Electronic Journal of Statistics
Vol. 14 (2020) 769–800
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1676

Estimation of a semiparametric transformation model: A novel approach based on least squares minimization

Benjamin Colling and Ingrid Van Keilegom

Research Centre for Operations Research and Statistics (ORSTAT), KU Leuven,
Naamsestraat 69, 3000 Leuven, Belgium
e-mail: benjamin.colling@uclouvain.be*;* ingrid.vankeilegom@kuleuven.be

Abstract: Consider the following semiparametric transformation model $\Lambda_\theta(Y) = m(X) + \varepsilon$, where X is a d-dimensional covariate, Y is a univariate response variable and ε is an error term with zero mean and independent of X. We assume that m is an unknown regression function and that $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions. Our goal is to develop two new estimators of the transformation parameter θ. The main idea of these two estimators is to minimize, with respect to θ, the L_2-distance between the transformation Λ_θ and one of its fully nonparametric estimators. We consider in particular the nonparametric estimator based on the least-absolute deviation loss constructed in Colling and Van Keilegom (2019). We establish the consistency and the asymptotic normality of the two proposed estimators of θ. We also carry out a simulation study to illustrate and compare the performance of our new parametric estimators to that of the profile likelihood estimator constructed in Linton et al. (2008).

Keywords and phrases: Asymptotic properties, estimation, L_2-distance minimization, parametric transformation, semiparametric regression.

Received July 2019.

Contents

1. Introduction

Transforming the data is a very common practice in statistics in order to improve the performance of a model or to interpret in an easier way a model. Transformation models can be encountered in a lot of various contexts, like in survival analysis and in quantile regression for example. In survival analysis, we mention the seminal works of Cox (1972) and Bennett (1983), who introduced respectively the Cox proportional hazards model and the proportional odds model to examine the effect of covariates on the survival time. The 'Box-Cox quantile regression model', based on the Box and Cox (1964) transform, is very popular in quantile regression, see Buchinsky (1995), Machado and Mata (2000), Mu and He (2007), and Fitzenberger et al. (2010), among others.

Historically speaking, transformations of the response variable go back to the simple linear regression model $Y = X^t\beta + \varepsilon$, where Y is a dependent variable, X is a vector of explanatory variables, β is a vector of unknown regression parameters and ε is the error term. This model relies on heavy assumptions and the violation of one or several of these assumptions could lead to inconsistent or inefficient estimation of the corresponding parameters and also to wrong predictions of the response Y. As a possible solution to this problem, Box and Cox (1964) introduced a parametric family of power transformations and suggested that this power transformation, when it is applied to the response variable Y, might induce additivity of the effects, homoscedasticity and normality of the new error term and reduce skewness and hence satisfy as much as possible the assumptions of the new linear regression model. Note that the Box and Cox (1964) transformation also includes as special cases the logarithm, the square root, the inverse and the identity.

The class of transformations introduced by Box and Cox (1964) has been generalized, see for example the Yeo and Johnson (2000) transform. We also mention the book of Carroll and Ruppert (1988), the review paper of Sakia (1992) and the papers of Zellner and Revankar (1969), John and Draper (1980), Bickel and Doksum (1981) and MacKinnon and Magee (1990) for more classes of transformations and more details on this topic.

In the literature on transformation models the regression function and the transformation of the response can be either parametric or nonparametric. The above mentioned papers all consider regression models which assume a parametric form for both functions. In the context of nonparametric transformations and parametric regression functions, we mention the work of Horowitz (1996), who proposed nonparametric estimators of the transformation and the cumulative distribution function of the error term and the work of Chen (2002), who proposed a rank-based estimator of the transformation that has the advantage of not involving nonparametric smoothing.

Next, in the context of fully nonparametric transformation models of the form $\Lambda(Y) = m(X) + \varepsilon$, where $\Lambda(\cdot)$ and $m(\cdot)$ are respectively an unknown transformation and an unknown regression function, Chiappori et al. (2015) and more recently Colling and Van Keilegom (2019) constructed fully nonparametric estimators of the transformation Λ. The main motivation of the estimators constructed in Colling and Van Keilegom (2019) with respect to the ones constructed in Chiappori et al. (2015), was to avoid kernel smoothing on Y since this can work badly in practice if the distribution of Y is very skewed. Their main idea was to rewrite the transformation $\Lambda(Y)$ as $\Gamma(U)$, where Γ is an increasing function, $U = [F_Y(Y) - F_Y(0)]/[F_Y(1) - F_Y(0)]$ and $F_Y(\cdot)$ is the distribution function of Y, and to construct estimators based on kernel smoothing of U, which works globally better since U is uniformly distributed. We also mention the work of Breiman and Friedman (1985) who constructed an algorithm for estimating the different components of the same model when the regression function m is supposed to be additive.

In fully nonparametric contexts that are slightly different from that of the previous model, we would also like to mention the works of Horowitz (2001) and Jacho-Chavez et al. (2010) among others, who proposed nonparametric estimators of a generalized additive model with an unknown link function.

In this paper, we will focus on a model that assumes a parametric form for the transformation function, while the regression function is left unspecified, i.e., we will consider a semiparametric transformation model of the following form:

$$\Lambda_\theta(Y) = m(X) + \varepsilon \ , \tag{1}$$

where $m(\cdot)$ is an unknown regression function, Λ_θ is a transformation belonging to a parametric family of strictly increasing functions and $\theta \in \Theta$ where Θ is a compact subset of $\mathbb{R}^k$. We will denote by θ_0 the true but unknown value of θ. Moreover, we assume that X is a d-dimensional covariate with compact support χ, Y is a univariate response variable with support $\mathcal{Y}$ and the error term ε has zero mean and is independent of X. We also introduce the following notations: $m(x,\theta) = E[\Lambda_\theta(Y)|X = x]$, $m(x,\theta_0) = m(x)$, $\varepsilon(\theta) = \Lambda_\theta(Y) - m(X,\theta)$ and $\varepsilon(\theta_0) = \varepsilon$. Finally, let F_X, $F_{\varepsilon(\theta)}$, f_X and $f_{\varepsilon(\theta)}$ be the distribution and density functions of X and $\varepsilon(\theta)$. We assume that we have randomly drawn an *iid* sample $(X_1,Y_1),\ldots,(X_n,Y_n)$ from model (1), where the components of X_i are denoted by $(X_{i1},\ldots,X_{id})$ for $i = 1,\ldots,n$.

Linton et al. (2008) extensively studied the semiparametric transformation model (1) and proposed two estimation methods for the unknown true parameter vector θ_0: a profile likelihood method and a mean squared distance from independence method. Moreover, they established the asymptotic properties of these two estimators and showed in their simulation study that the profile likelihood estimator outperforms the other one. The main idea of the profile likelihood method is to maximize the log-likelihood function of the vector (X,Y) with respect to θ, after having replaced all unknown functions in the likelihood by nonparametric estimators. Then, the profile likelihood estimator of θ is defined by

$$\widehat{\theta}_{PL} = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \left\{ \log \widehat{f}_{\varepsilon(\theta)}(\Lambda_\theta(Y_i) - \widehat{m}(X_i,\theta)) + \log \Lambda'_\theta(Y_i) \right\} , \tag{2}$$

where $\widehat{m}(\cdot,\theta)$ and $\widehat{f}_{\varepsilon(\theta)}(\cdot)$ are suitable nonparametric estimators of $m(\cdot,\theta)$ and $f_{\varepsilon(\theta)}(\cdot)$ respectively and $\Lambda'_\theta(y) = \frac{\partial}{\partial y}\Lambda_\theta(y)$.

In the literature we can find several other contributions on the semiparametric transformation model (1). First, we mention the work of Vanhems and Van Keilegom (2019) who studied the estimation of this model when some of the regressors are supposed to be endogenous. Next, we refer to Colling and Van Keilegom (2016), Colling and Van Keilegom (2017) and Kloodt and Neumeyer (2017), who developed tests for the parametric form of the regression function based on the error distribution function, the integrated regression function and a L_2-distance between the nonparametric and the parametric fits of m, respectively, while Allison et al. (2018) and Kloodt and Neumeyer (2017) constructed significance tests for the explanatory variables in the model based on Fourier-type conditional expectations and on U-statistics, respectively. Moreover, Hušková et al. (2018) proposed tests for the validity of the model involving characteristic functions and Colling et al. (2015) and Heuchenne et al. (2015) studied nonparametric estimators of the error density and the error distribution respectively. Finally, we also mention the work of Neumeyer et al. (2016) who introduced estimators of the different components of a heteroscedastic transformation model and proved the asymptotic normality of these estimators.

In this paper, our goal is to construct two new estimators of the transformation parameter θ_0 in the context of a semiparametric transformation model of the form (1). These estimators will be competitors of the profile likelihood estimator $\widehat{\theta}_{PL}$ introduced in (2). The main idea of the new estimators of θ_0 is to minimize, with respect to θ, the L_2-distance between the transformation Λ_θ and one of its fully nonparametric estimators. In Section 2, we will explain in more detail the intuition behind our two new estimators of θ_0, while we will give their exact definitions in Section 3. Next, in Section 4, we will present the theorems that establish the consistency and the asymptotic normality of these two estimators. A simulation study comparing the performance of our new estimators with that of the profile likelihood estimator is performed in Section 5. Some general conclusions and discussion about the contexts in which the proposed method is effective, is given in Section 6. Finally, Appendix A contains the technical assumptions and the proofs of the main results.

2. Main idea of the new estimators

As mentioned in the introduction, the main idea of the new estimators of the transformation parameter θ_0 is to minimize, with respect to θ, the L_2-distance between the transformation Λ_θ and one of its fully nonparametric estimators. Nonparametric estimators of the transformation have already been constructed in the literature, see Chiappori et al. (2015) and Colling and Van Keilegom (2019). Moreover, we explained in the introduction why the estimators constructed in Colling and Van Keilegom (2019) perform globally better than those constructed in Chiappori et al. (2015). The simulation studies performed in Chiappori et al. (2015) and Colling and Van Keilegom (2019) also show that a nonparametric estimator of the transformation based on the least absolute deviation loss performs better than a corresponding estimator based on the least squares loss, since the former is less sensitive to outliers. Consequently, we will use here the nonparametric estimator based on the least absolute deviation loss constructed by Colling and Van Keilegom (2019), which is, as far as we know, the nonparametric estimator of the transformation that performs globally the best.

To construct this estimator we need to assume that the true transformation $\Lambda = \Lambda_{\theta_0}$ satisfies $\Lambda(0) = 0$ and $\Lambda(1) = 1$. However, as we will see later, other identifiability constraints are possible as well. The latter condition on Λ fixes the location and the scale of the model, which is sufficient to identify the model. See Chiappori et al. (2015) and Colling and Van Keilegom (2019) for more details about the identification of the model. Following the same idea as in Colling and Van Keilegom (2019), we rewrite the transformation $\Lambda(Y)$ as $\Lambda(Y) = \Gamma(U)$, where Γ is an increasing function, and

$$U = T(Y) = \frac{F_Y(Y) - F_Y(0)}{F_Y(1) - F_Y(0)} \ . \tag{3}$$

Note that $T(0) = 0$ and $T(1) = 1$, and hence combined with the imposed condition on Λ, we find that $\Gamma(0) = 0$ and $\Gamma(1) = 1$, i.e. Γ satisfies the same identification constraints as Λ. We estimate the variable U by

$$\widehat{U} = \widehat{T}(Y) = \frac{\widehat{F}_Y(Y) - \widehat{F}_Y(0)}{\widehat{F}_Y(1) - \widehat{F}_Y(0)} \ , \tag{4}$$

where $\widehat{F}_Y(y) = n^{-1}\sum_{i=1}^n 1_{\{Y_i \leq y\}}$ is the empirical distribution function of $Y_1, \ldots, Y_n$. Next, to estimate the transformation Γ, first note that for all $x \in \chi$,

$$\Gamma(u) = \lambda_1(u,x) = \frac{S_1(u,x)}{S_1(1,x)} \quad \text{with} \quad S_1(u,x) = \int_0^u \frac{\frac{\partial}{\partial w}\varphi(w,x)}{\frac{\partial}{\partial x_1}\varphi(w,x)}\, dw, \tag{5}$$

where $\varphi(u,x) = P(U \leq u | X = x)$ is the conditional distribution of U given X, and x_1 is the first component of the vector $x = (x_1, \ldots, x_d)^t$. The proof of (5) is the same as the proof of Theorem 1 in Chiappori et al. (2015) and is therefore omitted. Hence, we can write $\Gamma(u)$ as

$$\Gamma(u) = \text{argmin}_{q_m \in \mathbb{R}} \int_\chi v(x)\ell\big(\lambda_1(u,x) - q_m\big)\, dx,$$

for any positive weight function $v(\cdot)$ and loss function $\ell(\cdot)$ satisfying $\ell(0) = 0$. In particular, we can work with the loss $\ell(u) = u\big(2L_b(u) - 1\big)$, where $L_b(\cdot) = L(\cdot/b)$, L is a given distribution function and $b > 0$ is a bandwidth sequence. This loss function is a smooth approximation of the absolute deviation loss $\ell(u) = |u|$ for b small. To estimate $\Gamma(u)$, we will replace the unknown function $\lambda_1(u,x)$ by an appropriate estimator. Let

$$\widehat{\lambda}_1(u,x) = \frac{\widehat{S}_1(u,x)}{\widehat{S}_1(1,x)} \quad \text{with} \quad \widehat{S}_1(u,x) = \int_0^u \frac{\frac{\partial}{\partial w}\widehat{\varphi}(w,x)}{\frac{\partial}{\partial x_1}\widehat{\varphi}(w,x)}\, dw,$$

where

$$\widehat{\varphi}(u,x) = \frac{\sum_{i=1}^n \mathcal{K}_{h_u}(u - \widehat{U}_i)\mathbf{K}_{h_x}(X_i - x)}{\sum_{i=1}^n \mathbf{K}_{h_x}(X_i - x)},$$

$\mathbf{K}_{h_x}(x) = \mathbf{K}(x/h_x)/h_x^d$, $\mathbf{K}(x)$ is a multivariate product kernel of the form $\mathbf{K}(x) = \prod_{i=1}^d K(x_i)$, $\mathcal{K}_{h_u} = \mathcal{K}(u/h_u)$, $\mathcal{K}(u) = \int_{-\infty}^u K(w)\,dw$, K is a univariate kernel and h_u and h_x are bandwidth sequences. Finally, define

$$\widehat{\Gamma}_{LAD,b}(u) = \mathrm{argmin}_{q_m \in \mathbb{R}} \int_{\mathcal{X}} v(x)\big(\widehat{\lambda}_1(u,x) - q_m\big)\big\{2L_b\big(\widehat{\lambda}_1(u,x) - q_m\big) - 1\big\}\,dx.$$

Consequently, a natural estimator of θ_0 is given by

$$\mathrm{argmin}_{\theta\in\Theta}\, n^{-1}\sum_{i=1}^n \Big(\widehat{\Gamma}_{LAD,b}(\widehat{T}(Y_i)) - \Lambda_\theta(Y_i)\Big)^2 \ . \tag{6}$$

However, it is important to remind that the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ has been constructed under the particular identification conditions $\Lambda(0) = 0$ and $\Lambda(1) = 1$. Certain classes of transformations do not satisfy these identification constraints. The class of Yeo and Johnson (2000) transformations, for example, satisfies $\Lambda_\theta(0) = 0$ and $\Lambda'_\theta(0) = 1$ instead of $\Lambda_\theta(1) = 1$ for all $\theta \in \Theta$. Expression (6) will then lead to an inconsistent estimator of θ. In the next section we will explain in detail how we can adjust the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ with additive and multiplicative constants so that the corresponding adjusted estimator in (6) is consistent under identification conditions that are more general than $\Lambda_\theta(0) = 0$ and $\Lambda_\theta(1) = 1$ for all $\theta \in \Theta$.

Another possibility to allow for other identification conditions would be to consider $\widehat{\Gamma}^*_{LAD,b}(\widehat{T}^*(\cdot))$ instead of $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$, where $\widehat{\Gamma}^*_{LAD,b}(\cdot)$ and $\widehat{T}^*(\cdot)$ are estimators of some suitable adaptations $\Gamma^*(\cdot)$ and $T^*(\cdot)$ of $\Gamma(\cdot)$ and $T(\cdot)$, depending on the particular identification conditions considered. However, the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ has several advantages. First, its asymptotic properties have already been developed in Colling and Van Keilegom (2019), which will facilitate the proofs in this paper. Second, we know that this estimator avoids kernel smoothing of Y, and the latter is known to work badly in practice if the distribution of Y is skewed. This is not necessarily the case for $\widehat{\Gamma}^*_{LAD,b}(\widehat{T}^*(\cdot))$. Indeed, if we consider the Yeo and Johnson (2000) transform for example, with $\Lambda_\theta(0) = 0$ and $\Lambda'_\theta(0) = 1$ as identification conditions, then $T^*(Y) = [F_Y(Y) - F_Y(0)]/f_Y(0)$, and so the estimation of $T^*(Y)$ will require kernel smoothing of Y since it depends on the density function f_Y of Y. Finally, the expression of the estimators $\widehat{\Gamma}^*_{LAD,b}(\cdot)$ and $\widehat{T}^*(\cdot)$ depends on the imposed identification conditions, whereas our goal is to construct an estimator of θ_0 that is consistent under general identification conditions.

3. Definition of the estimators

The estimators of θ that we will propose in this section, will work under the following identifiability conditions on the class $\{\Lambda_\theta : \theta \in \Theta\}$:

(I1) $\Lambda_\theta(\alpha_1) = a_1$ and $\Lambda_\theta(\alpha_2) = a_2$ for all $\theta \in \Theta$ and for some $\alpha_1 < \alpha_2$ and $a_1 < a_2$,

(I2) The set $\{x \in \chi : \frac{\partial}{\partial x_1} m(x) \neq 0\}$ has nonempty interior,

(the derivative $\frac{\partial}{\partial x_1} m(x)$ can also be replaced by $\frac{\partial}{\partial x_j} m(x)$ for some $j = 1, \dots, d$). The set of conditions (I1) can be replaced by the following alternative set of conditions:

(I1') $\Lambda_\theta(\alpha_1) = a_1$ and $\Lambda'_\theta(\alpha_3) = a_3$ for all $\theta \in \Theta$ and for some $a_1, a_3, \alpha_1, \alpha_3$.

The following proposition will be on the basis of the adjustment of the expression (6) and takes into account the set of identification conditions (I1) and (I1'). The assumptions (A1)–(A4) are given in the Appendix. Let $\mathcal{U}_0$ be a compact subset in the interior of the support $\mathcal{U}$ of U, and let $\mathcal{Y}_0$ be a compact set strictly included in $T^{-1}(\mathcal{U}_0)$.

Proposition 3.1. *Assume (A1)–(A4). Then, under either (I1) or (I1') and (I2), we have for all $x \in \chi$ and $y \in \mathcal{Y}_0$,*

$$\Lambda(y) = \big(\Lambda(1) - \Lambda(0)\big)\frac{S_1(T(y), x)}{S_1(1, x)} + \Lambda(0), \tag{7}$$

where T and S_1 are defined in (3) and (5) respectively. Moreover, the right hand side of (7) does not depend on x.

The proof of this proposition is given in Section A.2. Consequently, using Proposition 3.1 and the fact that $\widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$ is a nonparametric estimator of $S_1(T(y), x)/S_1(1, x)$, it is natural to define the following estimator of θ_0:

$$\widehat{\theta}_1 = \operatorname{argmin}_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} w(Y_i)\Big(\big(\Lambda_\theta(1) - \Lambda_\theta(0)\big)\widehat{\Gamma}_{LAD,b}(\widehat{T}(Y_i)) + \Lambda_\theta(0) - \Lambda_\theta(Y_i)\Big)^2, \tag{8}$$

where w is a certain positive weight function with support included in $\mathcal{Y}_0$, that has been added to facilitate the proofs of the main asymptotic results that will be presented in the next section. Moreover, if the transformation satisfies in particular $\Lambda_\theta(0) = 0$ and $\Lambda_\theta(1) = 1$ for all $\theta \in \Theta$, expression (8) equals expression (6) up to the weight function w that we have added in the meantime.

An alternative estimator can be obtained by letting the constants $\Lambda_\theta(1) - \Lambda_\theta(0)$ and $\Lambda_\theta(0)$ in the expression of $\widehat{\theta}_1$ be free parameters that do not depend on θ. In that way we will have $k + 2$ parameters over which we minimize our weighted L_2-distance (k being the dimension of θ) instead of just k. This could lead to a better estimator of θ. Therefore, we define a second estimator of θ_0, which is obtained by replacing $\Lambda_\theta(1) - \Lambda_\theta(0)$ and $\Lambda_\theta(0)$ respectively by constants $c_1 \in [c_{1L}, c_{1U}] \subset \mathbb{R}^+$ and $c_2 \in [c_{2L}, c_{2U}] \subset \mathbb{R}$ that do not depend on θ, i.e.

$$\widehat{\gamma}_2 = \big(\widehat{c}_1, \widehat{c}_2, \widehat{\theta}_2\big)^t = \operatorname{argmin}_{\gamma \in \Theta_\gamma} n^{-1} \sum_{i=1}^{n} w(Y_i)\Big(c_1\widehat{\Gamma}_{LAD,b}(\widehat{T}(Y_i)) + c_2 - \Lambda_\theta(Y_i)\Big)^2, \tag{9}$$

where $\gamma = (c_1, c_2, \theta)^t$ and $\Theta_\gamma = [c_{1L}, c_{1U}] \times [c_{2L}, c_{2U}] \times \Theta \subset \mathbb{R}^{k+2}$. The true but unknown value of γ equals $\gamma_0 = (\Lambda_{\theta_0}(1) - \Lambda_{\theta_0}(0), \Lambda_{\theta_0}(0), \theta_0)^t$. Note that if the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ performs well, $\widehat{c_1}$ and $\widehat{c_2}$ should be approximately equal to $\Lambda_{\widehat{\theta}_1}(1) - \Lambda_{\widehat{\theta}_1}(0)$ and $\Lambda_{\widehat{\theta}_1}(0)$ and then $\widehat{\theta_2}$ should perform similarly as $\widehat{\theta}_1$. Otherwise, $\widehat{c_1}$ and $\widehat{c_2}$ could compensate a bad estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ by taking some other values far away from $\Lambda_{\widehat{\theta}_1}(1) - \Lambda_{\widehat{\theta}_1}(0)$ and $\Lambda_{\widehat{\theta}_1}(0)$.

4. Asymptotic results

4.1. Notations and definitions

Before establishing the main asymptotic results, we need to introduce several notations. First, $\theta = (\theta_1, \dots, \theta_k)^t$, $\dot{\Lambda}_{\theta,j}(y) = \frac{\partial}{\partial \theta_j}\Lambda_\theta(y)$ for $j = 1, \dots, k$, and $\dot{\Lambda}_\theta(y) = (\dot{\Lambda}_{\theta,1}(y), \dots, \dot{\Lambda}_{\theta,k}(y))^t$ is the vector of partial derivatives of $\Lambda_\theta(y)$ with respect to the components of θ. Next, let $C_c^1(\mathcal{U}_0)$ be the set of functions $f : \mathcal{U}_0 \to \mathbb{R}$ for which $\|f\|_{\infty,1} \le c < \infty$, where

$$\|f\|_{\infty,1} = \sup_{u \in \mathcal{U}_0} |f(u)| + \sup_{u \neq u', u \in \mathcal{U}_0, u' \in \mathcal{U}_0} \frac{|f(u') - f(u)|}{|u' - u|}.$$

Moreover, we define

$$\mathcal{H} = \Big\{ h : \mathcal{Y}_0 \to \mathbb{R} : h = f \circ g, f \in C_c^1(\mathcal{U}_0), g : \mathcal{Y}_0 \to \mathcal{U}_0 \text{ is monotone} \Big\},$$

and let $\|h\|_{\mathcal{H}} = [E(h^2(Y))]^{1/2}$. Note that the derivative of $\varphi(u, x)$ with respect to a component of x is zero for all x if u is at the lower boundary of the support $\mathcal{U}$ of U. Hence, Assumptions (A4) and (A8), given in Section A.1, cannot be fulfilled for all $u \in \mathcal{U}$. Consequently, we have to work with the set $\mathcal{U}_0$, defined in Section 3, and similarly we have to work with the set $\mathcal{Y}_0$ instead of $\mathcal{Y}$.

The following notations are related to the first estimator $\widehat{\theta}_1$. First, we introduce the following four vectors of dimension k:

$$A(\theta) = \big(A_j(\theta)\big)_{j=1,\dots,k}^t \text{ with } A_j(\theta) = \big(\dot{\Lambda}_{\theta,j}(1) - \dot{\Lambda}_{\theta,j}(0)\big)\big(\Lambda_\theta(1) - \Lambda_\theta(0)\big)$$
$$B(\theta, y) = \big(B_j(\theta, y)\big)_{j=1,\dots,k}^t \text{ with } B_j(\theta, y) = \big(\dot{\Lambda}_{\theta,j}(1) - \dot{\Lambda}_{\theta,j}(0)\big)\big(\Lambda_\theta(0) - \Lambda_\theta(y)\big)$$
$$C(\theta, y) = \big(C_j(\theta, y)\big)_{j=1,\dots,k}^t \text{ with } C_j(\theta, y) = \big(\dot{\Lambda}_{\theta,j}(0) - \dot{\Lambda}_{\theta,j}(y)\big)\big(\Lambda_\theta(1) - \Lambda_\theta(0)\big)$$
$$D(\theta, y) = \big(D_j(\theta, y)\big)_{j=1,\dots,k}^t \text{ with } D_j(\theta, y) = \big(\dot{\Lambda}_{\theta,j}(0) - \dot{\Lambda}_{\theta,j}(y)\big)\big(\Lambda_\theta(0) - \Lambda_\theta(y)\big).$$

Moreover, for $h \in \mathcal{H}$ and $\theta \in \Theta$, we consider $\ell_1(y, \theta, h) = (\ell_{1,j}(y, \theta, h))_{j=1,\dots,k}^t$, where, for $j = 1, \dots, k$,

$$\ell_{1,j}(y, \theta, h) = w(y)\Big[A_j(\theta)h^2(y) + B_j(\theta, y)h(y) + C_j(\theta, y)h(y) + D_j(\theta, y)\Big]. \quad (10)$$

Finally, let $M_1(\theta, h) = E[\ell_1(Y, \theta, h)]$ and $M_{n,1}(\theta, h) = n^{-1}\sum_{i=1}^n \ell_1(Y_i, \theta, h)$, where the k components of the vectors $M_1(\theta, h)$ and $M_{n,1}(\theta, h)$ are respectively denoted by $M_{1,j}(\theta, h)$ and $M_{n,1,j}(\theta, h)$ for $j = 1, \dots, k$.

Similarly, the following notations are related to the second estimator $\widehat{\theta}_2$. For $h \in \mathcal{H}$ and $\gamma = (c_1, c_2, \theta)^t \in \Theta_\gamma$, we consider $\ell_2(y, \gamma, h) = (\ell_{2,j}(y, \gamma, h))^t_{j=1,\ldots,k+2}$, where

$$\begin{aligned} \ell_{2,1}(y,\gamma,h) &= w(y)\Big[c_1 h^2(y) + c_2 h(y) - \Lambda_\theta(y) h(y)\Big] \\ \ell_{2,2}(y,\gamma,h) &= w(y)\Big[c_1 h(y) + c_2 - \Lambda_\theta(y)\Big] \\ \ell_{2,j}(y,\gamma,h) &= -w(y)\dot{\Lambda}_{\theta,j-2}(y)\Big[c_1 h(y) + c_2 - \Lambda_\theta(y)\Big], \end{aligned}$$

for $j = 3, \ldots, k+2$. Moreover, let $M_2(\gamma, h) = E[\ell_2(Y, \gamma, h)]$ and $M_{n,2}(\gamma, h) = n^{-1}\sum_{i=1}^n \ell_2(Y_i, \gamma, h)$, where the $k+2$ components of the vectors $M_2(\gamma, h)$ and $M_{n,2}(\gamma, h)$ are respectively denoted by $M_{2,j}(\gamma, h)$ and $M_{n,2,j}(\gamma, h)$ for $j = 1, \ldots, k+2$.

Note that the vectors $M_{n,1}(\theta, \widehat{h})$ and $M_{n,2}(\gamma, \widehat{h})$, where $\widehat{h}(y) = \widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$ estimates $h_0(y) = \frac{S_1(T(y),x)}{S_1(1,x)}$, are simply the derivatives of the expressions that we minimize in (8) and (9) with respect to the components of the vectors $\theta = (\theta_1, \ldots, \theta_k)^t$ and $\gamma = (c_1, c_2, \theta_1, \ldots, \theta_k)^t$ respectively. Using Proposition 3.1, note also that $\ell_1(y, \theta_0, h_0) = 0$ and $\ell_2(y, \gamma_0, h_0) = 0$ for all $y \in \mathcal{Y}_0$, which implies that $M_{n,1}(\theta_0, h_0) = M_1(\theta_0, h_0) = 0$ and $M_{n,2}(\gamma_0, h_0) = M_2(\gamma_0, h_0) = 0$.

Finally, the reason for defining all these functions comes from the article of Chen et al. (2003). Indeed, Chen et al. (2003) proposed sufficient high-level conditions for the consistency and asymptotic normality of a class of semiparametric optimization estimators, that we will verify here for our estimators $\widehat{\theta}_1$ and $\widehat{\gamma}_2$; see the proofs of Theorems 4.1 and 4.2 in Section A.2. These sufficient conditions are mainly conditions on the class of functions $\mathcal{H}$ and either on the functions ℓ_1, M_1 and $M_{n,1}$ for the estimator $\widehat{\theta}_1$ or on the functions ℓ_2, M_2 and $M_{n,2}$ for the estimator $\widehat{\gamma}_2$.

4.2. Consistency and asymptotic normality

The following theorems establish respectively the consistency and the asymptotic normality of $\widehat{\theta}_1$ and $\widehat{\gamma}_2$. The assumptions under which these results are valid can be found in the Appendix.

Theorem 4.1. *Assume (A1)–(A13). Then, under either (I1) or (I1') and (I2),*

$$(i)\ \widehat{\theta}_1 - \theta_0 = o_P(1) \quad \textit{and} \quad (ii)\ \widehat{\gamma}_2 - \gamma_0 = o_P(1)\ .$$

Theorem 4.2. *Assume (A1)–(A13). Then, under either (I1) or (I1') and (I2),*

(i)

$$\sqrt{n}(\widehat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, \Omega_1),$$

where $\Omega_1 = \Delta_1^{-1} V_1 \Delta_1^{-1}$, $\Delta_1 = \Delta_1(\theta_0, h_0)$, $\Delta_1(\theta, h)$ is the $k \times k$ matrix of partial derivatives of $M_1(\theta, h)$ with respect to the components of θ, and *the matrix* V_1 *is given by*

$$V_1 = Var\Big(E\Big[w(Y')\varphi^v_{X,Y}(Y')\Big(2A(\theta_0)h_0(Y') + B(\theta_0, Y') + C(\theta_0, Y')\Big)\Big|X, Y\Big]\Big),$$

where Y' *is an i.i.d. copy of* Y, $h_0(y) = \frac{S_1(T(y),x)}{S_1(1,x)}$, *and* $\varphi^v_{X,Y}$ *is defined in Section A.2.*

(ii)

$$\sqrt{n}(\widehat{\gamma}_2 - \gamma_0) \xrightarrow{d} N(0, \Omega_2),$$

where $\Omega_2 = \Delta_2^{-1} V_2 \Delta_2^{-1}$, $\Delta_2 = \Delta_2(\gamma_0, h_0)$, $\Delta_2(\gamma, h)$ *is the* $(k+2) \times (k+2)$ *matrix of partial derivatives of* $M_2(\gamma, h)$ *with respect to the components of* γ, *and the matrix* V_2 *is given by*

$$V_2 = Var\begin{pmatrix} E\Big[w(Y')\varphi^v_{X,Y}(Y')\Big(2c_{1,0}h_0(Y') + c_{2,0} - \Lambda_{\theta_0}(Y')\Big)\Big|X, Y\Big] \\ E\Big[w(Y')c_{1,0}\varphi^v_{X,Y}(Y')\Big|X, Y\Big] \\ -E\Big[w(Y')c_{1,0}\dot{\Lambda}_{\theta_0,1}(Y')\varphi^v_{X,Y}(Y')\Big|X, Y\Big] \\ \vdots \\ -E\Big[w(Y')c_{1,0}\dot{\Lambda}_{\theta_0,k}(Y')\varphi^v_{X,Y}(Y')\Big|X, Y\Big] \end{pmatrix},$$

where $c_{1,0} = \Lambda_{\theta_0}(1) - \Lambda_{\theta_0}(0)$ *and* $c_{2,0} = \Lambda_{\theta_0}(0)$.

The proofs of these two theorems are given in Section A.2. Note that the covariance matrices V_1 and V_2 are derived from the pathwise derivatives of the vectors $M_1(\theta_0, h_0)$ and $M_2(\gamma_0, h_0)$ in the direction $\widehat{h} - h_0$. The exact expressions of these pathwise derivatives, as well as of their i.i.d. representations, are given in the proof of Theorem 4.2.

Note also that Theorem 4.1(ii) implies that $\widehat{\theta}_2$ is consistent for θ_0, and that Theorem 4.2(ii) implies that $\widehat{\theta}_2$ is asymptotically normally distributed with variance-covariance matrix given by the lower $k \times k$ submatrix of the matrix Ω_2.

5. Simulations

In this section, we perform simulations in order to compare the performance of our new estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of the transformation parameter with that of the profile likelihood estimator $\widehat{\theta}_{PL}$ proposed by Linton et al. (2008) and defined in (2).

The six simulated models are

Model 1: $\Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n \sim_{iid} N(0, 1)$,
Model 2: $\Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n \sim_{iid} N(0, 0.5^2)$,
Model 3: $\Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n \sim_{iid} \frac{2t_{10}}{\sqrt{5}}$,
Model 4: $\Lambda_\theta(Y_i) = 6X_i - 3 + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n \sim_{iid} N(0, 1)$,

Model 5: $\Lambda_\theta(Y_i) = 10X_{i1}^2 - 3 + 4\sin(2\pi X_{i2}) + \varepsilon_i$ where $\varepsilon_1, \dots, \varepsilon_n \sim_{iid} N(0,1)$,
Model 6: $\Lambda_\theta(Y_i) = 20X_{i1}^2 - 6 + 8\sin(2\pi X_{i2}) + \varepsilon_i$ where $\varepsilon_1, \dots, \varepsilon_n \sim_{iid} N(0,1)$.

Note that $d = 1$ in Models 1 to 4 and $d = 2$ in Models 5 and 6. In each model, $\Lambda_\theta(Y)$ represents the Yeo and Johnson (2000) transformation, i.e.,

$$\Lambda_\theta(Y) = \begin{cases} \frac{(Y+1)^\theta - 1}{\theta} & \text{if } \; Y \geq 0, \theta \neq 0 \\ \log(Y+1) & \text{if } \; Y \geq 0, \theta = 0 \\ \frac{-[(-Y+1)^{2-\theta} - 1]}{2-\theta} & \text{if } \; Y < 0, \theta \neq 2 \\ -\log(-Y+1) & \text{if } \; Y < 0, \theta = 2, \end{cases}$$

$X_1, \dots, X_n$ are independent uniform random variables on $[0,1]$ for Models 1 to 4 and (X_{11}, X_{12}), ..., (X_{n1}, X_{n2}) are independent and uniformly distributed on the unit square for Models 5 and 6. For each model we will also consider six sample sizes: $n = 50$, $n = 100$, $n = 200$, $n = 300$, $n = 400$ and $n = 500$ and four values of the transformation parameter: $\theta_0 = 0$ which corresponds to a logarithmic transformation, $\theta_0 = 0.5$ which corresponds to a square root transformation, $\theta_0 = 1$ which corresponds to the identity and $\theta_0 = 1.5$.

The goal will be to analyze the influence of the sample size n, the value of the transformation parameter θ_0, the dimension d of X (by comparing Models 1 and 4 to Models 5 and 6), the variability of the regression function $m(x)$ (by comparing Model 4 to Model 1 and Model 6 to Model 5), the variability of the error term (by comparing Model 2 to Model 1) and the distribution of the error term (by comparing Model 3 to Model 1) on the bias and variance of the different estimators. Note that, in Model 3, we consider $\varepsilon \sim \frac{2t_{10}}{\sqrt{5}}$ instead of $\varepsilon \sim t_{10}$ to ensure that $V(\varepsilon) = 1$, exactly as in Model 1. In that case, if we observe some significative difference in the performance of the estimators between Models 1 and 3, we will be sure that it comes from the distribution, and not from the variability, of the error term.

Next, exactly as in Colling and Van Keilegom (2019), we will work with the unsmoothed estimator of the transformation $\Lambda(\cdot)$. This estimator is arbitrarily close to the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ when the bandwidth b is close to zero, and is defined as follows:

$$\widehat{\Gamma}_{LAD}(\widehat{T}(y)) = \operatorname{argmin}_{q_m \in \mathbb{R}} \int_{\mathcal{X}} v(x) \left| \widehat{\lambda}_1(\widehat{T}(y), x) - q_m \right| dx.$$

The smoothed estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$ that we use in the theory is used in order to facilitate the proofs of the asymptotic properties, see Colling and Van Keilegom (2019) for more on the comparison between the two estimators. The non-smoothed estimator has in particular the advantage that it does not rely on the delicate choice of the bandwidth parameter b. Again as in Colling and Van Keilegom (2019), we approximate the expression $\widehat{\Gamma}_{LAD}(u)$ in practice by

$$\widetilde{\Gamma}_{LAD}(u) = \text{ median}\left(\widehat{\lambda}_1(u, \widetilde{x}_1), \dots, \widehat{\lambda}_1(u, \widetilde{x}_{n_x}) \right) \; ,$$

where $\widehat{\lambda}_1(u,x) = \widehat{S}_1(u,x)/\widehat{S}_1(1,x)$. For Models 1 to 4, the values $\widetilde{x}_1, \ldots, \widetilde{x}_{n_x}$ are the n_x remaining values from an original grid of $N_x = 100$ equidistant points $x_1^*, \ldots, x_{N_x}^*$ generated between $\min_{1\leq j\leq n} X_j$ and $\max_{1\leq j\leq n} X_j$, from which we remove the values x_ℓ^* for which the expression $\widehat{S}_1(u,x_\ell^*)$ diverges. This can happen if $\frac{\partial}{\partial x_1}\widehat{\varphi}(w,x_\ell^*)$ is very close to 0 for some w. Moreover, we also remove the x^*-values that are within 0.01 of the values x_ℓ^* for which $\widehat{S}_1(u,x_\ell^*)$ diverges, even if the corresponding integrals do not diverge. Note that removing some x^*-values is allowed since $\widehat{S}_1(u,x)/\widehat{S}_1(1,x)$ estimates $\Gamma(u) = S_1(u,x)/S_1(1,x)$ for all $x \in \chi$.

For Models 5 and 6, we proceed exactly in the same way except that we generate first 20 equidistant points $x_{1,1}^*, \ldots, x_{20,1}^*$ between $\min_{1\leq j\leq n} X_{j1}$ and $\max_{1\leq j\leq n} X_{j1}$ and 20 other equidistant points $x_{1,2}^*, \ldots, x_{20,2}^*$ between $\min_{1\leq j\leq n} X_{j2}$ and $\max_{1\leq j\leq n} X_{j2}$ which gives us $N_x = 400$ couples of points on the unit square. Next, we remove the couples of x^*-values according to the same rules as in dimension $d = 1$ and we can evaluate the function $\widehat{\lambda}_1(u,x)$ in the remaining x^*-values.

We consider the Epanechnikov kernel $K(x) = \frac{3}{4}(1-x^2)1_{\{|x|\leq 1\}}$ for Models 1 to 4 and the product of 2 Epanechnikov kernels for Models 5 and 6, i.e. $\mathbf{K}(x_1,x_2) = K(x_1)K(x_2) = \frac{9}{16}(1-x_1^2)(1-x_2^2)1_{\{|x_1|\leq 1\}}1_{\{|x_2|\leq 1\}}$. Moreover, for Models 1 to 4, we select the bandwidths h_x and h_u by the classical normal reference rule for kernel density estimation, i.e., $\widehat{h}_x = (40\sqrt{\pi})^{1/5}\widehat{\sigma}_x n^{-1/5}$ and $\widehat{h}_u = (40\sqrt{\pi})^{1/5}\widehat{\sigma}_u n^{-1/5}$, where $\widehat{\sigma}_x$ and $\widehat{\sigma}_u$ are the classical estimators of the standard deviation of X and U respectively. Similarly, for Models 5 and 6, the bandwidth h_x is simply selected by $\widehat{h}_x = (\widehat{h}_{x_1}, \widehat{h}_{x_2})$ where $\widehat{h}_{x_j} = (40\sqrt{\pi})^{1/5}\widehat{\sigma}_{x_j} n^{-1/5}$ and $\widehat{\sigma}_{x_j}$ is the classical estimator of the standard deviation of X_j for $j = 1, 2$. Finally, we take $v(x) = 1$ for all x such that $\widehat{S}_1(u,x)$ does not diverge.

Consequently, as the Yeo and Johnson (2000) transformation satisfies $\Lambda_\theta(0) = 0$ for all $\theta \in \Theta$, we approximate $\widehat{\theta}_1$ and $\widehat{\gamma}_2$ respectively by

$$\widetilde{\theta}_1 = \operatorname{argmin}_{\theta\in\Theta} n^{-1} \sum_{i=1}^n \Big(\Lambda_\theta(1)\widetilde{\Gamma}_{LAD}(\widehat{T}(Y_i)) - \Lambda_\theta(Y_i)\Big)^2 ,$$

and

$$\widetilde{\gamma}_2 = (\widetilde{c}_1, \widetilde{c}_2, \widetilde{\theta}_2) = \operatorname{argmin}_{\gamma\in\Theta_\gamma} n^{-1} \sum_{i=1}^n \Big(c_1\widetilde{\Gamma}_{LAD}(\widehat{T}(Y_i)) + c_2 - \Lambda_\theta(Y_i)\Big)^2 .$$

We have chosen to work with $w(Y) = 1_{\{Y\in\mathcal{Y}_0\}}$, where the compact set $\mathcal{Y}_0$ is chosen large enough such that it contains (quasi) all values in the sample.

Moreover, to compute the profile likelihood estimator $\widehat{\theta}_{PL}$ introduced in (2), we use a Nadaraya-Watson estimator to estimate $m(\cdot,\theta)$, with the same Epanechnikov kernel K as above and a bandwidth estimated by a cross-validation procedure, and we use a classical kernel density estimator to estimate $f_{\varepsilon(\theta)}(\cdot)$, with the Epanechnikov kernel K and a bandwidth estimated by the classical normal reference rule for kernel density estimation. The estimator $\widehat{\theta}_{PL}$ is then

Table 1

Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 2x - 1$ and $\varepsilon \sim N(0,1)$ (model 1).

θ_0		0	0.5	1	1.5	0	0.5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.1977	.0823	-.0109	-.0823	.1065	.0454	.0061	-.0481
	Var	.1399	.1302	.1342	.1329	.0485	.0489	.0566	.0564
.	MSE	.1789	.1370	.1343	.1396	.0599	.0509	.0567	.0587
$\widetilde{\theta}_1$	Bias	-.0706	-.0068	.0402	.0584	-.0288	.0224	.0554	.0708
	Var	.1085	.1065	.1084	.0934	.0347	.0407	.0451	.0442
	MSE	.1135	.1065	.1100	.0968	.0355	.0412	.0482	.0492
$\widetilde{\theta}_2$	Bias	-.0515	-.0280	-.0015	.0230	-.0413	-.0193	.0045	.0308
	Var	.0431	.0571	.0633	.0609	.0206	.0284	.0314	.0288
	MSE	.0458	.0579	.0633	.0614	.0223	.0288	.0314	.0297
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.0493	.0232	-.0023	-.0204	.0344	.0172	.0009	-.0225
	Var	.0176	.0223	.0228	.0209	.0129	.0157	.0161	.0158
	MSE	.0200	.0228	.0228	.0213	.0141	.0160	.0161	.0163
$\widetilde{\theta}_1$	Bias	-.0338	.0103	.0457	.0673	-.0321	.0087	.0403	.0559
	Var	.0152	.0194	.0222	.0202	.0131	.0188	.0204	.0193
	MSE	.0163	.0195	.0243	.0248	.0141	.0189	.0221	.0224
$\widetilde{\theta}_2$	Bias	-.0538	-.0319	-.0071	.0187	-.0499	-.0287	-.0040	.0220
	Var	.0086	.0117	.0133	.0123	.0081	.0114	.0122	.0113
	MSE	.0115	.0128	.0133	.0126	.0106	.0123	.0122	.0118
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.0358	.0150	.0037	-.0069	.0157	.0057	-.0038	-.0140
	Var	.0101	.0121	.0130	.0113	.0054	.0075	.0082	.0082
	MSE	.0114	.0123	.0130	.0114	.0057	.0075	.0082	.0084
$\widetilde{\theta}_1$	Bias	-.0281	.0087	.0398	.0604	-.0282	.0081	.0388	.0551
	Var	.0110	.0161	.0175	.0172	.0081	.0116	.0124	.0119
	MSE	.0117	.0162	.0191	.0208	.0089	.0116	.0139	.0150
$\widetilde{\theta}_2$	Bias	-.0428	-.0214	.0001	.0240	-.0483	-.0290	-.0059	.0187
	Var	.0070	.0099	.0104	.0094	.0053	.0078	.0084	.0077
	MSE	.0088	.0104	.0104	.0100	.0076	.0086	.0084	.0081

obtained iteratively with the function *optimize* in R. We refer to Colling and Van Keilegom (2016) for more details on the implementation of this estimator.

Tables 1 to 6 show the bias, variance and mean squared error of the profile likelihood estimator $\widehat{\theta}_{PL}$ and of our new estimators $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for all considered values of n and θ_0 when Models 1 to 6 are generated respectively, each of them obtained on the basis of 200 samples.

First, when the sample size n increases, the mean squared error of all estimators decreases, especially due to a significant decrease of their variance, which is an expected outcome. Next, we observe that $\widetilde{\theta}_2$ outperforms $\widetilde{\theta}_1$ in all scenarios in terms of variance and in most of the scenarios in terms of bias, which could be expected since $\widetilde{\theta}_2$ offers more flexibility and freedom than $\widetilde{\theta}_1$ as explained in Section 3. Hence, we will concentrate our following analysis on the comparison between $\widehat{\theta}_{PL}$ and $\widetilde{\theta}_2$.

Table 2

Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 2x - 1$ and $\varepsilon \sim N(0, 0.5^2)$ (model 2).

θ_0		0	0.5	1	1.5	0	0.5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.1286	.0476	.0038	-.0301	.0555	.0232	-.0109	-.0394
	Var	.1283	.1526	.1618	.1366	.0487	.0516	.0615	.0594
	MSE	.1448	.1549	.1618	.1375	.0518	.0521	.0616	.0610
$\widetilde{\theta}_1$	Bias	-.0254	.0327	.0622	.0555	-.0065	.0449	.0680	.0634
	Var	.1294	.1393	.1229	.1100	.0414	.0483	.0504	.0468
	MSE	.1301	.1404	.1268	.1131	.0414	.0504	.0550	.0508
$\widetilde{\theta}_2$	Bias	-.0402	-.0149	.0075	.0354	-.0392	-.0142	.0125	.0380
	Var	.0546	.0690	.0757	.0747	.0237	.0299	.0321	.0305
	MSE	.0562	.0692	.0758	.0759	.0253	.0301	.0323	.0320
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.0251	.0128	.0008	-.0209	.0216	.0066	-.0007	-.0071
	Var	.0200	.0221	.0241	.0280	.0154	.0177	.0188	.0168
	MSE	.0206	.0223	.0241	.0284	.0159	.0177	.0188	.0169
$\widetilde{\theta}_1$	Bias	-.0083	.0365	.0653	.0601	-.0131	.0322	.0583	.0544
	Var	.0179	.0236	.0252	.0245	.0152	.0199	.0205	.0196
	MSE	.0179	.0250	.0295	.0281	.0154	.0209	.0239	.0226
$\widetilde{\theta}_2$	Bias	-.0494	-.0249	.0016	.0283	-.0515	-.0283	-.0027	.0224
	Var	.0111	.0151	.0162	.0156	.0096	.0130	.0141	.0131
	MSE	.0136	.0157	.0162	.0164	.0122	.0138	.0141	.0136
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.0288	.0180	.0071	-.0010	.0137	.0029	-.0090	-.0166
	Var	.0100	.0117	.0132	.0106	.0072	.0085	.0092	.0089
	MSE	.0108	.0120	.0132	.0106	.0074	.0085	.0092	.0091
$\widetilde{\theta}_1$	Bias	-.0067	.0338	.0614	.0590	.0117	.0237	.0421	.0422
	Var	.0109	.0152	.0164	.0158	.0087	.0128	.0144	.0136
	MSE	.0109	.0164	.0201	.0193	.0088	.0134	.0162	.0154
$\widetilde{\theta}_2$	Bias	-.0429	-.0186	.0061	.0308	-.0481	-.0259	-.0034	.0199
	Var	.0067	.0088	.0096	.0090	.0058	.0080	.0088	.0082
	MSE	.0086	.0092	.0097	.0099	.0081	.0086	.0088	.0086

Next, for models 1 to 4, when θ_0 increases from 0 to 1, we observe globally that the bias of $\widehat{\theta}_{PL}$ and $\widetilde{\theta}_2$ tend to decrease in absolute value while their variance tends to increase, which leads to a general increase in their mean squared error. Conversely, if θ_0 increases from 1 to 1.5, the mean squared error of $\widehat{\theta}_{PL}$ and $\widetilde{\theta}_2$ tends to decrease due to a decrease of their variance, even if their bias tends to increase in absolute value. This suggests that the parametric transformation is more difficult to estimate when the response Y is less variable. Indeed, a logarithmic transformation $\theta_0 = 0$ will be easier to detect due to the presence of very high values in the sample $Y_1, \ldots, Y_n$ in comparison with the identity transformation $\theta_0 = 1$ for instance. For Models 5 and 6, these conclusions are the same for $\widetilde{\theta}_2$ except that both bias and variance (and not only variance) of $\widetilde{\theta}_2$ are deteriorating when θ_0 moves from 0 to 1 and are improving when θ_0 moves from 1 to 1.5. For $\widehat{\theta}_{PL}$ we remark the poor behavior of the estimator for $\theta_0 = 0$ compared to the other values of θ_0, both in terms of variance and in terms of bias.

Table 3

Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 2x - 1$ and $\varepsilon \sim \frac{2t_{10}}{\sqrt{5}}$ (model 3).

θ_0		0	0.5	1	1.5	0	0..5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.2278	.1065	-.0045	-.1223	.1163	.0381	-.0236	-.0825
	Var	.1523	.1713	.2025	.1699	.0777	.0800	.0941	.1003
	MSE	.2042	.1827	.2026	.1849	.0912	.0815	.0947	.1071
$\widetilde{\theta}_1$	Bias	-.0644	-.0010	.0389	.0644	-.0555	.0038	.0473	.0655
	Var	.0877	.0858	.0927	.0850	.0353	.0443	.0479	.0452
	MSE	.0918	.0858	.0943	.0891	.0383	.0443	.0501	.0495
$\widetilde{\theta}_2$	Bias	-.0324	-.0041	.0247	.0499	-.0593	-.0320	.0007	.0289
	Var	.0431	.0593	.0655	.0638	.0228	.0328	.0376	.0355
	MSE	.0442	.0593	.0662	.0663	.0264	.0339	.0376	.0363
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.0939	.0450	-.0007	-.0520	.0687	.0408	.0104	-.0232
	Var	.0352	.0412	.0445	.0421	.0208	.0299	.0321	.0304
	MSE	.0440	.0433	.0445	.0448	.0256	.0315	.0322	.0309
$\widetilde{\theta}_1$	Bias	-.0368	.0120	.0482	.0747	-.0328	.0170	.0590	-.0803
	Var	.0174	.0246	.0283	.0261	.0098	.0152	.0177	.0185
	MSE	.0188	.0247	.0306	.0316	.0109	.0155	.0212	.0250
$\widetilde{\theta}_2$	Bias	-.0553	-.0263	.0064	.0368	-.0555	-.0273	.0045	.0352
	Var	.0131	.0197	.0222	.0201	.0084	.0132	.0155	.0151
	MSE	.0162	.0204	.0222	.0214	.0115	.0140	.0155	.0163
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.0387	.0163	-.0081	-.0293	.0560	.0264	.0126	-.0076
	Var	.0138	.0167	.0184	.0168	.0117	.0120	.0123	.0110
	MSE	.0153	.0169	.0185	.0177	.0148	.0127	.0124	.0110
$\widetilde{\theta}_1$	Bias	-.0315	.0162	.0542	.0750	-.0275	.0201	.0553	.0758
	Var	.0077	.0115	.0132	.0123	.0059	.0088	.0103	.0101
	MSE	.0087	.0118	.0161	.0180	.0066	.0092	.0133	.0158
$\widetilde{\theta}_2$	Bias	-.0554	-.0262	.0600	.0364	-.0542	-.0279	.0023	.0334
	Var	.0067	.0100	.0110	.0099	.0050	.0075	.0085	.0079
	MSE	.0098	.0107	.0111	.0112	.0080	.0083	.0085	.0090

Using the same reasoning, it seems logical that $\widehat{\theta}_{PL}$ and $\widetilde{\theta}_2$ perform globally both better under Model 4 ($m(x) = 6x - 3$) than under Model 1 ($m(x) = 2x - 1$), and under Model 6 ($m(x) = 20x_1^2 - 6 + 8\sin(2\pi x_2)$) than under Model 5 ($m(x) = 10x_1^2 - 3 + 4\sin(2\pi x_2)$), in terms of bias and variance. Indeed, in Models 4 and 6, the regression functions $m(x)$ are more variable than in Models 1 and 5 respectively, which helps for estimating θ. Similarly, if we compare the results obtained under Model 1 to the ones obtained under Model 2, $\widehat{\theta}_{PL}$ and $\widetilde{\theta}_2$ perform both better when $\varepsilon \sim N(0, 1)$ than when $\varepsilon \sim N(0, 0.5^2)$, especially in terms of variance while the results seem globally comparable in terms of bias. Consequently, a more variable error term also helps for estimating θ.

Next, if we compare Models 1 and 4 to Models 5 and 6, it is clear that a model with $d = 2$ is more difficult to estimate than a model with $d = 1$ especially for small sample sizes. Indeed, the bias and the variance of $\widehat{\theta}_{PL}$ are really poor under Models 5 and 6 for $n = 50$ and $n = 100$ and especially for $\theta_0 = 0$. However, even

Table 4
Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 6x - 3$ and $\varepsilon \sim N(0, 1)$ (model 4).

θ_0		0	0.5	1	1.5	0	0.5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.0714	.0646	.0469	.0044	.0221	.0154	.0083	-.0043
	Var	.0434	.0680	.0768	.0473	.0106	.0131	.0141	.0117
	MSE	.0485	.0721	.0790	.0473	.0111	.0134	.0141	.0117
$\widetilde{\theta}_1$	Bias	-.1426	-.0933	-.0486	.0106	-.0759	-.0515	-.0209	.0037
	Var	.1586	.1125	.0814	.0443	.0580	.0362	.0250	.0204
	MSE	.1789	.1212	.0838	.0444	.0638	.0388	.0254	.0204
$\widetilde{\theta}_2$	Bias	-.0202	-.0074	.0077	.0221	-.0165	-.0034	.0109	.0231
	Var	.0098	.0155	.0173	.0153	.0042	.0069	.0078	.0067
	MSE	.0102	.0156	.0174	.0158	.0045	.0069	.0079	.0073
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.0038	.0004	-.0047	-.0112	.0022	.0006	.0010	-.0009
	Var	.0030	.0034	.0047	.0100	.0015	.0028	.0040	.0027
	MSE	.0031	.0034	.0047	.0102	.0015	.0028	.0040	.0027
$\widetilde{\theta}_1$	Bias	-.0409	-.0215	-.0054	.0132	-.0264	-.0171	.0017	.0184
	Var	.0163	.0118	.0113	.0092	.0043	.0077	.0089	.0069
	MSE	.0179	.0122	.0113	.0094	.0050	.0080	.0089	.0073
$\widetilde{\theta}_2$	Bias	-.0188	-.0063	.0065	.0202	-.0211	-.0094	.0043	.0175
	Var	.0015	.0026	.0031	.0030	.0014	.0024	.0027	.0024
	MSE	.0019	.0027	.0032	.0034	.0019	.0025	.0027	.0027
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.0044	.0077	.0040	.0025	-.0008	-.0023	-.0023	-.0022
	Var	.0010	.0031	.0021	.0018	.0007	.0013	.0014	.0012
	MSE	.0010	.0031	.0022	.0019	.0007	.0013	.0014	.0012
$\widetilde{\theta}_1$	Bias	-.0104	.0036	.0206	.0324	-.0218	-.0121	.0057	.0188
	Var	.0018	.0036	.0045	.0044	.0016	.0039	.0049	.0039
	MSE	.0019	.0036	.0050	.0054	.0021	.0041	.0049	.0043
$\widetilde{\theta}_2$	Bias	-.0167	-.0045	.0096	.0227	-.0236	-.0132	-.0004	.0124
	Var	.0014	.0024	.0026	.0023	.0009	.0016	.0019	.0017
	MSE	.0017	.0024	.0027	.0028	.0015	.0018	.0019	.0019

if the bias of $\widetilde{\theta}_2$ is poor under Models 5 and 6, this estimator performs globally well due to its small variance even for small sample sizes.

Moreover, if we compare the results obtained under Model 1 to the ones obtained under Model 3, it is clear that all the estimators perform better in terms of bias and variance when $\varepsilon \sim N(0, 1)$ than when $\varepsilon \sim \frac{2t_{10}}{\sqrt{5}}$ for all considered values of n and θ_0. Consequently, the distribution of the residuals also has an impact on the quality of the estimations of θ_0, which is again an expected conclusion. In particular, residuals that are normally distributed help for estimating θ_0. However, when $\varepsilon \sim \frac{2t_{10}}{\sqrt{5}}$, the estimator $\widetilde{\theta}_2$ clearly outperforms the profile likelihood estimator $\widehat{\theta}_{PL}$. Indeed, $\widehat{\theta}_{PL}$ suffers considerably under Model 3.

Despite the fact that $\widehat{\theta}_{PL}$ slightly outperforms $\widetilde{\theta}_2$ when $n = 500$ in Models 1 and 2 and when $n = 400$ and $n = 500$ in Models 4 to 6 (except for $\theta_0 = 0$ in Models 5 and 6), $\widetilde{\theta}_2$ clearly outperforms $\widehat{\theta}_{PL}$ under Model 3 and under Models 1,

TABLE 5

Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 10x_1^2 - 3 + 4\sin(2\pi x_2)$ and $\varepsilon \sim N(0, 1)$ (model 5).

θ_0		0	0.5	1	1.5	0	0.5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.3006	.0873	-.0503	-.1147	.2228	.0392	-.0064	-.0388
	Var	.3238	.2395	.1879	.1996	.1904	.0689	.0304	.0474
	MSE	.4141	.2471	.1904	.2127	.2400	.0704	.0305	.0489
$\widetilde{\theta}_1$	Bias	-.2278	-.1762	-.1292	-.0776	-.1411	-.1044	-.0800	-.0414
	Var	.1118	.0694	.0473	.0280	.0728	.0221	.0143	.0086
	MSE	.1637	.1004	.0640	.0340	.0927	.0330	.0206	.0104
$\widetilde{\theta}_2$	Bias	-.0769	-.1080	-.1079	-.0883	-.0678	-.0998	-.1013	-.0844
	Var	.0027	.0064	.0079	.0068	.0009	.0024	.0028	.0023
	MSE	.0086	.0180	.0196	.0146	.0055	.0123	.0131	.0094
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.1856	.0260	-.0021	-.0102	.1524	.0166	.0030	-.0007
	Var	.1406	.0366	.0175	.0150	.1142	.0173	.0032	.0035
	MSE	.1751	.0373	.0175	.0151	.1375	.0176	.0032	.0035
$\widetilde{\theta}_1$	Bias	-.0693	-.0942	-.0805	-.0515	-.0657	-.0914	-.0768	-.0475
	Var	.0062	.0042	.0038	.0038	.0067	.0048	.0036	.0032
	MSE	.0110	.0130	.0103	.0064	.0110	.0132	.0095	.0055
$\widetilde{\theta}_2$	Bias	-.0688	-.1071	-.1130	-.0985	-.0646	-.1004	-.1061	-.0921
	Var	.0005	.0012	.0014	.0012	.0004	.0010	.0012	.0010
	MSE	.0052	.0127	.0142	.0109	.0046	.0111	.0125	.0095
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.1315	.0028	-.0003	-.0021	.0997	-.0008	-.0010	-.0019
	Var	.1020	.0018	.0020	.0018	.0794	.0011	.0017	.0013
	MSE	.1193	.0018	.0020	.0018	.0893	.0011	.0017	.0013
$\widetilde{\theta}_1$	Bias	-.0581	-.0846	-.0726	-.0452	-.0575	-.0856	-.0728	-.0465
	Var	.0008	.0016	.0020	.0019	.0006	.0015	.0019	.0018
	MSE	.0041	.0088	.0072	.0040	.0039	.0088	.0072	.0040
$\widetilde{\theta}_2$	Bias	-.0638	-.1001	-.1065	-.0929	-.0632	-.1004	-.1070	-.0945
	Var	.0003	.0008	.0011	.0010	.0003	.0008	.0010	.0008
	MSE	.0044	.0108	.0124	.0096	.0043	.0109	.0124	.0098

2, 4, 5, 6 for all values of θ_0 and especially for small sample sizes. This suggests that the profile likelihood estimator of Linton et al. (2008) suffers more than our new estimator when the model becomes more difficult to estimate.

Finally, we also study the convergence of the optimization algorithms used to compute the three estimators. Figure 1 shows that for Model 1 and for $n = 50$, 200 iterations is sometimes a bit too small to attain convergence. Moreover, our estimator $\widetilde{\theta}_2$ is the one for which the MSE converges the fastest, which gives this estimator an additional advantage with respect to the two other estimators.

In conclusion, the new estimator $\widetilde{\theta}_2$ outperforms globally speaking the estimators $\widetilde{\theta}_1$ and $\widehat{\theta}_{PL}$, and especially when the model becomes more difficult to estimate (Models 1 to 3 and Models 5 and 6 for smaller sample sizes). In the latter case, the performance of the profile likelihood estimator drops significantly and $\widetilde{\theta}_1$ also outperforms $\widehat{\theta}_{PL}$.

Table 6

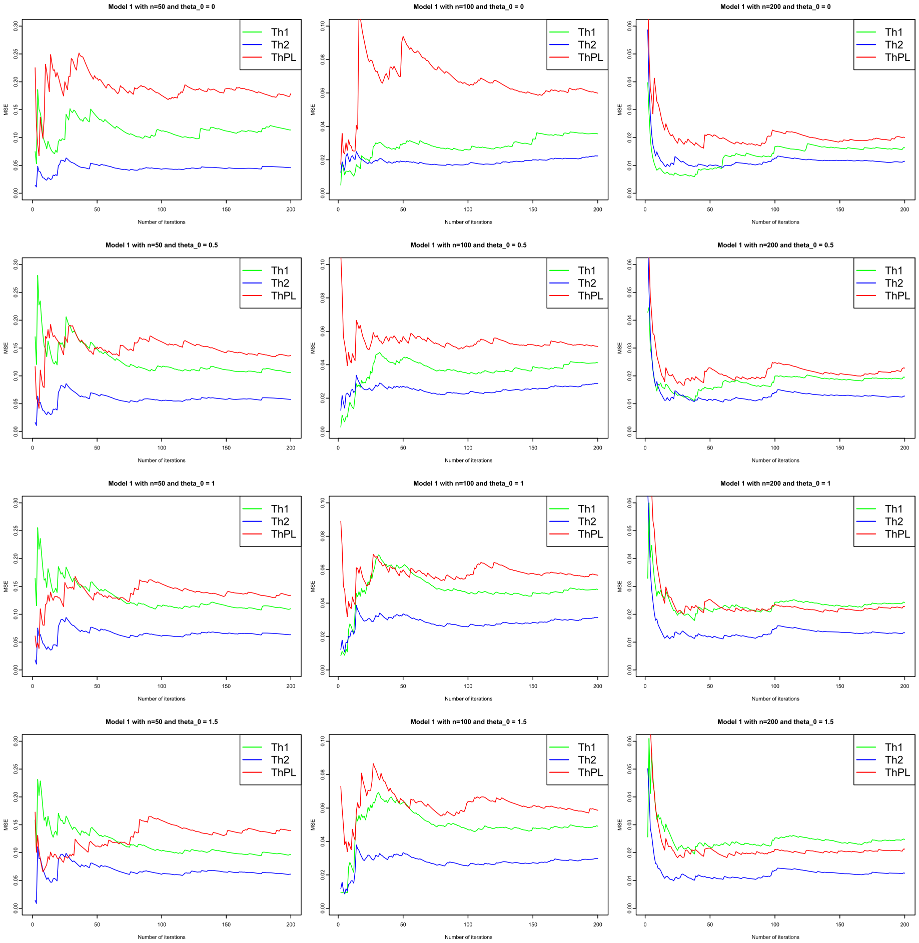
Bias, variance (Var) and mean squared error (MSE) of the estimators $\widehat{\theta}_{PL}$, $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ for different sample sizes and values of θ_0 when $m(x) = 20x_1^2 - 6 + 8\sin(2\pi x_2)$ and $\varepsilon \sim N(0,1)$ (model 6).

θ_0		0	0.5	1	1.5	0	0.5	1	1.5
		$n = 50$				$n = 100$			
$\widehat{\theta}_{PL}$	Bias	.2667	.1121	-.0597	-.2190	.2060	.0493	-.0281	-.0721
	Var	.1825	.2751	.1673	.2904	.1037	.0714	.0319	.0591
	MSE	.2536	.2877	.1709	.3384	.1461	.0739	.0327	.0643
$\widetilde{\theta}_1$	Bias	-.1411	-.1436	-.0927	-.0759	-.1210	-.1145	-.0816	-.0384
	Var	.0825	.0429	.0275	.0196	.0587	.0371	.0171	.0078
	MSE	.1024	.0636	.0361	.0254	.0733	.0502	.0237	.0093
$\widetilde{\theta}_2$	Bias	-.0477	-.0876	-.0867	-.0680	-.0408	-.0798	-.0836	-.0683
	Var	.0008	.0034	.0041	.0032	.0003	.0013	.0017	.0013
	MSE	.0031	.0110	.0116	.0079	.0020	.0077	.0087	.0059
		$n = 200$				$n = 300$			
$\widehat{\theta}_{PL}$	Bias	.1951	.0053	-.0204	-.0539	.1242	.0174	-.0089	-.0072
	Var	.0671	.0355	.0086	.0421	.0518	.0178	.0023	.0018
	MSE	.1052	.0356	.0090	.0450	.0673	.0180	.0024	.0019
$\widetilde{\theta}_1$	Bias	-.1097	-.0997	-.0818	-.0440	-.0679	-.0653	-.0593	-.0300
	Var	.0517	.0206	.0111	.0043	.0319	.0055	.0032	.0020
	MSE	.0638	.0305	.0178	.0062	.0366	.0097	.0067	.0029
$\widetilde{\theta}_2$	Bias	-.0379	-.0794	-.0867	-.0745	-.0343	-.0716	-.0786	-.0669
	Var	.0001	.0006	.0008	.0006	.0001	.0005	.0007	.0005
	MSE	.0016	.0070	.0083	.0062	.0013	.0057	.0068	.0050
		$n = 400$				$n = 500$			
$\widehat{\theta}_{PL}$	Bias	.1036	-.0064	-.0087	-.0087	.0817	-.0019	-.0087	-.0056
	Var	.0371	.0015	.0017	.0014	.0316	.0012	.0013	.0008
	MSE	.0479	.0016	.0018	.0015	.0382	.0012	.0014	.0008
$\widetilde{\theta}_1$	Bias	-.0402	-.0618	-.0595	-.0342	-.0393	-.0604	-.0533	-.0313
	Var	.0085	.0027	.0024	.0019	.0077	.0025	.0010	.0010
	MSE	.0101	.0065	.0060	.0030	.0093	.0062	.0038	.0019
$\widetilde{\theta}_2$	Bias	-.0333	-.0705	-.0780	-.0674	-.0319	-.0682	-.0756	-.0652
	Var	.0001	.0005	.0006	.0005	.0001	.0004	.0005	.0004
	MSE	.0012	.0054	.0067	.0051	.0011	.0051	.0063	.0047

6. Conclusions

In this paper we proposed two new estimators of the transformation parameter θ in a transformation model of the form $\Lambda_\theta(Y) = m(X) + \varepsilon$, where ε and X are independent and $m(\cdot)$ is completely unspecified. We showed the asymptotic normality of both estimators. Intensive simulations showed that the second proposed estimator ($\widetilde{\theta}_2$) outperforms in general the first one ($\widetilde{\theta}_1$), and in many cases both outperform the estimator of Linton et al. (2008) (denoted by $\widehat{\theta}_{PL}$).

Although the simulations showed that our estimators $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ perform well in practice, they also have a few drawbacks. First of all, they are quite computer intensive compared to the estimator $\widehat{\theta}_{PL}$. This is not surprising since our estimators rely on the nonparametric estimator of the transformation in Colling and Van Keilegom (2019), and the latter estimator is quite computer intensive. Another disadvantage of our estimators is that for dimensions d equal to 3,



Model 1 with n=50 and theta_0 = 0

Model 1 with n=100 and theta_0 = 0

Model 1 with n=200 and theta_0 = 0

Model 1 with n=50 and theta_0 = 0.5

Model 1 with n=100 and theta_0 = 0.5

Model 1 with n=200 and theta_0 = 0.5

Model 1 with n=50 and theta_0 = 1

Model 1 with n=100 and theta_0 = 1

Model 1 with n=200 and theta_0 = 1

Model 1 with n=50 and theta_0 = 1.5

Model 1 with n=100 and theta_0 = 1.5

Model 1 with n=200 and theta_0 = 1.5

FIG 1. *Convergence results for Model 1 for the estimators $\widetilde{\theta}_1$, $\widetilde{\theta}_2$ and $\widehat{\theta}_{PL}$, denoted by Th1, Th2 and ThPL.*

4 or larger, the estimators $\widetilde{\theta}_1$ and $\widetilde{\theta}_2$ will suffer in some degree from curse-of-dimensionality problems. This is because the nonparametric estimator of Colling and Van Keilegom (2019), on which they rely, is based on kernel smoothing of the covariate vector. This curse-of-dimensionality problem is however also present for the estimator $\widehat{\theta}_{PL}$ of Linton et al. (2008). A possible solution is to use (semi)-parametric estimators for $m(\cdot)$ in that case.

Appendix A: Proofs

A.1. Assumptions

The following conditions are needed for the main results of this paper. They are related to the distribution of ε, the transformations Γ and Λ_θ, the regression function m, the kernels K and L, the bandwidths h_x, h_u and b, the joint density function of U and X, the weight functions v and w, the functions M_1 and M_2, and the matrices Δ_1 and Δ_2, defined in the statement of Theorem 4.2.

(A1) The distribution function F_ε of ε is absolutely continuous and has a density f_ε that is continuous on its support. Moreover, X and ε are independent and the support $\mathcal{Y}$ of Y is a connected subset of $\mathbb{R}$.

(A2) The transformation Γ is strictly increasing and twice continuously differentiable on $\mathcal{U}_0$, where $\mathcal{U}_0$ is a compact subset in the interior of $\mathcal{U}$.

(A3) The regression function m is continuously differentiable.

(A4) The set $\mathcal{A}_1 = \{x \in \chi : \frac{\partial}{\partial x_1}\varphi(u,x) \neq 0 \ \forall u \in \mathcal{U}_0\}$ is nonempty.

(A5) The kernel K is symmetric, has support $[-1,1]$, $K(-1) = K(1) = 0$, K is of order s, i.e. $\int K(z)\,dz = 1$, $\int z^\ell K(z)\,dz = 0$ for $\ell = 1,\ldots,s-1$ and $\int z^s K(z)\,dz < \infty$, K is s-times continuously differentiable and K and K' are of bounded variation. Moreover, L is a twice continuously differentiable distribution function with uniformly bounded derivatives and with median at 0.

(A6) The bandwidths h_x, h_u and b satisfy $\sqrt{n}\max(h_x,h_u)^s \to 0$, $nb^4 \to \infty$ and $b\sqrt{n}h_x^d \min(h_x,h_u)^2(\log n)^{-1} \to \infty$.

(A7) The joint density function $f_{Y,X}$ of (Y,X) is uniformly bounded and $s+2$-times continuously differentiable on $\mathcal{Y}_0 \times \chi_0$, where $\chi_0 \subseteq \mathcal{A}_1$ is the compact support of the weight function $v(x)$ defined in (A9). We also assume that $\inf_{y:T(y)\in\mathcal{U}_0} f_Y(y) > 0$, where f_Y is the density function of Y.

(A8) $\inf_{x\in\chi_0} f_X(x) > 0$, $\inf_{(u,x)\in\mathcal{U}_0\times\chi_0} |\frac{\partial}{\partial x_1}\varphi(u,x)| > 0$ and $\inf_{x\in\chi_0} |S_1(1,x)| > 0$.

(A9) The weight function v has compact support $\chi_0 \subseteq \mathcal{A}_1$ with nonempty interior and satisfies $\int_{\chi_0} v(x)\,dx = 1$. Moreover, v is continuous on χ and is s-times continuously differentiable on χ_0.

(A10) The weight function w is positive, has support included in $\mathcal{Y}_0$ and satisfies $\sup_{y\in\mathcal{Y}_0} w(y) < \infty$.

(A11) The transformation Λ_θ is twice continuously differentiable in θ and $\Lambda_\theta(y) = \Lambda_{\theta'}(y)$ for all $y \in \mathcal{Y}_0$ implies that $\theta = \theta'$. Moreover,

$$E\left(\left[\sup_{\theta\in\Theta}\left|\left|\frac{\partial^l}{\partial\theta^l}\Lambda_\theta(Y)\right|\right|\right]^4\right) < \infty \ ,$$

for $l \in \{0,1,2\}$ and

$$E\left(\left[\sup_{\theta\in\Theta}\left|\left|\frac{\partial^i}{\partial\theta^i}\Lambda_\theta(Y)\left(\frac{\partial^j}{\partial\theta^j}\Lambda_\theta(Y)\right)^t\right|\right|\right]^4\right) < \infty \ ,$$

for $i,j \in \{0,1,2\}$ such that $0 \le i+j \le 2$.

(A12) For all $\delta_1 > 0$, there exists $\epsilon_1(\delta_1)$ such that $\inf_{||\theta-\theta_0||>\delta_1} ||M_1(\theta,h_0)|| \ge \epsilon_1(\delta_1) > 0$. Similarly, for all $\delta_2 > 0$, there exists $\epsilon_2(\delta_2)$ such that $\inf_{||\gamma-\gamma_0||>\delta_2} ||M_2(\gamma,h_0)|| \ge \epsilon_2(\delta_2) > 0$.

(A13) The matrices Δ_1 and Δ_2 are of full rank.

Assumptions (A1)–(A4) are standard and weak regularity conditions on the functions F_ε, Γ, m and φ. Assumptions (A5) and (A6) imply that s has to be strictly larger than $d+2$ (since $nh_x^{2s} \to 0$ and $nh_x^{2d+4} \to \infty$), and hence higher order kernels are required. Assumption (A7) can be expressed in terms of the joint density $f_{U,X}$ of (U,X) by noting that

$$f_{U,X}(u,x) = \frac{f_{Y,X}(T^{-1}(u),x)}{f_Y(T^{-1}(u))}\left(F_Y(1) - F_Y(0)\right).$$

Assumption (A8) is a technical assumption, needed to control denominators in various expansions used in the proofs. As for assumption (A9), note that it implies that $\int_\chi v(x)dx = 1$ and that $v(x) = 0$ for values of x at the boundary of χ, which will be needed in the proofs. Note that assumptions (A1)–(A9) are basically the same as in Colling and Van Keilegom (2019) and are required since our proofs rely on the weak convergence of the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$ that is established in the latter paper. We refer to the latter paper for a more detailed discussion of these assumptions. Assumptions (A10)–(A11) are technical conditions that are related to the fact that we have to restrict to the compact subset $\mathcal{Y}_0$ of $\mathcal{Y}$. Finally, assumptions (A12)–(A13) are required for the application of Theorems 1 and 2 in Chen et al. (2003).

A.2. Proofs of the main results

In this section, we will prove the three main results of this paper. The first one justifies the definitions of our new estimators of θ_0 and the second and the third results establish respectively the consistency and the asymptotic distributions of these new estimators.

Before giving these three proofs, let us first define the function $\varphi^v_{X,Y}$ that was introduced in the statement of Theorem 4.2. Define

$$\varphi^v_{X,Y}(y) = \delta^v_{X,Y}(T(y)) + \frac{\Gamma'(T(y))}{F_Y(1)-F_Y(0)}\Bigg(1_{\{Y\le y\}} - 1_{\{Y\le 0\}} - F_Y(y) + F_Y(0) \\ - T(y)\Big[1_{\{0\le Y\le 1\}} - F_Y(1) + F_Y(0)\Big]\Bigg), \qquad (11)$$

where $\delta^v_{X,Y}(u) = \delta^{v_1}_{X,Y}(1,u) - \delta^{v_2}_{X,Y}(u,1)$, $v_1(u_0,x) = v(x)/S_1(u_0,x)$, $v_2(u_0,x) = v(x)S_1(u_0,x)/S_1^2(1,x)$ and

$$\begin{aligned}
&\delta^{\widetilde{v}}_{X,Y}(u_0,u) \\
&= \int_{\max(0,U)}^{u}\left\{\widetilde{v}(u_0,X)D_{p,0}(w,X) - \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,x)D_{p,1}(w,x)\Big]\Big|_{x=X}\right\}dw \\
&\quad + \int_0^u\left\{\widetilde{v}(u_0,X)D_{f,0}(w,X) - \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,x)D_{f,1}(w,x)\Big]\Big|_{x=X}\right\}dw \\
&\quad + (1_{\{U\le u\}} - 1_{\{U\le 0\}})\widetilde{v}(u_0,X)D_{p,u}(U,X)
\end{aligned}$$

$$
\begin{aligned}
&+\int_0^u \left[\frac{1_{\{U\le w\}}-1_{\{U\le 0\}}}{F_Y(1)-F_Y(0)}-w\right]\int_\chi \Bigg(\Bigg\{\widetilde{v}(u_0,x)D_{p,0}(w,x)\\
&\quad+\frac{\partial}{\partial x_1}[\widetilde{v}(u_0,x)D_{p,1}(w,x)]\Bigg\}f_{U,X}(w,x)\\
&\quad+\widetilde{v}(u_0,x)D_{p,u}(w,x)\frac{\partial}{\partial w}f_{U,X}(w,x)\Bigg)\,dx\,dw\\
&-\left(\frac{1_{\{Y\le 1\}}-1_{\{Y\le 0\}}}{F_Y(1)-F_Y(0)}-1\right)\int_0^u w\int_\chi \Bigg\{\widetilde{v}(u_0,x)D_{p,0}(w,x)\\
&\quad-\widetilde{v}(u_0,x)\frac{\partial}{\partial w}D_{p,u}(w,x)+\frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,x)D_{p,1}(u,x)\Big]\Bigg\}f_{U,X}(w,x)\,dx\,dw\\
&-\left(\frac{1_{\{Y\le 1\}}-1_{\{Y\le 0\}}}{F_Y(1)-F_Y(0)}-1\right)u\int_{\mathcal{X}}\widetilde{v}(u_0,x)D_{p,u}(u,x)f_{U,X}(u,x)\,dx, \qquad (12)
\end{aligned}
$$

where $u_0\in\mathcal{U}$, $\widetilde{v}$ equals either v_1 or v_2, $f_{U,X}$ is the joint density of U and X, and

$$
D_{p,0}(u,x)=\frac{\varphi_u(u,x)f_{X,1}(x)}{\varphi_1^2(u,x)f_X^2(x)},\qquad D_{p,u}(u,x)=\frac{1}{f_X(x)\varphi_1(u,x)},
$$

$$
D_{p,1}(u,x)=\frac{-\varphi_u(u,x)}{f_X(x)\varphi_1^2(u,x)},\qquad D_{f,0}(u,x)=\frac{-\varphi_u(u,x)\varphi(u,x)f_{X,1}(x)}{\varphi_1^2(u,x)f_X^2(x)},
$$

and

$$
D_{f,1}(u,x)=\frac{\varphi_u(u,x)\varphi(u,x)}{\varphi_1^2(u,x)f_X(x)},
$$

with $\varphi_1(u,x)=\frac{\partial}{\partial x_1}\varphi(u,x)$, $\varphi_u(u,x)=\frac{\partial}{\partial u}\varphi(u,x)$, and $f_{X,1}(x)=\frac{\partial}{\partial x_1}f_X(x)$.

Proof of Proposition 3.1. The proof consists mainly in rewriting the function $S_1(u,x)$. First, note that for $u\in\mathcal{U}_0$, the conditional distribution of U given X can be rewritten as

$$
\begin{aligned}
\varphi(u,x) &= P\Big(\Gamma(U)\le\Gamma(u)\Big|X=x\Big)\\
&= P\Big(m(X)+\varepsilon\le\Gamma(u)\Big|X=x\Big)=F_\varepsilon\Big(\Gamma(u)-m(x)\Big),
\end{aligned}
$$

since $\Gamma(u)=\Lambda(T^{-1}(u))$ is strictly increasing for $u\in T(\mathcal{Y}_0)$, $\Gamma(U)=\Lambda(Y)$ and X and ε are independent. Hence,

$$
\frac{\partial}{\partial u}\varphi(u,x)=\Gamma'(u)f_\varepsilon\Big(\Gamma(u)-m(x)\Big),\ \frac{\partial}{\partial x_1}\varphi(u,x)=-\frac{\partial}{\partial x_1}m(x)f_\varepsilon\Big(\Gamma(u)-m(x)\Big),
$$

and

$$
S_1(u,x)=\int_0^u\frac{\Gamma'(w)}{-\frac{\partial}{\partial x_1}m(x)}\,dw=\frac{\Gamma(u)-\Gamma(0)}{-\frac{\partial}{\partial x_1}m(x)}\ .
$$

Consequently, $S_1(T(y),x)/S_1(1,x)$ is independent of x for $y\in\mathcal{Y}_0$, and is equal to

$$
\frac{S_1(T(y),x)}{S_1(1,x)}=\frac{\Gamma(T(y))-\Gamma(0)}{\Gamma(1)-\Gamma(0)}=\frac{\Lambda(y)-\Lambda(0)}{\Lambda(1)-\Lambda(0)}\ ,
$$

since $T(0) = 0$, $T(1) = 1$ and $\Gamma(T(y)) = \Lambda(y)$. This concludes the proof. □

Before proving the two main asymptotic results of this paper, we need to consider a technical lemma regarding the estimator $\widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$ and regarding the bracketing number $N_{[]}(\epsilon, \mathcal{H}, ||\cdot||_{L_2})$ of the space $\mathcal{H}$ defined in Section 4.1, i.e. the smallest number of ϵ-brackets needed to cover the space $\mathcal{H}$ with respect to the norm $||h||_{L_2} = [E(h^2(Y))]^{1/2}$.

Proposition A.1. *Assume (A1)–(A9). Then,*

(i) The function $\mathcal{Y}_0 \to \mathbb{R} : y \to \widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$ *belongs to* $\mathcal{H}$ *with probability tending to 1.*

(ii) $\log N_{[]}(\epsilon, \mathcal{H}, ||\cdot||_{L_2}) \leq K\epsilon^{-1}$ *for some* $K < \infty$.

Proof. First, we will prove (i). It is clear that $\widehat{T}$ is monotone and that $\widehat{T}(y) \in \mathcal{U}_0$ for n large and for $y \in \mathcal{Y}_0$, since $\mathcal{Y}_0$ is a compact set that is strictly included in $T^{-1}(\mathcal{U}_0)$ and since $\sup_y |\widehat{T}(y) - T(y)| = o_P(1)$. Hence, it suffices to show that $\widehat{\Gamma}_{LAD,b} \in C_c^1(\mathcal{U}_0)$ with probability tending to 1. Since $\sup_{u\in\mathcal{U}_0} |\Gamma(u)| < \infty$ and $\sup_{u\in\mathcal{U}_0} |\Gamma'(u)| < \infty$ thanks to assumption (A2), the result follows if we can show that

$$\sup_{u\in\mathcal{U}_0} |\widehat{\Gamma}_{LAD,b}(u) - \Gamma(u)| = o_P(1) \quad \text{and} \quad \sup_{u\in\mathcal{U}_0} |\widehat{\Gamma}'_{LAD,b}(u) - \Gamma'(u)| = o_P(1).$$

The former follows from Theorem 5.2 in Colling and Van Keilegom (2019), which establishes the weak convergence of the process $\sqrt{n}\big(\widehat{\Gamma}_{LAD,b}(\cdot) - \Gamma(\cdot)\big)$ as a process defined on $\mathcal{U}_0$. For the latter result, define

$$R_n(q_m, u) = \int_{\chi} v(x)\big(\widehat{\lambda}_1(u,x) - q_m\big)\big\{2L_b\big(\widehat{\lambda}_1(u,x) - q_m\big) - 1\big\}\,dx,$$

and note that by construction $Q_n(\widehat{\Gamma}_{LAD,b}(u), u) = 0$ for all $u \in \mathcal{U}_0$, where $Q_n(q_m, u) = \frac{\partial}{\partial q_m} R_n(q_m, u)$. Hence, the derivative of $Q_n(\widehat{\Gamma}_{LAD,b}(u), u)$ with respect to u is also equal to 0, i.e.

$$P_{1n}\big(\widehat{\Gamma}_{LAD,b}(u), u\big) + P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u), u\big)\widehat{\Gamma}'_{LAD,b}(u) = 0,$$

where $P_{1n}(q_m, u) = \frac{\partial}{\partial u} Q_n(q_m, u)$ and $P_{2n}(q_m, u) = \frac{\partial}{\partial q_m} Q_n(q_m, u)$. Similarly, defining

$$R(q_m, u) = \int_{\chi} v(x)\big(\lambda_1(u,x) - q_m\big)\big\{2L_b\big(\lambda_1(u,x) - q_m\big) - 1\big\}\,dx,$$

and

$$\begin{aligned} Q(q_m, u) &= \frac{\partial}{\partial q_m} R(q_m, u) \\ &= \int_{\chi} v(x)\Big[- 2L_b\Big(\lambda_1(u,x) - q_m\Big) \end{aligned}$$

$$+1 - 2\Big(\lambda_1(u,x) - q_m\Big) L_b'\Big(\lambda_1(u,x) - q_m\Big)\Big] \, dx,$$

where $L_b'(\cdot) = L'(\cdot/b)/b$, we have that $Q(\Gamma(u),u) = 0$ for all $u \in \mathcal{U}_0$ since $L(0) = 1/2$, and hence $P_1\big(\Gamma(u),u\big) + P_2\big(\Gamma(u),u\big)\Gamma'(u) = 0$, where $P_1(q_m,u) = \frac{\partial}{\partial u} Q(q_m,u)$ and $P_2(q_m,u) = \frac{\partial}{\partial q_m} Q(q_m,u)$. It follows that

$$\begin{aligned} & P_{1n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big) - P_1\big(\Gamma(u),u\big) \\ & \quad + \Big[P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big) - P_2\big(\Gamma(u),u\big)\Big]\Gamma'(u) \\ & \quad + P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big)\big(\widehat{\Gamma}'_{LAD,b}(u) - \Gamma'(u)\big) = 0. \end{aligned}$$

It is easily seen that $\sup_{u\in\mathcal{U}_0} \big|P_{1n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big) - P_1\big(\Gamma(u),u\big)\big| = O_P\big(b^{-2} \sup_{u\in\mathcal{U}_0} \big|\widehat{\Gamma}_{LAD,b}(u) - \Gamma(u)\big|\big)$ which is $O_P(n^{-1/2}b^{-2}) = o_P(b^{-1})$ by Theorem 5.2 in Colling and Van Keilegom (2019) and since $nb^2 \to \infty$, and that $P_2\big(\Gamma(u),u\big) = 4L_b'(0) = O(b^{-1})$. Moreover,

$$\begin{aligned} & P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big) - P_2\big(\Gamma(u),u\big) \\ & = \int_{\mathcal{X}} v(x) \Big[4L_b'\big(\widehat{\lambda}_1(u,x) - \widehat{\Gamma}_{LAD,b}(u)\big) - 4L_b'(0) \\ & \qquad\qquad + 2\big(\widehat{\lambda}_1(u,x) - \widehat{\Gamma}_{LAD,b}(u)\big) L_b''\big(\widehat{\lambda}_1(u,x) - \widehat{\Gamma}_{LAD,b}(u)\big)\Big] \, dx \\ & = O_p\Big(b^{-2} \sup_{u\in\mathcal{U}_0, x\in\mathcal{X}_0} \big|\widehat{\lambda}_1(u,x) - \lambda_1(u,x)\big|\Big) , \end{aligned}$$

where $L_b''(\cdot) = L''(\cdot/b)/b^2$. Using the same arguments as at the end of the proof of Theorem 5.2 in Colling and Van Keilegom (2019), the last expression is $o_P(b^{-2}n^{-1/4}b^{1/2})$ which is also $o_P(b^{-1})$ since $nb^2 \to \infty$. In particular, this also implies that

$$\begin{aligned} & \inf_{u\in\mathcal{U}_0} \Big|P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big)\Big| \\ & \geq \inf_{u\in\mathcal{U}_0} \Big|P_2\big(\Gamma(u),u\big)\Big| - \sup_{u\in\mathcal{U}_0} \Big|P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big) - P_2\big(\Gamma(u),u\big)\Big| \\ & = 4L_b'(0) + o_P(b^{-1}) \ . \end{aligned}$$

In conclusion,

$$\sup_{u\in\mathcal{U}_0} \big|\widehat{\Gamma}'_{LAD,b}(u) - \Gamma'(u)\big| \leq \Big(\inf_{u\in\mathcal{U}_0} \Big|P_{2n}\big(\widehat{\Gamma}_{LAD,b}(u),u\big)\Big|\Big)^{-1} o_P(b^{-1}) = o_P(1) \ ,$$

which shows (i).

We will now prove (ii). Every function h in the class $\mathcal{H}$ can be written as $h = f \circ g$ for some $f \in C_c^1(\mathcal{U}_0)$ and some monotone function g that maps $\mathcal{Y}_0$ into the bounded set $\mathcal{U}_0$. We will construct brackets for the space $\mathcal{H}$ by combining brackets for $C_c^1(\mathcal{U}_0)$ with brackets for the space of monotone and bounded functions. First, it follows from Corollary 2.7.2 in Van der Vaart and Wellner (1996) with $d = 1$, $\alpha = 1$ and $r = \infty$ that $N_{1\epsilon} = N_{[\,]}(\epsilon, C_c^1(\mathcal{U}_0), ||\cdot||_{L_\infty}) \leq \exp(K_1\epsilon^{-1})$ for some $K_1 < \infty$, where $||\cdot||_{L_\infty}$ is the supremum norm on $\mathcal{U}_0$. Let $f_{1,\ell} \leq f_{1,u}, \ldots, f_{N_{1\epsilon},\ell} \leq f_{N_{1\epsilon},u}$ be the $N_{1\epsilon}$ brackets for the space $C_c^1(\mathcal{U}_0)$. Next, Theorem 2.7.5 in Van der Vaart and Wellner (1996) shows that the ϵ-bracketing number for the space of monotone and bounded functions with respect to the L_2-norm on $\mathcal{Y}_0$ is bounded by $N_{2\epsilon} \leq \exp(K_2\epsilon^{-1})$ for some $K_2 < \infty$. Let $g_{1,\ell} \leq g_{1,u}, \ldots, g_{N_{2\epsilon},\ell} \leq g_{N_{2\epsilon},u}$ be the $N_{2\epsilon}$ brackets for the latter space. We will show that the bracketing number $N_{[\,]}(\epsilon, \mathcal{H}, ||\cdot||_{L_2})$ of the space $\mathcal{H}$ is bounded by $N_{1\epsilon} \times N_{2\epsilon} \leq \exp((K_1 + K_2)\epsilon^{-1})$.

For a given $1 \leq j \leq N_{1\epsilon}$ and $1 \leq k \leq N_{2\epsilon}$, let

$$h_{j,k,\ell}(y) = \min_{g_{k,\ell}(y) \leq u \leq g_{k,u}(y)} f_{j,\ell}(u) \quad \text{and} \quad h_{j,k,u}(y) = \max_{g_{k,\ell}(y) \leq u \leq g_{k,u}(y)} f_{j,u}(u).$$

Then, it is clear that for all $h \in \mathcal{H}$ there exist $1 \leq j \leq N_{1\epsilon}$ and $1 \leq k \leq N_{2\epsilon}$ such that $h_{j,k,\ell} \leq h \leq h_{j,k,u}$. Moreover,

$$\begin{aligned} &E\Big[\big(h_{j,k,u}(Y) - h_{j,k,\ell}(Y)\big)^2\Big] \\ &\leq E\Big[\big(h_{j,k,u}(Y) - f_{j,u}(g_{k,u}(Y))\big)^2\Big] + E\Big[\big(f_{j,u}(g_{k,u}(Y)) - f_{j,u}(g_{k,\ell}(Y))\big)^2\Big] \\ &\quad + E\Big[\big(f_{j,u}(g_{k,\ell}(Y)) - f_{j,\ell}(g_{k,\ell}(Y))\big)^2\Big] + E\Big[\big(f_{j,\ell}(g_{k,\ell}(Y)) - h_{j,k,\ell}(Y)\big)^2\Big], \end{aligned}$$

and each of these four terms is easily seen to be bounded by a finite multiple of ϵ^2. This finishes the proof. □

Proof of Theorem 4.1. The proof consists in verifying Conditions (1.1) to (1.5) in Theorem 1 in Chen et al. (2003). These conditions are mainly conditions on the functions ℓ_1, M_1 and $M_{n,1}$ in case (i) and on the functions ℓ_2, M_2 and $M_{n,2}$ in case (ii). All these functions are defined in Section 4.1.

Condition (1.1) in Chen et al. (2003) is satisfied since $M_{n,1}(\widehat{\theta}_1, \widehat{h}_b) = 0$ and $M_{n,2}(\widehat{\gamma}_2, \widehat{h}_b) = 0$ by construction, where $\widehat{h}_b(\cdot) = \widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$.

Condition (1.2) in Chen et al. (2003) is ensured in both cases by assumption (A12).

Next, to verify Condition (1.3) in Chen et al. (2003), for $h \in \mathcal{H}$, we have to prove that $\forall\varepsilon > 0$, $\exists\delta > 0$ such that $||h - h_0||_{\mathcal{H}} < \delta$ implies $\sup_{\theta\in\Theta} ||M_1(\theta, h) - M_1(\theta, h_0)|| < \varepsilon$ in case (i) and $\sup_{\gamma\in\Theta_\gamma} ||M_2(\gamma, h) - M_2(\gamma, h_0)|| < \varepsilon$ in case (ii), where $h_0(\cdot) = \Gamma(T(\cdot))$. First, in case (i), we have for $j = 1, \ldots, k$:

$$\begin{aligned} &\Big|M_{1,j}(\theta, h) - M_{1,j}(\theta, h_0)\Big| \\ &= \Bigg|E\Big[w(Y)\Big\{A_j(\theta)\Big(h^2(Y) - h_0^2(Y)\Big) + B_j(\theta, Y)\Big(h(Y) - h_0(Y)\Big) \\ &\qquad + C_j(\theta, Y)\Big(h(Y) - h_0(Y)\Big)\Big\}\Big]\Bigg| . \end{aligned}$$

Consequently, reminding that $||h-h_0||_{\mathcal{H}} = E[(h(Y)-h_0(Y))^2]^{1/2}$ and using Cauchy-Schwarz inequality, we obtain:

$$\sup_{\theta\in\Theta}\big|M_{1,j}(\theta,h)-M_{1,j}(\theta,h_0)\big| < \delta K_{1,j}\ ,$$

where we used $||h-h_0||_{\mathcal{H}} < \delta$ and where

$$\begin{aligned} K_{1,j} \quad &= \quad \sup_{y\in\mathcal{Y}_0}\big(w(y)\big)\Big[\sup_{\theta\in\Theta}\big|A_j(\theta)\big|E\Big(\big(h(Y)+h_0(Y)\big)^2\Big)^{1/2} \\ &\qquad + E\Big(\sup_{\theta\in\Theta}\big|B_j^2(\theta,Y)\big|\Big)^{1/2} + E\Big(\sup_{\theta\in\Theta}\big|C_j^2(\theta,Y)\big|\Big)^{1/2}\Big], \end{aligned}$$

which is finite for all $j = 1,\dots,k$ by assumptions (A10) and (A11) and since $E(h^4(Y)) < \infty$ for $h \in \mathcal{H}$. Taking $\delta = \varepsilon/\sum_{j=1}^k K_{1,j}$, it follows that $\sup_{\theta\in\Theta}\big|\big|M_1(\theta,h)-M_1(\theta,h_0)\big|\big| < \varepsilon$. In case (ii), we use exactly the same reasoning.

Condition (1.4) in Chen et al. (2003) is directly ensured by the fact that $||\widehat{h}_b - h_0||_{\mathcal{H}} \le ||\widehat{h}_b - h_0||_{L_\infty} = O_P(n^{-1/2}) = o_P(1)$ by Corollary 5.1 in Colling and Van Keilegom (2019), where $||\cdot||_{L_\infty}$ is the supremum norm over $\mathcal{Y}_0$.

Finally, for Condition (1.5) in Chen et al. (2003) it suffices by Lemma 1 in the latter paper to show that:

(C1) The class

$$\mathcal{F}_i = \big\{y \to \ell_i(y,\theta,h) : \theta\in\Theta, h\in\mathcal{H}\big\}$$

is P-Donsker $(i = 1,2)$, where P is the probability measure of Y.

(C2) The function $\ell_i(\cdot,\theta,h)$ $(i = 1,2)$ is $L_2(P)$-continuous at (θ_0,h_0).

To show Condition (C1), note that it suffices to show that

$$\int_0^\infty \sqrt{\log N_{[\,]}(\epsilon,\mathcal{F}_i,||\cdot||_{L_2})}\,d\epsilon < \infty. \tag{13}$$

We will show (13) for $i = 1$. The derivation for $i = 2$ is very similar. Since Θ is a compact subspace of the set $\mathbb{R}^k$, at most $N_{\epsilon,\Theta} = C\epsilon^{-k}$ brackets are needed to cover the space Θ for some $C < \infty$. Proposition A.1(ii) shows that $N_{\epsilon,\mathcal{H}} = N_{[\,]}(\epsilon,\mathcal{H},||\cdot||_{L_2}) \le \exp(K\epsilon^{-1})$ for some $K < \infty$. Note that

$$\ell_{1,j}(y,\theta,h) = w(y)\Big[A_j(\theta)h^2(y) + B_j(\theta,y)h(y) + C_j(\theta,y)h(y) + D_j(\theta,y)\Big]$$

for $j = 1,\dots,k$. Fix $\theta\in\Theta$ and $h\in\mathcal{H}$, and suppose that $\theta_{k_1,\ell} \le \theta \le \theta_{k_1,u}$ (where the inequality should be understood componentwise), where $(\theta_{k_1,\ell},\theta_{k_1,u})$ is one of the $N_{\epsilon,\Theta}$ brackets for the space Θ, and that $h_{k_2,\ell} \le h \le h_{k_2,u}$, where $(h_{k_2,\ell},h_{k_2,u})$ is one of the $N_{\epsilon,\mathcal{H}}$ brackets for the space $\mathcal{H}$. Let

$$\ell_{1,j,k_1,k_2,\ell}(y) = \inf_{\theta_{k_1,\ell}\le\theta\le\theta_{k_2,u},h_{k_2,\ell}\le h\le h_{k_2,u}} \ell_{1,j}(y,\theta,h)$$

and let

$$\ell_{1,j,k_1,k_2,u}(y) = \sup_{\theta_{k_1,\ell} \le \theta \le \theta_{k_2,u}, h_{k_2,\ell} \le h \le h_{k_2,u}} \ell_{1,j}(y,\theta,h).$$

Then, it is clear that $\ell_{1,j,k_1,k_2,\ell}(y) \le \ell_{1,j}(y,\theta,h) \le \ell_{1,j,k_1,k_2,u}(y)$ for all y. Next, consider

$$\begin{aligned}
&\left|\ell_{1,j,k_1,k_2,u}(y) - \ell_{1,j,k_1,k_2,\ell}(y)\right| \\
&\le w(y)\Big[c^2 \sup_{\theta\in\Theta} ||\dot{A}_j(\theta)||\, ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + 2c \sup_{\theta\in\Theta} |A_j(\theta)|\, |h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
&\qquad + c \sup_{\theta\in\Theta} ||\dot{B}_j(\theta,y)||\, ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + \sup_{\theta\in\Theta} |B_j(\theta,y)|\, |h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
&\qquad + c \sup_{\theta\in\Theta} ||\dot{C}_j(\theta,y)||\, ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + \sup_{\theta\in\Theta} |C_j(\theta,y)|\, |h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
&\qquad + \sup_{\theta\in\Theta} ||\dot{D}_j(\theta,y)||\, ||\theta_{k_1,u} - \theta_{k_1,\ell}||\Big],
\end{aligned}$$

where c is defined at the beginning of Section 4.1 and is an uniform upper bound for $h \in \mathcal{H}$, $\dot{A}_j(\theta) = \frac{\partial}{\partial\theta} A_j(\theta)$ and similarly for the other functions. Hence, using Cauchy-Schwarz inequality,

$$\begin{aligned}
&E\Big[\left|\ell_{1,j,k_1,k_2,u}(Y) - \ell_{1,j,k_1,k_2,\ell}(Y)\right|^2\Big] \qquad (14) \\
&\le 2^6 \Big(\sup_{y\in\mathcal{Y}_0} w(y)\Big)^2 \Big[c^4 \sup_{\theta\in\Theta} ||\dot{A}_j(\theta)||^2\, ||\theta_{k_1,u} - \theta_{k_1,\ell}||^2 \\
&\quad + 4c^2 \sup_{\theta\in\Theta} |A_j(\theta)|^2\, ||h_{k_2,u} - h_{k_2,\ell}||_{L_2}^2 + c^2 \sup_{\theta\in\Theta} ||\dot{B}_j(\theta,\cdot)||_{L_2}^2\, ||\theta_{k_1,u} - \theta_{k_1,\ell}||^2 \\
&\quad + \sup_{\theta\in\Theta} |B_j(\theta,\cdot)|^2\, ||h_{k_2,u} - h_{k_2,\ell}||_{L_2}^2 + c^2 \sup_{\theta\in\Theta} ||\dot{C}_j(\theta,\cdot)||_{L_2}^2\, ||\theta_{k_1,u} - \theta_{k_1,\ell}||^2 \\
&\quad + \sup_{\theta\in\Theta} |C_j(\theta,\cdot)|^2\, ||h_{k_2,u} - h_{k_2,\ell}||_{L_2}^2 + \sup_{\theta\in\Theta} ||\dot{D}_j(\theta,\cdot)||_{L_2}^2\, ||\theta_{k_1,u} - \theta_{k_1,\ell}||^2\Big],
\end{aligned}$$

and this is bounded by a finite multiple of ϵ^2 by assumptions (A10) and (A11). It now follows that the integral in (13) is finite.

Finally, for Condition (C2) we need to show that $E\|\ell_i(Y,\theta,h) - \ell_i(Y,\theta_0,h_0)\|^2$ converges to 0 as $||\theta-\theta_0|| \to 0$ and $||h-h_0||_{\mathcal{H}} \to 0$ $(i = 1,2)$. This is easily seen to hold true thanks to assumptions (A10) and (A11) and calculations similar to those leading to (14) above. □

Proof of Theorem 4.2. The proof consists in verifying Conditions (2.1) to (2.6) in Theorem 2 in Chen et al. (2003). Condition (2.1) in Chen et al. (2003) is satisfied since $M_{n,1}(\widehat{\theta}_1, \widehat{h}_b) = 0$ and $M_{n,2}(\widehat{\gamma}_2, \widehat{h}_b) = 0$ by construction, where $\widehat{h}_b(\cdot) = \widehat{\Gamma}_{LAD,b}(\widehat{T}(\cdot))$, while Condition (2.2) in Chen et al. (2003) is ensured by assumptions (A11) and (A13).

Next, Condition (2.3) involves the pathwise derivative of the functions $M_1(\theta,h_0)$ and $M_2(\gamma,h_0)$ in the direction $[h-h_0]$. These pathwise derivatives are respectively given by

$$\Gamma_1(\theta,h_0)[h-h_0] = \lim_{\tau\to 0} \tau^{-1}\Big(M_1(\theta, h_0 + \tau(h-h_0)) - M_1(\theta,h_0)\Big)$$

$$= \Big(\Gamma_{1,j}(\theta, h_0)[h - h_0]\Big)_{j=1,\dots,k},$$

where $\Gamma_{1,j}(\theta, h_0)[h-h_0] = E\big[w(Y)\big\{2A_j(\theta)h_0(Y)+B_j(\theta,Y)+C_j(\theta,Y)\big\}\big(h(Y)-h_0(Y)\big)\big]$, and

$$\begin{aligned}\Gamma_2(\gamma, h_0)[h - h_0] &= \lim_{\tau\to 0}\tau^{-1}\Big(M_2(\gamma, h_0 + \tau(h - h_0)) - M_2(\gamma, h_0)\Big)\\ &= E\begin{pmatrix} w(Y)\big(2c_1h_0(Y) + c_2 - \Lambda_\theta(Y)\big)\big(h(Y) - h_0(Y)\big)\\ w(Y)c_1\big(h(Y) - h_0(Y)\big)\\ -w(Y)\dot\Lambda_{\theta,1}(Y)c_1\big(h(Y) - h_0(Y)\big)\\ \vdots\\ -w(Y)\dot\Lambda_{\theta,k}(Y)c_1\big(h(Y) - h_0(Y)\big)\end{pmatrix}.\end{aligned}$$

The jth component of the vector $\Gamma_2(\gamma, h_0)[h - h_0]$ will be denoted by $\Gamma_{2,j}(\gamma, h_0)[h - h_0]$ for $j = 1,\dots,k + 2$.

We will now verify Condition (2.3)(i) in Chen et al. (2003). In case (i) and for $j = 1,\dots,k$, we have:

$$\begin{aligned}&\Big|M_{1,j}(\theta, h) - M_{1,j}(\theta, h_0) - \Gamma_{1,j}(\theta, h_0)[h - h_0]\Big|\\ &= \Big|E\Big\{w(Y)A_j(\theta)\Big[h^2(Y) - h_0^2(Y) - 2h_0(Y)(h(Y) - h_0(Y))\Big]\Big\}\Big|\\ &= \Big|E\Big\{w(Y)A_j(\theta)\Big(h(Y) - h_0(Y)\Big)^2\Big\}\Big|\\ &\le \beta_{1,j}||h - h_0||^2_{\mathcal{H}}\ ,\end{aligned}$$

where $\beta_{1,j} = \sup_{\theta\in\Theta}|A_j(\theta)|\sup_{y\in\mathcal{Y}_0} w(y) < \infty$ by assumption (A10). Hence, $||M_1(\theta, h) - M_1(\theta, h_0) - \Gamma_1(\theta, h_0)[h - h_0]|| \le \beta_1||h - h_0||^2_{\mathcal{H}}$ with $\beta_1 = \sum_{j=1}^k \beta_{1,j}$. The derivation is similar in case (ii).

Next, to verify Condition (2.3)(ii) in Chen et al. (2003), consider a positive sequence $\delta_n = o(1)$, $\theta\in\Theta_{\delta_n}$ and $h \in \mathcal{H}_{\delta_n}$, where $\Theta_{\delta_n} = \{\theta \in \Theta : ||\theta - \theta_0|| \le \delta_n\}$ and $\mathcal{H}_{\delta_n} = \{h \in \mathcal{H} : ||h - h_0||_{\mathcal{H}} \le \delta_n\}$. In case (i) and for $j = 1,\dots,k$, we have

$$\begin{aligned}&\Big|\Gamma_{1,j}(\theta, h_0)[h - h_0] - \Gamma_{1,j}(\theta_0, h_0)[h - h_0]\Big|\\ &\le \Big|E\Big\{2w(Y)\Big(A_j(\theta) - A_j(\theta_0)\Big)h_0(Y)\Big(h(Y) - h_0(Y)\Big)\Big\}\Big|\\ &\quad+\Big|E\Big\{w(Y)\Big(B_j(\theta, Y) - B_j(\theta_0, Y)\Big)\Big(h(Y) - h_0(Y)\Big)\Big\}\Big|\\ &\quad+\Big|E\Big\{w(Y)\Big(C_j(\theta, Y) - C_j(\theta_0, Y)\Big)\Big(h(Y) - h_0(Y)\Big)\Big\}\Big|\,. \qquad (15)\end{aligned}$$

Using the mean value theorem and since $E[h_0(Y)(h(Y)-h_0(Y))] \le ||h_0||_{\mathcal{H}}||h-h_0||_{\mathcal{H}}$ by Cauchy-Schwarz inequality, the first term on the right hand side of (15) is bounded by

$$2\sup_{y\in\mathcal{Y}_0}(w(y))\sup_{\theta\in\Theta}\big|\big|\dot A_j(\theta)\big|\big|\,||\theta - \theta_0||\,||h_0||_{\mathcal{H}}\,||h - h_0||_{\mathcal{H}},$$

where $\dot{A}_j(\theta) = \frac{\partial}{\partial\theta} A_j(\theta)$. Similarly, the second and third terms on the right hand side of (15) are bounded by

$$\sup_{y\in\mathcal{Y}_0} (w(y)) \left[E\Big(\sup_{\theta\in\Theta} \big|\big|\dot{B}_j(\theta, Y)\big|\big|^2\Big)\right]^{1/2} ||\theta - \theta_0||\, ||h - h_0||_{\mathcal{H}},$$

and

$$\sup_{y\in\mathcal{Y}_0} (w(y)) \left[E\Big(\sup_{\theta\in\Theta} \big|\big|\dot{C}_j(\theta, Y)\big|\big|^2\Big)\right]^{1/2} ||\theta - \theta_0||\, ||h - h_0||_{\mathcal{H}},$$

where $\dot{B}_j(\theta, y) = \frac{\partial}{\partial\theta} B_j(\theta, y)$ and $\dot{C}_j(\theta, y) = \frac{\partial}{\partial\theta} C_j(\theta, y)$. These last three terms are bounded by $o(1)\delta_n$ since $||h - h_0||_{\mathcal{H}} \le \delta_n = o(1)$, $||\theta - \theta_0|| \le \delta_n$, $E(h_0^2(Y)) < \infty$ since $h_0 \in \mathcal{H}$, $\sup_{y\in\mathcal{Y}_0} w(y) < \infty$ by assumption (A10) and $\sup_{\theta\in\Theta} ||\dot{A}_j(\theta)||$, $E(\sup_{\theta\in\Theta} ||\dot{B}_j(\theta, Y)||^2)$ and $E(\sup_{\theta\in\Theta} ||\dot{C}_j(\theta, Y)||^2)$ are finite by assumption (A11). The proof of Condition (2.3)(ii) in Chen et al. (2003) in case (ii) follows exactly the same way.

Moreover, Condition (2.4) in Chen et al. (2003) is satisfied in both cases using Proposition A.1(i) and since we have shown in the proof of Theorem 4.1 that $||\widehat{h}_b - h_0||_{\mathcal{H}} = O_P(n^{-1/2}) = o_P(n^{-1/4})$.

Next, in the proof of Theorem 4.1 we showed that Lemma 1 in Chen et al. (2003) is verified in our case. This lemma is not only sufficient for Condition (1.5) but also for Condition (2.5), see Remark 2 in Chen et al. (2003).

Finally, we verify Condition (2.6) in Chen et al. (2003). At the end of Section 4.1, we justified that $M_{n,1}(\theta_0, h_0) = 0$ and $M_{n,2}(\gamma_0, h_0) = 0$. Moreover, in case (i) and for $j = 1, \dots, k$, we have:

$$\begin{aligned}&\Gamma_{1,j}(\theta_0, h_0)[\widehat{h}_b - h_0]\\ &= E\Big[w(Y)\Big\{2A_j(\theta_0)h_0(Y) + B_j(\theta_0, Y) + C_j(\theta_0, Y)\Big\}(\widehat{h}_b(Y) - h_0(Y))\Big].\end{aligned}$$

Using Corollary 5.1 in Colling and Van Keilegom (2019), we have $\widehat{h}_b(y) - h_0(y) = n^{-1}\sum_{i=1}^n \varphi^v_{X_i,Y_i}(y) + o_P(n^{-1/2})$ uniformly in $y \in \mathcal{Y}_0$. Consequently,

$$\begin{aligned}&M_{n,1}(\theta_0, h_0) + \Gamma_{1,j}(\theta_0, h_0)[\widehat{h}_b - h_0]\\ &= n^{-1}\sum_{i=1}^n E\Big[w(Y)\Big\{2A_j(\theta_0)h_0(Y) + B_j(\theta_0, Y) + C_j(\theta_0, Y)\Big\}\varphi^v_{X_i,Y_i}(Y)\Big| X_i, Y_i\Big]\\ &\quad + o_P(n^{-1/2}),\end{aligned}$$

by assumptions (A10) and (A11) and the fact that $E(h_0(Y)) < \infty$. The last expression is a sum of i.i.d. terms. Hence, we conclude the proof of case (i) using the multivariate central limit theorem and the fact that $E(\varphi^v_{X,Y}(y)) = 0$ for all y by Corollary 5.1 in Colling and Van Keilegom (2019). Similarly, in case (ii), we have

$$M_{n,2}(\gamma_0, h_0) + \Gamma_2(\gamma_0, h_0)[\widehat{h}_b - h_0]$$

$$= n^{-1}\sum_{i=1}^{n}\begin{pmatrix} E\big\{w(Y)\big(2c_{1,0}h_0(Y)+c_{2,0}-\Lambda_{\theta_0}(Y)\big)\varphi^v_{X_i,Y_i}(Y)\big|X_i,Y_i\big\} \\ E\big\{w(Y)c_{1,0}\varphi^v_{X_i,Y_i}(Y)\big|X_i,Y_i\big\} \\ -E\big\{w(Y)\dot{\Lambda}_{\theta_0,1}(Y)c_{1,0}\varphi^v_{X_i,Y_i}(Y)\big|X_i,Y_i\big\} \\ \vdots \\ -E\big\{w(Y)\dot{\Lambda}_{\theta_0,k}(Y)c_{1,0}\varphi^v_{X_i,Y_i}(Y)\big|X_i,Y_i\big\} \end{pmatrix}$$
$$+o_P(n^{-1/2}).$$

It suffices to apply again the multivariate central limit theorem to conclude the proof of this theorem. □

Acknowledgements

B. Colling and I. Van Keilegom are financially supported by the European Research Council (2016–2021, Horizon 2020/ERC grant agreement No. 694409).

References

Allison, J., Hušková, M., and Meintanis, S. (2018). Testing the adequacy of semiparametric transformation models. *TEST*, 27(1):70–94. MR3764024

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2(2):273–277.

Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311. MR0624332

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society – Series B*, 26(2):211–252. MR0192611

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598. MR0803258

Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the US wage structure, 1963–1987. *Journal of Econometrics*, 65(1):109–154. MR1323055

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*, volume 30. CRC Press. MR1014890

Chen, S. (2002). Rank estimation of transformation models. *Econometrica*, 70(4):1683–1697. MR1929984

Chen, X., Linton, O., and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608. MR2000259

Chiappori, P.-A., Komunjer, I., and Kristensen, D. (2015). Nonparametric identification and estimation of transformation models. *Journal of Econometrics*, 188(1):22–39. MR3371659

Colling, B., Heuchenne, C., Samb, R., and Van Keilegom, I. (2015). Estimation of the error density in a semiparametric transformation model. *Annals of the Institute of Statistical Mathematics*, 67(1):1–18. MR3297856

Colling, B. and Van Keilegom, I. (2016). Goodness-of-fit tests in semiparametric transformation models. *TEST*, 25(2):291–308. MR3493520

Colling, B. and Van Keilegom, I. (2017). Goodness-of-fit tests in semiparametric transformation models using the integrated regression function. *Journal of Multivariate Analysis*, 160:10–30. MR3688687

Colling, B. and Van Keilegom, I. (2019). Estimation of fully nonparametric transformation models. *Bernoulli*, 25:3762–3795. MR4010972

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society – Series B*, 34:187–220. MR0341758

Fitzenberger, B., Wilke, R. A., and Zhang, X. (2010). Implementing Box–Cox quantile regression. *Econometric Reviews*, 29(2):158–181. MR2747497

Heuchenne, C., Samb, R., and Van Keilegom, I. (2015). Estimating the residual distribution in semiparametric transformation models. *Electronic Journal of Statistics*, 9:2391–2419. MR3417187

Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. *Econometrica*, 64(1):103–137. MR1366143

Horowitz, J. L. (2001). Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica*, 69(2):499–513. MR1819761

Hušková, M., Meintanis, S., Neumeyer, N., and Pretorius, C. (2018). Independence tests in semiparametric transformation models. *South African Statistical Journal*, 52(1):1–13. MR3793070

Jacho-Chavez, D., Lewbel, A., and Linton, O. (2010). Identification and nonparametric estimation of a transformation additively separable model. *Journal of Econometrics*, 156(2):392–407. MR2609941

John, J. and Draper, N. (1980). An alternative family of transformations. *Journal of the Royal Statistical Society – Series C*, 29(2):190–197.

Kloodt, N. and Neumeyer, N. (2017). Specification tests in semiparametric transformation models. *arXiv preprint arXiv:1709.06855*. MR3793070

Linton, O., Sperlich, S., and Van Keilegom, I. (2008). Estimation of a semiparametric transformation model. *The Annals of Statistics*, 36(2):686–718. MR2396812

Machado, J. A. and Mata, J. (2000). Box-Cox quantile regression and the distribution of firm sizes. *Journal of Applied Econometrics*, 15(3):253–274.

MacKinnon, J. G. and Magee, L. (1990). Transforming the dependent variable in regression models. *International Economic Review*, 31(2):315–339.

Mu, Y. and He, X. (2007). Power transformation toward a linear regression quantile. *Journal of the American Statistical Association*, 102(477):269–279. MR2293308

Neumeyer, N., Noh, H., and Van Keilegom, I. (2016). Heteroscedastic semiparametric transformation models: estimation and testing for validity. *Statistica Sinica*, 26:925–954. MR3559937

Sakia, R. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistical Society – Series D*, 41(2):169–178.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and Empir-*

ical Processes. Springer. MR1385671

Vanhems, A. and Van Keilegom, I. (2019). Estimation of a semiparametric transformation model in the presence of endogeneity. *Econometric Theory*, 35:73–110. MR3904172

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959. MR1813988

Zellner, A. and Revankar, N. S. (1969). Generalized production functions. *The Review of Economic Studies*, 36(2):241–250.