Submitted to Bernoulli
arXiv: arXiv:0000.0000

Estimation of fully nonparametric transformation models

BENJAMIN COLLING[1] and INGRID VAN KEILEGOM[1,**]

[1] *Research Centre for Operations Research and Business Statistics (ORSTAT), KU Leuven.*
E-mail: [**] ingrid.vankeilegom@kuleuven.be

Consider the following nonparametric transformation model $\Lambda(Y) = m(X) + \varepsilon$, where X is a d-dimensional covariate, Y is a continuous univariate dependent variable and ε is an error term with zero mean and which is independent of X. We assume that the unknown transformation Λ is strictly increasing and that m is an unknown regression function. Our goal is to develop two new nonparametric estimators of the transformation Λ, the first one based on the least squares loss and the second one based on the least absolute deviation loss, and to compare their performance with that of the estimators developed by Chiappori et al. (2015). Our proposed estimators are based on an estimator of the conditional distribution of U given X, where U is an appropriate transformation of Y that is uniformly distributed. The main motivation for working with U instead of Y is that, in transformation models, the response Y is often skewed with very long tails, and so kernel smoothing based on Y does not work well. Hence, we expect to obtain better estimators if we pre-transform Y before applying kernel smoothing. We establish the asymptotic normality of the two proposed estimators. We also carry out a simulation study to illustrate the performance of our estimators, to compare these new estimators with the ones of Chiappori et al. (2015) and to see under which model conditions which estimators behave the best.

Keywords: Asymptotic properties, Identification, Kernel smoothing, Nonparametric regression, Nonparametric transformation.

1. Introduction

Transforming the response variable Y of a simple linear regression model is a very common practice when we want to induce additivity of the effects, homoscedasticity and normality of the new error term, or reduce skewness, in order to satisfy the assumptions that are made on this model. The most commonly used family of transformations is the parametric family of power transformations, which was introduced by Box and Cox (1964) and which includes as special cases the logarithm and the identity. This class of transformations has been generalized, see for example the Yeo and Johnson (2000) transform.

The literature related to transformation models is very large. It contains papers in which the transformation of the response and the regression function are either parametric or nonparametric, depending on the particular objectives of the considered paper.

The two above mentioned papers, as well as the papers of Zellner and Revankar (1969), John and Draper (1980), Bickel and Doksum (1981), Carroll and Ruppert (1988), MacKinnon and Magee (1990) and Sakia (1992) among others, consider a model where both the transformation and the regression function are parametric.

We can also find papers in which the transformation is parametric and the regression function is nonparametric. In particular, Linton et al. (2008) proposed different estimators of the transformation parameter and established their asymptotic properties. We also refer to Colling et al. (2015) and Heuchenne et al. (2015) who introduced and studied respectively nonparametric estimators for the error density function and the error distribution function. Moreover, Colling and Van Keilegom (2016) and Colling and Van Keilegom (2017) developed tests for the parametric form of the regression function respectively based on the error distribution function and on the integrated regression function. Finally, we also mention the work of Vanhems and Van Keilegom (2018) who studied the estimation of this model supposing that some of the regressors are endogenous and the work of Neumeyer et al. (2016) who introduced estimators for the different components of a heteroscedastic transformation model and proved the asymptotic normality of these estimators.

Moreover, we refer to the important work of Horowitz (1996) who proposed estimators of the different components of a regression model, in which the transformation is nonparametric and the regression function is parametric. For the same model, Chen (2002) proposed a rank-based estimator for the transformation that has the advantage of not involving nonparametric smoothing.

In this paper, we will consider a model in which both the transformation and the regression function are nonparametric, i.e., we will consider a nonparametric transformation model of the form

$$\Lambda(Y) = m(X) + \varepsilon \ , \tag{1.1}$$

where $\Lambda(\cdot)$ and $m(\cdot)$ are respectively an unknown transformation and an unknown regression function. We also assume that $\Lambda(\cdot)$ is a strictly increasing function and X has compact support $\chi \subset \mathbb{R}^d$. Moreover, we assume that Y is a continuous univariate response variable, X is a d-dimensional covariate and the error term ε has zero mean and is independent of X. We denote by F_X, f_X, F_ε and f_ε the distribution and the probability density functions of X and ε respectively.

First, we would like to mention the works of Breiman and Friedman (1985), Horowitz (2001) and Jacho-Chavez et al. (2010) among others. They developed fully nonparametric estimators but in contexts that are slightly different from model (1.1). Breiman and Friedman (1985) constructed an algorithm for estimating the different components of model (1.1) when the regression function m is supposed to be additive. Moreover, Horowitz (2001) proposed a nonparametric estimation of a generalized additive model with an unknown link function and Jacho-Chavez et al. (2010) discussed the identification and the estimation of the unknown functions H, M, G and F, where $r(x,z) = H\{M(x,z)\}$ and $M(x,z) = G(x) + F(z)$.

Next, we refer to the paper of Chiappori et al. (2015) who developed sufficient conditions so that a nonparametric transformation model, allowing endogenous and exogenous

regressors, is identifiable. They also constructed nonparametric estimators of the transformation, the regression function and the conditional distribution function of ε given X^*, where X^* is the vector containing the potentially endogenous component, and analyzed their asymptotic properties. Finally, they developed a test to conclude about the exogeneity of a regressor or not.

In this paper, we will focus on the problem of nonparametric estimation of the transformation Λ. We will define two new nonparametric estimators of the transformation Λ in the case where all the components of the vector X are exogenous, and we will show that it often outperforms the ones constructed by Chiappori et al. (2015).

The main idea of the nonparametric estimators of Λ developed in Chiappori et al. (2015) is to consider the conditional distribution function of Y given X, denoted by $\varphi_C(y,x)$, and to divide its derivative with respect to y, denoted by $\varphi_{C,y}(y,x)$, by its derivative with respect to x_1, denoted by $\varphi_{C,1}(y,x)$, where x_1 is the first component of the vector x. This particular way of constructing a nonparametric estimator of the transformation was already considered in Horowitz (1996) but in the case of a parametric regression function. More precisely, under the normalization conditions (I1) introduced in Section 2 and under appropriate conditions, Chiappori et al. (2015) proved in their Theorem 1 that Λ is globally identified as $\Lambda(y) = \frac{S_{C,1}(y,x)}{S_{C,1}(1,x)}$ for any $x \in \chi$ where $S_{C,1}(y,x) = \int_0^y \frac{\varphi_{C,y}(w,x)}{\varphi_{C,1}(w,x)}\,dw$. This result is at the basis of the definition of the two following estimators that they proposed for $\Lambda(y)$:

$$\widehat{\Lambda}_{C,LS}(y) = \int_\chi v(x) \frac{\widehat{S}_{C,1}(y,x)}{\widehat{S}_{C,1}(1,x)}\,dx\ , \tag{1.2}$$

and

$$\widehat{\Lambda}_{C,LAD}(y) = \arg\min_{m\in\mathbb{R}} \int_\chi v(x) \left| \frac{\widehat{S}_{C,1}(y,x)}{\widehat{S}_{C,1}(1,x)} - m \right| dx\ , \tag{1.3}$$

where v is a weighting function, $\widehat{S}_{C,1}(y,x) = \int_0^y \frac{\widehat{\varphi}_{C,y}(w,x)}{\widehat{\varphi}_{C,1}(w,x)}\,dw$ and $\widehat{\varphi}_{C,y}(y,x)$ and $\widehat{\varphi}_{C,1}(y,x)$ are kernel estimators of $\varphi_{C,y}(y,x)$ and $\varphi_{C,1}(y,x)$.

Hence, to estimate the transformation in a nonparametric way, Chiappori et al. (2015) introduced kernel estimators and in particular smoothing on Y. However, the distribution of Y can be very asymmetric, depending on which transformation Λ is used. If ε is for instance normal and Λ is the logarithmic tranform, then Y is log-normal and hence smoothing for large values of Y will not work well at all. In that case, the kernel estimators based on Y can work very badly in practice. This naturally impacts the quality of the estimator of the conditional distribution function of Y given X and consequently also those of its derivatives and of the nonparametric estimators of the transformation as we can see in the simulations in Section 6.

In this paper, we will propose a way to overcome this problem. The main idea will be to rewrite the transformation $\Lambda(Y)$ as $\Gamma(T(Y))$ where Γ is an increasing function and T is a certain function such that we can control the distribution of $T(Y)$. This will be presented more precisely in Section 2 as well as the identification of model (1.1). In

Section 3, we explain how we can identify the function $\Gamma(\cdot)$. Sections 4 and 5 contain respectively the definitions of the new nonparametric estimators of the transformation Λ and the theorems that establish their asymptotic normality, as well as the technical assumptions we need to obtain these results. In Section 6, we realize a simulation study in order to compare the performance of our estimators to the ones developed in Chiappori et al. (2015). Section 7 contains the proofs of the main results. Finally, the proofs of the technical propositions, which are useful to establish the main results, and some additional simulation results, are given in the supplementary material.

2. Main idea of the estimation procedure

One way to identify model (1.1) is to impose one of the following normalization conditions:

(I1) either $\Lambda(\alpha_1) = a_1$ and $\Lambda(\alpha_2) = a_2$ for some $\alpha_1 < \alpha_2$ and $a_1 < a_2$,
(I2) or $\Lambda(\alpha_1) = a_1$ and $\Lambda'(\alpha_3) = a_3$ for some $a_1, a_3, \alpha_1, \alpha_3$.

The main idea of these two sets of conditions is to fix the location and the scale of the model. Condition $\Lambda(\alpha_1) = a_1$ fixes the location of model (1.1) in both cases and (I1) and (I2) differ in their second condition, i.e., in the way they fix the scale of model (1.1). We refer to Chiappori et al. (2015), Vanhems and Van Keilegom (2018) and Colling and Van Keilegom (2016) among others for more details about these sets of conditions. In this paper, we will work under (I1) for theoretical reasons and we will consider $\alpha_1 = 0$, $a_1 = 0$, $\alpha_2 = 1$ and $a_2 = 1$ without loss of generality. Note that the values 0 and 1 can be replaced by any other values as long as they are in the interior of the support $\mathcal{Y}$ of Y, where $\mathcal{Y}$ is a connected subset of $\mathbb{R}$.

As explained in the introduction, kernel smoothing based on Y can work very badly in practice. We will now explain how this can be avoided. The main idea is to rewrite the transformation $\Lambda(Y)$ as $\Gamma(T(Y))$, i.e., $\Lambda = \Gamma \circ T$ where Γ is an increasing function and

$$T(Y) = \frac{F_Y(Y) - F_Y(0)}{F_Y(1) - F_Y(0)} \ ,$$

where $F_Y(\cdot)$ is the distribution function of Y. Note that we can always write $\Lambda(Y)$ in this way, since $T(Y)$ is invertible. Moreover, we have to work with $T(Y)$ instead of the more simple transformation $F_Y(Y)$, since we want to guarantee that $T(0) = 0$ and $T(1) = 1$, and hence $\Gamma(0) = \Lambda(T^{-1}(0)) = 0$ and $\Gamma(1) = \Lambda(T^{-1}(1)) = 1$. Consequently, the set of normalization conditions (I1) does not change and we will now consider the following nonparametric transformation model:

$$\Gamma(U) = m(X) + \varepsilon \ , \tag{2.1}$$

where $U = T(Y)$ and $\Gamma(U) = \Lambda(Y)$. We will denote by $\mathcal{U} \subset \mathbb{R}$ the compact support of U. As $F_Y(Y)$ is uniformly distributed on the interval $[0, 1]$, it is clear that U is uniformly distributed on the interval $\left[\frac{-F_Y(0)}{F_Y(1)-F_Y(0)}, \frac{1-F_Y(0)}{F_Y(1)-F_Y(0)}\right]$. The advantage of working with U instead of Y is that U has a uniform distribution, and hence kernel smoothing based on U should work much better than when it is based on Y, which can have very long tails.

3. Identification of $\Gamma(\cdot)$

In this section, we will explain how we can identify the transformation $\Gamma(\cdot)$ defined in (2.1). The way we will do it is similar as in Chiappori et al. (2015) except that we will now work with $U = T(Y)$ instead of Y.

Let $x = (x_1, \ldots, x_d)^t$ and define $\varphi(u, x) = F_{U|X}(u, x)$ the conditional distribution function of U given X. Using the independence between X and ε, we have

$$\varphi(u, x) = P(U \leq u | X = x) = P(m(X) + \varepsilon \leq \Gamma(u) | X = x) = F_\varepsilon(\Gamma(u) - m(x)) \ .$$

Next, Conditions (A2) and (A3) in Section 5.1 imply that the derivatives of $\varphi(u, x)$ with respect to u and x_ρ for some $\rho \in \{1, \ldots, d\}$ exist. The derivative of $\varphi(u, x)$ with respect to u is denoted by $\varphi_u(u, x)$ and is given by

$$\varphi_u(u, x) = \frac{\partial}{\partial u}\varphi(u, x) = \Gamma'(u) f_\varepsilon(\Gamma(u) - m(x)) \ , \tag{3.1}$$

where $\Gamma'(u) = \frac{\partial}{\partial u}\Gamma(u)$ exists by Condition (A2). Similarly, the derivative of $\varphi(u, x)$ with respect to x_ρ is denoted by $\varphi_\rho(u, x)$ and is given by

$$\varphi_\rho(u, x) = \frac{\partial}{\partial x_\rho}\varphi(u, x) = -\left(\frac{\partial}{\partial x_\rho} m(x)\right) f_\varepsilon(\Gamma(u) - m(x)) \ . \tag{3.2}$$

We have used the fact that the function $m(x)$ is differentiable with respect to x_ρ by condition (A3), see Section 5.1. Moreover, we choose ρ such that the set $\mathcal{A}_\rho = \{x \in \chi : \varphi_\rho(u, x) \neq 0 \ \forall u \in \mathcal{U}_0\}$ is non empty, where $\mathcal{U}_0$ is a compact subset in the interior of $\mathcal{U}$. This last condition is imposed in (A4), see Section 5.1. Hence, dividing (3.1) by (3.2), we obtain

$$\Gamma'(u) = -\left(\frac{\partial}{\partial x_\rho} m(x)\right) \times \frac{\varphi_u(u, x)}{\varphi_\rho(u, x)} \ . \tag{3.3}$$

This last expression will be very important to prove the following Theorem which states how we can identify the transformation $\Gamma(u)$.

Theorem 3.1. *Assume (A1)-(A4). Then, for any $\rho \in \{1, \ldots, d\}$ such that $\mathcal{A}_\rho$ is non empty, Γ is identified under (I1) as*

$$\Gamma(u) = \lambda_\rho(u, x) = \frac{S_\rho(u, x)}{S_\rho(1, x)} \ ,$$

where $u \in \mathcal{U}_0$, $s_\rho(u, x) = \frac{\varphi_u(u,x)}{\varphi_\rho(u,x)}$ and $S_\rho(u, x) = \int_0^u s_\rho(w, x)dw$. Moreover, the expression $\lambda_\rho(u, x)$ does not depend on ρ nor x.

The proof of this theorem is the same as the proof of Theorem 1 in Chiappori et al. (2015) and is therefore omitted. This theorem implies that we can arbitrarily choose the component x_ρ of the vector $x = (x_1, \ldots, x_d)^t$ with respect to which we will derive to

construct the nonparametric estimators of the transformation in Section 4 as long as $\mathcal{A}_\rho$ is not empty. In practice, we recommend to choose ρ and x such that $\frac{\partial}{\partial x_\rho}\widehat{\varphi}(u,x)$ is as far as possible from 0, where $\widehat{\varphi}(u,x)$ is defined in the next section. In the following, to construct the estimators of $\Gamma(u)$, we will consider the derivative with respect to x_1 without loss of generality.

Note that we established Theorem 3.1 under the particular identification conditions $\Lambda(0)=0$ and $\Lambda(1)=1$, which requires that 0 and 1 are in the interior of the support $\mathcal{Y}$ of Y. However, if we consider for instance the square root transform $\Lambda(Y)=\sqrt{Y}$ with $\mathcal{Y}=[0,+\infty[$, it is clearly not the case since the point 0 is at the boundary of $\mathcal{Y}$. To overcome this problem, as we mentioned at the beginning of Section 2, the points 0 and 1 can be replaced by any other values as long as they are in the interior of the support $\mathcal{Y}$ of Y. In general, if $\Lambda(\alpha_1)=a_1$ and $\Lambda(\alpha_2)=a_2$ for some $\alpha_1<\alpha_2$ and $a_1<a_2$ such that α_1 and α_2 are in the interior of $\mathcal{Y}$, it suffices to define Γ such that $\Lambda=\Gamma\circ\widetilde{T}$ where

$$\widetilde{U}=\widetilde{T}(Y)=(\alpha_2-\alpha_1)\frac{F_Y(Y)-F_Y(\alpha_1)}{F_Y(\alpha_2)-F_Y(\alpha_1)}+\alpha_1\ .$$

This ensures that $T(\alpha_1)=\alpha_1$ and $T(\alpha_2)=\alpha_2$ and Γ satisfies consequently exactly the same identification conditions as Λ. Indeed, $\Gamma(\alpha_1)=\Lambda(T^{-1}(\alpha_1))=a_1$ and $\Gamma(\alpha_2)=\Lambda(T^{-1}(\alpha_2))=a_2$. In this situation, following exactly the same reasoning as in the proof of Theorem 3.1, Γ is identified under (I1) as

$$\Gamma(u)=a_1+(a_2-a_1)\frac{\widetilde{S}_1(u,x)}{\widetilde{S}_1(\alpha_2,x)}\ ,$$

where $\widetilde{S}_1(u,x)=\int_{\alpha_1}^{u}\frac{\widetilde{\varphi}_{\widetilde{u}(w,x)}}{\widetilde{\varphi}_1(w,x)}dw$ and $\widetilde{\varphi}_{\widetilde{u}}$ and $\widetilde{\varphi}_1$ are the derivatives of the conditional distribution function of $\widetilde{U}$ given X with respect to $\widetilde{u}$ and x_1 respectively.

4. Definitions of the new estimators

In this section, we will construct our new nonparametric estimators of the transformation $\Gamma(u)$ that will directly lead to estimators of $\Lambda(y)$ by replacing F_Y by its empirical distribution function $\widehat{F}_Y$. The idea is to construct an estimator of the function $\lambda_1(u,x)$ introduced in Theorem 3.1.

First, for $i=1,\dots,n$, let $X_i=(X_{i1},\dots,X_{id})$ and assume that we have randomly drawn an *iid* sample $(X_1,Y_1),\dots,(X_n,Y_n)$ from the nonparametric transformation model (1.1). Moreover, we define $U_i=T(Y_i)=\frac{F_Y(Y_i)-F_Y(0)}{F_Y(1)-F_Y(0)}$.

We can rewrite the conditional distribution function of U given X defined in Section 3 in the following way:

$$\varphi(u,x)=\frac{p(u,x)}{f_X(x)}\ , \tag{4.1}$$

where

$$p(u,x)=\int_{-\infty}^{u}f_{U,X}(w,x)\,dw\qquad,\qquad f_X(x)=\int_{\mathcal{U}}f_{U,X}(w,x)\,dw\ , \tag{4.2}$$

and $f_{U,X}(u,x)$ is the joint density function of U and X. Let K be a univariate kernel in $\mathbb{R}$ that satisfies Condition (A5) in Section 5.1, $\mathcal{K}(u) = \int_{-\infty}^{u} K(w)\,dw$ and $\mathbf{K}(x)$ be a multivariate product kernel of the form $\mathbf{K}(x) = \prod_{i=1}^{d} K(x_i)$ with $x = (x_1, \ldots, x_d)^t$. We also introduce $h_u > 0$ and $h_x > 0$ two univariate bandwidths that satisfy Condition (A6) in Section 5.1. Consequently, a natural kernel estimator of $\varphi(u,x)$ is given by

$$\widehat{\varphi}(u,x) = \frac{\widehat{p}(u,x)}{\widehat{f}_X(x)} \ ,$$

where

$$\widehat{p}(u,x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_{h_u}(u - \widehat{U}_i)\mathbf{K}_{h_x}(X_i - x) \quad \text{and} \quad \widehat{f}_X(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{K}_{h_x}(X_i - x) \ , \tag{4.3}$$

with $\mathcal{K}_{h_u} = \mathcal{K}(u/h_u)$ and $\mathbf{K}_{h_x}(x) = \mathbf{K}(x/h_x)/h_x^d$. Moreover, $\widehat{U}_i$ is defined by

$$\widehat{U}_i = \widehat{T}(Y_i) = \frac{\widehat{F}_Y(Y_i) - \widehat{F}_Y(0)}{\widehat{F}_Y(1) - \widehat{F}_Y(0)} \ ,$$

where $\widehat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Y_i \le y\}}$ is the empirical distribution of $Y_1, \ldots, Y_n$. It is important to notice that we have to work with $\widehat{U}_1, \ldots, \widehat{U}_n$ instead of $U_1, \ldots, U_n$ since the U_i's depend on the distribution function of Y which is an unknown function. This will make an important difference when we will develop the asymptotic properties of our estimators in comparison with Chiappori et al. (2015) who constructed their estimators on the basis of $Y_1, \ldots, Y_n$.

As mentioned in Chiappori et al. (2015), we have to use $\mathcal{K}_{h_u}(u - \widehat{U}_i)$ instead of $1_{\{\widehat{U}_i \le u\}}$ because we need $\widehat{\varphi}(u,x)$ to be differentiable with respect to u. Also, note that we could work with a separate bandwidth for each component X_i of the vector X but wanted to keep the presentation simple.

Next, we consider

$$\varphi_u(u,x) = \frac{\partial}{\partial u}\varphi(u,x) = \frac{p_u(u,x)}{f_X(x)} \ , \tag{4.4}$$

and

$$\varphi_1(u,x) = \frac{\partial}{\partial x_1}\varphi(u,x) = \frac{p_1(u,x)}{f_X(x)} - \frac{p(u,x)f_{X,1}(x)}{f_X^2(x)} \ , \tag{4.5}$$

where $p_u(u,x) = \frac{\partial}{\partial u}p(u,x)$, $p_1(u,x) = \frac{\partial}{\partial x_1}p(u,x)$ and $f_{X,1}(x) = \frac{\partial}{\partial x_1}f_X(x)$. We estimate these quantities respectively by

$$\widehat{\varphi}_u(u,x) = \frac{\widehat{p}_u(u,x)}{\widehat{f}_X(x)} \quad \text{and} \quad \widehat{\varphi}_1(u,x) = \frac{\widehat{p}_1(u,x)}{\widehat{f}_X(x)} - \frac{\widehat{p}(u,x)\widehat{f}_{X,1}(x)}{\widehat{f}_X^2(x)} \ ,$$

where $\widehat{p}_u(u,x) = \frac{\partial}{\partial u}\widehat{p}(u,x)$, $\widehat{p}_1(u,x) = \frac{\partial}{\partial x_1}\widehat{p}(u,x)$ and $\widehat{f}_{X,1}(x) = \frac{\partial}{\partial x_1}\widehat{f}_X(x)$ and $\widehat{p}(u,x)$ and $\widehat{f}_X(x)$ are defined in (4.3). Finally, we define the following estimator of $\lambda_1(u,x)$:

$$\widehat{\lambda}_1(u,x) = \frac{\widehat{S}_1(u,x)}{\widehat{S}_1(1,x)} \quad \text{where} \quad \widehat{S}_1(u,x) = \int_0^u \frac{\widehat{\varphi}_u(w,x)}{\widehat{\varphi}_1(w,x)}\,dw \ . \tag{4.6}$$

It is important to remark that the estimator $\widehat{\lambda}_1(u,x)$ depends on x despite the fact that it is not the case for the true function $\lambda_1(u,x)$ by Theorem 3.1. Consequently, we will have to integrate the function $\widehat{\lambda}_1(u,x)$ over χ. We also consider a weighting function $v(x)$ that satisfies Condition (A9) in Section 5.1. Moreover, Theorem 3.1 implies that $\Gamma(u) = \lambda_1(u,x)$ and then

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

where $\ell(\cdot)$ is a given loss function. In this paper, we will consider $\ell(z) = z^2$ and $\ell(z) = |z|$ which gives more precisely the two following nonparametric estimators for the transformation $\Gamma(u)$:

$$\widehat{\Gamma}_{LS}(u) = \int_\chi v(x)\widehat{\lambda}_1(u,x)\, dx \ , \tag{4.7}$$

and

$$\widehat{\Gamma}_{LAD}(u) = \arg\min_{q_m \in \mathbb{R}} \int_\chi v(x) \left|\widehat{\lambda}_1(u,x) - q_m\right| dx. \tag{4.8}$$

Next, if we want to estimate e.g. $\Lambda(2)$, we just have to calculate $\widehat{\Gamma}_{LS}(\widehat{T}(2))$ or $\widehat{\Gamma}_{LAD}(\widehat{T}(2))$ where $\widehat{T}(y) = \frac{\widehat{F}_Y(y) - \widehat{F}_Y(0)}{\widehat{F}_Y(1) - \widehat{F}_Y(0)}$.

To simplify the theoretical analysis, we follow Chiappori et al. (2015) and introduce a smoothed version of the above LAD estimator :

$$\widehat{\Gamma}_{LAD,b}(u) = \arg\min_{q_m \in \mathbb{R}} Q_b(q_m, \widehat{\lambda}_1(u,\cdot)) \ , \tag{4.9}$$

where

$$Q_b(q_m, \lambda_1(u,\cdot)) = \int_\chi v(x)\left(\lambda_1(u,x) - q_m\right)\left[2L_b(\lambda_1(u,x) - q_m) - 1\right] dx,$$

$L_b(\cdot) = L(\cdot/b)$, L is a given distribution function and $b > 0$ is a bandwidth sequence that satisfy Condition (A10) in Section 5.1. The main reason for working with $\widehat{\Gamma}_{LAD,b}(u)$ instead of working with $\widehat{\Gamma}_{LAD}(u)$ is that the function Q_b is twice differentiable with respect to q_m, which will be needed in the proof. Note that the two estimators can be made arbitrarily close to each other by letting b tend to zero.

In summary, the construction of our estimators is similar as in Chiappori et al. (2015) except that we proposed a way to overcome the facts that the distribution of Y can be very asymmetric and can have very long tails, which can lead to bad estimations of $\Lambda(y)$ due to the poor quality of the kernel smoothing. In Section 6, we will compare the performance of our new estimators to those of Chiappori et al. (2015) and we will see that the transformation of $U = T(Y)$ is recommended, especially for large values of Y. However, the disadvantage of transforming Y and considering kernel smoothing on $\widehat{U}$ is the asymptotic theory that will be considerably more difficult in comparison with Chiappori et al. (2015). Indeed, we have a lot of technical results to prove before establishing the limiting distributions of the estimators $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD,b}(u)$, see Section A in the supplementary material.

5. Asymptotic results

5.1. Notations and assumptions

First, we need to introduce the following notations:

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

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

and

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

For any functions $\widetilde{p}(u,x)$ and $\widetilde{f}(x)$, let $\widetilde{p}_u(u,x)$ and $\widetilde{p}_1(u,x)$ be respectively the derivatives of $\widetilde{p}(u,x)$ with respect to u and x_1, where x_1 is the first component of the vector $x \in \mathbb{R}^d$, and let $\widetilde{f}_1(x)$ be the derivative of the function $\widetilde{f}(x)$ with respect to x_1. We also define the following functionals

$$\nabla_p S_1(u,x)[\widetilde{p}] = \int_0^u \left\{ D_{p,0}(w,x)\widetilde{p}(w,x) + D_{p,u}(w,x)\widetilde{p}_u(w,x) + D_{p,1}(w,x)\widetilde{p}_1(w,x) \right\} dw \ , \tag{5.1}$$

and

$$\nabla_f S_1(u,x)[\widetilde{f}] = \int_0^u \left\{ D_{f,0}(w,x)\widetilde{f}(x) + D_{f,1}(w,x)\widetilde{f}_1(x) \right\} dw \ . \tag{5.2}$$

The main results of the asymptotic theory require the following regularity conditions on the distributions of ε given X, the transformation $\Gamma(u)$, the regression function $m(x)$, the kernels K and L, the bandwidths b, h_x and h_u, the joint density function of U and X and the weighting function v:

(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 $\mathcal{A}_\rho = \{x \in \chi : \varphi_\rho(u,x) \neq 0 \ \forall u \in \mathcal{U}_0\}$ is nonempty for some $\rho \in \{1,\dots,d\}$. In the following, we will take $\rho = 1$ without loss of generality.

(A5) The kernel K is symmetric, has support [-1,1], $K(-1) = K(1) = 0$, $\int K(z)\,dz = 1$, $\int z^\ell K(z)\,dz = 0$ for $\ell = 1,\dots,m-1$ and $\int z^m K(z)\,dz < \infty$. Moreover, K is m-times continuously differentiable and K and K' are of bounded variation.

(A6) The bandwidths h_x and h_u satisfy $\sqrt{n}h_x^m \to 0$, $\sqrt{n}h_u^m \to 0$, $\frac{\sqrt{n}h_x^{d+2}}{\log n} \to \infty$ and $\frac{\sqrt{n}h_u^2 h_x^d}{\log n} \to \infty$.

(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) $\inf_{x\in\chi_0} f_X(x) > 0$, $\inf_{(u,x)\in\mathcal{U}_0\times\chi_0} |\frac{\partial}{\partial x_1}\varphi(u,x)| > 0$ 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 m-times continuously differentiable on χ_0.

(A10) L is a twice continuously differentiable distribution function with uniformly bounded derivatives and with median at 0 and $b > 0$ is a bandwidth sequence that satisfies $nb^4 \to \infty$ and $\frac{b\sqrt{n}h_x^d(\min(h_x,h_u))^2}{\log n} \to \infty$.

Note that several of these assumptions are the same as in Chiappori et al. (2015). It is important to remark that $m > d+2$ is a consequence of Assumption (A6), which means that a higher-order kernel is required to ensure that the four conditions of Assumption (A6) are satisfied when the dimension d of X becomes large.

Next, Condition (A7) implies that the joint density $f_{U,X}(u,x)$ is uniformly bounded, $m+2$-times continuously differentiable and all these derivatives are uniformly bounded, i.e.,

$$\sup_{(u,x)\in\mathcal{U}_0\times\chi_0} \left| \frac{\partial^{|\alpha|}}{\partial u^{\alpha_0}\partial x_1^{\alpha_1}\dots\partial x_d^{\alpha_d}} f_{U,X}(u,x) \right| < \infty \ ,$$

since $f_{U,X}(u,x)$ can be rewritten as

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

Moreover, Condition (A7) implies that the functions $p(u,x)$, $p_u(u,x)$, $p_1(u,x)$, $f_X(x)$ and $f_{X,1}(x)$ are uniformly bounded, at least m-times continuously differentiable and all these derivatives are uniformly bounded. Indeed, these five functions can be rewritten as functions depending only of $f_{U,X}(u,x)$, see (4.2). As a consequence, using (4.1), (4.4), (4.5) and the fact that $\inf_{x\in\chi_0} f_X(x) > 0$ by condition (A8), the functions $\varphi(u,x)$, $\varphi_u(u,x)$ and $\varphi_1(u,x)$ are also uniformly bounded, at least m-times continuously differentiable and all these derivatives are uniformly bounded. Finally, these conclusions are again the same for the functions $D_{p,0}(u,x)$, $D_{p,u}(u,x)$, $D_{p,1}(u,x)$, $D_{f,0}(u,x)$, $D_{f,1}(u,x)$ and $S_1(u,x)$ using the definitions at the beginning of Section 5, the definition of $S_1(u,x)$ in Theorem 3.1 and the facts that $\inf_{x\in\chi_0} f_X(x) > 0$ and $\inf_{(u,x)\in\mathcal{U}_0\times\chi_0} |\varphi_1(u,x)| > 0$ by condition (A8). All these consequences of Condition (A7) will be constantly used in the different proofs in the supplementary material and in Section 7.

Finally, since $v(x) = 0$ for $x \notin \chi_0$, Condition (A9) implies that $\int_\chi v(x)\,dx = \int_{\chi_0} v(x)\,dx = 1$ and also $v(x) = 0$ for values of x at the boundary of χ_0 since v is assumed to be continuous on χ. The last two properties of v will be necessary to prove Theorem 5.1. Indeed, to prove Theorem 5.1, we will need to apply Propositions 6 and 7 given in the supplementary material with $v_p(u_0,x) = v_f(u_0,x) = v(x)/S_1(u_0,x)$ and with $v_p(u_0,x) = v_f(u_0,x) = v(x)S_1(u_0,x)/S_1^2(1,x)$ and these two propositions will require

that the functions v_p and v_f are continuous on χ and equal 0 for values of x on the boundary of χ_0.

5.2. Main theorems

In this section, we introduce two theorems and one corollary. The two first theorems establish respectively the limiting distributions of $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD,b}(u)$. However, recall that $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD,b}(u)$ are nonparametric estimators of $\Gamma(u)$ and that our original objective was to construct nonparametric estimators of $\Lambda(y)$. Hence, in addition to these two theorems, we introduce a corollary that establishes the limiting distributions of $\widehat{\Gamma}_{LS}(\widehat{T}(y))$ and $\widehat{\Gamma}_{LAD,b}(\widehat{T}(y))$, that are nonparametric estimators of $\Lambda(y)$.

Before stating these results, note that the derivative of $\varphi(u,x)$ with respect to a component of x is zero for all $x \in \chi$ if u is at the lower boundary of the support $\mathcal{U}$ of U. Hence, Assumptions (A4) and (A8), given in Section 5.1, cannot be fulfilled for all $u \in \mathcal{U}$. Consequently, we have to restrict Assumptions (A4) and (A8) and also the following two theorems to any $u \in \mathcal{U}_0$, where $\mathcal{U}_0$ is a compact subset in the interior of $\mathcal{U}$.

Similarly, the convergence established in the next corollary is restricted to any $y \in \mathcal{Y}_0$, where $\mathcal{Y}_0$ is some compact subset strictly included in the support $\mathcal{Y}$ of Y.

In particular, if 0 is the lower boundary of $\mathcal{Y}$, it can be replaced by any other value in the interior of $\mathcal{Y}$, see the remark at the end of Section 3. Note that this will also avoid boundary problems when we will do kernel smoothing on U.

Theorem 5.1. *Assume (A1)-(A9). Then, under (I1), the process* $\sqrt{n}(\widehat{\Gamma}_{LS}(u) - \Gamma(u))$ *converges weakly to* $\mathcal{N}(u)$*, where* $u \in \mathcal{U}_0$ *and* $\mathcal{N}$ *is a centered Gaussian process with covariance function given by* $\Delta(u_1,u_2) = E[\delta^v_{X_i,Y_i}(u_1)\delta^v_{X_i,Y_i}(u_2)]$*, where* $\delta^v_{X_i,Y_i}(u) = \delta^{v_1}_{X_i,Y_i}(1,u) - \delta^{v_2}_{X_i,Y_i}(u,1)$*,* $v_1(u_0,x) = \frac{v(x)}{S_1(u_0,x)}$*,* $v_2(u_0,x) = \frac{v(x)S_1(u_0,x)}{S_1^2(1,x)}$*,*

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

$$-\left(\frac{1_{\{Y_i\leq 1\}}-1_{\{Y_i\leq 0\}}}{F_Y(1)-F_Y(0)}-1\right)u\int_{\mathcal{X}}\widetilde{v}(u_0,x)D_{p,u}(u,x)f_{U,X}(u,x)\,dx, \tag{5.3}$$

for $u_0\in\mathcal{U}$ and $\widetilde{v}$ equals either v_1 or v_2.

***Remark* 5.1.** Note that the estimator $\widehat{\Gamma}_{LS}(u)$ is not necessarily monotone is u. We can make the estimator monotone in a second step by following e.g. the rearrangement principle introduced by Dette et al. (2006). In Chernozhukov et al. (2010) it was shown that the rearrangement operator is Hadamard diffferentiable and hence the weak convergence of the estimator obtained by rearranging the estimator $\widehat{\Gamma}_{LS}(u)$ follows immediately.

Theorem 5.2. *Assume (A1)-(A10). Then, under (I1), the process $\sqrt{n}(\widehat{\Gamma}_{LAD,b}(u)-\Gamma(u))$ converges weakly to $\mathcal{N}(u)$, where $u\in\mathcal{U}_0$ and $\mathcal{N}(u)$ is the same centered Gaussian process as in Theorem 5.1.*

Corollary 5.1. *Assume (A1)-(A10). Then, the processes $\sqrt{n}(\widehat{\Gamma}_{LS}(\widehat{T}(y))-\Lambda(y))$ and $\sqrt{n}(\widehat{\Gamma}_{LAD,b}(\widehat{T}(y))-\Lambda(y))$ converge weakly to $\widetilde{\mathcal{N}}(y)$ for $y\in\mathcal{Y}_0$, where $\mathcal{Y}_0$ is a compact subset strictly included in $T^{-1}(\mathcal{U}_0)$ such that $\sup_{y\in\mathcal{Y}_0}|\Gamma^{(r)}(T(y))|<\infty$ for $r=1,2$ and $\widetilde{\mathcal{N}}$ is a centered Gaussian process with covariance function given by $Cov(\widetilde{N}(y_1),\widetilde{N}(y_2))=E[\varphi^v_{X_i,Y_i}(y_1)\varphi^v_{X_i,Y_i}(y_2)]$, where*

$$\begin{aligned}\varphi^v_{X_i,Y_i}(y) &= \delta^v_{X_i,Y_i}(T(y))+\frac{\Gamma'(T(y))}{F_Y(1)-F_Y(0)}\Big(1_{\{Y\leq y\}}-1_{\{Y\leq 0\}}-F_Y(y)+F_Y(0)\\ &\quad -T(y)\Big[1_{\{0\leq Y\leq 1\}}-F_Y(1)+F_Y(0)\Big]\Big)\,,\end{aligned} \tag{5.4}$$

and $\delta^v_{X_i,Y_i}(\cdot)$ is defined in Theorem 5.1.

The proofs of the two theorems and the corollary are given in Section 7. It is important to notice that the function $\delta^v_{X_i,Y_i}(u)$ is not the same as the one obtained in Chiappori et al. (2015). The first three terms of the expression $\delta^v_{X_i,Y_i}(u)$ are the same as in Chiappori et al. (2015) and the last three terms in $\delta^v_{X_i,Y_i}(u)$ are new contributions to the final asymptotic representation that are due to the estimation of U by $\widehat{U}$. We refer to Proposition 3 in the supplementary material that shows in detail where these last three terms come from. Chiappori et al. (2015) did not need this proposition since they worked directly with Y.

***Remark* 5.2.** Note that the estimators $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD,b}(u)$ have exactly the same asymptotic distribution, which might seem surprising at first sight. As noted by a referee, this is caused by the fact that the bandwidth b used to smooth the LAD estimator is sufficiently large, so that the convergence rate of $\widehat{\lambda}_1(u,x)-\Gamma(u)$ is smaller than b. The current proof will not work if the convergence speeds are reversed (i.e. if b is smaller than the convergence rate of $\widehat{\lambda}_1(u,x)-\Gamma(u)$), and in that case it is expected that the

asymptotic distributions differ. To illustrate this point we consider a simplified version of the aggregated estimators where the mean/median are only taken over a fixed and finite number of points $x_1, \ldots, x_\ell$:

$$\begin{aligned}\widehat{\Gamma}^{alt}_{LS}(u) &= \frac{1}{\ell}\sum_{j=1}^{\ell}\widehat{\lambda}_1(u,x_j)\\ \widehat{\Gamma}^{alt}_{LAD}(u) &= \operatorname{argmin}_{q_m\in\mathbb{R}}\frac{1}{\ell}\sum_{j=1}^{\ell}|\widehat{\lambda}_1(u,x_j)-q_m|\\ &= \operatorname{median}(\widehat{\lambda}_1(u,x_1),\ldots,\widehat{\lambda}_1(u,x_\ell)).\end{aligned}$$

From equation (7.1) and Proposition 5 in the supplementary material it follows that $(nh_x^d)^{1/2}(\widehat{\lambda}_1(u,x_1)-\Gamma(u),\ldots,\widehat{\lambda}_1(u,x_\ell)-\Gamma(u))$ converges in distribution to a ℓ-variate normal random vector $W_u(x_1,\ldots,x_\ell)=(W_u(x_1),\ldots,W_u(x_\ell))$ of mean zero and diagonal variance-covariance matrix $\operatorname{diag}(\sigma^2_{1u},\ldots,\sigma^2_{\ell u})$, since the components of this vector are independent for n large. Hence,

$$(nh_x^d)^{1/2}(\widehat{\Gamma}^{alt}_{LS}(u)-\Gamma(u))=\frac{1}{\ell}\sum_{j=1}^{\ell}(nh_x^d)^{1/2}(\widehat{\lambda}_1(u,x_j)-\Gamma(u))\xrightarrow{d}\frac{1}{\ell}\sum_{j=1}^{\ell}W_u(x_j),$$

and the latter variable has a normal distribution with mean zero and variance given by $\ell^{-2}\sum_{j=1}^{\ell}\sigma^2_{ju}$. On the other hand,

$$\begin{aligned}&(nh_x^d)^{1/2}(\widehat{\Gamma}^{alt}_{LAD}(u)-\Gamma(u))\\ &=\operatorname{median}\big((nh_x^d)^{1/2}(\widehat{\lambda}_1(u,x_1)-\Gamma(u)),\ldots,(nh_x^d)^{1/2}(\widehat{\lambda}_1(u,x_\ell)-\Gamma(u))\big),\end{aligned}$$

and this converges to the median of $W_u(x_1),\ldots,W_u(x_\ell)$, which will be different from the limit of the LS estimator.

6. Simulations

In this section, we perform simulations in order to compare the performance of the new nonparametric estimators $\widehat{\Gamma}_{LS}(\widehat{T}(y))$ and $\widehat{\Gamma}_{LAD}(\widehat{T}(y))$ of the transformation with the estimators $\widehat{\Lambda}_{C,LS}(y)$ and $\widehat{\Lambda}_{C,LAD}(y)$ defined in (1.2) and (1.3) and given in Chiappori et al. (2015).

We consider $d=1$ and the simulated model is $\Lambda(Y_i)=6X_i-3+\varepsilon_i$, where $X_1,\ldots,X_n$ are independent uniform random variables on $[0,1]$ and $\varepsilon_1,\ldots,\varepsilon_n$ are independent standard normal random variables truncated on $[-3,3]$. Here, we will consider the four following different transformations:

$$\Lambda_1(Y)=\begin{cases}\frac{\log(Y+1)}{\log 2} & \text{if } Y\geq 0\\ \frac{1-(1-Y)^2}{2\log 2} & \text{if } Y<0,\end{cases},\quad \Lambda_2(Y)=\begin{cases}\frac{\sqrt{Y+1}-1}{\sqrt{2}-1} & \text{if } Y\geq 0\\ \frac{1-(1-Y)^{3/2}}{3(\sqrt{2}-1)} & \text{if } Y<0,\end{cases},$$

$$\Lambda_3(Y) = Y \quad \text{and} \quad \Lambda_4(Y) = \begin{cases} \frac{(Y+1)^{3/2}-1}{2\sqrt{2}-1} & \text{if } Y \geq 0 \\ \frac{3(1-\sqrt{1-Y})}{2\sqrt{2}-1} & \text{if } Y < 0, \end{cases} .$$

Hence, we can easily verify that Y takes values on the intervals $[-2.05, 63]$, $[-3.15, 11.15]$, $[-6, 6]$ and $[-20.69, 4.23]$ when we use the above simulated model with the transformations $\Lambda_j(Y)$, for $j = 1, \ldots, 4$, respectively.

Note that these four transformations are particular cases of the Yeo and Johnson (2000) transformation up to a multiplicative constant. Indeed, the Yeo and Johnson (2000) is given by

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

and $\Lambda_1(Y) = \frac{\Lambda_{\theta=0}(Y)}{\log 2}$, $\Lambda_2(Y) = \frac{\Lambda_{\theta=0.5}(Y)}{2(\sqrt{2}-1)}$, $\Lambda_3(Y) = \Lambda_{\theta=1}(Y)$ and $\Lambda_4(Y) = \frac{3\Lambda_{\theta=1.5}(Y)}{2(2\sqrt{2}-1)}$. These different multiplicative constants were introduced in order to ensure that $\Lambda_j(1) = 1$ for $j = 1, \ldots, 4$. Moreover, the Yeo and Johnson transformation is such that $\Lambda_\theta(0) = 0$ for all θ. Hence, $\Lambda_j(0) = 0$ for $j = 1, \ldots, 4$ and the normalization conditions (I1) are satisfied.

We will now explain how we approximate the expressions $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD}(u)$ in practice. The reasoning will be the same for $\widehat{\Lambda}_{C,LS}(y)$ and $\widehat{\Lambda}_{C,LAD}(y)$. Note that $K(x) = \frac{3}{4}(1-x^2)1_{\{|x|\leq 1\}}$ is the Epanechnikov kernel and we select the bandwidths h_x and h_u by a classical bandwidth selection rule for kernel density estimation. For simplicity, we choose here the normal reference rule, i.e., $\widehat{h}_x = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_x$ and $\widehat{h}_u = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_u$ where $\widehat{\sigma}_x$ and $\widehat{\sigma}_u$ are the classical estimators of the standard deviations of X and U respectively. First, we generate a grid $x_1^*, \ldots, x_{N_x}^*$ of N_x equidistant points between $\min_{1\leq j\leq n} X_j$ and $\max_{1\leq j\leq n} X_j$. Hence, we can compute

$$\widehat{S}_1(u, x_\ell^*) = \int_0^u \frac{\widehat{\varphi}_u(w, x_\ell^*)}{\widehat{\varphi}_1(w, x_\ell^*)}\, dw \quad \text{and} \quad \widehat{\lambda}_1(u, x_\ell^*) = \frac{\widehat{S}_1(u, x_\ell^*)}{\widehat{S}_1(1, x_\ell^*)} ,$$

for each $\ell = 1, \ldots, N_x$. However, in practice, the expression $\widehat{\varphi}_1(w, x_\ell^*)$ could be very close to 0 for some $\ell = 1, \ldots, N_x$, and consequently, the integral $\int_0^u \frac{\widehat{\varphi}_u(w,x_\ell^*)}{\widehat{\varphi}_1(w,x_\ell^*)}\, dw$ could diverge. To overcome this problem, we will remove from the original sample $x_1^*, \ldots, x_{N_x}^*$ the values x_ℓ^* such that $\widehat{S}_1(u, x_\ell^*)$ diverges. We will also remove the x^*-values that are within 0.01 of the values x_ℓ^* such that $\widehat{S}_1(u, x_\ell^*)$ diverges, even if the corresponding integrals do not diverge. The purpose is to avoid too small values of $\widehat{\varphi}_1(w, x_\ell^*)$ and thus too large values of $\widehat{\lambda}_1(\cdot)$ that would deteriorate the final estimation of the transformation. This is allowed since $\widehat{\lambda}_1(u, x)$ estimates $\Gamma(u)$ for all $x \in \chi$. We assume finally that n_x values remain, denoted by $\widetilde{x}_1, \ldots, \widetilde{x}_{n_x}$, from our original sample $x_1^*, \ldots, x_{N_x}^*$. In the following, we will use $N_x = 100$. Finally, we consider the weighting function $v(x) = 1$, for all x such

that the above integral does not diverge, and $\widehat{\Gamma}_{LS}(u)$ and $\widehat{\Gamma}_{LAD}(u)$ are consequently approximated by

$$\widetilde{\Gamma}_{LS}(u) = \frac{1}{n_x}\sum_{i=1}^{n_x}\widehat{\lambda}_1(u,\widetilde{x}_i) \text{ and } \widetilde{\Gamma}_{LAD}(u) = \text{median}(\widehat{\lambda}_1(u,\widetilde{x}_1),\ldots,\widehat{\lambda}_1(u,\widetilde{x}_{n_x})) .$$

Then, the transformations $\Lambda_1(y)$, $\Lambda_2(y)$, $\Lambda_3(y)$ and $\Lambda_4(y)$ will be estimated by $\widetilde{\Gamma}_{LS}(\widehat{T}(y))$ and $\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$. Using a similar procedure, we can construct $\widetilde{\Lambda}_{C,LS}(y)$ and $\widetilde{\Lambda}_{C,LAD}(y)$ the approximations of $\widehat{\Lambda}_{C,LS}(y)$ and $\widehat{\Lambda}_{C,LAD}(y)$.

Each graph of Figure 1 shows the true transformation $\Lambda_1(y)$ in black and each line in grey represents one estimation of $\Lambda_1(y)$ obtained with one sample of size $n = 200$. The top graphs on the left and on the right of Figure 1 were respectively obtained with the estimators $\widetilde{\Gamma}_{LS}(\widehat{T}(y))$ and $\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$ developed in this paper and the bottom graphs on the left and on the right of Figure 1 were respectively obtained with the estimators $\widetilde{\Lambda}_{C,LS}(y)$ and $\widetilde{\Lambda}_{C,LAD}(y)$, each of the four cases on the basis of 200 samples of size $n = 200$. Figures 2 to 4 are constructed exactly in the same way but for $\Lambda_2(y)$, $\Lambda_3(y)$ and $\Lambda_4(y)$ respectively.

First, we observe on these four figures that the estimators constructed with the median, i.e., the LAD-estimators, outperform in all tried scenarios the corresponding estimators constructed with the mean, i.e., the LS-estimators. This can be explained by the fact that the mean version of the estimator is more sensitive to outliers in $\widehat{\lambda}_1(u,\widetilde{x})$ as we vary $\widetilde{x}$. Hence, it is not surprising if the LAD-estimators behave better than the corresponding LS-estimators from the biases and the variances points of view. In the following, we will thus concentrate our conclusions on the LAD versions of the estimators.

We also observed in Figures 1 to 4 that the estimations of the transformation Λ are not necessarily monotone and especially the ones obtained with the mean versions of the estimator. As already mentioned in Remark 5.1, the estimators could be monotonized by applying e.g. the rearrangement principle proposed by Dette et al. (2006).

Next, we also introduce Tables 1 to 4 that show the estimations of the biases, variances and mean squared errors of the LAD-estimators of the transformations $\Lambda_j(y)$, $j = 1,\ldots,4$ evaluated at some specific values of y.

By construction, $\widetilde{\Gamma}_{LAD}(\widehat{T}(0)) = \widetilde{\Gamma}_{LAD}(0) = 0$, $\widetilde{\Lambda}_{C,LAD}(0) = 0$, $\widetilde{\Gamma}_{LAD}(\widehat{T}(1)) = \widetilde{\Gamma}_{LAD}(1) = 1$ and $\widetilde{\Lambda}_{C,LAD}(1) = 1$, which justifies the fact that all the grey lines from Figures 1 to 4 cross the points $(0,0)$ and $(1,1)$. At these points, the variances of all estimators are equal to 0 and we can see in Tables 1 to 4 that these variances increase for both LAD-estimators when y gets away from 0 in negative values and from 1 in positive values.

Moreover, when y is close to 0 and 1, the estimations of $\Lambda_j(y)$, for $j = 1,\ldots,4$, are slightly better in terms of variances and biases when we use $\widetilde{\Lambda}_{C,LAD}(y)$, see for example $\Lambda_j(0.5)$. However, when we estimate $\Lambda_1(y)$, $\Lambda_2(y)$ and $\Lambda_4(y)$ and when y gets away from 0 in negative values and from 1 in positive values, the absolute value of the biases of the $\widetilde{\Lambda}_{C,LAD}(y)$ estimator become significantly bigger and bigger, in comparison with the biases of $\widetilde{\Gamma}_{LAD}(y)$ that remain globally constant. This gives very poor estimations of the transformations $\Lambda_j(y)$ for large values of $|y|$ when we use $\widetilde{\Lambda}_{C,LAD}(y)$, see for

examples the estimations of $\Lambda_1(10)$, $\Lambda_2(5)$ and $\Lambda_4(-4)$, even if the estimations of $\Lambda_j(y)$, for $j = 1, \ldots, 4$, are globally better in terms of variance when we use $\widetilde{\Lambda}_{C,LAD}(y)$. As explained in the introduction, it is due to the fact that $\widetilde{\Lambda}_{C,LAD}(y)$ is constructed by doing smoothing on $Y_1, \ldots, Y_n$ which are drawn from a distribution with possibly a long tail. As a consequence, the kernel estimators behaves not always well in that cases and especially for large values of y.

In conclusion, as we can see in Figures 1 to 4 and in Tables 1 to 4, $\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$ clearly outperforms $\widetilde{\Lambda}_{C,LAD}(y)$ when we consider $\Lambda_1(y)$, $\Lambda_2(y)$ and $\Lambda_4(y)$ and large values of $|y|$ and $\widetilde{\Lambda}_{C,LAD}(y)$ only outperforms $\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$ when we consider the identity transformation $\Lambda_3(y)$. Moreover, we observe that $\widetilde{\Lambda}_{C,LAD}(y)$ gives very poor estimations of $\Lambda_1(y)$, $\Lambda_2(y)$ and $\Lambda_4(y)$ when $|y|$ increases, which suggests that $\widetilde{\Lambda}_{C,LAD}(y)$ is clearly inappropriate when the transformation of the response is not the identity. Indeed, the only case where $\widetilde{\Lambda}_{C,LAD}(y)$ performed well corresponds to a nonparametric regression model without transformation of the response.

Finally, we have also done a simulation in which we work with fixed bandwidths for h_y and h_u instead of the above stochastic rule-of-thumb bandwidths. The bandwidth h_x is estimated by $\widehat{h}_x = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_x$ as before. The results for model 1 and 2 are given in Section B of the supplementary material (the results for the other models being similar). The graphs in the supplement show that in most cases the MSE under the stochastic rule-of-thumb bandwidth is close to the optimal MSE that is obtained over a range of fixed bandwidths. This shows that the proposed bandwidth selection rule performs well in practice.

An optimal nonparametric estimator of the transformation could be obtained by combining the estimator given in Chiappori et al. (2015) when y is close to the points that define the identification conditions and our new estimator when y is far away from these points, so that we get the best from the two estimators. However, important questions arise such as the determination of the "critical" point(s), depending on the identification conditions and the true transformation, from which we have to switch from one estimator to the other. This could be subject to future research.

Finally, another way to improve the performance of these estimators could be to fix the points that define the identification conditions according to the interval in which we are interested to estimate the transformation.

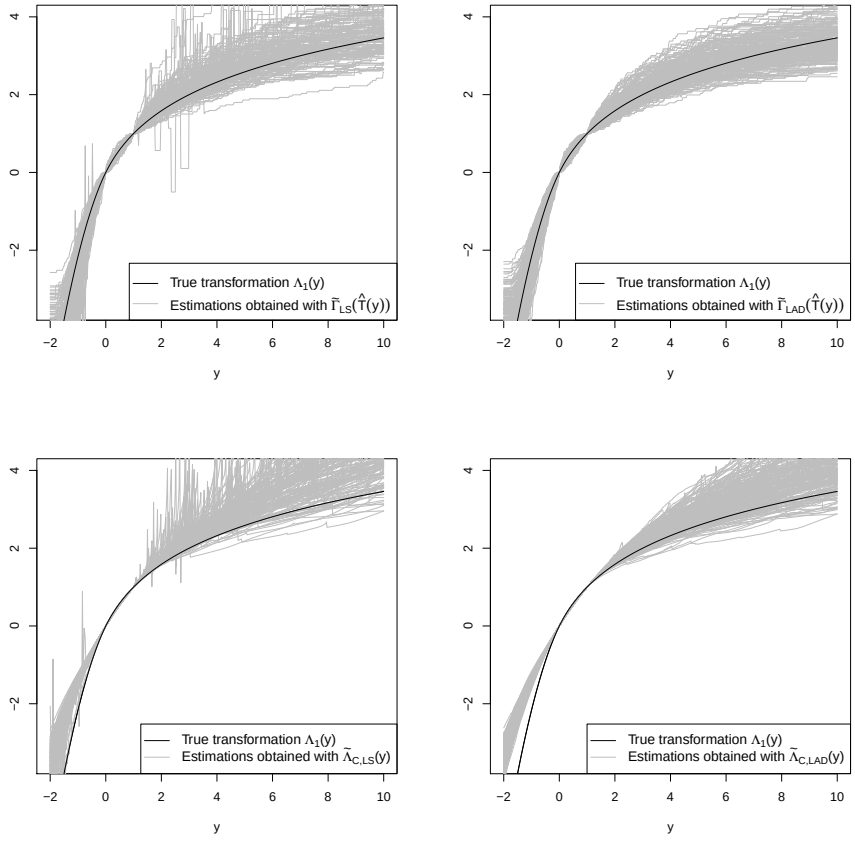

Figure 1. Comparison of the different nonparametric estimators of the transformation $\Lambda_1(Y)$.

Some	$\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$			$\widetilde{\Lambda}_{C,LAD}(y)$		
values of y	Bias	Var	MSE	Bias	Var	MSE
-1.5	0.2440	0.4682	0.5277	1.5102	0.0278	2.3087
-1	-0.1338	0.2151	0.2330	0.7600	0.0073	0.5850
0.5	0.0012	0.0076	0.0076	-0.0365	0.0001	0.0014
2	0.0577	0.0364	0.0397	0.0809	0.0033	0.0099
6	0.1121	0.1706	0.1831	0.3777	0.1420	0.2846
8	0.0222	0.2096	0.2101	0.5501	0.3234	0.6260
10	-0.0682	0.2466	0.2513	0.7581	0.5713	1.1460

Table 1. Estimations of the biases, variances and mean squared errors of the two LAD-estimators of $\Lambda_1(y)$ evaluated at some specific values of y.

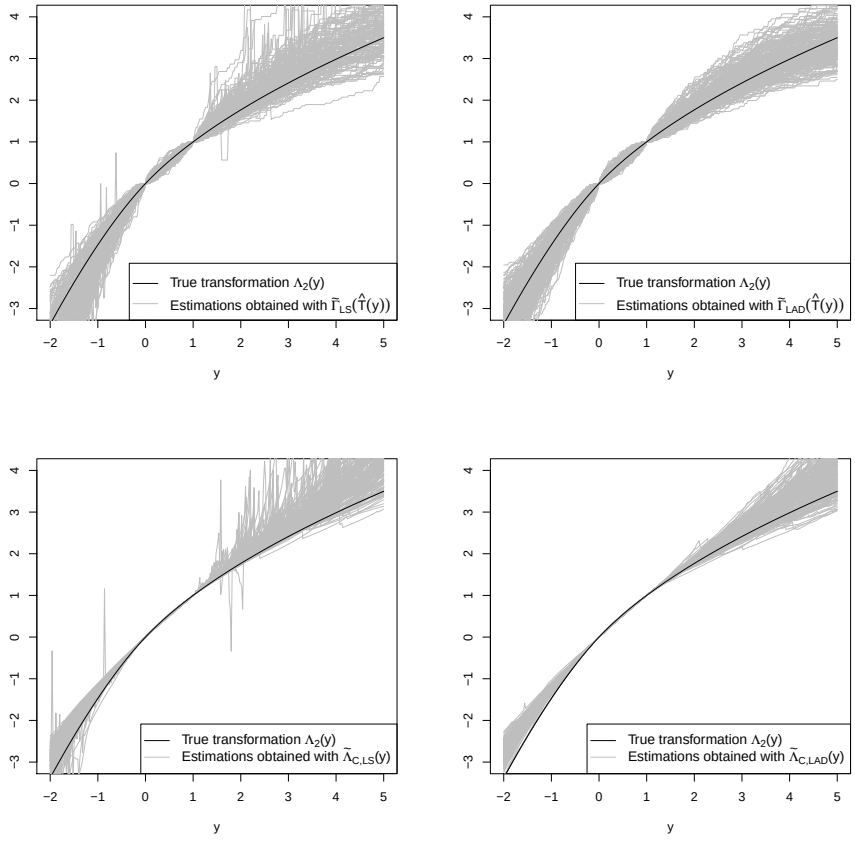

Figure 2. Comparison of the different nonparametric estimators of the transformation $\Lambda_2(Y)$.

Some	$\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$			$\widetilde{\Lambda}_{C,LAD}(y)$		
values of y	Bias	Var	MSE	Bias	Var	MSE
-2	0.0348	0.4137	0.4149	0.6138	0.0418	0.4185
-1.5	-0.1346	0.2417	0.2598	0.4054	0.0144	0.1788
0.5	-0.0470	0.0079	0.0079	-0.0156	0.0001	0.0003
2	0.0775	0.0479	0.0539	0.0642	0.0042	0.0083
4.5	-0.0012	0.2204	0.2204	0.3747	0.1665	0.3069
5	-0.0897	0.2440	0.2520	0.4769	0.2568	0.4842
6	-0.3565	0.2980	0.4251	0.7784	0.6872	1.2931

Table 2. Estimations of the biases, variances and mean squared errors of the two LAD-estimators of $\Lambda_2(y)$ evaluated at some specific values of y.

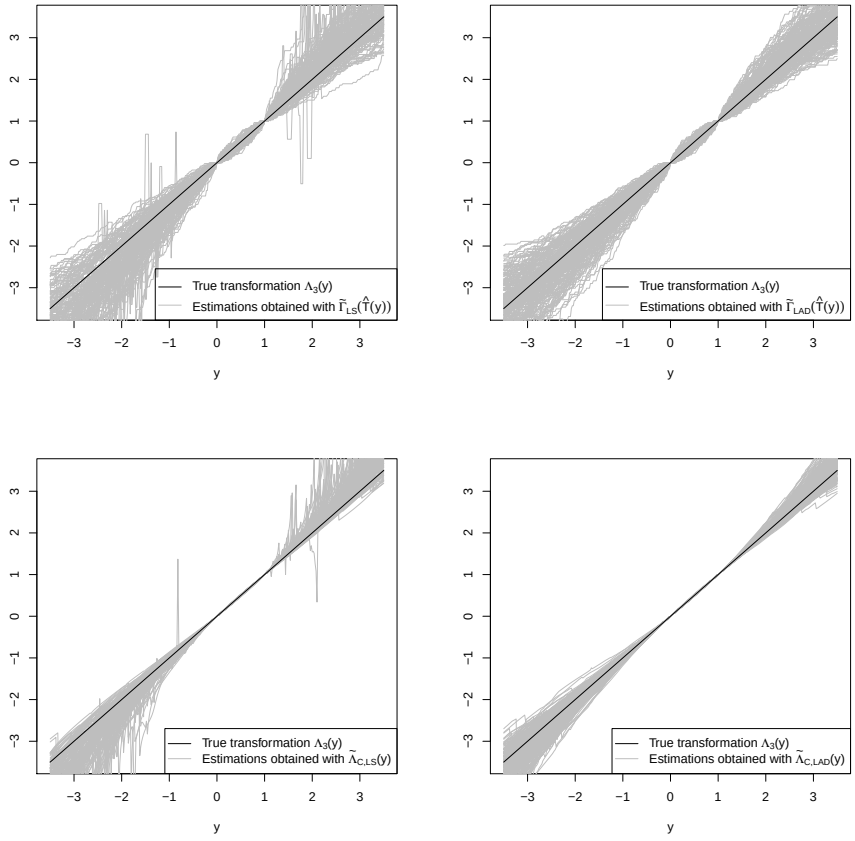

Figure 3. Comparison of the different nonparametric estimators of the transformation $\Lambda_3(Y)$.

Some	$\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$			$\widetilde{\Lambda}_{C,LAD}(y)$		
values of y	Bias	Var	MSE	Bias	Var	MSE
-3	-0.0830	0.3663	0.3732	-0.1350	0.1078	0.1260
-2	-0.1262	0.1836	0.1995	-0.0433	0.0298	0.0317
-1	-0.0530	0.0590	0.0618	-0.0063	0.0038	0.0039
0.5	-0.0062	0.0080	0.0080	-0.0016	0.0001	0.0001
1.5	0.0497	0.0324	0.0348	0.0107	0.0007	0.0008
2	0.1039	0.0783	0.0891	0.0360	0.0049	0.0062
3	0.0686	0.1915	0.1962	0.1361	0.0400	0.0585

Table 3. Estimations of the biases, variances and mean squared errors of the two LAD-estimators of $\Lambda_3(y)$ evaluated at some specific values of y.

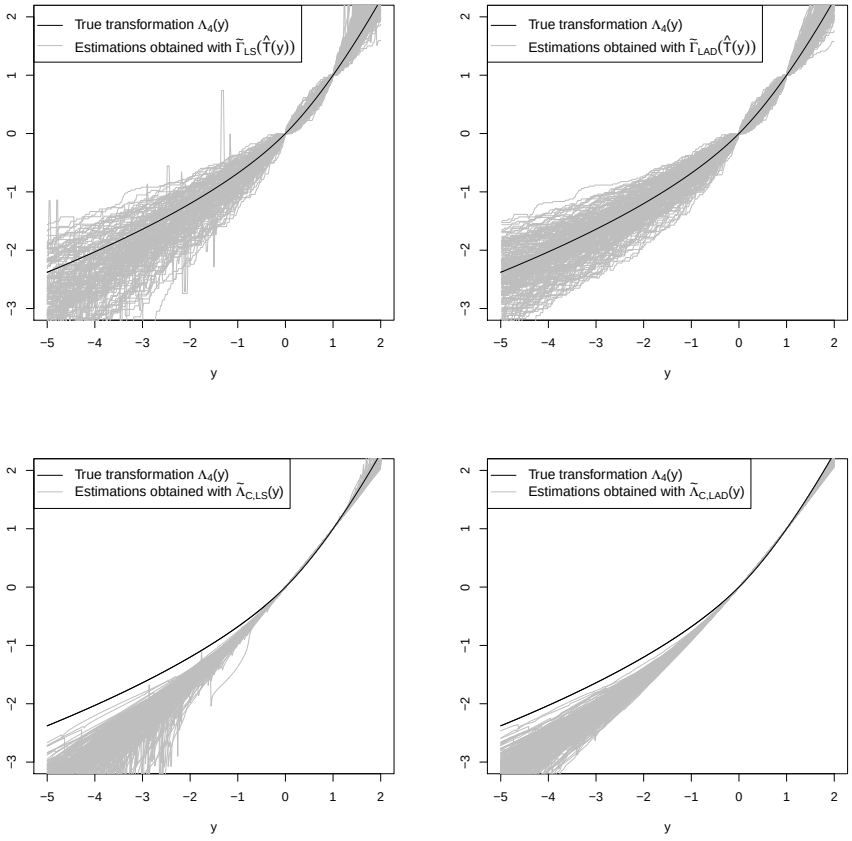

Figure 4. Comparison of the different nonparametric estimators of the transformation $\Lambda_4(Y)$.

Some	$\widetilde{\Gamma}_{LAD}(\widehat{T}(y))$			$\widetilde{\Lambda}_{C,LAD}(y)$		
values of y	Bias	Var	MSE	Bias	Var	MSE
-5	-0.1344	0.2411	0.2591	-0.8359	0.0917	0.7904
-4	-0.1235	0.1884	0.2037	-0.6624	0.0503	0.4892
-3	-0.0989	0.1357	0.1455	-0.5121	0.0225	0.2847
-2	-0.0640	0.0755	0.0796	-0.3534	0.0062	0.1311
-1	-0.0341	0.0349	0.0361	-0.1689	0.0008	0.0293
0.5	-0.0054	0.0082	0.0082	0.0230	0.00002	0.0006
1.5	0.0595	0.0381	0.0416	-0.0575	0.0002	0.0035

Table 4. Estimations of the biases, variances and mean squared errors of the two LAD-estimators of $\Lambda_4(y)$ evaluated at some specific values of y.

7. Appendix: Proofs of the main results

In this section, we give the proofs of the two main theorems and the corollary introduced in Section 5.2, regarding the limiting distributions of our new nonparametric estimators of the transformations $\Gamma(u)$ and $\Lambda(y)$ respectively.

Proof of Theorem 5.1. First, using successively Theorem 3.1, equation (4.6) and Proposition 1 in the supplementary material, we have

$$\begin{aligned}
\widehat{\lambda}_1(u,x) - \Gamma(u) &= \widehat{\lambda}_1(u,x) - \lambda_1(u,x) \\
&= \frac{\widehat{S}_1(u,x)}{\widehat{S}_1(1,x)} - \frac{S_1(u,x)}{S_1(1,x)} \\
&= \frac{1}{S_1(1,x)}\left\{\widehat{S}_1(u,x) - S_1(u,x)\right\} - \frac{S_1(u,x)}{S_1^2(1,x)}\left\{\widehat{S}_1(1,x) - S_1(1,x)\right\} \\
&\quad - \frac{1}{\widehat{S}_1(1,x)S_1(1,x)}\left\{\widehat{S}_1(1,x) - S_1(1,x)\right\}\left\{\widehat{S}_1(u,x) - S_1(u,x)\right\} \\
&\quad + \frac{S_1(u,x)}{\widehat{S}_1(1,x)S_1^2(1,x)}\left\{\widehat{S}_1(1,x) - S_1(1,x)\right\}^2 . \qquad (7.1)
\end{aligned}$$

Next, using the facts that

$$\left|\widehat{S}_1(1,x) - S_1(1,x)\right| \leq \sup_{(u,x)\in\mathcal{U}_0\times\chi_0} |\widehat{S}_1(u,x) - S_1(u,x)| = ||\widehat{S}_1 - S_1||_\infty ,$$

$$\left|\widehat{S}_1(u,x) - S_1(u,x)\right| \leq \sup_{(u,x)\in\mathcal{U}_0\times\chi_0} |\widehat{S}_1(u,x) - S_1(u,x)| = ||\widehat{S}_1 - S_1||_\infty ,$$

and $\sup_{(u,x)\in\mathcal{U}_0\times\chi_0} |S_1(u,x)| < \infty$ and $\inf_{x\in\chi_0} |S_1(1,x)| > 0$ by Conditions (A7), (A8) and the fact that $||\widehat{S}_1 - S_1||_\infty = o_P(1)$, we can conclude that the third and the fourth terms on the right hand side of (7.1) are $O(||\widehat{S}_1 - S_1||_\infty^2)$. These two terms are also $o_P(n^{-1/2})$ uniformly in $(u,x) \in \mathcal{U}_0 \times \chi_0$ using successively Propositions 5 and 4(ii) in the supplementary material.

Moreover, using Proposition 5 in the supplementary material and the facts that $\sup_{x\in\chi_0} |v(x)| < \infty$ and $\inf_{x\in\chi_0} |S_1(1,x)| > 0$ by Conditions (A8) and (A9), we have

$$\begin{aligned}
&\int_\chi \frac{v(x)}{S_1(1,x)}\left\{\widehat{S}_1(u,x) - S_1(u,x)\right\} dx \\
&= \int_\chi \frac{v(x)}{S_1(1,x)}\left\{\nabla_p S_1(u,x)[\widehat{p} - p] + \nabla_f S_1(u,x)[\widehat{f}_X - f_X]\right\} dx + o_P(n^{-1/2}) .
\end{aligned}$$

Hence, we can apply Propositions 6 and 7 in the supplementary material with $u_0 = 1$ and where $v_1(u_0,x) = \frac{v(x)}{S_1(u_0,x)}$ is the common function both for $v_p(\cdot)$ and $v_f(\cdot)$. We obtain:

$$\int_\chi \frac{v(x)}{S_1(1,x)}\left\{\widehat{S}_1(u,x) - S_1(u,x)\right\} dx = \frac{1}{n}\sum_{i=1}^n \delta_{X_i,Y_i}^{v_1}(1,u) + o_P(n^{-1/2}) , \qquad (7.2)$$

uniformly in $(u,x) \in \mathcal{U}_0 \times \chi_0$, where $\delta^{v_1}_{X_i,Y_i}(u_0,u)$ is defined in (5.3). Similarly, using successively Proposition 5 in the supplementary material, the facts that $\sup_{(u,x)\in\mathcal{U}_0\times\chi_0} |S_1(u,x)| < \infty$, $\inf_{x\in\chi_0} |S_1(1,x)| > 0$ and $\sup_{x\in\chi_0} |v(x)| < \infty$ as consequences of Conditions (A7), (A8) and (A9) and Propositions 6 and 7 in the supplementary material with $u_0 = u$ and where $v_2(u_0,x) = \frac{v(x)S_1(u_0,x)}{S_1^2(1,x)}$ is the common function both for $v_p(\cdot)$ and $v_f(\cdot)$, we obtain

$$\begin{aligned}
&\int_\chi \frac{v(x)S_1(u,x)}{S_1^2(1,x)} \left\{ \widehat{S}_1(1,x) - S_1(1,x) \right\} dx \\
&= \int_\chi \frac{v(x)S_1(u,x)}{S_1^2(1,x)} \left\{ \nabla_p S_1(1,x)[\widehat{p} - p] + \nabla_f S_1(1,x)[\widehat{f}_X - f_X] \right\} dx + o_P(n^{-1/2}) \\
&= \frac{1}{n} \sum_{i=1}^n \delta^{v_2}_{X_i,Y_i}(u,1) + o_P(n^{-1/2}) \,, \qquad (7.3)
\end{aligned}$$

uniformly in $(u,x) \in \mathcal{U}_0 \times \chi_0$, where $\delta^{v_2}_{X_i,Y_i}(u_0,u)$ is defined in (5.3). Next, using the definition of $\widehat{\Gamma}_{LS}(u)$ in (4.7) and the equations (7.1), (7.2) and (7.3), we have

$$\begin{aligned}
\sqrt{n}(\widehat{\Gamma}_{LS}(u) - \Gamma(u)) &= \sqrt{n} \int_\chi v(x) \{ \widehat{\lambda}_1(u,x) - \Gamma(u) \} \, dx \\
&= \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^n \delta^{v_1}_{X_i,Y_i}(1,u) - \frac{1}{n} \sum_{i=1}^n \delta^{v_2}_{X_i,Y_i}(u,1) + o_P(n^{-1/2}) \right\} \\
&= n^{-1/2} \sum_{i=1}^n \delta^{v}_{X_i,Y_i}(u) + o_P(1) \,,
\end{aligned}$$

where $\delta^{v}_{X_i,Y_i}(u) = \delta^{v_1}_{X_i,Y_i}(1,u) - \delta^{v_2}_{X_i,Y_i}(u,1)$. The pointwise weak convergence of $\sqrt{n}(\widehat{\Gamma}_{LS}(u) - \Gamma(u))$ to a Gaussian distribution is obtained using the central limit theorem. Next, to extend this result to weak functional convergence, we will prove that the class

$$\begin{aligned}
\mathcal{F} = \; & \left\{ (a,b,c) \to \int_{\max(0,a)}^u \left\{ \widetilde{v}(u_0,b) D_{p,0}(w,b) - \frac{\partial}{\partial x_1} \left[\widetilde{v}(u_0,x) D_{p,1}(w,x) \right] \Big|_{x=b} \right\} dw \right. \\
& + \int_0^u \left\{ \widetilde{v}(u_0,b) D_{f,0}(w,b) - \frac{\partial}{\partial x_1} \left[\widetilde{v}(u_0,x) D_{f,1}(w,x) \right] \Big|_{x=b} \right\} dw \\
& + (1_{\{a\le u\}} - 1_{\{a\le 0\}}) \widetilde{v}(u_0,b) D_{p,u}(a,b) \\
& + \int_0^u \left[\frac{1_{\{a\le w\}} - 1_{\{a\le 0\}}}{F_Y(1) - F_Y(0)} - w \right] \int_\chi \left(\left\{ \widetilde{v}(u_0,x) D_{p,0}(w,x) \right. \right. \\
& \left. \left. + \frac{\partial}{\partial x_1} [\widetilde{v}(u_0,x) D_{p,1}(w,x)] \right\} f_{U,X}(w,x) + \widetilde{v}(u_0,x) D_{p,u}(w,x) \frac{\partial}{\partial w} f_{U,X}(w,x) \right) dx \, dw \\
& - \left(\frac{1_{\{c\le 1\}} - 1_{\{c\le 0\}}}{F_Y(1) - F_Y(0)} - 1 \right) \int_0^u w \int_\chi \left\{ \widetilde{v}(u_0,x) D_{p,0}(w,x) \right.
\end{aligned}$$

$$-\widetilde{v}(u_0,x)\frac{\partial}{\partial w}D_{p,u}(w,x)+\frac{\partial}{\partial x_1}\left[\widetilde{v}(u_0,x)D_{p,1}(u,x)\right]\Big\}f_{U,X}(w,x)\,dx\,dw$$
$$-\left(\frac{1_{\{c\le 1\}}-1_{\{c\le 0\}}}{F_Y(1)-F_Y(0)}-1\right)u\int_{\mathcal{X}}\widetilde{v}(u_0,x)D_{p,u}(u,x)f_{U,X}(u,x)\,dx : u,u_0\in\mathcal{U}_0\Big\}\,,$$

where $\widetilde{v}$ will represent either v_1 or v_2, is Donsker. Indeed, $\mathcal{F}=\{(a,b,c)\to\delta^{\widetilde{v}}_{a,b}(u_0,u), u\in\mathcal{U}_0, u_0\in\mathcal{U}_0\}$. To obtain this result, it suffices to prove that the classes

$$\begin{aligned}\mathcal{F}_1 &= \Big\{(a,b)\to\int_{\max(0,a)}^{u}\Big\{\widetilde{v}(u_0,b)D_{p,0}(w,b)-\frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,x)D_{p,1}(w,x)\Big]\Big|_{x=b}\Big\}dw\\ &\quad+\int_0^u\Big\{\widetilde{v}(u_0,b)D_{f,0}(w,b)-\frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,x)D_{f,1}(w,x)\Big]\Big|_{x=b}\Big\}dw : u,u_0\in\mathcal{U}_0\Big\}\,,\end{aligned}$$

$$\mathcal{F}_2=\{a\to 1_{\{a\le u\}}-1_{\{a\le 0\}} : u\in\mathcal{U}_0\}\,,$$
$$\mathcal{F}_3=\{(a,b)\to\widetilde{v}(u_0,b)D_{p,u}(a,b) : u,u_0\in\mathcal{U}_0\},$$

$$\begin{aligned}\mathcal{F}_4 &= \Big\{a\to\frac{1}{F_Y(1)-F_Y(0)}\int_{\max(0,a)}^{u}\int_{\chi}\Big(\Big\{\widetilde{v}(u_0,x)D_{p,0}(w,x)\\ &\quad+\frac{\partial}{\partial x_1}[\widetilde{v}(u_0,x)D_{p,1}(w,x)]\Big\}f_{U,X}(w,x)\\ &\quad+\widetilde{v}(u_0,x)D_{p,u}(w,x)\frac{\partial}{\partial w}f_{U,X}(w,x)\Big)\,dx\,dw : u,u_0\in\mathcal{U}_0\Big\}\,,\end{aligned}$$

$$\mathcal{F}_5=\left\{a\to\frac{1_{\{a\le 0\}}}{F_Y(1)-F_Y(0)}\right\}$$

$$\begin{aligned}\mathcal{F}_6 &= \Big\{-\int_0^u\int_{\chi}\Big(\Big\{\widetilde{v}(u_0,x)D_{p,0}(w,x)+\frac{\partial}{\partial x_1}[\widetilde{v}(u_0,x)D_{p,1}(w,x)]\Big\}f_{U,X}(w,x)\\ &\quad+\widetilde{v}(u_0,x)D_{p,u}(w,x)\frac{\partial}{\partial w}f_{U,X}(w,x)\Big)\,dx\,dw : u,u_0\in\mathcal{U}_0\Big\}\,,\end{aligned}$$

$$\begin{aligned}\mathcal{F}_7 &= \Big\{-\int_0^u w\int_{\chi}\Big(\Big\{\widetilde{v}(u_0,x)D_{p,0}(w,x)+\frac{\partial}{\partial x_1}[\widetilde{v}(u_0,x)D_{p,1}(w,x)]\Big\}f_{U,X}(w,x)\\ &\quad+\widetilde{v}(u_0,x)D_{p,u}(w,x)\frac{\partial}{\partial w}f_{U,X}(w,x)\Big)\,dx\,dw : u,u_0\in\mathcal{U}_0\Big\}\,,\end{aligned}$$

$$\mathcal{F}_8=\left\{c\to-\frac{1_{\{c\le 1\}}-1_{\{c\le 0\}}}{F_Y(1)-F_Y(0)}+1\right\}\,,$$

$$\begin{aligned}\mathcal{F}_9 &= \Big\{\int_0^u w\int_{\chi}\Big\{\widetilde{v}(u_0,x)D_{p,0}(w,x)-\widetilde{v}(u_0,x)\frac{\partial}{\partial w}D_{p,u}(w,x)\end{aligned}$$

$$+\frac{\partial}{\partial x_1}\left[\widetilde{v}(u_0,x)D_{p,1}(u,x)\right]\Bigg\}f_{U,X}(w,x)\,dx\,dw : u,u_0\in\mathcal{U}_0\Bigg\}\,,$$

and

$$\mathcal{F}_{10}=\left\{\int_{\mathcal{X}}\widetilde{v}(u_0,x)D_{p,u}(u,x)f_{U,X}(u,x)\,dx : u,u_0\in\mathcal{U}_0\right\},$$

are Donsker and to apply Examples 2.10.7 and 2.10.8 in Van der Vaart and Wellner (1996) since $\mathcal{F}=\mathcal{F}_1+\mathcal{F}_2\mathcal{F}_3+\mathcal{F}_4+\mathcal{F}_5\mathcal{F}_6+\mathcal{F}_7+\mathcal{F}_8\mathcal{F}_9+\mathcal{F}_8\mathcal{F}_{10}$.

The reasoning that follows holds both for $\widetilde{v}=v_1$ and $\widetilde{v}=v_2$. First, note that all the functions that compose $\mathcal{F}_1$, $\mathcal{F}_3$, $\mathcal{F}_4$, $\mathcal{F}_6$, $\mathcal{F}_7$, $\mathcal{F}_9$ and $\mathcal{F}_{10}$ are at least m times continuously differentiable by Conditions (A7) and (A9) which implies that m is the smoothness of these classes. Hence, applying Corollary 2.7.2 in Van der Vaart and Wellner (1996), we get that $\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_1,L_2(P_1))\leq C_1\widetilde{\varepsilon}^{-(d+1)/m}$, where C_1 is some positive constant, $N_{[]}(\widetilde{\varepsilon},\mathcal{F}_1,L_2(P_1))$ is the $\widetilde{\varepsilon}$-bracketing number of the class $\mathcal{F}_1$, P_1 is the probability measure corresponding to the distribution of (U,X), and $L_2(P_1)$ is the L_2-norm. Consequently,

$$\int_0^{+\infty}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_1,L_2(P_1))}\,d\widetilde{\varepsilon}=\int_0^{2C}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_1,L_2(P_1))}\,d\widetilde{\varepsilon}\leq\frac{\sqrt{C_1}(2C)^{\frac{-d-1}{2m}+1}}{\frac{-d-1}{2m}+1}\,,$$

where C is a uniform upper bound for f_1 when $f_1\in\mathcal{F}_1$. The last expression is finite provided $m>(d+1)/2$, which is the case since $m>d+2$ by condition (A6). We obtained the first equality since only one $\widetilde{\varepsilon}$-bracket suffices to cover $\mathcal{F}_1$ if $\widetilde{\varepsilon}>2C$. Using exactly the same reasoning, we can prove for $j=3,4,6,7,9,10$ that $\int_0^{+\infty}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_j,L_2(P_j))}\,d\widetilde{\varepsilon}<\infty$ where $P_j\equiv P_1$.

Moreover, $\mathcal{F}_2$ is a class of increasing functions with image in $[0,1]$. Hence, applying Theorem 2.7.5 in Van der Vaart and Wellner (1996), we get that $\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_2,L_2(P_2))\leq C_2\widetilde{\varepsilon}^{-1}$ where C_2 is some positive constant and P_2 is the probability measure corresponding to the distribution of U. Hence,

$$\int_0^{+\infty}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_2,L_2(P_2))}\,d\widetilde{\varepsilon}=\int_0^{1}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_2,L_2(P_2))}\,d\widetilde{\varepsilon}\leq 2\sqrt{C_2}<\infty\,,$$

where the first equality was obtained since only one $\widetilde{\varepsilon}$-bracket suffices to cover $\mathcal{F}_2$ if $\widetilde{\varepsilon}>1$. Finally, only one bracket suffices to cover $\mathcal{F}_5$ and $\mathcal{F}_8$ since these classes of functions do not depend on u and u_0. Then, $\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_5,L_2(P_5))=0$ and $\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_8,L_2(P_8))=0$ where P_5 and P_8 are the probability measures corresponding to the distribution of Y.

Moreover, for $j=1,\ldots,10$, we will now prove that the envelope functions F_j of the classes $\mathcal{F}_j$ possess a weak second moment, i.e., $P(F_j>w)=o(w^{-2})$ as $w\to\infty$. The envelope function of $\mathcal{F}_j$ is some function F_j such that $|f_j(u,x)|\leq F_j(u,x)$ for all $(u,x)\in\mathcal{U}_0\times\chi_0$ and $f_j\in\mathcal{F}_j$.

We start with $\mathcal{F}_2$ which is the easier case. Indeed, for all functions $f_2\in\mathcal{F}_2$, we have $|f_2|\leq F_2=1$ uniformly in $(u,x)\in\mathcal{U}_0\times\chi_0$. Hence, using Markov's inequality, $P(F_2>w)\leq\frac{E(F_2^3)}{w^3}=\frac{1}{w^3}$, which is also $O(w^{-3})$ and $o(w^{-2})$. Next, if $f_5\in\mathcal{F}_5$ and $f_8\in\mathcal{F}_8$, we

necessarily have $|f_5| \leq F_5 = \frac{1}{F_Y(1)-F_Y(0)}$, and $|f_8| \leq F_8 = \frac{1}{F_Y(1)-F_Y(0)} + 1$, which are constants. Then, we have again $P(F_5 > w) = o(w^{-2})$ and $P(F_8 > w) = o(w^{-2})$.

Finally, the reasoning that follows is the same for the remaining classes and we explain it in detail only for $\mathcal{F}_1$. If $f_1 \in \mathcal{F}_1$, we have $|f_1| \leq F_1$ where F_1 is such that

$$\begin{aligned} F_1^3 \quad \leq \quad & C_4 \Bigg\{ \sup_{(u,u_0,x)\in\mathcal{U}_0^2\times\chi_0} \left|\widetilde{v}(u_0,x) D_{p,0}(u,x)\right|^3 + \sup_{(u,u_0,x)\in\mathcal{U}_0^2\times\chi_0} \left|\frac{\partial}{\partial x_1}(\widetilde{v}(u_0,x) D_{p,1}(u,x))\right|^3 \\ & + \sup_{(u,u_0,x)\in\mathcal{U}_0^2\times\chi_0} \left|\widetilde{v}(u_0,x) D_{f,0}(u,x)\right|^3 + \sup_{(u,u_0,x)\in\mathcal{U}_0^2\times\chi_0} \left|\frac{\partial}{\partial x_1}(\widetilde{v}(u_0,x) D_{f,1}(u,x))\right|^3 \Bigg\} , \end{aligned} \tag{7.4}$$

where C_4 is some positive constant. All the functions on the right hand side of (7.4) are continuous and the set $\mathcal{U}_0$ and χ_0 are compacts. Hence, $E(F_1^3) < \infty$ which implies that $P(F_1 > w)$ is $o(w^{-2})$ using Markov's inequality.

Consequently, for $j = 1, \ldots, 10$, we have proved that $\int_0^{+\infty} \sqrt{\log N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_2, L_2(P_j))}\, d\widetilde{\varepsilon} < \infty$ and the envelope functions F_j of $\mathcal{F}_j$ possess a weak second moment, which implies that the classes $\mathcal{F}_j$ are Donsker by Theorem 2.5.6 in Van der Vaart and Wellner (1996). In particular, the classes $\{(a,b) \to \delta_{a,b}^{v_1}(1,u), u \in \mathcal{U}_0\}$ and $\{(a,b) \to \delta_{a,b}^{v_2}(u,1), u \in \mathcal{U}_0\}$ are Donsker and thus also the class $\{(a,b) \to \delta_{a,b}^{v}(u), u \in \mathcal{U}_0\}$. As a consequence, the process $\sqrt{n}(\widehat{\Gamma}_{LS}(u) - \Gamma(u))$ converges to a Gaussian process.

Finally, we prove that $E\{\delta_{X_i,Y_i}^{\widetilde{v}}(u_0,u)\} = 0$, where $\delta_{X_i,Y_i}^{\widetilde{v}}(u_0,u)$ is defined in (5.3), which implies that $E\{\delta_{X_i,Y_i}^{v}(u)\} = 0$. First, we calculate the expectation of the first part of the first term on the right hand side of (5.3):

$$\begin{aligned} & E\left(\int_{\max(0,U_i)}^{u} \widetilde{v}(u_0,X_i) D_{p,0}(w,X_i)\, dw \right) \\ & = \int_{\chi}\int_{\mathcal{U}} \left[\int_{\max(0,t)}^{u} \widetilde{v}(u_0,x) D_{p,0}(w,x)\, dw \right] f_{U,X}(t,x)\, dt\, dx \\ & = \int_{\chi} \widetilde{v}(u_0,x) \int_{\mathcal{U}} \left[\int_{0}^{u} D_{p,0}(w,x) 1_{\{t\leq w\}}\, dw \right] f_{U,X}(t,x)\, dt\, dx \ . \end{aligned}$$

Rearranging the different integrals, the last expression is equal to

$$\begin{aligned} & \int_{\chi} \widetilde{v}(u_0,x) \int_{0}^{u} D_{p,0}(w,x) \left[\int_{\mathcal{U}} f_{U,X}(t,x) 1_{\{t\leq w\}}\, dt \right] dw\, dx \\ & = \int_{\chi} \widetilde{v}(u_0,x) \int_{0}^{u} D_{p,0}(w,x) \left[\int_{-\infty}^{w} f_{U,X}(t,x)\, dt \right] dw\, dx \end{aligned}$$

Using the definition of $p(u,x)$ given in (4.2), we obtain

$$E\left(\int_{\max(0,U_i)}^{u} \widetilde{v}(u_0,X_i) D_{p,0}(w,X_i)\, dw \right) = \int_{\chi} \widetilde{v}(u_0,x) \int_{0}^{u} D_{p,0}(w,x) p(w,x)\, dw\, dx \ . \tag{7.5}$$

Similarly, the expectation of the second part of the first term on the right hand side of (5.3) is equal to

$$\begin{aligned}
&E\bigg(-\int_{\max(0,U_i)}^{u} \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,\widetilde{x})D_{p,1}(w,\widetilde{x})\Big]\Big|_{\widetilde{x}=X_i} dw\bigg)\\
&= -\int_{\chi}\int_{0}^{u} \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,\widetilde{x})D_{p,1}(w,\widetilde{x})\Big]\Big|_{\widetilde{x}=x} p(w,x)\,dw\,dx\\
&= -\int_{\chi_{-1}}\int_{0}^{u} \Big[\widetilde{v}(u_0,x)D_{p,1}(w,x)p(w,x)\Big]_{\chi_1} dw\,dx_{-1}\\
&\quad +\int_{\chi}\int_{0}^{u} \widetilde{v}(u_0,x)D_{p,1}(w,x)p_1(w,x)\,dw\,dx\ , \qquad (7.6)
\end{aligned}$$

where χ_1 is the compact support of X_1, χ_{-1} is the compact support of $(X_2,\ldots,X_d)^t$, $x_{-1}=(x_2,\ldots,x_d)^t$, and the last expression was obtained using integration by parts. Using the same reasoning, the expectations of the first and second parts of the second term on the right hand side of (5.3) are respectively equal to

$$E\bigg(\int_0^u \widetilde{v}(u_0,X_i)D_{f,0}(w,X_i)\,dw\bigg) = \int_{\chi}\widetilde{v}(u_0,x)\int_0^u D_{f,0}(w,x)f_X(x)\,dw\,dx\ , \qquad (7.7)$$

and

$$\begin{aligned}
&E\bigg(-\int_{0}^{u} \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,\widetilde{x})D_{f,1}(w,\widetilde{x})\Big]\Big|_{\widetilde{x}=X_i} dw\bigg)\\
&= -\int_{\chi}\int_{0}^{u} \frac{\partial}{\partial x_1}\Big[\widetilde{v}(u_0,\widetilde{x})D_{f,1}(w,\widetilde{x})\Big]\Big|_{\widetilde{x}=x} f_X(x)\,dw\,dx\\
&= -\int_{\chi_{-1}}\int_{0}^{u} \Big[\widetilde{v}(u_0,x)D_{f,1}(w,x)f_X(x)\Big]_{\chi_1} dw\,dx_{-1}\\
&\quad +\int_{\chi}\widetilde{v}(u_0,x)\int_{0}^{u} D_{f,1}(w,x)f_{X,1}(x)\,dw\,dx\ , \qquad (7.8)
\end{aligned}$$

where the last equality was obtained using integration by parts. Moreover, the expectation of the third term on the right hand side of (5.3) is equal to

$$\begin{aligned}
&E\Big((1_{\{U_i\le u\}}-1_{\{U_i\le 0\}})\widetilde{v}(u_0,X_i)D_{p,u}(U_i,X_i)\Big)\\
&=\int_{\chi}\int_{\mathcal{U}}(1_{\{t\le u\}}-1_{\{t\le 0\}})\widetilde{v}(u_0,x)D_{p,u}(t,x)f_{U,X}(t,x)\,dt\,dx\\
&=\int_{\chi}\widetilde{v}(u_0,x)\int_0^u D_{p,u}(w,x)p_u(w,x)\,dw\,dx\ , \qquad (7.9)
\end{aligned}$$

where the last equality was obtained because $p_u(u,x)=f_{U,X}(u,x)$. Next, the expectation

of the fourth term on the right hand side of (5.3) is equal to 0 since

$$E\left[\frac{1_{\{U_i\le w\}}-1_{\{U_i\le 0\}}}{F_Y(1)-F_Y(0)}-w\right] = \frac{F_U(w)-F_U(0)}{F_Y(1)-F_Y(0)}-w = \frac{w\{F_Y(1)-F_Y(0)\}+F_Y(0)-F_U(0)}{F_Y(1)-F_Y(0)}-w=0\ .$$

We have used the fact that $U \sim \text{Un}\left[\frac{-F_Y(0)}{F_Y(1)-F_Y(0)}, \frac{1-F_Y(0)}{F_Y(1)-F_Y(0)}\right]$ which implies that $F_U(w) = w\{F_Y(1)-F_Y(0)\}+F_Y(0)$ and also $F_U(0) = F_Y(0)$. Similarly, the expectation of the fifth and sixth terms on the right hand side of (5.3) is also equal to 0. Indeed, it suffices to use the fact that $E(\widehat{F}_Y(1)-\widehat{F}_Y(0)) = F_Y(1)-F_Y(0)$. Finally, combining (7.5) to (7.9), we obtain

$$\begin{aligned}
&E(\delta^{\widetilde{v}}_{X_i,Y_i}(u_0,u))\\
&=\int_\chi \widetilde{v}(u_0,x)\int_0^u\Big\{D_{f,0}(w,x)f_X(x)+D_{f,1}(w,x)f_{X,1}(x)\\
&\quad+D_{p,0}(w,x)p(w,x)+D_{p,u}(w,x)p_u(w,x)+D_{p,1}(w,x)p_1(w,x)\Big\}\,dw\,dx\\
&\quad-\int\int_0^u\Big[\widetilde{v}(u_0,x)\{D_{p,1}(w,x)p(w,x)+D_{f,1}(w,x)f_X(x)\}\Big]_{\chi_1}\,dw\,dx_{-1}\\
&=\int_\chi \widetilde{v}(u_0,x)\Big\{\nabla_p S_1(u,x)[f_X]+\nabla_p S_1(u,x)[p]\Big\}\,dx=0\ .
\end{aligned}$$

Indeed, $\varphi(w,x)=\frac{p(w,x)}{f_X(x)}$ which implies that

$$\begin{aligned}
&D_{p,1}(w,x)p(w,x)+D_{f,1}(w,x)f_X(x)\\
&=\frac{-\varphi_u(w,x)}{f_X(x)\varphi_1^2(w,x)}\cdot p(w,x)+\frac{\varphi_u(w,x)\varphi(w,x)}{\varphi_1^2(w,x)f_X(x)}\cdot f_X(x)\\
&=\frac{-\varphi_u(w,x)\varphi(w,x)}{\varphi_1^2(w,x)}+\frac{\varphi_u(w,x)\varphi(w,x)}{\varphi_1^2(w,x)}=0\ ,
\end{aligned}$$

and we have also shown that $\nabla_p S_1(u,x)[p]=0$ and $\nabla_p S_1(u,x)[f_X]=0$ at the beginning of the proofs of Propositions 6 and 7 in the supplementary material respectively. □

Proof of Theorem 5.2. Recall that

$$\widehat{\Gamma}_{LAD,b}(u)=\arg\min_{q_m\in\mathbb{R}} Q_b(q_m,\widehat{\lambda}_1(u,\cdot))\ ,$$

where

$$Q_b(q_m,\lambda_1(u,\cdot))=\int_\chi v(x)\,(\lambda_1(u,x)-q_m)\,[2L_b(\lambda_1(u,x)-q_m)-1]\,dx\ . \tag{7.10}$$

We will prove that $\widehat{\Gamma}_{LAD,b}(u) = \widehat{\Gamma}_{LS}(u) + o_P(n^{-1/2})$ uniformly in u. First, the derivative of Q_b with respect to q_m is equal to

$$\begin{aligned}\frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m} &= -\int_\chi v(x)\left\{2L_b\left(\widehat{\lambda}_1(u,x) - q_m\right) - 1\right\} dx \\ &\quad -2\int_\chi v(x)\left\{\widehat{\lambda}_1(u,x) - q_m\right\} L_b'\left(\widehat{\lambda}_1(u,x) - q_m\right) dx \ ,\end{aligned}$$

where $L_b'(\cdot) = L'(\cdot/b)/b$. Consequently, $\frac{\partial Q_b(q_m,\Gamma(u))}{\partial q_m}|_{q_m=\Gamma(u)} = 0$ since $L(0) = 1/2$. Using this result and the fact that $\widehat{\Gamma}_{LAD,b}(u) = \arg\min_{q_m\in\mathbb{R}} Q_b(q_m, \widehat{\lambda}_1(u,\cdot))$ by definition, we have

$$\begin{aligned}0 &= \frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\widehat{\Gamma}_{LAD,b}(u)} - \frac{\partial Q_b(q_m,\Gamma(u))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} \\ &= \frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\widehat{\Gamma}_{LAD,b}(u)} - \frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} \\ &\quad + \frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} - \frac{\partial Q_b(q_m,\Gamma(u))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} .\end{aligned}$$

Next, using the mean value theorem with some $\widetilde{\Gamma}(u)$ between $\Gamma(u)$ and $\widehat{\Gamma}_{LAD,b}(u)$, the last expression becomes

$$\begin{aligned}&\frac{\partial^2 Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m^2}\bigg|_{q_m=\widetilde{\Gamma}(u)}\left(\widehat{\Gamma}_{LAD,b}(u) - \Gamma(u)\right) \\ &+ \frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} - \frac{\partial Q_b(q_m,\Gamma(u))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} = 0 \ ,\end{aligned}$$

Consequently,

$$\begin{aligned}&\widehat{\Gamma}_{LAD,b}(u) - \Gamma(u) \\ &= -\left(\frac{\partial^2 Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m^2}\bigg|_{q_m=\widetilde{\Gamma}(u)}\right)^{-1}\left(\frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} - \frac{\partial Q_b(q_m,\Gamma(u))}{\partial q_m}\bigg|_{q_m=\Gamma(u)}\right) .\end{aligned} \tag{7.11}$$

Next, recall the definition of the function Q_b in (7.10). The second derivative of Q_b with respect to q_m is equal to

$$\begin{aligned}\frac{\partial^2 Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m^2} &= 4\int_\chi v(x) L_b'\left(\widehat{\lambda}_1(u,x) - q_m\right) dx \\ &\quad +2\int_\chi v(x)\left\{\widehat{\lambda}_1(u,x) - q_m\right\} L_b''\left(\widehat{\lambda}_1(u,x) - q_m\right) dx \ ,\end{aligned}$$

where $L_b''(\cdot) = L''(\cdot/b)/b^2$. Consequently, using $\int_\chi v(x)\,dx = 1$, we have

$$\frac{\partial^2 Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m^2}\bigg|_{q_m=\widetilde{\Gamma}(u)} = 4L_b'(0) + 4\int_\chi v(x)\Big\{L_b'\left(\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right) - L_b'(0)\Big\}\,dx$$
$$+2\int_\chi v(x)\left\{\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right\} L_b''\left(\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right)\,dx\ .$$

Using the mean value theorem with some $\delta_1(u,x)$ between $\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)$ and 0, the last expression is equal to

$$4L_b'(0) + 4\int_\chi v(x) L_b''(\delta_1(u,x))\left(\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right)\,dx$$
$$+2\int_\chi v(x)\left\{\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right\} L_b''\left(\widehat{\lambda}_1(u,x) - \widetilde{\Gamma}(u)\right)\,dx\ .$$

Finally, we obtain

$$\frac{\partial^2 Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m^2}\bigg|_{q_m=\widetilde{\Gamma}(u)} = 4L_b'(0) + o_P(b^{-1})\ , \tag{7.12}$$

uniformly in u, since $||\widehat{\lambda}_1 - \widetilde{\Gamma}||_\infty \le ||\widehat{\lambda}_1 - \Gamma||_\infty = o_P(n^{-1/4})$ (which can be easily shown by combining the beginning of the proof of Theorem 5.1 with Propositions 4 and 5 in the supplementary material), and since $o_P(n^{-1/4})/b^2$ is $o_P(b^{-1})$ thanks to $nb^4 \to \infty$ by Condition (A10). Moreover, for some $\xi(u,x)$ in between $\Gamma(u)$ and $\widehat{\lambda}_1(u,x)$, we have

$$\frac{\partial Q_b(q_m, \widehat{\lambda}_1(u,\cdot))}{\partial q_m}\bigg|_{q_m=\Gamma(u)} - \frac{\partial Q_b(q_m, \Gamma(u))}{\partial q_m}\bigg|_{q_m=\Gamma(u)}$$
$$= -2\int_\chi v(x) L_b'\Big(\xi(u,x) - \Gamma(u)\Big)\left(\widehat{\lambda}_1(u,x) - \Gamma(u)\right)\,dx$$
$$-2\int_\chi v(x) L_b'\Big(\widehat{\lambda}_1(u,x) - \Gamma(u)\Big)\left(\widehat{\lambda}_1(u,x) - \Gamma(u)\right)\,dx$$
$$= -4L_b'(0)\int_\chi v(x)\left(\widehat{\lambda}_1(u,x) - \Gamma(u)\right)\,dx$$
$$-2\int_\chi v(x)\Big(L_b'\Big\{\xi(u,x) - \Gamma(u)\Big\} - L_b'(0)\Big)\left(\widehat{\lambda}_1(u,x) - \Gamma(u)\right)dx$$
$$-2\int_\chi v(x)\Big(L_b'\Big\{\widehat{\lambda}_1(u,x) - \Gamma(u)\Big\} - L_b'(0)\Big)\left(\widehat{\lambda}_1(u,x) - \Gamma(u)\right)\,dx$$

.

Similarly as before, using the mean value theorem with some $\delta_2(u,x)$ between $\xi(u,x) - \Gamma(u)$ and 0 and some $\delta_3(u,x)$ between $\widehat{\lambda}_1(u,x) - \Gamma(u)$ and 0, the last expression is equal to

$$-4L_b'(0)\Big(\widehat{\Gamma}_{LS}(u) - \Gamma(u)\Big) - 2\int_\chi v(x)\Big[L_b''(\delta_2(u,x))\Big\{\xi(u,x) - \Gamma(u)\Big\} \tag{7.13}$$

$$+L_b''\Big(\delta_3(u,x)\Big)\Big\{\widehat{\lambda}_1(u,x)-\Gamma(u)\Big\}\Big]\Big(\widehat{\lambda}_1(u,x)-\Gamma(u)\Big)\,dx.$$

We will show that the latter expression equals

$$-4L_b'(0)\big(\widehat{\Gamma}_{LS}(u)-\Gamma(u)\big)+o_P(n^{-1/2}b^{-1}), \tag{7.14}$$

since combining this with (7.11) and (7.12), leads to

$$\widehat{\Gamma}_{LAD,b}(u)-\Gamma(u) \;=\; \widehat{\Gamma}_{LS}(u)-\Gamma(u)+o_P(n^{-1/2}),$$

which is what we need to show. To prove (7.14), note that the second term of (7.13) is of the order $O_P(b^{-2}\,\|\widehat{\lambda}_1-\Gamma\|^2)$. It follows from the proof of Proposition 4 in the supplementary material that $\|\widehat{\lambda}_1-\Gamma\|^2$ is of the order $O_P(\max(h_x,h_u)^{2m})+O_P\big(\log n/\{nh_x^d\min(h_x,h_u)^2\}\big)$. Some straightforward algebra shows that this is $o_P(n^{-1/2}b)$ provided $n^{1/2}\max(h_x,h_u)^m = o(1)$, $nb^2\to\infty$ and $\log n/\{\sqrt{n}h_x^d\min(h_x,h_u)^2\}=o(b)$, which is guaranteed by assumptions (A6) and (A10). This concludes the proof. □

Proof of Corollary 5.1. Let $\widehat{\Gamma}$ be either $\widehat{\Gamma}_{LS}$ or $\widehat{\Gamma}_{LAD,b}$. First, it is shown in the proof of Theorem 5.1 that the class $\{(a,b)\to\delta_{a,b}^v(u), u\in\mathcal{U}_0\}$ is Donsker. Hence, since $\widehat{T}(y)-T(y)$ is $O_P(n^{-1/2})$ uniformly in $y\in\mathcal{Y}_0$ and then also $o_P(1)$, see the beginning of the proof of Proposition 3 in the supplementary material, we obtain using equation (2.1.8) in Van der Vaart and Wellner (1996):

$$\sup_{y\in\mathcal{Y}_0}\left|n^{-1}\sum_{i=1}^n\Big(\delta^v_{X_i,Y_i}(\widehat{T}(y))-\delta^v_{X_i,Y_i}(T(y))\Big)\right| = o_P(n^{-1/2})\;.$$

Consequently, using the fact that $\sqrt{n}(\widehat{\Gamma}(u)-\Gamma(u)) = n^{-1/2}\sum_{i=1}^n\delta^v_{X_i,Y_i}(u)+o_P(1)$ uniformly in $u\in\mathcal{U}_0$ by Theorems 5.1 and 5.2, we obtain:

$$\sup_{y\in\mathcal{Y}_0}\left|\widehat{\Gamma}(\widehat{T}(y))-\Gamma(\widehat{T}(y))-\Big(\widehat{\Gamma}(T(y))-\Gamma(T(y))\Big)\right| = o_P(n^{-1/2})\;,$$

and also uniformly in $y\in\mathcal{Y}_0$:

$$\begin{aligned}\widehat{\Gamma}(\widehat{T}(y))-\Gamma(T(y)) &= \widehat{\Gamma}(\widehat{T}(y))-\Gamma(\widehat{T}(y))+\Gamma(\widehat{T}(y))-\Gamma(T(y))\\ &= \widehat{\Gamma}(T(y))-\Gamma(T(y))+\Gamma(\widehat{T}(y))-\Gamma(T(y))+o_P(n^{-1/2})\;.\end{aligned}$$

Next, $\widehat{\Gamma}(T(y))-\Gamma(T(y)) = n^{-1}\sum_{i=1}^n\delta^v_{X_i,Y_i}(T(y))+o_P(n^{-1/2})$ and using a Taylor expansion with some ξ_y between $\widehat{T}(y)$ and $T(y)$, we have:

$$\Gamma(\widehat{T}(y))-\Gamma(T(y)) = \Gamma'(T(y))\Big(\widehat{T}(y)-T(y)\Big)+\frac{1}{2}\Gamma''(\xi_y)\Big(\widehat{T}(y)-T(y)\Big)^2\;.$$

Moreover, using the result given in Proposition 1 in the supplementary material, we have

$$\begin{aligned}\widehat{T}(y)-T(y) &= \frac{1}{F_Y(1)-F_Y(0)}\Big[\widehat{F}_Y(y)-\widehat{F}_Y(0)-F_Y(y)+F_Y(0)\Big]\\ &\quad -\frac{T(y)}{F_Y(1)-F_Y(0)}\Big[\widehat{F}_Y(1)-\widehat{F}_Y(0)-F_Y(1)+F_Y(0)\Big]\,.\end{aligned}$$

Hence, since $\widehat{F}_Y(y)=n^{-1}\sum_{i=1}^n 1_{\{Y_i\le y\}}$, $\sup_{y\in\mathcal{Y}_0}|\Gamma''(T(y))|<\infty$ and $(\widehat{T}(y)-T(y))^2$ is $o_P(n^{-1/2})$, we obtain uniformly in $y\in\mathcal{Y}_0$:

$$\widehat{\Gamma}(\widehat{T}(y))-\Gamma(T(y))=n^{-1}\sum_{i=1}^n\varphi^v_{X_i,Y_i}(y)+o_P(n^{-1/2})\,,$$

where $\varphi^v_{X_i,Y_i}(y)$ is defined in (5.4). The pointwise weak convergence of $\sqrt{n}(\widehat{\Gamma}(\widehat{T}(y))-\Lambda(y))$, where $\Lambda(y)=\Gamma(T(y))$, to a Gaussian distribution is obtained using the central limit theorem. Next, to extend this result to weak functional convergence, we will prove that the class $\mathcal{F}=\{(a,b)\to\varphi^v_{a,b}(y),y\in\mathcal{Y}_0\}$ is Donsker. To obtain this result, it suffices to prove that the classes

$$\mathcal{F}_1=\{(a,b)\to\delta^v_{a,b}(T(y)),y\in\mathcal{Y}_0\}\quad,\quad\mathcal{F}_2=\Big\{\frac{\Gamma'(T(y))}{F_Y(1)-F_Y(0)}:y\in\mathcal{Y}_0\Big\}\,,$$

$$\mathcal{F}_3=\Big\{b\to 1_{\{b\le y\}}-1_{\{b\le 0\}}:y\in\mathcal{Y}_0\Big\}\quad,\quad\mathcal{F}_4=\Big\{-(F_Y(y)-F_Y(0)):y\in\mathcal{Y}_0\Big\}\,,$$

and

$$\mathcal{F}_5=\Big\{b\to -T(y)\Big[1_{\{0\le b\le 1\}}-F_Y(1)+F_Y(0)\Big]:y\in\mathcal{Y}_0\Big\}\,,$$

are Donsker and to apply Examples 2.10.7 and 2.10.8 in Van der Vaart and Wellner (1996) since $\mathcal{F}=\mathcal{F}_1+\mathcal{F}_2(\mathcal{F}_3+\mathcal{F}_4+\mathcal{F}_5)$. In the proof of Theorem 5.1, we proved that the class $\mathcal{F}_1$ is Donsker. Moreover, for $j=2,\dots,5$, we have to prove that

$$\int_0^{+\infty}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_j,L_2(P))}\,d\widetilde{\varepsilon}<\infty\,,$$

where P is the probability measure corresponding to the distribution of Y and $L_2(P)$ is the L_2-norm, that the envelope functions F_j of $\mathcal{F}_j$ possess a weak second moment and to apply Theorem 2.5.6 in Van der Vaart and Wellner (1996).

First, for $j=2,\dots,5$, we have $\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_j,L_2(P))\le K_j\widetilde{\varepsilon}^{-1}$, for some constants $K_j>0$, using Corollary 2.7.2 in Van der Vaart and Wellner (1996) with $\alpha=1$ and $d=1$ for $j=2$ and Theorem 2.7.5 in Van der Vaart and Wellner (1996) for $j=3,4,5$ since $\mathcal{F}_3$, $\mathcal{F}_4$ and $\mathcal{F}_5$ are classes of monotone and bounded functions. Hence, if C_j denotes a uniform upper bound for $f_j\in\mathcal{F}_j$,

$$\int_0^{+\infty}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_j,L_2(P))}\,d\widetilde{\varepsilon}=\int_0^{2C_j}\sqrt{\log N_{[]}(\widetilde{\varepsilon},\mathcal{F}_j,L_2(P))}\,d\widetilde{\varepsilon}\le\int_0^{2C_j}\sqrt{K_j\widetilde{\varepsilon}^{-1}}\,d\widetilde{\varepsilon}\,.$$

We obtained the first equality since only one $\widetilde{\varepsilon}$-bracket suffices to cover $\mathcal{F}_j$ if $\widetilde{\varepsilon} > 2C_j$ and the last expression is equal to $2\sqrt{2C_jK_j}$ and is finite for $j = 2, \ldots, 5$.

Next, for $j = 2, \ldots, 5$, if $f_j \in \mathcal{F}_j$, we have $|f_j| \leq F_j$ where $F_2 = \frac{\sup_{y\in\mathcal{Y}_0} |\Gamma'(T(y))|}{F_Y(1)-F_Y(0)}$, $F_3 = 1$, $F_4 = 1$ and $F_5 = \frac{1}{F_Y(1)-F_Y(0)}$. Since all F_j are finite, we have $P(F_j > w) = o(w^{-2})$.

Finally, $E(\varphi^v_{X,Y}(y)) = 0$ since $E(\delta^v_{X,Y}(T(y))) = 0$ by Theorem 5.1, $E(1_{\{Y\leq y\}} - 1_{\{Y\leq 0\}}) = F_Y(y) - F_Y(0)$ and $E(1_{\{0\leq Y\leq 1\}}) = F_Y(1) - F_Y(0)$ which concludes the proof. □

Acknowledgements

We would like to thank the referees, the Associate Editor and the Editor for their valuable remarks on an earlier version of this paper. Moreover, we would like to thank Nick Kloodt (University of Hamburg) and a referee, who found a mistake in one of our proofs.
B. Colling and I. Van Keilegom are financially supported by the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409), and by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy).

References

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 (Methodological)*, 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.

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.

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

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.

Dette, H., Neumeyer, N., and Pilz, K. F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, 12(3):469–490.

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.

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 (Applied Statistics)*, 29(2):190–197.

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

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

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 (The Statistician)*, 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*, 98(4):954–959.

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