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

Colling B, Van Keilegom I.
Estimation of a semiparametric
transformation model: a novel approach
based on least squares minimization

Benjamin COLLING\(^1\) and Ingrid VAN KEILEGOM\(^*\),\(^1\)

\(^1\)Research Centre for Operations Research and Business Statistics (ORSTAT), KU Leuven

August 21, 2018

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 \( \varepsilon \) 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 \( \theta \). The main idea of these two estimators is to minimize, with respect to \( \theta \), the \( L^2 \)-distance between the transformation \( \Lambda_\theta \) 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 (2018). We establish the consistency and the asymptotic normality of the two proposed estimators of \( \theta \). 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).

Key Words: Asymptotic properties; Estimation; \( L^2 \)-distance minimization; Parametric transformation; Semiparametric regression.

\(^*\)Corresponding author: ingrid.vankeilegom@kuleuven.be
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, \( \beta \) is a vector of unknown regression parameters and \( \varepsilon \) 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 \( \Lambda(Y) = m(X) + \epsilon \), 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 (2018) constructed fully nonparametric estimators of the transformation \( \Lambda \). The main motivation of the estimators constructed in Colling and Van Keilegom (2018) 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 \( \Gamma \) 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:

\[
\Lambda_\theta(Y) = m(X) + \epsilon ,
\]

where \( m(\cdot) \) is an unknown regression function, \( \Lambda_\theta \) is a transformation belonging to a parametric family of strictly increasing functions and \( \theta \in \Theta \) where \( \Theta \) is a compact subset of \( \mathbb{R}^k \). We will denote by \( \theta_0 \) the true but unknown value of \( \theta \). Moreover, we assume that \( X \) is a \( d \)-dimensional covariate with compact support \( \chi \), \( Y \) is a univariate response variable with support \( \mathcal{Y} \) and the error term \( \epsilon \) has zero mean and is independent of \( X \). We also introduce the following notations: \( m(x, \theta) = E[\Lambda_\theta(Y)|X = x] \), \( m(x, \theta_0) = m(x) \), \( \epsilon(\theta) = \Lambda_\theta(Y) - m(X, \theta) \) and \( \epsilon(\theta_0) = \epsilon \). Finally, let \( F_X, F_{\epsilon(\theta)}, f_X \) and \( f_{\epsilon(\theta)} \) be the distribution and density functions of \( X \) and \( \epsilon(\theta) \). We assume that we have randomly drawn an iid sample \( (X_1, Y_1), \ldots, (X_n, Y_n) \).
from model (1.1), 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 (1.1) and proposed two estimation methods for the unknown true parameter vector $\theta_0$: 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 $\theta$, after having replaced all unknown functions in the likelihood by nonparametric estimators. Then, the profile likelihood estimator of $\theta$ is defined by

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

(1.2)

where $\hat{m}(\cdot, \theta)$ and $\hat{f}_{\epsilon(\theta)}(\cdot)$ are suitable nonparametric estimators of $m(\cdot, \theta)$ and $f_{\epsilon(\theta)}(\cdot)$ 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 (1.1). First, we mention the work of Vanhems and Van Keilegom (2018) 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 $\theta_0$ in the context of a semiparametric transformation model of the form (1.1). These estimators will be competitors of the profile likelihood estimator $\hat{\theta}_{PL}$ introduced in (1.2). The main idea of the new estimators of $\theta_0$ is to minimize, with respect to $\theta$, the $L^2$-distance
between the transformation $\Lambda_\theta$ and one of its fully nonparametric estimators. In Section 2, we will explain in more detail the intuition behind our two new estimators of $\theta_0$, 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. Finally, Section 6 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 $\theta_0$ is to minimize, with respect to $\theta$, the $L_2$-distance between the transformation $\Lambda_\theta$ 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 (2018). Moreover, we explained in the introduction why the estimators constructed in Colling and Van Keilegom (2018) 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 (2018) 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 (2018), 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 $\Lambda$ 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 (2018) for more details about the identification of the model. Following the same idea as in Colling and Van Keilegom (2018), we rewrite the transformation $\Lambda(Y)$ as $\Lambda(Y) = \Gamma(U)$, where $\Gamma$ is an increasing function, and

$$U = T(Y) = \frac{F_Y(Y) - F_Y(0)}{F_Y(1) - F_Y(0)}.$$ (2.1)
Note that \( T(0) = 0 \) and \( T(1) = 1 \), and hence combined with the imposed condition on \( \Lambda \), we find that \( \Gamma(0) = 0 \) and \( \Gamma(1) = 1 \), i.e. \( \Gamma \) satisfies the same identification constraints as \( \Lambda \).

We estimate the variable \( U \) by

\[
\hat{U} = \hat{T}(Y) = \frac{\hat{F}_Y(Y) - \hat{F}_Y(0)}{\hat{F}_Y(1) - \hat{F}_Y(0)},
\]

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

\[
\Gamma(u) = \lambda_1(u, x) = \frac{S_1(u, x)}{S_1(1, x)} \text{ with } S_1(u, x) = \int_0^u \frac{\partial}{\partial x_1} \phi(w, x) \, dw,
\]

as shown in Theorem 3.1 in Colling and Van Keilegom (2018), where \( \phi(u, x) = P(U \leq u | X = x) \) is the conditional distribution of \( U \) given \( X \), and \( x_1 \) is the first component of the vector \( x = (x_1, \ldots, x_d) \). Hence, we can write \( \Gamma(u) \) as

\[
\Gamma(u) = \arg\min_{q_m \in \mathbb{R}} \int_{\chi} v(x) \ell(\lambda_1(u, x) - q_m) \, dx,
\]

for any positive weight function \( v(\cdot) \) and loss function \( \ell(\cdot) \) satisfying \( \ell(0) = 0 \). In particular, we can work with the loss \( \ell(u) = u(2L_b(u) - 1) \), 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

\[
\hat{\lambda}_1(u, x) = \frac{\hat{S}_1(u, x)}{\hat{S}_1(1, x)} \text{ with } \hat{S}_1(u, x) = \int_0^u \frac{\partial}{\partial x_1} \hat{\phi}(w, x) \, dw,
\]

where

\[
\hat{\phi}(u, x) = \frac{\sum_{i=1}^{n} \mathcal{K}_{h_u}(u - \hat{U}_i) \mathcal{K}_{h_x}(X_i - x)}{\sum_{i=1}^{n} \mathcal{K}_{h_x}(X_i - x)},
\]

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

\[
\hat{\Gamma}_{\text{LAD}, b}(u) = \arg\min_{q_m \in \mathbb{R}} \int_{\chi} v(x) (\hat{\lambda}_1(u, x) - q_m) \{2L_b(\hat{\lambda}_1(u, x) - q_m) - 1\} \, dx.
\]
Consequently, a natural estimator of $\theta_0$ is given by
\[
\argmin_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} \left( \hat{\Gamma}_{LAD,b}(\hat{T}(Y_i)) - \Lambda_\theta(Y_i) \right)^2 .
\] (2.4)

However, it is important to remind that the estimator $\hat{\Gamma}_{LAD,b}(\hat{T}())$ 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 (2.4) will then lead to an inconsistent estimator of $\theta$. In the next section we will explain in detail how we can adjust the estimator $\hat{\Gamma}_{LAD,b}(\hat{T}())$ with additive and multiplicative constants so that the corresponding adjusted estimator in (2.4) 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 $\hat{\Gamma}^*_LAD,b(\hat{T}^*(\cdot))$ instead of $\hat{\Gamma}_{LAD,b}(\hat{T}()$, where $\hat{\Gamma}^*_LAD,b(\cdot)$ and $\hat{T}^*(\cdot)$ 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 $\hat{\Gamma}_{LAD,b}(\hat{T}())$ has several advantages. First, its asymptotic properties have already been developed in Colling and Van Keilegom (2018), 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. 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 $\hat{\Gamma}^*_LAD,b(\cdot)$ and $\hat{T}^*(\cdot)$ depends on the imposed identification conditions, whereas our goal is to construct an estimator of $\theta_0$ that is consistent under general identification conditions.

3 Definition of the estimators

The estimators of $\theta$ 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 \( j = 1, \ldots, d \). 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 (2.4) and takes into account the set of identification conditions (I1) and (I1'). The assumptions (A1)-(A4) are given in the Appendix. Let \( U_0 \) be a compact subset in the interior of the support \( U \) of \( U \), and let \( Y_0 \) be a compact set strictly included in \( T^{-1}(U_0) \).

**Proposition 3.1.** Assume (A1)-(A4). Then, under either (I1) or (I1') and (I2), we have for all \( x \in \chi \) and \( y \in Y_0 \),

\[
\Lambda(y) = \left( \Lambda(1) - \Lambda(0) \right) \frac{S_1(T(y), x)}{S_1(1, x)} + \Lambda(0),
\]

where \( T \) and \( S_1 \) are defined in (2.1) and (2.3) respectively. Moreover, the right hand side of (3.1) does not depend on \( x \).

The proof of this proposition is given in Section 6.2. Consequently, using Proposition 3.1 and the fact that \( \hat{\Gamma}_{LAD,b}(\hat{T}(y)) \) is a nonparametric estimator of \( S_1(T(y), x) / S_1(1, x) \), it is natural to define the following estimator of \( \theta_0 \) :

\[
\hat{\theta}_1 = \arg\min_{\theta \in \Theta} n^{-1} \sum_{i=1}^{n} w(Y_i) \left( (\Lambda_\theta(1) - \Lambda_\theta(0)) \hat{\Gamma}_{LAD,b}(\hat{T}(Y_i)) + \Lambda_\theta(0) - \Lambda_\theta(Y_i) \right)^2,
\]

where \( w \) is a certain positive weight function with support included in \( Y_0 \), 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 (3.2) equals expression (2.4) 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 \( \hat{\theta}_1 \) be free parameters that do not depend on \( \theta \). In that way we will have \( k + 2 \) parameters over which we minimize our weighted \( L_2 \)-distance (\( k \) being the dimension of \( \theta \)) instead of just \( k \). This could lead to a better estimator of \( \theta \). Therefore, we define a second estimator of \( \theta_0 \), which is obtained by replacing \( \Lambda_\theta(1) - \Lambda_\theta(0) \) and \( \Lambda_\theta(0) \)
Moreover, we define 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 \( \theta \), i.e.

\[
\hat{\gamma}_2 = (\hat{c}_1, \hat{c}_2, \hat{\theta}_2)^t = \arg\min_{\gamma \in \Theta} n^{-1} \sum_{i=1}^n w(Y_i) \left( c_1 \hat{\Gamma}_{LAD,b}(\hat{T}(Y_i)) + c_2 - \Lambda_\theta(Y_i) \right)^2,
\]

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 \( \gamma \) equals \( \gamma_0 = (\Lambda_{\theta_0}(1) - \Lambda_{\theta_0}(0), \Lambda_{\theta_0}(0), \theta_0)^t \). Note that if the estimator \( \hat{\Gamma}_{LAD,b}(\hat{T}(\cdot)) \) performs well, \( \hat{c}_1 \) and \( \hat{c}_2 \) should perform similarly as \( \gamma_0 \), i.e. \( \hat{\gamma}_2 \) should perform similarly as \( \gamma_0 \) and then \( \hat{\theta}_2 \) should perform similarly as \( \hat{\theta}_1 \). Otherwise, \( \hat{c}_1 \) and \( \hat{c}_2 \) could compensate a bad estimator \( \hat{\Gamma}_{LAD,b}(\hat{T}(\cdot)) \) by taking some other values far away from \( \Lambda_{\hat{\theta}_1}(1) - \Lambda_{\hat{\theta}_1}(0) \) and \( \Lambda_{\hat{\theta}_1}(0) \).

### 4 Asymptotic results

#### 4.1 Notations and definitions

Before establishing the main asymptotic results, we need to introduce several notations. First, \( \theta = (\theta_1, \ldots, \theta_k)^t \), \( \dot{\Lambda}_{\theta,j}(y) = \frac{\partial}{\partial \theta_j} \Lambda_\theta(y) \) for \( j = 1, \ldots, k \), and \( \dot{\Lambda}_\theta(y) = (\dot{\Lambda}_{\theta,1}(y), \ldots, \dot{\Lambda}_{\theta,k}(y))^t \) is the vector of partial derivatives of \( \Lambda_\theta(y) \) with respect to the components of \( \theta \). Next, let \( C^1_c(U_0) \) be the set of functions \( f : U_0 \rightarrow \mathbb{R} \) for which \( \|f\|_{\infty,1} \leq c < \infty \), where

\[
\|f\|_{\infty,1} = \sup_{u \in U_0} |f(u)| + \sup_{u \neq u', u \in U_0, u' \in U_0} \frac{|f(u') - f(u)|}{|u' - u|}.
\]

Moreover, we define

\[
\mathcal{H} = \left\{ h : \mathcal{Y}_0 \rightarrow \mathbb{R} : h = f \circ g, f \in C^1_c(U_0), g : \mathcal{Y}_0 \rightarrow U_0 \text{ is monotone} \right\},
\]

and let \( \|h\|_{\mathcal{H}} = [E(h^2(Y))]^{1/2} \). For technical reasons we have to work with the restricted sets \( \mathcal{Y}_0 \) and \( U_0 \) instead of \( \mathcal{Y} \) and \( U \). See Colling and Van Keilegom (2018) for more details.

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

\[
A(\theta) = (A_j(\theta))^t_{j=1,\ldots,k} \quad \text{with} \quad A_j(\theta) = (\dot{\Lambda}_{\theta,j}(1) - \dot{\Lambda}_{\theta,j}(0)) (\Lambda_\theta(1) - \Lambda_\theta(0))
\]

\[
B(\theta, y) = (B_j(\theta, y))^t_{j=1,\ldots,k} \quad \text{with} \quad B_j(\theta, y) = (\dot{\Lambda}_{\theta,j}(1) - \dot{\Lambda}_{\theta,j}(0)) (\Lambda_\theta(0) - \Lambda_\theta(y))
\]

\[
C(\theta, y) = (C_j(\theta, y))^t_{j=1,\ldots,k} \quad \text{with} \quad C_j(\theta, y) = (\dot{\Lambda}_{\theta,j}(0) - \dot{\Lambda}_{\theta,j}(y)) (\Lambda_\theta(1) - \Lambda_\theta(0))
\]

\[
D(\theta, y) = (D_j(\theta, y))^t_{j=1,\ldots,k} \quad \text{with} \quad D_j(\theta, y) = (\dot{\Lambda}_{\theta,j}(0) - \dot{\Lambda}_{\theta,j}(y)) (\Lambda_\theta(0) - \Lambda_\theta(y))
\]
Moreover, for \( h \in \mathcal{H} \) and \( \theta \in \Theta \), we consider \( \ell_1(y, \theta, h) = (\ell_{1,j}(y, \theta, h))_{j=1,\ldots,k} \), where, for \( j = 1, \ldots, k \),
\[
\ell_{1,j}(y, \theta, h) = w(y) \left[ A_j(\theta)h^2(y) + B_j(\theta, y)h(y) + C_j(\theta, y)h(y) + D_j(\theta, y) \right].
\]

Finally, let \( M_1(\theta, h) = \mathbb{E}[\ell_1(Y, \theta, h)] \) 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 \( j = 1, \ldots, k \).

Similarly, the following notations are related to the second estimator \( \hat{\theta}_2 \). For \( h \in \mathcal{H} \) and \( \gamma = (c_1, c_2, \theta)^t \in \Theta_\gamma \), we consider \( \ell_2(y, \gamma, h) = (\ell_{2,j}(y, \gamma, h))_{j=1,\ldots,k+2} \), where
\[
\ell_{2,1}(y, \gamma, h) = w(y) \left[ c_1 h^2(y) + c_2 h(y) - \Lambda_\theta(y)h(y) \right],
\]
\[
\ell_{2,2}(y, \gamma, h) = w(y) \left[ c_1 h(y) + c_2 - \Lambda_\theta(y) \right],
\]
\[
\ell_{2,j}(y, \gamma, h) = -w(y) \dot{\Lambda}_{\theta,j-2}(y) \left[ c_1 h(y) + c_2 - \Lambda_\theta(y) \right],
\]
for \( j = 3, \ldots, k+2 \). Moreover, let \( M_2(\gamma, h) = \mathbb{E}[\ell_2(Y, \gamma, h)] \) 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 \( M_{n,1}(\theta, \hat{h}) \) and \( M_{n,2}(\gamma, \hat{h}) \), where \( \hat{h}(y) = \hat{\Gamma}_{LAD,h}(\hat{T}(y)) \) estimates \( h_0(y) = \frac{S_1(T(y), x)}{S_1(T, x)} \), are simply the derivatives of the expressions that we minimize in (3.2) and (3.3) 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 \( \hat{\theta}_1 \) and \( \hat{\gamma}_2 \); see the proofs of Theorems 4.1 and 4.2 in Section 6.2. These sufficient conditions are mainly conditions on the class of functions \( \mathcal{H} \) and either on the functions \( \ell_1, M_1 \) and \( M_{n,1} \) for the estimator \( \hat{\theta}_1 \) or on the functions \( \ell_2, M_2 \) and \( M_{n,2} \) for the estimator \( \hat{\gamma}_2 \).

### 4.2 Consistency and asymptotic normality

The following theorems establish respectively the consistency and the asymptotic normality of \( \hat{\theta}_1 \) and \( \hat{\gamma}_2 \). 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),

\[(i) \hat{\theta}_1 - \theta_0 = o_P(1) \quad \text{and} \quad (ii) \hat{\gamma}_2 - \gamma_0 = o_P(1) .\]

Theorem 4.2. Assume (A1)-(A13). Then, under either (I1) or (I1’) and (I2),

\[(i) \sqrt{n} (\hat{\theta}_1 - \theta_0) \xrightarrow{d} N(0, \Omega_1),\]

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 \( \theta \), and the matrix \( V_1 \) is given by

\[ V_1 = \text{Var} \left( E\left[ w(Y') \varphi_{X,Y}^v(Y') \left( 2A(\theta_0)h_0(Y') + B(\theta_0, Y') + C(\theta_0, Y') \right) \bigg| X, Y \right] \right), \]

where \( Y' \) is an i.i.d. copy of \( Y \), \( h_0(y) = \frac{S_1(T(y), x)}{S_1(1, x)} \), and \( \varphi_{X,Y}^v \) is defined in Corollary 5.1 in Colling and Van Keilegom (2018).

\[(ii) \sqrt{n} (\hat{\gamma}_2 - \gamma_0) \xrightarrow{d} N(0, \Omega_2),\]

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 \( \gamma \), and the matrix \( V_2 \) is given by

\[ V_2 = \text{Var} \left( \begin{pmatrix} E\left[ w(Y') \varphi_{X,Y}^v(Y') \left( 2c_{1,0} h_0(Y') + c_{2,0} - \Lambda_{\theta_0}(Y') \right) \bigg| X, Y \right] \\ E\left[ w(Y') \varphi_{X,Y}^v(Y') \bigg| X, Y \right] \\ -E\left[ w(Y') c_{1,0} \hat{\Lambda}_{\theta_0,1}(Y') \varphi_{X,Y}^v(Y') \bigg| X, Y \right] \\ \vdots \\ -E\left[ w(Y') c_{1,0} \hat{\Lambda}_{\theta_0,k}(Y') \varphi_{X,Y}^v(Y') \bigg| X, Y \right] \end{pmatrix}, \]

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 6.2. 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 \( \hat{h} - h_0 \). 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 \( \hat{\theta}_2 \) is consistent for \( \theta_0 \), and that Theorem 4.2(ii) implies that \( \hat{\theta}_2 \) is asymptotically normally distributed with variance-covariance matrix given by the lower \( k \times k \) submatrix of the matrix \( \Omega_2 \).

5 Simulations

In this section, we perform simulations in order to compare the performance of our new estimators \( \hat{\theta}_1 \) and \( \hat{\theta}_2 \) of the transformation parameter with that of the profile likelihood estimator \( \hat{\theta}_{PL} \) proposed by Linton et al. (2008) and defined in (1.2).

The four simulated models are

Model 1: \( \Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i \) where \( \varepsilon_1, \ldots, \varepsilon_n \sim \text{iid } N(0, 1) \),

Model 2: \( \Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i \) where \( \varepsilon_1, \ldots, \varepsilon_n \sim \text{iid } N(0, 0.5^2) \),

Model 3: \( \Lambda_\theta(Y_i) = 2X_i - 1 + \varepsilon_i \) where \( \varepsilon_1, \ldots, \varepsilon_n \sim \text{iid } \frac{2\mu_\varepsilon}{\sqrt{5}} \),

Model 4: \( \Lambda_\theta(Y_i) = 6X_i - 3 + \varepsilon_i \) where \( \varepsilon_1, \ldots, \varepsilon_n \sim \text{iid } N(0, 1) \).

In each model, \( \Lambda_\theta(Y) \) represents the Yeo and Johnson (2000) transformation, i.e,

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

and \( X_1, \ldots, X_n \) are independent uniform random variables on \([0, 1]\). For each model we will also consider three sample sizes: \( n = 100 \), \( n = 200 \) and \( n = 300 \) 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 \( \theta_0 \), the variability of the regression function \( m(x) \) (by comparing Model 4 to Model 1), 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 \( \varepsilon \sim \frac{2 \mu}{\sqrt{\delta}} \) 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 (2018), we will work with the unsmoothed estimator of the transformation \( \Lambda(\cdot) \). This estimator is arbitrarily close to the estimator \( \hat{\Gamma}_{LAD,b}(\hat{T}(\cdot)) \) when the bandwidth \( b \) is close to zero, and is defined as follows:

\[
\hat{\Gamma}_{LAD}(\hat{T}(y)) = \min_{q_m \in \mathbb{R}} \int \chi v(x) \left| \hat{\lambda}_1(\hat{T}(y), x) - q_m \right| dx.
\]

The smoothed estimator \( \hat{\Gamma}_{LAD,b}(\hat{T}(y)) \) that we use in the theory is used in order to facilitate the proofs of the asymptotic properties, see Colling and Van Keilegom (2018) 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 (2018), we approximate the expression \( \hat{\Gamma}_{LAD}(u) \) in practice by

\[
\hat{\Gamma}_{LAD}(u) = \text{median} \left( \hat{\lambda}_1(u, \tilde{x}_1), \ldots, \hat{\lambda}_1(u, \tilde{x}_{n_x}) \right),
\]

where \( \hat{\lambda}_1(u, x) = \hat{S}_1(u, x)/\hat{S}_1(1, x) \). The values \( \tilde{x}_1, \ldots, \tilde{x}_{n_x} \) 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 \( x_i^* \) for which the expression \( \hat{S}_1(u, x_i^*) \) diverges. This can happen if \( \frac{\partial}{\partial x_1} \hat{S}(w, x_i^*) \) is very close to 0 for some \( w \). Moreover, we also remove the \( x_i^* \)-values that are within 0.01 of the values \( x_i^* \) for which \( \hat{S}_1(u, x_i^*) \) diverges, even if the corresponding integrals do not diverge. Note that removing some \( x_i^* \)-values is allowed since \( \hat{S}_1(u, x)/\hat{S}_1(1, x) \) estimates \( \Gamma(u) = S_1(u, x)/S_1(1, x) \) for all \( x \in \chi \).

We consider the Epanechnikov kernel \( K(x) = \frac{3}{4}(1 - x^2)1_{\{|x| \leq 1\}} \) and we select the bandwidths \( h_x \) and \( h_u \) by the classical normal reference rule for kernel density estimation, i.e., \( \hat{h}_x = (40\sqrt{\pi})^{1/5}\hat{\sigma}_x n^{-1/5} \) and \( \hat{h}_u = (40\sqrt{\pi})^{1/5}\hat{\sigma}_u n^{-1/5} \), where \( \hat{\sigma}_x \) and \( \hat{\sigma}_u \) are the classical estimators of the standard deviation of \( X \) and \( U \) respectively. Finally, we take \( v(x) = 1 \) for all \( x \) such that \( \hat{S}_1(u, x) \) does not diverge.

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

\[
\tilde{\theta}_1 = \underset{\theta \in \Theta}{\text{argmin}} \ n^{-1} \sum_{i=1}^{n} \left( \Lambda_\theta(1) \Gamma_{LAD}(\hat{T}(Y_i)) - \Lambda_\theta(Y_i) \right)^2,
\]

and

\[
\tilde{\gamma}_2 = (\tilde{c}_1, \tilde{c}_2, \tilde{\theta}_2) = \underset{\gamma \in \Theta_\gamma}{\text{argmin}} \ n^{-1} \sum_{i=1}^{n} \left( c_1 \Gamma_{LAD}(\hat{T}(Y_i)) + c_2 - \Lambda_\theta(Y_i) \right)^2.
\]

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 \( \hat{\theta}_{PL} \) introduced in (1.2), 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 \( f_\epsilon(\theta)(\cdot) \), with the Epanechnikov kernel \( K \) and a bandwidth estimated by the classical normal reference rule for kernel density estimation. The estimator \( \hat{\theta}_{PL} \) is then obtained iteratively with the function \textit{optimize} in R. We refer to Colling and Van Keilegom (2016) for more details on the implementation of this estimator.

Tables 1 to 4 show the bias, variance and mean squared error of the profile likelihood estimator \( \hat{\theta}_{PL} \) and of our new estimators \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) for all considered values of \( n \) and \( \theta_0 \) when Models 1 to 4 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 \( \tilde{\theta}_2 \) outperforms \( \tilde{\theta}_1 \) in all scenarios in terms of variance and in most of the scenarios in terms of bias, which could be expected since \( \tilde{\theta}_2 \) offers more flexibility and freedom than \( \tilde{\theta}_1 \) as explained in Section 3. Hence, we will concentrate our following analysis on the comparison between \( \hat{\theta}_{PL} \) and \( \tilde{\theta}_2 \).

Next, when \( \theta_0 \) increases from 0 to 1, we observe globally that the bias of \( \hat{\theta}_{PL} \) and \( \tilde{\theta}_2 \) 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 \( \theta_0 \) increases from 1 to 1.5, the mean squared error of \( \hat{\theta}_{PL} \) and \( \tilde{\theta}_2 \) 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.
Using the same reasoning, it seems logical that $\hat{\theta}_{PL}$ and $\tilde{\theta}_2$ perform both better under Model 4 ($m(x) = 6x - 3$) than under Model 1 ($m(x) = 2x - 1$) in terms of bias and variance. Indeed, in Model 4, the regression function $m(x)$ is more variable than in Model 1, which helps for estimating $\theta$. Similarly, if we compare the results obtained under Model 1 to the ones obtained under Model 2, $\hat{\theta}_{PL}$ and $\tilde{\theta}_2$ 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 $\theta$.

Moreover, under Model 4, $\tilde{\theta}_2$ slightly outperforms $\hat{\theta}_{PL}$ for $n = 100$ and $n = 200$, whereas $\hat{\theta}_{PL}$ and $\tilde{\theta}_2$ perform equivalently for $n = 300$. However, despite the fact that $\hat{\theta}_{PL}$ and $\tilde{\theta}_2$ perform globally equivalently under Model 4, $\tilde{\theta}_2$ clearly outperforms $\hat{\theta}_{PL}$ under Models 1 and 2 for all considered values of $n$ and $\theta_0$, which 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, 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 $\varepsilon \sim \frac{2\pi^{\frac{1}{t}}}{\sqrt{5}}$ for all considered values of $n$ and $\theta_0$. Consequently, the distribution of the residuals also has an impact on the quality of the estimations of $\theta_0$, which is again an expected conclusion. In particular, residuals that are normally distributed help for estimating $\theta_0$. However, when $\varepsilon \sim \frac{2\pi^{\frac{1}{t}}}{\sqrt{5}}$, the estimator $\tilde{\theta}_2$ clearly outperforms the profile likelihood estimator $\hat{\theta}_{PL}$. Indeed, $\hat{\theta}_{PL}$ suffers considerably under Model 3.

In conclusion, the new estimator $\tilde{\theta}_2$ outperforms globally speaking the estimators $\tilde{\theta}_1$ and $\hat{\theta}_{PL}$, and especially when the model becomes more difficult to estimate (Models 1 to 3). In the latter case, the performance of the profile likelihood estimator drops significantly and $\tilde{\theta}_1$ also outperforms $\hat{\theta}_{PL}$.
| Model 1 | | $\theta_0 = 0$ | $\theta_0 = 0.5$ | $\theta_0 = 1$ | $\theta_0 = 1.5$ |
|--------|--------|-----------------|-----------------|-----------------|-----------------|
|        | $\hat{\theta}_{PL}$ Bias | 0.1065 | 0.0454 | 0.0061 | -0.0481 |
|        | Var | 0.0485 | 0.0489 | 0.0566 | 0.0564 |
|        | MSE | 0.0599 | 0.0509 | 0.0567 | 0.0587 |
|        | $\tilde{\theta}_1$ Bias | -0.0288 | 0.0224 | 0.0554 | 0.0708 |
|        | Var | 0.0347 | 0.0407 | 0.0451 | 0.0442 |
|        | MSE | 0.0355 | 0.0412 | 0.0482 | 0.0492 |
| $n = 100$ | $\tilde{\theta}_2$ Bias | -0.0413 | -0.0193 | 0.0045 | 0.0308 |
|        | Var | 0.0206 | 0.0284 | 0.0314 | 0.0288 |
|        | MSE | 0.0223 | 0.0288 | 0.0314 | 0.0297 |
|        | $\hat{\theta}_{PL}$ Bias | 0.0493 | 0.0232 | -0.0023 | -0.0204 |
|        | Var | 0.0176 | 0.0223 | 0.0228 | 0.0209 |
|        | MSE | 0.0200 | 0.0228 | 0.0228 | 0.0213 |
|        | $\tilde{\theta}_1$ Bias | -0.0338 | 0.0103 | 0.0457 | 0.0673 |
|        | Var | 0.0152 | 0.0194 | 0.0222 | 0.0202 |
|        | MSE | 0.0163 | 0.0195 | 0.0243 | 0.0248 |
|        | $\tilde{\theta}_2$ Bias | -0.0538 | -0.0319 | -0.0071 | 0.0187 |
|        | Var | 0.0086 | 0.0117 | 0.0133 | 0.0123 |
|        | MSE | 0.0115 | 0.0128 | 0.0133 | 0.0126 |
|        | $\hat{\theta}_{PL}$ Bias | 0.0344 | 0.0172 | 0.0009 | -0.0225 |
|        | Var | 0.0129 | 0.0157 | 0.0161 | 0.0158 |
|        | MSE | 0.0141 | 0.0160 | 0.0161 | 0.0163 |
| $n = 200$ | $\tilde{\theta}_1$ Bias | -0.0321 | 0.0087 | 0.0403 | 0.0559 |
|        | Var | 0.0131 | 0.0188 | 0.0204 | 0.0193 |
|        | MSE | 0.0141 | 0.0189 | 0.0221 | 0.0224 |
|        | $\tilde{\theta}_2$ Bias | -0.0499 | -0.0287 | -0.0040 | 0.0220 |
|        | Var | 0.0081 | 0.0114 | 0.0122 | 0.0113 |
|        | MSE | 0.0106 | 0.0123 | 0.0122 | 0.0118 |
|        | $\hat{\theta}_{PL}$ Bias | 0.0344 | 0.0172 | 0.0009 | -0.0225 |
|        | Var | 0.0129 | 0.0157 | 0.0161 | 0.0158 |
|        | MSE | 0.0141 | 0.0160 | 0.0161 | 0.0163 |
| $n = 300$ | $\tilde{\theta}_1$ Bias | -0.0321 | 0.0087 | 0.0403 | 0.0559 |
|        | Var | 0.0131 | 0.0188 | 0.0204 | 0.0193 |
|        | MSE | 0.0141 | 0.0189 | 0.0221 | 0.0224 |
|        | $\tilde{\theta}_2$ Bias | -0.0499 | -0.0287 | -0.0040 | 0.0220 |
|        | Var | 0.0081 | 0.0114 | 0.0122 | 0.0113 |
|        | MSE | 0.0106 | 0.0123 | 0.0122 | 0.0118 |

Table 1: Bias, variance (Var) and mean squared error (MSE) of the estimators $\hat{\theta}_{PL}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ for different sample sizes and values of $\theta_0$ when $m(x) = 2x - 1$ and $\varepsilon \sim N(0, 1)$ (model 1).
| Model 2 | \( \theta_0 = 0 \) | \( \theta_0 = 0.5 \) | \( \theta_0 = 1 \) | \( \theta_0 = 1.5 \) |
|---------|-----------------|-----------------|-----------------|-----------------|
| \( \hat{\theta}_{PL} \) Bias  | 0.0555 | 0.0232 | -0.0109 | -0.0394 |
| Var     | 0.0487 | 0.0516 | 0.0615 | 0.0594 |
| MSE     | 0.0518 | 0.0521 | 0.0616 | 0.0610 |
| \( \tilde{\theta}_1 \) Bias  | -0.0065 | 0.0449 | 0.0680 | 0.0634 |
| Var     | 0.0414 | 0.0483 | 0.0504 | 0.0468 |
| MSE     | 0.0414 | 0.0504 | 0.0550 | 0.0508 |
| \( \tilde{\theta}_2 \) Bias  | -0.0392 | -0.0142 | 0.0125 | 0.0380 |
| Var     | 0.0237 | 0.0299 | 0.0321 | 0.0305 |
| MSE     | 0.0253 | 0.0301 | 0.0323 | 0.0320 |

| \( \hat{\theta}_{PL} \) Bias  | 0.0251 | 0.0128 | 0.0008 | -0.0209 |
| Var     | 0.0200 | 0.0221 | 0.0241 | 0.0280 |
| MSE     | 0.0206 | 0.0223 | 0.0241 | 0.0284 |
| \( \tilde{\theta}_1 \) Bias  | -0.0083 | 0.0365 | 0.0653 | 0.0601 |
| Var     | 0.0179 | 0.0236 | 0.0252 | 0.0245 |
| MSE     | 0.0179 | 0.0250 | 0.0295 | 0.0281 |
| \( \tilde{\theta}_2 \) Bias  | -0.0494 | -0.0249 | 0.0016 | 0.0283 |
| Var     | 0.0111 | 0.0151 | 0.0162 | 0.0156 |
| MSE     | 0.0136 | 0.0157 | 0.0162 | 0.0164 |

| \( \hat{\theta}_{PL} \) Bias  | 0.0216 | 0.0066 | -0.0007 | -0.0071 |
| Var     | 0.0154 | 0.0177 | 0.0188 | 0.0168 |
| MSE     | 0.0159 | 0.0177 | 0.0188 | 0.0169 |
| \( \tilde{\theta}_1 \) Bias  | -0.0131 | 0.0322 | 0.0583 | 0.0544 |
| Var     | 0.0152 | 0.0199 | 0.0205 | 0.0196 |
| MSE     | 0.0154 | 0.0209 | 0.0239 | 0.0226 |
| \( \tilde{\theta}_2 \) Bias  | -0.0515 | -0.0283 | -0.0027 | 0.0224 |
| Var     | 0.0096 | 0.0130 | 0.0141 | 0.0131 |
| MSE     | 0.0122 | 0.0138 | 0.0141 | 0.0136 |

Table 2: Bias, variance (Var) and mean squared error (MSE) of the estimators \( \hat{\theta}_{PL} \), \( \tilde{\theta}_1 \) and \( \tilde{\theta}_2 \) for different sample sizes and values of \( \theta_0 \) when \( m(x) = 2x - 1 \) and \( \varepsilon \sim N(0, 0.5^2) \) (model 2).
| Model 3 | $\theta_0 = 0$ | $\theta_0 = 0.5$ | $\theta_0 = 1$ | $\theta_0 = 1.5$ |
|---------|----------------|-----------------|----------------|----------------|
|        | $\hat{\theta}_{PL}$ | Bias | 0.1163 | 0.0381 | -0.0236 | -0.0825 | |
|        |                  | Var | 0.0777 | 0.0800 | 0.0941 | 0.1003 | |
|        |                  | MSE | 0.0912 | 0.0815 | 0.0947 | 0.1071 | |
| $n = 100$ | $\tilde{\theta}_1$ | Bias | -0.0555 | 0.0038 | 0.0473 | 0.0655 | |
|        |                  | Var | 0.0353 | 0.0443 | 0.0479 | 0.0452 | |
|        |                  | MSE | 0.0383 | 0.0443 | 0.0501 | 0.0495 | |
|        | $\tilde{\theta}_2$ | Bias | -0.0593 | -0.0320 | 0.0007 | 0.0289 | |
|        |                  | Var | 0.0228 | 0.0328 | 0.0376 | 0.0355 | |
|        |                  | MSE | 0.0264 | 0.0339 | 0.0376 | 0.0363 | |
| $n = 200$ | $\hat{\theta}_{PL}$ | Bias | 0.0939 | 0.0450 | -0.0007 | -0.0520 | |
|        |                  | Var | 0.0352 | 0.0412 | 0.0445 | 0.0421 | |
|        |                  | MSE | 0.0440 | 0.0433 | 0.0445 | 0.0448 | |
|        | $\tilde{\theta}_1$ | Bias | -0.0368 | 0.0120 | 0.0482 | 0.0747 | |
|        |                  | Var | 0.0174 | 0.0246 | 0.0283 | 0.0261 | |
|        |                  | MSE | 0.0188 | 0.0247 | 0.0306 | 0.0316 | |
|        | $\tilde{\theta}_2$ | Bias | -0.0553 | -0.0263 | 0.0064 | 0.0368 | |
|        |                  | Var | 0.0131 | 0.0197 | 0.0222 | 0.0201 | |
|        |                  | MSE | 0.0162 | 0.0204 | 0.0222 | 0.0214 | |
| $n = 300$ | $\hat{\theta}_{PL}$ | Bias | 0.0687 | 0.0408 | 0.0104 | -0.0232 | |
|        |                  | Var | 0.0208 | 0.0299 | 0.0321 | 0.0304 | |
|        |                  | MSE | 0.0256 | 0.0315 | 0.0322 | 0.0309 | |
|        | $\tilde{\theta}_1$ | Bias | -0.0328 | 0.0170 | 0.0590 | -0.0803 | |
|        |                  | Var | 0.0098 | 0.0152 | 0.0177 | 0.0185 | |
|        |                  | MSE | 0.0109 | 0.0155 | 0.0212 | 0.0250 | |
|        | $\tilde{\theta}_2$ | Bias | -0.0555 | -0.0273 | 0.0045 | 0.0352 | |
|        |                  | Var | 0.0084 | 0.0132 | 0.0155 | 0.0151 | |
|        |                  | MSE | 0.0115 | 0.0140 | 0.0155 | 0.0163 | |

Table 3: Bias, variance (Var) and mean squared error (MSE) of the estimators $\hat{\theta}_{PL}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ for different sample sizes and values of $\theta_0$ when $m(x) = 2x - 1$ and $\epsilon \sim \mathcal{N}(0, 5)$ (model 3).
Table 4: Bias, variance (Var) and mean squared error (MSE) of the estimators $\hat{\theta}_{PL}$, $\tilde{\theta}_1$ and $\tilde{\theta}_2$ for different sample sizes and values of $\theta_0$ when $m(x) = 6x - 3$ and $\varepsilon \sim N(0, 1)$ (model 4).
6 Appendix: Proofs

6.1 Assumptions

The following conditions are needed for the main results of this paper. They are related to the distribution of \( \varepsilon \), the transformations \( \Gamma \) and \( \Lambda_{\theta} \), 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 \( \Delta_1 \) and \( \Delta_2 \), defined in the statement of Theorem 4.2.

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

(A2) The transformation \( \Gamma \) 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 \( m \), i.e. \( \int K(z) \, dz = 1 \), \( \int z^\ell K(z) \, dz = 0 \) for \( \ell = 1, \ldots, m-1 \) and \( \int z^m K(z) \, dz < \infty \), \( K \) is \( m \)-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 \( \sqrt{n} \max(h_x, h_u)^m \rightarrow 0 \), \( nb^4 \rightarrow \infty \) and \( b\sqrt{nh_x^d \min(h_x, h_u)^2} \log n^{-1} \rightarrow \infty \).

(A7) The joint density function \( f_{Y,X} \) of \( (Y, X) \) is uniformly bounded and \( m+2 \)-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) \( \inf_{x \in \chi_0} f_X(x) > 0 \), \( \inf_{(u,x) \in \mathcal{U}_0 \times \chi_0} \left| \frac{\partial}{\partial x_1} \varphi(u, x) \right| > 0 \) and \( \inf_{x \in \chi_0} |S_1(1, x)| > 0 \).
(A9) The weight function \( v \) has compact support \( \chi_0 \subseteq A \) with nonempty interior and satisfies \( \int_{\chi_0} v(x) \, dx = 1 \). Moreover, \( v \) is continuous on \( \chi \) and is \( m \)-times continuously differentiable on \( \chi_0 \).

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

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

for \( l \in \{0, 1, 2\} \) and

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

for \( i, j \in \{0, 1, 2\} \) such that \( 0 \leq i + j \leq 2 \).

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

(A13) The matrices \( \Delta_1 \) and \( \Delta_2 \) are of full rank.

Assumptions (A1)-(A9) are basically the same as in Colling and Van Keilegom (2018) and are required since our proofs rely on the weak convergence of the estimator \( \tilde{\Gamma}_{LAD,b}(\tilde{T}(y)) \) that is established in the latter paper. 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).

### 6.2 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 \( \theta_0 \) and the second and the third results establish respectively the consistency and the asymptotic distributions of these new estimators.
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

$$\varphi(u, x) = P\left(\Gamma(U) \leq \Gamma(u) \mid X = x\right) = P\left(m(X) + \varepsilon \leq \Gamma(u) \mid X = x\right) = F_\varepsilon\left(\Gamma(u) - m(x)\right),$$

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

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

and

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

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. \hfill \Box

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

Proposition 6.1. Assume (A1)-(A9). Then,

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

(ii) $\log N(\varepsilon, \mathcal{H}, \|\cdot\|_{L_2}) \leq K \varepsilon^{-1}$ for some $K < \infty$.

Proof. First, we will prove (i). It is clear that $\hat{T}$ is monotone and that $\hat{T}(y) \in \mathcal{U}_0$ 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 $\sup_y |\hat{T}(y) - T(y)| = o_P(1)$. Hence, it suffices to show that $\hat{\Gamma}_{LAD,b} \in C^1_c(\mathcal{U}_0)$ 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

$$\sup_{u \in \mathcal{U}_0} |\hat{\Gamma}_{LAD,b}(u) - \Gamma(u)| = o_P(1) \quad \text{and} \quad \sup_{u \in \mathcal{U}_0} |\hat{\Gamma}'_{LAD,b}(u) - \Gamma'(u)| = o_P(1).$$
The former follows from Theorem 5.2 in Colling and Van Keilegom (2018), which establishes the weak convergence of the process $\sqrt{n}(\hat{\Gamma}_{LAD,b}(\cdot) - \Gamma(\cdot))$ as a process defined on $\mathcal{U}_0$. For the latter result, define

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

and note that by construction $Q_n(\hat{\Gamma}_{LAD,b}(u), u) = 0$ 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 $Q_n(\hat{\Gamma}_{LAD,b}(u), u)$ with respect to $u$ is also equal to 0, i.e.

$$P_1n(\hat{\Gamma}_{LAD,b}(u), u) + P_2n(\hat{\Gamma}_{LAD,b}(u), u)\hat{\Gamma}_{LAD,b}(u) = 0,$$

where $P_1n(q_m, u) = \frac{\partial}{\partial q_m} Q_n(q_m, u)$ and $P_2n(q_m, u) = \frac{\partial}{\partial q_m} Q_n(q_m, u)$. Similarly, defining

$$Q(q_m, u) = \frac{\partial}{\partial q_m} R(q_m, u)$$

and

$$Q(q_m, u) = \int_{\chi} v(x)\left[ -2L_b(\lambda_1(u, x) - q_m) + 1 - 2(\lambda_1(u, x) - q_m)\right] dx,$$

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 $P_1(\Gamma(u), u) + P_2(\Gamma(u), u)\Gamma'(u) = 0$, where $P_1(q_m, u) = \frac{\partial}{\partial q_m} Q(q_m, u)$ and $P_2(q_m, u) = \frac{\partial}{\partial q_m} Q(q_m, u)$. It follows that

$$P_1n(\hat{\Gamma}_{LAD,b}(u), u) - P_1(\Gamma(u), u)$$

$$+ \left[ P_2n(\hat{\Gamma}_{LAD,b}(u), u) - P_2(\Gamma(u), u) \right] \Gamma'(u) + P_2n(\hat{\Gamma}_{LAD,b}(u), u)\left(\hat{\Gamma}_{LAD,b}(u) - \Gamma'(u)\right) = 0.$$
where \( L_b^o(\cdot) = L^o(\cdot/b)/b^2 \). Using the same arguments as at the end of the proof of Theorem 5.2 in Colling and Van Keilegom (2018), 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

\[
\inf_{u \in \mathcal{U}_0} |P_{2n}(\hat{\Gamma}_{LAD,b}(u), u)| \geq \inf_{u \in \mathcal{U}_0} |P_2(\Gamma(u), u)| - \sup_{u \in \mathcal{U}_0} |P_{2n}(\hat{\Gamma}_{LAD,b}(u), u) - P_2(\Gamma(u), u)|
= 4L_b^o(0) + o_P(b^{-1}).
\]

In conclusion,

\[
\sup_{u \in \mathcal{U}_0} |\hat{\Gamma}_{LAD,b}'(u) - \Gamma'(u)| \leq \left( \inf_{u \in \mathcal{U}_0} |P_{2n}(\hat{\Gamma}_{LAD,b}(u), u)| \right)^{-1} o_P(b^{-1}) = o_P(1), \]

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 \( f \in C^1_b(\mathcal{U}_0) \) 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 \( C^1_b(\mathcal{U}_0) \) 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 \( N_{1\epsilon} = N_{1\epsilon}(\epsilon, C^1_b(\mathcal{U}_0), \| \cdot \|_{L_\infty}) \leq \exp(K_1\epsilon^{-1}) \) 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 \( C^1_b(\mathcal{U}_0) \). Next, Theorem 2.7.5 in Van der Vaart and Wellner (1996) shows that the \( \epsilon \)-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 \( N_{1\epsilon}(\epsilon, \mathcal{H}, \| \cdot \|_{L_2}) \) 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,

\[
E\left[ (h_{j,k,u}(Y) - h_{j,k,\ell}(Y))^2 \right] \\
\leq E\left[ (h_{j,k,u}(Y) - f_{j,u}(g_{k,u}(Y)))^2 \right] + E\left[ (f_{j,u}(g_{k,u}(Y)) - f_{j,u}(g_{k,\ell}(Y)))^2 \right] \\
+ E\left[ (f_{j,u}(g_{k,\ell}(Y)) - f_{j,\ell}(g_{k,\ell}(Y)))^2 \right] + E\left[ (f_{j,\ell}(g_{k,\ell}(Y)) - h_{j,k,\ell}(Y))^2 \right];
\]

24
and each of these four terms is easily seen to be bounded by a finite multiple of $\epsilon^2$. 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 $\ell_1, M_1$ and $M_{n,1}$ in case (i) and on the functions $\ell_2, M_2$ 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 $M_{n,1}(\hat{\theta}_1, \hat{h}_b) = 0$ and $M_{n,2}(\hat{\gamma}_2, \hat{h}_b) = 0$ by construction, where $\hat{h}_b(\cdot) = \hat{\Gamma}_{LAD,b}(\hat{T}(\cdot))$.

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 $\sup_{\gamma \in \Theta} ||M_2(\gamma, h) - M_2(\gamma, h_0)|| < \varepsilon$ in case (ii), where $h_0(\cdot) = \Gamma(T(\cdot))$. First, in case (i), we have for $j = 1, \ldots, k$:

$$
\left| M_{1,j}(\theta, h) - M_{1,j}(\theta, h_0) \right| = \left| E \left[ w(Y) \left\{ A_j(\theta) \left( h^2(Y) - h_0^2(Y) \right) + B_j(\theta, Y) \left( h(Y) - h_0(Y) \right) \right\} \right] \right|.
$$

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

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

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

$$
K_{1,j} = \sup_{y \in \mathcal{Y}_0} \left( \sup_{\theta \in \Theta} \left| A_j(\theta) \right| \left( \left| E \left( \left( h(Y) + h_0(Y) \right) \right)^2 \right|^{1/2} + E \left( \sup_{\theta \in \Theta} \left| B_j^2(\theta, Y) \right| \right)^{1/2} \right) + E \left( \sup_{\theta \in \Theta} \left| C_j^2(\theta, Y) \right| \right)^{1/2} \right),
$$

which is finite for all $j = 1, \ldots, k$ 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 $\sup_{\theta \in \Theta} \left| M_1(\theta, h) - M_1(\theta, h_0) \right| < \varepsilon$. In case (ii), we use exactly the same reasoning.

Condition (1.4) in Chen et al. (2003) is directly ensured by the fact that $||\hat{h}_b - h_0||_{\mathcal{H}} \leq ||\hat{h}_b - h_0||_{L_{\infty}} = O_P(n^{-1/2}) = o_P(1)$ by Corollary 5.1 in Colling and Van Keilegom (2018), where $|| \cdot ||_{L_{\infty}}$ is the supremum norm over $\mathcal{Y}_0$. 

25
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

\[ \mathcal{F}_i = \{ y \rightarrow \ell_i(y, \theta, h) : \theta \in \Theta, h \in \mathcal{H} \} \]

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 \((\theta_0, h_0)\).

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

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

(6.1)

We will show (6.1) for \( i = 1 \). The derivation for \( i = 2 \) is very similar. Since \( \Theta \) 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 \( \Theta \) for some \( C < \infty \). Proposition 6.1(ii) shows that \( N_{\epsilon, \mathcal{H}} = N_{||| \cdot |||_{L_2}} \leq \exp(K\epsilon^{-1}) \) for some \( K < \infty \). Note that

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

for \( j = 1, \ldots, k \). Fix \( \theta \in \Theta \) and \( h \in \mathcal{H} \), and suppose that \( \theta_{k_1,\ell} \leq \theta \leq \theta_{k_1,u} \) (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 \( \Theta \), and that \( h_{k_2,\ell} \leq h \leq h_{k_2,u} \), where \((h_{k_2,\ell}, h_{k_2,u})\) is one of the \( N_{\epsilon, \mathcal{H}} \) brackets for the space \( \mathcal{H} \). Let

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

and let

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

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

\[
| \ell_{1,j,k_1,k_2,u}(y) - \ell_{1,j,k_1,k_2,\ell}(y) | \\
\leq w(y) \left[ c^2 \sup_{\theta \in \Theta} ||A_j(\theta)|| ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + 2c \sup_{\theta \in \Theta} |A_j(\theta)||h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
+ c \sup_{\theta \in \Theta} ||B_j(\theta, y)|| ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + \sup_{\theta \in \Theta} |B_j(\theta, y)||h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
+ c \sup_{\theta \in \Theta} ||C_j(\theta, y)|| ||\theta_{k_1,u} - \theta_{k_1,\ell}|| + \sup_{\theta \in \Theta} |C_j(\theta, y)||h_{k_2,u}(y) - h_{k_2,\ell}(y)| \\
+ \sup_{\theta \in \Theta} ||D_j(\theta, y)|| ||\theta_{k_1,u} - \theta_{k_1,\ell}|| \right].
\]
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,
\[
E\left[ \left| \ell_{1,j,k_1,k_2,u}(Y) - \ell_{1,j,k_1,k_2,\ell}(Y) \right|^2 \right]
\leq 2^6 \left( \sup_{y \in Y_0} w(y) \right)^2 \left[ c^4 \sup_{\theta \in \Theta} \left| \dot{A}_j(\theta) \right|^2 \left| \theta_{k_1,u} - \theta_{k_1,\ell} \right|^2 + 4c^2 \sup_{\theta \in \Theta} \left| A_j(\theta) \right|^2 \left| h_{k_2,u} - h_{k_2,\ell} \right|^2 \right.
\]
\[
+ c^2 \sup_{\theta \in \Theta} \left| \dot{B}_j(\theta, \cdot) \right|^2 \left| \theta_{k_1,u} - \theta_{k_1,\ell} \right|^2 + \sup_{\theta \in \Theta} \left| B_j(\theta, \cdot) \right|^2 \left| h_{k_2,u} - h_{k_2,\ell} \right|^2
\]
\[
+ c^2 \sup_{\theta \in \Theta} \left| \dot{C}_j(\theta, \cdot) \right|^2 \left| \theta_{k_1,u} - \theta_{k_1,\ell} \right|^2 + \sup_{\theta \in \Theta} \left| C_j(\theta, \cdot) \right|^2 \left| h_{k_2,u} - h_{k_2,\ell} \right|^2
\]
\[
+ \sup_{\theta \in \Theta} \left| \dot{D}_j(\theta, \cdot) \right|^2 \left| \theta_{k_1,u} - \theta_{k_1,\ell} \right|^2 \right],
\]
and this is bounded by a finite multiple of $\epsilon^2$ by assumptions (A10) and (A11). It now follows that the integral in (6.1) is finite.

Finally, for Condition (C2) we need to show that $E[\left| \ell_i(Y, \theta, h) - \ell_i(Y, \theta_0, h_0) \right|^2] \to 0$ as $\|\theta - \theta_0\| \to 0$ and $\|h - h_0\|_H \to 0$ ($i = 1, 2$). This is easily seen to hold true thanks to assumptions (A10) and (A11) and calculations similar to those leading to (6.2) 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
\[
\Gamma_1(\theta, h_0)[h - h_0] = \lim_{\tau \to 0} \tau^{-1} \left( M_1(\theta, h_0 + \tau(h - h_0)) - M_1(\theta, h_0) \right) = \left( \Gamma_{1,j}(\theta, h_0)[h - h_0] \right)_{j = 1, \ldots, k},
\]
where $\Gamma_{1,j}(\theta, h_0)[h - h_0] = E \left[ w(Y) \left\{ 2A_j(\theta)h_0(Y) + B_j(\theta, Y) + C_j(\theta, Y) \right\} (h(Y) - h_0(Y)) \right]$, and
\[
\Gamma_2(\gamma, h_0)[h - h_0] = \lim_{\tau \to 0} \tau^{-1} \left( M_2(\gamma, h_0 + \tau(h - h_0)) - M_2(\gamma, h_0) \right) = E \left( \begin{array}{c}
w(Y) \left( 2c_1h_0(Y) + c_2 - \Lambda_\theta(Y) \right) (h(Y) - h_0(Y)) \\
w(Y)c_1(h(Y) - h_0(Y)) \\
-w(Y)\Lambda_{\theta,1}(Y)c_1(h(Y) - h_0(Y)) \\
\vdots \\
-w(Y)\Lambda_{\theta,k}(Y)c_1(h(Y) - h_0(Y)) \end{array} \right).
\]
The \( j \)th 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 \( j = 1, \ldots, k + 2 \).

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

\[
\begin{align*}
| M_{1,j}(\theta, h) - M_{1,j}(\theta, h_0) - \Gamma_{1,j}(\theta, h_0)[h - h_0] |
&= \left| E\left\{ w(Y) A_j(\theta) \left[ h^2(Y) - h_0^2(Y) - 2h_0(Y)(h(Y) - h_0(Y)) \right] \right\} \right| \\
&= \left| E\left\{ w(Y) A_j(\theta) \left( h(Y) - h_0(Y) \right)^2 \right\} \right| \\
&\leq \beta_{1,j} ||h - h_0||_H^2,
\end{align*}
\]

where \( \beta_{1,j} = \sup_{\theta \in \Theta} |A_j(\theta)| \sup_{y \in \mathcal{Y}_0} w(y) < \infty \) by assumption (A10). Hence, \( ||M_1(\theta, h) - M_1(\theta, h_0) - \Gamma_1(\theta, h_0)[h - h_0]|| \leq \beta_1 ||h - h_0||_H \) 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 \( \Theta_{\delta_n} = \{ \theta \in \Theta : ||\theta - \theta_0|| \leq \delta_n \} \) and \( \mathcal{H}_{\delta_n} = \{ h \in \mathcal{H} : ||h - h_0||_H \leq \delta_n \} \). In case (i) and for \( j = 1, \ldots, k \), we have

\[
\begin{align*}
| \Gamma_{1,j}(\theta, h_0)[h - h_0] - \Gamma_{1,j}(\theta_0, h_0)[h - h_0] |
&\leq \left| E\left\{ 2w(Y) \left( A_j(\theta) - A_j(\theta_0) \right) h_0(Y) \left( h(Y) - h_0(Y) \right) \right\} \right| \\
&+ \left| E\left\{ w(Y) \left( B_j(\theta,Y) - B_j(\theta_0,Y) \right) \left( h(Y) - h_0(Y) \right) \right\} \right| \\
&+ \left| E\left\{ w(Y) \left( C_j(\theta,Y) - C_j(\theta_0,Y) \right) \left( h(Y) - h_0(Y) \right) \right\} \right| . \tag{6.3}
\end{align*}
\]

Using the mean value theorem and since \( E[h_0(Y)(h(Y) - h_0(Y))] \leq ||h_0||_H ||h - h_0||_H \) by Cauchy-Schwarz inequality, the first term on the right hand side of (6.3) is bounded by

\[ 2 \sup_{y \in \mathcal{Y}_0} (w(y)) \sup_{\theta \in \Theta} ||\dot{A}_j(\theta)|| ||\theta - \theta_0|| ||h_0||_H ||h - h_0||_H, \]

where \( \dot{A}_j(\theta) = \frac{\partial}{\partial \theta} A_j(\theta) \). Similarly, the second and third terms on the right hand side of (6.3) are bounded by

\[ \sup_{y \in \mathcal{Y}_0} (w(y)) \left[ E \left( \sup_{\theta \in \Theta} ||\dot{B}_j(\theta,Y)||^2 \right) \right]^{1/2} ||\theta - \theta_0|| ||h - h_0||_H. \]
and
\[
\sup_{y \in \mathcal{Y}_0} (w(y)) \left[ E \left( \sup_{\theta \in \Theta} \left| \hat{C}_j(\theta, Y) \right| \right)^2 \right]^{1/2} \| \theta - \theta_0 \| \| h - h_0 \|_H,
\]
where \( \hat{B}_j(\theta, y) = \frac{\partial}{\partial \theta} B_j(\theta, y) \) and \( \hat{C}_j(\theta, y) = \frac{\partial}{\partial \theta} C_j(\theta, y) \). These last three terms are bounded by \( o(1)\delta_n \) since \( \| h - h_0 \|_H \leq \delta_n = o(1), \| \theta - \theta_0 \| \leq \delta_n, E(h_0^2(Y)) < \infty \) since \( h_0 \in H \), \( \sup_{y \in \mathcal{Y}_0} w(y) < \infty \) by assumption (A10) and \( \sup_{\theta \in \Theta} \| \hat{A}_j(\theta) \|, E(\sup_{\theta \in \Theta} ||\hat{B}_j(\theta, Y)||^2) \) and \( E(\sup_{\theta \in \Theta} ||\hat{C}_j(\theta, Y)||^2) \) are finite by assumption (A11). The proof of Condition (2.3)(ii) in Chen et al. (2003) is satisfied in both cases using Proposition 6.1(i) and since we have shown in the proof of Theorem 4.1 that \( \| \hat{h}_b - h_0 \|_H = O_P(n^{-1/2}) = o_P(n^{-1/4}) \).

Moreover, Condition (2.4) in Chen et al. (2003) is satisfied in both cases using Proposition 6.1(ii) and since we have shown in the proof of Theorem 4.1 that \( \| \hat{h}_b - h_0 \|_H = O_P(n^{-1/2}) = o_P(n^{-1/4}) \).

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 \( j = 1, \ldots, k \), we have:
\[
\Gamma_{1,j}(\theta_0, h_0)[\hat{h}_b - h_0] = E \left[ w(Y) \left\{ 2A_j(\theta_0)h_0(Y) + B_j(\theta_0, Y) + C_j(\theta_0, Y) \right\} (\hat{h}_b(Y) - h_0(Y)) \right].
\]
Using Corollary 5.1 in Colling and Van Keilegom (2018), we have \( \hat{h}_b(y) - h_0(y) = n^{-1} \sum_{i=1}^n \varphi_{X_i,Y_i}^v(y) + o_P(n^{-1/2}) \) uniformly in \( y \in \mathcal{Y}_0 \). Consequently,
\[
M_{n,1}(\theta_0, h_0) + \Gamma_{1,j}(\theta_0, h_0)[\hat{h}_b - h_0] = n^{-1} \sum_{i=1}^n E \left[ w(Y) \left\{ 2A_j(\theta_0)h_0(Y) + B_j(\theta_0, Y) + C_j(\theta_0, Y) \right\} \varphi_{X_i,Y_i}^v(Y) \right| X_i, Y_i] + o_P(n^{-1/2}),
\]
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 \( E(\varphi_{X,Y}^v(y)) = 0 \) for all \( y \) by Corollary 5.1 in Colling and.
Van Keilegom (2018). Similarly, in case (ii), we have

\[
M_{n,2}(\gamma_0, h_0) + \Gamma_2(\gamma_0, h_0)[\hat{h}_b - h_0] = n^{-1} \sum_{i=1}^{n} \left( \begin{array}{c}
E\{w(Y)(2c_{1,0}h_0(Y) + c_{2,0} - \Lambda_{\theta_0}(Y))\varphi_{X_i,Y_i}^v(Y)|X_i,Y_i\} \\
E\{w(Y)c_{1,0}\varphi_{X_i,Y_i}^v(Y)|X_i,Y_i\} \\
- E\{w(Y)\hat{\Lambda}_{\theta_0,1}(Y)c_{1,0}\varphi_{X_i,Y_i}^v(Y)|X_i,Y_i\} \\
\vdots \\
- E\{w(Y)\hat{\Lambda}_{\theta_0,k}(Y)c_{1,0}\varphi_{X_i,Y_i}^v(Y)|X_i,Y_i\}
\end{array} \right) + o_P(n^{-1/2}).
\]

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

\[\square\]

References

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

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.

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

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.

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

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

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

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

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.

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

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.

Colling, B. and Van Keilegom, I. (2018). Estimation of fully nonparametric transformation models. *Bernoulli (under revision)*.

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

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

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

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

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

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

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*.

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

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.

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

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 Empirical Processes*. Springer.

Vanhems, A. and Van Keilegom, I. (2018). Semiparametric transformation model with endogeneity: a control function approach. *Econometric Theory (to appear)*.

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

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