



## Goodness-of-fit tests for the error distribution in nonparametric regression

Cédric Heuchenne<sup>a,b,\*</sup>, Ingrid Van Keilegom<sup>b</sup>

<sup>a</sup> QuantOM,<sup>1</sup> HEC-Management School of University of Liège, Belgium

<sup>b</sup> Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium

### ARTICLE INFO

#### Article history:

Received 15 July 2009

Received in revised form 12 February 2010

Accepted 13 February 2010

Available online 23 February 2010

#### Keywords:

Bandwidth selection

Bootstrap

Error distribution

Goodness-of-fit tests

Local polynomial estimation

Nonparametric regression

### ABSTRACT

Suppose the random vector  $(X, Y)$  satisfies the regression model  $Y = m(X) + \sigma(X)\varepsilon$ , where  $m(\cdot)$  is the conditional mean,  $\sigma^2(\cdot)$  is the conditional variance, and  $\varepsilon$  is independent of  $X$ . The covariate  $X$  is  $d$ -dimensional ( $d \geq 1$ ), the response  $Y$  is one-dimensional, and  $m$  and  $\sigma$  are unknown but smooth functions. Goodness-of-fit tests for the parametric form of the error distribution are studied under this model, without assuming any parametric form for  $m$  or  $\sigma$ . The proposed tests are based on the difference between a nonparametric estimator of the error distribution and an estimator obtained under the null hypothesis of a parametric model. The large sample properties of the proposed test statistics are obtained, as well as those of the estimator of the parameter vector under the null hypothesis. Finally, the finite sample behavior of the proposed statistics, and the selection of the bandwidths for estimating  $m$  and  $\sigma$  are extensively studied via simulations.

© 2010 Published by Elsevier B.V.

### 1. Introduction

Suppose the  $d$ -dimensional random vector  $X$  and the random variable  $Y$  satisfy the following heteroscedastic regression model:

$$Y = m(X) + \sigma(X)\varepsilon, \quad (1.1)$$

where  $\varepsilon$  is independent of the random vector  $X$ ,  $E(\varepsilon) = 0$  and  $\text{Var}(\varepsilon) = 1$ . Hence,  $m(X) = E[Y|X]$  and  $\sigma^2(X) = \text{Var}[Y|X]$ .

In the literature, many papers have been devoted to testing the form of  $m(\cdot)$  and  $\sigma(\cdot)$ . Goodness-of-fit tests for  $m(\cdot)$  were studied by e.g. Härdle and Mammen (1993), Stute (1997), Dette and Munk (1998), Alcalá et al. (1999), Dette (1999), Fan et al. (2001), Deschepper et al. (2006) and Van Keilegom et al. (2008), among many others. Testing parametric models for  $\sigma(\cdot)$  has also been investigated, see e.g., among others, Dette et al. (2007). In the above papers the form of the error distribution is left unspecified.

In this paper, we are interested in testing hypotheses concerning the form of this error distribution, without making any assumption regarding the form of the regression function  $m(\cdot)$  and the variance function  $\sigma^2(\cdot)$ , except for smoothness assumptions. Knowing that  $\varepsilon$  follows a certain parametric distribution offers important advantages when doing inference for the functions  $m(\cdot)$  and  $\sigma^2(\cdot)$ .

Consider the parametric class  $\mathcal{F} = \{F_{\varepsilon\theta} : \theta \in \Theta\}$  of distribution functions, where  $\Theta$  is a compact subset of  $\mathbb{R}^{\kappa}$  and  $\kappa$  is a positive integer. We denote the true value of  $\theta$  by  $\theta_0$ , when  $H_0$  is true, and the distribution of the error  $\varepsilon$  by  $F_{\varepsilon}(y) = P(\varepsilon \leq y)$ .

\* Corresponding address: HEC-Management School of University of Liège, Boulevard du Rectorat 7, Building B31, Office 2.53, B-4000 Liège, Belgium. Tel.: +32 43662720; fax: +32 43662767.

E-mail address: [C.Heuchenne@ulg.ac.be](mailto:C.Heuchenne@ulg.ac.be) (C. Heuchenne).

<sup>1</sup> Centre for Quantitative Methods and Operations Management.

The aim of this paper is to test the hypothesis

$$H_0 : F_\varepsilon \in \mathcal{F} \quad \text{versus} \quad H_1 : F_\varepsilon \notin \mathcal{F}. \quad (1.2)$$

This testing problem has been studied by Jiménez Gamero et al. (2005) for linear regression models and by Neumeyer et al. (2006) for linear and nonparametric regression models. See also Huskova and Meintanis (2007, in press), where this problem is considered for nonparametric models using an approach based on the comparison of characteristic functions, and Mora and Pérez-Alonso (2009) for an approach based on martingale transformations. In Neumeyer et al. (2006), the authors work with a generic estimator for  $\theta_0$  that satisfies a certain asymptotic representation, and they obtain the asymptotic theory for their proposed test statistic under the assumption that such an estimator of  $\theta_0$  exists (see their assumption B5). In this paper, we work with a specific estimator for  $\theta_0$ , obtained by using a maximum likelihood approach, and we work out in detail the asymptotic properties of this estimator, and of the proposed test statistics. Another important difference between both procedures lies in the fact that heteroscedastic models are considered in the present paper. Hence, this leads to a local estimator of the variance function and to bandwidth selection procedures that depend on this local estimator. These procedures can clearly be adapted to the homoscedastic case by replacing the local estimator of the variance function by a global estimator. The bandwidth could be chosen by (2.4) and the obtained results should in that case be very similar to those in Neumeyer et al. (2006). Besides the fact that our methodology extends to the heteroscedastic case, the multiple regression case is also investigated and a formal study of the bandwidth selection procedure is achieved.

The test statistics we propose in this paper are based on a Kolmogorov–Smirnov and a Cramér–von Mises distance between an estimator of the error distribution obtained under  $H_0$  and a completely nonparametric estimator. Under the null hypothesis  $H_0$ , we show that the estimator of  $\theta_0$  and the test statistics reach the same rate of convergence as in the usual case where  $m(\cdot)$  and  $\sigma(\cdot)$  are parametric functions. The asymptotic results are largely based on the work of Neumeyer and Van Keilegom (2010) regarding the estimation of the error distribution under model (1.1), and on the results obtained by Chen et al. (2003) regarding inference for general semiparametric models involving non-smooth criterion functions.

In practice, the power of the tests are somewhat sensitive to the choice of the bandwidths used to estimate  $m(\cdot)$  and  $\sigma(\cdot)$ . We therefore study six different procedures to select these bandwidths. As it turns out, the power seems to be higher when loss functions using (a cross validation version of) the residuals themselves are involved.

The paper is organized as follows. In the next section, the estimator of  $\theta_0$  and the test statistic are described in detail. Section 3 summarizes the main asymptotic results including the asymptotic normality of the estimator of  $\theta_0$  and the weak convergence of the proposed test statistics under  $H_0$ . In Section 4, we study the finite sample behavior of the test procedure and the selection of the smoothing parameters for  $m(\cdot)$  and  $\sigma(\cdot)$  through extended simulations, while the Appendix contains the assumptions and the proofs of the asymptotic results of Section 3.

## 2. Description of the method

Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be an i.i.d. random sample generated from model (1.1), where the components of  $X_i$  are denoted by  $(X_{i1}, \dots, X_{id})(i = 1, \dots, n)$ . The distributions of  $\varepsilon$  and  $X$  are assumed to be absolutely continuous (with respect to the Lebesgue measure). We denote them by  $F_\varepsilon$  and  $F_X$  respectively, and their probability density functions by  $f_\varepsilon$  and  $f_X$ . We start by estimating the regression function  $m(x)$  and the variance function  $\sigma^2(x)$  at an arbitrary point  $x = (x_1, \dots, x_d)$  in the support  $R_X$  of  $X$ , which we suppose to be a compact subset of  $\mathbb{R}^d$ . We estimate  $m(x)$  by a local polynomial estimator of degree  $p$ , i.e.  $\hat{m}(x) = \hat{\beta}_0$ , where  $\hat{\beta}_0$  is the first component of  $\hat{\beta}$ , which is the solution of the local minimization problem

$$\min_{\beta} \sum_{i=1}^n \{Y_i - P_i(\beta, x, p)\}^2 K_h(X_i - x), \quad (2.1)$$

where  $P_i(\beta, x, p)$  is a polynomial of order  $p$  built up with all  $0 \leq s \leq p$  products of factors of the form  $X_{ij} - x_j (j = 1, \dots, d)$ , and  $\beta$  is the vector of all coefficients of this polynomial. Here, for  $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ ,  $K(u) = \prod_{j=1}^d k(u_j)$  is a  $d$ -dimensional product kernel,  $k$  is a univariate kernel function,  $h = (h_1, \dots, h_d)$  is a  $d$ -dimensional bandwidth vector converging to zero when  $n$  tends to infinity and  $K_h(u) = \prod_{j=1}^d k(u_j/h_j)/h_j$ . In the same way,  $\hat{\sigma}^2(x) = \hat{\gamma}_0$  is the first component of  $\hat{\gamma}$ , which is the solution of the local minimization problem

$$\min_{\gamma} \sum_{i=1}^n \{(Y_i - \hat{m}(X_i))^2 - P_i(\gamma, x, q)\}^2 K_g(X_i - x), \quad (2.2)$$

where  $\hat{m}(X_i)$ ,  $i = 1, \dots, n$ , is a local polynomial estimator obtained from (2.1) (except that  $h$  is replaced by  $l$ , which will be specified in Remark 2.1 below) and  $P_i(\gamma, x, q)$ ,  $\gamma$ ,  $K_g(u)$  and  $g = (g_1, \dots, g_d)$  are defined in a similar way as  $P_i(\beta, x, p)$ ,  $\beta$ ,  $K_h(u)$  and  $h$ . An estimator for  $\sigma(x)$ ,  $\hat{\sigma}(x)$ , will be simply obtained by taking the square root of  $\hat{\gamma}_0$ .

The nonparametric residuals can then be introduced into the likelihood function which will provide a vector of parameter estimators  $\theta_n$  for  $\theta_0$  by solving the maximization problem

$$\max_{\theta \in \Theta} \sum_{i=1}^n \log f_{\varepsilon\theta}(\hat{\varepsilon}_i), \quad (2.3)$$

where  $\hat{\varepsilon}_i = (Y_i - \hat{m}(X_i))/\hat{\sigma}(X_i)$  ( $i = 1, \dots, n$ ), where  $f_{\varepsilon\theta}(y) = \frac{d}{dy} F_{\varepsilon\theta}(y)$ .

**Remark 2.1** (*Choice of the Smoothing Parameters*). The objective is to provide an easy and data-driven way to select the smoothing parameters in (2.1) and (2.2). To this end, we propose to compare six procedures in Section 4. Four are based on different least squares cross-validation ideas and two on maximum likelihood cross-validation ideas. More precisely, let

$$h_n = \arg \min_h \sum_{j=1}^n (Y_j - \hat{m}_{h,-j}(X_j))^2, \tag{2.4}$$

$$g_{n1} = \arg \min_g \sum_{j=1}^n \{ (Y_j - \hat{m}_{g,-j}(X_j))^2 - \hat{\sigma}_{g,-j}^2(X_j) \}^2, \tag{2.5}$$

and

$$g_{n2} = \arg \min_g \sum_{j=1}^n \{ (Y_j - \hat{m}_{h_n}(X_j))^2 - \hat{\sigma}_{g,h,-j}^2(X_j) \}^2. \tag{2.6}$$

Here,  $\hat{m}_{l,-j}(X_j)$  is a local polynomial estimator obtained by an expression of the type (2.1) for  $x = X_j$  and  $h$  replaced by  $l = g$  or  $h$ , but based on a sample for which the  $j$ th data point has been removed. Moreover,  $\hat{\sigma}_{g,-j}^2(X_j)$  is a local polynomial estimator obtained from an expression of the type (2.2) for  $x = X_j$ , but based on the couples  $(X_i, (Y_i - \hat{m}_{g,-i}(X_i))^2)$  ( $i = 1, \dots, j-1, j+1, \dots, n$ ). Similarly,  $\hat{\sigma}_{g,h,-j}^2(X_j)$  is obtained from the pairs  $(X_i, (Y_i - \hat{m}_{h_n}(X_i))^2)$  ( $i = 1, \dots, j-1, j+1, \dots, n$ ). The two first procedures that we will consider in detail in Section 4 are based on choosing  $h_n$  for  $h$ ,  $g_{n1}$  (resp.  $h_n$ ) for  $l$  and  $g_{n1}$  (resp.  $g_{n2}$ ) for  $g$ .

An alternative idea for choosing  $h$ ,  $g$  and  $l$  is to transform (2.3) in a joint maximization problem over  $\theta$ ,  $h$ ,  $g$  and  $l$ . This could be simply achieved by jointly maximizing a cross-validation version of (2.3) with respect to  $\theta$ ,  $h$ ,  $g$  and  $l$ . More precisely, we propose to obtain  $h_{ns}$  (for  $h = g = l$ ) by solving

$$\max_{\theta, h} \sum_{j=1}^n \log f_{\varepsilon\theta}(\hat{\varepsilon}_{j,-j,h,s}), \tag{2.7}$$

for  $s = 1, 2$ , and where

$$\hat{\varepsilon}_{j,-j,h,s} = \frac{Y_j - \hat{m}_{h,-j}(X_j)}{\hat{\sigma}_{h,s}(X_j)}, \tag{2.8}$$

with  $\hat{\sigma}_{h,2}(X_j) = \hat{\sigma}_{h,-j}(X_j)$ , and with  $\hat{\sigma}_{h,1}(X_j)$  a local polynomial estimator obtained from an expression of the form (2.2) for  $x = X_j$  and  $g = h$ , but based on the couples  $(X_i, (Y_i - \hat{m}_{h,-i}(X_i))^2)$  ( $i = 1, \dots, n$ ).

Finally, an intermediate idea consists in directly using the residuals in cross-validation least squares minimization problems. That leads to choosing  $h'_{ns}$  (for  $h = g = l$ ) by solving

$$\min_h \sum_{i=1}^n \hat{\varepsilon}_{i,-i,h,s}^2, \tag{2.9}$$

for  $s = 1, 2$ .

To summarize, the six selection procedures that we propose for the bandwidths  $h$  and  $g$  are: (a)  $(h, l, g) = (h_n, g_{n1}, g_{n1})$ , (b)  $(h, l, g) = (h_n, h_n, g_{n2})$ , (c)  $h = g = l = h_{n1}$ , (d)  $h = g = l = h_{n2}$ , (e)  $h = g = l = h'_{n1}$  and (f)  $h = g = l = h'_{n2}$ . Their practical performance will be studied in Section 4.

**Remark 2.2** (*Order of Local Polynomials*). Apart from the practical choices for  $h$ ,  $l$  and  $g$  discussed above, it is important to mention here the dependency between  $h$  ( $l$  and  $g$ ) and the dimension  $d$  of  $X$  required by condition (C4) in the Appendix. Indeed,  $p$  and  $q$  have to increase when  $d$  increases in order that all bandwidth conditions in (C4) are simultaneously satisfied. This can also be interpreted as follows. The sample size  $n$  should increase exponentially with  $d$  to preserve the convergence rates (curse of dimensionality). Consequently, for fixed sample size  $n$ , in order to compensate for this curse of dimensionality, the bandwidths  $h_j$ ,  $l_j$  and  $g_j$  ( $j = 1, \dots, d$ ) should increase exponentially with  $1/d$ . In (C4), this implies that the degree of the polynomials  $P_i(\beta, x, p)$  and  $P_i(\gamma, x, q)$  should increase when  $d$  increases. For example, we will have to choose  $p$  and  $q$  at least equal to 2 when  $d = 2$  (local quadratic estimators), and at least equal to 4 when  $d = 3$  (order 4 local polynomial estimators).

Next, the test statistics are constructed from the difference between  $F_{\varepsilon\theta_n}(y)$  and the nonparametric estimator of  $F_\varepsilon(y)$ :

$$\hat{F}_\varepsilon(y) = \frac{1}{n} \sum_{i=1}^n I(\hat{\varepsilon}_i \leq y). \tag{2.10}$$

This estimator was first studied by Akritas and Van Keilegom (2001) and then extended to the case where  $X$  is  $d$ -dimensional by Neumeyer and Van Keilegom (2010). Consider the process

$$W_n(y) = n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y)), \quad -\infty < y < \infty, \tag{2.11}$$

and define the following test statistics of the Kolmogorov–Smirnov and Cramér-von Mises types:

$$T_{KS} = n^{1/2} \sup_{-\infty < y < \infty} |\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y)|, \tag{2.12}$$

and

$$T_{CM} = n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 d\hat{F}_\varepsilon(y). \tag{2.13}$$

### 3. Asymptotic results

We now turn to the analysis of the asymptotic properties of the estimator  $\theta_n$  and of the test statistics  $T_{KS}$  and  $T_{CM}$ . The assumptions under which these properties are valid are given in the [Appendix](#). The following notations will be used. Let  $\dot{f}_{\varepsilon\theta}(y) = \left(\frac{\partial}{\partial\theta_1}f_{\varepsilon\theta}(y), \dots, \frac{\partial}{\partial\theta_k}f_{\varepsilon\theta}(y)\right)^t$  and  $f'_{\varepsilon\theta}(y) = \frac{d}{dy}f_{\varepsilon\theta}(y)$ . Moreover, we will use the abbreviated notation  $f_\varepsilon \equiv f_{\varepsilon\theta_0}$ ,  $f'_\varepsilon \equiv f'_{\varepsilon\theta_0}$ ,  $\dot{f}_\varepsilon \equiv \dot{f}_{\varepsilon\theta_0}$ , and similarly for  $F_\varepsilon$  and  $\dot{F}_\varepsilon$ .

**Theorem 3.1.** Assume (C1)–(C7). Then, under  $H_0$ ,

$$\theta_n - \theta_0 = -\Omega^{-1}n^{-1} \sum \xi(\varepsilon_i) + o_p(n^{-1/2}),$$

where

$$\xi(t) = \frac{\dot{f}_\varepsilon(t)}{f_\varepsilon(t)} + \int \frac{\dot{f}_\varepsilon(y)f'_\varepsilon(y)}{f_\varepsilon(y)} \left\{t + \frac{y}{2}(t^2 - 1)\right\} dy,$$

and

$$\Omega = E \left[ \frac{\dot{f}_\varepsilon(\varepsilon)\dot{f}_\varepsilon^t(\varepsilon)}{f_\varepsilon^2(\varepsilon)} \right].$$

Moreover,

$$n^{1/2}(\theta_n - \theta_0) \xrightarrow{d} N(0, \Omega^{-1}V\Omega^{-1}),$$

where  $V = E[\xi(\varepsilon)\xi^t(\varepsilon)]$ .

**Theorem 3.2.** Assume (C1)–(C7). Then, under  $H_0$ ,

$$\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y) = n^{-1} \sum_{i=1}^n [I(\varepsilon_i \leq y) - F_\varepsilon(y) + \varphi(\varepsilon_i, y) + \dot{F}_\varepsilon^t(y) \Omega^{-1} \xi(\varepsilon_i)] + R_n(y),$$

where  $\sup_{-\infty < y < \infty} |R_n(y)| = o_p(n^{-1/2})$ , and

$$\varphi(z, y) = f_\varepsilon(y) \left\{z + \frac{y}{2}(z^2 - 1)\right\}.$$

Moreover, the process  $n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))$  ( $-\infty < y < \infty$ ) converges weakly to a zero-mean Gaussian process  $W(y)$  with covariance function

$$\begin{aligned} \text{Cov}(W(y_1), W(y_2)) &= E \left\{ [I(\varepsilon \leq y_1) - F_\varepsilon(y_1) + \varphi(\varepsilon, y_1) + \dot{F}_\varepsilon^t(y_1) \Omega^{-1} \xi(\varepsilon)] \right. \\ &\quad \times \left. [I(\varepsilon \leq y_2) - F_\varepsilon(y_2) + \varphi(\varepsilon, y_2) + \dot{F}_\varepsilon^t(y_2) \Omega^{-1} \xi(\varepsilon)] \right\}. \end{aligned}$$

As a consequence of the above result, we now obtain the asymptotic limit of the test statistics  $T_{KS}$  and  $T_{CM}$  under  $H_0$ .

**Corollary 3.3.** Assume (C1)–(C7). Then, under  $H_0$ ,

$$T_{KS} \xrightarrow{d} \sup_{-\infty < y < \infty} |W(y)|,$$

and

$$T_{CM} \xrightarrow{d} \int W^2(y) dF_\varepsilon(y).$$

**Remark 3.4** (Convergence Under Fixed Alternatives). Note that if the error distribution  $F_\varepsilon$  is a fixed distribution (independent of the sample size  $n$ ) that does not belong to the class  $\mathcal{F}$ , it can be easily seen that the test statistics  $T_{KS}$  and  $T_{CM}$  converge to infinity. In fact, the estimators  $\hat{F}_\varepsilon$  and  $F_{\varepsilon\theta_n}$  do not converge to the same distribution in that case, and hence the process  $n^{1/2}(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))$ ,  $-\infty < y < \infty$ , diverges.

**Remark 3.5** (Bootstrap Approximation). To estimate the distributions of the statistics  $T_{KS}$  and  $T_{CM}$  under  $H_0$ , the asymptotic result given in Corollary 3.3 could in principle be used, although in practice the estimation of the unknown quantities in the asymptotic limit could be cumbersome. Alternatively, resampling techniques can provide very good precision. Here, the method we propose to use is as follows. For  $B$  fixed and for  $b = 1, \dots, B$ ,

1. Let  $\{\varepsilon_{1,b}^*, \dots, \varepsilon_{n,b}^*\}$  be an i.i.d. random sample from the distribution  $F_{\varepsilon\theta_n}(\cdot)$ .
2. Define new responses

$$Y_{i,b}^* = \hat{m}(X_i) + \hat{\sigma}(X_i)\varepsilon_{i,b}^*, \quad i = 1, \dots, n.$$

3. Let  $T_{KS,b}^*$  and  $T_{CM,b}^*$  be the test statistics obtained from the bootstrap sample  $\{(X_1, Y_{1,b}^*), \dots, (X_n, Y_{n,b}^*)\}$ .

Then, if we denote  $T_{KS,(b)}^*$  for the  $b$ -th order statistic of  $T_{KS,1}^*, \dots, T_{KS,B}^*$  and analogously for  $T_{CM,(b)}^*$ , then  $T_{KS,((1-\alpha)B+1)}^*$  and  $T_{CM,((1-\alpha)B+1)}^*$  approximate the  $(1 - \alpha)$ -quantiles of the distributions of  $T_{KS}$  and  $T_{CM}$  respectively (where  $[\cdot]$  denotes the integer part). See also Neumeyer et al. (2006) and Neumeyer (2009), where respectively specific parametric and smooth residual bootstrap (for estimating error distribution processes similar to (2.11)) are considered, and their consistency is proved. The proof of the consistency of the bootstrap procedure we propose above can be studied by using these two papers as starting point. Although this is an important theoretical open question, we do not investigate it in this paper. The proofs are in fact expected to be quite lengthy and technical, making them more appropriate for a separate publication in a theoretical journal.

#### 4. Practical implementation and simulations

In this section, we study the finite sample behavior of the proposed test statistics focusing on some main practical aspects of the implementation and the resulting recommendations.

##### 4.1. Practical implementation

In the one-dimensional case ( $d = 1$ ), we generate i.i.d. data from the model

$$Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \alpha_3 X^3 + (\beta_0 + \beta_1 X)\varepsilon, \tag{4.1}$$

where  $X \sim U[0, 1]$ ,  $\beta_0$  and  $\beta_1$  are such that  $\min(\beta_0, \beta_0 + \beta_1) > 0$ , and  $\varepsilon$  is independent of  $X$  and is a mixture of two normal random variables, i.e.  $F_\varepsilon(y) = p_1 P(Z_1 \leq y) + (1 - p_1) P(Z_2 \leq y)$ , where  $Z_j$  is normal with mean  $\mu_j$  and variance  $\sigma_j^2$  ( $j = 1, 2$ ). Clearly, the parameters  $p_1, \mu_1, \sigma_1^2, \mu_2$  and  $\sigma_2^2$  need to satisfy  $p_1\mu_1 + (1 - p_1)\mu_2 = 0$  and  $p_1(\sigma_1^2 + \mu_1^2) + (1 - p_1)(\sigma_2^2 + \mu_2^2) = 1$ , since  $\varepsilon$  has zero mean and unit variance. Hence,  $\mu_2 = -\mu_1 p_1 / (1 - p_1)$ , and  $\sigma_2^2 = \{\sigma_1^2 p_1^2 - (1 + \sigma_1^2 + \mu_1^2)p_1 + 1\} / (1 - p_1)^2$ . We consider several values of  $p_1, \mu_1$  and  $\sigma_1^2$ .

In the two-dimensional case, we generate i.i.d. data from the model

$$Y = \alpha_{10} X_1 + \alpha_{01} X_2 + \alpha_{11} X_1 X_2 + \alpha_{20} X_1^2 + \alpha_{30} X_1^3 + (\beta_0 + \beta_{11} X_1 X_2)\varepsilon, \tag{4.2}$$

where  $X_1 \sim U[0, 1]$ ,

$$X_2 | X_1 \sim \begin{cases} \beta(0.5, 0.5) & \text{if } X_1 \leq 0.1 \text{ or } X_1 \geq 0.9, \\ U[0, 1] & \text{if } 0.1 < X_1 < 0.9, \end{cases}$$

and for any  $a, b > 0$ ,  $\beta(a, b)$  is the beta-distribution with parameters  $a$  and  $b$ . The parameters  $\beta_0$  and  $\beta_{11}$  satisfy  $\min(\beta_0, \beta_{11}) > 0$ , the error  $\varepsilon$  is independent of  $X = (X_1, X_2)$  and has the same distribution as for model (4.1).

For each model, we simulate 500 samples of size  $n = 100$  or  $200$ . For the estimation of  $m(\cdot)$  and  $\sigma(\cdot)$ , we use the biweight kernel  $k(u) = (15/16)(1 - u^2)^2 I(-1 \leq u \leq 1)$  when  $d = 1$  and the product kernel  $k(u_1)k(u_2)$  when  $d = 2$ . The smoothing parameters  $h, l$  and  $g$  and the estimator  $\theta_n$  are obtained from one of the six procedures described in Remark 2.1, namely by solving either (2.4), (2.5) and (2.3) (hereafter abbreviated by (a)), or (2.4), (2.6) and (2.3) (method (b)), or (2.7) and (2.8) for  $s = 1$  (method (c)), or for  $s = 2$  (method (d)), or (2.9) and (2.3) for  $s = 1$  (method (e)), or for  $s = 2$  (method (f)). When  $d = 2$ , both components of each bandwidth vector are chosen to be equal. The number of bootstrap replications is 200 and the level of the tests is 5%. Following Remark 2.2, we take  $p = q = 1$ , respectively  $p = q = 2$  when  $X$  is one-dimensional, respectively two-dimensional.

**Table 1**

Rejection proportions under  $H_{01}$  for different mixtures of normal distributions and for model (4.1). The parameters determining  $m(\cdot)$  and  $\sigma(\cdot)$  are chosen as  $\alpha_0 = 1, \alpha_1 = 1, \alpha_2 = -2, \alpha_3 = 1.5, \beta_0 = 0.1$  and  $\beta_1 = 0.1$ .

$p_1$	$\mu_1$	$\sigma_1$		(a)	(b)	(c)–(e)	(d)–(f)
0	0	1	$T_{KS}$	0.046	0.044	0.040	0.058
			$T_{CM}$	0.038	0.034	0.042	0.064
			$T_{AD}$	0.038	0.036	0.042	0.062
0.4	0.9	0.7	$T_{KS}$	0.232	0.238	0.226	0.308
			$T_{CM}$	0.284	0.330	0.288	0.382
			$T_{AD}$	0.242	0.290	0.266	0.320
0.45	0.3	0.4	$T_{KS}$	0.514	0.550	0.530	0.548
			$T_{CM}$	0.568	0.598	0.558	0.652
			$T_{AD}$	0.642	0.690	0.620	0.756
0.44	0.9	0.7	$T_{KS}$	0.630	0.680	0.652	0.736
			$T_{CM}$	0.740	0.778	0.756	0.826
			$T_{AD}$	0.708	0.760	0.738	0.792
0.5	0.7	0.1	$T_{KS}$	1	1	1	1
			$T_{CM}$	1	1	1	1
			$T_{AD}$	1	1	1	1

**Table 2**

Rejection proportions under  $H_{02}$  for different mixtures of normal distributions and for model (4.1). The parameters determining  $m(\cdot)$  and  $\sigma(\cdot)$  are chosen as  $\alpha_0 = 1, \alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0, \beta_0 = 0.3$  and  $\beta_1 = 0.3$ .

$p_1$	$\mu_1$	$\sigma_1$		(a)	(b)	(c)	(d)	(e)	(f)
0.44	0.9	0.7	$T_{KS}$	0.044	0.036	0.038	0.038	0.046	0.044
			$T_{CM}$	0.038	0.034	0.046	0.036	0.044	0.032
			$T_{AD}$	0.042	0.036	0.050	0.040	0.046	0.034
0	0	1	$T_{KS}$	0.046	0.042	0.048	0.054	0.044	0.056
			$T_{CM}$	0.044	0.054	0.042	0.056	0.038	0.062
			$T_{AD}$	0.046	0.048	0.044	0.052	0.038	0.056
0.45	0.3	0.4	$T_{KS}$	0.756	0.784	0.686	0.834	0.682	0.836
			$T_{CM}$	0.788	0.826	0.732	0.882	0.734	0.896
			$T_{AD}$	0.788	0.822	0.718	0.892	0.724	0.894
0.5	0.7	0.1	$T_{KS}$	1	1	1	1	1	1
			$T_{CM}$	1	1	1	1	1	1
			$T_{AD}$	1	1	1	1	1	1

We consider three different tests for  $F_\varepsilon$ :

1.  $H_{01} : \varepsilon \sim \Phi(0, 1)$ ,
2.  $H_{02} : \varepsilon \sim p_1\Phi(0.9, 0.49) + (1 - p_1)\Phi\left(\frac{-0.9p_1}{1-p_1}, \frac{0.49p_1^2 - 2.3p_1 + 1}{(1-p_1)^2}\right)$ ,
3.  $H_{03} : \varepsilon \sim p_1\Phi(\mu_1, 0.49) + (1 - p_1)\Phi\left(\frac{-\mu_1 p_1}{1-p_1}, \frac{0.49p_1^2 - (1.49 + \mu_1^2)p_1 + 1}{(1-p_1)^2}\right)$ ,

where  $\Phi(\mu, \sigma^2)$  stands for a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . Clearly, hypothesis  $H_{01}$  is satisfied for  $p_1 = 0, \mu_1 = 0$  and  $\sigma_1 = 1$ , hypothesis  $H_{02}$  is satisfied for any value of  $p_1$  and for  $\mu_1 = 0.9$  and  $\sigma_1 = 0.7$ , and hypothesis  $H_{03}$  is satisfied for any  $p_1$  and  $\mu_1$  and for  $\sigma_1 = 0.7$ .

Tables 1–6 present the rejection proportions for the above three null hypotheses. In order to illustrate the results in a more complete way, we added a test based on a statistic of the Anderson–Darling type given by

$$T_{AD} = n \int \frac{(\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2}{\hat{F}_\varepsilon(y)(1 - \hat{F}_\varepsilon(y))} d\hat{F}_\varepsilon(y). \tag{4.3}$$

A ‘parametric’ version of this statistic, obtained by replacing  $d\hat{F}_\varepsilon(y)$  by  $dF_{\varepsilon\theta_n}(y)$  and  $\hat{F}_\varepsilon(y)(1 - \hat{F}_\varepsilon(y))$  by  $F_{\varepsilon\theta_n}(y)(1 - F_{\varepsilon\theta_n}(y))$  in the expression above, was also studied but its results were slightly worse than those of  $T_{AD}$ . In the same way, the ‘parametric’ version of  $T_{CM}$  leads to quite similar results as  $T_{CM}$ . Rejection proportions for those ‘parametric’ statistics are therefore not reported here.

In general, (a), (b), (c) and (e) often seem to overfit more or less the data inducing more variability of the distributions of the corresponding test statistics and their bootstrap estimators. In (d) and (f), extrema of (2.7) and (2.9) are obtained for larger values of the bandwidth parameters than in (c) and (e) (and (a) and (b)). Indeed, adding globally increasing values of  $(Y_i - \hat{m}_{h,-i}(X_i))^2$  to the weighted average  $\hat{\sigma}_{h,2}^2(X_j)$  (for increasing values of  $h$ ) is more likely to increase the likelihood function (or decrease the least squares), since it does not consider the value of  $(Y_j - \hat{m}_{h,-j}(X_j))^2$ . Under the null, (c) seems to provide the closest resamples to the true model, leading to good bootstrap approximations (and also suggesting that (c)

**Table 3**

Rejection proportions under  $H_{03}$  for different mixtures of normal distributions and for model (4.1). The parameters determining  $m(\cdot)$  and  $\sigma(\cdot)$  are chosen as  $\alpha_0 = 1, \alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0, \beta_0 = 0.3$  and  $\beta_1 = 0.3$ .

$p_1$	$\mu_1$	$\sigma_1$		(a)	(b)	(c)	(d)	(e)	(f)
0.44	0.9	0.7	$T_{KS}$	0.012	0.034	0.024	0.010	0.030	0.018
			$T_{CM}$	0.014	0.018	0.032	0.016	0.026	0.010
			$T_{AD}$	0.012	0.016	0.022	0.016	0.024	0.010
0	0	1	$T_{KS}$	0.022	0.018	0.028	0.026	0.018	0.032
			$T_{CM}$	0.016	0.018	0.024	0.014	0.020	0.020
			$T_{AD}$	0.016	0.022	0.018	0.018	0.030	0.026
0.45	0.3	0.4	$T_{KS}$	0.074	0.080	0.126	0.136	0.104	0.164
			$T_{CM}$	0.088	0.116	0.162	0.192	0.116	0.210
			$T_{AD}$	0.100	0.136	0.158	0.206	0.112	0.252
0.5	0.7	0.1	$T_{KS}$	0.978	0.994	0.998	1	0.984	1
			$T_{CM}$	0.962	0.974	0.996	1	0.990	1
			$T_{AD}$	0.974	0.984	0.996	1	0.992	1

**Table 4**

Rejection proportions under  $H_{01}$  for different mixtures of normal distributions and for model (4.2). The smoothing parameter is obtained from (f),  $n = 100, \alpha_{10} = 1, \alpha_{01} = 1, \alpha_{11} = 1, \alpha_{20} = 4, \alpha_{30} = -3, \beta_0 = 0.3$  and  $\beta_{11} = 0.3$ .

$p_1$	$\mu_1$	$\sigma_1$	$T_{KS}$	$T_{CM}$	$T_{AD}$
0	0	1	0.062	0.054	0.074
0.4	0.9	0.7	0.230	0.244	0.208
0.45	0.3	0.4	0.432	0.428	0.446
0.44	0.9	0.7	0.548	0.634	0.598
0.5	0.7	0.1	1	0.996	1

**Table 5**

Rejection proportions under  $H_{02}$  for different mixtures of normal distributions and for model (4.2). The bandwidth parameter is obtained by (f),  $n = 100, \alpha_{10} = 1, \alpha_{01} = 1, \alpha_{11} = 1, \alpha_{20} = 4, \alpha_{30} = -3, \beta_0 = 0.3$  and  $\beta_{11} = 0.3$ .

$p_1$	$\mu_1$	$\sigma_1$	$T_{KS}$	$T_{CM}$	$T_{AD}$
0.44	0.9	0.7	0.032	0.040	0.038
0	0	1	0.028	0.022	0.034
0.45	0.3	0.4	0.448	0.446	0.454
0.5	0.7	0.1	1	0.992	0.998

**Table 6**

Rejection proportions under  $H_{02}$  for one mixture of normal distributions and for model (4.2). The bandwidth parameter is obtained by (f),  $n = 200, \alpha_{10} = 1, \alpha_{01} = 1, \alpha_{11} = 1, \alpha_{20} = 4, \alpha_{30} = -3, \beta_0 = 0.3$  and  $\beta_{11} = 0.3$ .

$p_1$	$\mu_1$	$\sigma_1$	$T_{KS}$	$T_{CM}$	$T_{AD}$
0.45	0.3	0.4	0.918	0.934	0.924

is a good estimation procedure for the parameters of  $F_{\varepsilon\theta}$ ). Under  $H_1$ , the best results are usually obtained by (f) since  $h'_{n2}$  is obtained independently of  $\theta_n$  and without assuming  $H_0$ .

When  $\kappa$  increases (so when we move from  $H_{01}$  to  $H_{02}$  and  $H_{03}$ ), the above considerations stay true, but obviously the bootstrap approximations are based on samples generated closer and closer to the original samples and the distances between  $F_\varepsilon$  and  $\mathcal{F}$  become considerably smaller, inducing less and less rejections. However, as it turns out, (c) globally seems to keep results closer to 0.05 under  $H_0$  and (f) loses less power under  $H_1$ .

In Tables 4–6 we study the behavior of the different statistics for  $d = 2$ . Only the results for (f) are displayed. Indeed, the same characteristics as in the one-dimensional case apply. As can be expected, the power is smaller than for the one-dimensional case, but the results stay reasonably good. An increase of  $\kappa$  seems to affect the results more than the same increase of  $d$ , and a small increase of  $n$  (from 100 to 200 in Table 6) already improves the results significantly.

## 4.2. Conclusions

In general, method (f) can be recommended in practice. However, there are as always exceptions on this general rule. For example, if strongly different behaviors of  $m(\cdot)$  and  $\sigma(\cdot)$  are detected (e.g. if  $m(\cdot)$  is much more wiggly than  $\sigma(\cdot)$ ), it seems better to use method (b). Moreover, method (c) behaves well under the null hypothesis, suggesting it would be interesting to investigate its estimation behavior, i.e., the study of  $\theta_n$  obtained with method (c) when a parametric distribution is assumed

for  $\varepsilon$ . We like to note here that the above recommendations are only based on simulations for small samples and not on any theoretical argument. The theoretical investigation of the proposed bandwidth selection procedures is a challenging problem, which the authors plan to study in the future. Finally, since the different statistics are based on a difference of distributions, we think that the Cramér-von Mises statistic should be recommended or possibly the Anderson–Darling statistic if discrepancies in the tails are suspected.

**Acknowledgements**

Both authors acknowledge financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy). The research of the second author is in addition financed by the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007–2013)/ERC Grant agreement No. 203650.

**Appendix**

In this appendix we state the assumptions under which the asymptotic results of Section 3 are valid, and we also give the proofs of these results. Throughout this appendix we will denote the true error by  $\varepsilon_0$ , the true regression function by  $m_0$  and the true scale function by  $\sigma_0$ .

For an arbitrary  $\theta$ , and for arbitrary functions  $m$  and  $\sigma > 0$  defined on  $R_X$ , let

$$G(\theta, m, \sigma) = E \left[ \frac{\dot{f}_{\varepsilon\theta} \left( \frac{Y-m(X)}{\sigma(X)} \right)}{f_{\varepsilon\theta} \left( \frac{Y-m(X)}{\sigma(X)} \right)} \right].$$

- (C1) For all  $\delta > 0$ , there exists an  $\varepsilon > 0$  such that  $\inf_{\|\theta-\theta_0\|>\delta} \|G(\theta, m_0, \sigma_0)\| > \varepsilon$ .
- (C2) Uniformly for all  $\theta \in \Theta$ ,  $G(\theta, m, \sigma)$  is continuous with respect to the supremum norm in  $(m, \sigma)$  at  $(m, \sigma) = (m_0, \sigma_0)$ . Moreover,  $\Omega$  is non-singular.
- (C3)  $k$  is a symmetric probability density function supported on  $[-1, 1]$ ,  $k$  is  $d$  times continuously differentiable, and  $k^{(j)}(\pm 1) = 0$  for  $j = 0, \dots, d - 1$ .
- (C4)  $h_j, l_j$  and  $g_j$  are of the same order ( $j = 1, \dots, d$ ) and satisfy  $h_j/h^* \rightarrow b_j, l_j/h^* \rightarrow c_j$  and  $g_j/h^* \rightarrow d_j$  for some  $0 < b_j, c_j, d_j < \infty$  and some baseline bandwidth  $h^*$ . Moreover, for  $r = p$  or  $q$ ,  $h^*$  satisfies  $nh^{*2r+4} \rightarrow 0$  when  $r$  is even,  $nh^{*2r+2} \rightarrow 0$  when  $r$  is odd and  $nh^{*3d+\delta} \rightarrow \infty$  for some small  $\delta > 0$ .
- (C5) All partial derivatives of  $F_X$  up to order  $2d + 1$  exist on the interior of  $R_X$ , they are uniformly continuous and  $\inf_{x \in R_X} f_X(x) > 0$ .
- (C6) All partial derivatives of  $m_0$  and  $\sigma_0$  up to order  $p + 2$  exist on the interior of  $R_X$ , they are uniformly continuous and  $\inf_{x \in R_X} \sigma_0(x) > 0$ .
- (C7) All (mixed) derivatives up to order 3 of  $F_{\varepsilon\theta}(y)$  with respect to  $y$  and the components of  $\theta$  exist and are continuous. Moreover,  $\sup_y |y^2 f'_\varepsilon(y)| < \infty$  and  $E(\varepsilon_0^6) < \infty$ .

**Proof of Theorem 3.1.** For an arbitrary  $\theta$ , and for arbitrary functions  $m : R_X \rightarrow \mathbb{R}$  and  $\sigma : R_X \rightarrow \mathbb{R}^+$ , let

$$G_n(\theta, m, \sigma) = n^{-1} \sum_{i=1}^n \frac{\dot{f}_{\varepsilon\theta} \left( \frac{Y_i-m(X_i)}{\sigma(X_i)} \right)}{f_{\varepsilon\theta} \left( \frac{Y_i-m(X_i)}{\sigma(X_i)} \right)}.$$

Then,  $\theta_n$  is the solution of  $G_n(\theta, \hat{m}, \hat{\sigma}) = 0$ . Also note that  $\theta_0$  satisfies  $G(\theta_0, m_0, \sigma_0) = 0$ . We will start by showing the consistency of  $\theta_n$ . This will be done by checking the conditions of Theorem 1 in Chen et al. (2003) (CLV hereafter). Condition (1.1) holds by definition of  $\theta_n$ , while the second and third condition are guaranteed by assumptions (C1) and (C2). Condition (1.4) follows from Lemma A.1 in Neumeyer and Van Keilegom (2010), and from conditions (C3)–(C7). Finally, condition (1.5) is very similar to condition (2.5) of Theorem 2 of CLV, and we will verify both conditions below. So, the conditions of Theorem 1 are verified, up to condition (1.5) which we postpone to later. Next, we verify conditions (2.1)–(2.6) of Theorem 2 in CLV in order to obtain an i.i.d. representation for  $\theta_n - \theta_0$  and to prove the asymptotic normality of  $\theta_n$ . Condition (2.1) is, as for condition (1.1), valid by construction of the estimator  $\theta_n$ . For condition (2.2), write for any  $\theta \in \Theta$ ,

$$\begin{aligned} \Gamma_1(\theta) &:= \frac{\partial}{\partial \theta} G(\theta, m_0, \sigma_0) = E \left[ \frac{\ddot{f}_{\varepsilon\theta}(\varepsilon_0) f_{\varepsilon\theta}(\varepsilon_0) - \dot{f}_{\varepsilon\theta}(\varepsilon_0) \dot{f}_{\varepsilon\theta}^t(\varepsilon_0)}{f_{\varepsilon\theta}^2(\varepsilon_0)} \right] \\ &= \int \ddot{f}_{\varepsilon\theta}(y) dy - \int \frac{\dot{f}_{\varepsilon\theta}(y) \dot{f}_{\varepsilon\theta}^t(y)}{f_{\varepsilon\theta}(y)} dy \\ &= \frac{\partial^2}{\partial \theta \partial \theta^t} \int f_{\varepsilon\theta}(y) dy - E \left[ \frac{\dot{f}_{\varepsilon\theta}(\varepsilon_0) \dot{f}_{\varepsilon\theta}^t(\varepsilon_0)}{f_{\varepsilon\theta}^2(\varepsilon_0)} \right] = -E \left[ \frac{\dot{f}_{\varepsilon\theta}(\varepsilon_0) \dot{f}_{\varepsilon\theta}^t(\varepsilon_0)}{f_{\varepsilon\theta}^2(\varepsilon_0)} \right], \end{aligned}$$

where  $\ddot{f}_{\varepsilon\theta}(y) = \frac{\partial^2}{\partial\theta\partial\theta\tau}f_{\varepsilon\theta}(y)$ . Hence, condition (2.2) is easily seen to hold thanks to condition (C2) and (C7). As for condition (2.3), we need to calculate

$$\Gamma_2(\theta, m_0, \sigma_0)[m - m_0, \sigma - \sigma_0] := \lim_{\tau \rightarrow 0} \frac{1}{\tau} \{G(\theta, m_0 + \tau(m - m_0), \sigma_0 + \tau(\sigma - \sigma_0)) - G(\theta, m_0, \sigma_0)\}.$$

For notational convenience, in a first stage, we ignore the  $\sigma$  component in the above expression and calculate

$$\begin{aligned} \Gamma_2(\theta, m_0)[m - m_0] &= \lim_{\tau \rightarrow 0} \frac{1}{\tau} \left\{ E \left[ \frac{\dot{f}_{\varepsilon\theta}(Y - m_0(X) - \tau(m - m_0)(X))}{f_{\varepsilon\theta}(Y - m_0(X) - \tau(m - m_0)(X))} - \frac{\dot{f}_{\varepsilon\theta}(Y - m_0(X))}{f_{\varepsilon\theta}(Y - m_0(X))} \right] \right\} \\ &= -E \left[ \frac{\dot{f}'_{\varepsilon\theta}(\varepsilon_0)f_{\varepsilon\theta}(\varepsilon_0) - \dot{f}_{\varepsilon\theta}(\varepsilon_0)f'_{\varepsilon\theta}(\varepsilon_0)}{f_{\varepsilon\theta}^2(\varepsilon_0)}(m - m_0)(X) \right] \\ &= - \left\{ \int \dot{f}'_{\varepsilon\theta}(y)dy - \int \frac{\dot{f}_{\varepsilon\theta}(y)f'_{\varepsilon\theta}(y)}{f_{\varepsilon\theta}(y)} dy \right\} E[(m - m_0)(X)] \\ &= \int \frac{\dot{f}_{\varepsilon\theta}(y)f'_{\varepsilon\theta}(y)}{f_{\varepsilon\theta}(y)} dy E[(m - m_0)(X)], \end{aligned}$$

since  $\frac{\partial}{\partial\theta} \int f'_{\varepsilon\theta}(y) dy = 0$ . Hence, taking the dependence on  $\sigma$  into account, we get that

$$\Gamma_2(\theta, m_0, \sigma_0)[m - m_0, \sigma - \sigma_0] = \int \frac{\dot{f}_{\varepsilon\theta}(y)f'_{\varepsilon\theta}(y)}{f_{\varepsilon\theta}(y)} dy E \left[ \frac{(m - m_0)(X)}{\sigma_0(X)} \right] + \int y \frac{\dot{f}_{\varepsilon\theta}(y)f'_{\varepsilon\theta}(y)}{f_{\varepsilon\theta}(y)} dy E \left[ \frac{(\sigma - \sigma_0)(X)}{\sigma_0(X)} \right],$$

for all  $\theta, m$  and  $\sigma > 0$ . Now, it is easily seen that condition (2.3)(i) is satisfied, if we define for any function  $h$  defined on  $R_X$ ,

$$\|h\|_{d+\alpha} = \max_{k \leq d} \sup_{x \in R_X} |D^k h(x)| + \max_{k=d} \sup_{x, x' \in R_X} \frac{|D^k h(x) - D^k h(x')|}{\|x - x'\|^\alpha},$$

$$k = (k_1, \dots, k_d),$$

$$D^k = \frac{\partial^k}{\partial x_1^{k_1} \dots \partial x_d^{k_d}},$$

$k_j = \sum_{j=1}^d k_j$ , and  $\|\cdot\|$  is the Euclidean norm on  $\mathbb{R}^d$ . As for (2.3)(ii), note that it follows from the proof of Theorem 2 in CLV that it suffices to show that

$$\|\Gamma_2(\hat{\theta}, m_0, \sigma_0)[\hat{m} - m_0, \hat{\sigma} - \sigma_0] - \Gamma_2(\theta_0, m_0, \sigma_0)[\hat{m} - m_0, \hat{\sigma} - \sigma_0]\| \leq \|\hat{\theta} - \theta_0\| o_p(1),$$

and this can be easily shown, using condition (C7). For condition (2.4), let

$$\mathcal{M} = \{m : R_X \rightarrow \mathbb{R} : \|m\|_{d+\alpha} \leq M_1\},$$

and

$$\mathcal{J} = \left\{ \sigma : R_X \rightarrow \mathbb{R} : \|\sigma\|_{d+\alpha} \leq M_1, \inf_{x \in R_X} \sigma(x) > M_0 \right\},$$

for some  $0 < M_0 < M_1 < \infty$ . Then, we apply once more (the proof of) Lemma A.1 in Neumeyer and Van Keilegom (2010) for the rate of convergence of  $\hat{m}$  and  $\hat{\sigma}$ . The same lemma ensures that  $P(\hat{m} \in \mathcal{M}) \rightarrow 1$  and  $P(\hat{\sigma} \in \mathcal{J}) \rightarrow 1$ . Next, for verifying condition (2.5) we check the conditions of Theorem 3 in CLV. It suffices to check either condition (3.1) or (3.2). Condition (3.2) is verified for  $s_j = 1$  and  $j = 1, \dots, \kappa$ , whereas condition (3.3) follows from Theorem 2.7.1 in Van der Vaart and Wellner (1996). It remains to verify condition (2.6), which is immediate after applying the i.i.d. representation for  $E \left[ \frac{(\hat{m} - m_0)(X)}{\sigma_0(X)} \right]$  and  $E \left[ \frac{(\hat{\sigma} - \sigma_0)(X)}{\sigma_0(X)} \right]$ , given in Lemma A.2 in Neumeyer and Van Keilegom (2010). This finishes the proof.  $\square$

**Proof of Theorem 3.2.** Write

$$\begin{aligned} \hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y) &= \left[ \hat{F}_\varepsilon(y) - F_\varepsilon(y) \right] - \left[ F_{\varepsilon\theta_n}(y) - F_\varepsilon(y) \right] \\ &= n^{-1} \sum_{i=1}^n I(\varepsilon_i \leq y) - F_\varepsilon(y) + n^{-1} \sum_{i=1}^n \varphi(\varepsilon_i, y) + \hat{F}_\varepsilon^t(y)(\theta_n - \theta_0) + o_p(n^{-1/2}), \end{aligned}$$

uniformly in  $y$ , where the latter equality follows from Theorem 2.1 in Neumeyer and Van Keilegom (2010). The result now follows from Theorem 3.1.  $\square$

**Proof of Corollary 3.3.** The convergence of the Kolmogorov–Smirnov statistic  $T_{KS}$  follows directly from the continuous mapping theorem. For the Crámer-von Mises statistic  $T_{CM}$  it suffices to show that  $d\hat{F}_\varepsilon(y)$  can be replaced by  $dF_\varepsilon(y)$ . Write

$$\left| n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 d[\hat{F}_\varepsilon(y) - F_\varepsilon(y)] \right| \leq \left| n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 d[\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y)] \right| + \left| n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 d[F_{\varepsilon\theta_n}(y) - F_\varepsilon(y)] \right|.$$

It suffices to consider the second term above, which can be written as

$$\left| n \int (\hat{F}_\varepsilon(y) - F_{\varepsilon\theta_n}(y))^2 \hat{f}_{\varepsilon\tilde{\theta}}^t(y) dy \Omega^{-1} n^{-1} \sum_{i=1}^n \xi(\varepsilon_i) \right| + o_p(1) = o_p(1),$$

(with  $\tilde{\theta}$  on the line segment between  $\theta_0$  and  $\theta_n$ ) which follows from [Theorems 3.1](#) and [3.2](#).  $\square$

## References

- Akritas, M.G., Van Keilegom, I., 2001. Nonparametric estimation of the residuals distribution. *Scand. J. Statist.* 28, 549–567.
- Alcalá, J.T., Cristóbal, J.A., González Manteiga, W., 1999. Goodness-of-fit test for linear models based on local polynomials. *Statist. Probab. Lett.* 42, 39–46.
- Chen, X., Linton, O., Van Keilegom, I., 2003. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71, 1591–1608.
- Deschepper, E., Thas, O., Ottoy, J.P., 2006. Regional residual plots for assessing the fit of linear regression models. *Comput. Statist. Data Anal.* 50, 1995–2013.
- Dette, H., 1999. A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.* 27, 1012–1040.
- Dette, H., Munk, A., 1998. Validation of linear regression models. *Ann. Statist.* 26, 778–800.
- 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.
- Fan, J., Zhang, C., Zhang, J., 2001. Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* 29, 153–193.
- Härdle, W., Mammen, E., 1993. Comparing non parametric versus parametric regression fits. *Ann. Statist.* 21, 1926–1947.
- Huskova, M., Meintanis, S.G., 2007. Omnibus tests for the error distribution in the linear regression model. *Statistics* 41, 363–376.
- Huskova, M., Meintanis, S.G., 2010. Tests for the error distribution in nonparametric possibly heteroscedastic regression models, *TEST* (in press).
- Jiménez Gamero, M.D., Muñoz García, J., Pino Mejías, R., 2005. Testing goodness of fit for the distribution of errors in multivariate linear models. *J. Multivariate Anal.* 95, 301–322.
- Mora, J., Pérez-Alonso, A., 2009. Specification tests for the distribution of errors in nonparametric regression: A martingale approach. *J. Nonparametr. Stat.* 21, 441–452.
- Neumeyer, N., 2009. Smooth residual bootstrap for empirical processes of nonparametric regression residuals. *Scand. J. Statist.* 36, 204–228.
- Neumeyer, N., Dette, H., Nagel, E.-R., 2006. Bootstrap tests for the error distribution in linear and nonparametric regression models. *Aust. N. Z. J. Stat.* 48, 129–156.
- Neumeyer, N., Van Keilegom, I., 2010. Estimating the error distribution in nonparametric multiple regression with applications to model testing. *J. Multivariate Anal.* 101, 1067–1078.
- Stute, W., 1997. Nonparametric model checks for regression. *Ann. Statist.* 25, 613–641.
- Van der Vaart, A.W., Wellner, J.A., 1996. *Weak Convergence and Empirical Processes*. Springer, New York.
- Van Keilegom, I., González Manteiga, W., Sánchez Selloero, C., 2008. Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *TEST* 17, 401–415.