DISCUSSION PAPER

2016/31

Goodness-of-fit tests in semiparametric transformation models using the integrated regression function

Colling, B. and I. Van Keilegom

Goodness-of-fit tests in semiparametric transformation models using the integrated regression function

Benjamin Colling [*] Ingrid Van Keilegom [*,§]

July 22, 2016

Abstract

Consider the following semiparametric transformation model $\Lambda_\theta(Y) = m(X) + \varepsilon$, where X is a d-dimensional covariate, Y is a univariate dependent variable and ε is an error term with zero mean and which is 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. We use a profile likelihood estimator for the parameter θ and a local polynomial estimator for m. Our goal is to develop a new test for the parametric form of the regression function m, which has power against all local alternatives that converge to the null model at parametric rate, and to compare its performance to that of the test proposed by Colling and Van Keilegom (2016). The idea of the new test is to compare the integrated regression function estimated in a semiparametric way to the integrated regression function estimated under the null hypothesis. We consider two different test statistics, a Kolmogorov-Smirnov and a Cramér-von Mises type statistic, and establish the limiting distributions of these two test statistics under the null hypothesis and under a local alternative. We use a bootstrap procedure to approximate the critical values of the test statistics under the null hypothesis. Finally, a simulation study is carried out to illustrate the performance of our testing procedure, to compare this new test to the previous one and to see under which model conditions which test behaves the best. We also apply both methods on a real data set.

Key Words: Bootstrap; Goodness-of-fit; Integrated regression function; Local polynomial smoothing; Profile likelihood; Semiparametric regression; Transformation model.

---

[*] Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain. Research supported by IAP research network grant nr. P7/06 of the Belgian government (Belgian Science Policy), and by the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

[§] Research also supported by the European Research Council (2016-2021, Horizon 2020 / ERC grant agreement No. 694409).

1 Introduction

The simple linear regression model is the most commonly used model in statistics when we want to explain the relationship between a dependent variable Y and a vector of explanatory variables denoted X. However, this model relies on heavy assumptions that are not always satisfied in practice, namely the structure of this model is additive and linear, the variance of the error term ε is constant and ε is normally distributed. 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. Note that the Box and Cox (1964) transformation also includes as special cases the logarithm and the identity.

This class of transformation has been generalized, see for example the Yeo and Johnson (2000) transform. Other types of transformations have also been introduced in the literature, e.g. the Zellner and Revankar (1969), the John and Draper (1980), the Bickel and Doksum (1981) and the MacKinnon and Magee (1990) transforms among others. We also refer to the book of Carroll and Ruppert (1988).

All the above mentioned papers consider a model where both the transformation and the regression function are parametric. In the literature, we can also find papers where both the transformation and the regression function are nonparametric, e.g. Breiman and Friedman (1985), Horowitz (2001) and Jacho-Chavez, Lewbel and Linton (2008), and papers where the transformation is nonparametric and the regression function is parametric, e.g. Horowitz (1996).

In this paper, we will focus on a model where the transformation is parametric and the regression function is nonparametric, i.e. we will consider a semiparametric transformation model of the following form :

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

where $m(\cdot)$ is an unknown regression function, $\Lambda_\theta(\cdot)$ is some parametric transformation of the response variable Y and $\theta \in \Theta$ where Θ is a finite dimensional compact subset of $\mathbb{R}^k$. We will denote by θ_0 and $m_0(\cdot)$ the true but unknown values of θ and $m(\cdot)$. Moreover, we assume that X is a d-dimensional covariate, Y is a univariate response variable and the error term ε has zero mean and is independent of X.

Linton, Sperlich and Van Keilegom (2008) have extensively studied the semiparametric transformation model (1.1). Their main objective was to propose different estimators of the transformation parameter θ and to establish the asymptotic properties of these estimators. Vanhems and Van Keilegom (2016) have also studied the estimation of this model supposing that some of the regressors are endogenous as a result of e.g. omitted variables, measurement error or simultaneous equations. We also like to mention the works of Colling, Heuchenne, Samb and Van Keilegom (2015) and Heuchenne, Samb and Van Keilegom (2015) who introduced and studied respectively nonparametric estimators for the error density function and the error distribution function. Moreover, Colling and Van Keilegom (2016) developed a test for the following null hypothesis :

$$H_0 : m \in \mathcal{M} \ , \tag{1.2}$$

where $\mathcal{M} = \{m_\beta : \beta \in \mathcal{B}\}$ is some parametric class of regression functions and $\mathcal{B} \subset \mathbb{R}^q$. The main idea of their test was to compare the distribution function of the error term estimated in a semiparametric way to the distribution function of the error term estimated under H_0.

We also like to mention the work of Neumeyer, Noh and Van Keilegom (2016). Recently, they introduced estimators for the different components of a heteroscedastic transformation model and proved the asymptotic normality of these estimators. They also proposed a test for the validity of this model.

The main objective of this paper is to develop a second test for the null hypothesis (1.2), which has power against all local alternatives that converge to the null model at parametric rate, and to compare this new test to the previous one developed by Colling and Van Keilegom (2016). The basic idea of the new test is to compare the integrated regression function estimated in a semiparametric way to the integrated regression function estimated under H_0. The idea of testing the form of the regression function using the integrated regression function has been studied among others by Bierens (1982), Stute (1997) and Escanciano (2006a). These three articles worked in a context of a nonparametric regression model without transformation of the response variable. The first consistent integrated test was proposed by Bierens (1982). He defined the following Cramér-von Mises test statistic :

$$\int [n^{-1/2} \sum_{j=1}^{n} e^{ix^t X_j} (Y_j - m_{\widehat{\beta}}(X_j))]^2 \Phi(x) \, dx \ ,$$

where i is the imaginary unit, $\widehat{\beta}$ is the least squares estimator of β and Φ is a positive integrating function, for example a d-variate normal density. The test of Stute (1997) was based on the following residual process :

$$n^{-1/2}\sum_{j=1}^{n} 1_{\{X_j\le x\}}(Y_j - m_{\widehat{\beta}}(X_j)) \ ,$$

where $1_{\{X\le x\}}$ is a component by component indicator. Finally, the test of Escanciano (2006a) was based on the following residual process :

$$n^{-1/2}\sum_{j=1}^{n} 1_{\{\gamma^t X_j\le x\}}(Y_j - m_{\widehat{\beta}}(X_j)) \ ,$$

where γ is a d-dimensional vector. The main difference between these three approaches is the weigthing function that each author uses to construct his residual process. More generally, this class of tests is based on the equivalence

$$E(\varepsilon|X) = 0 \text{ a.s.} \Longleftrightarrow E(\varepsilon w(X,x,\gamma)) = 0 \quad \forall (x,\gamma)\in\Pi \ , \tag{1.3}$$

where Π is a properly chosen space and $w(\cdot,x,\gamma)$ is a parametric family such that the equivalence (1.3) holds. Bierens and Ploberger (1997), Stinchcombe and White (1998) and Escanciano (2006b) among others propose some primitive conditions on the family of weighting functions $w(\cdot,x,\gamma)$ so that the equivalence (1.3) is satisfied, including $w(X,x,\gamma) = \exp(ix^tX)$, $w(X,x,\gamma) = 1_{\{X\le x\}}$ and $w(X,x,\gamma) = 1_{\{\gamma^tX\le x\}}$, the weighting functions used by Bierens (1992), Stute (1997) and Escanciano (2006a) respectively. Other possibilities are for example $w(X,x,\gamma) = \exp(x^tX)$, $w(X,x,\gamma) = (1+\exp(c-x^tX))^{-1}$ for some constant c, $w(X,x,\gamma) = \sin(x^tX)$ and $w(X,x,\gamma) = \sin(x^tX)+\cos(x^tX)$. In the context of nonparametric regression without transformation of the response, this class of tests, which is called "the integrated approach", avoids the use of smoothing methods which is an important advantage. In this paper, our goal is to extend this class of tests to the context of semiparametric transformation models.

Many papers in the literature use this integrated approach in other contexts, in time series for example. The most frequently used weighting functions are $w(X,x,\gamma) = \exp(ix^tX)$, see Bierens (1984) and Bierens (1990) for example, and $w(X,x,\gamma) = 1_{\{X\le x\}}$, see Koul and Stute (1999) and Whang (2000) for example. We also like to mention the work of Stute and Zhu (2002) who use a similar approach as Escanciano (2006a) except that Stute and Zhu (2002) estimate the parameter γ.

There are other ways to construct tests for (1.2) instead of using the idea based on the integrated regression function. We could for example define a test using the approach of

Härdle and Mammen (1993) among others. A recent overview on goodness-of-fit tests for regression models was given by González-Manteiga and Crujeiras (2013).

The paper is organized as follows. In Section 2, we explain in detail how we can estimate a semiparametric transformation model and we define our testing procedure. In Section 3, we present the main results of the asymptotic theory and in particular the limiting distributions of the proposed test statistics. Section 4 contains a simulation study that shows the performance of the proposed test and compares this new test to the previous test of Colling and Van Keilegom (2016). In Section 5, we apply our method to a real data set and Section 6 contains the conclusions. Finally, the Appendix contains the proofs of the main results.

2 The proposed test

2.1 Notations and definitions

We suppose that X has compact support $\chi \subset \mathbb{R}^d$. 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 semiparametric transformation model (1.1). We denote by F_X, f_X, F_ε and f_ε the distribution and the probability density functions of X and ε respectively. Moreover, let $\sigma^2 = V(\varepsilon) < \infty$ and define the function

$$m(x, \theta) = E[\Lambda_\theta(Y)|X = x] \ .$$

Note that $m(x, \theta_0) = m(x)$. We also denote

$$\frac{\partial}{\partial x} f_X(x) = \left(\frac{\partial}{\partial x_1} f_X(x), \dots, \frac{\partial}{\partial x_d} f_X(x) \right)^t \ ,$$

which is a $(d \times 1)$-vector where $x = (x_1, \dots, x_d)^t$, and let

$$\dot{\Lambda}_\theta(y) = \left(\frac{\partial}{\partial \theta_1} \Lambda_\theta(y), \dots, \frac{\partial}{\partial \theta_k} \Lambda_\theta(y) \right)^t$$

be a $(k \times 1)$-vector where $\theta = (\theta_1, \dots, \theta_k)^t$. Similar notations will be used for other functions. For any function φ, we define $\varphi'(u) = \partial \varphi / \partial u$. Finally, let $\varepsilon(\theta) = \Lambda_\theta(Y) - m(X, \theta)$ and let $F_{\varepsilon(\theta)}$ and $f_{\varepsilon(\theta)}$ be the distribution and the density function of $\varepsilon(\theta)$, respectively.

2.2 Estimation of the model

In this section, we will introduce the estimators of the transformation parameter θ and of the regression function $m(x, \theta)$ that we will use throughout this paper. We will proceed in exactly the same way as in Colling and Van Keilegom (2016). We will estimate θ by the profile likelihood estimator developed by Linton, Sperlich and Van Keilegom (2008).

The basic idea of the profile likelihood method is to calculate the log-likelihood function of Y given X and to replace unknown expressions by nonparametric estimators, which gives us the following estimator of θ :

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \left\{ \log \widehat{f}_{\varepsilon(\theta)}(\Lambda_\theta(Y_i) - \widehat{m}(X_i, \theta)) + \log \Lambda'_\theta(Y_i) \right\} , \tag{2.1}$$

where $\widehat{m}(x, \theta)$ and $\widehat{f}_{\varepsilon(\theta)}(y)$ are respectively nonparametric estimators of the unknown regression function $m(x, \theta)$ and of the error density function $f_{\varepsilon(\theta)}(y)$. More precisely, here we will estimate the unknown regression function by a local polynomial estimator of degree p (like in Neumeyer and Van Keilegom (2010)), i.e. let $h = (h_1, \ldots, h_d)^t$ be a d-dimensional bandwidth vector and let $K_1(u)$ be a d-dimensional product kernel of the form $K_1(u) = \prod_{j=1}^{d} k_1(u_j)$ where k_1 is a univariate kernel. Then, for an arbitrary point $x = (x_1, \ldots, x_d)^t$ in the support χ of X, $\widehat{m}(x, \theta) = \widehat{b}_0(\theta)$ where $\widehat{b}_0(\theta)$ is the first component of the vector $\widehat{b}(\theta)$, which is the solution of the following local minimization problem :

$$\min_b \sum_{i=1}^{n} (\Lambda_\theta(Y_i) - P_i(b, x, p))^2 K_1\left(\frac{X_i - x}{h}\right) ,$$

where $P_i(b, x, p)$ is a polynomial of order p built up with all products of $0 \leq l \leq p$ factors of the form $X_{ij} - x_j$ for $j = 1, \ldots, d$. We will use the notation $\widehat{m}(x) = \widehat{m}(x, \widehat{\theta})$ when there is no ambiguity. Moreover, $\widehat{f}_{\varepsilon(\theta)}(y)$ is the following kernel estimator of the error density function :

$$\widehat{f}_{\varepsilon(\theta)}(y) = \frac{1}{ng} \sum_{i=1}^{n} k_2\left(\frac{y - \widehat{\varepsilon}_i(\theta)}{g}\right) ,$$

where $\widehat{\varepsilon}_i(\theta) = \Lambda_\theta(Y_i) - \widehat{m}(X_i, \theta)$, k_2 is a kernel and g is a bandwidth.

It is important to remark that we assume a completely unspecified regression function $m(\cdot)$ which is slightly different from what Linton, Sperlich and Van Keilegom (2008) assume, since they assume an additive or multiplicative structure on $m(\cdot)$. Moreover, we have estimated this regression function by a local polynomial estimator whereas Linton, Sperlich and Van Keilegom (2008) used a higher order kernel estimator.

Finally, Colling and Van Keilegom (2016) proved that the following asymptotic representation for $\widehat{\theta} - \theta_0$ and the following limiting distribution of $n^{1/2}(\widehat{\theta} - \theta_0)$ obtained by Linton, Sperlich and Van Keilegom (1998) stay valid when $m(\cdot)$ is completely unspecified and is estimated by a local polynomial estimator :

$$\widehat{\theta} - \theta_0 = -n^{-1} \sum_{i=1}^{n} g(X_i, Y_i) + o_P(n^{-1/2}) \ ,$$

and

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, V(g(X, Y))) \ ,$$

where $g(X, Y) = \Gamma^{-1}\xi(\theta_0, X, Y)$,

$$\xi(\theta, X, Y) = \frac{1}{f_{\varepsilon(\theta)}(\varepsilon(\theta))}[f'_{\varepsilon(\theta)}(\varepsilon(\theta))(\dot{\Lambda}_\theta(Y) - \dot{m}(X, \theta)) + \dot{f}_{\varepsilon(\theta)}(\varepsilon(\theta))] + \frac{\dot{\Lambda}'_\theta(Y)}{\Lambda'_\theta(Y)} \ ,$$

and

$$\Gamma = \frac{\partial}{\partial \theta} E[\xi(\theta, X, Y)] \bigg|_{\theta=\theta_0} \ .$$

The assumptions under which these results are valid are given in Colling and Van Keilegom (2016).

2.3 The test statistics

We will introduce two new test statistics where the basic idea is to compare the integrated regression function estimated in a semiparametric way to the integrated regression function estimated under H_0. We consider the following integrated regression function :

$$M(x, \gamma, \theta) = \int w(t, x, \gamma) m(t, \theta) \, dF_X(t) = E[w(X, x, \gamma)\Lambda_\theta(Y)] \ ,$$

where w is some weighting function that depends on some parameter $\gamma \in \mathbb{R}^{d_\gamma}$ and that satisfies the equivalence (1.3). We assume that γ has compact support $\chi_\gamma \subset \mathbb{R}^{d_\gamma}$. The empirical analog of $M(x, \gamma, \theta)$ is given by

$$\widehat{M}(x, \gamma, \theta) = n^{-1} \sum_{i=1}^{n} w(X_i, x, \gamma)\Lambda_\theta(Y_i) \ .$$

Next, under H_0, $m \equiv m_{\beta_0}$ where β_0 is the true value of β under H_0. Then, the integrated regression function becomes

$$M_{\beta_0}(x, \gamma, \theta_0) = \int w(t, x, \gamma) m_{\beta_0}(t) \, dF_X(t) \ ,$$

and its empirical analog is given by

$$\widehat{M}_{\beta_0}(x,\gamma,\theta_0) = n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma) m_{\beta_0}(X_i) \ .$$

Hence, our test will be constructed on the basis of the following residual process :

$$R_n(x,\gamma,\theta_0,\beta_0) = \sqrt{n}(\widehat{M}(x,\gamma,\theta_0) - \widehat{M}_{\beta_0}(x,\gamma,\theta_0)) = n^{-1/2}\sum_{i=1}^{n} w(X_i,x,\gamma)(\Lambda_{\theta_0}(Y_i) - m_{\beta_0}(X_i)) \ .$$

Finally, as the parameters θ and β are unknown, we will estimate θ by the profile likelihood estimator defined in (2.1) and we will estimate β by a least squares estimator, i.e. we consider $\widehat{\beta}$ which is a minimizer over $\beta \in \mathcal{B}$ of the expression

$$S_n(\beta) = n^{-1}\sum_{i=1}^{n}(\Lambda_{\widehat{\theta}}(Y_i) - m_\beta(X_i))^2 \ . \tag{2.2}$$

This gives the following residual process :

$$R_n(x,\gamma,\widehat{\theta},\widehat{\beta}) = \sqrt{n}(\widehat{M}(x,\gamma,\widehat{\theta}) - \widehat{M}_{\widehat{\beta}}(x,\gamma,\widehat{\theta})) = n^{-1/2}\sum_{i=1}^{n} w(X_i,x,\gamma)(\Lambda_{\widehat{\theta}}(Y_i) - m_{\widehat{\beta}}(X_i,\widehat{\theta})) \ .$$

For an easier readability, we will use the notations $m_{\widehat{\beta}}(X_i) = m_{\widehat{\beta}}(X_i,\widehat{\theta})$, $R_n(x,\gamma) = R_n(x,\gamma,\theta_0,\beta_0)$ and $R_n^1(x,\gamma) = R_n(x,\gamma,\widehat{\theta},\widehat{\beta})$ when there is no ambiguity. It is important to remark that we will follow the idea of Escanciano (2006a) and we will not estimate the parameter γ unlike Stute and Zhu (2002) for example. We consider a process that depends both on x and γ. The test statistics that we will use are Kolmogorov-Smirnov and Cramér-von Mises type statistics defined by

$$D_n = \sup_{(x,\gamma)\in\Pi} |R_n^1(x,\gamma)| \quad \text{and} \quad W_n^2 = \int_\Pi [R_n^1(x,\gamma)]^2 \, d\Psi_n(x,\gamma) \ ,$$

where Π is a properly chosen compact space and $\Psi_n(x,\gamma)$ is a certain estimator of an arbitrary integrating function $\Psi(x,\gamma)$ that is absolutely continuous and that satisfies regularity condition (A10) given in the Appendix. The main advantage of putting a general weighting function w and an arbitrary integrating function Ψ_n in the definition of the Cramér-von Mises test statistic is that we can use the three main approaches in the literature based on the integrated regression function but in a context of a semiparametric transformation model :

1. Bierens (1982) : take $\Pi = \chi$, $w(X, x, \gamma) = \exp(ix^t X)$ where i is the imaginary unit, $\Psi_n(x, \gamma) = \Psi(x, \gamma)$ and $d\Psi(x, \gamma) = \Phi(x)dx$. Here, the function $\Phi(x)$ will be the standard d-variate normal density function, so that the imaginary part of the Cramér-von Mises test statistic is equal to 0. Using this particular function $\Phi(x)$ and doing similar calculations as in Bierens (1982), we find that :

$$W^2_{\exp_i} = n^{-1} \sum_{j=1}^{n} \sum_{k=1}^{n} e_j(\widehat{\theta}, \widehat{\beta}) e_k(\widehat{\theta}, \widehat{\beta}) \exp\left(-\frac{1}{2} \sum_{l=1}^{d} (X_{jl} + X_{kl})^2 \right) ,$$

   where $e_j(\widehat{\theta}, \widehat{\beta}) = \Lambda_{\widehat{\theta}}(Y_j) - m_{\widehat{\beta}}(X_j)$. We will use this particular expression of $W^2_{\exp_i}$ in our simulation study when the complex exponential weight will be used.

2. Stute (1997) : take $\Pi = \chi$, $w(X, x, \gamma) = 1_{\{X \le x\}}$ where $1_{\{X \le x\}}$ is a component by component indicator and $\Psi_n(x, \gamma) = \widehat{F}_X(x)$ where $\widehat{F}_X(x)$ is the empirical distribution function of the data $\{X_i\}_{i=1,\ldots,n}$. Moreover, $\Psi(x, \gamma) = F_X(x)$ is the true distribution function of X. Hence, the two test statistics take the following form :

$$D_1 = \max_{1 \le k \le n} \left| n^{-1/2} \sum_{j=1}^{n} e_j(\widehat{\theta}, \widehat{\beta}) 1_{\{X_j \le X_k\}} \right| ,$$

   and

$$W^2_1 = n^{-2} \sum_{k=1}^{n} \left(\sum_{j=1}^{n} e_j(\widehat{\theta}, \widehat{\beta}) 1_{\{X_j \le X_k\}} \right)^2 .$$

3. Escanciano (2006a) : take $d_\gamma = d$, $w(X, x, \gamma) = 1_{\{\gamma^t X \le x\}}$ and $d\Psi_n(x, \gamma) = d\widehat{F}_{n,\gamma}(x) d\gamma$ where $\widehat{F}_{n,\gamma}(x)$ is the empirical distribution function of the projected regressor $\{\gamma^t X_i\}_{i=1,\ldots,n}$ and $d\gamma$ is the uniform density on $\mathbb{S}_d$ which is the unit ball of dimension d and ensures that all directions are equally important. Then, $\Pi = [-\Delta, \Delta] \times \mathbb{S}_d$ where $\Delta = d \max_{1 \le i \le n} \sup_{t \in \chi} |t_i|$ and t_i is the i-th component of the vector $t \in \chi$. Moreover, $\Psi(x, \gamma) = F_\gamma(x)$ is the true cumulative distribution function of $\gamma^t X$. In that case, the two test statistics take the following form :

$$D_\gamma = \sup_{1 \le k \le n, \gamma \in \mathbb{S}_d} \left| n^{-1/2} \sum_{j=1}^{n} e_j(\widehat{\theta}, \widehat{\beta}) 1_{\{\gamma^t X_j \le \gamma^t X_k\}} \right| ,$$

and

$$\begin{aligned} W_\gamma^2 &= \int_\Pi \left(n^{-1/2} \sum_{j=1}^n e_j(\widehat{\theta}, \widehat{\beta}) 1_{\{\gamma^t X_j \leq u\}} \right)^2 d\widehat{F}_{n,\gamma}(u) d\gamma \\ &= n^{-1} \sum_{j=1}^n \sum_{k=1}^n e_j(\widehat{\theta}, \widehat{\beta}) e_k(\widehat{\theta}, \widehat{\beta}) \int_\Pi 1_{\{\gamma^t X_j \leq u\}} 1_{\{\gamma^t X_k \leq u\}} d\widehat{F}_{n,\gamma}(u) d\gamma \\ &= n^{-2} \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n e_j(\widehat{\theta}, \widehat{\beta}) e_k(\widehat{\theta}, \widehat{\beta}) \int_{\mathbb{S}_d} 1_{\{\gamma^t X_j \leq \gamma^t X_l\}} 1_{\{\gamma^t X_k \leq \gamma^t X_l\}} d\gamma \ . \end{aligned}$$

In practice, to compute these test statistics, we will consider a random sample $\gamma_1, \ldots, \gamma_{n_\gamma}$ from $\mathbb{S}_d$. Hence, we can approximate both test statistics respectively by

$$\widetilde{D}_\gamma = \sup_{1 \leq k \leq n, 1 \leq m \leq n_\gamma} \left| n^{-1/2} \sum_{j=1}^n e_j(\widehat{\theta}, \widehat{\beta}) 1_{\{\gamma_m^t X_j \leq \gamma_m^t X_k\}} \right| ,$$

and

$$\widetilde{W}_\gamma^2 = n^{-2} n_\gamma^{-1} \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n \sum_{m=1}^{n_\gamma} e_j(\widehat{\theta}, \widehat{\beta}) e_k(\widehat{\theta}, \widehat{\beta}) 1_{\{\gamma_m^t X_j \leq \gamma_m^t X_l\}} 1_{\{\gamma_m^t X_k \leq \gamma_m^t X_l\}} \ .$$

In the context of nonparametric regression without transformation of the response, the advantage of taking the indicator weight over the exponential weight is that it avoids the choice of an arbitrary function Ψ and the advantage of taking the exponential weight over the indicator weight of Stute (1997) is that the test is less sensitive to the dimension d. Moreover, note that the tests proposed by Stute (1997) and Escanciano (2006a) are equivalent when $d = 1$. However, in the context of goodness-of-fit in nonparametric regression without transformation, the method of Escanciano (2006a) is known to avoid the curse of dimensionality.

Note that our test can also be applied with several other weigthing functions we can find in the literature, see Stinchcombe and White (1998) and Escanciano (2007) among others. In this paper, we will also consider the three additional following approaches in our simulation study : $w(X, x, \gamma) = \exp(x^t X)$, $w(X, x, \gamma) = (1 + \exp(-x^t X))^{-1}$ and $w(X, x, \gamma) = \sin(x^t X)$ with $\Pi = \chi$, $\Psi_n(x, \gamma) = \widehat{F}_X(x)$ and $\Psi(x, \gamma) = F_X(x)$ in the three cases. The corresponding test statistics are given by :

$$D_n = \max_{1 \leq k \leq n} \left| n^{-1/2} \sum_{j=1}^n e_j(\widehat{\theta}, \widehat{\beta}) w(X_j, X_k, \gamma) \right| ,$$

and

$$W_n^2 = n^{-2} \sum_{k=1}^{n} \left(\sum_{j=1}^{n} e_j(\widehat{\theta}, \widehat{\beta}) w(X_j, X_k, \gamma) \right)^2 .$$

We will denote the three Kolmogorov-Smirnov test statistics respectively by $D_{\exp}$, $D_{1/\exp}$ and $D_{\sin}$ and the three Cramér-von Mises test statistics respectively by $W^2_{\exp}$, $W^2_{1/\exp}$ and $W^2_{\sin}$.

3 Asymptotic results

We first need to introduce the following notations :

$$\Omega = \left\{ E\left[\frac{\partial m_{\beta_0}(X)}{\partial \beta_r} \left(\frac{\partial m_{\beta_0}(X)}{\partial \beta_s} \right)^t \right] \right\}_{r,s=1,\ldots,q},$$

$$\eta_\beta(x,y) = \Omega^{-1} \frac{\partial m_\beta(x)}{\partial \beta} (\Lambda_{\theta_0}(y) - m_\beta(x)),$$

where

$$\frac{\partial m_\beta(x)}{\partial \beta} = \left(\frac{\partial m_\beta(x)}{\partial \beta_1}, \ldots, \frac{\partial m_\beta(x)}{\partial \beta_q} \right)^t$$

is a $(q \times 1)$-vector and $\beta = (\beta_1, \ldots, \beta_q)^t$. Finally, we consider $h(x, \beta) = \frac{\partial m_\beta(x)}{\partial \beta}$ and $H(x, \gamma, \beta) = E[w(X, x, \gamma)h(X, \beta)]$.

3.1 Results under H_0

To start, we introduce three theorems. The first one establishes the limiting process of R_n, the second one states that the process R_n^1 can be expressed in terms of the process R_n and a sum of iid terms up to a negligeable term, and the last one states that we can "replace" $\Psi_n(x, \gamma)$ by $\Psi(x, \gamma)$ in the definition of W_n^2 plus a negligible term. Combining these three results, we will next easily obtain the limiting process of R_n^1 and the limiting distributions of the test statistics. The assumptions under which these results are valid, as well as the proofs of these results, are given in the Appendix.

Theorem 3.1. *Assume (A1)-(A11). Then, under H_0, the process R_n converges weakly to R_∞, where R_∞ is a centered Gaussian process with covariance function given by*

$$C(x_1, \gamma_1, x_2, \gamma_2) = \sigma^2 E[w(X, x_1, \gamma_1) w(X, x_2, \gamma_2)] \ ,$$

for $(x_1,\gamma_1),(x_2,\gamma_2)\in\Pi$.

Theorem 3.2. *Assume (A1)-(A11). Then, under* H_0*,*

$$R_n^1(x,\gamma)=R_n(x,\gamma)-n^{-1/2}\sum_{i=1}^n G(x,\gamma,X_i,Y_i,\theta_0,\beta_0)+o_P(1)\ ,$$

where

$$\begin{aligned}G(x,\gamma,X,Y,\theta,\beta) &= H^t(x,\gamma,\beta)\eta_\beta(X,Y)+E[w(X,x,\gamma)(\dot\Lambda_\theta(Y))^t]g(X,Y)\\ &\quad -H^t(x,\gamma,\beta)\Omega^{-1}E\left[\frac{\partial m_\beta(X)}{\partial\beta}(\dot\Lambda_\theta(Y))^t\right]g(X,Y)\ .\end{aligned}\tag{3.1}$$

Theorem 3.3. *Assume (A1)-(A11). Then, under* H_0*,*

$$W_n^2=\int_\Pi[R_n^1(x,\gamma)]^2\,d\Psi_n(x,\gamma)=\int_\Pi[R_n^1(x,\gamma)]^2\,d\Psi(x,\gamma)+o_P(1)\ .$$

As a consequence of the previous theorems, we obtain successively the limiting process of R_n^1 and the limiting distributions of the Kolmogorov-Smirnov and Cramér-von Mises test statistics under the null hypothesis in the two following corollaries.

Corollary 3.1. *Assume (A1)-(A11). Then, under* H_0*, the process* R_n^1 *converges weakly to* R_∞^1*, where* R_∞^1 *is a centered Gaussian process with covariance function given by*

$$\begin{aligned}C_1(x_1,\gamma_1,x_2,\gamma_2) &= C(x_1,\gamma_1,x_2,\gamma_2)-E[G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)w(X,x_1,\gamma_1)\varepsilon]\\ &\quad -E[G(x_1,\gamma_1,X,Y,\theta_0,\beta_0)w(X,x_2,\gamma_2)\varepsilon]\\ &\quad +E[G(x_1,\gamma_1,X,Y,\theta_0,\beta_0)G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)]\ ,\end{aligned}$$

for $(x_1,\gamma_1),(x_2,\gamma_2)\in\Pi$.

Corollary 3.2. *Assume (A1)-(A11). Then, under* H_0*,*

$$D_n\xrightarrow{d}\sup_{(x,\gamma)\in\Pi}|R_\infty^1(x,\gamma)|\qquad \textit{and}\qquad W_n^2\xrightarrow{d}\int_\Pi[R_\infty^1(x,\gamma)]^2\,d\Psi(x,\gamma)\ .$$

3.2 Results under H_{1n}

We consider the following local alternative hypothesis in order to study the power of the test statistics :

$$H_{1n} : m(x) = m_{\beta_0}(x) + n^{-1/2} r(x) \text{ for all } x \tag{3.2}$$

for some fixed function $r \neq 0$. First, we obtain the analog of Theorem 3.2 under H_{1n}.

Theorem 3.4. *Assume (A1)-(A12). Then, under* H_{1n}*,*

$$R_n^1(x,\gamma) = R_n(x,\gamma) - n^{-1/2} \sum_{i=1}^n G(x,\gamma,X_i,Y_i,\theta_0,\widetilde{\beta}_{0n}) + b(x,\gamma) + o_P(1) \ ,$$

where $\widetilde{\beta}_{0n}$ *is a minimizer over* $\beta \in \mathcal{B}$ *of* $E[(m_\beta(X) - m(X))^2]$ *and*

$$b(x,\gamma) = -H^t(x,\gamma,\beta_0)\Omega^{-1} \int r(u) \frac{\partial m_{\beta_0}(u)}{\partial \beta} \, dF_X(u) + E[w(X,x,\gamma) r(X)] \ .$$

We remark the presence of an additional bias term b that depends on the deviation function r in comparison with Theorem 3.2. Note that this bias term is exactly the same as in the case where the transformation of the response would be known (see formula (3.3) in Stute, 1997, for the case where $w(t,x,\gamma) = 1_{\{t \leq x\}}$). In other words, the estimation of the transformation parameter θ has no impact on the asymptotic bias under H_{1n}. This is because θ is estimated based on a nonparametric estimator of $m(\cdot)$ (see formula (2.1)).

Finally, the following corollaries give respectively the limiting process of R_n^1 and the limiting distributions of the two test statistics under the local alternative.

Corollary 3.3. *Assume (A1)-(A12). Then, under* H_{1n}*, the process* R_n^1 *converges weakly to* $R_\infty^1 + b$*, where* R_∞^1 *is the same centered Gaussian process as in Corollary 3.1.*

Corollary 3.4. *Assume (A1)-(A12). Then, under* H_{1n}*,*

$$D_n \xrightarrow{d} \sup_{(x,\gamma)\in\Pi} |R_\infty^1(x,\gamma) + b(x,\gamma)| \qquad \text{and} \qquad W_n^2 \xrightarrow{d} \int_\Pi [R_\infty^1(x,\gamma) + b(x,\gamma)]^2 \, d\Psi(x,\gamma) \ .$$

Since the bias term $b(x,\gamma)$ is the same as in the case without transformation of the response, we can directly use the results that have been obtained in Stute (1997) and Escanciano (2006a), who studied the bias term when the response is not transformed and when $w(t,x,\gamma) = 1_{\{t\leq x\}}$ and $w(t,x,\gamma) = 1_{\{\gamma^t t\leq x\}}$, respectively. They showed that the tests have power against all alternatives r that are such that the projection of r onto the orthogonal complement of the functions $\int^{\cdot} \frac{\partial m_{\beta_0}}{\partial \beta_l}(x)dF_X(x)$, $l = 1,\ldots,q$ stays away from $\text{span}\{\frac{\partial m_{\beta_0}}{\partial \beta_1},\ldots,\frac{\partial m_{\beta_0}}{\partial \beta_q}\}$, which is a natural condition that can basically not be avoided. This is an important advantage of our new tests in comparison with those developed in Colling and Van Keilegom (2016). Indeed, in the latter paper an example is given of a local alternative that has no power.

In practice, we need to estimate the limiting distributions of D_n and W_n^2 obtained in Corollaries 3.2 and 3.4. However, this implies the estimation of f_ε, $\dot{f}_\varepsilon$, f'_ε, $\dot{m}$ and f_X and consequently the introduction of new bandwidths which is possible but not easy. Therefore, we prefer to use a bootstrap procedure in order to approximate the limiting distributions of D_n and W_n^2 under H_0 in practice. This bootstrap procedure is described in the next section.

4 Simulations

In this section, we perform some simulations in order to evaluate the performance of our test statistics for small samples and also to compare the results given by these tests to those given by the tests based on the error distribution function developed by Colling and Van Keilegom (2016).

The basic idea of the test developed by Colling and Van Keilegom (2016) was to compare the error distribution function estimated in a semiparametric way to the error distribution function estimated under H_0. This gave the following test statistics :

$$T_{KS} = n^{1/2} \sup_{y\in\mathbb{R}} |\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y)|$$

and

$$T_{CM} = n\int (\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y))^2\, d\widehat{F}_\varepsilon(y)\ ,$$

where $\widehat{F}_\varepsilon(y) = n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_i \leq y)$, $\widehat{\varepsilon}_i = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}(X_i)$ are the semiparametric residuals, $\widehat{F}_{\varepsilon_0}(y) = n^{-1}\sum_{i=1}^n I(\widehat{\varepsilon}_{i0} \leq y)$ and $\widehat{\varepsilon}_{i0} = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\beta}}(X_i)$ are the estimated residuals under H_0 where $\widehat{m}_{\widehat{\beta}}(X_i)$ is the local polynomial estimator of degree p of $m_{\widehat{\beta}}(X_i)$.

As the different Cramér-von Mises test statistics give in most cases similar or better results than the corresponding Kolmogorov-Smirnov test statistics, we decide in this section to compare only the results given by the Cramér-von Mises test statistics, i.e. T_{CM}, W_1^2, $W_{\exp_i}^2$, $\widetilde{W}_\gamma^2$, $W_{\exp}^2$, $W_{1/\exp}^2$ and $W_{\sin}^2$.

We will explain briefly now how we will estimate θ, β, h and g in practice. It is done similarly as in Colling and Van Keilegom (2016). For each value of θ, we obtain $h^*(\theta)$ the cross validation bandwidth estimator :

$$h^*(\theta) = \arg\min_h \sum_{i=1}^n (\Lambda_\theta(Y_i) - \widehat{m}_{-i}(X_i, \theta, h))^2 \ ,$$

where

$$\widehat{m}_{-i}(X_i, \theta, h) = \frac{\sum_{j=1, j\neq i}^n \Lambda_\theta(Y_j) k_1\left(\frac{X_j - X_i}{h}\right)}{\sum_{j=1, j\neq i}^n k_1\left(\frac{X_j - X_i}{h}\right)} \ ,$$

and $k_1(x) = k_2(x) = \frac{3}{4}(1 - x^2)\, 1_{\{|x| \leq 1\}}$ are the Epanechnikov kernel if we work in dimension $d = 1$. More generally, for $d > 1$, we use the product of d Epanechnikov kernels for the estimator of the regression function. Moreover, we estimate g by $\widehat{g}(\theta) = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_{\widehat{\varepsilon}(\theta, h^*(\theta))}$, where $\widehat{\sigma}_{\widehat{\varepsilon}(\theta, h^*(\theta))}$ is the classical estimator of the standard deviation of the error term $\widehat{\varepsilon}(\theta, h^*(\theta)) = \Lambda_\theta(Y) - \widehat{m}(X, \theta, h^*(\theta))$, where $\widehat{m}(x, \theta, h)$ denotes $\widehat{m}(x, \theta)$ constructed with a bandwidth h. Note that we have estimated g by the classical normal reference rule for kernel density estimation. Then, the optimal value of θ is given by

$$\widehat{\theta} = \arg\max_\theta l_\theta(h^*(\theta), \widehat{g}(\theta)) \ ,$$

where

$$l_\theta(h, g) = \sum_{i=1}^n \left\{ \log \widehat{f}_{\varepsilon(\theta), g}(\Lambda_\theta(Y_i) - \widehat{m}(X_i, \theta, h)) + \log \Lambda'_\theta(Y_i) \right\} \ ,$$

where $\widehat{f}_{\varepsilon(\theta), g}(y)$ denotes the estimator $\widehat{f}_{\varepsilon(\theta)}(y)$ constructed with a bandwidth g. The estimator of θ is obtained iteratively with the function *optimize* in R over the interval $[\theta_0 - 2, \theta_0 + 2]$. Finally, to estimate β, we minimize the following expression over the interval $[-20, 20]$:

$$\widehat{\beta} = \arg\min_\beta \sum_{i=1}^n (\Lambda_{\widehat{\theta}}(Y_i) - m_\beta(X_i))^2 \ .$$

The critical values of the different test statistics are obtained with the same residual bootstrap procedure as in Colling and Van Keilegom (2016). For fixed B and for $b = 1, \ldots, B$, we define the bootstrap sample (X_{ib}^*, Y_{ib}^*), $i = 1, \ldots, n$ where $X_{ib}^* = X_i$, $Y_{ib}^* = \Lambda_{\widehat{\theta}}^{-1}(m_{\widehat{\beta}}(X_{ib}^*) + \varepsilon_{ib}^*)$ are the new responses and $\varepsilon_{ib}^* = \zeta_{ib}^* + a_n \xi_{ib}$ are independent random errors where $\zeta^*_{1b}, \dots, \zeta^*_{nb}$ are bootstrap samples of the errors drawn with replacement from the empirical distribution of the zero mean residuals, a_n is some small bandwidth and $\xi_{1b}, \dots, \xi_{nb}$ are independent normally distributed random variables and independent from the original sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$. We choose here $a_n = 0.1$.

Hence, we can compute the different test statistics using the bootstrap sample (X^*_{ib}, Y^*_{ib}), $i = 1, \dots, n$ and the $(1-\alpha)$-th quantile of the distribution of each test statistic is estimated by the $[(1-\alpha)B]$-th order statistic of the corresponding test statistic obtained from these bootstrap samples. In our simulations, we take $B = 250$. We refer to Neumeyer (2009) for the consistency of this bootstrap procedure in the case where one is interested in the distribution of the estimator of the error distribution in a nonparametric location-scale model without transformation of the response.

First, we perform simulations in dimension $d = 1$. The simulated model is $\Lambda_\theta(Y_i) = \beta_1 + \beta_2 X_i + c(X_i) + \varepsilon_i$, where $\Lambda_\theta(Y)$ is the Yeo and Johnson (2000) transformation :

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

Note that the Yeo and Johnson (2000) transformation is an extension of the Box and Cox (1964) transformation that allows the response variable Y to be negative. We will consider three different values of the parameter transformation θ : $\theta_0 = 0$ which corresponds to a logarithmic transformation, $\theta_0 = 0.5$ which corresponds to a square root transformation and $\theta_0 = 1$ which corresponds to the identity. The true value of the parameter β is $\beta_0 = (\beta_{10}, \beta_{20}) = (3, 5)$. Moreover, $X_1, \dots, X_n$ are independent uniform random variables on [0,1] and $\varepsilon_1, \dots, \varepsilon_n$ are independent standard normal random variables truncated on [-3,3]. We will also consider the cases where $\varepsilon_1, \dots, \varepsilon_n$ are independent normal random variables with zero mean and standard deviation 0.5 truncated on [-3,3] and where $\varepsilon_1, \dots, \varepsilon_n$ are independent student-t random variables with degrees of freedom equal to 10. We consider the following null hypothesis :

$$H_0 : m(x) = \beta_1 + \beta_2 x \quad \text{for all } x \text{ and for some } (\beta_1, \beta_2) \in \mathbb{R}^2 .$$

We consider different deviation functions $c(x)$ from the null hypothesis : $c(x) = 2x^2$, $c(x) = 3x^2$, $c(x) = 4x^2$, $c(x) = 5x^2$, $c(x) = 2\exp(x)$, $c(x) = 3\exp(x)$, $c(x) = 4\exp(x)$, $c(x) = 5\exp(x)$, $c(x) = 0.25\sin(2\pi x)$, $c(x) = 0.5\sin(2\pi x)$, $c(x) = 0.75\sin(2\pi x)$, $c(x) = \sin(2\pi x)$.

Tables 1 to 6 show respectively the percentages of rejection obtained with the test statistics T_{CM}, W_1^2, $W^2_{\exp_i}$, $W^2_{\exp}$, $W^2_{1/\exp}$ and $W^2_{\sin}$ for 500 samples of size $n = 200$ under the null hypothesis and under the different deviations $c(x)$ we have introduced above. The nominal level is 10%. We also remind that the test statistics W_1^2 and $\widetilde{W}_\gamma^2$ are equivalent when $d = 1$, hence Tables 1 to 6 only include the results obtained with W_1^2 and not with $\widetilde{W}_\gamma^2$.

First, under H_0, we observe that the different estimations of the nominal level are globally good for all test statistics, choices of θ_0 and distributions of ε except if we use the test statistic W_1^2 when θ_0 is increasing. In that case, the estimation of the nominal level is increasingly small.

Next, under the alternative, the power is largest when $\varepsilon \sim N(0, 0.5^2)$, followed by $\varepsilon \sim N(0, 1)$ and finally by $\varepsilon \sim t_{10}$. This seems logical because we increase consequently the variance of ε when we change from one situation to another. In a similar way, we observe generally that the power is largest for $\theta_0 = 0$, followed by $\theta_0 = 0.5$ and finally by $\theta_0 = 1$. This conclusion is the same as in Colling and Van Keilegom (2016) and is logical with respect to the results obtained by Linton, Sperlich and Van Keilegom (2008).

Moreover, again under H_1, when the deviation from the null hypothesis is monotone, for example $c(x) = cx^2$ and $c(x) = c\exp(x)$, the highest power is generally obtained with the test statistics $W^2_{\exp}$, $W^2_{1/\exp}$ and $W^2_{\sin}$. This last conclusion is valid for all tested values of θ_0. On the other hand, if the deviation from the null hypothesis is non monotone, for example $c(x) = c\sin(2\pi x)$, the highest power is obtained with the test statistic $W^2_{\exp_i}$ when the deviation is close to the null hypothesis ($c = 0.25$ and sometimes $c = 0.5$) and is obtained either with the test statistic T_{CM} or with the test statistic W_1^2 when the deviation is less close to the null hypothesis ($c = 0.75$ and $c = 1$). This depends on the value of θ_0. If $\theta_0 = 0$, we will prefer W_1^2 and when θ_0 increases, we will prefer T_{CM} which is in line with what happens under H_0. Finally, note that the test statistics $W^2_{\exp}$, $W^2_{1/\exp}$ and $W^2_{\sin}$ give very small power when the deviation from H_0 is non monotone.

	$\theta_0 = 0$						$\theta_0 = 0.5$					
$c(x)$	T_{CM}	W_1^2	$W^2_{\exp_i}$	$W^2_{\exp}$	$W^2_{1/\exp}$	$W^2_{\sin}$	T_{CM}	W_1^2	$W^2_{\exp_i}$	$W^2_{\exp}$	$W^2_{1/\exp}$	$W^2_{\sin}$
0	10.2	10.2	13.4	12.2	11.4	12.4	10.4	7.2	12.4	12.8	11.4	12.4
$2x^2$	27.2	46.6	37.6	50.0	50.6	51.2	27.2	41.2	34.6	50.4	48.8	49.2
$3x^2$	48.8	75.4	68.0	78.8	79.0	78.6	49.2	72.0	62.6	78.8	78.2	78.8
$4x^2$	66.2	89.2	86.4	91.6	91.0	91.0	66.0	88.2	82.8	91.4	90.8	91.0
$5x^2$	78.2	94.4	91.8	94.4	94.4	94.6	77.4	95.2	91.2	97.0	96.4	96.4
$2\exp(x)$	23.6	33.2	26.4	38.2	39.2	38.8	22.8	22.8	19.8	32.2	33.0	31.6
$3\exp(x)$	34.2	55.0	42.0	60.8	60.2	60.0	34.6	36.2	27.4	52.8	53.6	53.0
$4\exp(x)$	49.0	69.6	60.0	71.6	73.2	72.2	47.0	52.4	36.4	64.8	67.6	66.0
$5\exp(x)$	62.8	78.2	72.2	81.0	81.2	80.8	61.2	63.4	46.6	72.6	74.4	73.2
$0.25\sin(2\pi x)$	21.4	21.0	29.4	12.4	13.8	13.4	24.8	13.8	28.2	12.0	13.2	12.4
$0.5\sin(2\pi x)$	56.2	59.0	56.4	13.4	16.8	15.6	54.8	45.0	56.6	12.8	14.8	13.6
$0.75\sin(2\pi x)$	86.0	92.4	70.0	16.0	20.4	20.6	84.4	82.8	69.8	12.8	16.4	17.8
$\sin(2\pi x)$	98.4	99.8	76.8	17.8	26.0	27.6	98.0	94.0	76.2	13.8	21.2	23.4

Table 1: Percentage of rejection for $\theta_0 = 0$, 0.5 and for $\varepsilon \sim N(0,1)$.

	$\theta_0 = 1$					
$c(x)$	T_{CM}	W_1^2	$W^2_{\exp_i}$	$W^2_{\exp}$	$W^2_{1/\exp}$	$W^2_{\sin}$
0	9.8	5.0	10.8	10.6	9.6	10.4
$2x^2$	29.2	38.6	32.0	49.2	47.8	48.2
$3x^2$	47.6	67.8	57.8	77.4	78.8	78.4
$4x^2$	65.8	86.8	82.0	90.4	89.8	89.8
$5x^2$	76.8	93.2	87.8	96.4	96.2	95.8
$2\exp(x)$	22.0	17.8	16.4	29.4	30.0	29.4
$3\exp(x)$	34.4	30.8	20.8	47.0	47.6	45.6
$4\exp(x)$	47.6	44.6	30.0	61.8	62.4	61.6
$5\exp(x)$	61.6	56.0	36.4	67.6	69.2	68.8
$0.25\sin(2\pi x)$	22.8	10.6	27.8	10.4	10.6	11.2
$0.5\sin(2\pi x)$	55.4	34.0	55.4	10.6	12.8	11.8
$0.75\sin(2\pi x)$	84.0	69.2	68.8	11.2	14.0	14.8
$\sin(2\pi x)$	98.2	87.8	75.6	11.6	15.8	18.0

Table 2: Percentage of rejection for $\theta_0 = 1$ and for $\varepsilon \sim N(0,1)$.

	$\theta_0 = 0$						$\theta_0 = 0.5$					
$c(x)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	11.0	10.8	11.8	12.4	12.0	12.6	12.2	6.4	10.4	11.2	10.2	10.4
$2x^2$	42.4	63.6	52.2	66.4	66.8	66.6	43.0	55.4	44.8	64.0	64.4	64.2
$3x^2$	65.0	85.4	79.4	87.8	87.4	87.0	63.2	80.0	72.6	86.0	86.0	86.0
$4x^2$	83.4	95.6	91.2	96.2	95.4	95.4	82.4	92.8	87.2	94.8	94.2	94.4
$5x^2$	92.4	98.2	96.2	98.2	98.2	98.2	91.8	96.8	93.2	97.4	97.2	97.2
$2\exp(x)$	32.4	46.6	35.2	51.2	51.2	51.0	32.8	30.8	23.2	40.4	40.4	40.8
$3\exp(x)$	47.6	63.2	52.0	67.6	68.2	67.8	47.4	48.2	31.6	57.8	58.8	58.2
$4\exp(x)$	64.8	76.4	65.2	78.6	79.0	78.8	62.2	57.8	39.0	66.2	68.0	67.0
$5\exp(x)$	74.4	83.2	73.6	85.0	85.4	85.0	73.6	67.8	46.0	74.4	74.8	74.8
$0.25\sin(2\pi x)$	36.0	32.8	42.2	14.0	14.2	13.8	35.4	20.4	38.0	13.0	14.2	14.2
$0.5\sin(2\pi x)$	81.8	88.8	65.6	16.2	24.4	22.8	83.4	75.6	62.4	14.8	18.0	18.0
$0.75\sin(2\pi x)$	98.2	99.8	78.2	21.0	29.6	29.4	98.6	96.2	76.4	17.0	23.4	23.4
$\sin(2\pi x)$	99.8	99.8	83.4	24.4	34.6	36.6	100.0	99.4	80.2	19.0	27.0	29.4

Table 3: Percentage of rejection for $\theta_0 = 0$, 0.5 and for $\varepsilon \sim N(0, 0.5^2)$.

	$\theta_0 = 1$					
$c(x)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	12.2	4.4	8.6	8.8	8.2	9.0
$2x^2$	41.2	49.2	39.0	62.2	61.4	60.8
$3x^2$	63.8	80.8	68.2	85.6	86.2	86.2
$4x^2$	82.0	92.4	84.8	93.8	93.2	93.8
$5x^2$	91.6	96.8	92.8	97.2	96.6	96.6
$2\exp(x)$	31.2	27.4	17.6	37.6	38.6	37.8
$3\exp(x)$	47.2	38.2	24.0	51.6	53.0	52.2
$4\exp(x)$	62.8	51.0	35.0	60.2	61.4	60.6
$5\exp(x)$	73.4	61.4	36.4	69.2	69.8	69.0
$0.25\sin(2\pi x)$	34.4	12.0	38.8	9.8	11.4	11.6
$0.5\sin(2\pi x)$	82.0	57.4	61.2	12.0	15.6	16.0
$0.75\sin(2\pi x)$	98.6	90.6	75.0	14.4	20.4	20.6
$\sin(2\pi x)$	100.0	96.8	78.6	15.2	22.0	23.6

Table 4: Percentage of rejection for $\theta_0 = 1$ and for $\varepsilon \sim N(0, 0.5^2)$.

	$\theta_0 = 0$						$\theta_0 = 0.5$					
$c(x)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	11.0	10.6	11.4	12.6	12.2	12.8	11.2	7.4	10.4	13.0	12.2	12.8
$2x^2$	24.4	34.4	24.4	38.0	37.4	37.0	28.8	38.2	29.2	46.8	45.8	45.2
$3x^2$	36.2	54.8	44.8	61.2	59.0	58.8	42.4	60.8	48.8	69.2	68.6	68.8
$4x^2$	54.8	75.0	65.4	79.4	78.4	78.6	59.8	78.4	70.2	84.8	84.8	84.8
$5x^2$	71.0	87.4	80.4	90.2	89.6	90.0	74.2	88.8	81.4	91.2	91.4	91.4
$2\exp(x)$	21.4	30.6	22.2	34.0	35.0	34.8	23.0	18.6	16.0	28.6	30.4	29.6
$3\exp(x)$	31.2	46.4	34.0	50.4	51.4	50.4	31.4	30.2	22.4	41.8	42.6	42.0
$4\exp(x)$	44.6	60.0	49.4	63.0	64.4	63.4	43.6	40.6	27.8	51.8	52.0	50.8
$5\exp(x)$	54.4	66.8	55.0	68.4	68.8	68.6	54.6	49.4	31.8	58.4	59.2	58.4
$0.25\sin(2\pi x)$	18.6	17.8	25.2	13.2	12.6	12.6	18.6	12.8	24.6	12.6	12.2	12.8
$0.5\sin(2\pi x)$	46.6	47.2	52.0	13.4	15.0	15.0	47.6	32.4	51.2	13.0	15.2	15.6
$0.75\sin(2\pi x)$	81.6	85.6	67.4	16.2	18.4	18.8	81.0	64.2	64.8	15.2	17.4	15.4
$\sin(2\pi x)$	96.2	97.4	75.0	17.8	23.4	24.4	95.8	83.4	71.4	13.4	19.0	20.4

Table 5: Percentage of rejection for $\theta_0 = 0$, 0.5 and for $\varepsilon \sim t_{10}$.

	$\theta_0 = 1$					
$c(x)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	11.0	5.6	9.0	11.6	10.6	11.2
$2x^2$	26.0	34.6	30.0	43.4	42.8	43.6
$3x^2$	40.8	55.4	47.6	64.8	64.6	65.0
$4x^2$	55.0	74.8	65.4	81.8	81.6	82.0
$5x^2$	74.6	86.6	80.2	90.6	89.8	90.2
$2\exp(x)$	22.4	13.4	14.4	27.2	27.4	26.8
$3\exp(x)$	32.0	23.0	17.8	37.6	38.6	37.8
$4\exp(x)$	43.0	32.0	23.0	45.2	46.6	45.8
$5\exp(x)$	55.0	41.8	27.0	54.2	55.6	54.2
$0.25\sin(2\pi x)$	19.0	9.4	25.2	12.6	10.8	11.4
$0.5\sin(2\pi x)$	47.2	24.2	51.6	11.8	13.0	12.6
$0.75\sin(2\pi x)$	80.8	52.8	66.0	11.0	14.2	13.8
$\sin(2\pi x)$	95.4	68.6	70.4	11.6	16.6	17.2

Table 6: Percentage of rejection for $\theta_0 = 1$ and for $\varepsilon \sim t_{10}$.

Finally, we perform simulations in dimension $d = 2$. The simulated model is $\Lambda_\theta(Y_i) = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + c(X_{1i}, X_{2i}) + \varepsilon_i$, where $\Lambda_\theta(Y)$ is again the Yeo and Johnson (2000) transformation. We will consider the same three different values of the parameter transformation θ : $\theta_0 = 0$, $\theta_0 = 0.5$ and $\theta_0 = 1$. The true value of the parameter β is $\beta_0 = (\beta_{10}, \beta_{20}, \beta_{30}) = (3, 2, 1)$. Moreover, $X_{11}, \ldots, X_{1n}$ and $X_{21}, \ldots, X_{2n}$ are independent and uniformly distributed on the unit square and $\varepsilon_1, \ldots, \varepsilon_n$ are independent normal random variables with zero mean and standard deviation 0.5 truncated on [-3,3]. We consider the following null hypothesis :

$$H_0 : m(x) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 \quad \text{for all } x \text{ and for some } (\beta_1, \beta_2, \beta_3) \in \mathbb{R}^3 .$$

We consider different deviations $c(x_1, x_2)$ from the null hypothesis : $c(x_1, x_2) = 2x_1x_2$, $c(x_1, x_2) = 3x_1x_2$, $c(x_1, x_2) = 4x_1x_2$, $c(x_1, x_2) = 5x_1x_2$, $c(x_1, x_2) = 0.5x_1 \sin(2\pi x_2)$, $c(x_1, x_2) = x_1 \sin(2\pi x_2)$, $c(x_1, x_2) = 1.5x_1 \sin(2\pi x_2)$, $c(x_1, x_2) = 2x_1 \sin(2\pi x_2)$, for 300 samples of size $n = 300$. For the estimation of θ, β, h and g and the bootstrap procedure, we proceed exactly as described before. Table 7 shows the percentage of rejection obtained with the test statistics T_{CM}, W_1^2, $W_{\exp_i}^2$, $\widetilde{W}_\gamma^2$, $W_{\exp}^2$, $W_{1/\exp}^2$ and $W_{\sin}^2$ under the null hypothesis and under the different deviations $c(x_1, x_2)$ we have introduced above. Note that the results given by W_1^2 and $\widetilde{W}_\gamma^2$ will be different here since $d \neq 1$.

Table 7 shows that the estimations of the nominal level are generally too low, especially for $\theta_0 = 0$ when we consider the different new test statistics developed in this paper. This problem was already encountered in Colling and Van Keilegom (2016) and is due to the poor nonparametric 2-dimensional estimation of the function $m(\cdot)$. This suggests that the method suffers from curse-of-dimensionality problems, implying that samples of size $n = 300$ are not always large enough.

Next, under the alternative, the highest power is obtained with the test statistic $W_{\exp_i}^2$ when the deviation from the null hypothesis is monotone, for example $c(x_1, x_2) = cx_1x_2$, and is obtained with the test statistic T_{CM} when the deviation from the null hypothesis is non monotone, for example $c(x_1, x_2) = cx_1 \sin(2\pi x_2)$. We can also observe that the test statistics W_1^2 and $\widetilde{W}_\gamma^2$ give very poor power, even if it is a little bit better when the deviation is non monotone. This suggests that it is not a good idea to use indicator weights when the dimension d of X is increasing. Finally, the test statistics $W_{\exp}^2$, $W_{1/\exp}^2$ and $W_{\sin}^2$ give a higher power than the one obtained with the indicator weight, but smaller than the one obtained with T_{CM} and $W_{\exp_i}^2$ with a monotone deviation.

	$\theta_0 = 0$						
$c(x_1, x_2)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$\widetilde{W}_\gamma^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	10.0	2.3	3.7	2.0	2.3	2.3	2.3
$2x_1x_2$	9.7	2.0	16.7	1.3	9.3	8.3	9.3
$3x_1x_2$	17.3	2.3	33.3	2.0	13.7	13.7	14.0
$4x_1x_2$	28.3	3.0	55.3	1.7	17.7	18.0	18.3
$5x_1x_2$	31.0	4.0	69.7	3.0	26.3	26.3	26.3
$0.5x_1 \sin(2\pi x_2)$	33.0	5.0	8.7	6.3	7.3	7.3	7.7
$x_1 \sin(2\pi x_2)$	68.0	14.3	17.3	19.7	26.0	25.3	26.0
$1.5x_1 \sin(2\pi x_2)$	88.3	16.0	16.7	24.0	38.3	37.0	39.0
$2x_1 \sin(2\pi x_2)$	90.0	17.3	17.7	27.7	44.0	41.3	45.3
	$\theta_0 = 0.5$						
$c(x_1, x_2)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$\widetilde{W}_\gamma^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	7.0	7.7	10.7	9.0	9.0	9.3	9.0
$2x_1x_2$	6.3	6.7	21.7	6.3	15.3	15.7	16.0
$3x_1x_2$	13.0	4.7	38.3	5.0	16.7	16.7	16.7
$4x_1x_2$	18.7	8.7	59.0	6.7	21.3	21.0	21.0
$5x_1x_2$	23.0	9.7	73.7	7.0	20.7	22.0	20.3
$0.5x_1 \sin(2\pi x_2)$	19.0	14.3	19.3	16.7	18.0	18.7	18.3
$x_1 \sin(2\pi x_2)$	51.7	23.3	26.3	28.0	37.3	38.0	37.3
$1.5x_1 \sin(2\pi x_2)$	82.3	22.0	26.3	26.7	46.3	47.3	47.3
$2x_1 \sin(2\pi x_2)$	91.0	25.0	23.3	32.3	52.3	50.7	53.7
	$\theta_0 = 1$						
$c(x_1, x_2)$	T_{CM}	W_1^2	$W_{\exp_i}^2$	$\widetilde{W}_\gamma^2$	$W_{\exp}^2$	$W_{1/\exp}^2$	$W_{\sin}^2$
0	8.7	8.7	8.7	10.3	7.3	7.0	7.7
$2x_1x_2$	5.3	5.0	20.0	4.0	15.0	15.0	15.3
$3x_1x_2$	10.3	4.3	30.7	4.0	18.3	19.7	19.0
$4x_1x_2$	19.3	9.7	51.0	8.3	22.3	23.3	23.0
$5x_1x_2$	22.0	10.3	67.7	9.0	23.0	24.0	24.0
$0.5x_1 \sin(2\pi x_2)$	30.0	12.0	14.3	15.0	16.0	16.0	16.7
$x_1 \sin(2\pi x_2)$	65.7	19.3	24.0	22.3	32.7	34.3	33.3
$1.5x_1 \sin(2\pi x_2)$	86.3	20.7	22.3	24.3	41.7	41.0	42.7
$2x_1 \sin(2\pi x_2)$	92.3	19.3	18.7	25.3	42.7	42.3	44.3

Table 7: Percentage of rejection for $\theta_0 = 0$, 0.5, 1 and for $\varepsilon \sim N(0, 0.5^2)$.

5 Application

We apply our testing procedure to a ultrasonic calibration data set composed of 214 observations. The data can be found on the website http://www.itl.nist.gov/div898/handbook/pmd /section6/pmd631.htm and comes from the NIST/SEMATECH e-Handbook of Statistical Methods. The response Y is ultrasonic response and the covariate X is metal distance.

This data set has already been analyzed in the e-Handbook and in Neumeyer, Noh and Van Keilegom (2016). In the e-Handbook, we can find that the data satisfy the model $\sqrt{Y_i} = m(X_i) + \varepsilon_i$, $i = 1, \ldots, n$. The goal of Neumeyer, Noh and Van Keilegom (2016) was to verify such validity with their own procedure and without using the square root transformation. They estimated the transformation and they concluded that this data set satisfies the assumption of a homoscedastic transformation model using a Box-Cox transformation.

In the e-Handbook, we can find that $\widehat{\sqrt{Y}} = \frac{\exp(-0.015X)}{0.0807+0.0639X}$ is the fitted transformed model. In this paper, we consider a Box-Cox transformation of the response variable. The estimated transformation parameter is equal to $\widehat{\theta} = 0.43$. Note that this transformation is very similar to the square root transformation of the e-Handbook. We will check the following natural goodness-of-fit :

$$H_0 : m(x) = \frac{\exp(\beta_1 x)}{\beta_2 + \beta_3 x} \text{ for all } x \quad \text{(test 1)} \ .$$

We will also check the following exponential, inverse linear, linear and quadratic goodness-of-fits :

- $H_0 : m(x) = \beta_1 + \beta_2 \exp(\beta_3 x)$ for all x (test 2) ,
- $H_0 : m(x) = \frac{\beta_1}{\beta_2+\beta_3 x}$ for all x (test 3) ,
- $H_0 : m(x) = \beta_1 + \beta_2 x$ for all x (test 4) ,
- $H_0 : m(x) = \beta_1 + \beta_2 x + \beta_3 x^2$ for all x (test 5) .

We use the Cramér-von Mises test statistic defined in Colling and Van Keilegom (2016) and the ones in this paper. The distributions and p-values of these test statistics are approximated by the bootstrap on the basis of 1000 replicates. The results are given in Table 8.

Table 8 indicates that there is no evidence against the fitted transformed model introduced in the e-Handbook when $\alpha = 0.01$ whereas only the test statistic W^2_{exp} rejects

	test 1	test 2	test 3	test 4	test 5
T_{CM}	0.375	0.525	0.090	0	0
W_1^2	0.271	0.093	0.017	0.001	0.004
$W_{\exp_i}^2$	0.130	0.103	0.033	0	0.004
$W_{\exp}^2$	0.024	0.835	0.291	0.002	0.011
$W_{1/\exp}^2$	0.176	0.085	0.028	0	0.004
$W_{\sin}^2$	0.379	0.067	0.036	0	0.004

Table 8: p-values of the different goodness-of-fit tests for the ultrasonic calibration data.

this model when $\alpha = 0.05$. Note that, in this case, our estimations of β_i, $i = 1, 2, 3$ are $\widehat{\beta}_1 = -0.0569$, $\widehat{\beta}_2 = 0.0568$ and $\widehat{\beta}_3 = 0.0367$, the small differences with respect to the e-Handbook are due to the use of a Box-Cox transformation with $\widehat{\theta} = 0.43$ instead of the square root transformation. Next, we observe that there is also no evidence against the exponential model introduced in test 2 when $\alpha = 0.05$. Finally, for most test statistics, the model introduced in test 3 is rejected as well as the linear and the quadratic fits of tests 4 and 5.

6 Conclusions

In this paper, we used the integrated approach to construct a new test for the parametric form of the regression function in a semiparametric transformation model. The main idea of our test was to compare the integrated regression function estimated in a semiparametric way to the one estimated under H_0. We defined a Kolmogorov-Smirnov and a Cramér-von Mises test statistic, both based on an empirical residual process depending on a general weighting function that allowed us to apply several approaches we found in the literature. We established the limiting distribution of these two test statistics under the null hypothesis and under a local alternative and we noticed that our tests have power against all local alternatives that converge to the null model at parametric rate. We compared the performance of this new test to the previous test developed by Colling and Van Keilegom (2016) by means of a large simulation study. Finally, we applied our test on a real data set.

7 Appendix: Proofs

The Appendix is structured as follows. In Subsection 7.1, we introduce a number of notations and we state the different assumptions under which the main results of this paper are valid. Note that these assumptions are in majority the same as in Colling and Van Keilegom (2016). Then, in Subsections 7.2, 7.3, we prove the main results of the asymptotic theory under the null hypothesis and under the local alternative respectively.

7.1 Notations and technical assumptions

For $0 < \alpha < \delta/2 < 1$, where δ is defined as in assumption $(A2)$ (see below), let $C_1^{d+\alpha}(\chi)$ be the set of d-times differentiable functions $f : \chi \to \mathbb{R}$ such that :

$$||f||_{d+\alpha} := \max_{j.\leq d} \sup_{x\in\chi} |D^j f(x)| + \max_{j.=d} \sup_{x,x'\in\chi} \frac{|D^j f(x) - D^j f(x')|}{||x - x'||^\alpha} \leq 1 \ ,$$

where $j = (j_1, \dots, j_d)$, $j. = \sum_{i=1}^d j_i$, $D^j = \frac{\partial^{j.}}{\partial x_1^{j_1} \dots \partial x_d^{j_d}}$ and $||.||$ is the Euclidean norm on $\mathbb{R}^d$.

The main results of the asymptotic theory require the following regularity conditions on the kernels, the bandwidths, the distributions of X and ε, the transformation Λ_θ, the weihting function w, the integrating functions Ψ_n and Ψ and the functions $m_\beta(x)$, $m(x)$ and $r(x)$:

(A1) The functions k_j $(j = 1, 2)$ are symmetric, have support [-1,1], $\int k_1(u)\,du = 1$, $\int u^k k_2(u)\,du = 0$ for $k = 1, \dots, q_2 - 1$ and $\int u^{q_2} k_2(u)\,du \neq 0$ for some $q_2 \geq 4$. Moreover, k_1 is d-times continuously differentiable, $k_1^{(l)}(\pm 1) = 0$ for $l = 0, \dots, d-1$ and k_2 is twice continuously differentiable.

(A2) h_l (for $l = 1, \dots, d$) satisfies $h_l/h \to c_l$ for some $0 < c_l < \infty$ and the bandwidths h and g satisfy $nh^{2p+2} \to 0$ for some $p \geq 3$, $nh^{3d+\delta} \to \infty$ for some $\delta > 0$, $ng^6(\ln g^{-1})^{-2} \to \infty$ and $ng^{2q_2} \to 0$ when $n \to \infty$, where q_2 is defined in condition $(A1)$.

(A3) (i) The support χ of the covariate X is a compact subset of $\mathbb{R}^d$.

(ii) The distribution function F_X is $2d + 1$-times continuously differentiable.

(iii) $\inf_{x\in\chi} f_X(x) > 0$.

(A4) The distribution function $F_{\varepsilon(\theta)|X}(y|x)$ is three times continuously differentiable with respect to y and θ, and

$$\sup_{\theta,y,x}\left|\frac{\partial^{i+j}}{\partial y^i \partial\theta_1^{j_1}\dots\partial\theta_k^{j_k}}F_{\varepsilon(\theta)|X}(y|x)\right| < \infty$$

for all i and j such that $0 \le i+j \le 2$ where $j = \sum_{l=1}^k j_l$.

(A5) (i) The transformation $\Lambda_\theta(y)$ is three times continuously differentiable with respect to both y and θ, and there exists $\alpha > 0$ such that :

$$E\left[\sup_{\theta':||\theta'-\theta||\le\alpha}\left|\left|\frac{\partial^{i+j}}{\partial y^i\partial\theta^j}\Lambda_{\theta'}(Y)\right|\right|\right] < \infty$$

for all $\theta \in \Theta$ and all i and j such that $0 \le i+j \le 3$.

(ii) $\sup_{x\in\chi} E[(\dot\Lambda_{\theta_0}(Y))_l^2|X = x] < \infty$.

(A6) (i) $\mathcal{B}$ is a compact subset of $\mathbb{R}^q$ and β_0 is an interior point of $\mathcal{B}$.

(ii) All partial derivatives of $m_\beta(x)$ with respect to the components of x and β of order 0, 1, 2 and 3 exist and are continuous in (x,β) for all x and β.

(iii) For all $\varepsilon > 0$:

$$\inf_{||\beta-\beta_0||>\varepsilon} E[(m_\beta(X) - m_{\beta_0}(X))^2] > 0 \ .$$

(iv) Ω is non singular.

(A7) The functions $m(x)$ and $\frac{\partial}{\partial\theta}m(x,\theta) := \dot m(x)$ are $p+2$ times continuously differentiable with respect to the components of x on $\chi \times N(\theta_0)$, where $N(\theta_0)$ is a neighbourhood of θ_0 and all derivatives up to order $p+2$ are bounded, uniformly in (x,θ) in $\chi \times N(\theta_0)$.

(A8) (i) For all $\eta > 0$, there exists $\varepsilon(\eta) > 0$ such that

$$\inf_{||\theta-\theta_0||>\eta} ||E(\xi(\theta, X, Y))|| \ge \varepsilon(\eta) > 0 \ .$$

(ii) The matrix Γ is of full rank.

(A9) (i) The class of functions $\mathcal{W} = \{u \to w(u, x, \gamma), (x, \gamma) \in \Pi\}$ is Donsker.

(ii) The weighting function w satisfies $\sup_{(x,\gamma)\in\Pi,t\in\chi} |w(t, x, \gamma)| < \infty$.

(iii) The weighting function w satisfies $\sup_{t\in\chi} V[w(t,\cdot,\cdot)] < \infty$, where $V[w(t,\cdot,\cdot)]$ is the total variation of $w(t,\cdot,\cdot)$ defined on the compact set Π.

(A10) The integrating functions $\Psi_n(x,\gamma)$ and $\Psi(x,\gamma)$ satisfy the two following regularity conditions : there exists a random function $\widetilde{\Psi}$ such that

$$\sup_{(x,\gamma)\in\Pi} \left|\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)\right| = o_P(n^{-1/2}) \ ,$$

and such that

$$\sup_{(x,\gamma)\in\Pi} \left|\widetilde{\Phi}(x,\gamma) - \Phi(x,\gamma)\right| = o_P(1) \ ,$$

where $\Phi(x,\gamma) = \frac{\partial^{d+d_\gamma}}{\partial x_1 \ldots \partial x_d \partial\gamma_1 \ldots \partial\gamma_{d_\gamma}} \Psi(x,\gamma)$ and $\widetilde{\Phi}(x,\gamma) = \frac{\partial^{d+d_\gamma}}{\partial x_1 \ldots \partial x_d \partial\gamma_1 \ldots \partial\gamma_{d_\gamma}} \widetilde{\Psi}(x,\gamma)$ if $d_\gamma \neq 0$ and $\Phi(x,\gamma) = \frac{\partial^{d}}{\partial x_1 \ldots \partial x_d} \Psi(x,\gamma)$ and $\widetilde{\Phi}(x,\gamma) = \frac{\partial^{d}}{\partial x_1 \ldots \partial x_d} \widetilde{\Psi}(x,\gamma)$ if $d_\gamma = 0$.

(A11) $\Lambda_{\theta_0}(\alpha) = a$ and $\Lambda_{\theta_0}(\beta) = b$ for some $\alpha < \beta$ and $a < b$, and the set $\{x \in \chi : \frac{\partial}{\partial x} m(x) \neq 0\}$ has nonempty interior.

(A12) $E(r^2(X)) < \infty$ and $r(x)$ is twice continuously differentiable for all x.

Note that conditions (A1) and (A2), which are assumptions on the different kernels and bandwidths and condition (A7) come partially from Linton, Sperlich and Van Keilegom (2008), partially from Neumeyer and Van Keilegom (2010) and partially from Colling and Van Keilegom (2016). Moreover, condition (A3)(ii) comes from Neumeyer and Van Keilegom (2010), conditions (A4), (A5) and (A8) come from Linton, Sperlich and Van Keilegom (2008) and conditions (A6) and (A12) come from Van Keilegom, González-Manteiga and Sánchez Sellero (2008) and Colling and Van Keilegom (2016). Condition (A11) is needed for identifying the model, see Vanhems and Van Keilegom (2016) and Colling and Van Keilegom (2016). Finally, conditions (A9) and (A10) are new conditions on the weighting function w and the integrating functions Ψ_n and Ψ respectively.

We will prove now that conditions (A9)(i) and (A10) are satisfied for the different weighting functions w and the corresponding integrating functions Ψ_n and Ψ that we have used in the simulations and in the application.

Proposition 7.1. *Condition (A9)(i) is satisfied for the weighting functions* $w(X,x,\gamma) = 1_{\{X\leq x\}}$, $w(X,x,\gamma) = 1_{\{\gamma^t X\leq x\}}$, $w(X,x,\gamma) = \exp(ix^tX)$, $w(X,x,\gamma) = \sin(x^tX)$, $w(X,x,\gamma) = \exp(x^tX)$ *and* $w(X,x,\gamma) = (1+\exp(-x^tX))^{-1}$.

Proof. Consider first $w(X, x, \gamma) = \sin(x^t X)$, $w(X, x, \gamma) = \exp(x^t X)$ and $w(X, x, \gamma) = (1 + \exp(-x^t X))^{-1}$. These three weighting functions are infinitely differentiable and all derivatives are uniformly bounded on the compact set χ. Hence, applying Corollary 2.7.2 in Van der Vaart and Wellner (1996), we get that $\log N_{[\,]}(\widetilde{\varepsilon}, \mathcal{W}, L_2(P)) \leq C\widetilde{\varepsilon}^{-d/\kappa}$, where C is some positive constant, κ is the smoothness of the class which can be taken arbitrarily large, $N_{[\,]}(\widetilde{\varepsilon}, \mathcal{W}, L_2(P))$ is the $\widetilde{\varepsilon}$-bracketing number of the class $\mathcal{W}$, P is the probability measure corresponding to the distribution of X, and $L_2(P)$ is the L_2-norm. This gives

$$\begin{aligned}
\int_0^{+\infty} \sqrt{\log N_{[\,]}(\widetilde{\varepsilon}, \mathcal{W}, L_2(P))}\, d\widetilde{\varepsilon} &= \int_0^{2T} \sqrt{\log N_{[\,]}(\widetilde{\varepsilon}, \mathcal{W}, L_2(P))}\, d\widetilde{\varepsilon} \\
&\leq \int_0^{2T} \sqrt{C\widetilde{\varepsilon}^{-d/\kappa}}\, d\widetilde{\varepsilon} \\
&= \frac{\sqrt{C}(2T)^{\frac{-d}{2\kappa}+1}}{\frac{-d}{2\kappa}+1} \\
&< \infty\ ,
\end{aligned}$$

provided $\kappa > d/2$, and where T is a uniform upper bound for the above three w-functions. We obtained the first equality since only one $\widetilde{\varepsilon}$-bracket suffices to cover $\mathcal{W}$ if $\widetilde{\varepsilon} > 2T$. Finally, it suffices to apply Theorem 2.5.6 in Van der Vaart and Wellner (1996) to get that $\mathcal{W}$ is Donsker. Next, if $w(X, x, \gamma) = \exp(ix^t X)$, the proof is exactly the same, taking into account that $\exp(ix^t X) = \cos(x^t X) + i\sin(x^t X)$.

Second, consider $w(X, x, \gamma) = 1_{\{X \leq x\}}$. This function is not differentiable, so we can not apply the same proof as above. However, in this case, we can apply Example 2.5.4 in Van der Vaart and Wellner (1996), which states that the set of all indicator functions of type $1_{\{X \leq x\}}$ in $\mathbb{R}^d$ is Donsker for any dimension d.

Finally, the proof for $w(X, x, \gamma) = 1_{\{\gamma^t X \leq x\}}$ is similar to the proof of Lemma 1 in Akritas and Van Keilegom (2001). We refer to their paper for more details. □

Proposition 7.2. *Condition (A10) is satisfied for the three following cases :*

- *(a)* $\Psi_n(x, \gamma) = \Psi(x, \gamma)$*, when the weighting function used in Bierens (1982) is considered.*
- *(b)* $\Psi_n(x, \gamma) = \widehat{F}_X(x)$ *and* $\Psi(x, \gamma) = F_X(x)$*, when the weighting function used in Stute (1997) is considered and also used when* $w(X, x, \gamma) = \sin(x^t X)$*,* $w(X, x, \gamma) = \exp(x^t X)$ *and* $w(X, x, \gamma) = (1 + \exp(-x^t X))^{-1}$.

(c) $d\Psi_n(x,\gamma) = d\widehat{F}_{n,\gamma}(x)d\gamma$ *and* $\Psi(x,\gamma) = F_\gamma(x)$*, when the weighting function used in Escanciano (2006a) is considered .*

Proof. First, the result is trivial in case (a), since we can take $\widetilde{\Psi} = \Psi_n = \Psi$. Second, consider case (b). In that case, we define

$$\widetilde{\Psi}(x) = \widetilde{F}_X(x) = b_n^{-d} \int \widehat{F}_X(t) L\Big(\frac{x-t}{b_n}\Big)\, dt \ ,$$

where $t = (t_1, \ldots, t_d)$, $dt = (dt_1, \ldots, dt_d)$, $\frac{x-t}{b_n} = (\frac{x_1-t_1}{b_n}, \ldots, \frac{x_d-t_d}{b_n})$, L is a kernel of order $\tau > \frac{d}{2}$ and b_n is a bandwidth such that $nb_n^d(\log n)^{-1} \to \infty$ and $nb_n^{2\tau} \to 0$. By definition of the kernel L, we know that $b_n^{-d}\int L(\frac{x-t}{b_n})\, dt = 1$. Consequently $\widehat{F}_X(x) = b_n^{-d}\int \widehat{F}_X(x) L(\frac{x-t}{b_n})\, dt$ and

$$\begin{aligned}
\widetilde{F}_X(x) - F_X(x) &= b_n^{-d} \int \widehat{F}_X(t) L\Big(\frac{x-t}{b_n}\Big)\, dt - F_X(x) \\
&= b_n^{-d} \int (\widehat{F}_X(t) - F_X(t) - \widehat{F}_X(x)) L\Big(\frac{x-t}{b_n}\Big)\, dt - F_X(x) \\
&\quad + \widehat{F}_X(x) + b_n^{-d} \int F_X(t) L\Big(\frac{x-t}{b_n}\Big)\, dt \ .
\end{aligned}$$

Next, using a Taylor expansion, we have $F_X(t) = F_X(x) + \sum_{j=1}^d \frac{\partial F_X(x)}{\partial x_j}(x_j - t_j) + \ldots + O(b_n^\tau)$ and also

$$b_n^{-d} \int F_X(t) L\Big(\frac{x-t}{b_n}\Big)\, dt = b_n^{-d} \int F_X(x) L\Big(\frac{x-t}{b_n}\Big)\, dt + O(b_n^\tau) \ ,$$

since F_X is τ times continuously differentiable by condition (A3)(ii) (if we take $\tau \leq 2d+1$), and since L is a kernel of order τ. We get

$$\widetilde{F}_X(x) - F_X(x) = \widehat{F}_X(x) - F_X(x) + b_n^{-d} \int (\widehat{F}_X(t) - F_X(t) - \widehat{F}_X(x) + F_X(x)) L\Big(\frac{x-t}{b_n}\Big)\, dt + O(b_n^\tau) \ .$$

We know that

$$b_n^{-d} \int (\widehat{F}_X(t) - F_X(t) - \widehat{F}_X(x) + F_X(x)) L\Big(\frac{x-t}{b_n}\Big)\, dt = O_P(n^{-1/2} b_n^{1/2}) = o_P(n^{-1/2}) \ ,$$

uniformly in $x \in \chi$. Note also that $O(b_n^\tau)$ is $o(n^{-1/2})$ since $nb_n^{2\tau} \to 0$. This implies that

$$\widetilde{F}_X(x) - F_X(x) = \widehat{F}_X(x) - F_X(x) + o_P(n^{-1/2}) \ ,$$

uniformly in $x \in \chi$, i.e. $\sup_{x\in\chi} |\widehat{F}_X(x) - \widetilde{F}_X(x)| = o_P(n^{-1/2})$. We will now check the second condition defined in (A10) in case (b). The function $\widetilde{\Phi}(x,\gamma) = \widetilde{f}_X(x)$ is here given by

$$\widetilde{f}_X(x) = \frac{\partial^d}{\partial x_1 \dots \partial x_d} \widetilde{F}_X(x) = (nb_n)^{-d} \sum_{i=1}^{n} L\left(\frac{X_i - x}{b_n}\right) ,$$

and $\Phi(x,\gamma) = f_X(x)$. Using the asymptotic properties of the classical kernel estimator of a density function, we obtain

$$\sup_{x\in\chi} |\widetilde{f}_X(x) - f_X(x)| = O_P((nb_n^d)^{-1/2}(\log n)^{1/2}) + O(b_n^{2\tau}).$$

The last expression is $o_P(1)$ since $nb_n^d(\log n)^{-1} \to \infty$. This concludes the proof for case (b). Finally, the proof of case (c) follows the same way as the proof of case (b). □

7.2 Proofs of the results under H_0

Proof of Theorem 3.1. First, we will prove that the class

$$\mathcal{F}_1 = \{(u,v) \to w(u,x,\gamma)(\Lambda_{\theta_0}(v) - m_{\beta_0}(u)), (x,\gamma) \in \Pi\} \ , \tag{7.1}$$

is Donsker. Using assumption (A9)(i), we know that there exists a finite number of $\widetilde{\varepsilon}-$brackets, say M, to cover $\mathcal{W}$. Let $w_1^L \leq w_1^U, \dots, w_M^L \leq w_M^U$ be the functions defining the M $\widetilde{\varepsilon}-$brackets to cover $\mathcal{W}$.

Next, the functions $y_1^L \leq y_1^U, \dots, y_M^L \leq y_M^U$, where $y_j^L(u,v) = \min(w_j^L(u)(\Lambda_{\theta_0}(v) - m_{\beta_0}(u)), w_j^U(u)(\Lambda_{\theta_0}(v) - m_{\beta_0}(u)))$, and $y_j^U(u,v) = \max(w_j^L(u)(\Lambda_{\theta_0}(v) - m_{\beta_0}(u)), w_j^U(u)(\Lambda_{\theta_0}(v) - m_{\beta_0}(u)))$ for $j = 1,\dots,M$, define the M $\widetilde{\varepsilon}-$brackets to cover $\mathcal{F}_1$. Indeed, for $j = 1,\dots,M$, we have

$$\begin{aligned} \left|\left| y_j^U(X,Y) - y_j^L(X,Y) \right|\right|_2^2 &= E\left[(w_j^U(X) - w_j^L(X))^2(\Lambda_{\theta_0}(Y) - m_{\beta_0}(X))^2\right] \\ &= E\left[(w_j^U(X) - w_j^L(X))^2\right] E(\varepsilon^2) \ , \end{aligned}$$

where the last equality is obtained by independence between X and ε. The last expression is $O(\widetilde{\varepsilon}^2)$ since $E(\varepsilon^2) = \sigma^2 < \infty$ and $E[(w_j^U(X) - w_j^L(X))^2] \leq \widetilde{\varepsilon}^2$ by definition of the brackets $w_j^L \leq w_j^U$ for $j = 1,\dots,M$. Hence, $N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_1, L_2(P)) < \infty$.

Consequently, if $f \in \mathcal{F}_1$, then $f(u,v)$ is bounded by $K|\Lambda_{\theta_0}(v) - m_{\beta_0}(u)|$ uniformly in $(x,\gamma) \in \Pi$, for some constant $K < \infty$, since the function w is uniformly bounded in $(x,\gamma) \in \Pi$ by condition (A9)(ii). This implies that

$$\begin{aligned}
&\int_0^{+\infty} \sqrt{\log N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_1, L_2(P))}\, d\widetilde{\varepsilon} \\
&= \int_0^{2K\sigma} \sqrt{\log N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_1, L_2(P))}\, d\widetilde{\varepsilon} + \int_{2K\sigma}^{+\infty} \sqrt{\log N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_1, L_2(P))}\, d\widetilde{\varepsilon} \\
&< \infty \ ,
\end{aligned}$$

as the first term is finite and the second term is equal to 0. Indeed, if $\widetilde{\varepsilon} > 2K\sigma$, only one bracket suffices to cover the class $\mathcal{F}_1$. Next, applying Theorem 2.5.6 in Van der Vaart and Wellner (1996), we get that the class $\mathcal{F}_1$ is Donsker. Hence, the process R_n converges weakly to a limiting Gaussian process.

Finally, by independence between X and ε, we have

$$E[w(X,x,\gamma)(\Lambda_{\theta_0}(Y) - m(X))] = E[w(X,x,\gamma)]E(\varepsilon) = 0 \ ,$$

as $\sup_{(x,\gamma)\in\Pi, t\in\mathcal{X}} |w(t,x,\gamma)| < \infty$ by assumption (A9)(ii). Finally,

$$\begin{aligned}
&\mathrm{Cov}[w(X,x_1,\gamma_1)(\Lambda_{\theta_0}(Y) - m(X)), w(X,x_2,\gamma_2)(\Lambda_{\theta_0}(Y) - m(X))] \\
&= E[w(X,x_1,\gamma_1)w(X,x_2,\gamma_2)\varepsilon^2] \ .
\end{aligned}$$

By independence between X and ε and since $E(\varepsilon^2) = \sigma^2$, this last expression is equal to $\sigma^2 E[w(X,x_1,\gamma_1)w(X,x_2,\gamma_2)]$. ☐

Proof of Theorem 3.2. First, note that

$$R_n^1(x,\gamma) = I + R_n(x,\gamma) - II \ , \tag{7.2}$$

where

$$I = n^{-1/2} \sum_{i=1}^n w(X_i,x,\gamma)(\Lambda_{\widehat{\theta}}(Y_i) - \Lambda_{\theta_0}(Y_i)) \ , \tag{7.3}$$

and

$$II = n^{-1/2} \sum_{i=1}^n w(X_i,x,\gamma)(m_{\widehat{\beta}}(X_i) - m(X_i)) \ . \tag{7.4}$$

Using successively a Taylor expansion with some ξ between $\widehat{\theta}$ and θ_0 and Proposition 1 in the supplementary material in Colling and Van Keilegom (2016), the expression I is equal to

$$\begin{aligned}
I &= n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(\dot{\Lambda}_{\xi}(Y_i))^t n^{1/2}(\widehat{\theta}-\theta_0) \\
&= n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y_i))^t\Big(-n^{1/2}\sum_{j=1}^{n} g(X_j,Y_j)+o_P(1)\Big) \\
&= -n^{-1/2}\sum_{j=1}^{n}[g(X_j,Y_j)]^t\Big(n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y_i))-E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))]\Big) \\
&\quad -n^{-1/2}\sum_{j=1}^{n} E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_j,Y_j)+o_P(1) \\
&= -n^{-1/2}\sum_{j=1}^{n} E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_j,Y_j)+o_P(1)\ . \qquad (7.5)
\end{aligned}$$

We have obtained the last two equalities because $n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y_i))^t = O_P(1)$ and $n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y_i)) - E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))] = O_P(n^{-1/2})$ uniformly in $(x,\gamma)\in\Pi$, and because $n^{-1/2}\sum_{j=1}^{n} g(X_j,Y_j) = O_P(1)$ by Proposition 1 in the supplementary material in Colling and Van Keilegom (2016). Indeed, for $l=1,\ldots,k$, we define the classes

$$\mathcal{G}_l = \{(u,v)\to w(u,x,\gamma)(\dot{\Lambda}_{\theta_0}(v))_l,(x,\gamma)\in\Pi\}\ ,$$

where $(\dot{\Lambda}_{\theta_0}(v))_l$ is the $l-$th component of the vector $\dot{\Lambda}_{\theta_0}(v)$. Define $z_{lj}^L(u,v) = \min(w_j^L(u)(\dot{\Lambda}_{\theta_0}(v))_l, w_j^U(u)(\dot{\Lambda}_{\theta_0}(v))_l)$ and $z_{lj}^U(u,v) = \max(w_j^L(u)(\dot{\Lambda}_{\theta_0}(v))_l, w_j^U(u)(\dot{\Lambda}_{\theta_0}(v))_l)$ for $j=1,\ldots,M$, where $w_1^L\le w_1^U$, …, $w_M^L\le w_M^U$ are the M functions defining the $\widetilde{\varepsilon}-$brackets to cover $\mathcal{W}$ that have been introduced in the proof of Theorem 3.1. Consequently, for $l=1,\ldots,k$ and $j=1,\ldots,M$, we have

$$\begin{aligned}
\Big|\Big|z_{lj}^U(X,Y)-z_{lj}^L(X,Y)\Big|\Big|_2^2 &= E\Big[\Big(w_j^U(X)-w_j^L(X)\Big)^2\Big(\dot{\Lambda}_{\theta_0}(Y)\Big)_l^2\Big] \\
&= E\Big[\Big(w_j^U(X)-w_j^L(X)\Big)^2 E\Big[\Big(\dot{\Lambda}_{\theta_0}(Y)\Big)_l^2\Big|X\Big]\Big] \\
&\le \sup_{x\in\chi} E\Big[\Big(\dot{\Lambda}_{\theta_0}(Y)\Big)_l^2\Big|X=x\Big]E\Big[\Big(w_j^U(X)-w_j^L(X)\Big)^2\Big]\ .
\end{aligned}$$

The last expression is $O(\widetilde{\varepsilon}^2)$ since $\sup_{x\in\chi} E[(\dot{\Lambda}_{\theta_0}(Y))_l^2|X=x]<\infty$ by condition (A5)(ii) and $E(w_j^U(X)-w_j^L(X))^2\le\widetilde{\varepsilon}^2$ by definition of the brackets $w_j^L\le w_j^U$ for $j=1,\ldots,M$. Hence, in a similar way as in the proof of Theorem 3.1, we conclude that $\mathcal{G}_l$, for $l = 1, \ldots, k$, is Donsker using Theorem 2.5.6 in Van der Vaart and Wellner (1996), and then (7.5) follows.

Next, as we are under H_0, $m(x) = m_{\beta_0}(x)$ and using a Taylor expansion for some value ζ between $\widehat{\beta}$ and β_0, expression II is equal to

$$\begin{aligned} II &= n^{-1/2}\sum_{i=1}^{n} w(X_i, x, \gamma)(\widehat{\beta} - \beta_0)^t \frac{\partial}{\partial \beta} m_\zeta(X_i) \\ &= n^{1/2}(\widehat{\beta} - \beta_0)^t n^{-1} \sum_{i=1}^{n} w(X_i, x, \gamma) h(X_i, \zeta) \\ &= n^{1/2}(\widehat{\beta} - \beta_0)^t n^{-1} \sum_{i=1}^{n} w(X_i, x, \gamma)(h(X_i, \zeta) - h(X_i, \beta_0)) \\ &\quad + n^{1/2}(\widehat{\beta} - \beta_0)^t n^{-1} \sum_{i=1}^{n} \Big[w(X_i, x, \gamma) h(X_i, \beta_0) - H(x, \gamma, \beta_0) \Big] \\ &\quad + n^{1/2}(\widehat{\beta} - \beta_0)^t H(x, \gamma, \beta_0) \ . \end{aligned} \tag{7.6}$$

Note that $n^{1/2}(\widehat{\beta} - \beta_0) = O_P(1)$ by Lemma 4 in the supplementary material in Colling and Van Keilegom (2016). Moreover, as $h(X, \beta)$ is a differentiable function in β with a uniformly bounded derivative (see condition (A6)(ii)), and as $\widehat{\beta} - \beta_0 = o_P(1)$ by Lemma 2 in the supplementary material in Colling and Van Keilegom (2016), we have $\sup_{x \in \chi} |h(x, \zeta) - h(x, \beta_0)| = o_P(1)$. Consequently,

$$\begin{aligned} \left| n^{-1} \sum_{i=1}^{n} w(X_i, x, \gamma)(h(X_i, \zeta) - h(X_i, \beta_0)) \right| &\leq n^{-1} \sum_{i=1}^{n} \Big| w(X_i, x, \gamma) \Big| \Big| h(X_i, \zeta) - h(X_i, \beta_0) \Big| \\ &= o_P(1) \ , \end{aligned}$$

since $\sup_{(x,\gamma) \in \Pi, t \in \chi} |w(t, x, \gamma)| < \infty$ by condition (A9)(ii). Hence, the first term on the right hand side of (7.6) is $o_P(1)$. Next, for $l = 1, \ldots, q$, we define the classes

$$\mathcal{H}_l = \{u \to w(u, x, \gamma) h_l(u, \beta_0), (x, \gamma) \in \Pi\} \ ,$$

where $h_l(u, \beta_0)$ is the $l-$th component of the vector $h(u, \beta_0)$, i.e. $h_l(u, \beta_0) = \frac{\partial m_{\beta_0}(X)}{\partial \beta_l}$, and $v_{lj}^L(u) = \min(w_j^L(u) h_l(u, \beta_0), w_j^U(u) h_l(u, \beta_0))$ and $v_{lj}^U(u) = \max(w_j^L(u) h_l(u, \beta_0), w_j^U(u) h_l(u, \beta_0))$

for $j = 1, \ldots, M$. Consequently, for $l = 1, \ldots, q$ and $j = 1, \ldots, M$, we have

$$\begin{aligned}
\left|\left|v_{lj}^U(X) - v_{lj}^L(X)\right|\right|_1 &= E\left[\left|w_j^U(X) - w_j^L(X)\right|\left|h_l(X, \beta_0)\right|\right] \\
&\leq \left(E[h_l^2(X, \beta_0)]\right)^{1/2}\left(E\left[(w_j^U(X) - w_j^L(X))^2\right]\right)^{1/2} \\
&\leq \sup_{x \in \mathcal{X}}\left|h_l(x, \beta_0)\right|\left(E\left[(w_j^U(X) - w_j^L(X))^2\right]\right)^{1/2} .
\end{aligned}$$

This last expression is finite by condition (A6)(ii) and since $E[(w_j^U(X) - w_j^L(X))^2] \leq \widetilde{\varepsilon}^2$. Hence, for $l = 1, \ldots, q$, $N_{[]}(\widetilde{\varepsilon}, \mathcal{H}_l, L_1(P)) < \infty$, which implies that $\mathcal{H}_l$ is Glivenko-Cantelli by Theorem 2.4.1 in Van der Vaart and Wellner (1996), i.e.

$$\sup_{f \in \mathcal{H}_l}\left|n^{-1}\sum_{i=1}^n f(X_i) - E(f(X))\right| = o_P(1) .$$

Then, the second term on the right hand side of (7.6) is $o_P(1)$. In conclusion, combining (7.2), (7.5) and (7.6), we get

$$R_n^1(x, \gamma) = R_n(x, \gamma) - n^{-1/2}\sum_{i=1}^n E[w(X, x, \gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_i, Y_i) - n^{1/2}(\widehat{\beta} - \beta_0)^t H(x, \gamma, \beta_0) + o_P(1) ,$$

uniformly in $(x, \gamma) \in \Pi$. Finally, we conlude the proof using Lemma 4 in the supplementary material in Colling and Van Keilegom (2016) (with $r \equiv 0$) :

$$\begin{aligned}
R_n^1(x, \gamma) &= R_n(x, \gamma) - n^{-1/2}\sum_{i=1}^n E[w(X, x, \gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_i, Y_i) \\
&\quad - n^{-1/2}\sum_{i=1}^n H^t(x, \gamma, \beta_0)\Big\{\eta_{\beta_0}(X_i, Y_i) - \Omega^{-1}E\left[\frac{\partial m_{\beta_0}(X)}{\partial \beta}(\dot{\Lambda}_{\theta_0}(Y))^t\right]g(X_i, Y_i)\Big\} \\
&\quad + o_P(1) .
\end{aligned}$$

The last equality was obtained using the fact that $H(x, \gamma, \beta_0)$ is bounded uniformly in $(x, \gamma) \in \Pi$. Indeed, for each $l = 1, \ldots, q$, if we denote by $H_l(x, \gamma, \beta)$ the $l-$th component of the vector $H(x, \gamma, \beta)$, i.e. $H_l(x, \gamma, \beta) = E[w(X, x, \gamma)\frac{\partial m_\beta(X)}{\partial \beta_l}]$, we have that

$$\sup_{(x,\gamma) \in \Pi}\left|H_l(x, \gamma, \beta_0)\right| \leq \sup_{(x,\gamma) \in \Pi, t \in \mathcal{X}}\left|w(t, x, \gamma)\right|\sup_{t \in \mathcal{X}}\left|\frac{\partial m_{\beta_0}(t)}{\partial \beta_l}\right| .$$

Using conditions (A9)(ii) and (A6)(ii), we conclude that the last expression is finite. $\square$

Proof of Theorem 3.3. First, note that

$$\left| \int_\Pi [R_n^1(x,\gamma)]^2 \, d(\Psi_n(x,\gamma) - \Psi(x,\gamma)) \right| \le A + B \ ,$$

where

$$A = \left| \int_\Pi [R_n^1(x,\gamma)]^2 \, d(\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)) \right| \ ,$$

and

$$B = \left| \int_\Pi [R_n^1(x,\gamma)]^2 \, d(\widetilde{\Psi}(x,\gamma) - \Psi(x,\gamma)) \right| \ .$$

We will first prove that the term A is $o_P(1)$ uniformly in $(x,\gamma) \in \Pi$. Using integration by parts, the term A is equal to

$$A = \left| \Big[(\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)) R_n^1(x,\gamma) \Big]_\Pi - 2 \int_\Pi (\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)) R_n^1(x,\gamma) \, dR_n^1(x,\gamma) \right| .$$

Note that the first term on the right hand side is $o_P(1)$, since $\sup_{(x,\gamma)\in\Pi} |R_n^1(x,\gamma)| = O_P(1)$ by Corollary 3.1 and since $\sup_{(x,\gamma)\in\Pi} |\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)| = o_P(n^{-1/2}) = o_P(1)$ by condition (A10). Consequently,

$$\begin{aligned} A &= 2 \left| \int_\Pi (\Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma)) R_n^1(x,\gamma) \, dR_n^1(x,\gamma) \right| \\ &\le 2 \sup_{(x,\gamma)\in\Pi} \left| \Psi_n(x,\gamma) - \widetilde{\Psi}(x,\gamma) \right| \sup_{(x,\gamma)\in\Pi} \left| R_n^1(x,\gamma) \right| \\ &\quad \times \sup_{t\in\chi} V\Big[w(t,\cdot,\cdot) \Big] n^{-1/2} \sum_{i=1}^n \left| \Lambda_{\widehat{\theta}}(Y_i) - m_{\widehat{\beta}}(X_i) \right| . \end{aligned} \tag{7.7}$$

Next, the first factor on the right hand side of (7.7) is $o_P(n^{-1/2})$ using condition (A10), the second factor is $O_P(1)$ as explained just above and $\sup_{t\in\chi} V[w(t,\cdot,\cdot)] < \infty$ using condition (A9)(iii). Moreover, it follows easily from the proofs of Theorem 3.2 and Corollary 3.1 that

$$n^{-1/2} \sum_{i=1}^n \left| \Lambda_{\widehat{\theta}}(Y_i) - m_{\widehat{\beta}}(X_i) \right| = O_P(n^{1/2}) \ ,$$

which implies that A is $o_P(1)$ uniformly in $(x,\gamma) \in \Pi$. Finally, we will prove that the term B is $o_P(1)$ uniformly in $(x,\gamma) \in \Pi$. As $\widetilde{\Psi}(x,\gamma) - \Psi(x,\gamma)$ is a differentiable function with respect to (x,γ), we have that

$$\begin{aligned} B &= \left| \int_\Pi [R_n^1(x,\gamma)]^2 (\widetilde{\Phi}(x,\gamma) - \Phi(x,\gamma)) \, dx d\gamma \right| \\ &\le K \sup_{(x,\gamma)\in\Pi} \left| R_n^1(x,\gamma) \right|^2 \sup_{(x,\gamma)\in\Pi} \left| \widetilde{\Phi}(x,\gamma) - \Phi(x,\gamma) \right| , \end{aligned} \tag{7.8}$$

for some $K < \infty$. We conclude the proof using the facts that the first factor on the right hand side of (7.8) is $O_P(1)$ as explained just above and that the second factor is $o_P(1)$ using condition (A10). $\square$

Proof of Corollary 3.1. First, we prove that $R_n^1(x,\gamma)$ converges to a limiting Gaussian process. In the proof of Theorem 3.1 we have shown that the class $\mathcal{F}_1$ (corresponding to the process $R_n(x,\gamma)$) is Donsker. Hence, using Theorem 3.2, it suffices to show that the class

$$\mathcal{F}_2 = \Big\{(u,v) \to G(x,\gamma,u,v,\theta_0,\beta_0), (x,\gamma) \in \Pi\Big\}$$

is Donsker. Recall that

$$\begin{aligned} G(x,\gamma,u,v,\theta_0,\beta_0) &= H^t(x,\gamma,\beta_0)\eta_{\beta_0}(u,v) + E[w(X,x,\gamma)(\dot\Lambda_{\theta_0}(Y))^t]g(u,v) \\ & \quad -H^t(x,\gamma,\beta_0)\Omega^{-1}E\left[\frac{\partial m_{\beta_0}(X)}{\partial\beta}(\dot\Lambda_{\theta_0}(Y))^t\right]g(u,v)\ . \end{aligned}$$

Hence, each term of $G(x,\gamma,u,v,\theta_0,\beta_0)$ can be decomposed in a factor that depends on (x,γ) but not on (u,v), and another factor that depends on (u,v) but not on (x,y). Hence, it can be easily seen using similar arguments as before that the class $\mathcal{F}_2$ is Donsker. In fact, we only need to prove that $\sup_{(x,\gamma)\in\Pi}|H(x,\gamma,\beta_0)| < \infty$ and that $\sup_{(x,\gamma)\in\Pi}|E[w(X,x,\gamma)(\dot\Lambda_{\theta_0}(Y))_l]| < \infty$ for $l = 1,\dots,k$. The former property has been shown at the end of the proof of Theorem 3.2, whereas for the latter note that for $l = 1,\dots,k$, we have

$$\begin{aligned} \sup_{(x,\gamma)\in\Pi}\left|E\Big[w(X,x,\gamma)(\dot\Lambda_{\theta_0}(Y))_l\Big]\right| &= \sup_{(x,\gamma)\in\Pi}\left|E\Big[w(X,x,\gamma)E\Big[(\dot\Lambda_{\theta_0}(Y))_l\Big|X\Big]\Big]\right| \\ &\le \sup_{(x,\gamma)\in\Pi,t\in\chi}\Big|w(t,x,\gamma)\Big|\sup_{x\in\chi}E\Big[\Big|(\dot\Lambda_{\theta_0}(Y))_l\Big|\Big|X=x\Big]\ . \end{aligned}$$

This last expression is finite by conditions (A9)(ii) and (A5)(ii). It now follows that $R_n^1(x,\gamma)$ converges to a Gaussian process R_∞^1. We can easily see that R_∞^1 has zero mean, because R_n converges to a centered Gaussian process by Theorem 3.1, $E(\eta_{\beta_0}(X,Y)) = 0$ and $E(g(X,Y)) = 0$ since $g(X,Y) = \Gamma^{-1}\xi(\theta_0,X,Y)$ and $\xi(\theta_0,X,Y)$ is the derivative of the likelihood. Moreover,

$$\begin{aligned}
&\mathrm{Cov}[w(X,x_1,\gamma_1)(\Lambda_{\theta_0}(Y)-m(X))-G(x_1,\gamma_1,X,Y,\theta_0,\beta_0),\\
&\qquad w(X,x_2,\gamma_2)(\Lambda_{\theta_0}(Y)-m(X))-G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)]\\
&=E[\{w(X,x_1,\gamma_1)(\Lambda_{\theta_0}(Y)-m(X))-G(x_1,\gamma_1,X,Y,\theta_0,\beta_0)\}\\
&\qquad\times\{w(X,x_2,\gamma_2)(\Lambda_{\theta_0}(Y)-m(X))-G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)\}]\\
&=E[w(X,x_1,\gamma_1)w(X,x_2,\gamma_2)\varepsilon^2]-E[G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)w(X,x_1,\gamma_1)\varepsilon]\\
&\quad-E[G(x_1,\gamma_1,X,Y,\theta_0,\beta_0)w(X,x_2,\gamma_2)\varepsilon]\\
&\quad+E[G(x_1,\gamma_1,X,Y,\theta_0,\beta_0)G(x_2,\gamma_2,X,Y,\theta_0,\beta_0)].
\end{aligned}$$

Note that the first term on the right hand side of the last expression is $C(x_1,\gamma_1,x_2,\gamma_2)$. $\square$

Proof of Corollary 3.2. First, to obtain the limiting distribution of D_n, it suffices to apply Corollary 3.1 and the continuous mapping theorem. Next, for W_n^2, we use Theorem 3.3, which states that we can replace $d\Psi_n(x,\gamma)$ by $d\Psi(x,\gamma)$ up to a negligible term. Hence, we obtain the limiting distribution of W_n^2 applying Corollary 3.1 and the continuous mapping theorem. $\square$

7.3 Proofs of the results under H_{1n}

Proof of Theorem 3.4. First, we remind that

$$R_n^1(x,\gamma)=I+R_n(x,\gamma)-II\ , \tag{7.9}$$

where I and II are given in (7.3) and (7.4) respectively. Exactly in the same way as in the proof of Theorem 3.2, expression I is equal to

$$I=-n^{-1/2}\sum_{i=1}^{n}E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_i,Y_i)+o_P(1)\ . \tag{7.10}$$

Next, as we are under H_{1n}, $m(x)=m_{\beta_0}(x)+n^{-1/2}r(x)$ and using a Taylor expansion for some value ζ between $\widehat{\beta}$ and β_0, expression II is equal to

$$\begin{aligned}
II &= n^{-1/2}\sum_{i=1}^{n} w(X_i,x,\gamma)(\widehat{\beta}-\beta_0)^t \frac{\partial}{\partial\beta} m_{\zeta}(X_i) - n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)r(X_i) \\
&= n^{1/2}(\widehat{\beta}-\beta_0)^t n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)h(X_i,\zeta) - n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)r(X_i) \\
&= n^{1/2}(\widehat{\beta}-\beta_0)^t n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(h(X_i,\zeta)-h(X_i,\widetilde{\beta}_{0n})) \\
&\quad + n^{1/2}(\widehat{\beta}-\beta_0)^t n^{-1}\sum_{i=1}^{n}\Big[w(X_i,x,\gamma)h(X_i,\widetilde{\beta}_{0n}) - H(x,\gamma,\widetilde{\beta}_{0n})\Big] \\
&\quad + n^{1/2}(\widehat{\beta}-\beta_0)^t H(x,\gamma,\widetilde{\beta}_{0n}) - n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)r(X_i) \ . \qquad (7.11)
\end{aligned}$$

Note that $n^{1/2}(\widehat{\beta}-\beta_0) = O_P(1)$ by Lemma 4 in the supplementary material in Colling and Van Keilegom (2016). Moreover, as the function h is a differentiable function in β with a uniformly bounded derivative (see condition (A6)(ii)) and $\widetilde{\beta}_{0n}-\beta_0 = o_P(1)$ using Lemma 3 in the supplementary material in Colling and Van Keilegom (2016), we can prove in exactly the same way as in the proof of Theorem 3.2 that expressions

$$n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)(h(X_i,\zeta)-h(X_i,\widetilde{\beta}_{0n})) \ ,$$

and

$$n^{-1}\sum_{i=1}^{n}\Big[w(X_i,x,\gamma)h(X_i,\widetilde{\beta}_{0n}) - H(x,\gamma,\widetilde{\beta}_{0n})\Big] \ ,$$

are $o_P(1)$ uniformly in $(x,\gamma)\in\Pi$, which implies that the first and the second terms on the right hand side of (7.11) are also $o_P(1)$ uniformly in $(x,\gamma)\in\Pi$. Then, combining (7.9), (7.10) and (7.11), we get

$$\begin{aligned}
R_n^1(x,\gamma) &= R_n(x,\gamma) - n^{-1/2}\sum_{i=1}^{n} E[w(X,x,\gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_i,Y_i) \\
&\quad - n^{1/2}(\widehat{\beta}-\beta_0)^t H(x,\gamma,\widetilde{\beta}_{0n}) + n^{-1}\sum_{i=1}^{n} w(X_i,x,\gamma)r(X_i) + o_P(1) \ .
\end{aligned}$$

Finally, we define the class

$$\mathcal{F}_3 = \{u \to w(u,x,\gamma)r(u), (x,\gamma)\in\Pi\} \ ,$$

and we let $r_j^L(u) = \min(w_j^L(u)r(u), w_j^U(u)r(u))$ and $r_j^U(u) = \max(w_j^L(u)r(u), w_j^U(u)r(u))$ for $j = 1, \ldots, M$, where $w_1^L \leq w_1^U$, $\ldots$, $w_M^L \leq w_M^U$ are the M functions defining the $\widetilde{\varepsilon}-$brackets to cover $\mathcal{W}$ that have been introduced in the proof of Theorem 3.1. Consequently, for $j = 1, \ldots, M$, we have

$$\begin{aligned}\left|\left|r_j^U(X) - r_j^L(X)\right|\right|_1 &= E\left[\left|w_j^U(X) - w_j^L(X)\right|\left|r(X)\right|\right] \\ &\leq \left(E[r^2(X)]\right)^{1/2}\left(E\left[(w_j^U(X) - w_j^L(X))^2\right]\right)^{1/2} .\end{aligned}$$

This last expression is finite by condition (A12) and since $E[(w_j^U(X) - w_j^L(X))^2] \leq \widetilde{\varepsilon}^2$. Hence, $N_{[]}(\widetilde{\varepsilon}, \mathcal{F}_3, L_1(P)) < \infty$, which implies that $\mathcal{F}_3$ is Glivenko-Cantelli by Theorem 2.4.1 in Van der Vaart and Wellner (1996), i.e.

$$\sup_{f \in \mathcal{F}_3}\left|n^{-1}\sum_{i=1}^n f(X_i) - E(f(X))\right| = o_P(1) .$$

This means that $n^{-1}\sum_{i=1}^n w(X_i, x, \gamma)r(X_i) = E[w(X, x, \gamma)r(X)] + o_P(1)$ uniformly in $(x, \gamma) \in \Pi$. We conlude the proof using Lemma 4 in the supplementary material in Colling and Van Keilegom (2016) :

$$\begin{aligned}R_n^1(x, \gamma) &= R_n(x, \gamma) - n^{-1/2}\sum_{i=1}^n E[w(X, x, \gamma)(\dot{\Lambda}_{\theta_0}(Y))^t]g(X_i, Y_i) \\ &\quad - n^{-1/2}\sum_{i=1}^n H^t(x, \gamma, \widetilde{\beta}_{0n})\left\{\eta_{\widetilde{\beta}_{0n}}(X_i, Y_i) - \Omega^{-1}E\left[\frac{\partial m_{\widetilde{\beta}_{0n}}(X)}{\partial \beta}(\dot{\Lambda}_{\theta_0}(Y))^t\right]g(X_i, Y_i)\right\} \\ &\quad - H^t(x, \gamma, \widetilde{\beta}_{0n})\Omega^{-1}\int r(u)\frac{\partial m_{\widetilde{\beta}_{0n}}(u)}{\partial \beta}\, dF_X(u) + E[w(X, x, \gamma)r(X)] + o_P(1).\end{aligned}$$

The last equality was obtained using the fact that $H(x, \gamma, \widetilde{\beta}_{0n})$ is bounded uniformly in $(x, \gamma) \in \Pi$, the proof is exactly the same as in Theorem 3.2. □

Proof of Corollary 3.3. To prove this result, we will show that expressions $R_n(x, \gamma) - n^{-1/2}\sum_{i=1}^n G(x, \gamma, X_i, Y_i, \theta_0, \beta_0)$ under H_0 and $R_n(x, \gamma) - n^{-1/2}\sum_{i=1}^n G(x, \gamma, X_i, Y_i, \theta_0, \widetilde{\beta}_{0n})$ under H_{1n} have the same limiting Gaussian process R_∞^1. Since the bias term $b(x, \gamma)$ was already obtained in Theorem 3.4, this will conclude the proof.

Recall the definition of $G(x, \gamma, X, Y, \theta_0, \beta_0)$ given in (3.1). First note that Theorem 3.2 in Colling and Van Keilegom (2016) shows that $n^{-1/2}\sum_{i=1}^n \eta_{\beta_0}(X_i, Y_i)$ under H_0 and $n^{-1/2}\sum_{i=1}^n \eta_{\widetilde{\beta}_{0n}}(X_i, Y_i)$ under H_{1n} have the same limiting distribution.

Moreover, combining assumption (A6)(ii) and the fact that $\widetilde{\beta}_{0n} - \beta_0 = o_P(1)$ by Lemma 3 in the supplementary material in Colling and Van Keilegom (2016), we have that $\sup_{t \in \mathcal{X}} \left| \frac{\partial m_{\widetilde{\beta}_{0n}}(t)}{\partial \beta} - \frac{\partial m_{\beta_0}(t)}{\partial \beta} \right| = o(1)$.

Next, for $l = 1, \ldots, q$, we denote by $H_l(x, \gamma, \beta)$ the $l-$th component of the vector $H(x, \gamma, \beta)$, i.e. $H_l(x, \gamma, \beta) = E[w(X, x, \gamma) \frac{\partial m_\beta(X)}{\partial \beta_l}]$. For $l = 1, \ldots, q$, we have

$$\begin{aligned}
\sup_{(x,\gamma)\in\Pi} \left| H_l(x, \gamma, \widetilde{\beta}_{0n}) - H_l(x, \gamma, \beta_0) \right| &= \sup_{(x,\gamma)\in\Pi} \left| E\left[w(X, x, \gamma) \left(\frac{\partial m_{\widetilde{\beta}_{0n}}(X)}{\partial \beta_l} - \frac{\partial m_{\beta_0}(X)}{\partial \beta_l} \right) \right] \right| \\
&\leq \sup_{(x,\gamma)\in\Pi, t\in\mathcal{X}} \left| w(t, x, \gamma) \right| \sup_{t\in\mathcal{X}} \left| \frac{\partial m_{\widetilde{\beta}_{0n}}(t)}{\partial \beta_l} - \frac{\partial m_{\beta_0}(t)}{\partial \beta_l} \right| .
\end{aligned}$$

The last expression is $o_P(1)$ using condition (A9)(ii). Hence, $\sup_{(x,\gamma)\in\Pi} |H_l(x, \gamma, \widetilde{\beta}_{0n}) - H_l(x, \gamma, \beta_0)| = o(1)$ for $l = 1, \ldots, q$.

This shows the above statement, and hence finishes the proof. □

Proof of Corollary 3.4. First, note that we can prove by very similar arguments the same result as in Theorem 3.3 but under H_{1n}. The proof of this Corollary is similar to the proof of Corollary 3.2. It suffices to apply Corollary 3.3 and the continuous mapping theorem. □

References

Akritas, M.G. and Van Keilegom, I. (2001). Nonparametric estimation of the residual distribution. *Scandinavian Journal of Statistics*, **28**, 549-568.

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

Bierens, H.J. (1982). Consistent model specification tests. *Journal of Econometrics*, **20**, 105-134.

Bierens, H.J. (1984). Model specification testing of time series regressions. *Journal of Econometrics*, **26**, 323-353.

Bierens, H.J. (1990). A consistent conditional moment test of functional form. *Econometrica*, **58**, 1443-1458.

Bierens, H.J. and Ploberger, W. (1997). Asymptotic theory of integrated conditional moment test. *Econometrica*, **65**, 1129-1151.

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

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

Carroll, R.J. and Ruppert, D. (1988). *Transformation and Weighting in Regression.* Chapman and Hall, New-York.

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**, 1-18.

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

Escanciano, J.C. (2006a). A consistent test for regression models using projections. *Econometric Theory*, **22**, 1030-1051.

Escanciano, J.C. (2006b). Goodness-of-fit tests for linear and nonlinear time series models. *Journal of the American Statistical Association*, **101**, 531-541.

Escanciano, J.C. (2007). Model checks using residual marked empirical processes. *Statistica Sinica*, **17**, 115-138.

González-Manteiga, W. and Crujeiras, R. (2013). An updated review of goodness-of-fit tests for regression models (with discussion). *TEST*, **22**, 361-447.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1926-1947.

Heuchenne, C., Samb, R. and Van Keilegom, I. (2015). Estimating the error 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**, 103-137.

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

Jacho-Chavez, D., Lewbel, A. and Linton, O. (2008). Identification and nonparametric estimation of a transformation additively separable model. Technical report.

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

Koul, H.L. and Stute, W. (1999). Nonparametric model checks for time series. *Annals of Statistics*, **27**, 204-236.

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

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

Neumeyer, N. (2009). Smooth residual bootstrap for empirical processes of non-parametric regression residuals. *Scandinavian Journal of Statistics*, **36**, 204-228.

Neumeyer, N. and Van Keilegom, I. (2010). Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, **101**, 1067-1078.

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

Stinchcombe, M. and White, H. (1998). Consistent specification testing with nuisance parameters present only under the alternative. *Econometric Theory*, **14**, 295-325.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613-641.

Stute, W. and Zhu, L.X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, **29**, 535-545.

Van der Vaart, A.W. and Wellner, J.A (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New-York.

Van Keilegom, I., González-Manteiga, W. and Sánchez Sellero, C. (2008). Goodness of fit tests in parametric regression based on the estimation of the error distribution. *TEST*, **17**, 401-415.

Vanhems, A. and Van Keilegom, I. (2016). Semiparametric transformation model with endogeneity : a control function approach. *Econometric Theory* (under revision).

Whang, Y.J. (2000). Consistent bootstrap tests of parametric regression functions. *Journal of Econometrics*, **98**, 27-46.

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

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