Journal of Business & Economic Statistics

ISSN: 0735-0015 (Print) 1537-2707 (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

A Smooth Nonparametric, Multivariate, Mixed-Data Location-Scale Test

Jeffrey S. Racine & Ingrid Van Keilegom

To cite this article: Jeffrey S. Racine & Ingrid Van Keilegom (2020) A Smooth Nonparametric, Multivariate, Mixed-Data Location-Scale Test, Journal of Business & Economic Statistics, 38:4, 784-795, DOI: 10.1080/07350015.2019.1574227

To link to this article: https://doi.org/10.1080/07350015.2019.1574227

A Smooth Nonparametric, Multivariate, Mixed-Data Location-Scale Test

Jeffrey S. RACINE
Department of Economics, McMaster University, Kenneth Taylor Hall, Room 426, 1280 Main Street West Hamilton, Ontario, Canada, L8S 4M4; Info-Metrics Institute, American University; Rimini Center for Economic Analysis; Center for Research in Econometric Analysis of TimE Series (CREATES), Aarhus University (*racinej@mcmaster.ca*)

Ingrid VAN KEILEGOM
ORSTAT, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium (*ingrid.vankeilegom@kuleuven.be*)

A number of tests have been proposed for assessing the location-scale assumption that is often invoked by practitioners. Existing approaches include Kolmogorov–Smirnov and Cramer–von Mises statistics that each involve measures of divergence between unknown joint distribution functions and products of marginal distributions. In practice, the unknown distribution functions embedded in these statistics are typically approximated using nonsmooth empirical distribution functions (EDFs). In a recent article, Li, Li, and Racine establish the benefits of smoothing the EDF for inference, though their theoretical results are limited to the case where the covariates are *observed* and the distributions unobserved, while in the current setting some covariates *and* their distributions are *unobserved* (i.e., the test relies on population error terms from a location-scale model) which necessarily involves a separate theoretical approach. We demonstrate how replacing the nonsmooth distributions of unobservables with their kernel-smoothed sample counterparts can lead to substantial power improvements, and extend existing approaches to the smooth multivariate and mixed continuous and discrete data setting in the presence of unobservables. Theoretical underpinnings are provided, Monte Carlo simulations are undertaken to assess finite-sample performance, and illustrative applications are provided.

KEY WORDS: Inference; Kernel smoothing; Kolmogorov–Smirnov.

1. INTRODUCTION

Assuming independence of the predictors and error in a location-scale regression model is a common assumption. The independence assumption is for instance needed for certain bootstrap procedures (see Neumeyer 2008, 2009; Neumeyer and Van Keilegom 2018). There is also an extensive literature on testing procedures that use the independence between the error and the covariates, and that are based on a comparison between a nonparametric estimator of the error distribution and an estimator under the null hypothesis. We refer for instance to Van Keilegom, González-Manteiga, and Sánchez-Sellero (2008) and Dette, Neumeyer, and Van Keilegom (2007) for goodness-of-fit tests for the parametric form of the regression and the variance function, respectively, to Pardo-Fernández, Van Keilegom, and González-Manteiga (2007) for tests for the equality of regression curves, and to Escanciano, Pardo-Fernández, and Van Keilegom (2018) for distribution-free tests in this context. A testing procedure for the location-scale structure having high power would be particularly appealing.

A variety of tests have been proposed for assessing the appropriateness of the location-scale assumption that is often invoked in applied settings; see by way of illustration Akritas and Van Keilegom (2001) and Racine and Li (2017), who adopt the location-scale framework, and see Einmahl and Van Keilegom (2008), Birke, Neumeyer, and Volgushev (2017), and Neumeyer, Noh, and Van Keilegom (2016) for various approaches that have been proposed to test the location-scale assumption in a range of settings. These approaches employ test statistics that are based on conditional mean models, in particular, the difference between the joint distribution of the predictor and error and the product of the marginal distributions of the predictor and error, and include the Kolmogorov–Smirnov (Kolmogorov 1933; Smirnov 1948), Cramer–von Mises (Cramér 1928; von Mises 1928), and Anderson–Darling (Anderson and Darling 1952) statistics, among others. In this literature, the unknown joint and marginal distributions are estimated using the respective nonsmooth empirical distribution functions (EDFs). However, it turns out that substantial power gains can be realized by replacing the nonsmooth EDFs with their kernel-smoothed counterparts. We demonstrate that we retain all of the desirable features of this testing framework yet can realize substantial improvements to existing procedures from the vantage point of finite-sample power without impacting size. Though we consider inference for location-scale models, the results contained herein are of broad applicability and ought to appeal to a wide audience, particularly practitioners concerned with

Journal of Business & Economic Statistics
October 2020, Vol. 38, No. 4
DOI: 10.1080/07350015.2019.1574227
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

power properties associated with this popular class of test statistics.

The remainder of this article proceeds as follows: Section 2 outlines the location-scale framework and the proposed smooth testing procedure; Sections 3 and 4 present the theoretical underpinnings of the proposed approach; Section 5 presents simulation evidence that demonstrates power gains achievable by a fully data-driven implementation of the proposed approach; Section 6 considers an illustrative application, while Section 7 presents some concluding remarks. The proofs of the main results are given in Appendix A (supplementary materials), while detailed tables outlining power gains are presented in Appendices B and C (supplementary materials).

2. METHODOLOGY

Consider a smooth location-scale model of the form

$$Y = \mu(X) + \sigma(X)\epsilon,$$

where $\mu(\cdot)$ and $\sigma(\cdot) \geq 0$ are unknown smooth location and scale functions, X is a vector of predictors, and ϵ has zero mean, unit variance, and is otherwise an unknown error process with distribution F_ϵ that is independent of X. We observe n independent copies of (X^{T}, Y), denoted by $(X_1^{\mathrm{T}}, Y_1), \ldots, (X_n^{\mathrm{T}}, Y_n)$.

The location-scale assumption is often invoked as it confers a number of useful properties on the resulting estimator, including (i) simpler asymptotic properties than its unstructured counterpart,[1] (ii) the ability to nonparametrically estimate the error distribution at a $\sqrt{n}$-rate (Akritas and Van Keilegom 2001; Escanciano and Jacho-Chávez 2012), and (iii) more efficient estimation of the conditional distribution of Y given X than its unstructured counterpart.

Even though the independence of the predictors X and error ϵ is a common assumption (see, e.g., Akritas and Van Keilegom 2001; Racine and Li 2017), particularly in econometrics, it might be too strong, hence a testing procedure having high power is particularly appealing. For what follows, we define $F_X(x) = P(X \leq x)$, $F_\epsilon(t) = P(\epsilon \leq t)$, and $F_{X,\epsilon}(x,t) = P(X \leq x, \epsilon \leq t)$, and we let

$$H_0 : X \text{ and } \epsilon \text{ are independent.}$$

Consider by way of illustration the test of Einmahl and Van Keilegom (2008) which can be used to assess the adequacy of the location-scale assumption. In essence, Einmahl and Van Keilegom (2008) test for independence between the predictors X and error ϵ in the location-scale model $Y = \mu(X) + \sigma(X)\epsilon$. Given kernel estimates of $\mu(x) = E(Y|X = x)$ and $\sigma^2(x) = V(Y|X = x)$, denoted $\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$, one tests for independence of X_i and $\hat{\epsilon}_i = (Y_i - \hat{\mu}(X_i))/\hat{\sigma}(X_i)$ using, for instance, a Kolmogorov–Smirnov test statistic of the form

$$T_{\mathrm{KS}} = \sqrt{n} \sup_{x,t} |\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)|, \tag{1}$$

where $\hat{F}_{X,\hat{\epsilon}}(x,t)$, $\hat{F}_X(x)$, and $\hat{F}_{\hat{\epsilon}}(t)$ are the respective EDFs. The empirical support, that is, the X_i and $\hat{\epsilon}_i$, are used when computing the supremum in applied settings. Due to the inadequacy of using the asymptotic distribution of T_{KS} for inference, a simple bootstrap procedure is used instead to obtain the null distribution from which nonparametric p-values can readily be obtained. This procedure can be easily modified to test for the validity of a homoscedastic model $Y_i = \mu(X_i) + \epsilon_i$ with ϵ_i being independent of X_i, which is also a common assumption in econometrics, or to test the validity of a transformation model of the form $\Delta(Y_i) = \mu(X_i) + \sigma(X_i)\epsilon_i$ as outlined by Neumeyer, Noh, and Van Keilegom (2016), where $\Delta(\cdot)$ is some parametric monotone transformation. See Neumeyer (2008) who demonstrated consistency of the bootstrap in the kernel-smoothed case, while Neumeyer and Van Keilegom (2018) addressed the open question of whether a classical nonsmooth residual bootstrap is asymptotically valid in this context, and show that the nonsmooth residual bootstrap is consistent.

We note in passing that the Cramer–von Mises statistic, which is also popular in applied settings, is given by

$$T_{\mathrm{CM}} = n \int \int \left(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\right)^2 d\hat{F}_X(x)\, d\hat{F}_{\hat{\epsilon}}(t).$$

We propose replacing the EDFs in these statistics with their kernel-smoothed counterparts and the unknown error term with its kernel estimate using an approach similar to Li, Li, and Racine (2017) that is briefly described below. One major difference between the results established here and those in Li, Li, and Racine (2017) is that, here, some covariates and distributions involve *unobserved* error terms that need to be estimated by $\hat{\epsilon}_i = (Y_i - \hat{\mu}(X_i))/\hat{\sigma}(X_i)$, which results in markedly different asymptotics and finite sample performance from that reported in Li, Li, and Racine (2017). Like Li, Li, and Racine (2017), our approach is multivariate in nature and allows for mixed datatypes, but the presence of unobservables leads to results that to the best of our knowledge have not been exploited in the literature. Related work includes Conover (1999, pp. 396–406) who considers a smooth two-sample Kolmogorov–Smirnov test,[2] Bowman, Hall, and Prvan (1998) who considered bandwidth selection for univariate kernel smoothed CDFs, and Wang, Cheng, and Yang (2013) who considered a plug-in bandwidth procedure for smooth univariate kernel smoothed CDFs with a focus on the construction of simultaneous confidence bands.

2.1. Kernel Estimation of $F_X(x)$, $F_\epsilon(t)$, and $F_{X,\epsilon}(x,t)$ With Mixed Data

Though Einmahl and Van Keilegom (2008) and others restricted attention to the scalar predictor case, in applied settings one would expect to encounter multivariate predictors that, in addition, might consist of both discrete and continuous datatypes. Though the Kolmogorov–Smirnov and Cramer–von Mises statistics were developed under the assumption that the random variables possessed continuous distributions $F(\cdot)$ and have been extended to instead admit discrete distributions (Conover 1972; Gleser 1985; Choulakian, Lockhart, and

[1] By "unstructured" we mean a model of the form $Y_i = \mu(X_i) + \epsilon_i$ with $E(\epsilon_i|X_i) = 0$.

[2] This approach compares two kernel smoothed *univariate* distributions; see the function `KS.test` in the R package Qiu (2014) which implements this procedure using Wang, Cheng, and Yang's (2013) plug-in bandwidth and uses the asymptotic distribution for critical values which is known to be problematic.

Stephens 1994; Lockhart, Spinelli, and Stephens 2007), to the best of our knowledge they are unable to handle the multivariate mix of continuous and discrete data often found in regression settings. Our approach tackles this shortcoming by leveraging recent work on nonparametric kernel estimation of distributions involving a mix of discrete and continuous variables.

Li, Li, and Racine (2017) proposed an estimator of a joint distribution function defined over a mix of observed continuous and discrete random variables, which we will explain by means of the vector of covariates X. We suppose that X_j $(j = 1, \ldots, n)$ is a $(q + r)$-dimensional vector of covariates, consisting of q continuous covariates denoted by $X_j^c = (X_{j1}^c, \ldots, X_{jq}^c)$ and r (ordered) discrete covariates denoted by $X_j^d = (X_{j1}^d, \ldots, X_{jr}^d)$. Likewise, X consists of a q-dimensional vector of continuous covariates $X_1^c, \ldots, X_q^c$ and an r-dimensional vector of discrete covariates $X_1^d, \ldots, X_r^d$. The support of X is denoted by $R_X = R_{X^c} \times R_{X^d}$, where R_{X^c} is supposed to be a compact subset of $\mathcal{R}^q$. We consider smooth kernel-based estimators of $F_X(x) = P(X \le x) = P(X^c \le x^c, X^d \le x^d)$ (where inequalities should be understood componentwise).

We consider discrete variables distributed over a finite grid, and without loss of generality assume that X_{js}^d takes values in $\{0, 1, \ldots, c_s - 1\}$ $(s = 1, \ldots, r)$, where $c_s \ge 2$ is a positive integer. Let λ_s denote the bandwidth for the sth discrete variable. We use the kernel function $l(x_s^d, X_{js}^d, \lambda_s) = \eta_s \sum_{z_s^d \le x_s^d} \lambda_s^{|X_{js}^d - z_s^d|}$, with $\lambda_s^0 = 1$, $0^0 = 1$, $\lambda_s \in [0, 1]$, and η_s a normalizing factor such that $l(c_s - 1, c_s - 1, \lambda_s) = 1$. Write the product (discrete variable) cumulative kernel function as $L_\lambda(x^d, X_j^d) = \prod_{s=1}^r l(x_s^d, X_{js}^d, \lambda_s)$.

Let h_s be the bandwidth associated with X_s^c $(s = 1, \ldots, q)$. The product cumulative kernel function used for the continuous variables is given by $K_h(x^c, X_j^c) = \prod_{s=1}^q \int_{-\infty}^{x_s^c} h_s^{-1} k((z_s^c - X_{js}^c)/h_s)\, dz_s^c$, where $k(\cdot)$ is a univariate density kernel function for a continuous variable such as the standard Epanechnikov or Gaussian kernel function.[3] The cumulative kernel function for the vector of mixed variables is simply the product of $K_h(\cdot)$ and $L_\lambda(\cdot)$ defined above and is given by $G_\gamma(x, X_j) = K_h(x^c, X_j^c) \times L_\lambda(x^d, X_j^d)$, where $\gamma = (h, \lambda)$. Li, Li, and Racine (2017) consider the mixed-datatype kernel estimator of $F_X(x)$ defined by

$$\hat{F}_X(x) = \frac{1}{n} \sum_{j=1}^n G_\gamma(x, X_j). \tag{2}$$

Next, to estimate $F_\epsilon(t)$, we assume that ϵ is continuous and hence $F_\epsilon(t)$ can be estimated by a (univariate) continuous cumulative kernel estimator. However, unlike the case considered by Li, Li, and Racine (2017), ϵ is not observed, so we first need to estimate it. To estimate $\mu(x)$ and $\sigma(x)$, we use local polynomial smoothing for the continuous covariates (see Fan and Gijbels (1996) or Ruppert and Wand (1994), among others), and for the discrete covariates, we use a variation on Aitchison and Aitken's (1976) kernel function defined by

$$v(x_s^d, X_{js}^d, \nu_s) = \begin{cases} 1 & \text{if } x_s^d = X_{js}^d \\ \nu_s & \text{otherwise.} \end{cases}$$

The range of ν_s is $[0, 1]$. Note that when $\nu_s = 0$ the above kernel function becomes an indicator function, and when $\nu_s = 1$, it is a constant function. The product kernel function for the vector x^d of discrete covariates is then given by

$$V_\nu(x^d, X_j^d) = \prod_{s=1}^r \nu_s^{1 - 1(x_s^d = X_{js}^d)}.$$

Combining this with local polynomial smoothing of the continuous covariates (of which the order p will depend on the dimension q and will be determined later—see assumption (A2) in Appendix A in the supplementary materials), we define $\hat{\mu}(x) = \hat{\beta}_0$, where $\hat{\beta}_0$ is the first component of the vector $\hat{\beta}$, which is the solution of the local minimization problem

$$\min_\beta \sum_{j=1}^n \Big\{ Y_j - P_j(\beta, x^c, p) \Big\}^2 \prod_{s=1}^q \frac{1}{g_s} k\Big(\frac{x_s^c - X_{js}^c}{g_s} \Big) V_\nu(x^d, X_j^d), \tag{3}$$

where $P_j(\beta, x^c, p)$ is a polynomial of order p built up with all $0 \le k \le p$ products of factors of the form $X_{js}^c - x_s^c$ $(s = 1, \ldots, q)$. The vector β is the vector of length $\sum_{k=0}^p q^k$, consisting of all coefficients of this polynomial. Here, $g = (g_1, \ldots, g_q)$ is a q-dimensional bandwidth vector. To estimate $\sigma^2(x)$, define $\hat{\sigma}^2(x) = \hat{\gamma}_0$, where $\hat{\gamma}_0$ is defined in the same way as $\hat{\beta}_0$, but with Y_j replaced by $(Y_j - \hat{\mu}(X_j))^2$ in (3) $(j = 1, \ldots, n)$.

Then, let $\hat{\epsilon}_j = (Y_j - \hat{\mu}(X_j))/\hat{\sigma}(X_j)$ be the jth residual, and define

$$\hat{F}_{\hat{\epsilon}}(t) = \frac{1}{n} \sum_{j=1}^n \int_{-\infty}^t \frac{1}{b} k\Big(\frac{s - \hat{\epsilon}_j}{b} \Big)\, ds, \tag{4}$$

where $b = b_n$ is the bandwidth for smoothing the residuals. Finally, let

$$\hat{F}_{X,\hat{\epsilon}}(x, t) = \frac{1}{n} \sum_{j=1}^n G_\gamma(x, X_j) \int_{-\infty}^t \frac{1}{b} k\Big(\frac{s - \hat{\epsilon}_j}{b} \Big)\, ds \tag{5}$$

be an estimator of the joint distribution $F_{X,\epsilon}(x, t)$ of (X, ϵ).

Bandwidth selection proceeds via minimization of a cross-validation function, which we explain for the distribution F_X (for F_ϵ and $F_{X,\epsilon}$ similar ideas apply):

$$\mathrm{CV}(\gamma) = \frac{1}{n n_j} \sum_{i=1}^n \sum_{j=1}^{n_j} \Big\{ \mathbf{1}(X_i \le x_j^e) - \hat{F}_{X,-i}(x_j^e) \Big\}^2, \tag{6}$$

where x_j^e, $j = 1, \ldots, n_j$, denotes evaluation points, and where $\hat{F}_{X,-i}(x)$ is the estimator defined in (2) except that the ith data point is removed from the sample. The number of evaluation points can be fixed at, say, $n_j = 100$. This grid of evaluation points plays a role not unlike the number/position of points used for numerical integration. Under quite general conditions and

[3] As noted by a referee, by working with kernel estimators we obtain a test that is not invariant to transformations of the regressors. A remedy could be to replace $k((x_s^c - X_s^c)/h_s)$ by nearest neighbor windows of the form $k((\tilde{F}_{X_s^c}(x_s^c) - \tilde{F}_{X_s^c}(X_s^c))/h_s)$, where $\tilde{F}_{X_s^c}$ is the nonsmoothed empirical distribution of X_s^c. It is clear that this is invariant under any monotone transformation of X_s^c.

using the cross-validated bandwidths, Li, Li, and Racine (2017) obtained the result that

$$\sqrt{n}\left(\hat{F}_X(x) - F_X(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 \frac{\partial^2}{\partial (x_s^c)^2} F_X(x) - \sum_{s=1}^{r} \lambda_s B_s(x)\right) \xrightarrow{d} N(0, V),$$

where $V = F(x)(1 - F(x))$, $\kappa_2 = \int u^2 k(u)\, du$, and

$$B_s(x) = E_{X^d}\Big[\sum_{z^d \le x^d} \mathbf{1}(X^d_{-s} = z^d_{-s}) P(X^d_s = z^d_s) F_{X^c|X^d}(x^c|X^d)\Big], \tag{7}$$

with $F_{X^c|X^d}(x^c|x^d)$ the conditional distribution of X^c given X^d, X^d_{-s} contains all components of X^d except the sth component, and equalities and inequalities should be understood componentwise.

Li, Li, and Racine (2017) delivered a smooth nonparametric estimator that, like its nonsmooth EDF counterpart, achieves a dimension-free $\sqrt{n}$ rate of convergence. The important point to note is that when the underlying distribution is itself smooth, the kernel estimator is capable of delivering estimators that outperform their nonsmooth counterparts in finite-sample settings; see Li, Li, and Racine (2017) for details. In a typical location-scale model with a continuous response and predictor, smoothness of the joint and marginal distributions of the predictor and error term can be safely assumed in a wide range of applications.

3. ASYMPTOTIC PROPERTIES

We start with a preliminary result that gives an iid representation for the estimators $\hat{F}_X(x)$, $\hat{F}_{\hat{\epsilon}}(t)$ and $\hat{F}_{X,\hat{\epsilon}}(x,t)$. The regularity conditions mentioned below, as well as the proofs of the results of this section, can be found in Appendix A (supplementary materials).

Define $\kappa_2 = \int u^2 k(u)\, du$, and more generally for $p \ge 0$, let κ_{p+1} be the first element of the vector $S^{-1}(s_{p+1}, \ldots, s_{2p+1})^T$, where S is the $(p+1) \times (p+1)$ matrix whose (i,j)th entry is s_{i+j-2}, with $s_j = \int u^j k(u)\, du$. Also, let

$$C_s(x) = \sum_{z^d}\Big[1(z^d_s \ne x^d_s)\prod_{t\ne s} 1(z^d_t = x^d_t)\mu(x^c, z^d) - \mu(x)\Big] \times f_X(x^c, z^d),$$

$$D_s(x) = \sum_{z^d}\Big[1(z^d_s \ne x^d_s)\prod_{t\ne s} 1(z^d_t = x^d_t)\sigma^2(x^c, z^d) - \sigma^2(x)\Big] \times f_X(x^c, z^d),$$

for $s = 1, \ldots, r$, where $f_X(x)$ is the joint probability density function of X.

Theorem 1. Assume (A1)–(A6). Then, under H_0 and for any x and t,

$$\begin{aligned}\hat{F}_X(x) - F_X(x) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \le x) - F_X(x) \\ &\quad + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 \frac{\partial^2 F_X(x)}{\partial (x_s^c)^2} + \sum_{s=1}^{r} \lambda_s B_s(x) + R_{n,X}(x),\end{aligned}$$

$$\begin{aligned}\hat{F}_{\hat{\epsilon}}(t) - F_\epsilon(t) &= \frac{1}{n}\sum_{i=1}^{n}\Big[1(\epsilon_i \le t) - F_\epsilon(t) \\ &\quad + f_\epsilon(t)\Big\{\epsilon_i + \frac{t}{2}(\epsilon_i^2 - 1)\Big\}\Big] \\ &\quad + f_\epsilon(t)\int \frac{1}{\sigma(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} g_s^{p+1}\frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\mu(z) \\ &\quad + \sum_{s=1}^{r} \nu_s C_s(z)\Big\} dF_X(z) \\ &\quad + \frac{t}{2} f_\epsilon(t)\int \frac{1}{\sigma^2(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} g_s^{p+1}\frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\sigma^2(z) \\ &\quad + \sum_{s=1}^{r} \nu_s D_s(z)\Big\} dF_X(z) + \frac{\kappa_2}{2} b^2 f'_\epsilon(t) + R_{n,\epsilon}(t),\end{aligned}$$

$$\begin{aligned}\hat{F}_{X,\hat{\epsilon}}(x,t) - F_{X,\epsilon}(x,t) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbf{1}(X_i \le x, \epsilon_i \le t) - F_{X,\epsilon}(x,t) \\ &\quad + f_\epsilon(t)\mathbf{1}(X_i \le x)\Big\{\epsilon_i + \frac{t}{2}(\epsilon_i^2 - 1)\Big\}\Big] \\ &\quad + f_\epsilon(t)\int_{-\infty}^{x} \frac{1}{\sigma(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} g_s^{p+1}\frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\mu(z) \\ &\quad + \sum_{s=1}^{r} \nu_s C_s(z)\Big\} dF_X(z) \\ &\quad + \frac{t}{2} f_\epsilon(t)\int_{-\infty}^{x} \frac{1}{\sigma^2(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} g_s^{p+1}\frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\sigma^2(z) \\ &\quad + \sum_{s=1}^{r} \nu_s D_s(z)\Big\} dF_X(z) \\ &\quad + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 \frac{\partial^2 F_X(x)}{\partial (x_s^c)^2} F_\epsilon(t) + \sum_{s=1}^{r} \lambda_s B_s(x) F_\epsilon(t) \\ &\quad + \frac{\kappa_2}{2} b^2 F_X(x) f'_\epsilon(t) \\ &\quad + R_{n,X,\epsilon}(x,t),\end{aligned}$$

where $\sup_{x\in R_X} |R_{n,X}(x)| = o_P(n^{-1/2})$, $\sup_{t\in\mathcal{R}} |R_{n,\epsilon}(t)| = o_P(n^{-1/2})$, and $\sup_{x\in R_X, t\in\mathcal{R}} |R_{n,X,\epsilon}(x,t)| = o_P(n^{-1/2})$.

An immediate consequence of this theorem is the following corollary. Note that instead of assuming condition (A3) which says that h, λ, and b should tend to zero sufficiently fast, we only require that h, λ, and $b \to 0$. This is because the bias coming from the smoothing of the EDFs disappears in the formula of $\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)$.

Corollary 2. Assume $h, \lambda, b \to 0$, (A1)–(A2) and (A4)–(A6). Then, under H_0 and for any x and t,

$$\begin{aligned}\sqrt{n}\big(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} H(X_i, \epsilon_i, x, t) \\ &\quad + b(x,t) + R_n(x,t),\end{aligned}$$

where

$$H(X_i,\epsilon_i,x,t) = \mathbf{1}(X_i \le x, \epsilon_i \le t) - F_{X,\epsilon}(x,t) - F_\epsilon(t) \times \{\mathbf{1}(X_i \le x) - F_X(x)\} - F_X(x)\{1(\epsilon_i \le t) - F_\epsilon(t)\} + f_\epsilon(t)\{\mathbf{1}(X_i \le x) - F_X(x)\}\Big\{\epsilon_i + \frac{t}{2}(\epsilon_i^2 - 1)\Big\},$$

$$b(x,t) = f_\epsilon(t)\int \big(\mathbf{1}(z \le x) - F_X(x)\big)\frac{1}{\sigma(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} c_{g,s}^{p+1} \times \frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\mu(z) + \sum_{s=1}^{r} c_{v,s}C_s(z)\Big\}dF_X(z) + \frac{t}{2}f_\epsilon(t)\int \big(\mathbf{1}(z \le x) - F_X(x)\big)\frac{1}{\sigma^2(z)}\Big\{\frac{\kappa_{p+1}}{(p+1)!}\sum_{s=1}^{q} c_{g,s}^{p+1} \times \frac{\partial^{p+1}}{\partial (z_s^c)^{p+1}}\sigma^2(z) + \sum_{s=1}^{r} c_{v,s}D_s(z)\Big\}dF_X(z),$$

and $\sup_{x\in R_X, t\in\mathcal{R}} |R_n(x,t)| = o_P(1)$.

Remark 3. Note that some of the bias terms appearing in Theorem 1 cancel out in Corollary 2. Indeed, the biases arising from smoothing the distribution functions $\hat{F}_X(x)$, $\hat{F}_{\hat{\epsilon}}(t)$ and $\hat{F}_{X,\hat{\epsilon}}(x,t)$, disappeared in Corollary 2. Hence, the asymptotic representation of $\sqrt{n}\big(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\big)$ is the same as in the case where the empirical distributions are not smoothed (see Einmahl and Van Keilegom 2008). However, inspection of the proof of Theorem 2.2 in the latter paper reveals that their iid expansion of $\sqrt{n}\big(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\big)$ does not contain the term $f_\epsilon(t)\, n^{-1/2}\sum_{i=1}^n \big\{\mathbf{1}(X_i \le x) - F_X(x)\big\}\big\{\epsilon_i + \frac{t}{2}(\epsilon_i^2 - 1)\big\}$ that shows up in the formula of $n^{-1/2}\sum_{i=1}^n H(X_i,\epsilon_i,x,t)$ given above. This is because the second statement in their Lemma A.1 is wrong, in the sense that the expression on the left hand side is not centered, and so it is certainly not $o_P(n^{-1/2})$. The correct version of their Theorem 2.2 can be obtained from Corollary 4 by using undersmoothing of the bandwidth g_1 (and taking $q = 1$ and $r = 0$).

We are now ready to state the weak convergence of $\sqrt{n}\big(\hat{F}_{X,\hat{\epsilon}} - \hat{F}_X\hat{F}_{\hat{\epsilon}}\big)$ as a process in $\ell^\infty(R_X \times \mathcal{R})$, and the limiting distribution of our test statistics T_{KS} and T_{CM}. Here, $\ell^\infty(R_X \times \mathcal{R})$ is the set of bounded functions from $R_X \times \mathcal{R}$ to $\mathcal{R}$, equipped with the uniform norm.

Corollary 4. Assume $h, \lambda, b \to 0$, (A1)–(A2) and (A4)–(A6).

(i) Under H_0, the process $\sqrt{n}\big(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\big)$ converges weakly in $\ell^\infty(R_X \times \mathcal{R})$ to a Gaussian process $Z(x,t)$ with mean function $b(x,t)$ and covariance function

$$\mathrm{cov}\big(Z(x_1,t_1), Z(x_2,t_2)\big) = E\big[H(X,\epsilon,x_1,t_1)H(X,\epsilon,x_2,t_2)\big]$$

for $x_1, x_2 \in R_X$ and $t_1, t_2 \in \mathcal{R}$.

(ii) Under H_0,

$$T_{\mathrm{KS}} \xrightarrow{d} \sup_{x\in R_X, t\in\mathcal{R}} |Z(x,t)| \quad \text{and} \quad T_{\mathrm{CM}} \xrightarrow{d} \int\int Z^2(x,t)dF_X(x)\,dF_\epsilon(t).$$

4. SECOND ORDER PROPERTIES

We now analyze the source of the power gains that arise from smoothing by considering higher order expansions (we are grateful to an anonymous referee who suggested that we address this issue). Below we demonstrate that the smoothed process has smaller MISE than the nonsmoothed process which arises primarily due to a reduction in variance, which in turn leads to the finite-sample power gains that are evident in the simulations reported in Section 5. For notational convenience we shall restrict attention to the case of one continuous covariate (so $q = 1$ and $r = 0$), but the result can be readily extended to the general case.

We use the notations $\tilde{F}_{\hat{\epsilon}}(t) = n^{-1}\sum_{i=1}^n 1(\hat{\epsilon}_i \le t)$ and $\tilde{F}_{X,\hat{\epsilon}}(x,t) = n^{-1}\sum_{i=1}^n 1(X_i \le x, \hat{\epsilon}_i \le t)$ for the nonsmoothed estimators of $F_\epsilon(t)$ and $F_{X,\epsilon}(x,t)$, respectively.

Theorem 5 demonstrates that the MISE of the smoothed process dominates that of the nonsmoothed process, and the reduction in MISE translates into improved power for our proposed procedure.

Theorem 5. Assume $h, b \to 0$, $ng^7(\log n)^{-1} \to \infty$, $p \ge 3$, (A1)–(A2) and (A4)–(A6). Then,

$$n^{1/2}\big(\hat{F}_{X,\hat{\epsilon}}(x,t) - \hat{F}_X(x)\hat{F}_{\hat{\epsilon}}(t)\big) = n^{-1/2}\sum_{i=1}^n \int\int H(X_i,\epsilon_i, x - uh, t - vb)k(u)k(v)\,du\,dv + b(x,t) + \hat{R}(x,t) := \hat{G}(x,t) + b(x,t) + \hat{R}(x,t)$$

$$n^{1/2}\big(\tilde{F}_{X,\hat{\epsilon}}(x,t) - \tilde{F}_X(x)\tilde{F}_{\hat{\epsilon}}(t)\big) = n^{-1/2}\sum_{i=1}^n H(X_i,\epsilon_i,x,t) + b(x,t) + \tilde{R}(x,t) := \tilde{G}(x,t) + b(x,t) + \tilde{R}(x,t),$$

where $\sup_{x,t}|\hat{R}(x,t)| = O_P((ng)^{-1/6}\log n)$ and $\sup_{x,t}|\tilde{R}(x,t)| = O_P((ng)^{-1/6}\log n)$.

In addition,

$$\int \mathrm{var}(\hat{G}(x,t))\,dx\,dt = \Big[A - h\Psi(k)\Big]\Big[B - b\Psi(k)\Big] + O(h^2 + b^2)$$

$$\int \mathrm{var}(\tilde{G}(x,t))\,dx\,dt = AB + O(h^2 + b^2),$$

where

$$A = \int F_X(x)(1 - F_X(x))\,dx > 0$$

$$B = \int \Big\{F_\epsilon(t)(1 - F_\epsilon(t)) + E\Big[2\big\{I(\epsilon \le t) - F_\epsilon(t)\big\} Q(t,\epsilon) + Q^2(t,\epsilon)\Big]\Big\}dt > 0$$

$$Q(t,\epsilon) = f_\epsilon(t)\Big\{\epsilon + \frac{t}{2}(\epsilon^2 - 1)\Big\}$$

$$\Psi(k) = 2\int uK(u)k(u)\,du > 0.$$

Note that the remainder terms $\hat{R}(x,t)$ and $\tilde{R}(x,t)$ are of the order $O_P(n^{-1/6}n^{1/[6(2p+2)]}\log n)$ uniformly in x and t if

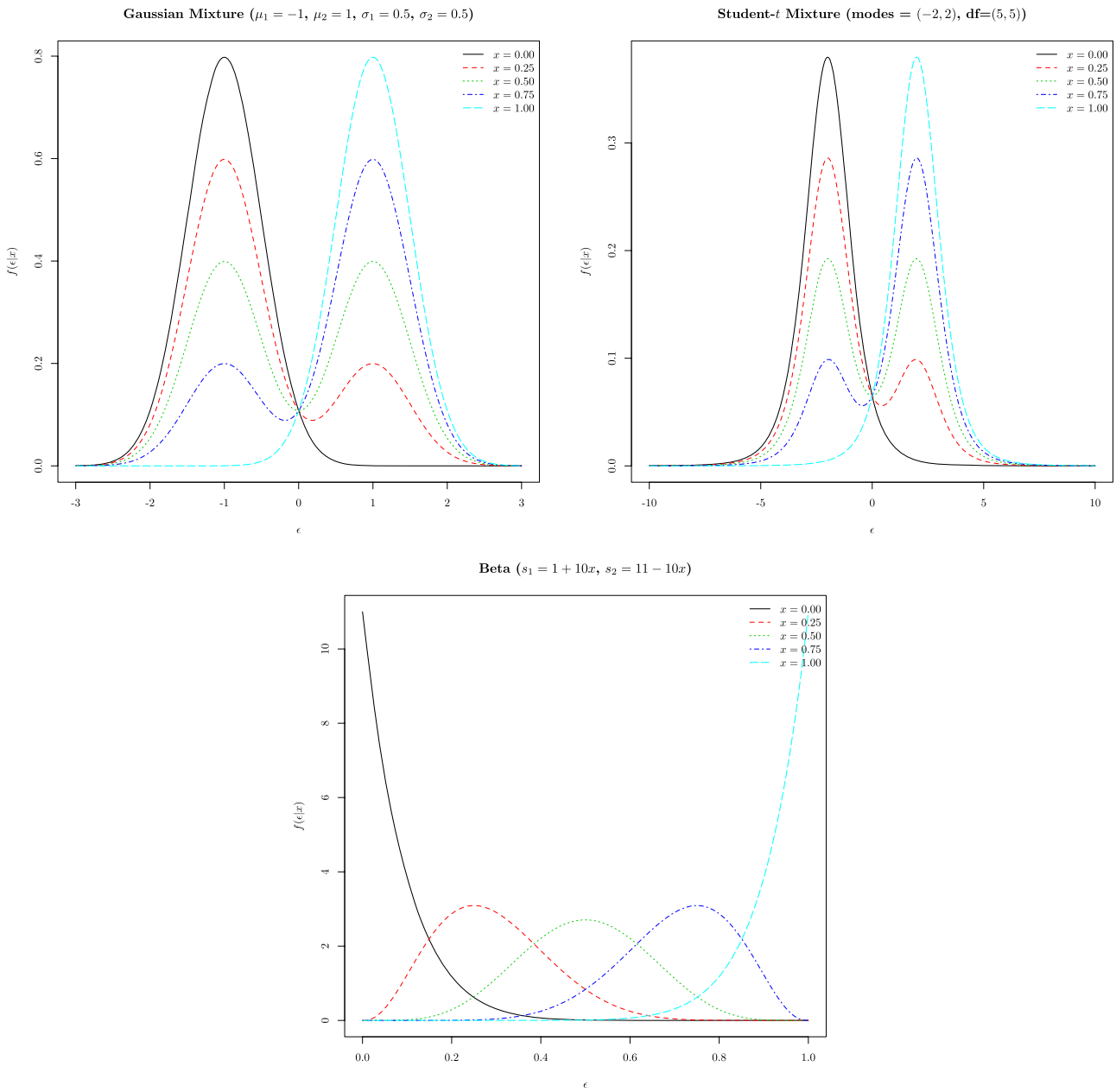

Figure 1. Density of $\epsilon(x)$ for various levels of $x \in [0, 1]$, $\delta = 1$, for the Gaussian mixture, Student's t mixture and Beta (the error density for computing size, $\delta = 0$, corresponds to the density when $x = 0$, i.e., the solid black density).

we take $c_{g,s} > 0$ in condition (A2). Since $p \geq 3$, this is $O_P(n^{-7/48} \log n)$ in the worst case (i.e., when $p = 3$). On the other hand, if h is, for example, proportional to $n^{-1/4}$ (but any larger bandwidth is also fine), then $h^{1/2} \sim n^{-1/8}$, and this is larger than $n^{-7/48} \log n$ (and similarly for $b^{1/2}$). Hence, the remainder terms are of negligible order. Since $\Psi(k), A, B > 0$ the above result shows that the smoothed process has smaller MISE than the nonsmoothed process. The proof of Theorem 5 can be found in Appendix A (supplementary materials).

The astute reader may have noted that the first term in the expansion of $\sqrt{n}(\hat{F}_{X,\hat{\epsilon}} - \hat{F}_X \hat{F}_{\hat{\epsilon}})$ differs from that in Corollary 2, and perhaps a few words are in order. In Corollary 2 we were only interested in first-order asymptotic properties, and so we only needed to demonstrate that the order of the remainder term is $o_P(1)$. In Theorem 5 however, we want to show that the smoothed estimator has better second-order properties than the nonsmoothed estimator, and that requires a finer analysis. The main term in the iid expansion in Theorem 5 is actually equal to the main term in the iid expansion in Corollary 2, except for some terms that are of smaller order and that are absorbed in the remainder term R_n in Corollary 2. In Theorem 5 we cannot absorb these terms in the remainder term simply because we need them to show that the smoothed estimator outperforms the nonsmoothed one in second order. This explains why the two representations are different.

5. FINITE-SAMPLE PERFORMANCE

5.1. The Univariate Continuous Predictor Setting

To assess the finite-sample performance of our proposed approach, we replace the EDFs in (1) with their kernel-

smoothed counterparts described in Section 2.1.[4] Bandwidth selection is obtained via cross-validation for h, b, and λ (i.e., minimization of Equation (6)) and via least squares cross-validation for g and ν, and the Epanechnikov kernel function is employed. There are various data-driven permutations one might consider for bandwidth selection, namely (a) separate bandwidths for the joint and marginal distributions, (b) common bandwidths taken from the joint distribution, and (c) common bandwidths taken from the marginal distributions. From the perspective of assessing size, one might expect that the same bandwidths used for estimating the joint distribution be used for the construction of the marginal distributions, otherwise the estimated joint distribution might systematically diverge from the product of the estimated marginals under the null. We investigate this issue empirically and on this basis recommend (b) to practitioners.

To assess finite-sample performance, we simulate data for which X is uniform $[0, 1]$ and Y has location $\mu(x) = \sin(2\pi x)$ and scale $\sigma(x)$ that is determined from the error distributions specified below, that is,

$$Y_i = \sin(2\pi X_i) + \sigma(X_i)\epsilon_i(X_i).$$

We consider three DGPs in the simulations that follow. For the first the errors are mixtures of two Gaussians, $N(-1, 0.5^2)$ and $N(1, 0.5^2)$ with mixing probabilities $1 - \delta x$ and δx. For the second the errors are drawn from a heavy-tail mixture of two t-distributions, one having a mode at 2 and the other at -2, both having 5 degrees of freedom with the same mixing probabilities as for the Gaussian mixture. For the third the errors are drawn from the Beta distribution with shape parameters $s_1 = 1 + \delta 10x$ and $s_2 = 11 - \delta 10x$ where $x \in [0, 1]$. These errors are then rescaled to have unconditional mean zero and unit variance thereby maintaining a constant signal-to-noise ratio across DGPs and across the range of values for δ considered. In all cases, when $\delta = 0$ the model is a location-scale DGP (i.e., the distribution of ϵ is not a function of x) while when $\delta > 0$ it is a location-shape DGP (i.e., the distribution of ϵ is a function of x). Figure 1 presents the density of $\epsilon(x)$ for various levels of $x \in [0, 1]$ when $\delta = 1$.

For what follows we consider $M = 1000$ Monte Carlo replications drawn from each DGP. For each Monte Carlo replication, we compute each test statistic T_{KS} (nonsmooth and smooth, respectively) along with the $B = 399$ (Davidson and MacKinnon 2000) null bootstrap replicates $T^*_{b,KS}$, $b = 1, \ldots, B$ (based on resamples drawn from the (nonsmooth) distribution functions), then we compute the empirical p-value as $B^{-1}\sum_{b=1}^{B} \mathbf{1}(T^*_{b,KS} > T_{\mathrm{KS}})$, where $\mathbf{1}(A)$ is the usual indicator function taking value one when A is true and zero otherwise. Finally, based on the $M = 1000$ p-values, we compute empirical rejection probabilities for the nonsmooth and smooth test statistics for nominal levels $\alpha = (0.01, 0.05, 0.10)$. To assess size we set $\delta = 0$, and to assess power we let $\delta \in (0, 1]$. Size (i.e., empirical rejection probability when $\delta = 0$) is summarized in Table 1.

Table 1. Size for the nonsmooth and smooth tests (empirical rejection probabilities under the null, $\delta = 0$, univariate continuous predictor setting, see Appendix B in the supplementary materials for power).

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
	Nonsmooth, Gaussian mixture		
100	0.008	0.041	0.101
200	0.008	0.045	0.089
400	0.006	0.048	0.089
800	0.006	0.045	0.102
	Smooth, Gaussian mixture		
100	0.011	0.045	0.104
200	0.006	0.046	0.093
400	0.007	0.057	0.110
800	0.010	0.047	0.092
	Nonsmooth, Student-t mixture		
100	0.009	0.051	0.081
200	0.009	0.050	0.104
400	0.005	0.039	0.095
800	0.006	0.038	0.089
	Smooth, Student-t mixture		
100	0.009	0.048	0.101
200	0.010	0.052	0.095
400	0.010	0.045	0.092
800	0.008	0.042	0.095
	Nonsmooth, Beta		
100	0.012	0.049	0.090
200	0.008	0.050	0.105
400	0.008	0.038	0.099
800	0.011	0.048	0.098
	Smooth, Beta		
100	0.011	0.042	0.074
200	0.012	0.051	0.095
400	0.008	0.054	0.110
800	0.013	0.054	0.104

Table 1 indicates that both the nonsmooth and smooth versions of the test appear to be correctly sized, hence we can proceed to compare their power curves.[5] for the Beta errors in Figure 4 (Figures 2 and 3 present results for the Gaussian and Student's t mixtures).

The percentage gain in power is reported in the tables in Appendix B in the supplementary materials, and these figures and tables reveal that the improvements in power arising from replacing the EDF with its kernel-smoothed counterpart can be upward of 100% or more, depending on the nominal size of the test, sample size, and degree of departure from the null. Furthermore, if anything the smooth test appears to be slightly conservative relative to its nonsmooth counterpart, particularly

[4]We proceed with the Kolmogorov–Smirnov statistic by way of illustration as the Cramer–von Mises statistic requires multivariate integration for its computation while the Kolmogorov–Smirnov statistic does not. However, as will be demonstrated for the Kolmogorov–Smirnov approach, replacing the nonsmooth distribution functions in the Cramer–von Mises statistic with their kernel-smoothed counterparts would be expected to reveal similar power gains.

[5]If anything, the nonsmooth version appears to be slightly over-sized for smaller n and the smooth version slightly under-sized for smaller n, but this admits a fair comparison of power curves. Next, we vary $\delta \in [0, 1]$, and present power curves.

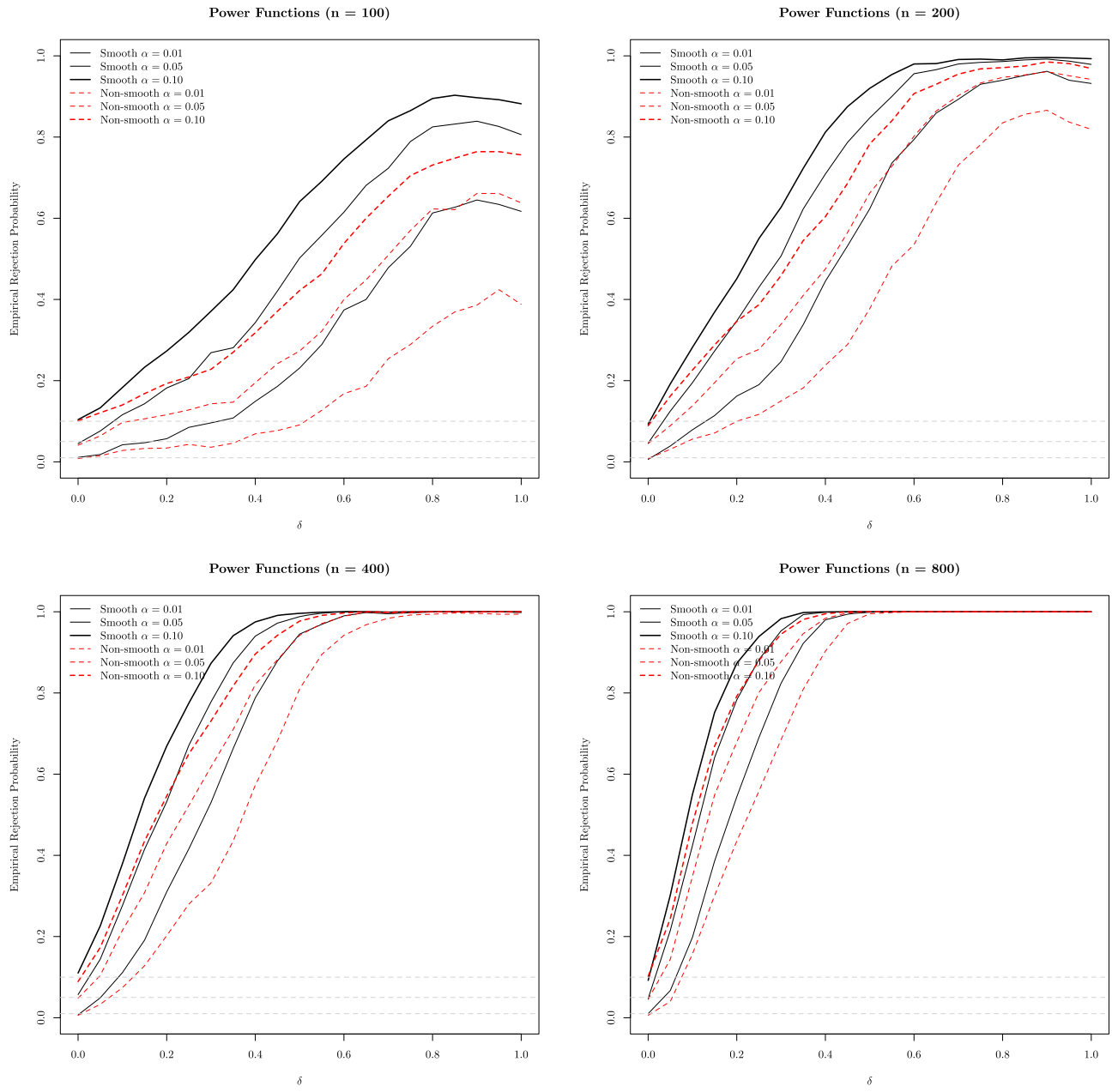

Figure 2. Power curves for the nonsmooth and smooth versions of the test, Gaussian mixture. Solid lines are for the smooth version, dashed the nonsmooth version.

for the Beta error case for the smaller sample sizes considered (a positive feature as it has a lower probability of a Type I error than its nonsmooth counterpart under the null yet higher power under the alternative).

5.2. The Multivariate Continuous Predictor Setting

To assess finite-sample performance in the multivariate continuous predictor setting, we simulate data for which X_1 and X_2 are uniform $[0, 1]$ and Y has location $\mu(x) = \sin(\pi(x_1 + x_2))$ and scale $\sigma(x_1 + x_2)/2$ that is determined from the error distributions specified below, that is,

$$Y_i = \sin(\pi(X_{i1} + X_{i2})) + \sigma(X_{i1} + X_{i2})\epsilon_i(X_{i1} + X_{i2}).$$

We consider three DGPs in the simulations that follow. For the first the errors are mixtures of two Gaussians, $N(-1, 0.5^2)$ and $N(1, 0.5^2)$ with mixing probabilities $1 - \delta(x_1 + x_2)/2$ and $\delta(x_1 + x_2)/2$. For the second the errors are drawn from a heavy-tail mixture of two t-distributions, one having a mode at 2 and the other at -2, both having 5 degrees of freedom, with the same mixing probabilities as for the Gaussian mixture. For the third the errors are drawn from the Beta distribution with shape parameters $s_1 = 1 + \delta 5(x_1 + x_2)$ and $s_2 = 11 - \delta 5(x_1 + x_2)$. Per above, these errors are then rescaled to have unconditional mean zero and unit variance.

Table 2 reveals that both the nonsmooth and smooth version of the test perform adequately and appear to possess

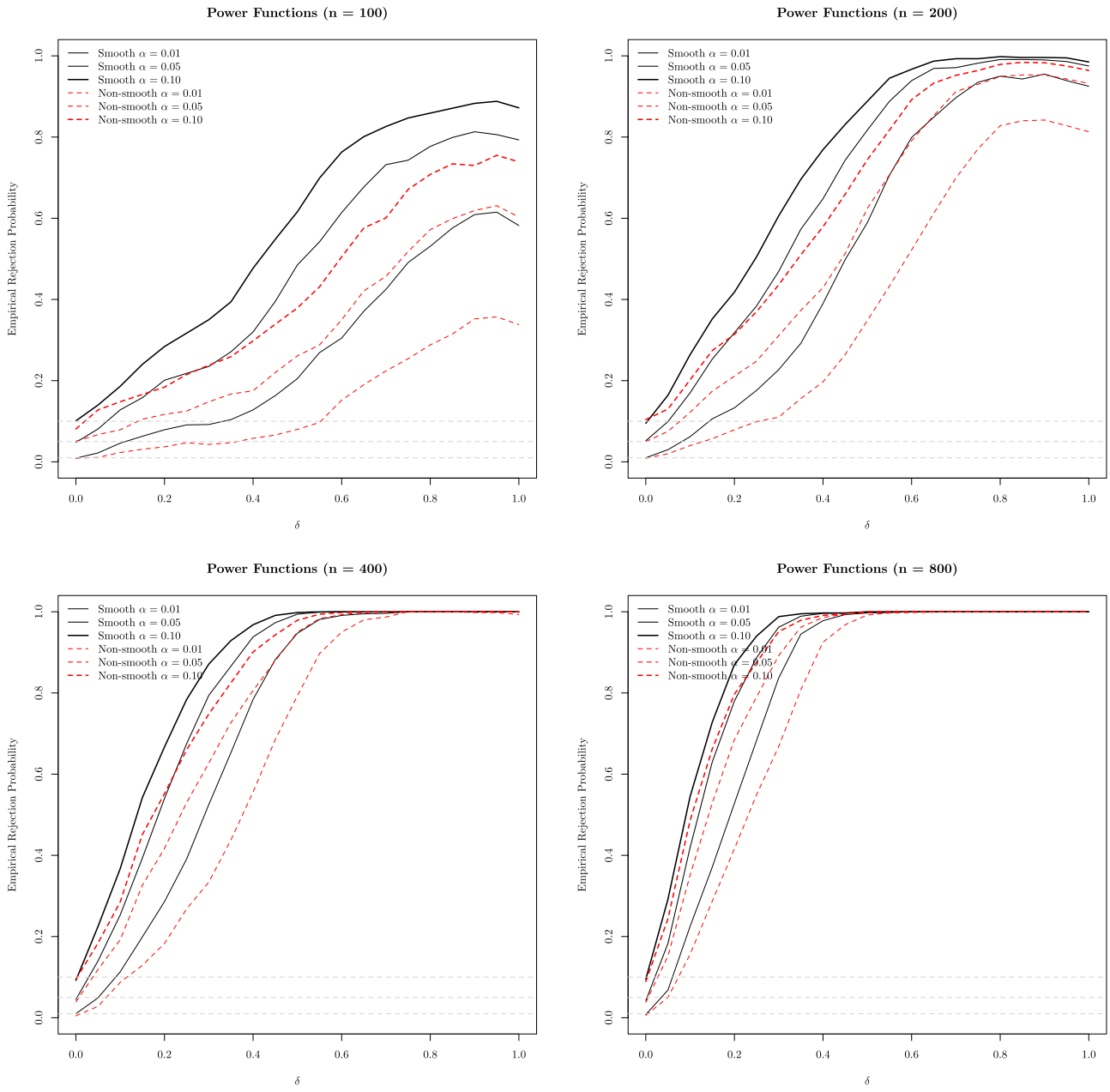

Figure 3. Power curves for the nonsmooth and smooth versions of the test, Student's t mixture. Solid lines are for the smooth version, dashed the nonsmooth version.

reasonable size when there exist multivariate continuous predictors.

5.3. The Multivariate Mixed Continuous and Discrete Predictor Setting

To assess finite-sample performance in the multivariate mixed continuous and discrete predictor setting, we simulate data for which X_1 is uniform $[0, 1]$, X_2 is a discrete uniform variable taking value $0, 1/10, 2/10, \ldots, 1$, Y has location $\mu(x) = \sin(2\pi x_1) + x_2$ and scale $\sigma(X_i)$ that is determined from the error distributions specified below, that is,

$$Y_i = \sin(2\pi X_{i1}) + X_{i2} + \sigma(X_{i1} + X_{i2})\epsilon_i(X_{i1} + X_{i2}).$$

We consider three DGPs in the simulations that follow. For the first the errors are mixtures of two Gaussians, $N(-1, 0.5^2)$ and $N(1, 0.5^2)$ with mixing probabilities $1 - \delta(x_1 + x_2)/2$ and $\delta(x_1 + x_2)/2$. For the second the errors are drawn from a heavy-tail mixture of two t-distributions, one having a mode at 2 and the other at -2, both having 5 degrees of freedom, with the same mixing probabilities as for the Gaussian mixture. For the third the errors are drawn from the Beta distribution with shape parameters $s_1 = 1 + \delta 5(x_1 + x_2)$ and $s_2 = 11 - \delta 5(x_1 + x_2)$. Per above, these errors are then rescaled to have unconditional mean zero and unit variance.

Table 3 reveals a curious feature of the nonsmooth test—in the presence of discrete predictors, the test displays extreme size distortions rendering it completely unsuited for

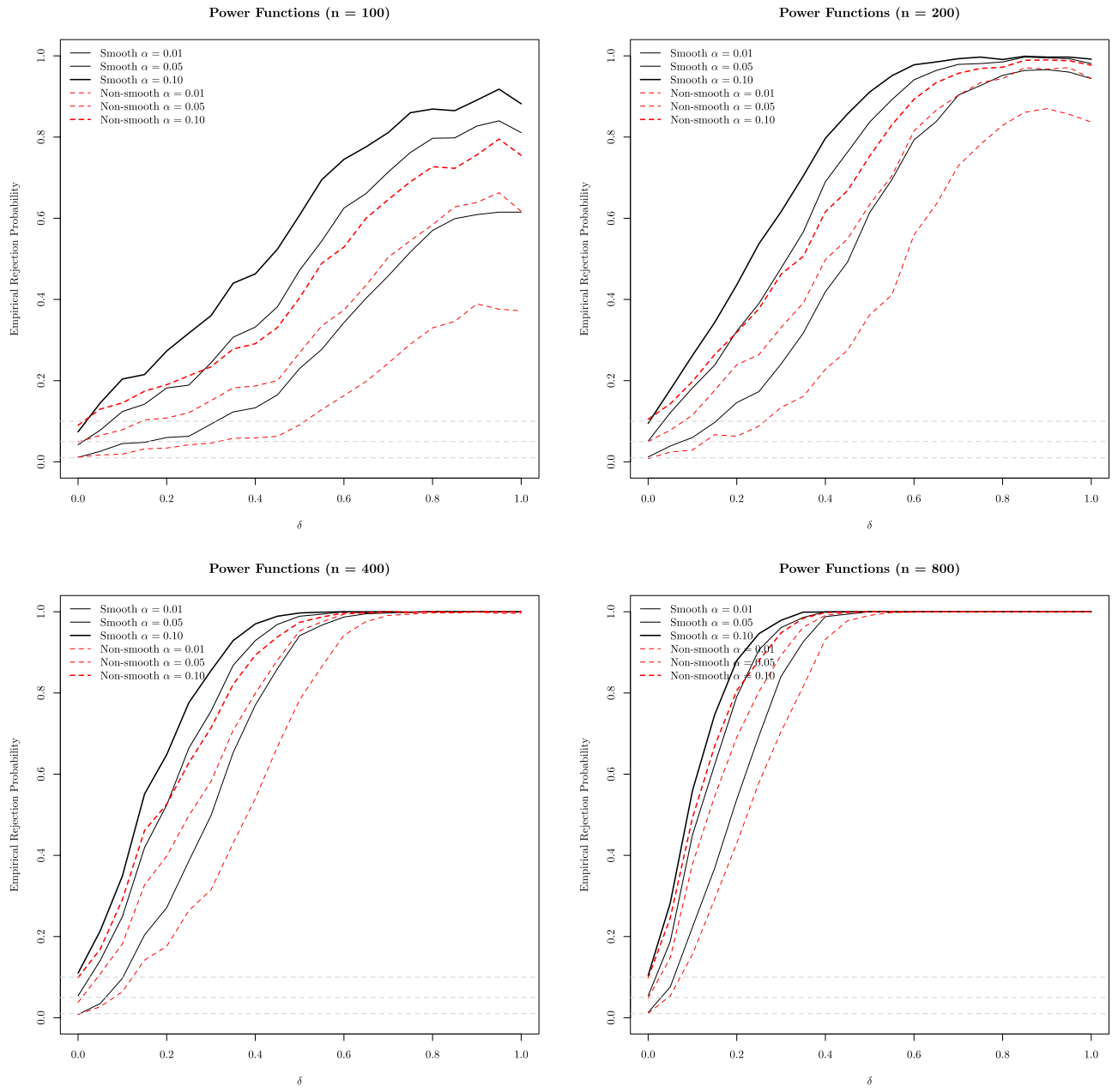

Figure 4. Power curves for the nonsmooth and smooth versions of the test, Beta errors. Solid lines are for the smooth version, dashed the nonsmooth version.

practical application (its empirical rejection probability under the null approaches 1 as n increases for all conventional levels). However, the smooth version of the test appears to be reasonably sized. The former is perhaps not too surprising given the literature on discrete/discontinuous distributions (Conover 1972; Gleser 1985; Choulakian, Lockhart, and Stephens 1994; Lockhart, Spinelli, and Stephens 2007). Given the extreme size distortions present, we make no attempt at power comparisons.

6. APPLICATION

Racine and Li (2017) imposed a location-scale quantile model structure on a novel nonparametric quantile estimator that is based on kernel smoothing of a parametric quantile function in a particular manner. A practitioner concerned with their imposition of the location-scale structure might wish to use a pretest approach, proceeding with the location-scale model if it is deemed appropriate versus an alternative model that does not rely on the location-scale assumption otherwise. They present two illustrative applications, one in which the covariate is continuous and one in which it is discrete, so these illustrations will serve to highlight the potential application of the proposed procedure.

We first consider an Italian gross domestic product (GDP) growth panel for 21 regions covering the period 1951–1998 (millions of Lire, 1990 = base). There are $n = 1008$ observations on 2 variables, "year" and "gdp," and "year" is

Table 2. Size for the nonsmooth and smooth tests (empirical rejection probabilities under the null, $\delta = 0$, multivariate continuous predictor setting, see Appendix C in the supplementary materials for power).

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
	Nonsmooth, Gaussian mixture		
100	0.007	0.042	0.100
200	0.007	0.042	0.094
400	0.012	0.038	0.075
800	0.005	0.037	0.075
	Smooth, Gaussian mixture		
100	0.005	0.033	0.091
200	0.008	0.048	0.087
400	0.006	0.032	0.072
800	0.006	0.037	0.073
	Nonsmooth, Student's t mixture		
100	0.018	0.056	0.116
200	0.018	0.080	0.121
400	0.012	0.053	0.107
800	0.008	0.055	0.095
	Smooth, Student's t mixture		
100	0.023	0.078	0.140
200	0.016	0.085	0.140
400	0.017	0.068	0.111
800	0.012	0.057	0.109
	Nonsmooth, Beta		
100	0.005	0.042	0.100
200	0.021	0.074	0.118
400	0.018	0.062	0.124
800	0.019	0.064	0.108
	Smooth, Beta		
100	0.011	0.057	0.104
200	0.015	0.059	0.121
400	0.020	0.070	0.127
800	0.018	0.060	0.112

Table 3. Size for the nonsmooth and smooth tests (empirical rejection probabilities under the null, $\delta = 0$, multivariate mixed continuous and discrete predictor setting).

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
	Nonsmooth, Gaussian mixture		
100	0.285	0.545	0.678
200	0.625	0.836	0.918
400	0.950	0.994	0.998
800	1.000	1.000	1.000
	Smooth, Gaussian mixture		
100	0.009	0.046	0.090
200	0.013	0.056	0.100
400	0.015	0.058	0.116
800	0.024	0.081	0.159
	Nonsmooth, Student's t mixture		
100	0.361	0.632	0.752
200	0.773	0.933	0.974
400	0.993	1.000	1.000
800	1.000	1.000	1.000
	Smooth, Student's t mixture		
100	0.014	0.054	0.096
200	0.013	0.061	0.108
400	0.014	0.063	0.125
800	0.012	0.074	0.140
	Nonsmooth, Beta		
100	0.518	0.782	0.877
200	0.951	0.991	0.998
400	1.000	1.000	1.000
800	1.000	1.000	1.000
	Smooth, Beta		
100	0.006	0.027	0.074
200	0.004	0.040	0.084
400	0.012	0.051	0.124
800	0.023	0.110	0.200

a discrete predictor and is treated as such in what follows. Next we consider Canadian cross-section wage data consisting of a random sample taken from the 1971 Canadian Census Public Use Tapes for male individuals having common education (grade 13). There are $n = 205$ observations in total on two variables, "logwage" (logarithm of the wages) and "age," age being treated as a continuous predictor. We report the test statistics and their bootstrapped p-values in Table 4 based on $B = 999$ bootstrap replications.

Table 4 reveals that the location-scale presumption is inappropriate for the Italian GDP data, but is appropriate for the Canadian wage data (though it is possible, given the small sample size of $n = 205$, that this is a Type II error). These results indicate that practitioners can use Racine and Li's (2017) quantile approach for the latter but ought to exercise caution when applying it to the former, which is consistent with the findings presented in Racine and Li (2017) which used instead the nonsmooth testing approach of Einmahl and Van Keilegom (2008).

Table 4. Application of the proposed test to the Italian GDP Dataset and to the Canadian Cross-Section Wage Dataset.

Dataset	T_{KS}	p-value
Canadian Wage	0.4637124	0.3413413
Italian GDP	1.174512	<2e−16

7. CONCLUDING REMARKS

Numerous tests have been proposed that rely on the nonsmooth EDF for their implementation; see Einmahl and Van Keilegom (2008), Birke, Neumeyer, and Volgushev (2017), and Neumeyer, Noh, and Van Keilegom (2016) by way of illustration. We demonstrate how the use of smooth estimators of distribution functions rather than their nonsmooth counterparts can deliver tests having superior power profiles, which ought to be particularly appealing for practitioners.

FUNDING

Racine would like to gratefully acknowledge support from the Natural Sciences and Engineering Research Council of Canada (NSERC: *www.nserc.ca*), the Social Sciences and Humanities Research Council of Canada (SSHRC: *www.sshrc.ca*), and the Shared Hierarchical Academic Research Computing Network (SHARCNET: *www.sharcnet.ca*). I. Van Keilegom acknowledges financial support from the European Research Council (2016–2021, Horizon 2020/ERC grant agreement no. 694409).

[Received February 2018. Revised December 2018.]

REFERENCES

Aitchison, J., and Aitken, C. G. G. (1976), "Multivariate Binary Discrimination by the Kernel Method," *Biometrika*, 63, 413–420. [786]

Akritas, M., and Van Keilegom, I. (2001), "Nonparametric Estimation of the Residual Distribution," *Scandinavian Journal of Statistics*, 28, 549–568. [784,785]

Anderson, T. W., and Darling, D. A. (1952), "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes," *The Annals of Mathematical Statistics*, 23, 193–212. [784]

Billingsley, P. (1999), *Convergence of Probability Measures*, New York: Wiley.

Birke, M., Neumeyer, N., and Volgushev, S. (2017), "The Independence Process in Conditional Quantile Location-Scale Models and an Application to Testing for Monotonicity," *Statistica Sinica*, 27, 1815–1839. [784,794]

Bowman, A. W., Hall, P., and Prvan, T. (1998), "Bandwidth Selection for the Smoothing of Distribution Functions," *Biometrika*, 85, 799–808. [785]

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994), "Cramer-von Mises Statistics for Discrete Distributions," *The Canadian Journal of Statistics*, 22, 125–137. [786,793]

Conover, W. J. (1972), "A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions," *Journal of the American Statistical Association*, 67, 591–596. [785,793]

——— (1999), *Practical Nonparametric Statistics* (3rd ed.), New York: Wiley. [785]

Cramér, H. (1928), "On the Composition of Elementary Errors," *Scandinavian Actuarial Journal*, 1, 13–74. [784]

Davidson, R., and MacKinnon, J. G. (2000), "Bootstrap Tests: How Many Bootstraps?," *Econometric Reviews*, 19, 55–68. [790]

Dette, H., Neumeyer, N., and Van Keilegom, I. (2007), "A New Test for the Parametric Form of the Variance Function in Non-parametric Regression," *Journal of the Royal Statistical Society*, Series B, 69, 903–917. [784]

Einmahl, J. H., and Van Keilegom, I. (2008), "Specification Tests in Nonparametric Regression," *Journal of Econometrics*, 143, 88–102. [784,785,788,794]

Escanciano, J., and Jacho-Chávez, D. T. (2012), "$\sqrt{n}$-Uniformly Consistent Density Estimation in Nonparametric Regression Models," *Journal of Econometrics*, 12, 305–361. [785]

Escanciano, J. C., Pardo-Fernández, J. C., and Van Keilegom, I. (2018), "Asymptotic Distribution Free Tests for Semiparametric Regressions With Dependent Data," *Annals of Statistics*, 46, 1167–1196. [784]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall. [786]

Gijbels, I., Van Keilegom, I., and Zhao, Y. (2018), "Gaussian Copulas Adjusted for Nonparametric Regression," Working Paper, KU Leuven.

Gleser, L. J. (1985), "Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions," *Journal of the American Statistical Association*, 80, 954–958. [785,793]

Jones, M. (1990), "The Performance of Kernel Density Functions in Kernel Distribution Function Estimation," *Statistics & Probability Letters*, 9, 129–132.

Kolmogorov, A. (1933), "Sulla determinazione empirica di una legge di distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, 4, 1–11. [784]

Li, C., Li, H., and Racine, J. S. (2017), "Cross-Validated Mixed-Datatype Bandwidth Selection for Nonparametric Cumulative Distribution/Survivor Functions," *Econometric Reviews*, 36, 970–987. [785,786,787]

Li, Q., and Racine, J. S. (2004), "Cross-Validated Local Linear Nonparametric Regression," *Statistica Sinica*, 14, 485–512.

Lockhart, R., Spinelli, J., and Stephens, M. (2007), "Cramér-von Mises Statistics for Discrete Distributions With Unknown Parameters," *Canadian Journal of Statistics*, 35, 125–133. [786,793]

Neumeyer, N. (2008), "A Bootstrap Version of the Residual-Based Smooth Empirical Distribution Function," *Journal of Nonparametric Statistics*, 20, 153–174. [784,785]

——— (2009), "Smooth Residual Bootstrap for Empirical Processes of Nonparametric Regression Residuals," *Scandinavian Journal of Statistics*, 36, 204–228. [784]

Neumeyer, N., Noh, H., and Van Keilegom, I. (2016), "Heteroscedastic Semiparametric Transformation Models: Estimation and Testing for Validity," *Statistica Sinica*, 26, 925–954. [784,785,794]

Neumeyer, N., and Van Keilegom, I. (2010), "Estimating the Error Distribution in Nonparametric Multiple Regression With Applications to Model Testing," *Journal of Multivariate Analysis*, 101, 1067–1078.

——— (2018), "Bootstrap of Residual Processes in Regression: To Smooth or Not to Smooth?," *Biometrika* (to appear). [784,785]

Pardo-Fernández, J. C., Van Keilegom, I., and González-Manteiga, W. (2007), "Testing for the Equality of k Regression Curves," *Statistica Sinica*, 17, 1115–1137. [784]

Qiu, D. (2014), "snpar: Supplementary Non-parametric Statistics Methods," R Package Version 1.0, available at *https://CRAN.R-project.org/package=snpar*. [785]

Racine, J. S., and Li, K. (2017), "Nonparametric Conditional Quantile Estimation: A Locally Weighted Quantile Kernel Approach," *Journal of Econometrics*, 201, 72–94. [784,785,793,794]

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *Annals of Statistics*, 22, 1346–1370. [786]

Smirnov, N. (1948), "Table for Estimating the Goodness of Fit of Empirical Distributions," *Annals of Mathematical Statistics*, 19, 279–281. [784]

Stute, W. (1982), "The Oscillation Behavior of Empirical Processes," *The Annals of Probability*, 10, 86–107.

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

Van Keilegom, I., González-Manteiga, W., and Sánchez-Sellero, C. (2008), "Goodness-of-Fit Tests in Parametric Regression Based on the Estimation of the Error Distribution," *TEST*, 17, 401–415. [784]

von Mises, R. E. (1928), *Wahrscheinlichkeit, Statistik und Wahrheit*, Vienna: Julius Springer. [784]

Wang, J., Cheng, F., and Yang, L. (2013), "Smooth Simultaneous Confidence Bands for Cumulative Distribution Functions," *Journal of Nonparametric Statistics*, 25, 395–407. [785]