Annals of the Institute of Statistical Mathematics (2020) 72:1287–1315
https://doi.org/10.1007/s10463-019-00731-5

Semi-parametric transformation boundary regression models

Natalie Neumeyer[1] · Leonie Selk[1] · Charles Tillier[1]

Received: 4 December 2018 / Revised: 13 June 2019 / Published online: 21 September 2019

Abstract

In the context of nonparametric regression models with one-sided errors, we consider parametric transformations of the response variable in order to obtain independence between the errors and the covariates. In view of estimating the transformation parameter, we use a minimum distance approach and show the uniform consistency of the estimator under mild conditions. The boundary curve, i.e., the regression function, is estimated applying a smoothed version of a local constant approximation for which we also prove the uniform consistency. We deal with both cases of random covariates and deterministic (fixed) design points. To highlight the applicability of the procedures and to demonstrate their performance, the small sample behavior is investigated in a simulation study using the so-called Yeo–Johnson transformations.

Keywords Box–Cox transformations · Frontier estimation · Minimum distance estimation · Local constant approximation · Boundary models · Nonparametric regression · Yeo–Johnson transformations

Financial support by the DFG (Research Unit FOR 1735 *Structural Inference in Statistics: Adaptation and Efficiency*) is gratefully acknowledged.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10463-019-00731-5) contains supplementary material, which is available to authorized users.

✉ Charles Tillier
charles.tillier@gmail.com

Natalie Neumeyer
neumeyer@math.uni-hamburg.de

Leonie Selk
leonie.selk@math.uni-hamburg.de

[1] Department of Mathematics, University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany

1 Introduction

Before fitting a regression model, it is very common in applications to transform the response variable. The aim of the transformation is to gain efficiency in the statistical inference, for instance, by reducing skewness or inducing a specific structure of the model, e.g., linearity of the regression function or homoscedasticity. In practice, often a parametric class of transformations is considered from which an 'optimal' one should be selected data dependently (with a specific purpose in mind). A classical example is the class of Box–Cox power transformations introduced for linear models by Box and Cox (1964). There is a vast literature on parametric transformation models in the context of mean regression, and we refer to the monograph by Carroll and Ruppert (1988). Powell (1991) introduced Box–Cox transformations in the context of linear quantile regression; see also Mu and He (2007) who considered transformations to obtain a linear quantile regression function. Horowitz (2009) reviewed estimation in transformation models with parametric regression in the cases where either the transformation or the error distribution or both are modeled nonparametrically. Linton et al. (2008) suggested parametric estimators for transformations, while the error distribution is estimated nonparametrically and the regression function is additive. In this paper, the aim of the transformation is to induce independence between the covariables and the errors. Linton et al. (2008) considered profile likelihood and minimum distance estimation for the transformation parameter. The results for the profile likelihood estimator were generalized for nonparametric regression models by Colling and Van Keilegom (2016).

All literature cited above is about mean or quantile regression. In contrast in the paper at hand, we consider boundary regression models. Such nonparametric regression models with one-sided errors have been considered, among others, by Hall and Van Keilegom (2009), Meister and Reiß (2013), Jirak et al. (2013), Jirak et al. (2014) and Drees et al. (2019). Relatedly, estimation of support boundaries has been considered, for instance, by Härdle et al. (1995), Hall et al. (1998), Girard and Jacob (2008) and Daouia et al. (2016). Such models naturally appear when analyzing auctions or records or production frontiers. Unlike conditional mean models, regression models with one-sided errors (as well as quantile regression models) have the attractive feature of equivariance under monotone transformations. Thus in such a model with monotone transformation of the response, one can recover the original functional dependence in an easy manner. Similar to Linton et al. (2008), the aim of our transformation is to induce a model where the error distribution does not depend on the covariates. Independence of errors and covariates is a very typical assumption in regression models. For boundary models, this assumption is met, e.g., by Müller and Wefelmeyer (2010), Meister and Reiß (2013) and Drees et al. (2019). A transformation inducing (approximate) independence between the covariable and the error would allow for a global bandwidth selection in the adaptive regression estimator suggested by Jirak et al. (2014). Wilson (2003) pointed out that in production frontier models, independence assumptions are needed for validity of bootstrap procedures for nonparametric frontier models (see Simar and Wilson 1998) and suggested some tests for independence of errors and covariates (see also Drees et al. 2019).

While Linton et al. (2008) found advantages of the profile likelihood approach over minimum distance estimation of the transformation parameter in corresponding mean regression transformation models, this is at the cost of strong regularity conditions, among others a bounded error density with bounded derivative. In the context of boundary models with error distribution which is regularly varying at zero and irregular, one needs to avoid assumptions on bounded densities. Thus, we investigate a minimum distance approach to estimate the transformation parameter and give mild model assumptions under which the estimator is consistent.

We consider the cases of random covariates and deterministic (fixed) design points, which are both meaningful. The equidistant fixed design—as well as its natural generalization to deterministic covariates—is often used in real-life applications when time is involved in the data set. This is the case for instance in Jirak et al. (2014) where the authors studied the monthly sunspot observations and the annual best running times of 1500 m. Besides, deterministic design is met across a number of papers in regression models, see for instance Brown and Low (1996), Meister and Reiß (2013) and the references within. The case of random covariates is obviously the most relevant and appears in essence in many applications in boundary models, among other, in insurance and financial risk modeling when analyzing the optimality of portfolios (see Markowitz (1952) for the seminal contribution).

The remaining part of the manuscript is organized as follows. In Sect. 2, the model is explained, while in Sect. 3, the estimation procedure is described. In Sect. 4, we show consistency of the transformation parameter estimator. In Sect. 5, we present simulation results. The proofs for the random covariate case are given in "Appendix," while supplementary material contains proofs for the fixed design case and some additional figures and simulation results.

2 Model

2.1 The random design case

Consider independent and identically distributed observations $(X_i, Y_i), i = 1, \dots, n$, with the same distribution as (X, Y), where Y is univariate and X is distributed on $[0, 1]$. Further, consider a family $\mathcal{L} = \{\Lambda_\vartheta | \vartheta \in \Theta\}$ of strictly increasing and continuous transformations. Throughout the paper, we assume existence of a transformation Λ_{ϑ_0} in the class $\mathcal{L}$ such that in the corresponding boundary regression model

$$\Lambda_{\vartheta_0}(Y) = h_{\vartheta_0}(X) + \varepsilon, \tag{1}$$

the errors and the covariates are stochastically independent. Note that for notational simplicity, we set $\Lambda_0 = \Lambda_{\vartheta_0}$ and $h_0 = h_{\vartheta_0}$. Further denote by F_0 the cumulative distribution function (cdf) of the independent and identically distributed (iid) $\varepsilon_i = \Lambda_0(Y_i) - h_0(X_i)$, $i = 1, \dots, n$. Then, we assume that $F_0(0) = 1$ and $F_0(-\Delta) < 1$ for all $\Delta > 0$. This identifies the function h_0 as the upper boundary curve of the observations since

$$\mathbb{P}(\Lambda_0(Y_i) \leq h_0(X_i) \mid X_i = x) = 1 \text{ for all } x \in [0,1]$$
$$\mathbb{P}(\Lambda_0(Y_i) - h_0(X_i) \leq -\Delta \mid X_i = x) < 1 \text{ for all } x \in [0,1], \Delta > 0.$$

The aim is to estimate ϑ_0 from the observations.

Remark 1 Note that even if the model does not hold exactly (i.e., there does not exist any $\vartheta_0 \in \Theta$ that leads to exact independence of the errors and covariates), the transformation can be useful in applications because it will reduce the dependence.

For each $\vartheta \in \Theta$, one can consider the transformed responses $\Lambda_\vartheta(Y_i)$. Note that those form a boundary regression model with boundary curve $h_\vartheta = \Lambda_\vartheta \circ \Lambda_0^{-1} \circ h_0$, because

$$\mathbb{P}(\Lambda_\vartheta(Y_i) \leq h_\vartheta(X_i) \mid X_i = x) = \mathbb{P}(\Lambda_0(Y_i) \leq h_0(X_i) \mid X_i = x) = 1,$$

and for each $\delta > 0$,

$$\begin{aligned}&\mathbb{P}\left(\Lambda_\vartheta(Y_i) - h_\vartheta(X_i) \leq -\delta \mid X_i = x\right)\\ &\quad = \mathbb{P}\Big(\Lambda_0(Y_i) \leq \Lambda_0(\Lambda_\vartheta^{-1}(h_\vartheta(x) - \delta)) \mid X_i = x\Big) < 1\end{aligned}$$

since $\Delta = h_0(x) - \Lambda_0(\Lambda_\vartheta^{-1}(\Lambda_\vartheta(\Lambda_0^{-1}(h_0(x))) - \delta)) > 0$ since each Λ_ϑ is strictly increasing. The conditional distribution of $\Lambda_\vartheta(Y_i)$ for some general $\vartheta \in \Theta$ reads as

$$\begin{aligned}\mathbb{P}(\Lambda_\vartheta(Y_i) \leq y \mid X_i = x) &= \mathbb{P}\Big(\Lambda_0(Y_i) \leq \Lambda_0(\Lambda_\vartheta^{-1}(y)) \mid X_i = x\Big)\\ &= F_0\Big(\Lambda_0(\Lambda_\vartheta^{-1}(y)) - h_0(x)\Big).\end{aligned}$$

Remark 2 It is important to give conditions under which the unknown components $\Lambda_0 = \Lambda_{\vartheta_0}$, $h_0 = h_{\vartheta_0}$ and F_0 in model (1) are identifiable. To this end, we impose the following assumptions.

- Assume that Y has a continuous distribution, and w.l.o.g. assume that 0 is in the data range (Otherwise shift the data).
- Assume that X is continuously distributed with support $[0, 1]$.
- Assume $\Lambda_\vartheta(0) = 0$ for all $\vartheta \in \Theta$, and Λ_ϑ is strictly increasing and continuous for each $\vartheta \in \Theta$.
- Assume that if for some $\vartheta_0, \vartheta_1 \in \Theta$ one has

  $$(\Lambda_{\vartheta_1} \circ \Lambda_{\vartheta_0}^{-1})(a - b) = (\Lambda_{\vartheta_1} \circ \Lambda_{\vartheta_0}^{-1})(a) - (\Lambda_{\vartheta_1} \circ \Lambda_{\vartheta_0}^{-1})(b)$$

  for all $a, b \in J$, where J is an interval of positive length, then it follows that $\vartheta_0 = \vartheta_1$.
- Assume that F_0 (the cdf of $\varepsilon = \Lambda_{\vartheta_0}(Y) - h_{\vartheta_0}(X)$) is strictly increasing.
- Assume that h_{ϑ_0} is not constant and is continuous.

Now assume that the model

$$\Lambda_\vartheta(Y) = h_\vartheta(X) + \varepsilon(\vartheta) \text{ with } X \text{ independent from } \varepsilon(\vartheta)$$

holds for $\vartheta = \vartheta_0$ (with our notations $\Lambda_{\vartheta_0} = \Lambda_0$, $h_{\vartheta_0} = h_0$, $\varepsilon(\vartheta_0) = \varepsilon$) and for $\vartheta = \vartheta_1$.

Note that from the assumption, it follows that $\mathbb{P}(\varepsilon(\vartheta) \leq 0) = 1, \mathbb{P}(\varepsilon(\vartheta) \leq -\Delta) < 1$ for each $\Delta > 0$, such that h_ϑ is the upper boundary curve in the model (for $\vartheta \in \{\vartheta_0, \vartheta_1\}$).

We show in Sect. B of "Appendix" that it follows that $\vartheta_0 = \vartheta_1$. Thus, the transformation is identifiable. Further, $h_{\vartheta_0}(x)$ is then the right endpoint of the conditional distribution of $\Lambda_{\vartheta_0}(Y)$, given $X = x$, and F_0 is identified as cdf of $\Lambda_{\vartheta_0}(Y) - h_{\vartheta_0}(X)$.

If the function class $\mathcal{L}$ contains the identity, then the assumptions rule out that it contains transformations which are linear on some interval with positive length. On the other hand, it is clear that linear transformations can never reduce the dependence between the covariate and the error distribution.

Example 1 In this example, we give two classes of transformations that fulfill the identifiability assumptions.

Yeo and Johnson (2000) generalized the Box–Cox transformations by suggesting

$$\Lambda_\vartheta(y) = \begin{cases} \frac{(y+1)^\vartheta - 1}{\vartheta}, & \text{if } y \geq 0, \vartheta \neq 0 \\ \log(y+1), & \text{if } y \geq 0, \vartheta = 0 \\ -\frac{(-y+1)^{2-\vartheta} - 1}{2-\vartheta}, & \text{if } y < 0, \vartheta \neq 2 \\ -\log(-y+1), & \text{if } y < 0, \vartheta = 2, \end{cases}$$

which are typically considered for $\vartheta \in \Theta = [0, 2]$ because then they are bijective maps $\Lambda_\vartheta : \mathbb{R} \to \mathbb{R}$. Note that $\Lambda_\vartheta(0) = 0$ for all $\vartheta \in \Theta$.

The class of sinh–arcsinh transformations, see Jones and Pewsey (2009), do shift the location, but they can be modified to fulfill $\Lambda_\vartheta(0) = 0$ for all $\vartheta \in \Theta$, e.g., consider

$$\Lambda_{(\vartheta_1,\vartheta_2)}(y) = \sinh(\vartheta_1 \sinh^{-1}(y) - \vartheta_2) - \sinh(-\vartheta_2).$$

Here, $\vartheta_1 > 0$ is the tail weight parameter and $\vartheta_2 \in \mathbb{R}$ the skewness parameter. These transformations define also bijective maps $\Lambda_{(\vartheta_1,\vartheta_2)} : \mathbb{R} \to \mathbb{R}$.

2.2 The fixed design case

In the fixed design case, we consider a triangular array of independent observations $Y_{i,n}$, $i = 1, \ldots, n$, and deterministic design points $0 < x_{1,n} < \cdots < x_{n,n} < 1$. Once again, we assume existence of a transformation $\Lambda_0 = \Lambda_{\vartheta_0}$ in the class $\mathcal{L}$ such that setting $h_0 = h_{\vartheta_0}$ in the corresponding regression model

$$\Lambda_0(Y_{i,n}) = h_0(x_{i,n}) + \varepsilon_{i,n}, \tag{2}$$

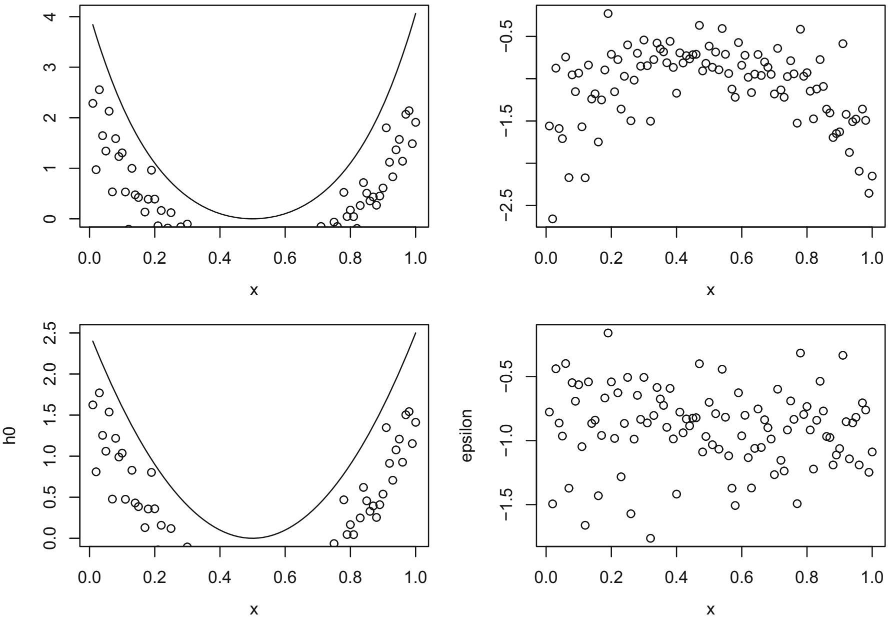

Fig. 1 Original data (upper panel) and transformed data (2) (lower panel) with $h_0(x) = 10(x - \frac{1}{2})^2$, $n = 100$ and $-\varepsilon_{i,n} \sim$ Weibull(1, 3) with Yeo and Johnson transformation $\Lambda_{0.5}$ as defined in Example 1. The design points are equidistant. The left panels show the data and regression functions, and the right panels show the errors

the cdf of the errors does not depend on the design points, i.e., $\varepsilon_{i,n} \sim F_0\ \forall i, n$. Note that as in the random design case, we assume $F_0(0) = 1$ and $F_0(-\Delta) < 1$ for all $\Delta > 0$ leading again to $h_\vartheta = \Lambda_\vartheta \circ \Lambda_0^{-1} \circ h_0$.

Remark 3 Identifiability can be shown under the same conditions as in Remark 2 as long as $\bar{\Delta}_n := \max_{1 \le i \le n+1} \left(x_{i,n} - x_{i-1,n} \right) \to 0$; see section D of the supplement.

Example 2 Figures 1 and 2 show realizations of the original data and the transformed data (2) using a Yeo and Johnson transformation; see Example 1. For each figure, in the upper left panel the original data $(x_{i,n}, Y_{i,n})$, $i = 1, \ldots, n = 100$, are depicted with their boundary curve $\Lambda_0^{-1} \circ h_0$, while the upper right panel shows the corresponding non-iid errors $Y_{i,n} - \Lambda_0^{-1} \circ h_0(x_{i,n})$. The lower left panel shows the transformed data with the curve h_0, while the lower right panel shows the iid errors $\varepsilon_{i,n}$, $i = 1, \ldots, n$.

3 Estimating the transformation

3.1 The random design case

If ϑ_0 were known, we could estimate the regression function (upper boundary curve) h_0 by a local constant approximation, i.e.,

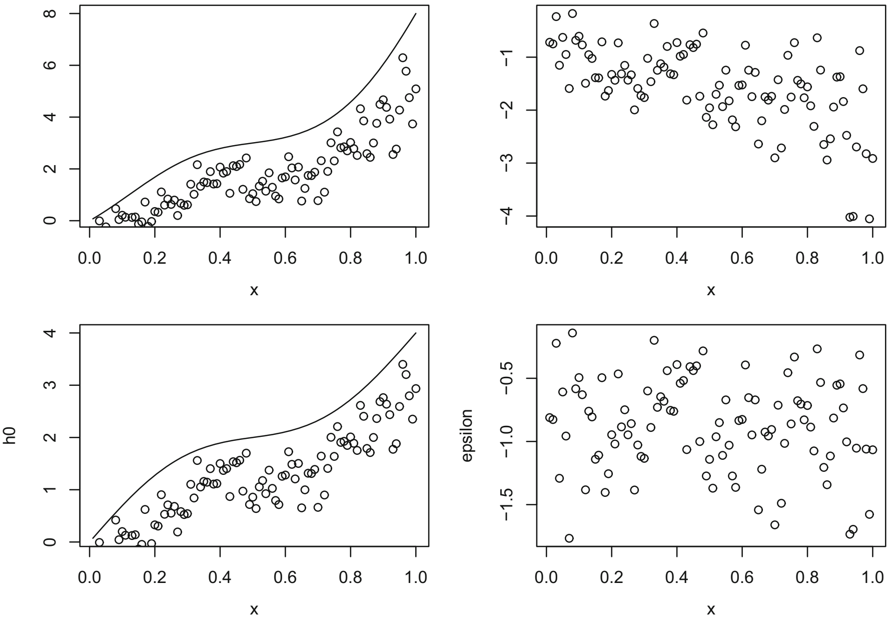

Fig. 2 Setting is similar to Fig. 1 but with $h_0(x) = \frac{1}{2}\sin(2\pi x) + 4x$

$$\tilde{h}_0(x) = \max\{\Lambda_0(Y_i)|i = 1, \ldots, n \text{ with } |X_i - x| \leq b_n\}, \tag{3}$$

where $b_n \searrow 0$ is a sequence of bandwidths. For this estimator, we will show uniform consistency under the following assumptions.

- **(A1)** Model (1) holds with iid $\varepsilon_1, \ldots, \varepsilon_n \sim F_0$ and $F_0(0) = 1$, $F_0(-\Delta) < 1$ for all $\Delta > 0$, and $\varepsilon_1, \ldots, \varepsilon_n$ are independent of $X_1, \ldots, X_n$.
- **(A2)** The covariates $X_1, \ldots, X_n$ are iid with cdf F_X and density f_X that is continuous and bounded away from zero on its support $[0, 1]$.
- **(A3)** The regression function h_0 is continuous on $[0, 1]$.
- **(A4)** Let $(b_n)_{n\in\mathbb{N}}$ be a sequence of positive bandwidths that satisfies $\lim_{n\to\infty} b_n = 0$ and $\lim_{n\to\infty}(\log n)/(nb_n) = 0$.

Note that we do not require any assumption on the error distribution. In particular, in the setup of regularly varying distributed errors, all the results hold for regular as well as irregular distributions. In what follows, let $\|\cdot\|_\infty$ denote the supremum norm and $I\{\cdot\}$ the indicator function.

Lemma 3 *Under model (1) with assumptions* **(A1)–(A4)***, we have* $\|\tilde{h}_0 – h_0\|_\infty = o_P(1)$.

The proof of the lemma is given in Sect. A.1 of "Appendix." The result applies for a model without transformation. Thus, as a by-product, we show uniform consistency of a boundary curve estimator in models with random covariates (and non-equidistant fixed design, see Lemma 5), while in contrast, Drees et al. (2019) assumed equidistant design and obtained rates of convergence under stronger assumptions on the error distribution F_0 and on the boundary curve h_0.

For general $\vartheta \in \Theta$, we define a simple boundary curve estimator accordingly as

$$\tilde{h}_\vartheta(x) = \max\{\Lambda_\vartheta(Y_i) | i = 1, \ldots, n \text{ with } |X_i - x| \leq b_n\},$$

and it holds that $\tilde{h}_\vartheta = \Lambda_\vartheta \circ \Lambda_0^{-1} \circ \tilde{h}_0$. Thus, $\tilde{h}_\vartheta$ consistently estimates h_ϑ. The local constant estimator can be improved by introducing slight smoothing. To this end, let K be a density with compact support and a_n some sequence of bandwidths that decreases to zero such that $na_n \to \infty$. Define

$$\hat{h}_\vartheta(x) = \frac{\sum_{i=1}^n \tilde{h}_\vartheta(X_i) K\left(\frac{x-X_i}{a_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{a_n}\right)}, \tag{4}$$

and then $\hat{h}_\vartheta$ is also uniformly consistent for h_ϑ; see Lemma 7.

Example 4 For data as in Example 2, Figures 5 and 6 in the online supplementary material demonstrate the smoothing of the estimator. We use the Epanechnikov kernel $K(x) = 0.75(1 - x^2)I_{[-1,1]}(x)$ and bandwidths $b_n = 0.5n^{-1/3}$, $a_n = 0.5b_n$ with $n = 100$.

Based on this estimator, we define the joint empirical distribution function of residuals and covariates as $\hat{F}_{n,\vartheta}(y, s) = \frac{1}{n}\sum_{i=1}^n I\{\Lambda_\vartheta(Y_i) - \hat{h}_\vartheta(X_i) \leq y\}I\{X_i \leq s\}$. For $\vartheta = \vartheta_0$, the covariate X_i and the error $\Lambda_\vartheta(Y_i) - h_\vartheta(X_i)$ are stochastically independent, and thus, the joint empirical distribution function minus the product of the marginals, namely $\hat{F}_{n,\vartheta}(y, s) - \hat{F}_{n,\vartheta}(y, 1)\hat{F}_{X,n}(s)$, estimates zero for $\vartheta = \vartheta_0$. Here, $\hat{F}_{X,n}(\cdot) = \hat{F}_{n,\vartheta}(\infty, \cdot)$ denotes the empirical distribution function of $X_1, \ldots, X_n$. We will use this idea to estimate the transformation parameter ϑ_0. To this end, for any function $h : [0, 1] \to \mathbb{R}$ define

$$G_n(\vartheta, h)(y, s) = \frac{1}{n}\sum_{i=1}^n I\{\Lambda_\vartheta(Y_i) - h(X_i) \leq y\}\big(I\{X_i \leq s\} - \hat{F}_{X,n}(s)\big), \tag{5}$$

and note that $G_n(\vartheta, \hat{h}_\vartheta)(y, s) = \hat{F}_{n,\vartheta}(y, s) - \hat{F}_{n,\vartheta}(y, 1)\hat{F}_{X,n}(s)$. Our criterion function will be

$$M_n(\vartheta) = \|G_n(\vartheta, \hat{h}_\vartheta)\|$$

for some semi-norm $\|\cdot\|$ as described in the following assumption.

(N1) $\|\cdot\|$ is a semi-norm such that $\|\Gamma\| \leq c \sup_{\substack{y \in C \\ s \in [0,1]}} |\Gamma(y, s)|$ for some constant $c > 0$ and some compact set $C = [c_1, c_2] \subset \mathbb{R}$ with $c_1, c_2 > 0$ and $0 \in C$, for all measurable functions $\Gamma : \mathbb{R} \times [0, 1] \to \mathbb{R}$.

For instance, one can consider one of the following semi-norms,

(i) $\|\Gamma(y,s)\| = \sup_{\substack{s\in[0,1]\\ y\in C}} |\Gamma(y,s)|$

(ii) $\|\Gamma(y,s)\| = \left(\int \Gamma(y,s)^2 w(y,s)\,\mathrm{d}(y,s)\right)^{1/2}$ for some integrable weight function $w : \mathbb{R}\times[0,1] \to \mathbb{R}_0^+$ with support included in $C\times[0,1]$

(iii) $\|\Gamma(y,s)\| = \sup_{s\in[0,1]} \left(\int \Gamma(y,s)^2 w(y)\,\mathrm{d}y\right)^{1/2}$ for some integrable weight function $w : \mathbb{R} \to \mathbb{R}_0^+$ with support included in C

(iv) $\|\Gamma(y,s)\| = \sup_{y\in C} \left(\int \Gamma(y,s)^2 w(s)\,\mathrm{d}s\right)^{1/2}$ for some integrable weight function $w : [0,1] \to \mathbb{R}_0^+$.

The first two semi-norms correspond to Kolmogorov–Smirnov and Cramér–von Mises distances, respectively, while the last two are mixtures of both.

Now we define the estimator $\hat{\vartheta}$ of ϑ_0 as the minimizer of $M_n(\vartheta)$ over Θ, i.e.,

$$\hat{\vartheta} = \arg\min_{\vartheta\in\Theta} M_n(\vartheta). \tag{6}$$

For the following theory also, the weaker condition $M_n(\hat{\vartheta}) \le \inf_{\theta\in\Theta} M_n(\vartheta) + o_P(1)$ is sufficient. Note that

$$\begin{aligned}\mathbb{P}\left(\Lambda_\vartheta(Y_i) - h(X_i) \le y \mid X_i = x\right) &= \mathbb{P}\left(\Lambda_0(Y_i) \le \Lambda_0(\Lambda_\vartheta^{-1}(y + h(x))) \mid X_i = x\right)\\ &= F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + h(x))) - h_0(x)\right),\end{aligned}$$

which reduces to $F_0(y)$ for $\vartheta = \vartheta_0$ and $h = h_0$. Now considering expectations, we define

$$\begin{aligned}G(\vartheta,h)(y,s) = &\int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + h(x))) - h_0(x)\right) I\{x \le s\} f_X(x)\,\mathrm{d}x\\ &- \int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + h(x))) - h_0(x)\right) f_X(x)\,\mathrm{d}x\, F_X(s)\end{aligned} \tag{7}$$

as deterministic counterpart of $G_n(\vartheta,h)$. Further, set

$$M(\vartheta) = \|G(\vartheta, h_\vartheta)\|,$$

and note that $M(\vartheta_0) = \|G(\vartheta_0, h_0)\| = 0$. In Sect. 4, we formulate assumptions under which $\hat{\vartheta}$ consistently estimates ϑ_0.

3.2 The fixed design case

In the fixed design model (2), we define the estimator for the boundary curve h_0 as

$$\tilde{h}_0(x) = \max\{\Lambda_0(Y_{i,n}) | i = 1, \ldots, n \text{ with } |x_{i,n} - x| \leq b_n\}$$

and obtain uniform consistency under the following modified assumptions. We set $x_{0,n} = 0$ and $x_{n+1,n} = 1$.

(A1') Model (2) holds with independent $\varepsilon_{1,n}, \ldots, \varepsilon_{n,n}$ with cdf F_0 $(\forall n)$ such that $F_0(0) = 1$, $F_0(-\Delta) < 1$ for all $\Delta > 0$.

(A2') The design points $0 < x_{1,n} < \cdots < x_{n,n} < 1$ are deterministic.

(A4') Let $(b_n)_{n\geq 0}$ be a sequence of positive bandwidths that satisfies $\lim_{n\to\infty} b_n = 0$ and $\lim_{n\to\infty} \bar{\Delta}_n \log(n)/b_n = 0$ for $\bar{\Delta}_n := \max_{1\leq i\leq n+1} \big(x_{i,n} - x_{i-1,n}\big)$.

Lemma 5 *Under model (2) with assumptions* **(A1')**, **(A2')**, **(A3)** *and* **(A4')**, *we have* $\|\tilde{h}_0 - h_0\|_\infty = o_P(1)$.

The proof is given in section C.1 of the online supplementary material. For general $\vartheta \in \Theta$, we define a consistent boundary curve estimator as

$$\tilde{h}_\vartheta(x) = \max\{\Lambda_\vartheta(Y_{i,n}) | i = 1, \ldots, n \text{ with } |x_{i,n} - x| \leq b_n\} = \Lambda_\vartheta(\Lambda_0^{-1}(\tilde{h}_0(x))).$$

In analogy to (5) we define, for any function $h : [0, 1] \to \mathbb{R}$,

$$G_n(\vartheta, h)(y, s) = \frac{1}{n} \sum_{i=1}^{n} I\{\Lambda_\vartheta(Y_{i,n}) - h(x_{i,n}) \leq y\}\big(I\{x_{i,n} \leq s\} - \hat{F}_{X,n}(s)\big), \quad (8)$$

where

$$\hat{F}_{X,n}(s) = \frac{1}{n} \sum_{i=1}^{n} I\{x_{i,n} \leq s\}.$$

The criterion function is again $M_n(\vartheta) = \|G_n(\vartheta, \hat{h}_\vartheta)\|$ where the smooth estimator $\hat{h}_\vartheta$ is defined accordingly as in (4), and with this the transformation parameter estimator is similar to (6). In order to consider the same deterministic G as in (7), an additional assumption is needed.

(A2'') The design points $0 < x_{1,n} < \cdots < x_{n,n} < 1$ are deterministic. There exists a cdf F_X with continuous density function $f_X : [0, 1] \to \mathbb{R}$ which is bounded away from zero such that

$$\max_{i=1,\ldots,n+1} \left| \int_{x_{i-1,n}}^{x_{i,n}} f_X(x)\,\mathrm{d}x - \frac{1}{n} \right| = o\left(\frac{1}{n}\right).$$

Assumption **(A2”)** is common in the literature on fixed design regression models. It allows the application of the mean value theorem for integrals to obtain, for some $\xi_{i,n} \in [x_{i-1,n}, x_{i,n}]$,

$$f_X(\xi_{i,n})(x_{i,n} - x_{i-1,n}) = \int_{x_{i-1,n}}^{x_{i,n}} f_X(x)\,\mathrm{d}x = \frac{1}{n} + o\left(\frac{1}{n}\right)$$

uniformly in $i = 1, \ldots, n$. Thus, it follows from **(A2”)** that $\bar{\Delta}_n$ in assumption **(A4’)** has the exact rate n^{-1}, and therefore assumption **(A4’)** reduces to **(A4)**. Further, the following Riemann sum approximations for bounded integrable functions φ can be applied to get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\varphi(x_{i,n}) &= \sum_{i=1}^{n}\varphi(x_{i,n})(f_X(\xi_{i,n})(x_{i,n} - x_{i-1,n}) + o(\tfrac{1}{n})) \\ &= \int \varphi(x) f_X(x)\,\mathrm{d}x + o(1). \qquad (9)\end{aligned}$$

In the next section, we state conditions under which $\hat{\vartheta} = \arg\min_{\vartheta\in\Theta} M_n(\vartheta)$ consistently estimates ϑ_0.

4 Main result

To prove consistency of the estimator for the transformation parameter, we need the following additional assumptions. Please note that assumption **(B1)** implies identifiability of the transformation Λ_0 in the class $\mathcal{L}$.

(B1) For every $\delta > 0$, there exists some $\epsilon > 0$ such that $\inf_{\|\vartheta-\vartheta_0\|>\delta} M(\vartheta) \geq \epsilon$.

(B2) $\mathcal{L} = \{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ is a class of strictly increasing continuous functions $\mathbb{R} \to \mathbb{R}$.

(B3) Let $S = \{\Lambda_0^{-1}(h_0(x)) \mid x \in [0, 1]\}$. Then, the class $\mathcal{L}_S = \{\Lambda_\vartheta|_S \mid \vartheta \in \Theta\}$ is pointwise bounded and uniformly equicontinuous, i.e., $\sup_{\vartheta\in\Theta} |\Lambda_\vartheta(y)| < \infty$ for all $y \in S$, and for every $\epsilon > 0$, there exists some $\delta > 0$ such that $\sup_{\vartheta\in\Theta} |\Lambda_\vartheta(y) - \Lambda_\vartheta(z)| < \epsilon$ for all $y, z \in S$ with $|y - z| \leq \delta$.

(B4) The class $\mathcal{L}^1_{\tilde{S}} = \{\Lambda_0 \circ \Lambda_\vartheta^{-1}|_{\tilde{S}} \mid \vartheta \in \Theta\}$ is pointwise bounded and uniformly equicontinuous for $\tilde{S} = \{z + h_\vartheta(x) \mid z \in C_\tau, \vartheta \in \Theta, x \in [0, 1]\}$ with $C_\tau = [c_1 - \tau, c_2 + \tau]$ (for $C = [c_1, c_2]$ from **(N1)**) for some $\tau > 0$, i.e., $\sup_{\vartheta\in\Theta} |\Lambda_0(\Lambda_\vartheta^{-1}(z))| < \infty$ for all $z \in \tilde{S}$, and for every $\delta > 0$, there exists some $\gamma > 0$ such that $\sup_{\vartheta\in\Theta} |\Lambda_0(\Lambda_\vartheta^{-1}(x)) - \Lambda_0(\Lambda_\vartheta^{-1}(z))| \leq \delta$ for all $x, z \in \tilde{S}$ with $|x - z| \leq \gamma$.

(B5) For some $\tau > 0$, F_0 is uniformly continuous on the set $\tilde{C} = \{\Lambda_0(\Lambda_\vartheta^{-1}(y + a + h_\vartheta(x))) - h_0(x) \mid y \in C, \vartheta \in \Theta, x \in [0, 1], |a| \leq \tau\}$ (with C from **(N1)**), i.e., for every $\epsilon > 0$, there is some $\delta > 0$ such that $|F_0(y) - F_0(z)| < \epsilon$ if $|y - z| \leq \delta$, $y, z \in \tilde{C}$.

(B6) K is a density with support $[-1, 1]$ and $b_n \searrow 0$, $nb_n \to \infty$.

Let us now make few comments regarding these assumptions.

- **(B1)** is a common assumption in M-estimation and needed for uniqueness of the true parameter.
- **(B2)** implies the existence of continuous inverse functions Λ_ϑ^{-1}. Further, note that uniform equicontinuity and pointwise boundedness imply totally boundedness by the Arzelà–Ascoli theorem. Thus for each ϵ, there is a finite covering of the classes $\mathcal{L}_S$ from **(B3)** and $\mathcal{L}^1_{\tilde{S}}$ from **(B4)** with balls of radius ϵ with respect to the sup norm. Thus, also the sup-norm bracketing numbers of those classes are finite, i.e.,
$$N_{[\,]}(\epsilon, \mathcal{L}_S, \|\cdot\|_\infty) < \infty, \quad N_{[\,]}(\epsilon, \mathcal{L}^1_{\tilde{S}}, \|\cdot\|_\infty) < \infty \text{ for all } \epsilon > 0 \tag{10}$$
(see, e.g., Lemma 9.21 in Kosorok 2008).
- **(B3)**–**(B5)** can be seen as minimal assumptions on the class $\mathcal{L} = \{\Lambda_\vartheta \mid \vartheta \in \Theta\}$ and F_0. As typically the sets S, $\tilde{S}$ and $\tilde{C}$ are unknown, the assumptions can be replaced by stronger assumptions that hold on all compact sets. Besides, working on compact set transformation parameter, assumptions **(B3)**–**(B4)** hold for most of transformations used in practice such as the Box and Cox transformations (see Box and Cox 1964) (suitably modified taking into account the data range), the exponential transformations (see Manly 1976) and the sinh–arcsinh transformations (see Jones and Pewsey 2009). For instance, with regard to Yeo–Johnson transformations, when $\vartheta \in \Theta = [0, 2]$, $\Lambda_\vartheta : \mathbb{R} \to \mathbb{R}$ defines a bijective map (see Remark 1) and both Λ_ϑ and Λ_ϑ^{-1} have uniform bounded derivatives on compact sets so that one may show that they fulfill assumptions **(B3)**–**(B4)** using the mean value theorem to Λ_ϑ and Λ_ϑ^{-1}. Further under stronger assumptions on the smoothness of F_0, Λ_ϑ and $\Lambda_0 \circ \Lambda_\vartheta^{-1}$, the theoretical results can be generalized to semi-norms that are not restricted to a compact $C \times [0, 1]$ as in assumption **(N1)**.
- **(B6)** is standard in kernel smoothing and is needed for the smoothed estimator $\hat{h}_\vartheta$ to be consistent. While we noticed in the simulations that slight smoothing improves the procedure, the following theorem still holds when $\hat{h}_\vartheta$ is replaced by the non-smooth estimator $\tilde{h}_\vartheta$. Assumption **(B5)** holds, e.g., for Hölder continuous distribution functions F_0.

The following theorem states consistency of the transformation parameter estimator.

Theorem 6 (i). *(The random design case.) Assume model (1) under assumptions* **(A1)**–**(A4)**, **(N1)**, **(B1)**–**(B6)**. *Then,* $\hat{\vartheta}$ *is a consistent estimator, i.e.,* $\hat{\vartheta} - \vartheta_0 = o_P(1)$.

(ii). *(The fixed design case.) Assume model (2) under assumptions* **(A1')**, **(A2")**, **(A3)**, **(A4)**, **(N1)**, **(B1)**–**(B6)**. *Then,* $\hat{\vartheta}$ *is a consistent estimator, i.e.,* $\hat{\vartheta} - \vartheta_0 = o_P(1)$.

The proof for the random design case is given in Sect. A.3 of "Appendix" and the proof for the fixed design case in section C.2 of the supplement. One basic ingredient is the following result, which is proven in Sect. A.2 of "Appendix" for the random design case. The proof for the fixed design case is analogous.

Lemma 7 (i). *(The random design case.) Under model (1) with assumptions* **(A1)**–**(A4)**, **(B2)**, **(B3)**, **(B6)**, *we have* $\sup_{\vartheta\in\Theta}\|\hat{h}_\vartheta - h_\vartheta\|_\infty = o_P(1)$.

(ii). *(The fixed design case.) Under model (2) with assumptions* **(A1')**, **(A2')**, **(A3)**, **(A4')**, **(B2)**, **(B3)**, **(B6)**, *we have* $\sup_{\vartheta\in\Theta}\|\hat{h}_\vartheta - h_\vartheta\|_\infty = o_P(1)$.

The consistency result in Theorem 6 should be seen as a first step in the analysis of transformation boundary regression models. An interesting and challenging topic for future research is to derive an asymptotic distribution of $\hat{\vartheta} - \vartheta_0$ (properly scaled) and to investigate the asymptotic influence of the estimation on subsequent procedures based on the transformed data. This is beyond the scope of the paper as yet there are no results on the uniform asymptotic distribution of $\tilde{h}_0 - h_0$ in the literature.

We finally highlight that under the further condition **(A3')** defined below regarding the regularity of the boundary curve, we obtain as a corollary of Theorem 6 the consistency of the estimator $\hat{h}_{\hat{\vartheta}}$ of the boundary curve.

(A3') ϑ_0 is an inner point of a convex parameter space Θ, and h_ϑ is continuously differentiable with respect to ϑ. Besides, we assume that there exists some $\delta > 0$ such that

$$\sup_{x\in[0,1]}\ \sup_{\|\vartheta-\vartheta_0\|<\delta}\left\|\frac{\partial h_\vartheta(x)}{\partial\vartheta}\right\| < \infty.$$

Corollary 8 (i). *(The random design case.) Assume model (1) holds under assumptions* **(A1)**, **(A2)**, **(A3')**, **(A4)**, **(N1)** *and* **(B1)**–**(B6)**. *Then,* $\hat{h}_{\hat{\vartheta}}$ *is a consistent estimator of* h_{ϑ_0}, *i.e.,* $\|\hat{h}_{\hat{\vartheta}} - h_{\vartheta_0}\|_\infty = o_P(1)$.

(ii). *(The fixed design case.) Assume model (2) holds under assumptions* **(A1')**, **(A2")**, **(A3')**, **(A4)**, **(N1)** *and* **(B1)**–**(B6)**. *Then,* $\hat{h}_{\hat{\vartheta}}$ *is a consistent estimator of* h_{ϑ_0}, *i.e.,* $\|\hat{h}_{\hat{\vartheta}} - h_{\vartheta_0}\|_\infty = o_P(1)$.

Proof We only prove **(i)** since the proof of **(ii)** is identical. Observe first that

$$\|\hat{h}_{\hat{\vartheta}} - h_{\vartheta_0}\|_\infty \le \|\hat{h}_{\hat{\vartheta}} - h_{\hat{\vartheta}}\|_\infty + \|h_{\hat{\vartheta}} - h_{\vartheta_0}\|_\infty.$$

The first term in the right-hand side of the above inequality goes to 0 in probability from Lemma 7 since the consistency holds uniformly over $\vartheta \in \Theta$. Regarding the second term, applying the mean value theorem, there exists some $\vartheta^*(x)$ on the line between $\hat{\vartheta}$ and ϑ_0 such that

$$\|h_{\hat{\vartheta}} - h_{\vartheta_0}\|_\infty = \sup_{x\in[0,1]}\left|\frac{\partial h_\vartheta(x)^T}{\partial\vartheta}|_{\vartheta=\vartheta^*(x)}(\hat{\vartheta} - \vartheta_0)\right|.$$

From Theorem 6, $\hat{\vartheta} - \vartheta_0 = o_P(1)$ which concludes the proof under assumption **(A3')**. □

5 Simulations

To study the small sample behavior, we generate data as $Y = \Lambda_{\vartheta_0}^{-1}(h_0(x) + \varepsilon)$ using the Yeo–Johnson transformation for different values of ϑ_0. We focus on the equidistant design framework and examine the two regression functions $h_0(x) = 10(x - \frac{1}{2})^2$ and $h_0(x) = \frac{1}{2}\sin(2\pi x) + 4x$ for two different error distributions, namely the Weibull distribution with scale parameter 1 and shape parameter 3 and the exponential distribution with mean $1/3$. We consider samples of size $n = 50$ and $n = 100$. It means that we investigate the following four models:

$$h_0(x) = 10\left(x - \tfrac{1}{2}\right)^2 \quad \text{with} \quad -\varepsilon \sim \text{Weibull}(1, 3) \tag{11}$$

$$h_0(x) = 10\left(x - \tfrac{1}{2}\right)^2 \quad \text{with} \quad -\varepsilon \sim \text{Exp}(3) \tag{12}$$

$$h_0(x) = \tfrac{1}{2}\sin(2\pi x) + 4x \quad \text{with} \quad -\varepsilon \sim \text{Weibull}(1, 3) \tag{13}$$

$$h_0(x) = \tfrac{1}{2}\sin(2\pi x) + 4x \quad \text{with} \quad -\varepsilon \sim \text{Exp}(3). \tag{14}$$

Figures 1 and 2 show realizations of models (11) and (13). The bandwidth $b_n = n^{-1/3}$ is chosen according to Drees et al. (2019), and simulations are based on 1000 iterations. We use the Epanechnikov kernel to smooth the boundary curve estimator and compare the results for two smoothing parameters $a_n = b_n/2$ and $a_n = b_n/20$. The transformation parameter estimator is as in (6) on the interval $[-0.5, 2.5]$, where the semi-norm in the criterion function $M_n(\vartheta)$ is chosen as in (i), (ii), (iii) and (iv) in the examples of condition **(N1)**. In the following, we denote the according estimators as TKS, TCM, TKSCM and TCMKS. Here, TKS and TCM refer to Kolmogorov–Smirnov and Cramér–von Mises distances, respectively, while TKSCM and TCMKS are mixtures of both. For simplicity, the weight functions are chosen identically equal to 1 in all the settings, i.e., $w(y, s) = 1$ for all $(y, s) \in \mathbb{R} \times [0, 1]$, $w(y) = 1$ for all $y \in \mathbb{R}$ and $w(s) = 1$ for all $s \in [0, 1]$ in (ii), (iii) and (iv), respectively (Although for the theory, we assumed a compact support).

We sum up the simulation results in the following eight tables. Tables 1, 3 and Tables 2, 4 deal with models (11) and (14) for $a_n = b_n/2$ and $a_n = b_n/20$, respectively, whereas Tables 6, 8 and Tables 7, 9 in the supplement show the results for models (12) and (13). In Fig. 3, we have represented the density function of each estimator for model (11) when $\vartheta_0 = 0.5$ with $n = 100$ and $a_n = b_n/20$, which corresponds to the settings of Table 2. To assess the performance of our estimates, we provide for each estimator the mean, the median and the mean integrated squared error (MISE) in brackets for five values of the true parameter $\vartheta_0 = 0, 0.5, 1, 1.5, 2$. The best-performing one regarding the mean (respectively the MISE) is highlighted in bold (respectively underlined).

Looking at the MISE, it turns out that the estimator using the Cramér–von Mises distance (TCM) out-performs in many cases even when it does not out-perform the mean; see Tables 1 and 2 when $n = 100$ for instance. Besides, as it is intended, results are better in most of the cases when the sample size n increases. However, this does not hold for every case. For instance, one may see in Table 2 that for the second estimator TCM, most of the results are better for $n = 50$ than for $n = 100$. This might relate

Table 1 Mean, median and MISE for model (11) for $n = 50$ and $n = 100$ with $a_n = b_n/2$

$n = 50$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.162 0.162 (0.120)	**0.095** 0.122 (0.060)	0.216 0.241 (0.149)	0.196 0.183 (0.141)
$\vartheta_0 = 0.5$	0.643 0.646 (0.142)	**0.576** 0.595 (0.080)	0.766 0.800 (0.250)	0.691 0.646 (0.204)
$\vartheta_0 = 1$	1.120 1.190 (0.232)	**1.100** 1.090 (0.121)	1.340 1.380 (0.335)	1.200 1.310 (0.287)
$\vartheta_0 = 1.5$	**1.610** 1.720 (0.228)	1.640 1.670 (0.133)	1.900 1.920 (0.336)	1.620 1.780 (0.282)
$\vartheta_0 = 2$	1.810 2.020 (0.357)	**2.110** 2.130 (0.076)	2.310 2.460 (0.179)	1.860 2.060 (0.297)
$n = 100$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.014 −0.039 (0.055)	−0.014 −0.029 (0.019)	0.098 0.092 (0.031)	**-0.006** −0.021 (0.026)
$\vartheta_0 = 0.5$	0.483 0.496 (0.041)	0.461 0.451 (0.026)	0.625 0.618 (0.049)	**0.503** 0.523 (0.047)
$\vartheta_0 = 1$	0.951 0.964 (0.062)	0.949 0.950 (0.034)	1.150 1.140 (0.059)	**1.000** 1.010 (0.071)
$\vartheta_0 = 1.5$	**1.500** 1.490 (0.050)	1.470 1.450 (0.040)	1.670 1.660 (0.078)	1.520 1.520 (0.077)
$\vartheta_0 = 2$	1.960 2.000 (0.071)	**1.970** 1.960 (0.034)	2.150 2.150 (0.065)	1.970 2.030 (0.074)

Table 2 Mean, median and MISE for model (11) for $n = 50$ and $n = 100$ with $a_n = b_n/20$

$n = 50$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.154 0.146 (0.126)	−**0.059** −0.051 (0.050)	0.198 0.205 (0.114)	0.173 0.171 (0.128)
$\vartheta_0 = 0.5$	0.613 0.635 (0.118)	**0.511** 0.513 (0.066)	0.717 0.735 (0.173)	0.658 0.646 (0.163)
$\vartheta_0 = 1$	1.120 1.190 (0.182)	**1.030** 1.030 (0.091)	1.280 1.300 (0.230)	1.190 1.270 (0.221)
$\vartheta_0 = 1.5$	1.630 1.690 (0.164)	**1.560** 1.540 (0.095)	1.830 1.810 (0.233)	1.620 1.740 (0.209)
$\vartheta_0 = 2$	1.880 2.040 (0.272)	**2.050** 2.080 (0.066)	2.280 2.370 (0.149)	1.870 2.040 (0.272)
$n = 100$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	**0.003** 0.046 (0.040)	0.045 0.057 (0.020)	0.078 0.067 (0.026)	0.005 0.022 (0.025)
$\vartheta_0 = 0.5$	0.454 0.448 (0.037)	0.418 0.406 (0.028)	0.581 0.566 (0.033)	**0.473** 0.478 (0.046)
$\vartheta_0 = 1$	0.938 0.950 (0.053)	0.913 0.918 (0.038)	1.100 1.080 (0.047)	**0.984** 0.986 (0.070)
$\vartheta_0 = 1.5$	1.450 1.420 (0.052)	1.410 1.390 (0.041)	1.600 1.590 (0.050)	**1.470** 1.440 (0.062)
$\vartheta_0 = 2$	**1.950** 1.960 (0.057)	1.930 1.910 (0.037)	2.100 2.100 (0.053)	1.940 1.980 (0.066)

to a sensitivity with respect to the choices of bandwidth and smoothing parameter. A lot of criteria may be used to judge the performance of the estimators. We deal here with the mean, the median and the MISE but we emphasize that using different criteria (e.g., median absolute deviation, mode or even graphical analysis) could give different results concerning the comparison of the methods. For instance, results in Table 3 for $n = 100$, $a_n = b_n/2$ and $\vartheta_0 = 1$ are quite not accurate regarding the mean (e.g., 0.845 for the TCM). Nevertheless, looking at Fig. 4, it appears that the plots of the densities look satisfactory.

It is clear that the TCM and the TKSCM out-perform in model (13) and in model (14), respectively; see for instance Table 4. Nonetheless, in a general setting, we are not able to state which estimator performs better since it depends first on the cri-

Table 3 Mean, median and MISE for model (14) for $n = 50$ and $n = 100$ with $a_n = b_n/2$

$n = 50$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.054 0.062 (0.218)	−0.029 −0.029 (0.005)	**0.000** 0.000 (0.000)	0.050 0.046 (0.182)
$\vartheta_0 = 0.5$	0.419 0.427 (0.038)	0.412 0.408 (0.027)	**0.469** 0.478 (0.025)	0.458 0.470 (0.030)
$\vartheta_0 = 1$	0.887 0.881 (0.080)	0.791 0.770 (0.104)	**0.942** 0.971 (0.077)	0.934 0.933 (0.069)
$\vartheta_0 = 1.5$	1.360 1.350 (0.130)	1.210 1.270 (0.214)	1.310 1.420 (0.319)	**1.410** 1.350 (0.111)
$\vartheta_0 = 2$	1.800 1.790 (0.165)	1.560 1.690 (0.504)	1.249 1.769 (1.739)	**1.810** 1.790 (0.141)
$n = 100$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.025 0.043 (0.004)	−0.037 −0.038 (0.003)	**0.012** 0.009 (0.002)	−0.034 −0.054 (0.004)
$\vartheta_0 = 0.5$	0.459 0.458 (0.017)	0.406 0.405 (0.016)	**0.520** 0.518 (0.008)	0.449 0.453 (0.017)
$\vartheta_0 = 1$	0.922 0.945 (0.038)	0.845 0.856 (0.046)	**1.030** 1.040 (0.017)	0.907 0.909 (0.044)
$\vartheta_0 = 1.5$	1.430 1.350 (0.056)	1.290 1.310 (0.080)	**1.540** 1.540 (0.032)	1.400 1.350 (0.063)
$\vartheta_0 = 2$	**1.930** 1.970 (0.072)	1.740 1.770 (0.126)	1.880 2.010 (0.419)	1.850 1.790 (0.090)

Table 4 Mean, median and MISE for model (14) for $n = 50$ and $n = 100$ with $a_n = b_n/20$

$n = 50$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.021 0.062 (0.177)	0.053 0.057 (0.007)	**0.013** 0.010 (0.007)	0.017 0.062 (0.141)
$\vartheta_0 = 0.5$	0.404 0.415 (0.044)	0.375 0.380 (0.034)	**0.449** 0.461 (0.029)	0.434 0.445 (0.034)
$\vartheta_0 = 1$	0.851 0.830 (0.086)	0.737 0.695 (0.119)	**0.915** 0.931 (0.074)	0.895 0.891 (0.076)
$\vartheta_0 = 1.5$	1.330 1.350 (0.150)	1.130 1.200 (0.260)	1.310 1.410 (0.293)	**1.380** 1.350 (0.119)
$\vartheta_0 = 2$	1.760 1.790 (0.176)	1.520 1.610 (0.435)	1.409 1.770 (1.250)	**1.770** 1.790 (0.156)
$n = 100$	TKS	TCM	TKSCM	TCMKS
$\vartheta_0 = 0$	0.027 0.043 (0.004)	−0.038 −0.040 (0.003)	**0.011** 0.008 (0.002)	0.037 0.062 (0.004)
$\vartheta_0 = 0.5$	0.462 0.466 (0.018)	0.411 0.407 (0.015)	**0.524** 0.522 (0.008)	0.453 0.460 (0.019)
$\vartheta_0 = 1$	0.930 0.950 (0.037)	0.849 0.856 (0.043)	**1.040** 1.030 (0.018)	0.917 0.915 (0.042)
$\vartheta_0 = 1.5$	1.440 1.390 (0.058)	1.300 1.320 (0.078)	**1.550** 1.540 (0.031)	1.420 1.350 (0.065)
$\vartheta_0 = 2$	**1.930** 1.960 (0.067)	1.730 1.760 (0.130)	1.830 2.010 (0.509)	1.840 1.790 (0.091)

teria selected to judge the performance but more importantly on the choice of the bandwidths and the smoothing parameter.

Finally, we recall that the aim of this work is to reduce the dependence between the covariates and the errors. As one can see in Table 5, although the estimation of ϑ_0 is less good than expected in model (14), the correlations between the covariates and the errors (after transformation) are very small; see also Table 10 in the supplement for the correlations in model (13). We obtain similar results for the random design case.

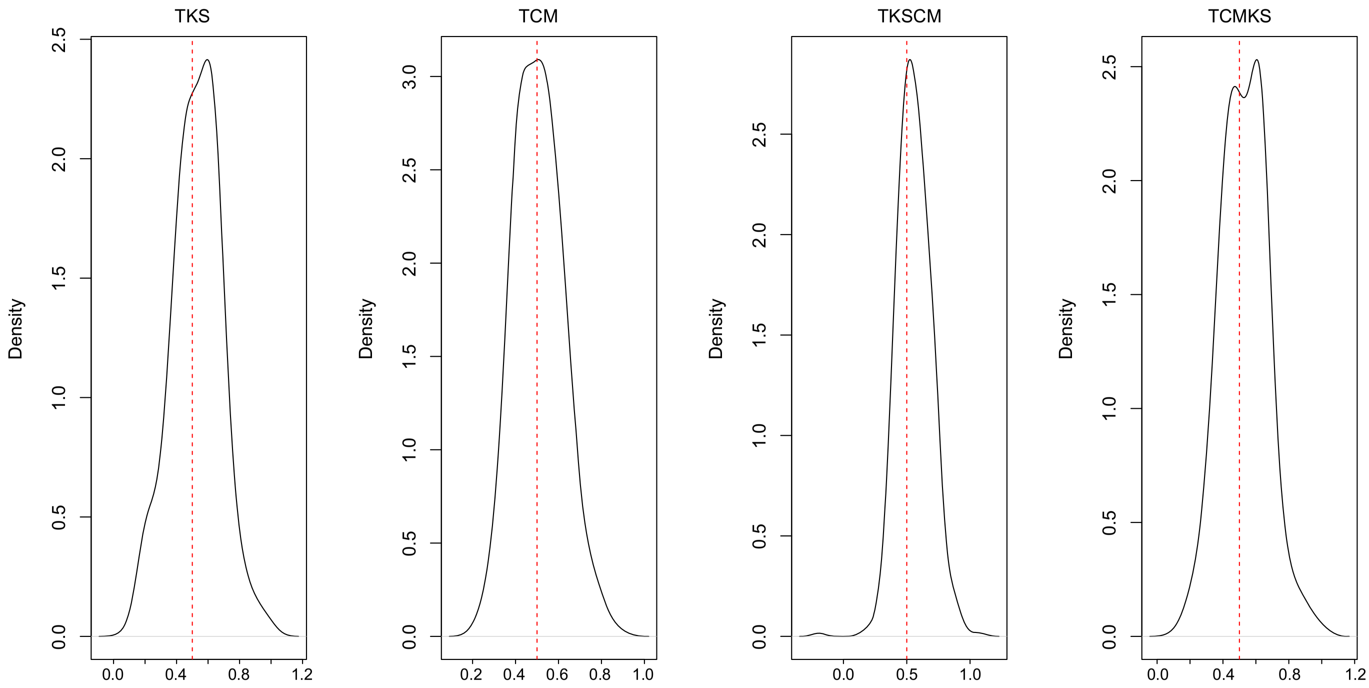

Fig. 3 Density function of the four estimators TKS, TCM, TCMKS and TKSCM for model (11) with a sample size $n = 100$ and bandwidths $a_n = b_n/20$ with $b_n = n^{-1/3}$. This corresponds to results in Table 2. The vertical dashed line corresponds to the true parameter $\vartheta_0 = 0.5$

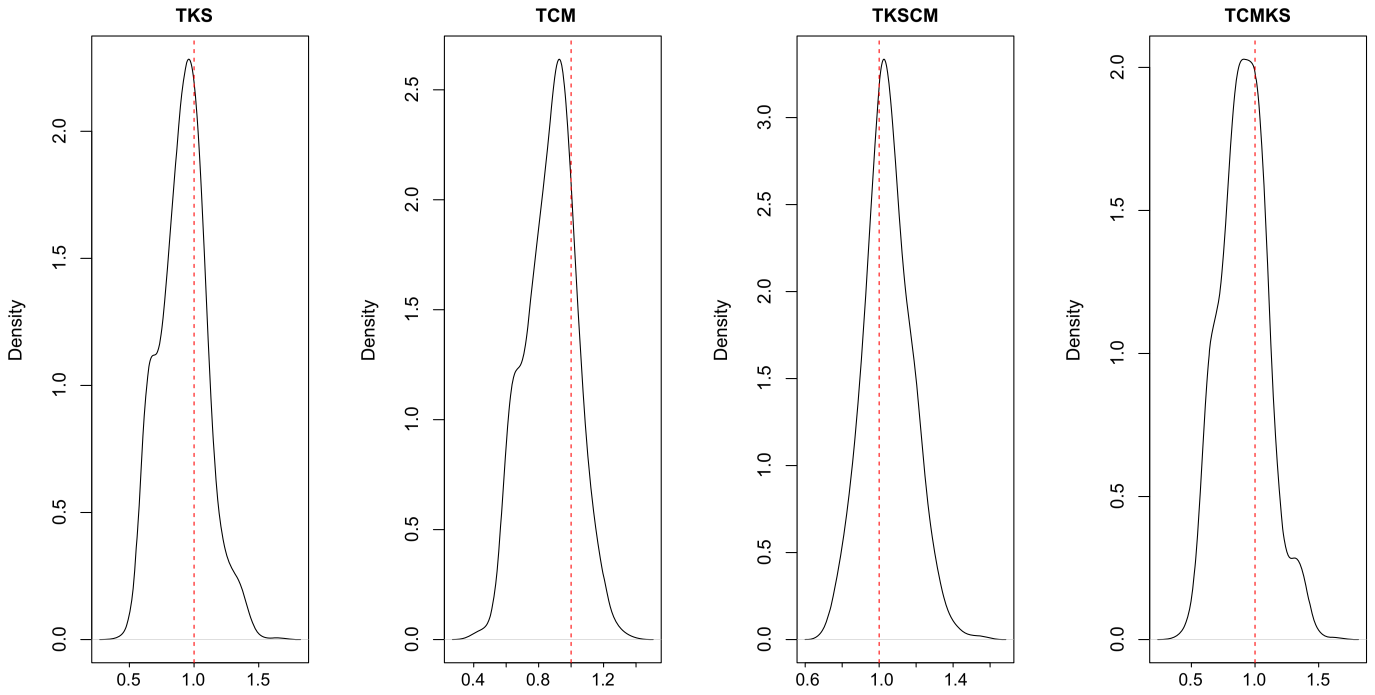

Fig. 4 Density function of the four estimators TKS, TCM, TCMKS and TKSCM for model (14) with a sample size $n = 100$ and bandwidths $a_n = b_n/2$ with $b_n = n^{-1/3}$. This corresponds to results in Table 3. The vertical dashed line corresponds to the true parameter $\vartheta_0 = 1$

Table 5 Pearson's, Kendall's and Spearman's correlation coefficients (the average over 1000 iterations) between the covariates and the errors for model (14) when $n = 100$

Method	Pearson	Kendall	Spearman
Original data	− 0.273	− 0.165	− 0.234
True parameter ϑ_0	0.005	0.003	0.004
TKS	0.008	0.004	0.007
TCM	0.024	0.014	0.021
TKSCM	0.011	0.007	0.009
TCMKS	0.003	0.001	0.002

The first line corresponds to the correlations for the original data while the second line is for the true transformation parameter ($\vartheta_0 = 0.5$). The last four lines correspond to the correlations for each estimator

A Proofs of asymptotic results in the random covariate case

For the proofs of the asymptotic results, let us fix some notation: $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ are the floor and ceiling functions, respectively; $\bar{F} = 1 - F$ denotes the survival function associated to a cdf F; $X_1 \overset{d}{=} X_2$ means that two random variables X_1, X_2 share the same distribution; $a_n \underset{n\to\infty}{\sim} b_n$ holds if there exists a constant $c > 0$ such that $\lim_{n\to\infty} a_n/b_n = c$ for two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ of nonnegative numbers; A^c is the complement of a set A.

In the following, we give the proofs of our results in the random design case, whereas the proofs for the fixed design case can be found in the online supplementary material.

A.1 Proof of Lemma 3

At first, we need the following intermediary lemma.

Lemma 9 *Assume model (1) holds with assumptions* **(A1)**, **(A2)** *and* **(A4)**. *Then, we have*

$$\sup_{x\in[0,1]} \min_{\substack{i\in\{1,\ldots,n\} \\ |X_i - x|\leq b_n}} |\varepsilon_i| = o_P(1).$$

Proof For $n \geq 1$ denote $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ the order statistics of the random design sample $X_1, X_2, \ldots, X_n$. Let π be the random permutation of $\{1, \ldots, n\}$ such that $X_{(i)} = X_{\pi(i)}, i = 1, \ldots, n$. Due to the independence between the errors and the covariates under **(A1)**, $\varepsilon_{\pi(1)}, \ldots, \varepsilon_{\pi(n)}$ are iid with cdf F_0. Let $Z_i = -\varepsilon_{\pi(i)}$, $i = 1, \ldots, n$, then $Z_1, \ldots, Z_n$ are iid with cdf U with $U(x) = 1 - F_0(-x)$, and we need to show that

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{x\in[0,1]} \min_{\substack{i\in\{1,\ldots,n\} \\ |X_{(i)} - x|\leq b_n}} Z_i > \epsilon \right) = 0, \quad \epsilon > 0. \tag{15}$$

Define for $n \geq 1$ the event

$$\Omega_n = \left\{ \inf_{x \in [0,1]} \sum_{i=1}^{n} I\{|X_i - x| \leq b_n\} \geq Cnb_n \right\}$$

for a suitable constant $C > 0$ specified later. Note that on Ω_n, there are at least Cnb_n covariates in each of the intervals $[x - b_n, x + b_n]$. We will first show that $\lim_{n\to\infty} \mathbb{P}(\Omega_n) = 1$. To this end, for $n \geq 1$ let $f_{n,x}(z) = I\{|x - z| \leq b_n\}$, and note that

$$\begin{aligned} &\inf_{x \in [0,1]} \frac{1}{n} \sum_{i=1}^{n} I\{|X_i - x| \leq b_n\} \\ &\quad \geq \inf_{x \in [0,1]} \mathbb{P}(|X_1 - x| \leq b_n) \\ &\qquad - \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} (f_{n,x}(X_i) - \mathbb{E}[f_{n,x}(X_i)]) \right|. \end{aligned} \tag{16}$$

Applying the mean value theorem of integration, it follows that

$$\begin{aligned} 2b_n \sup_{x \in [0,1]} f_X(x) &\geq \mathbb{P}(|X_1 - x| \leq b_n) \\ &= \int_{\max(0,x-b_n)}^{\min(1,x+b_n)} f_X(x)\,\mathrm{d}x \geq b_n \inf_{x \in [0,1]} f_X(x). \end{aligned} \tag{17}$$

Then, there exists a constant $C_1 > 0$, which actually corresponds to the lower bound of the density function f_X involved in assumption **(A2)** such that

$$\mathbb{P}(|X_1 - x| \leq b_n) \geq C_1 b_n, \tag{18}$$

uniformly over $x \in [0, 1]$.

Fix $n \geq 1$ and denote $P_n f_{n,x} := \frac{1}{n} \sum_{i=1}^{n} f_{n,x}(X_i)$ and $P f_{n,x} := \mathbb{E}[f_{n,x}(X_1)]$ so that P_n and P refer to the empirical measure and the distribution of the random design sample $X_1, \ldots, X_n$, respectively. By (17), $P f_{n,x}^2 = \mathbb{E}[I\{|X - x| \leq b_n\}] \leq 2C_2 b_n$, where $C_2 := \sup_{x\in[0,1]} f_X(x)$, which is finite under **(A2)**. Moreover, since $|f_{n,x}(X)| \leq 1$ and the assumption on the covering number is fulfilled (see Example 38 and Problem 28 to be convinced in Pollard 1984), Theorem 37 in Pollard (1984, p. 34) holds, and we have

$$\sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^{n} (f_{n,x}(X_i) - \mathbb{E}[f_{n,x}(X_i)]) \right| = o(b_n).$$

From this together with (16) and (18), it follows that $\lim_{n\to\infty} \mathbb{P}(\Omega_n) = 1$. It means that for any sub-interval $I_n := [x - b_n, x + b_n]$, there are at least Cnb_n random design points with probability converging to 1. Thus, for $n \geq 1$ let $d_n := \lceil Cnb_n \rceil$, $l_n = \lfloor \frac{n}{d_n} \rfloor$, and for $0 \leq l \leq l_n$, define

$$M_{n,l} = \max_{j \in \{ld_n+1,\ldots,(l+1)d_n\}} \min_{i \in \{j,\ldots,j+d_n\}} Z_i$$

so that

$$M_{n,0} = \max_{j \in \{1,\ldots,d_n\}} \min_{i \in \{j,\ldots,j+d_n\}} Z_i.$$

Then, for all $y > 0$, we have

$$\begin{aligned}
&\mathbb{P}\left(\sup_{x \in [0,1]} \min_{\substack{i \in \{1,\ldots,n\} \\ |X_{(i)}-x| \leq b_n}} Z_i > y \right) \\
&\leq \mathbb{P}\left(\left\{ \max_{j \in \{1,\ldots,n-d_n\}} \min_{i \in \{j,\ldots,j+d_n\}} Z_i > y \right\} \cap \Omega_n \right) + \mathbb{P}\left(\Omega_n^c\right) \\
&\leq \mathbb{P}\left(\max_{\substack{l \in \{0,\ldots,l_n\} \\ l \text{ even}}} M_{n,l} > y \right) + \mathbb{P}\left(\max_{\substack{l \in \{0,\ldots,l_n\} \\ l \text{ odd}}} M_{n,l} > y \right) + \mathbb{P}\left(\Omega_n^c\right).
\end{aligned}$$

Since $M_{n,l}$ are iid, it follows that, for l even (and similarly for l odd),

$$\mathbb{P}\left(\max_{\substack{l \in \{0,\ldots,l_n\} \\ l \text{ even}}} M_{n,l} > y \right) = 1 - \mathbb{P}\left(M_{n,0} < y\right)^{\lfloor l_n/2 \rfloor + 1}.$$

This leads to

$$\mathbb{P}\left(\sup_{x \in [0,1]} \min_{\substack{i \in \{1,\ldots,n\} \\ |X_{(i)}-x| \leq b_n}} Z_i > y \right) \leq 2\left(1 - \mathbb{P}\left(M_{n,0} < y\right)^{\lfloor l_n/2 \rfloor + 1}\right) + \mathbb{P}\left(\Omega_n^c\right). \quad (19)$$

Besides for $n \geq 1$, we also have

$$\begin{aligned}
\mathbb{P}\left(M_{n,0} > y\right) &= \mathbb{P}\left(\min_{i \in \{1,\ldots,d_n+1\}} Z_i > y \right) \\
&\quad + \sum_{j=2}^{d_n} \mathbb{P}\left(\{Z_{j-1} \leq y\} \bigcap \left\{ \min_{i \in \{j,\ldots,j+d_n\}} Z_i > y \right\} \right) \\
&= \overline{U}(y)^{d_n+1} + (d_n - 1)U(y)\overline{U}(y)^{d_n+1} \\
&\leq (1 + d_n U(y))\overline{U}(y)^{d_n}. \quad (20)
\end{aligned}$$

Until now, plugging inequality (20) in inequality (19), we have shown that for $\epsilon > 0$

$$\mathbb{P}\left(\sup_{x\in[0,1]} \min_{\substack{i\in\{1,\dots,n\}\\ |X_{(i)}-x|\le b_n}} Z_i > y\right) \le 2\left(\left(1-\left[1-(1+d_nU(\epsilon))\overline{U}(\epsilon)^{d_n}\right)\right]^{\lfloor l_n/2\rfloor+1}\right)+\mathbb{P}\left(\Omega_n^c\right).$$

To conclude the proof, it remains to prove that the right-hand side in the latest equation tends to 0 for all $\epsilon > 0$. We already know that $\lim_{n\to\infty}\mathbb{P}\left(\Omega_n^c\right) = 0$. Thus, since $l_n \underset{n\to\infty}{\sim} n/d_n$, it arises if

$$(1+d_nU(\epsilon))\overline{U}(\epsilon)^{d_n} = o\left(b_n\right). \tag{21}$$

Since $d_n \underset{n\to\infty}{\sim} nb_n$, Eq. (21) holds if

$$c(n) := nb_n \log(\overline{U}(\epsilon)) + \log\left(1+nb_nU(\epsilon)\right) - \log\left(b_n\right) \underset{n\to\infty}{\longrightarrow} -\infty.$$

Assumption **(A4)** implies

$$\frac{|\log(b_n)|}{nb_n} = O\left(\frac{\log(n)}{nb_n}\right) = o(1).$$

Let $c > 0$. Then, for n sufficiently large $-\log\left(b_n\right) \le cnb_n$. Choosing $c < |\log(\overline{U}(\epsilon))|$, we thus have

$$\begin{aligned} c(n) &\le nb_n(\log(\overline{U}(\epsilon))+c)+\log\left(2nb_nU(\epsilon)\right)\\ &= \log\left(nb_n\right).\left[\frac{nb_n}{\log\left(nb_n\right)}(\log(\overline{U}(\epsilon))+c)+\frac{\log(2U(\epsilon))}{\log\left(nb_n\right)}+1\right]\\ &\underset{n\to\infty}{\longrightarrow} -\infty, \end{aligned}$$

since $\overline{U}(\epsilon) < 1$ under **(A1)** and $nb_n \to \infty$ when $n \to \infty$ necessarily under **(A4)**. This proves Eq. (15) and concludes the proof. □

We are now ready to prove Lemma 3.

Proof of Lemma 3 On the one hand, we have

$$\begin{aligned} \sup_{x\in[0,1]}\left(\tilde{h}_0(x)-h_0(x)\right) &= \sup_{x\in[0,1]}\left(\max_{\substack{i\in\{1,\dots,n\}\\ |X_i-x|\le b_n}}\{h_0(X_i)+\varepsilon_i-h_0(x)\}\right)\\ &\le \sup_{x\in[0,1]}\left(\max_{\substack{i\in\{1,\dots,n\}\\ |X_i-x|\le b_n}}\{h_0(X_i)-h_0(x)\}\right) \end{aligned}$$

$$\begin{aligned}&\leq \sup_{|t-x|\leq b_n} |h_0(t)-h_0(x)|\\&= o(1), \qquad (22)\end{aligned}$$

since the errors $(\varepsilon_i)_{1\leq i\leq n}$ are nonpositive and h_0 is continuous on the compact set $[0, 1]$ and thereby uniformly continuous under **(A3)**. On the other hand,

$$\begin{aligned}\sup_{x\in[0,1]}\left(h_0(x)-\tilde{h}_0(x)\right) &= \sup_{x\in[0,1]}\left(h_0(x)-\max_{\substack{i\in\{1,\ldots,n\}\\|X_i-x|\leq b_n}}\left\{h_0(X_i)+\varepsilon_{i,n}\right\}\right)\\&=\sup_{x\in[0,1]}\left(\min_{\substack{i\in\{1,\ldots,n\}\\|X_i-x|\leq b_n}}\left\{h_0(x)-h_0(X_i)-\varepsilon_{i,n}\right\}\right)\\&\leq \sup_{x\in[0,1]}\left(\min_{\substack{i\in\{1,\ldots,n\}\\|X_i-x|\leq b_n}}\{-\varepsilon_i\}\right)+\sup_{|t-x|\leq b_n}|h_0(t)-h_0(x)|\\&=o_P(1), \qquad (23)\end{aligned}$$

with Lemma 9. Finally, combining Eqs. (22) and (23), it follows that

$$\begin{aligned}\|h_0-\tilde{h}_0\|_\infty &= \sup_{x\in[0,1]}\left|\tilde{h}_0(x)-h_0(x)\right|\\&=\sup_{x\in[0,1]}\left(\max\left\{\tilde{h}_0(x)-h_0(x), h_0(x)-\tilde{h}_0(x)\right\}\right)\\&\leq \max\left\{\sup_{x\in[0,1]}(\tilde{h}_0(x)-h_0(x)), \sup_{x\in[0,1]}(h_0(x)-\tilde{h}_0(x))\right\} = o_P(1),\end{aligned}$$

which is the desired result. □

A.2 Proof of Lemma 7

Let $\epsilon > 0$. Note that $\Lambda_0^{-1}\circ h_0$ is uniformly continuous due to assumptions **(A3)** and **(B2)**. Thus with $h_\vartheta = \Lambda_\vartheta \circ \Lambda_0^{-1}\circ h_0$ and assumption **(B3)**, it follows that there exists some $\delta > 0$ such that $\sup_{\vartheta\in\Theta}|h_\vartheta(x)-h_\vartheta(y)| \leq \epsilon$ if $|x-y|\leq\delta$. Now let n be large enough such that $b_n \leq \delta$. Then due to the definition of $\hat{h}_\vartheta$ and $\operatorname{supp}(K) = [-1, 1]$, one obtains

$$\begin{aligned}&\|\hat{h}_\vartheta - h_\vartheta\|_\infty \leq \|\tilde{h}_\vartheta - h_\vartheta\|_\infty\\&\quad+\sup_{x\in[0,1]}\left|\frac{\sum_{i=1}^n (h_\vartheta(X_i)-h_\vartheta(x))K\left(\frac{x-X_i}{b_n}\right)}{\sum_{i=1}^n K\left(\frac{x-X_i}{b_n}\right)}\right| \leq \|\tilde{h}_\vartheta - h_\vartheta\|_\infty + \epsilon.\end{aligned}$$

From Lemma 3, we have $\|\tilde{h}_0 - h_0\|_\infty = o_P(1)$, and thus with assumption **(B3)**, it follows that

$$\sup_{\vartheta\in\Theta} \|\tilde{h}_\vartheta - h_\vartheta\|_\infty = \sup_{\vartheta\in\Theta} \sup_{x\in[0,1]} |\Lambda_\vartheta(\Lambda_0^{-1}(\tilde{h}_0(x))) - \Lambda_\vartheta(\Lambda_0^{-1}(h_0(x)))| = o_P(1)$$

and therefore the assertion of the lemma. □

A.3 Proof of Theorem 6

By the argmax theorem applied to the criterion function $M_n(\vartheta)$ multiplied by (-1) and using assumption **(B1)**, it suffices to show that

$$\sup_{\vartheta\in\Theta} |M_n(\vartheta) - M(\vartheta)| = o_P(1)$$

(see Kosorok 2008, Theorem 2.12(i)). To obtain this, note that

$$\begin{aligned}\sup_{\vartheta\in\Theta} |M_n(\vartheta) - M(\vartheta)| \le& \sup_{\vartheta\in\Theta} \|G_n(\vartheta, \hat{h}_\vartheta) - \bar{G}_n(\vartheta, \hat{h}_\vartheta)\| \\ &+ \sup_{\vartheta\in\Theta} \|\bar{G}_n(\vartheta, \hat{h}_\vartheta) - G(\vartheta, \hat{h}_\vartheta)\| \\ &+ \sup_{\vartheta\in\Theta} \|G(\vartheta, \hat{h}_\vartheta) - G(\vartheta, h_\vartheta)\|,\end{aligned}$$

where

$$\begin{aligned}\bar{G}_n(\vartheta, h)(y, s) =& \frac{1}{n}\sum_{i=1}^{n} I\{\Lambda_\vartheta(Y_i) - h(X_i) \le y\} I\{X_i \le s\} \\ &- F_X(s)\frac{1}{n}\sum_{i=1}^{n} I\{\Lambda_\vartheta(Y_i) - h(X_i) \le y\}. \qquad (24)\end{aligned}$$

Note that for any deterministic function h, we have $\mathbb{E}[\bar{G}_n(\vartheta, h)] = G(\vartheta, h)$. The assertion of the theorem follows from

$$\sup_{\vartheta\in\Theta} \|G_n(\vartheta, \hat{h}_\vartheta) - \bar{G}_n(\vartheta, \hat{h}_\vartheta)\| \le \sup_{s\in[0,1]} |\hat{F}_{X,n}(s) - F_X(s)| = o_P(1)$$

and Lemmas 10 and 11. □

Lemma 10 *Under the assumptions of Theorem 6 (i),*

$$\sup_{\vartheta\in\Theta} \|\bar{G}_n(\vartheta, \hat{h}_\vartheta) - G(\vartheta, \hat{h}_\vartheta)\| = o_P(1).$$

Proof From Lemma 7 follows the existence of some deterministic sequence $a_n \searrow 0$ such that the probability of the event

$$\sup_{\vartheta\in\Theta} \|h_\vartheta - \hat{h}_\vartheta\|_\infty \leq a_n \tag{25}$$

converges to one. Thus, we assume in what follows that (25) holds.

We only consider the difference between the first sum in the definition of $\bar{G}_n(\vartheta, h)$ (see (24)) and the first integral in the definition of $G(\vartheta, h)$ (see (7)). The difference between the second sum and the second integral can be treated similarly. Applying (25) the first sum in $\bar{G}_n(\vartheta, \hat{h}_\vartheta)(y, s)$ can be nested as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n I\{\Lambda_\vartheta(Y_i) - h_\vartheta(X_i) \leq y - a_n\} I\{X_i \leq s\}\\
&\quad\leq \frac{1}{n}\sum_{i=1}^n I\{\Lambda_\vartheta(Y_i) - \hat{h}_\vartheta(X_i) \leq y\} I\{X_i \leq s\}\\
&\quad\leq \frac{1}{n}\sum_{i=1}^n I\{\Lambda_\vartheta(Y_i) - h_\vartheta(X_i) \leq y + a_n\} I\{X_i \leq s\}
\end{aligned}$$

while the first integral in $G(\vartheta, \hat{h}_\vartheta)(y, s)$ can be nested as

$$\begin{aligned}
&\int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y - a_n + h_\vartheta(x))) - h_0(x)\right) I\{x \leq s\} f_X(x)\,\mathrm{d}x\\
&\quad\leq \int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + \hat{h}_\vartheta(x))) - h_0(x)\right) I\{x \leq s\} f_X(x)\,\mathrm{d}x\\
&\quad\leq \int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + a_n + h_\vartheta(x))) - h_0(x)\right) I\{x \leq s\} f_X(x)\,\mathrm{d}x.
\end{aligned}$$

Thus, we have to consider

$$\begin{aligned}
&H_{n,\vartheta}^{(1)}(y, s)\\
&\quad= \frac{1}{n}\sum_{i=1}^n \Bigg(I\{\Lambda_\vartheta(Y_i) - h_\vartheta(X_i) \leq y + a_n\} I\{X_i \leq s\}\\
&\qquad - \int F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + a_n + h_\vartheta(x))) - h_0(x)\right) I\{x \leq s\} f_X(x)\,\mathrm{d}x\Bigg)\\
&H_{n,\vartheta}^{(2)}(y, s)\\
&\quad= \int \Bigg(F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + a_n + h_\vartheta(x))) - h_0(x)\right)\\
&\qquad - F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(y + h_\vartheta(x))) - h_0(x)\right)\Bigg) I\{x \leq s\} f_X(x)\,\mathrm{d}x,
\end{aligned}$$

and the same terms with $y + a_n$ replaced by $y - a_n$, which can be treated completely analogously. We have to show that $\sup_{\vartheta\in\Theta} \|H^{(1)}_{n,\vartheta}\| = o_P(1)$ and $\sup_{\vartheta\in\Theta} \|H^{(2)}_{n,\vartheta}\| = o(1)$.

Recall condition **(N1)** and note that $\sup_{\vartheta\in\Theta} \sup_{\substack{s\in[0,1]\\ y\in C}} |H^{(2)}_{n,\vartheta}(y,s)| = o(1)$ follows from uniform continuity of F_0 and of $\Lambda_0 \circ \Lambda_\vartheta^{-1}$ uniformly in ϑ (see **(B5)** and **(B4)**), from the representation $h_\vartheta = \Lambda_\vartheta \circ \Lambda_0^{-1} \circ h_0$ and uniform continuity of Λ_ϑ uniformly in ϑ (see **(B3)**), and $a_n \to 0$.

Let n be large enough such that $|a_n| \leq \tau$ for τ both from **(B5)** and **(B4)**. Now to prove $\sup_{\vartheta\in\Theta} \|H^{(1)}_{n,\vartheta}\| = o_P(1)$, note that

$$\sup_{\vartheta\in\Theta} \|H^{(1)}_{n,\vartheta}\| \leq \sup_{f\in\mathcal{F}} |P_n f - Pf|,$$

where P_n denotes the empirical measure of $(X_1, Y_1), \ldots, (X_n, Y_n)$ and P the measure of (X_1, Y_1) and

$$\mathcal{F} = \{(x,y) \mapsto I\{\Lambda_\vartheta(y) - h_\vartheta(x) \leq z\} I\{x \leq s\} \mid \vartheta \in \Theta, s \in [0,1], z \in C_\tau\}$$

with C_τ as in assumption **(B4)**. The assertion follows from the Glivenko–Cantelli theorem as stated in Theorem 2.4.1 in van der Vaart and Wellner (1996) if we show that the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, L_1(P))$ is finite for each $\epsilon > 0$. To this end let $\epsilon > 0$, and for the moment fix $s \in [0,1]$, $\vartheta \in \Theta$ and $z \in C_\tau$. Choose $\delta > 0$ corresponding to ϵ as in assumption **(B5)**.

Partition $[0,1]$ into finitely many intervals $[s_j, s_{j+1}]$ such that $F_X(s_{j+1}) - F_X(s_j) \leq \epsilon$ for all j. For the fixed s, denote the interval containing s by $[s_j, s_{j+1}] = [s^\ell, s^u]$.

Now choose a finite sup-norm bracketing of length γ for the class $\mathcal{L}_S = \{\Lambda_\vartheta|_S : \vartheta \in \Theta\}$ according to (10) with γ as in assumption **(B4)** corresponding to the above chosen δ. For the fixed ϑ, this gives a bracket $h^\ell \leq h_\vartheta \leq h^u$ of sup-norm length γ.

Choose a finite sup-norm bracketing of length δ for the class $\mathcal{L}^1_{\tilde{S}} = \{\Lambda_0 \circ \Lambda_\vartheta^{-1}|_{\tilde{S}} : \vartheta \in \Theta\}$ according to (10). For the fixed ϑ, this gives a bracket $V^\ell \leq \Lambda_0 \circ \Lambda_\vartheta^{-1} \leq V^u$.

Then, consider the bounded and increasing function

$$D(z) = \int F_0(V^\ell(z + h^\ell(x)) - h_0(x)) f_X(x)\,\mathrm{d}x,$$

and choose a finite partition of the compact C_τ in intervals $[z_k, z_{k+1}]$ such that $D(z_{k+1}) - D(z_k) < \epsilon$. For the fixed z, denote the interval containing z by $[z_k, z_{k+1}] = [z^\ell, z^u]$.

Now for the function $f \in \mathcal{F}$, that is, determined by ϑ, s and z, a bracket is given by $[f^\ell, f^u]$ with

$$f^\ell(x,y) = I\{\Lambda_0(y) \leq V^\ell(z^\ell + h^\ell(x))\} I\{x \leq s^\ell\}$$
$$f^u(x,y) = I\{\Lambda_0(y) \leq V^u(z^u + h^u(x))\} I\{x \leq s^u\}$$

with $L_1(P)$ norm

$$\begin{aligned}
&\mathbb{E}[I\{\Lambda_0(Y_i) \le V^u(z^u + h^u(X_i))\}I\{X_i \le s^u\}] \\
&\quad -\mathbb{E}[I\{\Lambda_0(Y_i) \le V^\ell(z^\ell + h^\ell(X_i))\}I\{X_i \le s^\ell\}] \\
&\le F_X(s^u) - F_X(s^\ell) \\
&\quad + \int \Big| F_0\left(V^u(z^u + h^u(x)) - h_0(x)\right) \\
&\quad - F_0\left(V^\ell(z^\ell + h^\ell(x)) - h_0(x)\right) \Big| f_X(x)\,\mathrm{d}x \\
&\le 2\epsilon + \int \Big| F_0\left(V^u(z^u + h^u(x)) - h_0(x)\right) \\
&\quad - F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(z^u + h^u(x))) - h_0(x)\right) \Big| f_X(x)\,\mathrm{d}x \\
&\quad + \int \Big| F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(z^u + h^\ell(x))) - h_0(x)\right) \\
&\quad - F_0\left(V^\ell(z^u + h^\ell(x)) - h_0(x)\right) \Big| f_X(x)\,\mathrm{d}x \\
&\quad + \int \Big| F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(z^u + h^u(x))) - h_0(x)\right) \\
&\quad - F_0\left(\Lambda_0(\Lambda_\vartheta^{-1}(z^u + h^\ell(x))) - h_0(x)\right) \Big| f_X(x)\,\mathrm{d}x \\
&\le 4\epsilon
\end{aligned}$$

by the definition of $[s^\ell, s^u]$ and $[z^\ell, z^u]$ and using the construction of brackets above. (Note that $\|V^u - \Lambda_0 \circ \Lambda_\vartheta^{-1}\|_\infty \le \delta$, $\|\Lambda_0 \circ \Lambda_\vartheta^{-1} - V^\ell\|_\infty \le \delta$, $\|h^u - h^\ell\|_\infty \le \gamma$ and recall assumptions **(B5)** and **(B4)**).

There are finitely many such brackets to cover $\mathcal{F}$, and thus the assertion follows. □

Lemma 11 *Under the assumptions of Theorem 6 (i),*

$$\sup_{\vartheta \in \Theta} \|G(\vartheta, h_\vartheta) - G(\vartheta, \hat{h}_\vartheta)\| = o_P(1).$$

Proof According to assumption **(N1)**, it suffices to show

$$\sup_{\vartheta \in \Theta} \sup_{\substack{s \in [0,1] \\ y \in C}} |G(\vartheta, h_\vartheta)(y, s) - G(\vartheta, \hat{h}_\vartheta)(y, s)| = o_P(1).$$

Recalling the definition of G in (7), we see that the assertion follows from Lemma 7 and uniform continuity of F_0 and of $\Lambda_0 \circ \Lambda_\vartheta^{-1}$ (uniformly in ϑ). □

B Identifiability of the model in the random design case

We prove the assertion of Remark 2. First note that $\varepsilon(\vartheta_1)$ is independent of X, and thus the conditional distribution of $\varepsilon(\vartheta_1)$, i.e.,

$$\begin{aligned}\mathbb{P}(\varepsilon(\vartheta_1) \leq y \mid X = x) &= \mathbb{P}(Y \leq \Lambda_{\vartheta_1}^{-1}(y + h_{\vartheta_1}(x))) \mid X = x) \\ &= F_0(\Lambda_{\vartheta_0}(\Lambda_{\vartheta_1}^{-1}(y + h_{\vartheta_1}(x))) - h_{\vartheta_0}(x)),\end{aligned}$$

does not depend on x. Further, $h_{\vartheta_0} = \Lambda_{\vartheta_0} \circ \Lambda_{\vartheta_1}^{-1} \circ h_{\vartheta_1}$, and for $y \leq 0$, we have $\Lambda_{\vartheta_0}(\Lambda_{\vartheta_1}^{-1}(y+h_{\vartheta_1}(x))) \leq \Lambda_{\vartheta_0}(\Lambda_{\vartheta_1}^{-1}(h_{\vartheta_1}(x)))$ because $\Lambda_{\vartheta_0} \circ \Lambda_{\vartheta_1}^{-1}$ is strictly increasing. As F_0 is strictly increasing by assumption, it follows that

$$H^{-1}(y + H(h_{\vartheta_0}(x))) - h_{\vartheta_0}(x)$$

does not depend on x for $y \in (-\infty, 0]$ and $x \in [0, 1]$, where for ease of presentation write $H := \Lambda_{\vartheta_1} \circ \Lambda_{\vartheta_0}^{-1}$. Thus,

$$H^{-1}(y + H(a)) - a = H^{-1}(y + H(b)) - b$$

for all $y \leq 0$, $a, b \in h_{\vartheta_0}([0, 1])$. Because Y may take the value 0 by assumption and $\varepsilon \leq 0$, one obtains $h_{\vartheta_0}([0, 1]) \cap \mathbb{R}_0^+ \neq \emptyset$. To conclude the proof, we distinguish two cases.

(1) Let $h_{\vartheta_0}([0, 1]) \cap \mathbb{R}_0^+ = \{0\}$. Set $a = 0$, since by assumption $\Lambda_\vartheta(0) = 0$ for all $\vartheta \in \Theta$, then

$$H^{-1}(y) = H^{-1}(y + H(b)) - b$$

for all $y \leq 0$, $b \in h_{\vartheta_0}([0, 1]) \subset \mathbb{R}_0^-$. Set $c = H^{-1}(y + H(b))$, then it follows that

$$H(c) - H(b) = H(c - b)$$

for all $b, c \in (-\delta, 0]$ for some $\delta > 0$, and from the assumptions it follows that $\vartheta_1 = \vartheta_0$ with $H = \mathrm{id}$.

(2) Let $h_{\vartheta_0}([0, 1]) \cap \mathbb{R}_0^+ = I$ be an interval of positive length. For $a \in I$, one has $y := -H(a) \leq 0$, and

$$0 = H^{-1}(0) = a + H^{-1}(-H(a) + H(b)) - b,$$

and thus

$$H(b - a) = H(b) - H(a)$$

for all $a, b \in I$. From the assumptions, it follows that $\vartheta_1 = \vartheta_0$ with $H = \mathrm{id}$ and thus identifiability of the model. □

References

Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*: *Series B*, *26*, 211–252.

Brown, L. D., Low, M. G. (1996). Asymptotic equivalenfce of nonparametric regression and white noise. *The Annals of Statistics*, *24*, 2384–2398.

Carroll, R. J., Ruppert, D. (1988). *Transformation and weighting in regression, monographs on statistics and applied probability*. New York: Chapman & Hall.

Colling, B., Van Keilegom, I. (2016). Goodness-of-fit tests in semiparametric transformation models. *TEST*, *25*, 291–308.

Daouia, A., Noh, H., Park, B. U. (2016). Data envelope fitting with constrained polynomial splines. *Journal of the Royal Statistical Society*: *Series B*, *78*, 3–30.

Drees, H., Neumeyer, N., Selk, L. (2019). Estimation and hypotheses testing in boundary regression models. *Bernoulli*, *25*, 424–463.

Girard, S., Jacob, P. (2008). Frontier estimation via kernel regression on high power-transformed data. *Journal of Multivariate Analysis*, *99*, 403–420.

Hall, P., Park, B. U., Stern, S. E. (1998). On polynomial estimators of frontiers and boundaries. *Journal of Multivariate Analysis*, *66*, 71–98.

Hall, P., Van Keilegom, I. (2009). Nonparametric “regression” when errors are positioned at end-points. *Bernoulli*, *15*, 614–633.

Härdle, W., Park, B. U., Tsybakov, A. B. (1995). Estimation of a non sharp support boundaries. *Journal of Multivariate Analysis*, *55*, 205–218.

Horowitz, J. L. (2009). *Semiparametric and nonparametric methods in econometrics*. Springer series in statistics. New York: Springer.

Jirak, M., Meister, A., Reiß, M. (2013). Asymptotic equivalence for nonparametric regression with non-regular errors. *Probability Theory and Relative Fields*, *155*, 201–229.

Jirak, M., Meister, A., Reiß, M. (2014). Adaptive estimation in nonparametric regression with one-sided errors. *Annals of Statistics*, *42*, 1970–2002.

Jones, M. C., Pewsey, A. (2009). Sinh–arcsinh distributions. *Biometrika*, *96*, 761–780.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer Series in Statistics. New York: Springer.

Linton, O., Sperlich, S., Van Keilegom, I. (2008). Estimation on a semiparametric transformation model. *Annals of Statistics*, *36*, 686–718.

Manly, B. F. J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society*: *Series D*, *25*, 37–42.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*, 77–91.

Meister, A., Reiß, M. (2013). Asymptotic equivalence for nonparametric regression with non-regular errors. *Probability Theory and Related Fields*, *155*, 201–229.

Mu, Y., He, X. (2007). Power transformation toward a linear regression quantile. *Journal of the American Statistical Association*, *102*, 269–279.

Müller, U. U., Wefelmeyer, W. (2010). Estimation in nonparametric regression with non-regular errors. *Commmunication in Statistics—Theory and Methods*, *39*, 1619–1629.

Pollard, D. (1984). *Convergence of stochastic processes*. New York: Springer.

Powell, J. (1991). Estimation of monotonic regression models under quantile restrictions. In W. Barnett, J. Powell & G. Tauchen (Eds.), *Nonparametric and semiparametric methods in econometrics*, pp. 357–384. New York: Cambridge University Press.

Simar, L., Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, *44*, 49–61.

van der Vaart, A. W., Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Wilson, P. W. (2003). Testing independence in models of productive efficiency. *Journal of Productivity Analysis*, *20*, 361–390.

Yeo, I.-K., Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*, 954–959.