice of λ. Concretely, we consider (i) the generalized cross-validation criterion of O'Sullivan et al. (1986) where $\boldsymbol{\beta}$ and λ are estimated separately (referred to as "CV"), (ii) the same generalized cross-validation criterion with interlaced estimation ("GCV") and (iii) the negative log restricted likelihood criterion with interlaced estimation ("REML"); whereat (ii) and (iii) are implemented in `mgcv`. As a benchmark, we combine the approach of Kemp and Santos Silva (2010) with quadratic penalties, too.

The adaptive tuning depends on the assumptions for the asymptotic theory for parametric models and requires a preceding median regression. As in some semiparametric settings the results of median regression differ substantially from those of the mode regression, we avoid adaptive tuning in the following. Instead, tuning parameters that summarize the experiences of the data-adaptive choice of k in the parametric settings in Section 2.4 are employed. For samples sizes $n = 100$ and $n = 500$, we consider $n_{rep} = 100$ replications of the model $y = f(x) + \varepsilon$. The errors ε are normally or log-normally distributed as in Section 2.4. In order to consider an extreme scenario where the mode should be found at the lower boundary of the data, exponential errors $\varepsilon \sim Exp(0.5)$ are included. The data generating function $f(x)$ is chosen as either

- a linear effect $f_1(x) = x$,
- a parabola $f_2(x) = x^2$,
- a cubic polynomial $f_3(x) = x^3$
- or a trigonometric function $f_4(x) = \sin(2(4x - 2)) + 2\exp(-16^2 \cdot (x - 0.5)^2)$.

The covariate x is uniformly distributed on $[-2, 2]$ for $f_1(x)$, $f_2(x)$, $f_3(x)$ and on $[0, 1]$ for $f_4(x)$. The functions are modeled with cubic B-spline bases with 20 equally spaced outer knots and second order differences in the penalty matrix.

Figures 4 and 5 show the fitted functions for exemplary data sets with sample size $n = 100$. In the left panels, the results of mean, median and mode regression are compared while in the right panels, $1 - \alpha = 0.95$ confidence intervals based on $B = 100$ bootstrap samples illustrate the variability of the estimation procedure.

Results To evaluate the results, the root mean squared errors (RMSE) for the fitted values are shown in Figures 6 and 7 for $n = 100$. We conclude:

- Combining penalized splines and the NILS approach seems to be very reasonable for the purpose of a nonlinear mode regression, especially, when the penalty parameter λ is chosen by GCV or REML. With skew errors and GCV/REML, the NILS approach results in the lowest RMSE in nearly all settings. For normal errors, it is obviously less efficient, but the loss is about of the same magnitude as from mean to median regression.

- The performance of the approach of Kemp and Santos Silva (2010) depends strongly on the set of starting values. Even though the boxplots are based on the best starting values we found (a preceding mean regression), the approach of Kemp and Santos Silva (2012) comes along with a distinctively larger RMSE in nearly all settings.

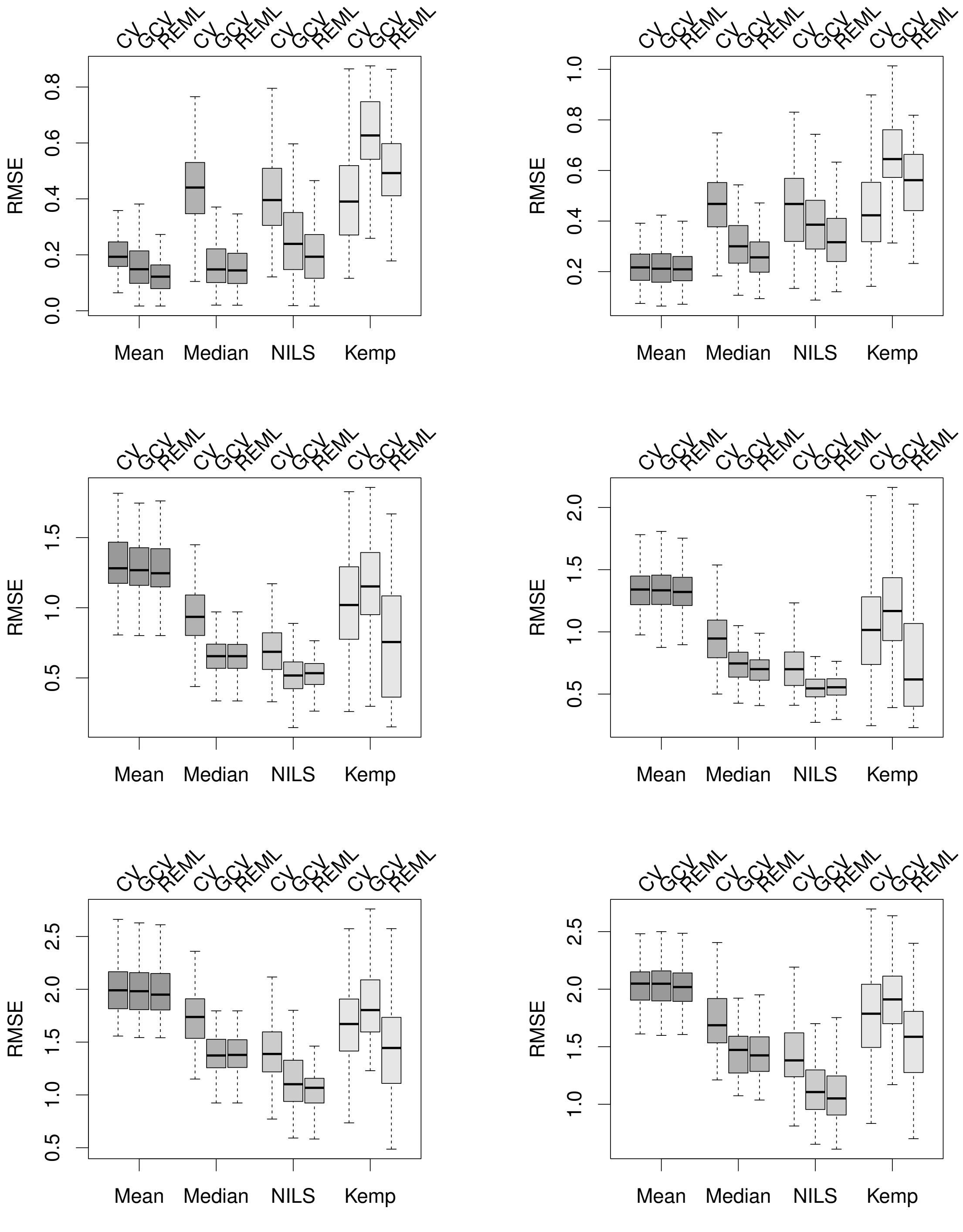

Figure 6: RMSE for function $f_1(x)$ (left) and function $f_2(x)$ (right) for normal (top), log-normal (middle) and exponential (bottom) errors; $n = 100$.

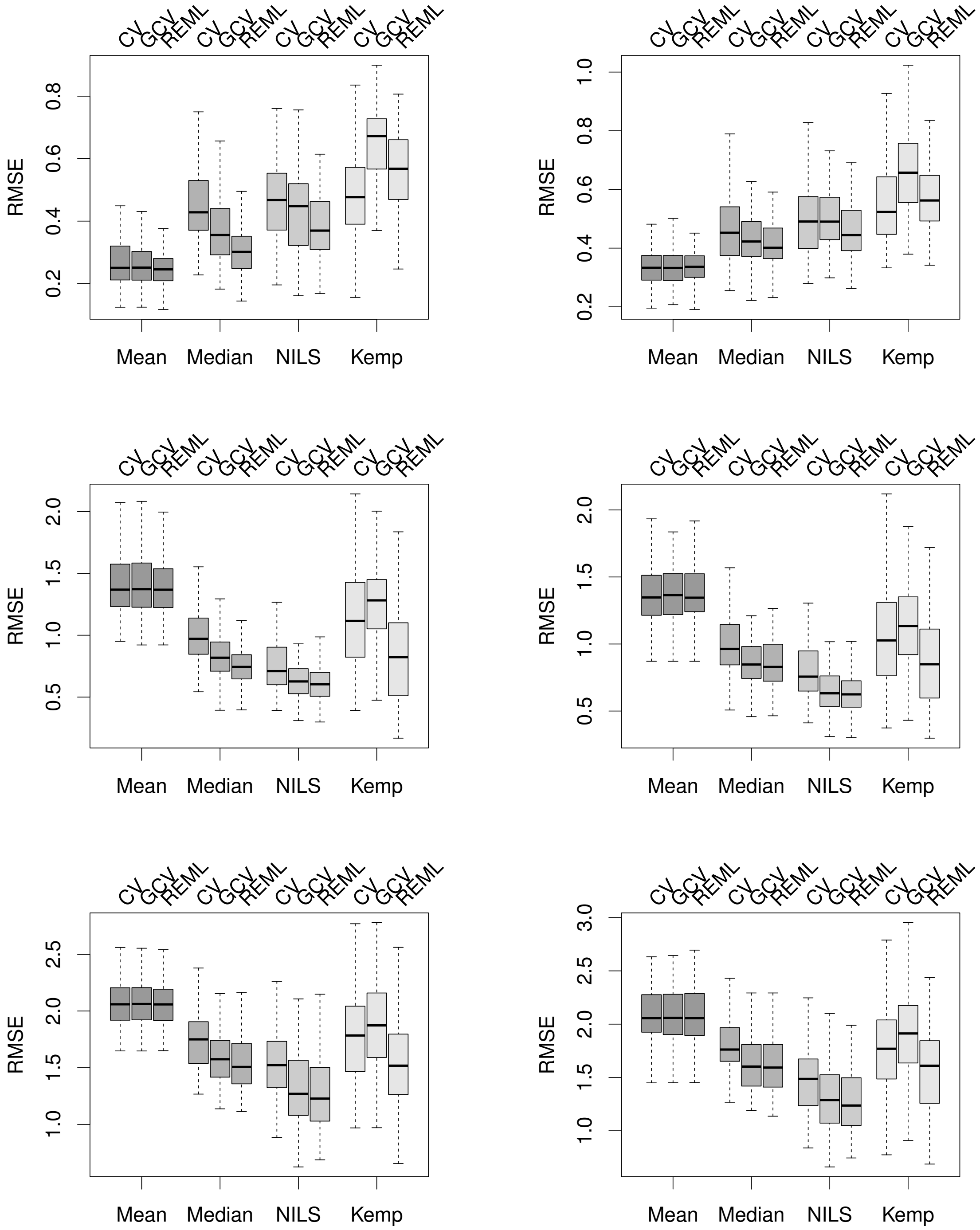

Figure 7: RMSE for function $f_3(x)$ (left) and function $f_4(x)$ (right) for normal (top), log-normal (middle) and exponential (bottom) errors; $n = 100$.

	Model (12)			Model (13)			Model (14)		
	Mean	Median	Mode Kemp	Mean	Median	Mode NILS	Mean	Median	Mode NILS
β_0	26.610	25.380	23.846	26.437	25.303	23.088	26.533	25.382	23.049
β_n	0.074	0.426	-0.354	0.074	0.431	0.009	0.075	0.444	-0.084
β_y	0.064	0.052	-0.028	0.064	0.051	-0.024	–	–	–
β_{a1}	3.051	3.549	4.095	–	–	–	–	–	–
β_{a2}	-0.342	0.565	0.088	–	–	–	–	–	–
β_{a3}	0.733	0.839	-2.059	–	–	–	–	–	–

Table 2: Estimated parametric effects for the mean, median and mode of the BMI in the considered models.

4 Mode Regression for the BMI Distribution in England

To explain the development of the body mass index (BMI) in England, we reanalyze a data set already used in Kemp and Santos Silva (2010) with a focus on non-pregnant women between the ages of 18 and 65 observed in the period between 1997 and 2006. This yields a data set of 44,651 observations with the age, the calendar year of the study and a binary factor indicating non-white women as available covariates. Our first model is in accordance with Kemp and Santos Silva (2010) where the effect of age is modeled by a polynomial while the other covariates are treated linearly:

$$\text{BMI} = \beta_0 + \beta_n\text{non-white} + \beta_y\text{year} + \beta_{a1}\log(\text{age}) + \beta_{a2}\log(\text{age})^2 + \beta_{a3}\log(\text{age})^3 + \varepsilon. \quad (12)$$

As seen in Table 2, one finds a slightly negative effect of the calendar year in mode regression while in mean regression, the effect of the calendar year is positive. In the left panel of Figure 8, the estimated effect of the age is shown.

A more flexible way to model the effect of the age is to replace the linear predictor above with

$$\text{BMI} = \beta_0 + \beta_n\text{non-white} + \beta_y\text{year} + f(\text{age}) + \varepsilon, \quad (13)$$

where $f(\text{age})$ is modeled by penalized cubic B-splines with 14 knots as the default choice of 20 knots caused some numerical instabilities. The set of tuning parameters is chosen as described in Section 3.2, and the penalty parameter λ is chosen by the REML criterion. Estimates of the parametric effects are given in Table 2, the estimate of the smooth function is shown in the right panel of Figure 8. At first sight, the effect of the age seems to be wigglier, but the trend does perfectly fit to the routines of a typical lifestyle and to typical biological changes: The

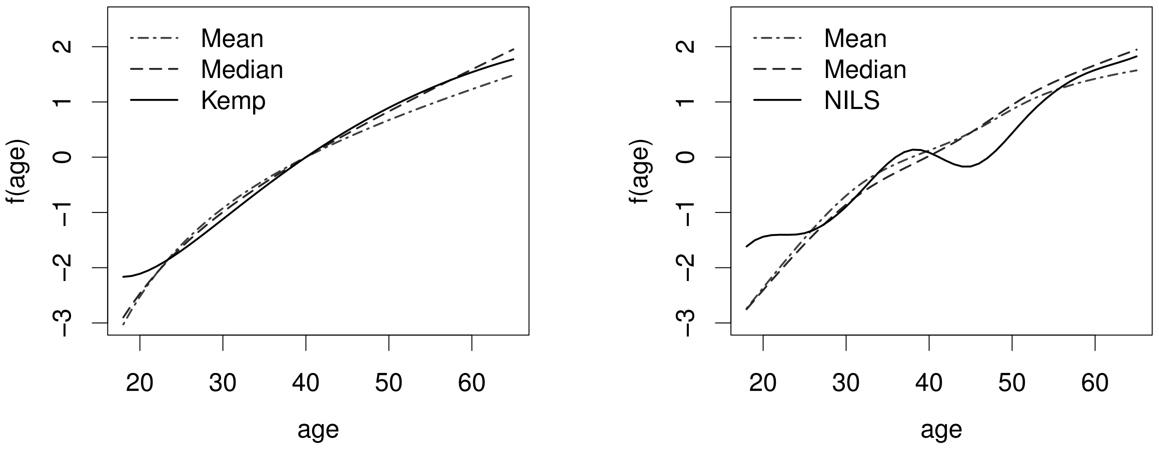

Figure 8: The estimated effect of age in model (12) (left panel) and in model (13) (right panel). For comparison, the results of mean and median regression are added.

effect of the BMI is relatively constant in early adulthood and increases around the age of 30. The second increase of the effect coincides with the typical age of the climacteric period.

In a third model, not only the effect of the age but of the age and of the calendar year are modeled smoothly:

$$\text{BMI} = \beta_0 + \beta_n \text{non-white} + f_1(\text{year}) + f_2(\text{age}) + \varepsilon, \tag{14}$$

Again, the estimates of parametric effects are given in Table 2. In Figure 9 (top), the estimates of f(age) and f(year) are plotted. For both effects, there is a clear difference between the fitted functions for mean, median and mode regression. For mean and median regression, the estimated effect of the age has the same functional form as in model (13), while the estimated effect of the calendar year is an increasing function suggesting an increasing BMI over time. However, for mode regression, the effect is ambiguous: There is a positive effect in the first two years of the study, but a negative one in the last two years. In Figure 9 (bottom), the results of the models fitted on $B = 50$ bootstrap samples are added confirming the shape of the effect of the age on the mode of the response (left panel). As before, the effect of the year cannot be clearly classified (right panel).

As seen in the empirical evaluation in Section 2.4, the differences in the estimated intercepts $\hat{\beta}_0$ as seen in Table 2 indicate a general skewness in the conditional distribution of the BMI. However, in Section 2.4, the effects of the covariates are estimated equally well for mean, median and mode regression even when the errors are skewed. As the estimated effects in Table 2 differ, this may be an indication for a more complex error structure.

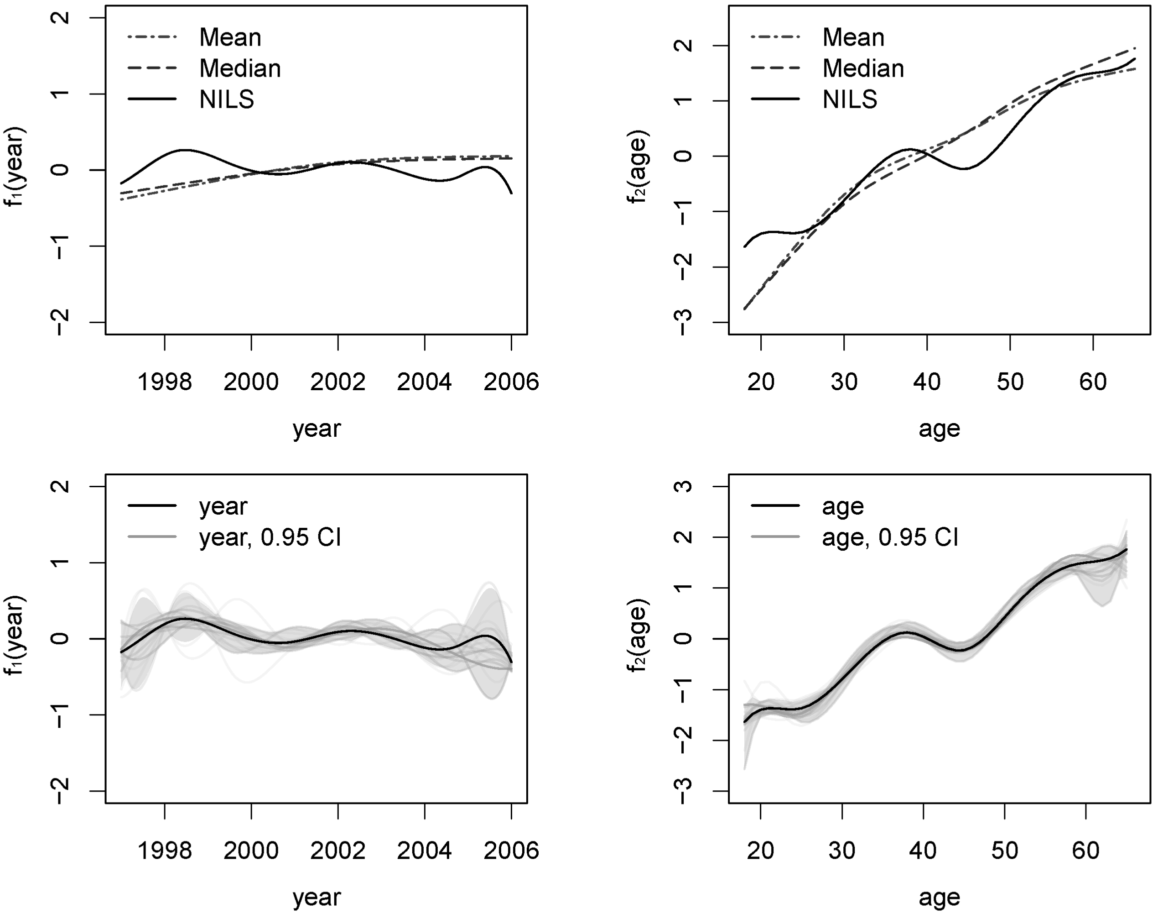

Figure 9: The estimated effects of the calendar year (left panel) and the age (right panel) in model (14). On top, the results of mean and median regression are added. On bottom, fitted functions for $B = 50$ bootstrap samples are added.

5 The Munich Rent Index

In a second application, we analyze data on the rents in Munich with mode regression. The data has been collected in 2003 and gives detailed information on the living conditions and the associated costs of 3051 flats in Munich. Previous analyses of this data set show strong nonlinear and spatial effects on the expected net rent as dependent variable, but also reveal the presence of heteroscedasticity and skewness (Kneib, 2013). Hence, we compare the results of mean, median and mode regression for the model equation

$$\text{rent} = \boldsymbol{x}^T\boldsymbol{\beta} + f_1(\text{year}) + f_2(\text{size}) + f_{\text{spat}}(\text{district}) + \varepsilon, \tag{15}$$

where $\boldsymbol{x}$ consists of 12 categorical covariates indicating several quality attributes of a flat such as the kitchen equipment or the type of heating (see Table 3 for a complete list). Functions

	Mean		Median		NILS	
Intercept	8.84	(8.81, 8.94)	8.91	(8.87, 9.01)	8.93	(8.87, 9.05)
Absence of bathroom	0.89	(0.52, 1.18)	0.71	(0.35, 1.20)	0.23	(-0.16, 0.85)
Presence of second bathroom	-0.64	(-0.79, -0.48)	-0.87	(-1.07, -0.70)	-0.97	(-1.28, -0.81)
Special features of bathroom	0.54	(0.37, 0.78)	0.53	(0.30, 0.72)	0.71	(0.47, 0.90)
Normal quality kitchen	0.70	(0.57, 0.93)	0.64	(0.56, 0.86)	0.70	(0.59, 0.80)
Good quality kitchen	1.05	(0.86, 1.20)	1.13	(0.91, 1.29)	1.03	(0.68, 1.16)
Absence of intercom	-0.47	(-0.66, -0.41)	-0.58	(-0.77, -0.51)	-0.62	(-0.86, -0.58)
Simple floor cover	-1.10	(-1.33, -0.96)	-1.05	(-1.28, -0.90)	-1.00	(-1.21, -0.84)
Absence of warm water supply	-1.39	(-1.81, -1.02)	-1.42	(-1.92, -1.01)	-1.82	(-2.70, -1.27)
Absence of central heating system	-1.16	(-1.29, -0.86)	-1.31	(-1.56, -0.96)	-1.31	(-1.50, -0.67)
Presence of storage heating system	-0.79	(-1.10, -0.57)	-0.66	(-0.96, -0.50)	-0.38	(-0.68, 0.09)
Simple and old building	-0.71	(-0.96, -0.54)	-0.83	(-1.18, -0.65)	-0.67	(-0.94, -0.27)
Simple and post world war building	-0.64	(-0.83, -0.31)	-0.84	(-1.05, -0.51)	-1.18	(-1.37, -0.74)

Table 3: List of categorical covariates in model (15) and the corresponding estimated effects for mean, median and mode regression. In parenthesis, there are 95% confidence intervals based on $B = 50$ bootstrap samples.

f_1(year) and f_2(size) represent nonlinear effects of the year of construction and of the size of the flat (in square meters). They are approximated by penalized cubic splines with 14 outer knots and with a second order difference penalty. The spatial effect f_{spat}(district) is defined by 100 districts in Munich and estimated by a Markov random field (Rue and Held, 2005).

The estimated coefficients for the categorical covariates obtained with mode, mean and median regression are given in Table 3. While, for the mean, a second bathroom makes the flat about 0.89 € per square meter more expensive, the effect of this covariate is stronger for mode regression. We also see that flats in simple post-war buildings are generally cheaper, but there is a clear difference between a price reduction of 0.64 € on average and a reduction of 1.18 € for mode regression. Overall, the table shows that the estimates for the mean and the median are very similar while we can find stronger differences in comparison to the mode regression estimates. Therefore, we might distinguish between average rents and typical rents. This is supported by the estimated nonlinear effect of the size of a flat in square meters as shown in Figure 10. Again, we find strong similarities between the mean and median while the estimated effect for mode regression is less extreme, especially for flats larger than $140\,m^2$. Similarly, the estimated spatial effects of mean and mode regression depicted in Figure 11 show similar patterns even though the results of the median regression are less extreme. The results of mode regression show that the typical rents of a few outlying districts differ immensely from those estimated by the other location measures. In fact, the variability of the spatial effect is the largest for mode regression and the smallest for median regression.

Overall, the estimated tendencies are roughly the same for mean, median and mode regression

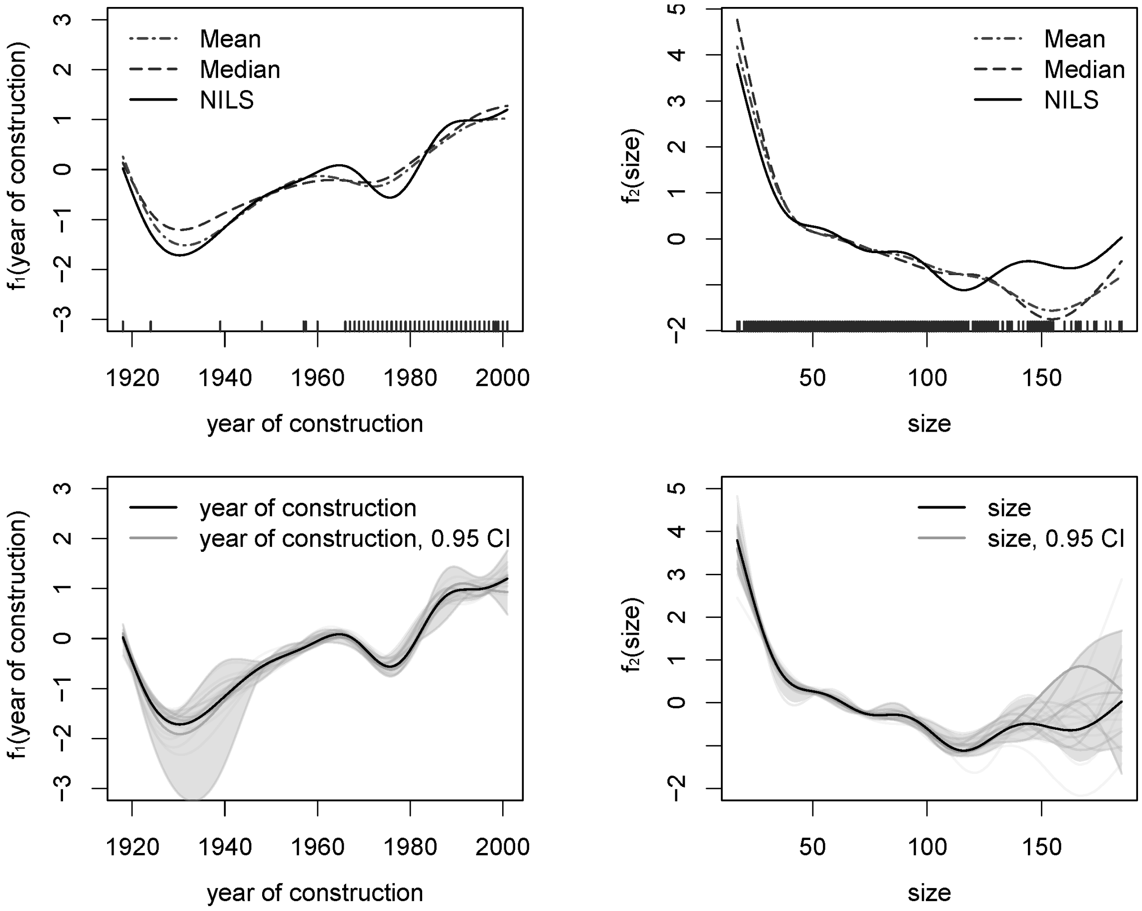

Figure 10: Estimated effects of year of construction and size in square meters in model (15).

– despite some effects that are remarkably different for average and for typical flats. The results of mode regression point out effects that should be investigated carefully in order to understand the pricing mechanism.

6 Remarks

We developed a new estimator for the conditional mode based on a local quadratic approximation $\mathcal{L}$ of the limiting case in (5) which can be determined iteratively with a nested interval approach. The properties of our kernel function allow to adapt asymptotic properties of the estimator in a parametric setting. However, similar to Kemp and Santos Silva (2012), the rate of convergence is rather slow such that depending on the error structure, confidence intervals are much to narrow. In most situations, this problem can be reduced with bootstrap methods. The main advantage of our approach is that it can easily be extended to semiparametric pre-

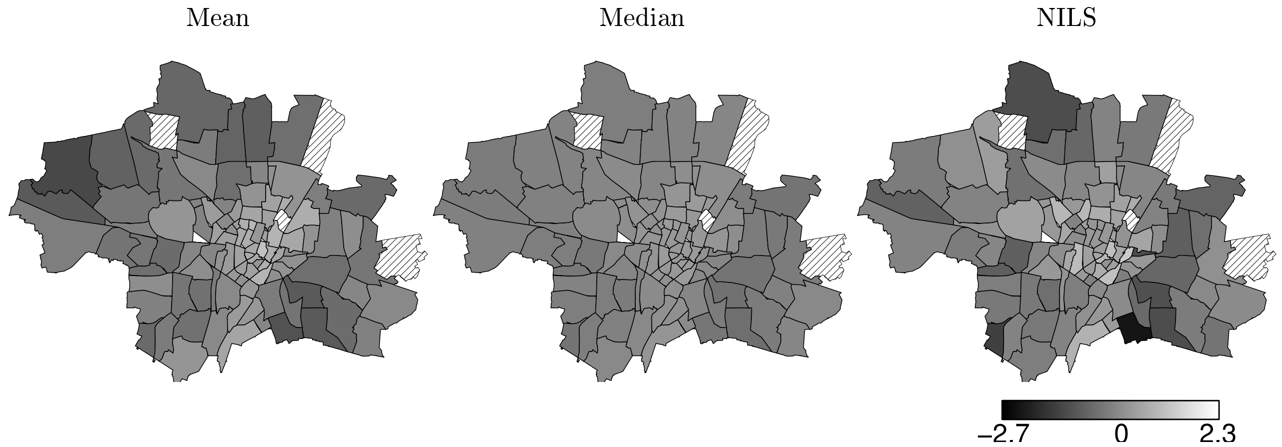

Figure 11: Estimated spatial effects of the districts in model (15) in mean, median and mode regression. For the hatched areas, there are no observations. The figure is created with the `R` package `BayesX` (Kneib et al., 2014).

dictor structures, yielding considerably expanded flexibility in the specification of conditional mode regression. The penalized IRLS framework also allows to borrow existing inferential tools from mean regression, for example, for the determination of smoothing parameters. An open question for future research is the extension of the asymptotic results to such semiparametric specifications, especially when including data-driven estimates for the smoothing parameter and/or basis dimensions increasing with the sample size.

In cases, where the error structure is additive, mean, median and mode regression should only differ in a shift of the intercepts such that mode regression (in addition to the appeals mentioned in the introduction) can be a helpful tool to draw conclusions about the underlying error structure. Although Taylor and Einbeck (2011) show that the true multivariate mode regression is hard to interpret due to the non-additivity of the predictor, extending our approach to bivariate problems would allow to study the mode of a joint bivariate distribution and is conceptually straight forward.

Acknowledgments

We thank Gerhard Tutz sincerely for his support and for his helpful comments. Financial support was given by the German Research Foundation (DFG) via the research project TU 62/5-1, the research training group 1644 on “Scaling Problems in Statistics” and the research project KN 922/4-1.

A Derivation of the IRLS Algorithm

In what follows, the iteratively re-weighted least squares (IRLS) algorithm from page 7 is derived. We start with a first order Taylor expansion of $\mathcal{M}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_{(l)}$:

$$\begin{aligned}
\mathcal{M}(\boldsymbol{\beta}) &\approx \mathcal{M}(\boldsymbol{\beta}_{(l)}) + \nabla\mathcal{M}(\boldsymbol{\beta}_{(l)})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) \\
\nabla\mathcal{M}(\boldsymbol{\beta}_{(l)})^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) &= \sum_{i=1}^{n}\left(\nabla\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})\right)^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) \\
\nabla\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}) &= \frac{\partial\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{\partial(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}\frac{\partial(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{\partial\boldsymbol{\beta}} \\
&= \mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})\cdot(-\boldsymbol{x}_i) \\
\text{with}\mathcal{D}(\xi) &= \frac{\partial\mathcal{L}(\xi)}{\partial\xi} = k((k\xi)^{2g}+c)^{\frac{1}{2g}-1}(k\xi)^{2g-1} \\
&\quad \cdot\exp(c^{\frac{1}{2g}}-((k\xi)^{2g}+c)^{\frac{1}{2g}}).
\end{aligned}$$

The local trick of Fan and Li (2001) gives

$$(\nabla\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}))^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) \approx \mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})\cdot\frac{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}}\cdot(-\boldsymbol{x}_i^T)\cdot(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}).$$

With the quadratic approximation of Ulbricht (2010), we obtain

$$\begin{aligned}
(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta})(-\boldsymbol{x}_i^T)(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) &= -y_i\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)})+\boldsymbol{x}_i^T\boldsymbol{\beta}\boldsymbol{x}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) \\
&\approx -y_i\boldsymbol{x}_i^T\boldsymbol{\beta}+y_i\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}+\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}+\boldsymbol{\beta}_{(l)}^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}).
\end{aligned}$$

Overall, we have

$$\begin{aligned}
(\nabla\mathcal{L}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}))^T(\boldsymbol{\beta}-\boldsymbol{\beta}_{(l)}) &\approx -\boldsymbol{a}_i^T\boldsymbol{\beta}+\boldsymbol{a}_i^T\boldsymbol{\beta}_{(l)}+\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}_i\boldsymbol{\beta}+\boldsymbol{\beta}_{(l)}^T\boldsymbol{A}_i\boldsymbol{\beta}_{(l)}) \\
\text{with } \boldsymbol{a}_i^T &= \frac{\mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}}y_i\boldsymbol{x}_i^T \\
\text{and } \boldsymbol{A}_i &= \frac{\mathcal{D}(y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)})}{y_i-\boldsymbol{x}_i^T\boldsymbol{\beta}_{(l)}}\boldsymbol{x}_i\boldsymbol{x}_i^T.
\end{aligned}$$

Hence, the approximated objective function can be written as

$$\mathcal{M}(\boldsymbol{\beta}) \approx \mathcal{M}(\boldsymbol{\beta}_{(l)})-\boldsymbol{a}^T\boldsymbol{\beta}+\boldsymbol{a}^T\boldsymbol{\beta}_{(l)}+\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{A}\boldsymbol{\beta}+\boldsymbol{\beta}_{(l)}^T\boldsymbol{A}\boldsymbol{\beta}_{(l)})=\mathcal{M}^{app}.$$

$\mathcal{M}^{app}$ has the derivatives

$$\begin{aligned} s(\boldsymbol{\beta}) &= \frac{\mathcal{M}^{app}}{\partial \boldsymbol{\beta}} = -\boldsymbol{a} + \boldsymbol{A}\boldsymbol{\beta} \\ \text{and } H(\boldsymbol{\beta}) &= \frac{\mathcal{M}^{app}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \boldsymbol{A}. \end{aligned}$$

Hence, the update of the current estimate $\boldsymbol{\beta}_{(l)}$ in iteration (l) is

$$\begin{aligned} \boldsymbol{\beta}_{(l+1)} &= \boldsymbol{\beta}_{(l)} - \nu H(\boldsymbol{\beta}_{(l)})^{-1} s(\boldsymbol{\beta}_{(l)}) \\ &= \boldsymbol{\beta}_{(l)} - \nu \boldsymbol{A}_{(l)}^{-1}(-\boldsymbol{a}_{(l)} + \boldsymbol{A}_{(l)}\boldsymbol{\beta}_{(l)}) \\ &= \boldsymbol{\beta}_{(l)} + \nu \boldsymbol{A}_{(l)}^{-1}\boldsymbol{a}_{(l)} - \nu \boldsymbol{A}_{(l)}^{-1}\boldsymbol{A}_{(l)}\boldsymbol{\beta}_{(l)} \\ &= (1-\nu)\boldsymbol{\beta}_{(l)} + \nu \boldsymbol{A}_{(l)}^{-1}\boldsymbol{a}_{(l)}. \end{aligned}$$

Apart from the update of the derivative $\mathcal{D}(\xi)$ of the approximated loss, the algorithm is as complex as usual IRLS algorithms. With $\boldsymbol{A}_{(l)}$ positive definite for all iterations l, the algorithm converges almost surely. However, as $\mathcal{L}(\xi) = 1 = const.$ and therewith $\mathcal{D}(\xi) = 0$ for sufficiently large $|\xi|$, the initial values of $\beta_{(l)}$ have to be chosen carefully. The algorithm can be easily implemented; we employ an R (R Core Team, 2013) implementation.

B Proofs for Section 2.2

B.1 Characteristics of K and Some Auxiliary Functions

Note that

$$\exp(-\sqrt{u^2+c}) \quad < \quad \exp(-|u|), \tag{16}$$

as $\sqrt{u^2+c} > |u|$ for all $u \in \mathbb{R}$.

$$\int_{-\infty}^{\infty} \frac{1}{2}\exp(-|u|)\mathrm{d}u \quad = \quad 1, \tag{17}$$

since $\frac{1}{2}\exp(-|u|)$ is the density of the Laplace distribution with $\mathbb{E}(u) = 0$, $\mathbb{V}(u) = 2$. Moreover,

$$\frac{u}{\sqrt{u^2+c}} \quad < \quad 1. \tag{18}$$

The derivatives of K are

$$\begin{aligned} K'(u) &= -\frac{1}{2}\exp(-\sqrt{u^2+c})\frac{u}{\sqrt{u^2+c}}, && (19)\\ K''(u) &= \frac{1}{2}\exp(-\sqrt{u^2+c})\left(\frac{u^2}{u^2+c}-\frac{c}{(u^2+c)^{3/2}}\right), && (20)\\ K'''(u) &= u\exp(-\sqrt{u^2+c})\left(\frac{1.5c}{(u^2+c)^{5/2}}-\frac{u^2\sqrt{u^2+c}-3c}{2(u^2+c)^2}\right). && (21) \end{aligned}$$

Note: $|K'|$, $|K''|$ and $|K'''|$ are symmetric around zero.

$$\int_0^\infty u^a \exp(-bu)\mathrm{d}u = \frac{a!}{b^{a+1}}, \qquad (22)$$

for $a \in \mathbb{Z}$, $b > 0$, see Gradshteyn and Ryzhik (2007; Section 3.326, equation $2.^{10}$).

Define

$$s : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto s(u) = \left(\frac{u^2}{u^2+c}-\frac{c}{(u^2+c)^{3/2}}\right), \qquad (23)$$

with $|s(u)| \le c^{-1/2}$ as:

- $$\frac{\partial s(u)}{\partial u} = \frac{cu(2\sqrt{u^2+c}+3)}{(u^2+c)^{5/2}},$$

  with roots at $u = 0$ and $u = \frac{1}{2}\sqrt{9-4c}$.

- $|\min_{u\ge 0} s(u)| = |s(0)| = |-c^{-1/2}| > |\max_{u\ge 0} s(u)| = |s(\frac{1}{2}\sqrt{9-4c})| = |1-\frac{20}{27}c|$.

$$\begin{aligned} \sup_{u\ge 0}\exp(-u) &= 1. && (24)\\ u\exp(-|u|) &< 1, \quad \text{for all} u \in \mathbb{R}. && (25) \end{aligned}$$

Define

$$t : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t(u) = \frac{1.5c}{(u^2+c)^{5/2}} - \frac{u^2\sqrt{u^2+c}-3c}{2(u^2+c)^2}, \qquad (26)$$

Later on, we need $|t(u)| \le \frac{3}{2}(c^{-1}+c^{-3/2})$. This follows from

-

$$\begin{aligned} t(u) &= t_{(+)}(u) - t_{(-)}(u) \quad \text{with} \\ t_{(+)} : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t_{(+)}(u) &= \frac{1.5c}{(u^2+c)^{5/2}} + \frac{3c}{2(u^2+c)^2}, \\ t_{(-)} : \mathbb{R}_0^+ \to \mathbb{R},\ u \mapsto t_{(-)}(u) &= \frac{u^2\sqrt{u^2+c}}{2(u^2+c)^2}. \end{aligned}$$

-

$$\frac{\partial t_{(+)}(u)}{\partial u} = -u\left(\frac{6c}{(u^2+c)^3} + \frac{7.5}{(u^2+c)^{7/2}}\right),$$

with one root at $u = 0$.

- $0 \leq t_{(+)}(u) \leq \max_{u\geq 0} t_{(+)}(u) = \frac{3}{2}(c^{-1} + c^{-3/2})$.

-

$$\frac{\partial t_{(-)}(u)}{\partial u} = \frac{2cu - u^3}{2(u^2+c)^{5/2}}$$

with roots at $u = 0$, $u = \sqrt{2c}$.

- $0 = \min_{u\geq 0} t_{(-)}(u) \leq t_{(-)}(u) \leq \max_{u\geq 0} t_{(-)}(u) = t_{(-)}(\sqrt{2c}) = 3^{-3/2}c^{-1/2}$.

- $\max_{u\geq 0} t_{(+)}(u) > \max_{u\geq 0} t_{(-)}(u) \Rightarrow |t(u)| \leq \frac{3}{2}(c^{-1} + c^{-3/2})$.

B.2 Proofs

Lemma 3. *The kernel function $K : \mathbb{R} \to \mathbb{R}$ defined in* (10) *is differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1$,

(ii) $\sup_{u\in\mathbb{R}} |K(u)| = c_0 < \infty$,

(iii) $\sup_{u\in\mathbb{R}} |K'(u)| = c_1 < \infty$, *where* $K'(u) = \mathrm{d}K(u)/\mathrm{d}u$.

Proof.

(i) $\int_{-\infty}^{\infty} K(u)\mathrm{d}u \overset{(16),(17)}{<} 1$ and $K(u) > 0$.
Hence, it exists a constant $\delta \in \mathbb{R}^+$ such that $\int_{-\infty}^{\infty} \delta K(u)\mathrm{d}u = 1$.

(ii) Using (16) and (24), it follows that

$$\sup_{u\in\mathbb{R}} |K(u)| < \max_{u\in\mathbb{R}} \frac{1}{2}\exp\left(-|u|\right) = \frac{1}{2} < \infty.$$

(iii) $|K'(u)|$ is symmetric around zero, see (19); hence, it is to prove that $\sup_{u\in\mathbb{R}_0^+} |K'(u)| < \infty$.

$$\begin{aligned}
|K'(u)| &= \left|-\frac{1}{2}\exp(-\sqrt{u^2+c})\frac{u}{\sqrt{u^2+c}}\right| \\
&\overset{(18)}{<} \frac{1}{2}\exp(-\sqrt{u^2+c}) \\
&\overset{(16)}{<} \frac{1}{2}\exp(-u) \\
&< \frac{1}{2}.
\end{aligned}$$

□

Lemma 4. *The kernel function* $K : \mathbb{R} \to \mathbb{R}$ *defined in* (10) *is three times differentiable and fulfills the following conditions:*

(i) $\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0$*,*

(ii) $\lim_{u\to\pm\infty} K(u) = 0$*,*

(iii) $\int_{-\infty}^{\infty} u^2|K(u)|\mathrm{d}u = M_0 < \infty$*,*

(iv) $\int_{-\infty}^{\infty} |K'(u)|^2\mathrm{d}u = M_1 < \infty$*,*

(v) $\sup_{u\in\mathbb{R}} |K''(u)| = M_2 < \infty$*,*

(vi) $\sup_{u\in\mathbb{R}} |K'''(u)| = M_3 < \infty$*,*

(vii) $\int_{-\infty}^{\infty} |K''(u)|^2\mathrm{d}u = M_4 < \infty$*.*

Proof. For differentiability, see (19), (20), (21).

(i)

$$\begin{aligned}
\int_{-\infty}^{\infty} uK(u)\mathrm{d}u &= \left[\frac{1}{2}\exp(-\sqrt{u^2+c})(-\sqrt{u^2+c}-1)\right]_{-\infty}^{\infty} \\
&= 0.
\end{aligned}$$

(ii) Follows directly from the definition.

(iii)

$$\begin{aligned}
\int_{-\infty}^{\infty} u^2 |K(u)| \mathrm{d}u &= 2 \cdot \frac{1}{2} \int_0^{\infty} u^2 \exp(-\sqrt{u^2+c}) \mathrm{d}u \\
&\overset{(16)}{<} \int_0^{\infty} u^2 \exp(-u) \mathrm{d}u \\
&\overset{(22)}{=} 2.
\end{aligned}$$

(iv)

$$\begin{aligned}
\int_{-\infty}^{\infty} |K'(u)|^2 \mathrm{d}u &\overset{(19)}{=} \int_{-\infty}^{\infty} \frac{1}{4} \exp(-2\sqrt{u^2+c}) \frac{u^2}{u^2+c} \mathrm{d}u \\
&= \frac{1}{2} \int_0^{\infty} \exp(-2\sqrt{u^2+c}) \frac{u^2}{u^2+c} \mathrm{d}u \\
&\overset{(18),(16)}{<} \frac{1}{2} \int_0^{\infty} \exp(-2u) \mathrm{d}u \\
&\overset{(22)}{=} \frac{1}{4}.
\end{aligned}$$

(v) $|K''(u)|$ is symmetric around zero, see (20); therefore, it is to prove that

$$\begin{aligned}
\sup_{u \in \mathbb{R}_0^+} |K''(u)| &< \infty. \\
|K''(u)| &= \left| \frac{1}{2} \exp(-\sqrt{u^2+c}) s(u) \right| \\
&\overset{(23)}{\leq} \frac{1}{2} \exp(-\sqrt{u^2+c}) c^{-1/2} \\
&\overset{(16)}{<} \frac{1}{2} \exp(-u) c^{-1/2} \\
&\overset{(24)}{\leq} \frac{1}{2\sqrt{c}}.
\end{aligned}$$

(vi) $|K'''(u)|$ is symmetric around zero, see (21); therefore, it is to prove that

$$
\begin{aligned}
\sup_{u\in\mathbb{R}_0^+} |K'''(u)| &< \infty. \\
|K'''(u)| &\overset{(21)}{=} \left|u\exp(-\sqrt{u^2+c})t(u)\right| \\
&\overset{(26)}{\leq} u\exp(-\sqrt{u^2+c})\frac{3}{2}(c^{-1}+c^{-3/2}) \\
&\overset{(16)}{<} u\exp(-u)\frac{3}{2}(c^{-1}+c^{-3/2}) \\
&\overset{(25)}{<} \frac{3}{2}(c^{-1}+c^{-3/2}).
\end{aligned}
$$

(vii)

$$
\begin{aligned}
\int_{-\infty}^{\infty} |K''(u)|^2 \mathrm{d}u &\overset{(20)}{=} 2\int_0^{\infty} |K''(u)|^2 \mathrm{d}u \\
&\overset{(20)}{=} 2\cdot\frac{1}{4}\int_0^{\infty} \exp(-2\sqrt{u^2+c})\left(\frac{u^2}{u^2+c}-\frac{c}{(u^2+c)^{3/2}}\right)^2 \mathrm{d}u \\
&= \frac{1}{2}\int_0^{\infty} \exp(-2\sqrt{u^2+c})\,(s(u))^2\,\mathrm{d}u \\
&\overset{(23)}{\leq} \frac{1}{2}\int_0^{\infty} \exp(-2\sqrt{u^2+c})c^{-1}\mathrm{d}u \\
&\overset{(16)}{<} \frac{1}{2c}\int_0^{\infty} \exp(-2u)\mathrm{d}u \\
&\overset{(22)}{=} \frac{1}{4c}.
\end{aligned}
$$

□

References

Collomb, G., W. Härdle, and S. Hassani (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference 15*, 227–236.

Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with *b*-splines and penalties. *Statistical Science 11*(2), 89–102.

Einbeck, J. and G. Tutz (2006). Modelling beyond regression functions: An application of multimodal regression to speed-flow data. *Journal of the Royal Statistical Society. Series C. Applied Statistics 55*(4), 461–475.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression – Models, Methods and Applications.* Heidelberg: Springer.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models.* SAGE Publications.

Gannoun, A., J. Saracco, and K. Yu (2010). On semiparametric mode regression estimation. *Communications in Statistics - Theory and Methods 39*(7), 1141–1157.

Gradshteyn, I. S. and I. M. Ryzhik (2007). *Table of Integrals, Series, and Products.* Table of Integrals, Series, and Products Series. Elsevier Science.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models.* Chapman & Hall, Ltd., London.

Held, L. (2008). *Methoden der statistischen Inferenz: Likelihood und Bayes.* Heidelberg: Spektrum Akademischer Verlag.

Kemp, G. C. and J. Santos Silva (2010). Regression towards the mode. Economics Discussion Papers 686, University of Essex, Department of Economics.

Kemp, G. C. and J. Santos Silva (2012). Regression towards the mode. *Journal of Econometrics 170*(1), 92–101.

Kneib, T. (2013). Beyond mean regression. *Statistical Modelling 13*(4), 275–303.

Kneib, T., F. Heinzl, A. Brezger, D. S. Bove, and N. Klein (2014). *BayesX: R utilities accompanying the software package BayesX.* R package versions 0.2-8/0.2-9.

Lee, M. (1989). Mode regression. *Journal of Econometrics 42*(3), 337–349.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics 57*(1–3), 1–19.

Manski, C. F. (1991). Regression. *Journal of Economic Literature 29*(1), 34–50.

O'Sullivan, F., B. S. Yandell, and W. J. Raynor, Jr. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association 81*(393), 96–103.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. R version 3.1.0 (2014-04-10).

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications.* Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Boca Raton, USA.

Taylor, J. and J. Einbeck (2011). Multivariate regression smoothing through the 'fallling net'. In D. Conesa, A. Forte, A. Lopez-Quilez, and F. Munoz (Eds.), *Proceedings of the 26th International Workshop on Statistical Modelling*, pp. 597–602.

Ulbricht, J. (2010). *Variable Selection in Generalized Linear Models.* Dissertation, Department of Statistics, Ludwig-Maximilians-Universität München: Verlag Dr. Hut.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society. Series B. Statistical Methodology 73*(1), 3–36. R package versions 1.7-29/1.8-2.

Wood, S. N. (2006). *Generalized Additive Models : An Introduction with R.* New York: Chapman & Hall.

Yu, K. and K. Aristodemou (2012). Bayesian mode regression. Technical report. `http://arxiv.org/abs/1208.0579v1`.