**Noname manuscript No.**
(will be inserted by the editor)

# Goodness-of-fit tests in semiparametric transformation models

**Benjamin Colling · Ingrid Van Keilegom**

Received: date / Accepted: date

**Abstract** Consider a semiparametric transformation model of the form $\Lambda_\theta(Y) = m(X) + \varepsilon$, where $Y$ is a univariate dependent variable, $X$ is a $d$-dimensional covariate, and $\varepsilon$ is independent of $X$ and has mean zero. We assume that $\{\Lambda_\theta : \theta \in \Theta\}$ is a parametric family of strictly increasing functions, while $m$ is an unknown regression function. The goal of the paper is to develop tests for the null hypothesis that $m(\cdot)$ belongs to a certain parametric family of regression functions. We propose a Kolmogorov-Smirnov and a Cramér-von Mises type test statistic, which measure the distance between the distribution of $\varepsilon$ estimated under the null hypothesis and the distribution of $\varepsilon$ without making use of this null hypothesis. The estimated distributions are based on a profile likelihood estimator of $\theta$ and a local polynomial estimator of $m(\cdot)$. The limiting distributions of these two test statistics are established 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 procedures, and we apply our tests to data on the scattering of sunlight in the atmosphere.

**Keywords** Bootstrap · Goodness-of-fit · Local polynomial smoothing · Profile likelihood · Semiparametric regression · Transformation model

B. Colling · I. Van Keilegom
Institut de statistique, biostatistique et sciences actuarielles, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium.

B. Colling
E-mail: benjamin.colling@uclouvain.be

I. Van Keilegom (✉)
E-mail: ingrid.vankeilegom@uclouvain.be

**Mathematics Subject Classification (2000)** 62G08 · 62G09 · 62G10 · 62G20

## 1 Introduction

Consider the following semiparametric transformation model :

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

where $\Lambda_\theta(\cdot)$ belongs to a parametric family of strictly increasing functions and the function $m(\cdot)$ is unknown. We assume that $X$ is a $d$-dimensional covariate, $Y$ is a univariate response variable, the error term $\varepsilon$ is independent of $X$, and $E(\varepsilon) = 0$. Let $\theta$ belong to a finite dimensional compact subset $\Theta$ of $\mathbb{R}^k$, and denote the true but unknown values of $\theta$ and $m(\cdot)$ by $\theta_0$ and $m_0(\cdot)$.

The motivation for considering this model comes from the rich literature on parametric transformations in regression, starting from the seminal paper by Box and Cox (1964). They proposed a parametric family of power transformations that includes as special cases the logarithm and the identity. They suggested that when this power transformation is applied to the response in a linear regression model, the regression function of the new model might have an additive structure, and the new error might be approximately normal and homoscedastic. Other transformations have been proposed in the literature, like for example, the Zellner and Revankar (1969) transform and the Bickel and Doksum (1981) transform. See also the book by Carroll and Ruppert (1988) and the review paper by Sakia (1992) for more details and references on this topic.

Whereas the above references restrict attention to models in which the regression function (as well as the transformation) is parametric, we will focus in this paper on model (1), which assumes that the regression function is nonparametric. The estimation of this (semiparametric) transformation model has been studied by Linton, Sperlich and Van Keilegom (2008). They proposed two different estimators of the transformation parameter $\theta$ and developed the asymptotic properties of these estimators. Moreover, Colling, Heuchenne, Samb and Van Keilegom (2013) and Heuchenne, Samb and Van Keilegom (2014) studied nonparametric estimators of the density and of the distribution function of the error term $\varepsilon$ under this model. Other papers that have studied the estimation of this model include Vanhems and Van Keilegom (2013), who suppose 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 work by Horowitz (1996), who worked with a nonparametric transformation $\Lambda$ and a parametric regression function $m$, and the papers by Horowitz (2001) and Jacho-Chavez, Lewbel and Linton (2008), who suppose that both $\Lambda$ and $m$ are nonparametric.

All the above papers focus on the problem of estimation of a transformation model (that can be of parametric, semiparametric or nonparametric nature). As far as we know, no paper has considered so far the problem of testing in a transformation model. Several aspects of the model can be tested, like the form of the transformation, the form of the regression function, the homoscedasticity of the error term, or the separability of the regression model. In this paper, we like to test the hypothesis

$$H_0 : m \in \mathcal{M}, \tag{2}$$

where $\mathcal{M} = \{m_\beta : \beta \in \mathcal{B}\}$ is some parametric class of regression functions and $\mathcal{B} \subset \mathbb{R}^q$. We will construct two test statistics, which measure a certain distance between the distribution function of $\varepsilon$ estimated in a semiparametric way and the distribution function of $\varepsilon$ estimated under the null hypothesis. We will show that the two distributions are equal if and only if the null hypothesis $H_0$ is true.

The idea of testing the form of the regression function by comparing two estimators of the distribution of the error term was introduced for the first time by Van Keilegom, González-Manteiga and Sánchez-Sellero (2008). Their test was developed for a nonparametric location-scale model without transforming the response variable. In the present paper we will see how their ideas and methodology can be carried over to a transformation model. For the same location-scale model, a similar testing approach was also used (among others) by Pardo-Fernández, Van Keilegom and González-Manteiga (2007) for testing the equality of regression curves, and by Dette, Neumeyer and Van Keilegom (2007) for testing the form of the variance function. All these papers build further on the work of Akritas and Van Keilegom (2001), who studied the asymptotic properties of a nonparametric estimator of the error distribution in a location-scale model without transforming the response variable.

Instead of using the idea based on the comparison of error distributions, other approaches could be used as well. We refer to the nice review paper by González-Manteiga and Crujeiras (2013) for a recent overview of developments on goodness-of-fit tests for regression models. Among the possible alternative testing procedures are the tests in the spirit of the seminal papers by Härdle and Mammen (1993) and Stute (1997). They will be considered in forthcoming papers.

The paper is organized as follows. In the section 2 we explain in detail the testing procedure. Section 3 contains the main asymptotic results concerning the proposed test statistics. In Section 4 we explain how the critical values of these test statistics can be obtained using a bootstrap procedure, and a simulation study is carried out to illustrate the performance of our tests. Section 5 is devoted to the application of our testing procedures to data on the scattering of sunlight in the atmosphere, and in Section 6 we give some general conclusions. Finally, Section 7 contains the technical assumptions and the supplementary material contains the proofs of the main results.

## 2 The proposed test

### 2.1 Notations and definitions

Suppose that we have randomly drawn an *iid* sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from model (1), where the components of $X_i$ are denoted by $(X_{i1}, \ldots, X_{id})$ for $i = 1, \ldots, n$. Denote by $F_X$ and $F_\varepsilon$ the distribution functions of $X$ and $\varepsilon$ respectively. The probability density functions of $X$ and $\varepsilon$ will be denoted respectively by $f_X$ and $f_\varepsilon$. Moreover, assume that $X$ has compact support $\chi \subset \mathbb{R}^d$, define the regression function

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

and let $\sigma^2 = V(\varepsilon) \ < \infty$. Note that $m(x, \theta_0) = m(x)$. Also, denote

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

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

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

be a $(k \times 1)$-vector where $\theta = (\theta_1, \ldots, \theta_k)^t$. Similar notations will be used for other functions. For any function $\varphi$, 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

We start by estimating the parameter $\theta$. Linton, Sperlich and Van Keilegom (2008) proposed two estimation methods for the unknown true parameter vector $\theta_0$ : a profile likelihood method and a mean squared distance from independence method. Here, we will use the profile likelihood estimator, since it was shown in the latter paper that it outperforms the other estimator. Note however that our model and estimation method are slightly different from what Linton, Sperlich and Van Keilegom (2008) did : we assume that $m(\cdot)$ is completely unspecified (whereas they assume an additive or multiplicative structure on $m(\cdot)$), and we will use local polynomial smoothing (instead of kernel smoothing based on higher order kernels). This has however no impact on how the profile likelihood estimator of $\theta$ is constructed.

The idea of the profile likelihood method is to calculate the log-likelihood function of $Y$ given $X$ and to replace all unknown expressions by nonparametric estimators. The log-likelihood function of $Y$ given $X$ is given by :

$$\sum_{i=1}^{n} \left\{ \log f_{\varepsilon(\theta_0)}(\Lambda_{\theta_0}(Y_i) - m(X_i, \theta_0)) + \log \Lambda'_{\theta_0}(Y_i) \right\} .$$

In this expression, $\Lambda_{\theta_0}(\cdot)$ and $\Lambda'_{\theta_0}(\cdot)$ are known (except for the parameter $\theta_0$), unlike $m(\cdot,\theta_0)$ and $f_{\varepsilon(\theta_0)}(\cdot)$. These two quantities will be replaced by nonparametric estimators. First, for an arbitrary point $x = (x_1,\ldots,x_d)^t$ in the support $\chi$ of $X$, we start by estimating the regression function $m(x,\theta)$ by a local polynomial estimator of degree $p$ (like in Neumeyer and Van Keilegom (2010)), i.e. $\widehat{m}(x,\theta) = \widehat{b}_0(\theta)$ 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 $j = 1,\ldots,d$. Moreover, $h = (h_1,\ldots,h_d)^t$ is a $d$-dimensional bandwidth vector and for $u = (u_1,\ldots,u_d)^t$, $K_1(u)$ is 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. Introduce also the following notation :

$$K_{1h}(u) = \prod_{j=1}^d k_1(u_j/h_j)/h_j \ .$$

Second, the error density function $f_{\varepsilon(\theta)}(y)$ is estimated by the classical kernel estimator of a 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 (which can be different from $k_1$) and $g$ is a bandwidth. Define $k_{2g}(u) = k_2(u/g)/g$. The profile likelihood estimator of $\theta$ is now defined by :

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

The asymptotic properties of this estimator have been established by Linton, Sperlich and Van Keilegom (2008). In their Theorem 4.1, they prove the following asymptotic representation for $\widehat{\theta} - \theta_0$ :

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

where

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

and

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

Also, denote

$$g(X,Y) = \Gamma^{-1}\xi(\theta_0, X, Y) \ . \tag{3}$$

However, as mentioned before, our model and estimation method are slightly different from those considered in Linton, Sperlich and Van Keilegom (2008). In the supplementary material, it will be shown that their Theorem 4.1 continues to hold true in our case, under appropriate regularity conditions. Finally, let (for reasons of simplicity of notation)

$$\widehat{m}(x) = \widehat{m}(x, \widehat{\theta}) \ .$$

2.3 The test statistics

The main idea of the test statistics is to compare the distribution function of the error term $\varepsilon = \Lambda_{\theta_0}(Y) - m(X)$ estimated in a semiparametric way with the distribution function of $\varepsilon$ estimated under $H_0$. That this leads to a valid testing procedure, is shown in the next theorem.

**Theorem 1** *Let $m$ be a continuous function. Then, $H_0$ is valid if and only if the random variables*

$$\Lambda_{\theta_0}(Y) - m(X) \quad \textit{and} \quad \Lambda_{\theta_0}(Y) - m_{\widetilde{\beta}_0}(X)$$

*have the same distribution, where $\widetilde{\beta}_0 = \arg\min_{\beta \in \mathcal{B}} E[(m(X) - m_\beta(X))^2]$.*

The proof is given in the supplementary material. Clearly, when $H_0$ is true, then $\widetilde{\beta}_0 = \beta_0$ where $\beta_0$ is the true value of $\beta$ under $H_0$. Remind that $F_\varepsilon(y) = F_{\varepsilon(\theta_0)}(y) = P(\Lambda_{\theta_0}(Y) - m(X) \leq y)$, and define $F_{\varepsilon_0}(y) = P(\Lambda_{\theta_0}(Y) - m_{\widetilde{\beta}_0}(X) \leq y)$. Next, we explain how to estimate $F_\varepsilon(\cdot)$ and $F_{\varepsilon_0}(\cdot)$ in order to construct the test statistics. First, define

$$\widehat{F}_\varepsilon(y) = n^{-1} \sum_{i=1}^n I(\widehat{\varepsilon}_i \leq y) \ , \tag{4}$$

where $\widehat{\varepsilon}_i = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}(X_i)$ are the semiparametric residuals. Second, estimate $m_{\widetilde{\beta}_0}(x)$ by the least squares method for nonlinear regression, i.e. estimate $m_{\widetilde{\beta}_0}(x)$ by $m_{\widehat{\beta}}(x)$, where $\widehat{\beta}$ is a minimizer over $\beta \in \mathcal{B}$ of the expression

$$S_n(\beta) = n^{-1} \sum_{i=1}^n (\Lambda_{\widehat{\theta}}(Y_i) - m_\beta(X_i))^2 \ . \tag{5}$$

Next, we follow the idea of Härdle and Mammen (1993) and we smooth the function $m_{\widehat{\beta}}(x)$ by a local polynomial estimator of degree $p$, i.e. we define $\widehat{m}_{\widehat{\beta}}(x) = \widehat{c}_0$, where $\widehat{c}_0$ is the first component of the vector $\widehat{c}$, which is the solution of the following local minimization problem :

$$\min_c \sum_{i=1}^n (m_{\widehat{\beta}}(X_i) - P_i(c, x, p))^2 K_1\left(\frac{X_i - x}{h}\right) \ ,$$

where $P_i(c, x, p)$ 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 $j = 1, \ldots, d$. Note that we use here the same $d$-dimensional kernel $K_1$, the same $d$-dimensional bandwidth $h$ and the same order $p$ of the local polynomial as in the local polynomial estimator of the regression function $m(x)$. This is to ensure that these two estimators have the same asymptotic bias under $H_0$. Hence, we obtain the following estimator of the distribution function of $\varepsilon$ under $H_0$ :

$$\widehat{F}_{\varepsilon_0}(y) = n^{-1} \sum_{i=1}^{n} I(\widehat{\varepsilon}_{i0} \leq y) \ , \tag{6}$$

where $\widehat{\varepsilon}_{i0} = \Lambda_{\widehat{\theta}}(Y_i) - \widehat{m}_{\widehat{\beta}}(X_i)$ are the residuals estimated under $H_0$. The test statistics that we will use are Kolmogorov-Smirnov and Cramér-von Mises type statistics defined by

$$T_{KS} = n^{1/2} \sup_{y \in \mathbb{R}} |\widehat{F}_{\varepsilon}(y) - \widehat{F}_{\varepsilon_0}(y)| \quad \text{and} \quad T_{CM} = n \int (\widehat{F}_{\varepsilon}(y) - \widehat{F}_{\varepsilon_0}(y))^2 \, d\widehat{F}_{\varepsilon}(y) \ .$$

Next, to study the power of the test statistics, consider the following local alternative hypothesis :

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

for some fixed function $r \neq 0$. Note that the local alternative $H_{1n}$ only affects the regression function $m(x)$ and not the error distribution.

## 3 Asymptotic results

Before stating the main results of this paper, we 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 $\beta = (\beta_1, \ldots, \beta_q)^t$. The regularity conditions under which the results of this section are valid, can be found in Section 7.

### 3.1 Results under $H_0$

First, under $H_0$, the following theorem states an asymptotic representation for $\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y)$ and gives the limiting distribution of the process $n^{1/2}(\widehat{F}_\varepsilon(\cdot) - \widehat{F}_{\varepsilon_0}(\cdot))$.

**Theorem 2** *Assume (A1)-(A9) and suppose that $H_0$ holds.*

*(i) Then,*

$$\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y) = f_\varepsilon(y)\ n^{-1}\sum_{i=1}^n H(X_i, Y_i, \theta_0, \beta_0) + R_n(y)\ ,$$

*where*

$$\begin{aligned} H(X,Y,\theta,\beta) = \Lambda_\theta(Y) - m(X) - \int \left(\frac{\partial m_\beta(x)}{\partial\beta}\right)^t dF_X(x)\ \eta_\beta(X,Y) \\ -E[(\dot{\Lambda}_\theta(Y))^t]g(X,Y) + \int \left(\frac{\partial m_\beta(x)}{\partial\beta}\right)^t dF_X(x) \\ \Omega^{-1}E\left[\frac{\partial m_\beta(X)}{\partial\beta}(\dot{\Lambda}_\theta(Y))^t\right]g(X,Y)\ , \end{aligned}$$

*where $\sup_{y\in\mathbb{R}}|R_n(y)| = o_P(n^{-1/2})$ and $g(X,Y)$ is defined in (3).*

*(ii) Moreover, the process $n^{1/2}(\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y))$ $(-\infty < y < +\infty)$ converges weakly to $f_\varepsilon(y)W$, where $W$ is a zero mean normal random variable with variance*

$$V(W) = E[H^2(X,Y,\theta_0,\beta_0)]\ .$$

This theorem states that the difference between the two empirical distribution functions factorizes in the error density function and a certain sum of *iid* terms, plus negligeable terms. Note that the second term in this asymptotic expansion is due to the estimation of $\beta$ and the third and the last terms are due to the estimation of $\theta$. If $\beta$ and $\theta$ would be known, then $V(W)$ would simply be equal to $\sigma^2$.

As a consequence, we obtain the following corollary, which gives the limiting distribution of the Kolmogorov-Smirnov and Cramér-von Mises statistics under $H_0$.

**Corollary 1** *Assume (A1)-(A9). Then, under $H_0$,*

$$T_{KS} \xrightarrow{d} \sup_{y\in\mathbb{R}}|f_\varepsilon(y)|\,|W| \qquad \text{and} \qquad T_{CM} \xrightarrow{d} \int f_\varepsilon^2(y)\,dF_\varepsilon(y)\,W^2\ .$$

### 3.2 Results under $H_1$

First, we define $S_0(\beta) = \sigma^2 + E[(m_\beta(X) - m_{\beta_0}(X))^2]$, and

$$\widetilde{S}_{0n}(\beta) = \sigma^2 + E[(m_\beta(X) - m(X))^2] \ , \tag{7}$$

and let $\widetilde{\beta}_{0n}$ be a minimizer over $\beta \in \mathcal{B}$ of $\widetilde{S}_{0n}(\beta)$, which depends on $n$ under $H_{1n}$. Similarly to Section 3.1, but now under $H_{1n}$, the following theorem states an asymptotic representation for $\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y)$ and gives the limiting distribution of the process $n^{1/2}(\widehat{F}_\varepsilon(\cdot) - \widehat{F}_{\varepsilon_0}(\cdot))$.

**Theorem 3** *Assume (A1)-(A10) and suppose that $H_{1n}$ holds.*

*(i) Then,*

$$\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y) = f_\varepsilon(y)\ n^{-1}\sum_{i=1}^n H(X_i, Y_i, \theta_0, \widetilde{\beta}_{0n}) + n^{-1/2} f_\varepsilon(y) b + R_n(y)\ ,$$

*where $\sup_{y\in\mathbb{R}} |R_n(y)| = o_P(n^{-1/2})$, $H(X, Y, \theta, \beta)$ is defined in Theorem 2 and*

$$b = -\int \left(\frac{\partial m_{\beta_0}(x)}{\partial \beta}\right)^t dF_X(x)\ \Omega^{-1} \int r(x)\ \frac{\partial m_{\beta_0}(x)}{\partial \beta}\, dF_X(x) + \int r(x)\, dF_X(x)\,.$$

*(ii) Moreover, the process $n^{1/2}(\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y))$ $(-\infty < y < +\infty)$ converges weakly to $f_\varepsilon(y)(W + b)$, where $W$ is the same normal random variable as in Theorem 2(ii).*

Note that the bias term $f_\varepsilon(y)b$ equals zero under $H_0$, i.e. when $r \equiv 0$. Finally, the following corollary states the limiting distribution of the two test statistics under the local alternative $H_{1n}$.

**Corollary 2** *Assume (A1)-(A10). Then, under $H_{1n}$,*

$$T_{KS} \xrightarrow{d} \sup_{y\in\mathbb{R}} |f_\varepsilon(y)|\,|W + b| \qquad \text{and} \qquad T_{CM} \xrightarrow{d} \int f_\varepsilon^2(y)\, dF_\varepsilon(y)\, (W + b)^2\ .$$

One advantage of our approach is that our tests can detect alternatives at the rate $n^{-1/2}$, which is faster than the rate $n^{-1/2}h^{-d/4}$ obtained by other approaches in the literature, see for example the seminal paper by Härdle and Mammen (1993), or any other paper based on their approach. Note however that there are situations in which the random variable $W$ defined in Theorem 2 is non degenerate, and the bias term $b$ in Theorem 3 is equal to zero. Take e.g. the case where $X$ is uniform on $[-1, 1]$, $r(x) = x$ and we are interested in testing $H_0 : m(x) = \beta x^2$ for all $x$. Then, it is easily seen that $b = 0$, whereas in general $H(X, Y, \theta_0, \beta_0)$ will be a.s. different from zero. Although this example shows that cases can be constructed where the tests have no power under the local alternative $H_{1n}$, there are very many cases where the test is consistent under $H_{1n}$. Similar features have been found in other papers, see e.g. Van Keilegom, González-Manteiga and Sánchez-Sellero (2008) and Pardo-Fernández, Van Keilegom and González-Manteiga (2007), among others. Also, remind that our test is consistent in the sense of Theorem 1.

In order to apply the result of Corollary 1 in practice, we need to estimate the limiting distribution of $T_{KS}$ and $T_{CM}$ by plugging in estimators of $f_\varepsilon$, $\dot{f}_\varepsilon$, $f'_\varepsilon$, $\dot{m}$ and $f_X$. Although this is in principle possible, it is not an easy task, as it requires the introduction of new bandwidths. Therefore, we prefer to approximate the distribution of the test statistics under $H_0$ by using a bootstrap procedure. This will be described in detail in the next section.

## 4 Simulations

In this section, we carry out simulations to evaluate the performance of our proposed tests for small samples. The simulated model is $\Lambda_\theta(Y_i) = 3 + \beta X_i + c(X_i) + \varepsilon_i$, where $\Lambda_\theta$ is the Box-Cox (1964) transformation

$$\Lambda_\theta(y) = \begin{cases} \frac{y^\theta - 1}{\theta}, & \theta \neq 0 \\ \log(y), & \theta = 0. \end{cases}$$

Moreover, $X_1, \ldots, X_n$ are independent, of dimension $d = 1$ and uniformly distributed on [0,1], and $\varepsilon_1, \ldots, \varepsilon_n$ are independent standard normal random variables truncated on [-3,3]. We consider the following null hypothesis :

$$H_0 : m(x) = 3 + \beta x \quad \text{for all } x \ .$$

We perform simulations for three different values of the parameter $\theta$ : $\theta_0 = 0$ 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 $\beta$ is $\beta_0 = 2$. The term $c(x)$ represents different deviations from the null hypothesis and we consider here $c(x) = 5x^2$, $c(x) = 7.5x^2$, $c(x) = 10x^2$, $c(x) = 0.5\exp(x)$, $c(x) = \exp(x)$, $c(x) = 2\exp(x)$, $c(x) = 3\exp(x)$, $c(x) = 0.25\sin(2\pi x)$, $c(x) = 0.5\sin(2\pi x)$, $c(x) = \sin(2\pi x)$ and $c(x) = 1.5\sin(2\pi x)$ for sample size $n = 200$.

Next, we use the Epanechnikov kernel $k_1(x) = k_2(x) = \frac{3}{4}\left(1 - x^2\right) 1_{\{|x| \leq 1\}}$ for both the estimator of the regression function and the density function. For the estimation of $\theta$, $h$, $g$ and $\beta$, we proceed as follows. We maximize the following function with respect to $\theta$ for some optimal values of $h$ and $g$ :

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

where $\widehat{m}(x,\theta,h)$ denotes $\widehat{m}(x,\theta)$ constructed with a bandwidth $h$. For each value of $\theta$, let $h^*(\theta)$ be the bandwidth obtained by least squares cross-validation :

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

where

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

Note that the kernels $k_1$ and $k_2$ and the bandwidths $h$ and $g$ do not satisfy some of the requirements in assumptions (A1) and (A2). However, we believe that these requirements are sufficient but not strictly necessary, and they help to simplify the technical arguments in the proofs in the supplementary material. Also, note that a bandwidth obtained by cross-validation may not be optimal for testing purposes. We therefore also consider the case where the bandwidth $h$ is fixed by the user in order to compare the results obtained with both methods. We choose here $h = 0.1$, $h = 0.15$ and $h = 0.2$. Moreover, we select $g$ by a classical bandwidth selection rule for kernel density estimation. For simplicity, we choose here the normal reference rule, i.e. $\widehat{g}(\theta) = (40\sqrt{\pi})^{1/5} n^{-1/5} \widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$, where $\widehat{\sigma}_{\widehat{\varepsilon}(\theta,h^*(\theta))}$ is the classical estimator of the standard deviation of the error term $\widehat{\varepsilon}(\theta,h^*(\theta)) = \Lambda_\theta(Y) - \widehat{m}(X,\theta,h^*(\theta))$. Consequently, the optimal value of $\theta$, $\widehat{\theta} = \arg\max_\theta l_\theta(h^*(\theta),\widehat{g}(\theta))$, is obtained iteratively with the function *optimize* in R over the interval $[\theta_0 - 2, \theta_0 + 2]$. Finally, to estimate $\beta$, 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 test statistics $T_{KS}$ and $T_{CM}$ will be approximated by means of the following bootstrap procedure. First, we standardize the residuals $\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n$ in order to have mean zero :

$$\breve{\varepsilon}_i = \widehat{\varepsilon}_i - \frac{1}{n}\sum_{k=1}^n \widehat{\varepsilon}_k \quad i = 1, \ldots, n$$

Let $\breve{F}_\varepsilon$ be the empirical distribution of these standardized residuals. Let $\zeta_1^*, \ldots,$ $\zeta_n^*$ be bootstrap samples of the errors drawn with replacement from this distribution. Moreover, let $\xi_1, \ldots, \xi_n$ be independent standard normally distributed random variables and independent from the original sample $\{(X_1, Y_1), \ldots,$ $(X_n, Y_n)\}$. We define the bootstrap errors by $\varepsilon_i^* = \zeta_i^* + b_n \xi_i$, $i = 1, \ldots, n$, where $b_n$ is some small bandwidth. We choose here $b_n = 0.1$. Finally, we can see that $\varepsilon_i^*$ has a smooth distribution function given by

$$\widetilde{F}_\varepsilon(y) = \frac{1}{n}\sum_{j=1}^n \Phi\left(\frac{y - \breve{\varepsilon}_j}{b_n}\right)$$

**Table 1** Percentage of rejection under the null hypothesis (nominal level 5%) when the bandwidth $h$ is obtained by cross-validation and for samples of size $n = 200$.

| | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|
| $c(x)$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| 0 | 5.2 | 7.8 | 5.8 | 8.0 | 6.6 | 7.4 |

**Table 2** Percentage of rejection under the null hypothesis (nominal level 5%) for fixed bandwidth $h$ and for samples of size $n = 200$.

| | | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|---|
| $c(x)$ | $h$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| 0 | 0.1 | 2.2 | 4.2 | 4.4 | 6.2 | 6.8 | 7.0 |
| 0 | 0.15 | 3.8 | 7.0 | 7.2 | 9.0 | 7.0 | 8.2 |
| 0 | 0.2 | 3.6 | 6.4 | 6.6 | 9.0 | 7.0 | 7.4 |

where $\Phi$ is the standard normal distribution function. Note that we have to work with a smoothed distribution in the bootstrap procedure, because the asymptotic representation of $\widehat{F}_\varepsilon(y) - \widehat{F}_{\varepsilon_0}(y)$ given in Theorem 2 involves the density $f_\varepsilon(y)$ (see e.g. Silverman and Young (1987) or Neumeyer (2009) for similar bootstrap procedures).

The bootstrap procedure can now be described as follows. For fixed $B$ and for $b = 1, \ldots, B$ :

1. Let $\varepsilon^*_{1b}, \ldots, \varepsilon^*_{nb}$ be independent random errors drawn from $\widetilde{F}_\varepsilon$, and let $X^*_{ib} = X_i$ $(i = 1, \ldots, n)$.
2. Define new responses $Y^*_{ib} = \Lambda^{-1}_{\widehat{\theta}}(m_{\widehat{\beta}}(X^*_{ib}) + \varepsilon^*_{ib})$, $i = 1, \ldots, n$, obtained under the null hypothesis.
3. Let $T^*_{KS,b}$ and $T^*_{CM,b}$ be the test statistics obtained from the bootstrap sample $(X^*_{ib}, Y^*_{ib})$, $i = 1, \ldots, n$.

Then, the $[(1-\alpha)B]$-th order statistic of $T^*_{KS,1}, \ldots, T^*_{KS,B}$ approximates the $(1-\alpha)$-th quantile of the distribution of $T_{KS}$, and similarly for $T_{CM}$. 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. In our simulations, we take $B = 250$.

Tables 1 and 3 show respectively the percentage of rejection under the null hypothesis and under the different deviations $c(x)$ we have introduced above when the bandwidth $h$ is obtained by cross-validation. Tables 2 and 4 show respectively the percentage of rejection under the null hypothesis and under different deviations $c(x)$ when the bandwidth $h$ is fixed by the user. These percentages of rejection are obtained with the test statistics $T_{KS}$ and $T_{CM}$ for 500 samples. The nominal level is 5%.

We see that the different estimations of the nominal level under $H_0$ are globally good and we note that the results for $T_{KS}$ are slightly better. Indeed, the percentages of rejection given by $T_{CM}$ are a little bit too high. Nevertheless,

**Table 3** Percentage of rejection under the alternative hypothesis (nominal level 5%) when the bandwidth $h$ is obtained by cross-validation and for samples of size $n = 200$.

| | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|
| $c(x)$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| $5x^2$ | 100.0 | 100.0 | 57.4 | 60.8 | 13.4 | 14.4 |
| $7.5x^2$ | 99.6 | 99.6 | 96.4 | 96.4 | 76.0 | 77.2 |
| $10x^2$ | 99.0 | 99.0 | 96.6 | 96.0 | 87.2 | 86.4 |
| $0.5\exp(x)$ | 30.8 | 38.6 | 17.2 | 19.0 | 10.0 | 12.6 |
| $\exp(x)$ | 86.8 | 92.0 | 36.4 | 39.8 | 20.2 | 22.6 |
| $2\exp(x)$ | 99.8 | 99.8 | 68.0 | 69.2 | 45.8 | 46.6 |
| $3\exp(x)$ | 98.6 | 98.4 | 72.8 | 72.0 | 54.8 | 56.2 |
| $0.25\sin(2\pi x)$ | 17.0 | 19.6 | 11.2 | 13.0 | 11.8 | 14.2 |
| $0.5\sin(2\pi x)$ | 29.2 | 32.4 | 19.8 | 22.0 | 18.8 | 20.8 |
| $\sin(2\pi x)$ | 49.6 | 50.6 | 25.4 | 27.4 | 23.8 | 26.0 |
| $1.5\sin(2\pi x)$ | 68.6 | 66.6 | 26.8 | 27.8 | 19.2 | 19.4 |

**Table 4** Percentage of rejection under the alternative hypothesis (nominal level 5%) for fixed bandwidth $h$ and for samples of size $n = 200$.

| | | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|---|
| $c(x)$ | $h$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| $5x^2$ | 0.1 | 100.0 | 100.0 | 57.8 | 59.8 | 9.6 | 9.4 |
| $5x^2$ | 0.15 | 100.0 | 100.0 | 53.2 | 53.8 | 12.6 | 11.4 |
| $5x^2$ | 0.2 | 100.0 | 100.0 | 41.6 | 43.6 | 8.8 | 8.8 |
| $7.5x^2$ | 0.1 | 100.0 | 100.0 | 98.0 | 98.2 | 79.2 | 79.8 |
| $7.5x^2$ | 0.15 | 100.0 | 100.0 | 96.2 | 96.4 | 69.6 | 73.0 |
| $7.5x^2$ | 0.2 | 100.0 | 100.0 | 85.8 | 87.8 | 50.4 | 53.2 |
| $10x^2$ | 0.1 | 100.0 | 100.0 | 99.2 | 99.2 | 91.6 | 91.4 |
| $10x^2$ | 0.15 | 100.0 | 100.0 | 99.4 | 99.4 | 91.6 | 91.8 |
| $10x^2$ | 0.2 | 100.0 | 100.0 | 96.6 | 97.0 | 75.8 | 78.0 |

we expect that the procedure can be finetuned so as to obtain more accurate nominal levels. For instance, increasing the number of bootstrap iterations (currently $B = 250$) will improve the level and it is also expected that the selection of the bandwidths can be further improved. The testing procedure relies on a complicated estimation procedure, involving many parameters and functions, and this definitely influences the precision of the nominal level. Next, under the alternative, the power is largest for $\theta_0 = 0$, followed by $\theta_0 = 0.5$ and then $\theta_0 = 1$. This result seems logical, because Linton, Sperlich and Van Keilegom (2008) showed that the mean squared error of the profile likelihood estimator of $\theta$ is largest for $\theta_0 = 1$, followed by $\theta_0 = 0.5$ and then $\theta_0 = 0$. Finally, we see that the percentages of rejection are slightly larger for $T_{CM}$ than for $T_{KS}$, which is in line with what happens under $H_0$.

Moreover, under the null hypothesis and when the bandwidth $h$ is fixed, the best results are in majority obtained for $h = 0.1$, both for $T_{KS}$ and $T_{CM}$. Under the alternative, the power decreases when $h$ increases and especially for $\theta_0 = 0.5$ and $\theta_0 = 1$, which suggests that among the three tested values of the bandwidth, $h = 0.1$ is the most adapted value for this test. Finally, we

**Table 5** Percentage of rejection under the null and the alternative hypothesis (nominal level 5%) for samples of size $n = 200$.

| | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|
| $c(x_1, x_2)$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| 0 | 0.8 | 0.4 | 2.0 | 2.0 | 5.2 | 5.4 |
| $2.5x_1x_2$ | 20.2 | 7.4 | 5.4 | 4.6 | 5.0 | 5.6 |
| $5x_1x_2$ | 52.6 | 39.8 | 11.4 | 10.8 | 6.2 | 6.6 |
| $7.5x_1x_2$ | 66.4 | 67.0 | 24.2 | 26.2 | 11.6 | 11.4 |
| $10x_1x_2$ | 67.6 | 74.2 | 41.0 | 41.6 | 17.8 | 18.6 |
| $2(x_1+1)/(x_2+1)$ | 94.8 | 96.0 | 9.0 | 10.0 | 6.8 | 7.0 |
| $3(x_1+1)/(x_2+1)$ | 96.6 | 97.6 | 34.6 | 38.4 | 11.4 | 13.2 |
| $4(x_1+1)/(x_2+1)$ | 97.0 | 97.8 | 50.6 | 53.8 | 21.2 | 23.0 |
| $5(x_1+1)/(x_2+1)$ | 94.6 | 95.0 | 65.8 | 67.8 | 33.4 | 35.4 |

notice that sometimes the power is better when the bandwidth $h$ is estimated by cross-validation, for example when $c(x) = 5x^2$, and sometimes the power is better when $h$ is fixed, for example when $c(x) = 7.5x^2$ and $h = 0.1$.

We finish this section by presenting the results of a simulation study that shows the performance of our test in higher dimensions. We consider $d = 2$, $q = 2$ and the simulated model is $\Lambda_\theta(Y_i) = 3+\beta_1X_{1i}+\beta_2X_{2i}+c(X_{1i}, X_{2i})+\varepsilon_i$, where $\Lambda_\theta$ is again the Box-Cox (1964) transformation. 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 standard normal random variables truncated on [-3,3]. We consider the following null hypothesis :

$$H_0 : m(x) = 3 + \beta_1 x_1 + \beta_2 x_2 \quad \text{for all } x \ .$$

We perform simulations for the same three different values of the parameter $\theta$ as before : $\theta_0 = 0$, $\theta_0 = 0.5$ and $\theta_0 = 1$. The true value of the parameter $\beta = (\beta_1, \beta_2)$ is $(\beta_{10}, \beta_{20}) = (3, 5)$. The term $c(x_1, x_2)$ represents different deviations from the null hypothesis and we consider here $c(x_1, x_2) = 2.5x_1x_2$, $c(x_1, x_2) = 5x_1x_2$, $c(x_1, x_2) = 7.5x_1x_2$, $c(x_1, x_2) = 10x_1x_2$, $c(x_1, x_2) = 2\frac{x_1+1}{x_2+1}$, $c(x_1, x_2) = 3\frac{x_1+1}{x_2+1}$, $c(x_1, x_2) = 4\frac{x_1+1}{x_2+1}$, $c(x_1, x_2) = 5\frac{x_1+1}{x_2+1}$, for samples of size $n = 200$. We use here the product of two Epanechnikov kernels for the estimator of the regression function and the Epanechnikov kernel for the estimator of the density function. For the estimation of $\theta$, $\beta$, $h$ and $g$ and the bootstrap procedure, we proceed exactly as before. Note that we estimate $h$ by cross validation. Table 5 shows the percentage of rejection under the null hypothesis and under the different deviations $c(x_1, x_2)$ we have introduced above.

Table 5 shows that the nominal level is in general too low, especially for $\theta_0 = 0$ and $\theta_0 = 0.5$. This bad behavior can be explained by the poor quality of the estimation of the function $m(\cdot)$. Since we are estimating $m(\cdot)$ in two dimensions in a completely nonparametric way, the method is clearly suffering from curse-of-dimensionality problems, which are very common in nonparametric regression. So, a sample size of $n = 200$ is not sufficient for the method to work well. For that reason, we perform an additional simulation under $H_0$ to check whether the problem disappears for a larger sample size. Table 6 shows

**Table 6** Percentage of rejection under the null hypothesis (nominal level 5%) for samples of size $n = 500$.

| | $\theta_0 = 0$ | | $\theta_0 = 0.5$ | | $\theta_0 = 1$ | |
|---|---|---|---|---|---|---|
| $c(x_1, x_2)$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ | $T_{KS}$ | $T_{CM}$ |
| 0 | 2.4 | 2.0 | 6.2 | 6.8 | 6.4 | 6.4 |

**Table 7** p-values of the $l$th degree polynomial fit for the sunlight data.

| | $l = 1$ | $l = 2$ | $l = 3$ | $l = 4$ |
|---|---|---|---|---|
| $T_{KS}$ | 0.333 | 0.734 | 0.737 | 0.475 |
| $T_{CM}$ | 0.208 | 0.748 | 0.666 | 0.549 |

the percentage of rejection under the null hypothesis for sample size $n = 500$. We see in this table that the nominal level is now much better approximated even if the precision can still be improved for the same reasons as explained in the analysis of the results in dimension 1.

## 5 Application

We apply our testing procedure to a data set composed of 355 observations resulting from an experiment on the scattering of sunlight in the atmosphere (see Bellver (1987)). The data can be found in Cleveland (1993). The response $Y$ is the scattering angle at which the polarization of sunlight vanishes, called the Babinet point. Note that the response is positive, which justifies the use of a Box-Cox transformation. Moreover, the covariate $X$ is the cube root of a measure of particulate concentration in the atmosphere and we standardize it.

This data set has already been analyzed, but without transformation of the response variable, in different articles, like in Hart (1997), in Zhang (2003) and in Van Keilegom, González-Manteiga and Sánchez-Sellero (2008). A test for linearity of the underlying regression function was realized in Hart (1997), while different tests for $l$th degree polynomial regression ($l = 1, 2, 3, 4$) were realized in Zhang (2003) and in Van Keilegom, González-Manteiga and Sánchez-Sellero (2008), both with their own testing procedure.

Here, a Box-Cox transformation of the response variable is considered and we check the goodness-of-fit of the $l$th degree polynomial regression ($l = 1, 2, 3, 4$) by using the Kolmogorov-Smirnov and the Cramér-von Mises test statistics defined in this paper. The distributions and p-values of these two test statistics are approximated by bootstrap on the basis of 1000 replicates. The results are given in Table 7.

First, note that the profile likelihood estimator of $\theta$ is equal to $\widehat{\theta} = 1.9428$. This implies that we transform the response variable $Y$ by taking approximatively its square. Table 7 indicates that there is no evidence against a polynomial fit of order $l = 2, 3, 4$, similarly as in Van Keilegom, González-Manteiga and Sánchez-Sellero (2008). Moreover, there is also no evidence against a linear fit, which is different from the conclusions in Hart (1997), Zhang (2003) and Van Keilegom, González-Manteiga and Sánchez-Sellero (2008). This can be explained by the transformation realized on the response variable $Y$, which has an important impact on the regression function.

## 6 Conclusions and future research

In this paper, we constructed a test for the parametric form of the regression function in a semiparametric transformation model. The transformation of the dependent variable in this model was supposed to belong to some parametric family of strictly increasing functions. We defined a Kolmogorov-Smirnov and a Cramér-von Mises test statistic, where the main idea was to compare the distribution of the error term estimated in a semiparametric way to the one estimated under $H_0$. We established the limiting distribution of these two test statistics under the null hypothesis and under a local alternative. We evaluated the performance of our test by means of some simulations and we applied our method on a real data set.

It would be interesting to extend the paper of Pardo-Fernández, Van Keilegom and González-Manteiga (2007) by constructing a test for the equality of regression curves in the case of semiparametric transformation models. Another possibility of future research could be the extension of this paper to the case of censored data. Finally, the extension of the methods of Härdle and Mammen (1993) and Stute (1997) to the context of transformation models would be an useful alternative for the tests developed in this paper, and it would then be informative to know under which model conditions which test behaves best.

## 7 Technical assumptions

In this section, we introduce a number of notations and state the assumptions under which the main results of this paper are valid. For $0 < \alpha < \delta/2$, where $\delta$ is defined as in condition $(A2)$ (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 $\varepsilon$, the transformation $\Lambda_\theta$ 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 $k = 1, \ldots, q_2 - 1$ 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 $l = 0, \ldots, d-1$ and $k_2$ is twice continuously differentiable.

(A2) $h_l$ (for $l = 1, \ldots, d$) 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 $\chi$ 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) (i) The error term $\varepsilon = \Lambda_{\theta_0}(Y) - m(X)$ has finite fourth moment and is independent of $X$.
(ii) The distribution function $F_{\varepsilon(\theta)|X}(y|x)$ is three times continuously differentiable with respect to $y$ and $\theta$, and

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

for all $i$ and $j$ such that $0 \leq i + j \leq 2$ 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 $\theta$, and there exists $\alpha > 0$ such that :

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

for all $\theta \in \Theta$ and all $i$ and $j$ such that $0 \leq i + j \leq 3$.
(ii) $\sup_{x \in \chi} ||E[\dot{\Lambda}_{\theta_0}^4(Y)|X = x]|| < \infty$.
(iii) $\sup_{\theta \in \Theta, x \in \chi} ||E[\dot{\Lambda}_\theta(Y)|X = x]|| < \infty$.
(iv) The density function of $(\dot{\Lambda}_\theta(Y), X)$ exists and is continuous for all $\theta \in \Theta$.

(A6) (i) $\mathcal{B}$ is a compact subset of $\mathbb{R}^q$ and $\beta_0$ is an interior point of $\mathcal{B}$.
(ii) All partial derivatives of $m_\beta(x)$ with respect to the components of $x$ and $\beta$ of order 0, 1, 2 and 3 exist and are continuous in $(x, \beta)$ for all $x$ and $\beta$.
(iii) For all $\varepsilon > 0$ :

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

(iv) $\Omega$ 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 $\theta_0$ and all derivatives up to order $p + 2$ are bounded, uniformly in $(x, \theta)$ in $\chi \times N(\theta_0)$.

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

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

(ii) The matrix $\Gamma$ is of full rank.

(A9) $\Lambda_{\theta_0}(\alpha) = a$ 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.

(A10) $E(r^2(X)) < \infty$ and $r(x)$ is twice continuously differentiable for all $x$.

Note that condition (A9) is needed for identifying the model (see Vanhems and Van Keilegom (2013)), because

- The two conditions $\Lambda_{\theta_0}(\alpha) = a$ and $\Lambda_{\theta_0}(\beta) = b$ for some $\alpha < \beta$ and $a < b$ are needed to fix the location and the scale of the model. More precisely, consider the following transformation $\overline{\Lambda}_{\theta_0}(Y) = c\Lambda_{\theta_0}(Y) + d$ such that $\overline{\Lambda}_{\theta_0}(\alpha) = a$, $\overline{\Lambda}_{\theta_0}(\beta) = b$ and where $c$ and $d$ are some constants. Then, $\overline{\Lambda}_{\theta_0}(\alpha) = c\Lambda_{\theta_0}(\alpha) + d = a$ and $\overline{\Lambda}_{\theta_0}(\beta) = c\Lambda_{\theta_0}(\beta) + d = b$, which implies that $c \cdot a + d = a$ and $c \cdot b + d = b$. We conclude that $c = 1$ and $d = 0$, and then $\overline{\Lambda} = \Lambda$.
- Suppose that $\frac{\partial m}{\partial x}(x) = 0$ for all $x$, then $m(x) = \gamma$ for some constant $\gamma$, and the semiparametric transformation model becomes $\Lambda_{\theta_0}(Y) = \gamma + \epsilon$ with $\epsilon$ independent of $X$. Next, consider another transformation $\widetilde{\Lambda}_{\theta_0}(Y)$ of $Y$. Then $\widetilde{\Lambda}_{\theta_0}(Y) = \widetilde{\Lambda}_{\theta_0}\Lambda_{\theta_0}^{-1}\Lambda_{\theta_0}(Y) = \widetilde{\Lambda}_{\theta_0}\Lambda_{\theta_0}^{-1}(\gamma+\epsilon) = Z - E(Z) + E(Z)$ where $Z = \widetilde{\Lambda}_{\theta_0}\Lambda_{\theta_0}^{-1}(\gamma+\epsilon)$. Finally $\widetilde{\Lambda}_{\theta_0}(Y) = \widetilde{\gamma} + \widetilde{\epsilon}$ where $\widetilde{\gamma} = E(Z)$, $\widetilde{\epsilon} = Z - E(Z)$, $\widetilde{\epsilon}$ has zero mean and $\widetilde{\epsilon}$ is independent of $X$. The model is thus not identified.

Moreover, conditions (A6) and (A10) come from Van Keilegom, González-Manteiga and Sánchez Sellero (2008), conditions (A4)(ii), (A5)(i) and (A8) come from Linton, Sperlich and Van Keilegom (2008), condition (A3)(ii) come from Neumeyer and Van Keilegom (2010) and conditions (A4)(i), (A5)(ii) and (A5)(iv) come from Heuchenne, Samb and Van Keilegom (2014). Finally, 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) and partially from Neumeyer and Van Keilegom (2010).

**Acknowledgements** B. Colling and I. Van Keilegom are financially 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'. I. Van Keilegom is also financially supported by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650. We also thank the two referees for their constructive remarks and suggestions.

## References

1. Akritas MG, Van Keilegom I (2001) Nonparametric estimation of the residual distribution. Scand J Stat 28:549-568
2. Bellver C (1987) Influence of particulate pollution on the positions of neutral points in the sky in Seville (Spain). Atmos Environ 21:699-702
3. Bickel PJ, Doksum K (1981) An analysis of transformations revisited. J Am Stat Assoc 76:296-311
4. Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc 26:211-252

5. Carroll RJ, Ruppert D (1988) Transformation and Weighting in Regression. Chapman and Hall, New-York
6. Cleveland WS (1993) Visualizing Data. Hobart Press, Summit
7. Colling B, Heuchenne C, Samb R, Van Keilegom I (2013) Estimation of the error density in a semiparametric transformation model. Ann Inst Stat Math (in press)
8. Dette H, Neumeyer N, Van Keilegom I (2007) A new test for the parametric form of the variance function in nonparametric regression. J R Stat Soc Ser B 69:903-917
9. González-Manteiga W, Crujeiras R (2013) An updated review of goodness-of-fit tests for regression models (with discussion). TEST 22:361-447
10. Härdle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21:1926-1947
11. Hart JD (1997) Nonparametric Smoothing and Lack-of-fit Tests. Springer, New-York
12. Heuchenne C, Samb R, Van Keilegom I (2014) Estimating the residual distribution in semiparametric transformation models (submitted, and available on http://www.uclouvain.be/en-369695.html, DP2014/11)
13. Horowitz JL (1996) Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Econometrica 64:103-137
14. Horowitz JL (2001) Nonparametric estimation of a generalized additive model with an unknown link function. Econometrica 69:499-513
15. Jacho-Chavez D, Lewbel A, Linton O (2008) Identification and nonparametric estimation of a transformation additively separable model. Technical report
16. Linton O, Sperlich S, Van Keilegom I (2008) Estimation of a semiparametric transformation model. Ann Stat 36:686-718
17. Neumeyer N (2009) Smooth residual bootstrap for empirical processes of non-parametric regression residuals. Scand J Stat 36:204-228
18. Neumeyer N, Van Keilegom I (2010) Estimating the error distribution in nonparametric multiple regression with applications to model testing. J Multivar Anal 101:1067-1078
19. Pardo-Fernández JC, Van Keilegom I, González-Manteiga W (2007) Testing for the equality of $k$ regression curves. Stat Sin 17:1115-1137
20. Sakia RM (1992) The Box-Cox transformation technique : a review. The Statistician 41:169-178
21. Silverman BW, Young GA (1987) The bootstrap : to smooth or not to smooth. Biometrika 74:469-479
22. Stute W (1997) Nonparametric model checks for regression. Ann Stat 25:613-641
23. Van Keilegom I, González-Manteiga W, Sánchez Sellero C (2008) Goodness of fit tests in parametric regression based on the estimation of the error distribution. TEST 17:401-415
24. Vanhems A, Van Keilegom I (2013) Semiparametric transformation model with endogeneity : a control function approach (submitted, and available on http://www.uclouvain.be/en-369695.html, DP2013/18).
25. Zellner A, Revankar NS (1969) Generalized production functions. Rev Econ Stud 36:241-250
26. Zhang CM (2003) Adaptative tests of regression functions via multiscale generalized likelihood ratio. Can J Stat 31:151-171