Journal of Econometrics 148 (2009) 186–200

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A nonparametric test for equality of distributions with mixed categorical and continuous data

Qi Li [a,b,*], Esfandiar Maasoumi [c], Jeffrey S. Racine [d]

[a] Department of Economics, Texas A&M University, College Station, TX 77843-4228, United States
[b] Department of Economics Tsinghua University, Beijing 100084, PR China
[c] Department of Economics, Emory University, Atlanta, GA 30322, United States
[d] Department of Economics, McMaster University, Hamilton, Ontario, Canada L8S 4M4

ARTICLE INFO

Article history:
Available online 28 October 2008

Keywords:
Mixed discrete and continuous variables
Density testing
Nonparametric smoothing
Cross-validation

ABSTRACT

In this paper we consider the problem of testing for equality of two density or two conditional density functions defined over mixed discrete and continuous variables. We smooth both the discrete and continuous variables, with the smoothing parameters chosen via least-squares cross-validation. The test statistics are shown to have (asymptotic) normal null distributions. However, we advocate the use of bootstrap methods in order to better approximate their null distribution in finite-sample settings and we provide asymptotic validity of the proposed bootstrap method. Simulations show that the proposed tests have better power than both conventional frequency-based tests and smoothing tests based on ad hoc smoothing parameter selection, while a demonstrative empirical application to the joint distribution of earnings and educational attainment underscores the utility of the proposed approach in mixed data settings.

1. Introduction

It is difficult to think of a more ubiquitous test in applied statistics than the test for equality of distributions and conditional distributions, sometimes conditioned on discrete covariates. The most popular variants are derivative since they involve testing equality of moments, such as means and/or variances, or perhaps quantiles. Examples include tests for 'regime change', heteroskedasticity, and 'symmetry'. Also, comparing distributions, or reconstructing indirectly observed distributions (such as the counter factuals in program evaluation) is implicit and ever present in almost all statistical/econometric work. However, moment-based tests, which only compare a finite number of moments from two distributions, are not consistent tests. The same can be said for parametric tests which require specification of the null distribution. When the parametric null distribution is misspecified, parametric tests can lead to erroneous conclusions. Generally, interest truly lies in detecting *any* potential difference between two distributions without having to specify a parametric family, not just their means or variances. When this is the case, nonparametric tests have obvious appeal.

A number of kernel-based tests of equality of distribution functions exist; however, existing kernel-based tests presume that the underlying variable is *continuous* in nature; see Ahmad and van Belle (1974), Mammen (1992), Fan and Gencay (1993), Li (1996) and Fan and Ullah (1999), and the references therein. It is widely known that a traditional 'frequency-based' kernel approach could be used to consistently estimate a joint probability function in the presence of mixed continuous and categorical variables, and hence one could readily construct a kernel-based test for the equality between two unknown density functions by simply employing the conventional frequency kernel method. In contrast we consider kernel 'smoothing' the discrete variables as well, following a rich literature in statistics on smoothing discrete variables and its potential benefits; see Aitchison and Aitken (1976), Hall (1981), Grund and Hall (1993), Scott (1992), Simonoff (1996), Li and Racine (2003) and Hall et al. (2004, 2007), among others. Though smoothing discrete variables in an appropriate manner may introduce some finite-sample bias, it simultaneously reduces finite-sample variance substantially, and leads to a reduction in the finite-sample mean square error of the nonparametric estimator relative to the frequency-based estimator. It turns out that, for testing purposes, this is also highly desirable. The tests developed herein are extensions of existing frequency-based 'smooth' kernel tests. 'Non-smooth' (i.e., empirical cumulative distribution function (CDF)) tests of distributional differences have recently been examined and reviewed in Anderson (2001).

* Corresponding author at: Department of Economics, Texas A&M University, College Station, TX 77843-4228, United States. Tel.: +1 979 845 9954.
E-mail address: qi@econmail.tamu.edu (Q. Li).

0304-4076/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2008.10.007

In this paper we first propose a kernel-based test for equality of distributions mounted on a square integral metric defined over mixed continuous/discrete variables. We then extend our result to the case of testing the equality of two conditional distributions. Conditional distributions, such as that of earnings given gender, age or education categories, are often the main targets of inference and policy analysis. As an alternative to our approach in this paper, entropy metrics have been used for testing equality of distributions, or hypotheses which may be cast as such. For a pioneering paper see Robinson (1991), as well as Hong and White (2005), Ahmad and Li (1997) and Racine and Maasoumi (2007). We use data-driven bandwidth selection methods, smooth both the continuous and discrete variables in a particular manner, and advocate a resampling method for obtaining the statistic's null distribution, though we also provide its limiting (asymptotic) null distribution and prove that the bootstrap works. It is well known that the selection of smoothing parameters is of crucial importance in nonparametric estimation, and it is now known that the selection of smoothing parameters also affects the power of nonparametric tests such as ours. When discrete variables are present, cross-validation has been shown to be an effective method of smoothing parameter selection. Not only is there a large sample optimality property associated with minimizing estimation mean square error, but we also avoid sample splitting in small sample applications. When one smooths *both* the discrete and continuous variables, cross-validation seems to be the only feasible way of selecting the smoothing parameters. Configuring plug-in rules for mixed data is an algebraically tedious task, and no general formulae are yet available. Additionally, plug-in rules, even after adaption to mixed data, require choice of 'pilot' smoothing parameters, and it is not clear how to best make that selection when both continuous and discrete variables are involved.

We believe the improved power of the proposed tests is partly due to smoothing over discrete variables. Heuristically, this type smoothing is equivalent to endowing "degrees of likelihood" to the set of potential values for a discrete variable. For a Bayesian interpretation see Kiefer and Racine (2008). This additional information increases power.

The paper is organized as follows. Section 2 presents a test for the equality of two unconditional distribution functions and examines the asymptotic distribution of the test statistic, Section 3 proposes a test for the equality of two conditional density functions, Section 4 presents two simulation experiments designed to assess the finite-sample performance of the estimator, while Section 5 presents a demonstrative empirical application that tests for differences in the joint distribution of earnings and educational attainment over time. Section 6 concludes, and all proofs are relegated to the Appendix.

2. A nonparametric test for the equality of unconditional density functions with mixed categorical and continuous data

2.1. Testing the equality of two density functions

We consider the case where we are faced with a mixture of discrete and continuous data. Let $X = (X^c, X^d) \in \mathbb{R}^q \times \mathbb{S}^r$, where X^c is the continuous variable having dimension q, and X^d is the discrete variable having dimension r. Let x_s^d and X_{is}^d denote the sth components of x^d and X_i^d respectively. Following Aitchison and Aitken (1976), for $x_s, X_{is}^d \in \mathbb{S}_s^r = \{a_1, a_2, \ldots, a_{c_s}\}$ (x_s^d takes c_s different values) so that X^d assumes values in $\mathbb{S}^r = \prod_{s=1}^r \{a_1, a_2, \ldots, a_{c_s}\}$. Similarly, $Y = (Y^c, Y^d)$, which has the same dimension as X. Let $f(\cdot)$ and $g(\cdot)$ denote the density functions of X and Y, respectively, and let $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$ be i.i.d. random draws from populations having density functions $f(\cdot)$ and $g(\cdot)$, respectively.[1] We are interested in testing the null hypothesis that

$$H_0: f(x) = g(x) \quad \text{for } x^d \in \mathbb{S}^r \text{ and for almost all } x^c \in \mathbb{R}^q$$

against the alternative hypothesis H_1 that $f(x) \neq g(x)$ for some $x^d \in \mathbb{S}^r$ or for some x^c on a set with positive measure. We first discuss how to estimate $f(\cdot)$ and $g(\cdot)$ and then outline the test statistic. We define a univariate kernel function

$$l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 1-\lambda_s & \text{if } X_{is}^d = x_s^d, \\ \lambda_s/(c_s-1) & \text{if } X_{is}^d \neq x_s^d, \end{cases} \tag{2.1}$$

where the range of the smoothing parameter λ_s is $[0, (c_s-1)/c_s]$. Note that when $\lambda_s = 0$, $l(X_{is}^d, x_s^d, 0) = I(X_{is}^d = x_s^d)$ becomes an indicator function. We shall use $I(\cdot)$ to denote an indicator function, i.e., $I(A) = 1$ if the event A holds true, otherwise $I(A) = 0$. Observe that if $\lambda_s = (c_s-1)/c_s$, then $l(X_{is}^d, x_s^d, \frac{c_s-1}{c_s}) = 1/c_s$ which is a constant for *all* values of X_{is}^d and x_s^d.

A product kernel function for the discrete variable components x^d is given by

$$L_{\lambda,x_i,x} = \prod_{t=1}^r l(X_{is}^d, x_s^d, \lambda_s) = \prod_{s=1}^r \{\lambda_s/(c_s-1)\}^{I_{x_{is}^d \neq x_s^d}} (1-\lambda_s)^{I_{x_{is}^d = x_s^d}}, \tag{2.2}$$

where $I_{x_{is}^d \neq x_s^d} = I(X_{is}^d \neq x_s^d)$ and $I_{x_{is}^d = x_s^d} = I(X_{is}^d = x_s^d)$. Here $I(A)$ is an indicator function which takes value one if A holds true, zero otherwise.

Let $w\left(\frac{X_{is}^c - x_s^c}{h_s}\right)$ be a univariate kernel function associated with the continuous variable x_s^c, where h_s is the associated smoothing parameter. The product kernel for the continuous variable components x^c is given by $W_{h,x_i,x} = \prod_{s=1}^q h_s^{-1} w\left(\frac{X_{is}^c - x_s^c}{h_s}\right)$.

The 'generalized' product kernel defined over both discrete and continuous variables is given by

$$K_{\gamma,x_i,x} = W_{h,x_i,x} L_{\lambda,x_i,x}, \tag{2.3}$$

where $\gamma = (h, \lambda)$. We estimate the density functions $f(x)$ and $g(x)$ by

$$\hat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} K_{\gamma,x_i,x} \quad \text{and} \quad \hat{g}(x) = \frac{1}{n_2}\sum_{i=1}^{n_2} K_{\gamma,y_i,x}. \tag{2.4}$$

A test statistic can be constructed based on the integrated squared density difference given by $I = \int [f(x) - g(x)]^2 \mathrm{d}x = \int [f(x)\mathrm{d}F(x) + g(x)\mathrm{d}G(x) - f(x)\mathrm{d}G(x) - g(x)\mathrm{d}F(x)]$, where $F(\cdot)$ and $G(\cdot)$ are the cumulative distribution functions for X and Y, respectively, and where $\int \mathrm{d}x = \sum_{x^d \in \mathbb{S}^d} \int \mathrm{d}x^c$. Replacing $f(\cdot)$ and $g(\cdot)$ by their kernel estimates, and replacing $F(\cdot)$ and $G(\cdot)$ by their empirical distribution functions, we obtain the following test statistic,

$$\begin{aligned} I_n^a &= \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{f}(X_i) + \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{g}(Y_i) - \frac{1}{n_1}\sum_{i=1}^{n_1}\hat{g}(X_i) - \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{f}(Y_i) \\ &= \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} K_{\gamma,x_i,x_j} + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} K_{\gamma,y_i,y_j} \\ &\quad - \frac{1}{n_1 n_2}\left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{\gamma,x_i,y_j} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} K_{\gamma,x_j,y_i}\right]. \end{aligned} \tag{2.5}$$

[1] For what follows, when we consider a distribution defined over mixed continuous and discrete variables, we shall use the word 'density' to mean that, for any given value of $x^d \in \mathbb{S}^r$, $f(x^c, x^d)$ is absolutely continuous with respect to x^c.

It can be shown that the test statistic I_n^a has a non-zero center term, say c_n, even under the null hypothesis. Li (1996) proposed a center-free test statistic which is obtained by removing the $i = j$ terms in the double summations appearing in I_n^a. However, this causes a new problem in that the test (with $i = j$ terms removed) depends on the ordering of the data. To see this, note that the $i = j$ terms in the third term of I_n^a is $\sum_{i=1}^{\min\{n_1,n_2\}} K_{\gamma,x_i,y_i}$, which depends on how one orders the data $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$. Below we propose a test statistic which does not have a non-zero center term (under H_0) and is also invariant to the ordering of the data. The test statistic we propose is given by

$$I_n = \frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} K_{\gamma,x_i,x_j} + \frac{1}{n_2(n_2-1)}\sum_{i=1}^{n_2}\sum_{j\neq i}^{n_2} K_{\gamma,y_i,y_j} - \frac{1}{n_1 n_2}\left[\sum_{i=1}^{n_1}\sum_{j=i}^{n_2} K_{\gamma,x_i,y_j} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} K_{\gamma,x_j,y_i}\right]. \tag{2.6}$$

Note that the double summation of the first two terms of I_n removes the $i = j$ terms, while the third term in I_n does not remove the $i = j$ terms. Clearly, the test statistic I_n is invariant to the ordering of the data because the terms removed from I_n^a (i.e., $\sum_{i=1}^{n_1} K_{\gamma,x_i,x_j}$ and $\sum_{i=1}^{n_2} K_{\gamma,y_i,y_j}$) are both invariant with respect to the ordering of the data. We will show in Theorem 2.1 that I_n is an asymptotic zero mean test statistic (under H_0).

The following conditions will be used to derive the asymptotic distribution of I_n.

(C1) The data $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$ are independent and identically distributed (i.i.d.) as X and Y respectively.

(C2) For all $x^d \in \mathbb{S}^r$, $f(\cdot, x^d)$ and $g(\cdot, x^d)$ are bounded (from above by some positive constants) and continuous functions (continuous with respect to x^c). The kernel function $w(\cdot)$ is a bounded, non-negative second order kernel, $\int w(v)v^4 \mathrm{d}v$ is finite, and it satisfies a Lipschitz condition: $|w(u) - w(v)| \leq \xi(v)|u - v|$, where $\xi(\cdot)$ is a bounded smooth function with $\int \xi(v)v^4 \mathrm{d}v < \infty$.

(C3) Letting $\delta_n = n_1/n_2$, then as $n = \min\{n_1, n_2\} \to \infty$, $\delta_n \to \delta \in (0, 1)$, $nh_1 \dots h_q \to \infty$, $h_s \to 0$ for $s = 1, \dots, q$ and $\lambda_s \to 0$ for $s = 1, \dots, r$.

Note that in (C1) we assume that X_i (Y_i) is independent of X_j (Y_j) for $j \neq i$. When $n_1 = n_2 = n$, however, we allow for the possibility that X_i and Y_i are correlated, as would be the case in panel or longitudinal settings where we have repeated measures on individuals. The i.i.d. assumption can be relaxed to weakly dependent (β-mixing) data processes, in which case one needs to apply the central limit theorem (CLT) for degenerate U-statistics with weakly dependent data as given in Fan and Li (1999) in order to derive the asymptotic distribution of the test statistic. Of course, with dependent data, the bootstrap procedure (see Theorem 2.3) will also need to be modified; block or stationary bootstrapping or subsampling methods would be appropriate. In the remaining part of this paper, we will only consider i.i.d. data as stated in (C1).

The other conditions under which Theorem 2.1 holds are quite weak. (C2) only requires that $f(\cdot)$ and $g(\cdot)$ are bounded and continuous, and (C3) is the minimum condition placed upon the smoothing parameters required for consistent estimation of $f(\cdot)$ and $g(\cdot)$. In addition, (C3) requires that the two sample sizes have the same order of magnitude.

The following theorem provides the asymptotic null distribution of the test statistic I_n.

Theorem 2.1. *Assuming that conditions* (C1) *through* (C3) *hold, we have, under* H_0, *that*

$$T_n = (n_1 n_2 \dots h_q)^{1/2} I_n/\sigma_n \to N(0, 1) \text{ in distribution},$$

where

$$\sigma_n^2 = 2(n_1 n_2 h_1 \dots h_q)\left[\frac{1}{n_1^2(n_1-1)^2}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}(K_{\gamma,x_i,x_j})^2 + \frac{1}{n_2^2(n_2-1)^2}\sum_{i=1}^{n_2}\sum_{j\neq i}^{n_2}(K_{\gamma,y_i,y_j})^2 + \frac{1}{n_1^2 n_2^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(K_{\gamma,x_i,y_j})^2 + \frac{1}{n_1^2 n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_1}(K_{\gamma,x_j,y_i})^2\right],$$

which is a consistent estimator of $\sigma_0^2 = 2[\delta^{-1} + \delta + 2][E[f(X_i)]][\int W^2(v)\mathrm{d}v]$, *the asymptotic variance of* $(n_1 n_2 h_1 \dots h_q)^{1/2} I_n$, *where* $\delta = \lim_{\min\{n_1,n_2\}\to\infty}(n_1/n_2)$.

The proof of Theorem 2.1 is given in the Appendix.

It can also be shown that, when H_0 is false, the test statistic T_n will diverge to $+\infty$ at the rate of $(n_1 n_2 h_1 \dots h_q)^{1/2}$. To see this, note that when H_0 is false, one can show that $I_n \to \int [f(x) - g(x)]^2 \mathrm{d}x \equiv C > 0$ (in probability), and that $\sigma_n = O_p(1)$. Hence, T_n will diverge to $+\infty$ at the rate of $(n_1 n_2 h_1 \dots h_q)^{1/2}$, and therefore it is a consistent test.

It is well known that the selection of smoothing parameters is of crucial importance in nonparametric estimation, and it is now known that the selection of smoothing parameters also affects the performance (particularly the power) of nonparametric tests such as the I_n test. Given the reasons outlined in the introduction as to why cross-validation methods seem to be the only feasible way of selecting the smoothing parameters in the presence of mixed discrete and continuous variables, we suggest using cross-validation methods for selecting (h, λ).

The cross-validation method we consider involves selecting smoothing parameters by minimizing a sample analogue of the integrated square error (ISE) of the density estimator. The ISE is defined by $\mathrm{ISE} = \int \left[\hat{f}(x) - f(x)\right]^2 \mathrm{d}x = \int \hat{f}(x)^2 \mathrm{d}x - 2\int \hat{f}(x) f(x) + \int f(x)^2$, where $\int \mathrm{d}x = \sum_{x^d} \int \mathrm{d}x^c$. The third term on the right-hand-side of ISE does not depend on the smoothing parameters. Therefore, in practice one chooses the smoothing parameters to minimize an estimator of $\int \hat{f}(x)^2 \mathrm{d}x - 2\int \hat{f}(x) f(x)$. Let $\{Z_i\}_{i=1}^N$ denote the pooled sample ($N = n_1 + n_2$), i.e., $Z_i = X_i$ for $1 \leq i \leq n_1$ and $Z_{n_1+i} = Y_i$ for $1 \leq i \leq n_2$. Let $\tilde{f}(Z_i) = (N-1)^{-1}\sum_{j\neq i}^N K_{\gamma,z_i,z_j}$ be the leave-one-out estimator of $f(Z_i)$. Then $\int \hat{f}(x)^2 \mathrm{d}x - 2\int \hat{f}(x) f(x)$ can be consistently estimated by the following cross-validation function:

$$CV(h, \lambda) = \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \bar{K}_{\gamma,z_i,z_j} - \frac{2}{N(N-1)}\sum_{i=1}^N\sum_{j\neq i}^N K_{\gamma,z_i,z_j}, \tag{2.7}$$

where $K_{\gamma,z_i,z_j} = W_{h,z_i,z_j} L_{\lambda,z_i,z_j}$, and $\bar{K}_{\gamma,z_i,z_j} = \bar{W}_{h,ij}\bar{L}_{\lambda,ij}$, $\bar{W}_{h,ij} = \int W_{h,z_i,z} W_{h,z_j,z}\mathrm{d}z$ and $\bar{L}_{\lambda,ij} = \sum_{z\in\mathbb{S}^r} L_{\lambda,z,z_i} L_{\lambda,z,z_j}$. It can be shown that $\bar{W}_{h,x_i,x_j} = \prod_{s=1}^q h_s^{-1}\bar{w}((X_{is} - X_{js})/h_s)$, where $\bar{w}(v) = \int w(u)w(v-u)\mathrm{d}u$ is the two-fold convolution kernel derived from $w(\cdot)$, which is also a standard second order kernel function. For example, if $w(v) = \mathrm{e}^{-v^2/2}/\sqrt{2\pi}$, i.e., a standard normal kernel, then $\bar{w}(v) = \mathrm{e}^{-v^2/4}/\sqrt{4\pi}$, a normal kernel with mean zero and variance two, which follows from the fact that two independent $N(0, 1)$ random variables sum to a $N(0, 2)$ random variable. Therefore, we select the smoothing parameters by minimizing the CV function defined in (2.7).

Letting $(\hat{h}_1, \ldots, \hat{h}_q)$ and $(\hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ denote the cross-validated values of $(h_1, \ldots, h_q)$ and $(\lambda_1, \ldots, \lambda_r)$, Li and Racine (2003)[2] and Ouyang et al. (2006) have proved the following, which we summarize in a condition below for ease of reference.

(C4) $\hat{h}_s/h_s^0 - 1 \to 0$ in probability, and $\hat{\lambda}_s/\lambda_s^0 - 1 \to 0$ in probability, where $h_s^0 = a_s^0 n^{-\zeta}$, and $\lambda_s^0 = b_s^0 n^{-2\zeta}$ for some $\zeta > 0$, where a_s^0 and b_s^0 are some finite constants.

h_s^0 and λ_s^0 in (C4) above are the non-stochastic optimal smoothing parameters that minimize the integrated mean squared difference $\int E[(\hat{f}(z) - f(z))^2]\mathrm{d}z$. To establish (C4), Li and Racine (2003) and Ouyang et al. (2006) assumed that (i) $f(x)$ is four times differentiable with respect to x^c; (ii) there exists $x^d, z^d \in \mathcal{S}^r$, such that $f(x^c, x^d) \neq f(x^c, z^d)$ for all x^c in a subset of the support of X^c with positive measure; (iii) $Pr(X_s^d = x_s^d)$ is not constant for all $x_s^d \in \{0, 1, \ldots, c_s - 1\}$. For details on the regularity conditions that ensure (C4) holds, see Li and Racine (2003) and Ouyang et al. (2006).

When $f(x)$ and $g(x)$ have unbounded support and non-vanishing second derivative functions with respect to x_s^c for all $s = 1, \ldots, q$, Li and Racine (2003) and Ouyang et al. (2006) show that $\zeta = 1/(4+q)$, i.e., $\hat{h}_s = O_p(n^{-1/(4+q)})$ and $\hat{\lambda} = O_p(n^{-2/(4+q)})$. When the support of x^c is bounded, the kernel estimator may suffer from the boundary bias problem. However, even in this case, the cross-validated smoothing parameters $\hat{h}_s$ still converge to the optimal smoothing parameter values in the sense that $\hat{h}_s/h_s^0 \to 1$ in probability. In fact, Stone (1984) showed that as long as the marginal density function for X_s is bounded (for all $s = 1, \ldots, q$), then $\hat{h}_s/h_s^0 \to 1$ almost surely. However, the rate at which $\hat{h}_s$ (or h_s^0) converges to zero may be different when the support of x^c is bounded. For example, for the case where $q = 1$ and where x^c is uniformly distributed, Ouyang et al. (2006, their Lemma 3.1) showed that $\zeta = 1/2$ so that $\hat{h} = O_p(n^{-1/2})$. The reason why $\hat{h} \to 0$ (or $\hat{h}_s/h_s^0 \to 1$) in probability even when x^c has bounded support (i.e., is uniformly distributed) is as follows. Consider the case where the support of X is $[0, 1]$. Then, at the boundary regions $x \in [0, h] \cup [1-h, 1]$, $\hat{f}(x) - f(x)$ usually does not converge to zero in probability due to the boundary bias problem. In this case, only when $h \to 0$ will the integrated (mean) squared error converge to zero since the boundary regions shrink to zero length as $h \to 0$. Since $\hat{h}_s$ asymptotically minimizes the integrated squared error, we know that $\hat{h}_s$ must converge to zero whether or not X has bounded support.

Let $\hat{T}_n$ ($\hat{I}_n$) denote the test statistic T_n (I_n) but with (h, λ) being replaced by $(\hat{h}, \hat{\lambda})$, the cross-validated smoothing parameters. The next theorem shows that the test statistic $\hat{T}_n$ has the same asymptotic distribution as T_n.

Theorem 2.2. *Assuming that conditions* (C1) *through* (C3) *hold, then under* H_0 *we have*

$$\hat{T}_n = (n_1 n_2 \hat{h}_1 \ldots \hat{h}_q)^{1/2} \hat{I}_n / \hat{\sigma}_n \to N(0, 1) \text{ in distribution},$$

where $\hat{\sigma}_n$ *is defined the same way as in* σ_n *but with* (h, λ) *replaced by* $(\hat{h}, \hat{\lambda})$.

The proof of Theorem 2.2 is given in the Appendix.

2.2. Comparison with non-smoothing tests

In this section we discuss the local power property of our $\hat{T}_n$ test and compare it with some non-smoothing tests. For ease of exposition we only consider the case where x (y) is a continuous variable of dimension q.[3] One class of non-smoothing tests involves fixing the value of h_s in a smoothing test, say letting $h_s = 1$ for all $s = 1, \ldots, q$; see Anderson et al. (1994), Fan (1998) and Fan and Li (2000) and the references therein. Another class of non-smoothing tests involves testing for the equality of two CDFs. There is a rich literature on testing the equality of two CDFs, i.e., where one tests the null hypothesis that $F(x) = G(x)$ for all x where $F(x)$ and $G(x)$ are two unknown CDFs. Anderson et al. (1994) show that one can set $h_1 = \cdots = h_q = 1$ in the I_n^a test to obtain a non-smoothing test of the form

$$I_{n,h=1} = \left[\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \frac{W(X_i - X_j)}{n_1^2} + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \frac{W(Y_i - Y_j)}{n_2^2} - \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{W(X_i - Y_j)}{n_1 n_2} - \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} \frac{W(X_j - Y_i)}{n_1 n_2} \right], \quad (2.8)$$

where $W(X_i - X_j) = \prod_{s=1}^{q} w(X_{is} - X_{js})$ which is obtained from W_{h,x_i,x_j} by setting $h_1 = \cdots = h_q = 1$. It can be shown that, for a wide class of kernel functions $w(\cdot)$, $I_{n,h=1}$ leads to a consistent test for the null hypothesis of equality of $f(x)$ and $g(x)$ for almost all $x \in \mathbb{R}^q$. It is well known that a non-smoothing test such as $I_{n,h=1}$ does not have an asymptotic normal distribution. It can be shown that $\sqrt{n_1 n_2} I_{n,h=1}$ has an asymptotic weighted χ^2 distribution of the form $\sum_{l=1}^{\infty} c_l \chi_l(1)$, where the c_l's are some constants, and the $\chi_l(1)$'s are independent chi-square random variables with one degree of freedom. The weight c_l depends on the unknown density functions $f(x)$ and $g(x)$. Therefore, it is impossible to tabulate this asymptotic distribution. However, bootstrap methods may be used to approximate the null distribution of $I_{n,h=1}$.

One can also test the null hypothesis of equality of two distributions based upon estimation of the unknown CDFs. For example, a Kolmogorov-Smirnov type test can be constructed based on $\sup_{x \in \mathbb{R}^q} |F(x) - G(x)|$. Let $F_{n_1}(\cdot)$ and $G_{n_2}(\cdot)$ be the empirical CDFs of $\{X_i\}_{i=1}^{n_1}$ and $\{Y_i\}_{i=1}^{n_2}$, respectively. Formally, we have

$$KS_n = \sup_{x \in \mathbb{R}^q} \left| \sqrt{\frac{2n_1 n_2}{n_1 + n_2}} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} I(X_i \leq x) - \frac{1}{n_2} \sum_{i=1}^{n_2} I(Y_i \leq x) \right] \right|.$$

It is easy to check that $\sqrt{2n_1 n_2/(n_1 + n_2)}[F_{n_1}(\cdot) - G_{n_2}(\cdot)]$ converges to a zero mean Gaussian process (under H_0), say $GP(\cdot)$. Then it follows from the continuous mapping theorem that $KS_n \to \sup_{x \in \mathbb{R}^q} |GP(x)|$ in distribution under H_0.

A Cramer–von Mises (CM) type statistic (based on $\int [F(x) - G(x)]^2 \mathrm{d}x$) can be constructed by

$$CM_n = \frac{2n_1 n_2}{n_1 + n_2} \int \left[\frac{1}{n_1} \sum_{i=1}^{n_1} I(X_i \leq x) - \frac{1}{n_2} \sum_{i=1}^{n_2} I(Y_i \leq x) \right] \times \left[\frac{1}{n_1} \sum_{j=1}^{n_1} I(X_j \leq x) - \frac{1}{n_2} \sum_{j=1}^{n_2} I(Y_j \leq x) \right] \mathrm{d}x.$$

It can be shown that $CM_n \to \int GP(x)^2 \mathrm{d}x$ in distribution under H_0.

[2] Li and Racine (2003) only consider the case for which $h_1 = \cdots = h_q = h$ and $\lambda_1 = \cdots = \lambda_r = \lambda$. It is straightforward to generalize the result of Li and Racine (2003) to the vector h and λ case, and the result should be modified as given here.

[3] Adding discrete components to x (y) will require more complex notation, but will not affect the result of the local power analysis.

We will use some bootstrap methods to approximate the null distributions of $I_{n,h=1}$, KS_n and CM_n. A simple bootstrap method involves resampling from the pooled sample $\{Z_i\}_{i=1}^{n_1+n_2}$ with replacement, where $Z_i = X_i$ for $i = 1, \ldots, n_1$, and $Z_{n_1+i} = Y_i$ for $i = 1, \ldots, n_2$. One then uses the bootstrap sample $\{X_i^*\}_{i=1}^{n_1} = \{Z_i^*\}_{i=1}^{n_1}$, and $\{Y_i^*\}_{i=1}^{n_2} = \{Z_{n_1+i}^*\}_{i=1}^{n_2}$ to compute $I_{n,h=1}^*$, KS_n^* and CM_n^*, respectively.

Note that since both the KS_n and the CM_n tests involve indicator functions, therefore, the sup operator in KS_n can be replaced by maximization over the $n_1 + n_2$ sample realizations as follows:

$$KS_n = \max_{1\le j\le n_1+n_2} \left| \sqrt{\frac{2n_1n_2}{n_1+n_2}} \left[\frac{1}{n_2}\sum_{i=1}^{n_1} I(X_i \le W_j) - \frac{1}{n_2}\sum_{i=1}^{n_2} I(Y_i \le W_j) \right] \right|. \tag{2.9}$$

Similarly, all integration required for the computation of CM_n can be computed easily leading to the following result:

$$CM_n = \frac{2n_1n_2}{n_1+n_2} \left\{ \frac{1}{n_1n_2} \left[\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \max\{X_i, Y_j\} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_1} \max\{Y_i, X_j\} \right] - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \max\{X_i, X_j\} - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} \max\{Y_i, Y_j\} \right\}. \tag{2.10}$$

We now compare the local power properties of smoothing and non-smoothing tests. We consider two types of local alternatives. One is a sequence of 'regular' or 'Pitman' alternatives given by

$$LH_r : f(x) = g(x) + \alpha_n \Delta(x),$$

where $\int \Delta(x)dx = 0$ and $\alpha_n \to 0$ as $n \to \infty$. The second is a sequence of so-called 'singular' local alternatives which was first introduced by Rosenblatt (1975) and is given by

$$LH_s : f(x) = g(x) + \alpha_n \Delta_n(x),$$

where $\int \Delta_n(x)dx = 0$, $\int \Delta_n^2(x)dx \to 0$, and $\alpha_n \to 0$ as $n \to \infty$. For example, one can have $\Delta_n(x) = \sum_{j=1}^{p} d_j((x - l_j)/\beta_n)$, where p is a positive integer, $l_1, \ldots, l_p$ are constant vectors in $\mathbb{R}^q$, $d_1(\cdot), \ldots, d_p(\cdot)$ are bounded smooth functions satisfying $\int d_j(x)dx = 0$ for $j = 1, \ldots, p$, and $\beta_n \to 0$ as $n \to \infty$. Then it is easy to see that $\int \Delta_n(x)^2 dx = O(\beta_n) = o(1)$.

In finite-sample applications, the 'singular' alternative corresponds to a 'rapidly changing' or a 'high frequency' density function. In the simulations reported in Section 4, we use some mixture normal distributions (densities with multiple peaks) to represent 'high frequency' density functions.

It is well established that non-smoothing tests can detect both the Pitman and the Rosenblatt local alternatives that approach the null at the rate of $n^{-1/2}$. In contrast, smoothing tests can detect Pitman local alternatives converging to the null at rate $n^2(h_1 \ldots h_q)^{-1/4}$, which is slower than $n^{-1/2}$ because $h_s \to 0$ for all $s = 1, \ldots, q$. Therefore, for Pitman local alternatives, a non-smoothing test is asymptotically more powerful than a smoothing test. However, it is also known that, for the class of 'singular' local alternatives, a smoothing test can detect local alternatives that approach the null at a rate of $o(n^{-1/2})$; see Ghosh and Huang (1991), Fan (1998) and Fan and Li (2000). Hence, a smoothing test is more powerful than a non-smoothing test for 'singular' local alternatives. Indeed the simulation evidence reported in Fan and Li (2000) reveals strong support for the above theoretical local power analysis.

The existing simulation comparisons between $I_{n,h=1}$ and I_n typically use some ad-hoc selection of h such as $h_{s,ad\text{-}hoc} = z_{s,sd}(n_1 + n_2)^{-1/(4+q)}$ when computing I_n, where $z_{s,sd}$ is the sample standard deviation of $\{Z_{is}\}_{i=1}^{n_1+n_2}$ $(s = 1, \ldots, q)$. In Section 4 we show that cross-validated (CV) selection of h_s results in a test that is often more powerful (in finite-sample applications) than either using $h_{s,ad\text{-}hoc}$ or using $h = 1$. The superior performance of the CV-based test arises because the CV method can automatically adapt to the smoothness of the underlying density functions. When $f(x)$ $(g(x))$ is a relatively smooth (i.e., unimodal and slowly changing) function of x_s, the CV method will select a relatively large value for h_s; when $f(x)$ $(g(x))$ is a relatively high frequency function of x_s (i.e., multimodal and peaked), the CV method will select a small value for h_s resulting in a test having high power against either low or high frequency alternatives. The simple ad-hoc rule of selecting $h_{s,ad\text{-}hoc}$ or even fixing $h = 1$ cannot possess such flexibility which can harm their power as will be seen.

The CDF-based tests have local power properties similar to non-smoothing tests $I_{n,h=1}$. Therefore, they are asymptotically more powerful than a smoothing test against Pitman local alternatives, and they may be less powerful against 'singular' local alternatives. In Section 4 we report simulation results that examine the finite-sample performance of our smoothing test versus some non-smoothing tests including the Kolmogorov-Smirnov test and the Cramer–von Mises test discussed above.

2.3. A bootstrap procedure

Theorems 2.1 and 2.2 show that T_n and $\hat{T}_n$ have asymptotic standard normal null distributions. However, existing simulation results suggest that this limiting normal distribution is in fact a poor approximation to the finite-sample distribution of T_n. Our experience also shows that the same holds true for the $\hat{T}_n$ statistic. Therefore, in order to better approximate the null distribution of $\hat{T}_n$, in applied settings we advocate the use of the following bootstrap procedure.

Let $Z_i = X_i$ for $i = 1, \ldots, n_1$ and $Z_{n_1+i} = Y_i$ for $i = 1, \ldots, n_2$. Randomly draw n_1 observations from the pooled sample $\{Z_j\}_{j=1}^{n_1+n_2}$ *with* replacement, and call the resulting sample $\{X_i^*\}_{i=1}^{n_1}$; then randomly draw another n_2 observations from $\{Z_j\}_{j=1}^{n_1+n_2}$ with replacement, and call the resulting sample $\{Y_i^*\}_{i=1}^{n_2}$. Compute the bootstrap test statistic given by $\hat{T}_n^* = (n_1n_2\hat{h}_1 \ldots \hat{h}_q)^{1/2}\hat{I}_n^*/\hat{\sigma}_n^*$, where $\hat{I}_n^*$ and $\hat{\sigma}_n^*$ are defined the same way as $\hat{I}_n$ and $\hat{\sigma}_n$ except that X_i and Y_i are replaced by X_i^* and Y_i^*, respectively. We repeat this procedure a large number of times, say $B = 399$ times (Davidson and MacKinnon, 2000), and we use the empirical distribution of the B bootstrap statistics $\{\hat{T}_{n,l}^*\}_{l=1}^{B}$ to approximate the null distribution of $\hat{T}_n$. Empirical P-values can be computed via $\hat{P} = B^{-1}\sum_{l=1}^{B} I(\hat{T}_{n,l}^* > \hat{T}_n)$, where $I(\cdot)$ is an indicator function, which is simply the proportion of resampled test statistics under the null that are more extreme than the statistic itself.

Note that we use the same smoothing parameters $(\hat{h}, \hat{\lambda})$ when computing $\hat{T}_n^*$, i.e., we do not re cross-validate for each bootstrap replication. Therefore, this bootstrap procedure is computationally less costly than the computation of $\hat{T}_n$, which involves a cross-validation procedure. The next theorem proves the validity of the proposed bootstrap method.

Theorem 2.3. *Define $\hat{T}_n^* = (n_1n_2\hat{h}_1 \ldots \hat{h}_q)^{1/2}\hat{I}_n^*/\hat{\sigma}_n^*$. Assuming that the same conditions given in Theorem 2.2 hold, but without imposing the null hypothesis, then we have*

$$\sup_{z\in\mathbb{R}} \left| P\left(\hat{T}_n^* \le z | \{X_i, Y_i\}_{i=1}^{n}\right) - \Phi(z) \right| = o_p(1), \tag{2.11}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

A sketch of the proof of Theorem 2.3 is given in the Appendix. To summarize, Theorem 2.3 states that $\hat{T}_n^*$ converges to $N(0, 1)$ in distribution in probability.

A former wording definition of *convergence in distribution in probability* can be given as follows: Let ξ_n denote a statistic that depends on the random sample $\{W_i\}_{i=1}^n$, we say that $(\xi_n|W_1, W_2, \ldots)$ converges to $(\xi|W_1, W_2, \ldots)$ *in distribution in probability* if for any subsequence $\xi_{n'}$, there exists a further subsequence $\xi_{n''}$ such that $(\xi_{n''}|W_1, W_2, \ldots)$ converges to $(\xi|W_1, W_2, \ldots)$ for almost every sequence $(W_1, W_2, \ldots)$.

3. A nonparametric test for the equality of conditional density functions with mixed categorical and continuous data

In this section we consider the problem of testing the equality of two conditional density functions. We will only consider the case for which the conditioning variable is categorical in nature. There are two reasons for this. First, technically it is difficult to handle the continuous conditioning variable case when the density function is not bounded below by a positive constant. The second consideration is a practical one. In empirical applications it is often the case that one is interested in knowing the distribution of a continuous variable, say the distribution of income conditional on a discrete variable such as a person's gender, or perhaps their level of education.

Given that we only consider a discrete conditioning variable in this section, we shall employ slightly different notation for what follows. We shall continue to use $x = (x^c, x^d) \in \mathbb{R}^q \times \mathbb{S}^r$ to denote a mixture of continuous and discrete variables, and we use w to denote the conditioning discrete variable. w can be a multivariate discrete variable. We use $\mathbb{S}_w$ to denote the support of W, and we assume that $P(w) = \Pr(W = w)$ is bounded below by a positive constant for all $w \in \mathbb{S}_w$. Suppose we have i.i.d data, $\{X_i, U_i\}_{i=1}^{n_1}$, which are random draws from the joint density function $f(x, w)$ along with i.i.d. draws of $\{Y_i, V_i\}_{i=1}^{n_2}$ from the joint density function $g(x, w)$. We use $f(x|w)$ $(g(x|w))$ to denote the conditional density function of X (Y) conditional on $U = w$ $(V = w)$. We use $\mathcal{S}_w$ to denote a subset of the support of w such that one is interested testing for $f(x|w) = g(x|w)$ for all $w \in \mathcal{S}_w$. Formally, we want to test the following null hypothesis.

$$H_0^c : f(x|w) = g(x|w) \quad \text{for all } w \in \mathcal{S}_w, x^d \in \mathbb{S}^r \text{ and for almost all } x^c \in \mathbb{R}^q, \tag{3.12}$$

against the alternative hypothesis, H_1^c, that $f(x|w) \neq g(x|w)$ on a set with positive measure.

Define $p_f(w) = Pr(U = w)$ and $p_g(w) = Pr(V = w)$. Note that $p_f(w)$ can differ from $p_g(w)$. For example, consider the case where x is income and w is a dummy variable equal to one for males, zero otherwise. Clearly the percentage of males in two populations can differ, i.e., $p_f(w)$ may not equal $p_g(w)$.

Using $f(x|w) = f(x, w)/p_f(w)$ and $g(x|w) = g(x, w)/p_g(w)$, we will construct a test statistic based on

$$\begin{aligned} J &= \sum_{w \in \mathcal{S}_w} \int [f(x|w) - g(x|w)]^2 \mathrm{d}x \\ &= \sum_{w \in \mathcal{S}_w} \int \left[\frac{f(x, w)^2}{p_f(w)^2} + \frac{g(x, w)^2}{p_g(w)^2} - \frac{2f(x, w)g(x, w)}{p_f(w)p_g(w)} \right] \mathrm{d}x, \end{aligned} \tag{3.13}$$

where $\int \mathrm{d}x = \sum_{x^d \in \mathbb{S}^r} \int \mathrm{d}x^c$.

Let $I_{u_i,w} = I(U_i = w)$ denote an indicator function which equals one if $U_i = w$ and zero otherwise. $I_{v_i,w}$ is similarly defined. We estimate the joint density of $f(x, w)$ and $g(x, w)$ by

$$\begin{aligned} \hat{f}(x, w) &= \frac{1}{n_1} \sum_{i=1}^{n_1} K_{\gamma,x_i,x} I_{u_i,w}, \quad \text{and} \\ \hat{g}(x, w) &= \frac{1}{n_2} \sum_{i=1}^{n_2} K_{\gamma,y_i,x} I_{v_i,w}. \end{aligned} \tag{3.14}$$

Also, we estimate $p_f(w)$ and $p_g(w)$ by

$$\hat{p}_f(w) = \frac{1}{n_1} \sum_{i=1}^{n_1} I(U_i = w) \quad \text{and} \quad \hat{p}_g(w) = \frac{1}{n_2} \sum_{i=1}^{n_2} I(V_i = w). \tag{3.15}$$

Define the leave-one-out empirical functions by $F_{n,-i}(x) = (n_1 - 1)^{-1} \sum_{j \neq i}^{n_1} I(X_j \leq x)$ and $G_{n,-i}(x) = (n_2 - 1)^{-1} \sum_{j \neq i}^{n_2} I(Y_j \leq x)$. Replacing f, g, p_f and p_g by their estimators in (3.13), and using the short-hand notation $\hat{p}_f = \hat{p}_f(w)$ and $\hat{p}_g = \hat{p}_g(w)$, we obtain a feasible test statistic given by

$$\begin{aligned} J_n &= \sum_{w \in \mathcal{S}_w} \int \left[\frac{\hat{f}(x, w)}{\hat{p}_f^2} \mathrm{d}F_{n,-i}(x) + \frac{\hat{g}(x, w)}{\hat{p}_g^2} \mathrm{d}G_{n,-i}(x) \right. \\ &\quad \left. - \frac{\hat{f}(x, w)}{\hat{p}_f \hat{p}_g} \mathrm{d}G_n(x) - \frac{\hat{g}(x, w)}{\hat{p}_f \hat{p}_g} \mathrm{d}F_n(x) \right] \\ &= \sum_{w \in \mathcal{S}_w} \left\{ \sum_i \sum_{j \neq i} \left[\frac{\bar{K}_{\gamma,x_i,x_j} I_{u_i,w} I_{u_j,z}}{n_1(n_1 - 1)\hat{p}_f^2} + \frac{\bar{K}_{\gamma,y_i,y_j} I_{v_i,w} I_{v_j,w}}{n_2(n_2 - 1)\hat{p}_g^2} \right] \right. \\ &\quad \left. - \frac{1}{n_1 n_2 \hat{p}_f \hat{p}_g} \sum_i \sum_j \left[\bar{K}_{\gamma,x_i,y_j} I_{u_i,w} I_{v_j,w} + \bar{K}_{\gamma,x_j,y_i} I_{u_j,w} I_{v_i,w} \right] \right\}, \end{aligned} \tag{3.16}$$

where $\bar{K}_{h,x_i,y_j} = \bar{W}_{h,x_i,x} \bar{L}_{\lambda,x_i,x_j}$, and $\bar{W}_{h,x_i,x}$ and $\bar{L}_{\lambda,x_i,x_j}$ are defined in Section 2. Also, as in Section 2, $\sum_i$ is $\sum_{i=1}^{n_1}$ if the summand has (X_i, U_i) as its argument, and $\sum_i$ is $\sum_{i=1}^{n_2}$ if the summand has (Y_i, V_i) as its argument. We choose $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_q)$ by the cross-validation method discussed in Section 2, and we use $(\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_q)$ to denote the cross-validated smoothing parameters. We will use $\hat{J}_n$ to denote our test statistic as defined in (3.16) but with $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_q)$ replaced by $(\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_q)$.

We make the following additional assumption.

(C5) For all (x^d, w) $(x^d \in \mathbb{S}^r)$, both $f(\cdot, x^d, w)$ and $g(\cdot, x^d, w)$ are bounded (from above by some positive constants) and continuous functions (continuous with respect to x^c).

The asymptotic null distribution of our test statistic is given in the next theorem.

Theorem 3.1. *Assuming that conditions* (C1)–(C5) *hold, then under* H_0^c, *we have*

$$\hat{T}_{n,c} \overset{def}{=} (n_1 n_2 \hat{h}_1 \ldots \hat{h}_q)^{1/2} \hat{J}_n / \hat{\sigma}_{n,c} \to N(0, 1) \text{ in distribution},$$

where

$$\begin{aligned} \hat{\sigma}_{n,c}^2 &= 2(n_1 n_2 \hat{h}_1 \ldots \hat{h}_q) \sum_{w \in \mathcal{S}_w} \left[\sum_{i=1}^{n_1} \sum_{j \neq i}^{n_1} \frac{(\bar{K}_{\hat{\gamma},x_i,x_j} I_{u_i,w} I_{u_j,w})^2}{n_1^4 \hat{p}_f(w)^4} \right. \\ &\quad + \sum_{i=1}^{n_2} \sum_{j \neq i}^{n_2} \frac{(\bar{K}_{\hat{\gamma},y_i,y_j} I_{v_i,w} I_{v_j,w})^2}{n_2^4 \hat{p}_g(w)^4} + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(\bar{K}_{\hat{\gamma},x_i,y_j} I_{u_i,w} I_{v_j,w})^2}{n_1^2 n_2^2 \hat{p}_f(w)^2 \hat{p}_g(w)^2} \\ &\quad \left. + \sum_{i=1}^{n_2} \sum_{j=1}^{n_1} \frac{(\bar{K}_{\hat{\gamma},x_j,y_i} I_{v_i,w} I_{u_j,w})^2}{n_1^2 n_2^2 \hat{p}_g(w)^2 \hat{f}_f(w)^2} \right]. \end{aligned}$$

The proof of Theorem 3.1 is given in the Appendix.

In practice we recommend the use of the following bootstrap procedure to approximate the null distribution of $\hat{T}_{n,c}$.

Let $Z_i = \{X_i, U_i\}$ for $i = 1, \ldots, n_1$, and $Z_{n_1+i} = \{Y_i, V_i\}$ for $i = 1, \ldots, n_2$. Then randomly draw n_1 observations from the pooled sample $\{Z_j\}_{j=1}^{n_1+n_2}$ *with* replacement, and call the resulting sample $\{X_i^*, U_i^*\}_{i=1}^{n_1}$, and then randomly draw another

n_2 observations from $\{Z_j\}_{j=1}^{n_1+n_2}$ with replacement, and call the resulting sample $\{Y_i^*, V_i^*\}_{i=1}^{n_2}$. Compute a test statistic $\hat{T}_{n,c}^* = (n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{J}_n^* / \hat{\sigma}_{n,c}^*$, where $\hat{J}_n^*$ and $\hat{\sigma}_{n,c}^*$ are defined the same way as $\hat{J}_n$ and $\hat{\sigma}_{n,c}$ except that (X_i, U_i) and (Y_i, V_i) are replaced by (X_i^*, U_i^*) and (Y_i^*, V_i^*), respectively. We repeat this procedure a large number of times (say $B = 399$), and we use the empirical distribution of the B bootstrap statistics $\{\hat{T}_{n,c,l}^*\}_{l=1}^B$ to approximate the null distribution of $\hat{T}_{n,c}$.

The next theorem states that the above bootstrap method can be used to approximate the null distribution of $\hat{T}_{n,c}$.

Theorem 3.2. *Define $\hat{T}_{n,c}^* = (n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{J}_n^* / \hat{\sigma}_{n,c}^*$. Assume the same conditions as in Theorem 3.1 except that we do not impose the null hypothesis H_0^c. Then we have*

$$\sup_{z\in\mathbb{R}} \left| P\left(\hat{T}_{n,c}^* \le z | \{X_i, U_i, Y_i, V_i\}_{i=1}^n\right) - \Phi(z) \right| = o_p(1), \tag{3.17}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

The proof of Theorem 3.2 is similar to the proof of Theorem 3.1 and is thus omitted.

Note that in constructing our conditional density test statistic $T_{n,c}$, we smooth both x^c and x^d, and we do not smooth over the conditional discrete covariate w. In practice one can also smooth the conditional discrete variable w when testing H_0^c. For expositional simplicity, we discuss the case where w is a scalar below. In this case, one replaces the indicator function, say, $I_{u_i,w} = I(U_i = w)$ by $l_{\lambda_0,u_i,w} = l(U_i, w, \lambda_0)$, which is defined in (2.1), and λ_0 is the smoothing parameter associated with w. The modified test statistic becomes

$$\begin{aligned} J_{n,\lambda_0} = \sum_{w\in\mathcal{S}_w} \Bigg\{ \sum_i \sum_{j\neq i} \Bigg[& \frac{\bar{K}_{\gamma,x_i,x_j} l_{\lambda_0,u_i,w} l_{\lambda_0,u_j,z}}{n_1(n_1-1)\tilde{p}_f^2} \\ & + \frac{\bar{K}_{\gamma,y_i,y_j} l_{\lambda_0,v_i,w} l_{\lambda_0,v_j,w}}{n_2(n_2-1)\tilde{p}_g^2} \Bigg] \\ & - \frac{1}{n_1 n_2 \tilde{p}_f \tilde{p}_g} \sum_i \sum_j \Big[\bar{K}_{\gamma,x_i,y_j} l_{\lambda_0,u_i,w} l_{\lambda_0,v_j,w} \\ & + \bar{K}_{\gamma,x_j,y_i} l_{\lambda_0,u_j,w} l_{\lambda_0,v_i,w} \Big] \Bigg\}, \end{aligned} \tag{3.18}$$

where $\tilde{p}_f = n_1^{-1} \sum_{i=1}^n l_{\lambda_0,u_i,w}$ and $\tilde{p}_g = n_2^{-1} \sum_{i=1}^n l_{\lambda_0,v_i,w}$. For the test statistic J_{n,λ_0} we also need a different method for selecting the smoothing parameters. In the framework of estimating a conditional density function, Hall et al. (2004) propose selecting the smoothing parameters by minimizing the sample analogue of $\sum_{w\in\mathcal{S}_w} \sum_{x^d\in\mathbb{S}^r} \int [\hat{f}(x|w) - f(x|w)]^2 \mu(x) \mathrm{d}x^c$, where $\mu(x)$ is a weight function. We suggest using the least squares cross-validation method proposed by Hall et al. (2004) for selecting the smoothing parameters $(\lambda_0, \lambda_1, \dots, \lambda_r)$ and $(h_1, \dots, h_q)$. In Section 4 we also compute the test statistic J_{n,λ_0} and compare it with the test statistic J_n. The bootstrap procedure for obtaining critical values for the J_{n,λ_0} test is similar to that for J_n, except that one replaces the indicator functions by the (discrete variable) corresponding kernel functions. The simulations reported there show that J_{n,λ_0} test has better power performance than that of J_n.

The asymptotic analysis of J_{n,λ_0} is much more involved than that of J_n. This is because the J_n statistic uses $\hat{p}_f(w)$ and $\hat{p}_g(w)$ to estimate $p_f(w)$ and $p_g(w)$; and that $\hat{p}_f(w) - p_f(w) = O_p(n_1^{-1/2})$ and $\hat{p}_g(w) - p_g(w) = O_p(n_2^{-1/2})$; they both have the parametric root-n convergence rate. In Appendix we show that the asymptotic distribution of $T_{n,c}$ is unaffected if one replaces $\hat{p}_f(w)$ and $\hat{p}_g(w)$ by $p_f(w)$ and $p_g(w)$. In contrast, the J_{n,λ_0} statistic uses kernel-smoothed probability estimators $\tilde{p}_f(w)$ and $\tilde{p}_g(w)$ to estimate $p_f(w)$ and $p_g(w)$; and $\tilde{p}_f(w) - p_f(w) = O_p\left(\sum_{s=1}^q h_s^2 + \sum_{s=0}^r \lambda_s + (nh_1 \dots h_q)^{-1/2}\right) = \tilde{p}_g(w) - p_g(w)$. They both have a (slow) nonparametric rate of convergence. Estimating $p_f(w)$ and $p_g(w)$ by the nonparametric estimators $\tilde{p}_f(w)$ and $\tilde{p}_g(w)$ may affect the asymptotic distribution of J_{n,λ_0}, rendering the asymptotic analysis much more complex than that of J_n. We leave the asymptotic analysis of J_{n,λ_0} as a future research topic. We use simulations to examine the finite-sample performance of J_{n,λ_0} based on bootstrap critical values.

4. Monte Carlo simulations

In this section we consider the finite-sample performance of the proposed tests in a variety of settings. We begin by comparing the performance of the proposed unconditional density test (T_n) with its frequency-based and ad-hoc smoothing parameter selection counterparts when there exist both continuous and discrete variables. Next, we compare the proposed unconditional density test with non-smooth tests under both 'high frequency' and 'low frequency' alternatives. The local power analysis of Section 2.2 suggests that a non-smoothing test is likely to be more powerful against low frequency (i.e., slowly changing) density functions, while a smoothing test is expected to be more powerful for high frequency (i.e., rapidly changing) density functions. Finally, we examine the finite-sample performance of the proposed conditional density (J_n and J_{n,λ_0}) tests.

4.1. Testing equality of unconditional density functions with mixed data

We consider a range of mixed data DGPs for $f(x)$ and $g(x)$ designed to examine empirical size and power of the proposed test. We allow X^c and X^d to be correlated, and vary the degree of correlation. We let $X^d \in \{0, 1, \dots, 3\}$ with probabilities $(0.125, 0.375, 0.375, 0.125)$. We let Y be a mixture of normals drawn from $N(-2, \sigma^2)$ and $N(2, \sigma^2)$ with equal probability. We first draw X^d and then let $X^c = \alpha X^d + Y$. When $\alpha = 0$, X^d and X^c are independent, while when $\alpha = 3/4$, $\rho_{x^c,x^d} = 1/4$. As the bootstrap test is correctly sized, we report size only once, and report power for a range of alternative DGPS. The null DGP is DGP0 for independent x^c and x^d where $f(\cdot) = g(\cdot)$. For DGP1, we let the continuous components of $f(\cdot)$ and $g(\cdot)$ differ in their means under the alternative with the difference in means equal to $1/2$, with $\rho_{x^c,x^d} = 0$. For DGP2, we again let the continuous components of $f(\cdot)$ and $g(\cdot)$ differ in their means under the alternative with the difference in means equal to $1/2$, with $\rho_{x^c,x^d} = 1/4$. For DGP3, the marginals of x^c and x^d are identical under the null and alternative, but the degree of correlation differs under the alternative ($\rho_{x^c,x^d} = 1/2$ under the null, $\rho_{x^c,x^d} = 1/4$ under the alternative). Finally, for DGP4 the standard deviation of x^c differs by $1/2$ under the null and alternative, while $\rho_{x^c,x^d} = 1/4$.[4]

We consider three tests of the hypothesis $H_0 : g(x) = f(x)$: (i) the proposed test with least-squares cross-validated h and λ (T_n), (ii) the conventional frequency test with cross-validated h *and* $\lambda = 0$ ($T_{n,\lambda=0}$), and (iii) the conventional ad hoc test with $h = 1.06\sigma n^{-1/5}$ and $\lambda = 0$ ($T_{n,h=1.06\sigma n^{-1/5},\lambda=0}$). Least-squares cross-validated bandwidth selection is used to obtain h and λ for each of the $M = 1000$ Monte Carlo replications, except where noted. The

[4] We are grateful to an anonymous referee for suggesting this rich range of DGPs.

Table 1
Unconditional density T_n test mixed data Monte Carlo.

n	T_n			$T_{n,\lambda=0}$			$T_{n,h=1.06\sigma n^{-1/5},\lambda=0}$		
	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.01$	$\alpha=0.05$	$\alpha=0.10$
Size (DGP0)									
50	0.013	0.058	0.106	0.011	0.050	0.105	0.010	0.034	0.073
100	0.006	0.059	0.102	0.006	0.053	0.113	0.008	0.043	0.089
200	0.009	0.052	0.106	0.007	0.045	0.104	0.005	0.036	0.083
400	0.013	0.052	0.103	0.010	0.055	0.104	0.008	0.042	0.097
Power (DGP1), mean of x differs under null and alternative, $\rho_{x,z}=0$									
50	0.045	0.146	0.257	0.026	0.102	0.187	0.016	0.067	0.148
100	0.093	0.255	0.389	0.062	0.195	0.316	0.042	0.146	0.265
200	0.221	0.499	0.646	0.174	0.424	0.592	0.143	0.374	0.531
400	0.633	0.867	0.936	0.564	0.830	0.920	0.534	0.850	0.925
Power (DGP2), mean of x differs under null and alternative, $\rho_{x,z}=1/4$									
50	0.036	0.135	0.218	0.029	0.120	0.213	0.017	0.085	0.162
100	0.066	0.216	0.343	0.064	0.208	0.340	0.039	0.154	0.280
200	0.176	0.437	0.605	0.158	0.435	0.604	0.121	0.389	0.551
400	0.532	0.830	0.913	0.508	0.828	0.909	0.485	0.805	0.906
Power (DGP3), marginals identical under null and alternative, correlation differs under alternative									
50	0.059	0.192	0.308	0.042	0.165	0.276	0.023	0.118	0.215
100	0.111	0.325	0.492	0.105	0.301	0.466	0.058	0.247	0.386
200	0.335	0.647	0.767	0.316	0.629	0.768	0.250	0.575	0.731
400	0.811	0.964	0.989	0.815	0.968	0.986	0.776	0.967	0.985
Power (DGP4), standard deviations of x differ under the null and alternative, $\rho_{x,z}=1/4$									
50	0.028	0.122	0.221	0.030	0.116	0.203	0.014	0.067	0.131
100	0.068	0.217	0.343	0.067	0.219	0.328	0.036	0.140	0.246
200	0.164	0.413	0.569	0.165	0.409	0.567	0.089	0.304	0.490
400	0.461	0.788	0.895	0.485	0.785	0.892	0.388	0.732	0.867

second order Gaussian kernel is used throughout. For each Monte Carlo replication we conduct $B = 399$ bootstrap replications, and then compute empirical P-values for each statistic. We then summarize the empirical rejection frequencies for each test at the 1%, 5%, and 10% levels. We vary the sample size from $n = 50$ through $n = 400$. Empirical size and power over the $M = 1000$ Monte Carlo replications is summarized in Table 1.

Table 1 suggests the following; (i) our test is correctly sized, while the other test sizes are reasonable as well, (ii) the proposed method often exhibits substantial power gains, especially in small sample situations relative to the conventional frequency test ($\lambda = 0$) and relative to the ad-hoc test in particular ($h = 1.06\sigma n^{-1/5}$, $\lambda = 0$), and (iii) the consistency of the tests is evident in the large sample experiments with power approaching one. As the ad-hoc test appears to be slightly undersized, we also computed size-adjusted power and the ranking of estimators in terms of power remains unchanged (the results of size-adjusted power are not reported here to save space).

4.2. Testing equality of unconditional density functions under 'low frequency' alternatives

First, we consider a Monte Carlo simulation designed to demonstrate how, under a 'low frequency' alternative, non-smoothing tests such as the KS_n and CM_n tests can perform better than smoothing tests such as the T_n test proposed in this paper. For what follows, we consider the case where the marginal density for the continuous variable (x^c) is a simple univariate normal density function. Here we treat a unimodal normal distribution as a low frequency (i.e., slowly changing) density function.

Specifically, for this experiment the null DGP is $N(0, 1)$ while the alternative DGP is $N(1/2, 1)$. Under the null, both X and Y are drawn from the $N(0, 1)$, while under the alternative X is drawn the $N(0, 1)$ while Y is drawn from the $N(1/2, 1)$. Least-squares cross-validated bandwidth selection is used for the T_n test, and is computed for each of the $M = 1000$ Monte Carlo replications. For each Monte Carlo replication we conduct $B = 399$ bootstrap replications, and then compute empirical P-values for each statistic. We then summarize the empirical rejection frequencies for each test at the 1%, 5%, and 10% levels. We vary the sample size from $n = 50$ through $n = 400$ at which point power is equal to one for all three tests considered. Results are reported in Table 2.

Table 2 reveals that each test is correctly sized while power is highest for the non-smoothed tests as expected under 'low frequency' alternatives with the CM_n test being most powerful for this DGP. We also computed the non-smoothing test $I_{n,h=1}$ in our simulations, and since results show that the $I_{n,h=1}$ test is correctly sized and has power similar to that of the KS_n test, detailed results are not reported here due to space limitation.

4.3. Testing equality of unconditional density functions under 'high frequency' alternatives

Next, we consider a Monte Carlo simulation designed to demonstrate how, under 'high frequency' alternatives, smoothing tests such as the T_n test proposed in this paper can perform better than non-smoothing tests such as the KS_n and CM_n tests, a fact that may not be appreciated by all readers.

We shall draw data from a mixture of normal distributions each having different locations and scales. Under the null, $f(x)$ (and $g(x)$) is a mixture of two normal distributions: $N(-1/2, 1)$ and $N(1/2, 4)$, with equal probability. That is, we draw data for X and Y from a $N(-1/2, 1)$ and $N(1/2, 4)$ with equal probability. It can be seen that the PDF of data drawn from this mixture has a bimodal and asymmetric distribution, the left peak being higher than the right. Under the alternative, however, $f(x)$ remains the same as above, while $g(x)$ is again a mixture but of two different normal distributions: $N(-1/2, 4)$ and $N(1/2, 1)$ with equal probability. That is, we reverse the peaks and draw data for Y from a $N(-1/2, 4)$ and $N(1/2, 1)$ with equal probability. All

Table 2
Monte Carlo comparison of the T_n, CM_n, and KS_n tests (low frequency data).

n	T_n			CM_n			KS_n		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Size (DGP0)									
50	0.013	0.061	0.118	0.011	0.049	0.097	0.018	0.063	0.110
100	0.014	0.057	0.113	0.011	0.041	0.093	0.011	0.051	0.090
200	0.015	0.060	0.111	0.006	0.043	0.088	0.009	0.062	0.101
400	0.007	0.052	0.130	0.014	0.054	0.103	0.011	0.055	0.109
Power (DGP1)									
50	0.170	0.416	0.548	0.344	0.637	0.745	0.297	0.583	0.718
100	0.440	0.715	0.823	0.719	0.922	0.958	0.643	0.874	0.931
200	0.795	0.959	0.982	0.976	0.999	0.999	0.936	0.994	0.999
400	0.990	0.999	0.999	1.000	1.000	1.000	0.999	1.000	1.000

Table 3
Monte Carlo comparison of the T_n, CM_n, and KS_n tests (high frequency data).

n	T_n			CM_n			KS_n		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Size (DGP0)									
50	0.012	0.053	0.100	0.007	0.037	0.095	0.011	0.069	0.116
100	0.006	0.047	0.091	0.008	0.035	0.073	0.007	0.047	0.100
200	0.006	0.059	0.106	0.006	0.041	0.086	0.011	0.051	0.114
400	0.009	0.050	0.095	0.008	0.043	0.081	0.006	0.041	0.098
Power (DGP1)									
50	0.107	0.269	0.395	0.017	0.092	0.175	0.054	0.155	0.258
100	0.214	0.452	0.586	0.036	0.159	0.287	0.078	0.233	0.337
200	0.539	0.756	0.850	0.129	0.409	0.614	0.208	0.433	0.570
400	0.895	0.986	0.994	0.428	0.823	0.943	0.439	0.734	0.835

remaining particulars are the same as for the Monte Carlo setting considered above. Results are reported in Table 3.

Table 3 reveals that each test is correctly sized while power is highest for the smoothing tests as expected under 'high frequency' alternatives with the T_n test being most powerful for this DGP. Note that our so-called 'high frequency' density function only has two modes. Simulation results (not reported here to save space) show that, for density functions with more than two modes, a smoothing test enjoys additional power gains relative to non-smoothing tests. Therefore, smoothing tests complement non-smoothing tests and should be part of all applied researchers' standard toolkit.

4.4. Testing equality of conditional density functions with mixed data

Finally, we consider a Monte Carlo simulation designed to examine the finite-sample performance of the proposed J_{n,λ_0} and J_n tests and compare them to a counterpart that smooths the continuous variable with an ad-hoc h and uses the frequency indicator function for the discrete conditional variable.

Under the null of $f(x|w) = g(x|w)$ we let the discrete variable, w, assume four values with equal probability, $\{0, 1, \ldots, 3\}$. Next, we create $Y = z/4 + N(0, 1)$ under the null so that $f(x|w),\ g(x|w) \sim N(w/4, 1)$. Under the alternative $Y = z/4 + N(1/2, 1)$ so that $g(Y|w) \sim N(w/4 + 1/2, 1)$. All remaining particulars are the same as for the Monte Carlo setting considered above. Results are reported in Table 4.

Table 4 reveals that the proposed tests J_{n,λ_0} and J_n are correctly sized and both are more powerful than their ad-hoc/frequency counterpart $J_{n,h=1.06\sigma n^{-1/5}}$. Also, the J_{n,λ_0} test is more powerful than the J_n test due to the fact that J_{n,λ_0} also smooths the conditional discrete variable z.

5. Earnings, educational attainment, and wage gaps

There exists a large literature in labor economics regarding how returns to a college education is instrumental in understanding the widening wage gap in the US economy. The first step of such analysis, however, would involve determining whether or not statistically significant differences between joint distributions defined over both continuous (income) and discrete (educational attainment) variables exist.

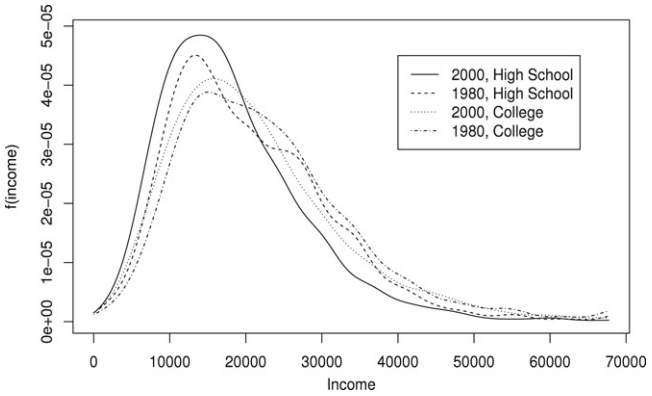

Fig. 1. Kernel-smoothed PDFs by year and educational attainment.

For what follows, we consider data spanning the years 1980–2000 constructed from the US Current Population Survey (CPS) March supplement on real incomes for white non-Hispanic workers aged 25 to 55 years who were full-time workers working at least 30 h a week and at least 40 weeks a year. Self-employed, farmers, unpaid family workers, and members of the Armed Forces are excluded. We consider the distribution of income for high school versus college graduates. Wage income is the income category considered, and figures are expressed in 2000 dollars.

Fig. 1 presents kernel smoothed PDF estimates for income by year and educational attainment.

Table 5 presents various moments for the income data, namely, measures of location and scale by year and educational attainment. We observe that average/median income for both high school

Table 4
Conditional density tests (J_{n,λ_0} and J_n).

n	J_{n,λ_0}			J_n			$J_{n,h=1.06\sigma n^{-1/5}}$		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
Size (DGP0)									
50	0.018	0.060	0.113	0.013	0.045	0.096	0.007	0.036	0.088
100	0.015	0.062	0.120	0.013	0.051	0.100	0.012	0.046	0.100
200	0.012	0.048	0.103	0.011	0.051	0.098	0.008	0.053	0.097
400	0.013	0.055	0.105	0.009	0.050	0.097	0.011	0.051	0.091
Power (DGP1)									
50	0.149	0.362	0.508	0.072	0.222	0.329	0.073	0.206	0.318
100	0.304	0.578	0.683	0.175	0.392	0.499	0.163	0.367	0.496
200	0.643	0.843	0.907	0.504	0.733	0.826	0.457	0.722	0.808
400	0.938	0.988	0.996	0.876	0.975	0.987	0.857	0.964	0.987

Table 5
Income location and scale summaries by year and educational attainment.

	1980		2000	
	High School	College	High School	College
Mean	$20,637.72	$22,838.88	$18,578.75	$22,104.75
Median	$18,880.17	$20,661.16	$16,647.06	$19,207.68
Stdev	$10,720.27	$11,767.74	$11,331.94	$14,547.50
IQR	$14,709.37	$14,876.03	$12,242.35	$13,805.52

and college graduates is lower in 2000 than it was in 1980. The interquartile range (IQR) has fallen for both groups from 1980 to 2000, while standard deviations have increased.

As noted in Section 1, moment-based tests, which only compare a finite number of moments from two distributions, are not consistent tests. By way of example, we test whether the joint distribution of earnings and educational attainment differ over time. We select two random samples, one for the year 1980 and one for the year 2000, each of size $n_1 = n_2 = 1000$, and apply the unconditional T_n and conditional J_n tests. We obtain $\hat{T}_n = 87.15$ with an associated bootstrap P-value of $P < 0.001$, while $\hat{J}_n = 54.63$ with an associated bootstrap P-value of $P < 0.001$. This suggests that there are indeed significant differences in the joint distribution of income and educational attainment between 1980 and 2000, and that there are significant differences in the distribution of income conditional upon educational attainment between 1980 and 2000.

6. Conclusion

We consider the problem of testing for equality of two density or two conditional density functions defined over mixed discrete and continuous data. Smoothing parameters are chosen via least squares cross-validation, and we smooth both the discrete and continuous variables in a particular manner. We advocate the use of bootstrap methods for obtaining the statistic's null distribution in finite-sample settings. Simulations show that the proposed tests enjoy power gains relative to both a conventional frequency-based test and a smoothing test based on ad hoc smoothing parameter selection. An application to testing for the equality of the joint distribution of income and educational attainments underscores the novelty and flexibility of the proposed approach in mixed data settings.

Our approach can be extended to testing the equality of two residual distributions. Hall et al. (2004) have shown that the cross-validation method has the remarkable ability of potentially removing *irrelevant* conditioning variables. In the testing framework we expect that this will lead to a more powerful test relative to peers that lack this ability. In this paper we only consider the case where the discrete variable has finite support. Extension of our approach to allow the discrete variable to be countably infinite will be a fruitful avenue for further investigation. We leave the exploration of these topics for future research.

Acknowledgements

We would like to thank three anonymous referees, an associate editor, and Peter Robinson for their numerous helpful comments that collectively led to a much improved version of this paper. Li's research is partially supported by the Private Enterprize Research Center, Texas A&M University. 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). We would also like to acknowledge Jose Galdo for his assistance with data management.

Appendix. Proofs of Theorems

Proof of Theorem 2.1. The test statistic I_n can be written as $I_n = I_{1n} + I_{2n}$, with

$$I_{1n} = -\frac{2}{n_1 n_2} \sum_{i=1}^{\min\{n_1,n_2\}} K_{\gamma,x_i,y_i}$$

and

$$I_{2n} = \sum_i \sum_{j\neq i} \left[\frac{1}{n_1(n_1-1)} K_{\gamma,x_i,x_j} + \frac{1}{n_2(n_2-1)} K_{\gamma,y_i,y_j} - \frac{1}{n_1 n_2} K_{\gamma,x_i,y_j} - \frac{1}{n_1 n_2} K_{\gamma,x_j,y_i} \right],$$

where $\sum_i = \sum_{i=1}^{n_1}$ if the summand contains x_i, and $\sum_i = \sum_{i=1}^{n_2}$ if the summand contains y_i. For example, $\sum_i \sum_{j\neq i} K_{\gamma,x_i,y_j} = \sum_{i=1}^{n_1} \sum_{j\neq i}^{n_2} K_{\gamma,x_i,y_j}$, and $\sum_i \sum_{j\neq i} K_{\gamma,x_j,y_i} = \sum_{i=1}^{n_2} \sum_{j\neq i}^{n_1} K_{\gamma,x_j,y_i}$.

Let $n = \min\{n_1, n_2\}$ and let $m(x, y)$ denote the joint density of (X_i^c, Y_i^c). By noting that $|L_{\lambda,x_i,y_i}| \leq 1$, we have that $E[|K_{\gamma,x_i,y_i}|] \leq E[|W_{h,x_i,y_i}|] = (h_1 \ldots h_q)^{-1} \int W((y_i^c - x_i^c)/h) m(x_i^c, y_i^c) \mathrm{d}x_i^c \mathrm{d}y_i^c = \int W(v) m(x_i^c, x_i^c + hv) \mathrm{d}v \mathrm{d}x_i^c = O(1)$. Then it follows that $E[\,|I_{1n}|\,] = (n_1 n_2)^{-1} O(\min\{n_1, n_2\} E|K_{\gamma,x_i,y_i}|) = O(n^{-1})$. This implies that

$$I_{1n} = O_p(n^{-1}). \tag{A.1}$$

Note that we obtain the above result by allowing for arbitrary correlation between X_i and Y_i. For example, for panel data with two time periods, X_i and Y_i can be repeated measures from the same individual i over two different time periods.

Next, we consider I_{2n}. We will write $A_n = B_n + (s.o.)$ to mean that B_n is the leading term of A_n, while $(s.o.)$ denotes terms having

orders smaller than B_n. Define $H_{n,ij} = K_{\gamma,x_i,x_j} + K_{\gamma,y_i,y_j} - K_{\gamma,x_i,y_j} - K_{\gamma,x_j,y_i}$. For $i \neq j$, we have

$$E[H_{n,ij}|X_i, Y_i] = E[K_{\gamma,x_i,x_j}|X_i] - E[K_{\gamma,x_i,y_j}|X_i] + E[K_{\gamma,y_i,y_j}|Y_i] - E[K_{\gamma,x_j,y_i}|Y_i] = 0,$$

which follows because $E[K_{\gamma,x_i,x_j}|X_i] = \int K_{\gamma,x_i,x_j} f(x_j)\mathrm{d}x_j = \int K_{\gamma,x_i,y_j} g(y_j)\mathrm{d}y_j = E[K_{\gamma,x_i,y_j}|X_i]$ (where $\int \mathrm{d}x = \sum_{x^d} \int \mathrm{d}x^c$) since $f(\cdot) = g(\cdot)$ under H_0.

Therefore, I_{2n} is a degenerate U-statistic. To save space we will write $(n_j - 1)$ by n_j $(j = 1, 2)$ as this will change any results asymptotically. Defining $H = h_1 \dots h_q$, then it is easy to show that

$$\begin{aligned}
\mathrm{var}(I_{2n}) &= E[(I_{2n})^2] \\
&= 2\sum_i \sum_{j\neq i} \left\{ n_1^{-4} E[(K_{\gamma,x_i,x_j})^2] + n_2^{-4} E[(K_{\gamma,y_i,y_j})^2] \right. \\
&\quad + (n_1 n_2)^{-2} E[(K_{\gamma,x_i,y_j})^2] \\
&\quad \left. + (n_1 n_2)^{-2} E[(K_{\gamma,x_j,y_i})^2] + (s.o.) \right\} \\
&= \frac{2}{n_1 n_2 H} \left\{ \left[\delta_n^{-1} + \delta_n + 2\right] \right. \\
&\quad \left. \times \left[E[f(X_i)] \left[\int W^2(v)\mathrm{d}v \right] + o(1) \right] \right\} \\
&\equiv (n_1 n_2 H)^{-1} \left\{ \sigma_0^2 + o(1) \right\},
\end{aligned}$$

where $\delta_n = n_1/n_2$, and $\sigma_0^2 = 2[\delta^{-1} + \delta + 2][E[f(X_i)]][\int W^2(v)\mathrm{d}v]$, and we have used

$$\begin{aligned}
E[(K_{\gamma,x_i,x_j})^2] &= H^{-2} \sum_{x_i^d, x_j^d} \int W^2((x_j^c - x_i^c)/h) L^2(x_i^d, x_j^d, \lambda) \\
&\quad \times f(x_i) f(x_j) \mathrm{d}x_i^c \mathrm{d}x_j^c \\
&= H^{-1} \sum_{x_i^d, x_j^d} \int W^2(v) L^2(x_i^d, x_j^d, \lambda) \\
&\quad \times f(x_i) f(x_i^c + hv, x_j^d) \mathrm{d}x_i^c \mathrm{d}v \\
&= H^{-1} \left\{ E[f(X_i)] \left[\int W^2(v)\mathrm{d}v \right] + o(1) \right\},
\end{aligned}$$

with $H = h_1 \dots h_q$, where $|h|^2 = \sum_{s=1}^q h_s^2$ and $|\lambda| = \sum_{s=1}^r \lambda_s$. Similarly, $E[(K_{\gamma,y_i,y_j})^2]$, $E[(K_{\gamma,x_i,y_j})^2]$, $E[(K_{\gamma,x_j,y_i})^2]$ all equal $H^{-1}\{E[f(X_i)][\int W^2(v)\mathrm{d}v] + o(1)\}$ under H_0 (since $E[f(X_i)] = E[g(Y_i)]$ under H_0).

It is straightforward, though tedious, to check that the conditions for the CLT of Hall (1984) for degenerate U-statistics holds. Thus, under H_0 we have

$$(n_1 n_2 H)^{1/2} I_{2n}/\sigma_0 \to N(0, 1) \text{ in distribution.} \tag{A.2}$$

Note that $E(\sigma_n^2) = \sigma_0^2 + o(1)$ (σ_n^2 is defined in Theorem 2.1), and by the U-statistic H-decomposition, it follows that $\sigma_n^2 = E(\sigma_n^2) + o_p(1) = \sigma_0^2 + o_p(1)$. Therefore, from (A.2) we obtain

$$(n_1 n_2 H)^{1/2} I_{2n}/\sigma_n \to N(0, 1) \text{ in distribution.} \tag{A.3}$$

In the above proof we implicitly assumed that the support of x^c is unbounded since we did not address the possible boundary bias problem. Below we show that, in fact, the above result holds true even when x^c has bounded support and the density is bounded below by a positive constant in its support. First, it is obvious that $E[H_{n,ij}|X_i, Y_i] = 0$ under H_0, because this follows from $f(\cdot) = g(\cdot)$, regardless of whether x^c has bounded or unbounded support. Next, we show that $\mathrm{var}(I_{2n}) = (n_1 n_2 H)^{-1}\{\sigma_0^2 + o(1)\}$ also holds true. It suffices to show that $E[(K_{\gamma,x_i,x_j})^2] = H^{-1}\{E[f(X_i)][\int W^2(v)\mathrm{d}v] + o(1)\}$. For expositional simplicity, we will only consider the univariate x^c case (and without x^d) where x^c is uniformly distributed in $[a, b]$ for some constants a, b with $b > a$. The proof for the general case is similar but much more tedious. Now let $\alpha \in (0, 1)$ be a constant. Then we have

$$\begin{aligned}
E[(W_{h,x_i,x_j})^2] &= H^{-2} \int_a^b \int_a^b W^2((x_j^c - x_i^c)/h) f(x_i^c) f(x_j^c) \mathrm{d}x_i^c \mathrm{d}x_j^c \\
&= H^{-1} \int_a^b \int_{(a-x_i)/h}^{(b-x_i)/h} W^2(v) f(x_i^c) f(x_i^c + hv) \mathrm{d}v \mathrm{d}x_i^c \\
&= H^{-1} \left[\int_a^{a+h^\alpha} + \int_{a+h^\alpha}^{b-h^\alpha} + \int_{b-h^\alpha}^b \right] \\
&\quad \times \int_{(a-x_i)/h}^{(b-x_i)/h} W^2(v) f(x_i^c) f(x_i^c + hv) \mathrm{d}v \mathrm{d}x_i^c \\
&= H^{-1} \left\{ \left[\int_{a+h^\alpha}^{b-h^\alpha} f(x_i^c)^2 \mathrm{d}x_i \right] \left[\int_{-\infty}^{\infty} W^2(v)\mathrm{d}v \right] + o(1) \right\} \\
&= H^{-1} \left\{ E[f(X_i^c)] \int W^2(v)\mathrm{d}v + o(1) \right\},
\end{aligned}$$

where we have used the fact that for $x_i \in (a + h^\alpha, b - h^\alpha)$, the interval $((a - x_i)/h, (b - x_i)/h) \supset (-h^\alpha/h, h^\alpha/h)$, which expands to $(-\infty, \infty)$ as $h \to 0$ since $0 < \alpha < 1$. Hence, $\int_{(a-x_i)/h}^{(b-x_i)/h} W(v^2)\mathrm{d}v \to \int_{-\infty}^{\infty} W^2(v)\mathrm{d}v$ as $h \to 0$. The basic idea underlying the proof above is as follows: We divide $[a, b]$ into three intervals: $[a, a + h^\alpha]$, $(a + h^\alpha, b - h^\alpha)$, and $[b - h^\alpha, b]$. Compared with the second interval, the first and third intervals are negligible since their lengths shrink to zero as $n \to \infty$. The second interval does not have a boundary problem as the boundary regions lie in the first and third intervals. The testing problem we consider here is different from pointwise estimation which may suffer from boundary bias issues. The leading term of the test statistic has zero mean and its asymptotic variance has the same expression, regardless of the nature of the support of x^c. Hence, the asymptotic null (normal) distribution presented in Theorem 2.1 is invariant to the nature of the support of x^c. Summarizing the above, (A.1) and (A.3) complete the proof of Theorem 2.1. □

Below we present a lemma which will be used in the proof of Theorem 2.2.

Lemma A.1. *Let $A_n(c) = (n_1 n_2 h_1 \dots h_q)^{1/2} I_{2n}(h, \lambda)$, where $h_s = a_s n^{-\zeta}$, $\lambda_s = b_s n^{-2\zeta}$, $c = (a_1, \dots, a_q, b_1, \dots, b_r)$, $c_s \in [C_{1s}, C_{2s}]$ with $0 < C_{1s} < C_{2s} < \infty$ $(s = 1, \dots, q + r)$.*

Then the stochastic process $A_n(c)$ indexed by c is tight under the sup-norm.

Proof. Writing $K_{\gamma,ij}$ as $(h_1 \dots h_q)^{-1} K_{c,ij}$ with $h_s = a_s n^{-\zeta}$ and $\lambda_s = b_2 n^{-2\zeta}$, where $K_{c,ij} = W\left(\frac{X_j - X_i}{h}\right) L(X_j^d, X_i^d, \lambda)$, and letting $\delta = q\zeta$, $H^{-1/2} = (h_1 \dots h_q)^{-1/2}$, $C_1 = (a_1, \dots, a_q)^{\mathrm{T}}$, $C_2 = (b_1, \dots, b_r)^{\mathrm{T}}$ (where the superscript T denotes transpose), $\bar{C}_1 = \prod_{s=1}^q a_s$, and $\bar{C}_2 = \prod_{s=1}^r b_s$. Then we have $H^{-1/2} K_{c,ij} = \bar{C}_1 n^{\delta/2} W_{C_1,ij} L_{C_2,ij}$. Let $C_2' = (b_1', \dots, b_r')^{\mathrm{T}}$, then note that $|L_{C_2',ij} - L_{C_2,ij}| \leq d_1 \sum_{s=1}^r |b_s - b_s'| \leq d_2 \|C_2 - C_2'\|$, where d_1 and d_2 are some finite positive constants.

Noting that the a_s are all bounded below by some positive constants, we have for all $s = 1, \dots, q$ that

$$\begin{aligned}
&\left| \frac{X_{is} - X_{js}}{h_s} - \frac{X_{is} - X_{js}}{h_s'} \right| \\
&\leq \left| X_{is} - X_{js} \right| \left| \frac{1}{h_s} - \frac{1}{h_s'} \right| = \left| \frac{X_{is} - X_{js}}{h_s} \right| \left| \frac{h_s' - h_s}{h_s'} \right|
\end{aligned}$$

$$= \left|\frac{X_{is}-X_{js}}{h_s}\right| \left|\frac{a_s'-a_s}{a_s'}\right| \le d_s \left|\frac{X_{is}-X_{js}}{h_s}\right| \left|a_s'-a_s\right|, \tag{A.4}$$

where d_s is a finite positive constant (a_s is bounded below and above by some positive constants).

By the Lipschitz condition (see (C2)) on the univariate kernel function, i.e., $|w(u)-w(v)| \le \xi(v)|u-v|$, it is easy to show that the product kernel also satisfies a Lipschitz condition, namely,

$$|W(u)-W(v)| \le M(v)\|u-v\|, \tag{A.5}$$

where $M(v) = c[\sum_{s=1}^{q}\xi(v_s)]$ and where c is a positive constant such that $\sup_{v\in\mathbb{R}} w(v)^{q-1} \le c$.

(A.4) and (A.5) yield

$$|W_{C_1',x_i,x_j} - W_{C_1,x_i,x_j}| \le d\, M_{C,x_i,x_j} \left\|\frac{X_i-X_j}{h}\right\| \times \|C_1'-C_1\|, \tag{A.6}$$

where d is a positive constant, and $M_{C_1,ij} = M((X_j-X_i)/h) = c[\sum_{s=1}^{q}\xi((X_{js}-X_{is})/h_s)]$.

Using (A.6) we have

$$\begin{aligned}
&\left|(H')^{-1/2}K_{C',ij} - H^{-1/2}K_{C,ij}\right| \\
&= \left|n^{\delta/2}\left\{(\bar{C}_1')^{-1/2}W_{C_1',ij}L_{C_2',ij} - \bar{C}_1^{-1/2}W_{C_1,ij}L_{C_2,ij}\right\}\right| \\
&= \left|n^{\delta/2}\left\{(\bar{C}_1')^{-1/2}W_{C_1',ij}\left[L_{C_2',ij}-L_{C_2,ij}\right]\right.\right. \\
&\quad \left.\left. + \left[(\bar{C}_1')^{-1/2}W_{C_1',ij} - \bar{C}_1^{-1/2}W_{C_1,ij}\right]L_{C_2,ij}\right\}\right| \\
&\le D_1\left\{(H')^{-1/2}W_{C_1',ij}\|C_2'-C_2\|\right. \\
&\quad \left. + H^{-1/2}M_{C_1,ij}\left\|\frac{X_j-X_i}{h}\right\| \times \|C_1'-C_1\|\right\},
\end{aligned} \tag{A.7}$$

where $D_1 > 0$ is a finite constant. In the last equality we used $|L_{C_2,ij}| \le 1$ and Assumption (C3), and we also replaced one of the $(\bar{C}_1')^{-1/2}$ by $\bar{C}_1^{-1/2}$ because $a_s \in [C_{1s}, C_{2s}]$ are all bounded from above and below. The difference can be absorbed into D_1.

By noting that $A_n(c') - A_n(c)$ is a degenerate U-statistic, and using (A.7), we have

$$\begin{aligned}
&E\left\{\left[A_n(c')-A_n(c)\right]^2\right\} = E\left\{\left[(H')^{-1/2}K_{c',ij} - H^{-1/2}K_{c,ij}\right]^2\right\} \\
&\le 4D_2E\left\{\left[(H')^{-1}W_{C_2',ij}^2\|C_2'-C_2\|^2\right.\right. \\
&\quad \left.\left. + H^{-1}M_{C,ij}^2\left\|\frac{X_j-X_i}{h}\right\|^2\|C_1'-C_1\|^2\right]\right\} \\
&\le D_3\left\{\left[\int\int f(x_i)f(x_i+hu)W^2(u)\mathrm{d}x_i\mathrm{d}u\right]\|C_2'-C_2\|^2\right. \\
&\quad \left. + \left[\int\int f(x_i)f(x_i+v)M^2(v)\|v\|^2\mathrm{d}x_i\mathrm{d}v\right]\|C_1'-C_1\|^2\right\} \\
&\le D_4\sup_x f(x)\left\{\left[\int W^2(u)\mathrm{d}u\right]\|C_2'-C_2\|^2\right. \\
&\quad \left. + \left[\int M^2(v)\|v\|^2\mathrm{d}v\right]\|C_1'-C_1\|^2\right\} \\
&\le D_5\|C'-C\|^2,
\end{aligned} \tag{A.8}$$

where D_j ($j = 2, 3, 4, 5$) are some finite positive constants. Therefore, $A_n(\cdot)$ (hence, $B_n(\cdot)$) is tight by Theorem 15.6 of Billingsley (1968, p. 128), or Theorem 3.1 of Ossiander (1987). □

Proof of Theorem 2.2. Theorem 2.1 implies that when $h_s = h_s^0 = a_s^0 n^{-\zeta}$ and $\lambda_s = \lambda_s^0 = b_s^0 n^{-2\zeta}$, the test statistic $\hat{T}_n(h^0, \lambda^0) \to N(0,1)$ in distribution. Therefore, it is sufficient to prove $\hat{T}_n(\hat{h}, \hat{\lambda}) - \hat{T}_n(h_0, \lambda_0) = o_p(1)$. For this, it suffices to show the following:

(i) $(n_1n_2\hat{h}_1\ldots\hat{h}_q)^{1/2}\hat{I}_{2n} = (n_1n_2h_1^0\ldots h_q^0)^{1/2}I_{2n} + o_p(1)$,
(ii) $(n_1n_2\hat{h}_1\ldots\hat{h}_q)^{1/2}\hat{I}_{1n} = (n_1n_2h_1^0\ldots h_q^0)^{1/2}I_{n1} = o_p(1)$, and
(iii) $\hat{\sigma}_n^2 = \sigma_0^2 + o_p(1)$, σ_0^2 is defined in Theorem 2.1 but with $(h_1,\ldots,h_q,\lambda_1,\ldots,\lambda_r)$ replaced by $(h_1^0,\ldots,h_q^0,\lambda_1^0,\ldots,\lambda_r^0)$.

Below we will only prove (i) since (ii) and (iii) are much easier to establish than (i). Write $\hat{h}_s = \hat{a}_s n^{-\zeta}$ and $\hat{\lambda}_s = \hat{b}_s n^{-2\zeta}$. From Theorem 3.1 of Li and Racine (2003), we know that $\hat{h}_s/h_s^0 - 1 \to 0$ and $\hat{\lambda}_s/\lambda_s^0 - 1 \to 0$ (in probability). This implies that $\hat{a}_s \to a_s^0$ and $\hat{b}_s \to b_s^0$ in probability. Let $\mathcal{C} = \prod_{s=1}^{q}[a_{1s}, a_{2s}] \times \prod_{t=1}^{r}[b_{1t}, b_{2t}]$, where a_{js} and b_{jt} ($j = 1, 2$) are some positive constants with $a_{1s} < a_s^0 < a_{2s}$ ($s = 1,\ldots,q$) and $b_{1t} < b_t^0 < b_{2t}$ ($t = 1,\ldots,r$). Let $c = (a_1,\ldots,a_q,b_1,\ldots,b_r)$, $c_0 = (a_1^0,\ldots,a_q^0,b_1^0,\ldots,b_r^0)$, and $\hat{c} = (\hat{a}_1,\ldots,\hat{a}_q,\hat{b}_1,\ldots,\hat{b}_r)$. Then Lemma A.1 shows that $A_n(c) \equiv (n_1n_2h_1\ldots h_q)^{1/2}I_{2n}(h,\lambda)$ (with $h_s = a_s n^{-\zeta}$ and $\lambda_s = b_s n^{-2\zeta}$) is tight in $c \in \mathcal{C}$.

Define $B_n(c) = A_n(c) - A_n(c_0)$. Then (i) becomes $B_n(\hat{c}) = o_p(1)$, i.e., we want to show that, for all $\epsilon > 0$,

$$\lim_{n\to\infty}\Pr\left[|B_n(\hat{c})| < \epsilon\right] = 1. \tag{A.9}$$

For any $\delta > 0$, denote the δ-ball centered at c_0 by $C_\delta = \{c : \|c-c_0\| \le \delta\}$, where $\|.\|$ denotes the Euclidean norm of a vector. By Lemma A.1 we know that $A_n(\cdot)$ is tight. By the Arzela–Ascoli Theorem (see Theorem 8.2 of Billingsley (1968, p. 55)) we know that tightness implies the following stochastic equicontinuous condition: for all $\epsilon > 0$, $\eta_1 > 0$, there exists a δ ($0 < \delta < 1$) and an N_1, such that

$$\Pr\left[\sup_{\|c'-c\|<\delta}\left|A_n(c')-A_n(c)\right| > \epsilon\right] < \eta_1 \tag{A.10}$$

for all $n \ge N_1$. (A.10) implies that

$$\Pr\left[|B_n(\hat{c})| > \epsilon, \hat{c} \in C_\delta\right] \le \Pr\left[\sup_{c\in C_\delta}|B_n(c)| > \epsilon\right] < \eta_1 \tag{A.11}$$

for all $n \ge N_1$. Also, from $\hat{c} \to c_0$ in probability we know that, for all $\eta_2 > 0$ and for the δ given above, there exists an N_2 such that

$$\Pr\left[\hat{c} \notin C_\delta\right] \equiv \Pr\left[\|\hat{c}-c_0\| > \delta\right] < \eta_2 \tag{A.12}$$

for all $n \ge N_2$. Therefore,

$$\begin{aligned}
\Pr\left[|B_n(\hat{c})| > \epsilon\right] &= \Pr\left[|B_n(\hat{c})| > \epsilon, \hat{c} \in C_\delta\right] \\
&\quad + \Pr\left[|B_n(\hat{c})| > \epsilon, \hat{c} \notin C_\delta\right] < \eta_1 + \eta_2
\end{aligned} \tag{A.13}$$

for all $n \ge \max\{N_1, N_2\}$ by (A.11) and (A.12), where we have also used the fact that $\{|B_n(\hat{c})| > \epsilon,\ \hat{c} \notin C_\delta\}$ is a subset of $\{\hat{c} \notin C_\delta\}$ (if A is a subset of B, then $P(A) \le P(B)$).

(A.13) is equivalent to (A.9). This completes the proof of (i). □

Proof of Theorem 2.3. In order to shorten the proof and to save space, we will only consider the case where $n_1 = n_2 = n$. Also, because the cumulative distribution function for the standard normal random variable is a continuous distribution, by Polyā's Theorem (Bhattacharya and Rao, 1986), we know that (3.17) is equivalent to, for a given value of $z \in \mathcal{R}$,

$$\left|P\left(\hat{T}_n^* \le z | \{X_i, Y_i\}_{i=1}^n\right) - \Phi(z)\right| = o_p(1). \tag{A.14}$$

First, we can write $\hat{I}_n^* = \hat{I}_{1n}^* + \hat{I}_{2n}^*$, where $\hat{I}_{jn}^*$ is the same as $\hat{I}_{jn}$ ($j = 1, 2$) except that $X_i(Y_i)$ is replaced by $X_i^*(Y_i^*)$ and (h, λ) is replaced by $(\hat{h}, \hat{\lambda})$. Let $E^*(\cdot)$ denote $E(\cdot|\{X_i, Y_i\}_{i=1}^n)$ and $P^*(\cdot)$ denote $P(\cdot|\{X_i, Y_i\}_{i=1}^n)$. Also, we write $B_n^* = o_p^*(1)$ to mean that, for all $\epsilon > 0$, $P^*[|B_n^*| > \epsilon] = o_p(1)$. It is easy to check that if $E^*[|B_n^*|] = o_p(1)$, then $B_n^* = o_p^*(1)$. Now from $I_{1n}^* = -\frac{2}{n(n-1)}\sum_{i=1}^n K_{\hat{\gamma},x_i^*,y_i^*}$, and noting that $E^*[|K_{\hat{\gamma},x_i^*,y_i^*}|] \le E^*[W_{\hat{h},x_i^*,y_i^*}] = (2n)^{-2}\sum_{i=1}^{2n}\sum_{j=1}^{2n} W_{\hat{h},z_i,z_j} = O_p(1)$, we obtain

$$I_{1n}^* = -\frac{2}{n(n-1)}\sum_{i=1}^n K_{\hat{\gamma},x_i^*,y_i^*} = O_p^*(n^{-1}). \tag{A.15}$$

Next, we consider

$$I_{2n}^* = \frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j>i}^n \left[K_{\hat{\gamma},x_i^*,x_j^*} + K_{\hat{\gamma},y_i^*,y_j^*} - K_{\hat{\gamma},x_i^*,y_j^*} - K_{\hat{\gamma},x_j^*,y_i^*}\right]$$
$$\equiv \frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j>i}^n H_{n,ij}^*,$$

where $H_{n,ij}^* = K_{\hat{\gamma},x_i^*,x_j^*} + K_{\hat{\gamma},y_i^*,y_j^*} - K_{\hat{\gamma},x_i^*,y_j^*} - K_{\hat{\gamma},x_j^*,y_i^*}$. I_{2n}^* is a degenerate U-statistic because

$E^*[H_{n,ij}^*|X_i^*, Y_i^*] = E[K_{\hat{\gamma},x_i^*,x_j^*}|X_i^*] - E[K_{\hat{\gamma},x_i^*,y_j^*}|X_i^*] + E[K_{\hat{\gamma},y_i^*,y_j^*}|Y_i^*] - E[K_{\hat{\gamma},x_j^*,y_i^*}|Y_i^*] = \frac{1}{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},x_i^*,z_j} - \frac{1}{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},x_i^*,z_j} + \frac{1}{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},y_i^*,z_j} - \frac{1}{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},z_j,y_i^*} = 0$ for almost all sample paths $(X_1, Y_1, X_2, Y_2, \ldots)$.

Let $U_{n,ij}^* = [2/n(n-1)]H_{n,ij}^*$, and define $U_n^* = \sum_i\sum_{j>i} U_{n,ij}^* \equiv I_{2n}^*$. We apply the CLT of de Jong (1987) for generalized quadratic forms to derive the asymptotic distribution of $U_n^*|\{X_i, Y_i\}_{i=1}^n$. By de Jong (1987, Proposition 3.2) we know that, conditional on $\{X_i, Y_i\}_{i=1}^n$, $U_n^*/S_n^* \to N(0, 1)$ in distribution if G_I^*, G_{II}^* and G_{IV}^* are all $o_p(S_n^{*4})$, where $S_n^{*2} = E^*[U_n^{*2}]$, $G_I^* = \sum_i\sum_{j>i} E^*[U_{n,ij}^{*4}]$, $G_{II}^* = \sum_i\sum_{j>i}\sum_{l>j>i}[E^*(U_{n,ij}^{*2}U_{n,il}^{*2}) + E^*(U_{n,ji}^{*2}U_{n,jl}^{*2}) + E^*(U_{n,li}^{*2}U_{n,lj}^{*2})]$, and $G_{IV}^* = (1/2)\sum_i\sum_{j>i}\sum_s\sum_{t>s} E^*(U_{n,is}^*U_{n,sj}^*U_{n,tl}^*U_{n,js}^*)$.

We will use the notation $A_n \sim B_n$ to denote that A_n and B_n have the same (probability) order of magnitude, and use the notation $A_n = O_e(a_n)$ to denote an exact order, i.e., it means that $A_n = O_p(a_n)$, but $A_n \ne o_p(a_n)$. Then we have

$$E^*[H_{n,ij}^{*2}] \sim E^*[K_{\hat{\gamma},x_i^*,x_j^*}^2] + E^*[K_{\hat{\gamma},y_i^*,y_j^*}^2] + E^*[K_{\hat{\gamma},x_i^*,y_j^*}^2] + E^*[K_{\hat{\gamma},x_j^*,y_i^*}^2]$$
$$= 4(2n)^{-2}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},z_i,z_j}^2 = O_e\left((\hat{h}_1\ldots\hat{h}_q)^{-1}\right).$$

Hence, $S_n^{*2} = \frac{4}{n^2(n-1)^2}\sum_{i=1}^{n-1}\sum_{j>i}^n E^*[H_{n,ij}^{*2}] \sim \frac{1}{n^2(n-1)^2}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},z_i,z_j}^2 = O_e(n^{-2}(\hat{h}_1\ldots\hat{h}_q)^{-1})$.

Thus we have, $1/S_n^{*2} = O_e(n^2(\hat{h}_1\ldots\hat{h}_q))$ and $1/S_n^{*4} = O_e(n^4(\hat{h}_1\ldots\hat{h}_q)^2)$.

Next,

$$E^*[H_{n,ij}^{*4}] \sim E^*[K_{\hat{\gamma},x_i^*,x_j^*}^4] + E^*[K_{\hat{\gamma},y_i^*,y_j^*}^4]$$
$$= 2(2n)^{-2}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},z_i,z_j}^4 = O_e((\hat{h}_1\ldots\hat{h}_q)^{-3}).$$

Hence, we have

$$G_I^* = [16/n^4(n-1)^4]\sum_i\sum_{j>i} E^*[U_{n,ij}^{*4}]$$
$$\sim [1/n^4(n-1)^4]\sum_{i=1}^{2n}\sum_{j=1}^{2n} K_{\hat{\gamma},z_i,z_j}^4 = O_p(n^{-6}(\hat{h}_1\ldots\hat{h}_q)^{-3}).$$

From the above calculation it should be apparent that the probability orders of G_I^*, G_{II}^* and G_{IV}^* are solely determined by the factor of n's and $(\hat{h}_1\ldots\hat{h}_p)$'s through $K_{ij,\hat{\gamma}}$. Therefore, tedious but straightforward calculations show that

$$G_{II}^* \sim n^{-8}\sum_{i=1}^{2n}\sum_{j=1}^{2n}\sum_{s=1}^{2n}[K_{ij,\hat{\gamma}}^2K_{is,\hat{\gamma}}^2 + K_{js,\hat{\gamma}}^2K_{ji,\hat{\gamma}}^2 + K_{si,\hat{\gamma}}^2K_{sj,\hat{\gamma}}^2]$$
$$= O_p(n^{-5}(\hat{h}_1\ldots\hat{h}_q)^{-2}),$$
$$G_{IV}^* \sim n^{-8}\sum_{i=1}^{2n}\sum_{j=1}^{2n}\sum_{s=1}^{2n}\sum_{t=1}^{2n}[K_{si,\hat{\gamma}}K_{sj,\hat{\gamma}}K_{ti,\hat{\gamma}}K_{tj,\hat{\gamma}}]$$
$$= O_p(n^{-4}(\hat{h}_1\ldots\hat{h}_q)^{-1}).$$

Therefore, $G_k^*/S_n^{*4} = o_p(1)$ for all $k = I, II, IV$. This means that for any subsequence $U_{n'}^*/S_{n'}^*$, there exists a further subsequence $U_{n''}^*/S_{n''}^*$ such that $(U_{n''}^*/S_{n''}^*|X_1, Y_1, X_2, Y_2, \ldots)$ converges to $(N(0, 1)|X_1, Y_1, X_2, Y_2, \ldots)$ for almost every sequence $(X_1, Y_1, X_2, Y_2, \ldots)$. Or equivalently, we have that

$$U_n^*/S_n^* \to N(0, 1) \text{ in distribution in probability.} \tag{A.16}$$

Next, define $V_{n,\hat{\gamma}}^* \overset{def}{=} \frac{2(\hat{h}_1\ldots\hat{h}_q)}{n(n-1)}\sum_i\sum_{j\ne i} E^*[H_{n,ij}^{*2}]$, and $\hat{V}_{n,\hat{\gamma}}^* \overset{def}{=} \frac{2(\hat{h}_1\ldots\hat{h}_q)}{n(n-1)}\sum_i\sum_{j\ne i} H_{n,ij}^{*2}$. Similar to the analysis of S_n^{*2}, one can show that $\hat{V}_{n,\hat{\gamma}}^* - V_{n,\hat{\gamma}}^* = o_p^*(1)$ and that $V_{n,\hat{\gamma}}^* - (n^2\hat{h}_1\ldots\hat{h}_q)S_n^{*2} = o_p^*(1)$. These results together with (A.16) tell us that

$$n(\hat{h}_1\ldots\hat{h}_q)^{1/2}I_{2n}^*/\sqrt{\hat{V}_{n,\hat{\gamma}}^*} \to N(0, 1) \text{ in distribution in probability.}$$

Since I_{2n}^* is the leading term of I_n^*, we conclude that $n(\hat{h}_1\ldots\hat{h}_q)^{1/2}I_n^*/\sqrt{\hat{V}_{n,\hat{\gamma}}^*}$ has the same asymptotic distribution as that of $n(\hat{h}_1\ldots\hat{h}_q)^{1/2}I_{2n}^*/\sqrt{\hat{V}_{n,\hat{\gamma}}^*}$. Hence, we have that

$$n(\hat{h}_1\ldots\hat{h}_q)^{1/2}I_n^*/\sqrt{\hat{V}_{n,\hat{\gamma}}^*} \to N(0, 1) \text{ in distribution in probability.}$$

□

Proof of Theorem 3.1. We know that $\hat{h}_s = h_s^0 + o_p(h_s^0)$ for $s = 1, \ldots, q$; and $\hat{\lambda}_s = \lambda_s^0 + o_p(\lambda_s^0)$ for $s = 1, \ldots, r$, where $h_0^s = a_s^0n^{-\zeta}$ and $\lambda_s^0 = b_s^0n^{-2\zeta}$ for some $\zeta > 0$. We will only prove Theorem 3.1 for the non-stochastic smoothing parameter case, i.e., for $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r) = (h_1^0, \ldots, h_q^0, \lambda_1^0, \ldots, \lambda_r^0)$, since the cross-validated smoothing parameter case follows by stochastic equicontinuity arguments analogous to those used in the proof of Theorem 2.2.

For expositional simplicity, we will assume that $n_1 = n_2 = n$.[5] We write

$$J_n = J_{1n} + J_{2n}, \tag{A.17}$$

where $J_{1n} = -\frac{2}{n^2\hat{p}_f\hat{p}_g}\sum_{w\in\mathcal{S}_w}\sum_{i=1}^n \bar{K}_{\gamma,x_i,y_i}I_{u_i,w}I_{v_i,w}$, and $J_{2n} = J_n - J_{1n}$. Using $n^{-2} = [n(n-1)]^{-1} + O(n^{-3})$ we can write J_{2n} as

$$J_{2n} = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\ne i}^n\sum_{w\in\mathcal{S}_w}\left\{\frac{\bar{K}_{\gamma,x_i,x_j}I_{u_i,w}I_{u_j,w}}{\hat{p}_f^2} + \frac{\bar{K}_{\gamma,y_i,y_j}I_{v_i,w}I_{v_j,w}}{\hat{p}_g^2} - \frac{1}{\hat{p}_f\hat{p}_g}\left[\bar{K}_{\gamma,x_i,y_j}I_{u_i,w}I_{v_j,w} + \bar{K}_{\gamma,x_j,y_i}I_{u_j,w}I_{v_i,w}\right]\right\} + O_p(n^{-1}). \tag{A.18}$$

[5] The proof for the general $n_1 \ne n_2$ case is similar, but much more tedious notationally.

Using the fact that $\hat{p}^{-1} = O_p(1)$ ($\hat{p} = \hat{p}_f(w)$ or $\hat{p} = \hat{p}_g(w)$) and by the same argument used in the proof of $I_{1n} = O_p(n^{-1})$, one can easily show that $J_{1n} = O_p(n^{-1})$.

Next, we consider J_{2n}. Using the fact that $\hat{p} - p = O_p(n^{-1/2})$ ($p = p_f(w)$ or $p = p_g(w)$), the following expansion immediately follows:

$$\begin{aligned}\frac{1}{\hat{p}} &= \frac{1}{p} + \frac{p-\hat{p}}{p^2} + O_p(n^{-1}),\\ \frac{1}{\hat{p}^2} &= \frac{1}{p^2} + \frac{2(p-\hat{p})}{p^3} + O_p(n^{-1}).\end{aligned} \tag{A.19}$$

Define the leave-two-out estimators $\hat{p}_{f,-(ij)} = (n-2)^{-1}\sum_{l\neq i,j}^{n} I_{u_l,w}$, and $\hat{p}_{g,-(ij)} = (n-2)^{-1}\sum_{l\neq i,j}^{n} I_{v_l,w}$. From $\hat{p}_f = n^{-1}\sum_{l=1}^{n} I_{u_l,w}$ and $\hat{p}_g = n^{-1}\sum_{l=1}^{n} I_{v_l,w}$, it is easy to see that $\hat{p}_f = \hat{p}_{f,-(ij)} + O_p(n^{-1})$ and $\hat{p}_g = \hat{p}_{g,-(ij)} + O_p(n^{-1})$ (uniformly in $i, j = 1, \ldots, n$). Hence, we can replace $\hat{p}_f$ and $\hat{p}_g$ in (A.18) by $\hat{p}_{f,-(ij)}$ and $\hat{p}_{g,-(ij)}$ without affecting the asymptotic behavior of J_{2n}.

Hence, substituting (A.19) into (A.18), and using the leave-two-out estimators $\hat{p}_{f,-(ij)}$ and $\hat{p}_{g,-(ij)}$ to replace $\hat{p}_f$ and $\hat{p}_g$, we obtain

$$J_{2n} = J^a_{2n} + J^b_{2n} + O_p(n^{-1}), \tag{A.20}$$

where J^a_{2n} is obtained from J_{2n} by replacing $\hat{p}_f$ and $\hat{p}_g$ by p_f and p_g in J_{2n}, i.e.,

$$\begin{aligned}J^a_{2n} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{w\in\mathcal{S}_w}\left[\frac{\bar{K}_{\gamma,x_i,x_j}I_{u_i,w}I_{u_j,w}}{p_f^2} + \frac{\bar{K}_{\gamma,y_i,y_j}I_{v_i,w}I_{v_j,w}}{p_g^2}\right.\\ &\quad\left. - \frac{1}{p_f p_g}\left[\bar{K}_{\gamma,x_i,y_j}I_{u_i,w}I_{v_j,w} + \bar{K}_{\gamma,x_j,y_i}I_{u_j,w}I_{v_i,w}\right]\right]\\ &\equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} H^a_{n,ij},\end{aligned}$$

where $H^a_{n,ij} = \sum_{w\in\mathcal{S}_w}\{\bar{K}_{\gamma,x_i,x_j}I_{u_i,w}I_{u_j,w}/p_f^2 + \bar{K}_{\gamma,y_i,y_j}I_{v_i,w}I_{v_j,w}/p_g^2 - [\bar{K}_{\gamma,x_i,y_j}I_{v_i,z}I_{v_j,w} + \bar{K}_{\gamma,x_j,y_i}I_{u_j,w}I_{v_i,w}]/(p_f p_g)\}$, and

$$\begin{aligned}J^b_{2n} &= \frac{1}{n(n-1)(n-2)}\\ &\quad\times\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{l\neq i,j}^{n}\sum_{w\in\mathcal{S}_w}\left\{\frac{2(p_f - I_{u_l,w})}{p_f^3}\bar{K}_{\gamma,x_i,x_j}I_{u_i,w}I_{u_j,w}\right.\\ &\quad + \frac{2(p_g - I_{v_l,w})}{p_g^3}\bar{K}_{\gamma,y_i,y_j}I_{v_i,w}I_{v_j,w} - \left[\frac{(p_g - I_{v_l,w})}{p_f p_g^2}\right.\\ &\quad\left.\left. + \frac{(p_f - I_{u_l,w})}{p_g p_f^2}\right]\left[\bar{K}_{\gamma,x_i,y_j}I_{u_i,w}I_{v_j,w} + \bar{K}_{\gamma,x_j,y_i}I_{u_j,w}I_{v_i,w}\right]\right\}.\end{aligned}$$

We will first analyze J^a_{2n}. J^a_{2n} is a second order U-Statistic. Below we show that it is a degenerate U-statistic. Letting $\mathcal{Z}_i = (X_i, V_i, Y_i, W_i)$, we have

$$\begin{aligned}E\left(H^a_{n,ij}|\mathcal{Z}_i\right) &= \sum_{w\in\mathcal{S}_w}\left\{\frac{I_{u_i,w}}{p_f}\left[p_f^{-1}E\left(\bar{K}_{\gamma,x_i,x_j}I_{u_j,w}|\mathcal{Z}_i\right)\right.\right.\\ &\quad - p_g^{-1}E\left(\bar{K}_{\gamma,x_i,y_j}I_{v_j,w}|\mathcal{Z}_i\right)\Big] + \frac{I_{v_i,w}}{p_g}\left[p_g^{-1}E\left(\bar{K}_{\gamma,y_i,y_j}I_{v_j,w}|\mathcal{Z}_i\right)\right.\\ &\quad\left.\left. - p_f^{-1}E\left(\bar{K}_{\gamma,x_j,y_i}I_{u_j,w}|\mathcal{Z}_i\right)\right]\right\} = 0,\end{aligned} \tag{A.21}$$

because $p_f^{-1}E(\bar{K}_{\gamma,x_i,x_j}I_{u_j,w}|\mathcal{Z}_i) - p_g^{-1}E(\bar{K}_{\gamma,x_i,y_j}I_{v_j,w}|\mathcal{Z}_i) = \int[f(x,w)/p_z(w)]\bar{K}_\gamma((x-x_i)/h)\mathrm{d}x - \int[g(x,w)/p_g(w)]\bar{K}_\gamma((x-x_i)/h)\mathrm{d}x = 0$ since $f(x,w)/p_f(w) = g(x,w)/p_g(w)$ under H_0^c.

By utilizing the CLT of Hall (1984) for degenerate U-statistics, one can show that

$$(n^2 h_1 \ldots h_q)^{-1/2} J^a_{2n}/\sigma_{n,J} \to N(0,1) \text{ in distribution,} \tag{A.22}$$

where

$$\begin{aligned}\sigma^2_{n,J} &= \frac{2(h_1\ldots h_q)}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{w\in\mathcal{S}_w} E\{(\bar{K}_{\gamma,x_i,x_j}I_{u_i,w}I_{u_j,w})^2 p_f(w)^{-4}\\ &\quad + (\bar{K}_{\gamma,y_i,y_j}I_{v_i,w}I_{v_j,w})^2 p_g(w)^{-4} + [(\bar{K}_{\gamma,x_i,y_j}I_{u_i,w}I_{v_j,w})^2\\ &\quad + (\bar{K}_{\gamma,y_i,x_j}I_{v_i,w}I_{u_j,w})^2](p_f(w)p_g(w))^{-2}\}.\end{aligned}$$

It is straightforward to show that $\hat{\sigma}^2_{n,J} = \sigma^2_{n,J} + o_p(1)$. Hence, we have

$$(n^2 h_1 \ldots h_q)^{-1/2} J^a_{2n}/\hat{\sigma}_{n,J} \to N(0,1) \text{ in distribution.} \tag{A.23}$$

Next, we consider J^b_{2n}. We will show that J^b_{2n} has an order smaller than that of J^a_{2n}. Because we only need to evaluate the order of J^b_{2n}, we will omit $\sum_{w\in\mathcal{S}_w}$ to simplify notation. Alternatively, one can observe that, for each $w \in \mathcal{S}_w$, we derive an upper bound for $J^b_{2n}(w)$ for any fixed value of w. Since $\mathcal{S}_w$ is a finite set, the same bound holds true for $\max_{w\in\mathcal{S}_w}|J^b_{2n}(w)|$.

Noting that J^b_{2n} contains three summations, it can therefore be written as a third order U-statistic:

$$\begin{aligned}J^b_{2n} &= \frac{1}{3n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{l\neq i,j}^{n}\left[J^b_{n,ijl} + J^b_{n,jil} + J^b_{n,lji}\right]\\ &\equiv \frac{1}{3n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{i\neq j,l}^{n} H^b_{n,ijl},\end{aligned}$$

where $H^b_{n,ijl} = J^b_{n,ijl} + J^b_{n,jil} + J^b_{n,lji}$,

$$\begin{aligned}J^b_{n,ijl} &= \frac{2}{p_f^3}(p_f - I_{u_i,w})\bar{K}_{\gamma,x_l,x_j}I_{u_l,w}I_{u_j,w}\\ &\quad + \frac{2}{p_g^3}(p_g - I_{v_i,w})\bar{K}_{\gamma,y_l,y_j}I_{v_l,w}I_{v_j,w}\\ &\quad - \left[\frac{(p_g - I_{v_i,w})}{p_f p_g^2} + \frac{(p_f - I_{u_i,z})}{p_g p_f^2}\right]\\ &\quad\times\left[\bar{K}_{\gamma,x_l,y_j}I_{u_l,w}I_{v_j,w} + \bar{K}_{\gamma,x_j,y_l}I_{u_j,w}I_{v_l,w}\right].\end{aligned}$$

Let $\mathcal{Z}_i = (X_i, Y_i, U_i, V_i)$. Below we show that $E(H^b_{n,ijl}|\mathcal{Z}_i) = 0$ under H_0^b. Note that

$$E(H^b_{n,ijl}|\mathcal{Z}_i) = E(J^b_{n,ijl}|\mathcal{Z}_i) + E(J^b_{n,jil}|\mathcal{Z}_i) + E(J^b_{n,lji}|\mathcal{Z}_i).$$

From $E(p_f - I_{v_j,z}) = 0$ we immediately have $E(J^b_{n,jil}|\mathcal{Z}_i) = 0$ and $E(J^b_{n,lji}|\mathcal{Z}_i) = 0$. Now,

$$\begin{aligned}E\left(J^b_{n,ijl}|\mathcal{Z}_i\right) &= (p_f - I_{u_i,w})p_f^{-2}\left[2p_f^{-1}E\left(\bar{K}_{\gamma,x_l,x_j}I_{u_l,w}I_{u_j,w}\right)\right.\\ &\quad - p_g^{-1}E\left(\bar{K}_{\gamma,x_l,y_j}I_{u_l,w}I_{v_j,w}\right) - p_g^{-1}E\left(\bar{K}_{\gamma,x_j,y_l}I_{u_j,w}I_{v_l,w}\right)\Big]\\ &\quad + (p_g - I_{v_i,w})p_g^{-2}\left[2p_g^{-1}E\left(\bar{K}_{\gamma,y_l,y_j}I_{v_l,w}I_{v_j,w}\right)\right.\\ &\quad\left. - p_f^{-1}E\left(\bar{K}_{\gamma,x_l,y_j}I_{u_l,w}I_{v_j,w}\right) - p_f^{-1}E\left(\bar{K}_{\gamma,x_j,y_l}I_{u_j,w}I_{v_l,w}\right)\right] = 0,\end{aligned}$$

by the same arguments as we used in the proof of $E(H^a_{n,ij}|\mathcal{Z}_i) = 0$ (since $f(x,w)/p_f(w) = g(x,w)/p_g(w)$). Hence, $E(H^a_{n,ijl}|\mathcal{Z}_i) = 0$, and J^b_n is a degenerate U-statistic. Define $H^b_{n,ij} = E(H^b_{n,ijl}|\mathcal{Z}_I, \mathcal{Z}_j)$. Then by a standard change-of-variables argument, one can show that

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} H^b_{n,ij}$$

$$= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} H^b_{n,ij,0} + O_p(|h|^2 n^{-1}), \tag{A.24}$$

where $|h|^2 = \sum_{s=1}^{q} h_s^2$, and

$$\begin{aligned} H^b_{n,ij,0} &= E(H^b_{n,ijl} | \mathcal{Z}_i, \mathcal{Z}_j) = 2p_f^{-3}(p_f - I_{u_i,w}) f(x_j, w) I_{u_j,w} \\ &\quad + 2p_g^{-3}(p_g - I_{v_i,w}) g(y_j, w) I_{v_j,w} \\ &\quad - \left[\frac{(p_g - I_{v_i,w})}{p_f p_g^2} + \frac{(p_f - I_{u_i,z})}{p_g p_f^2} \right] \\ &\quad \times \left[f(y_j, w) I_{v_j,w} + g(x_j, w) I_{u_j,w} \right]. \end{aligned}$$

Therefore, by the U-statistic H-decomposition (see Lee (1990)), we have $J^b_{2n} = J^b_{2n,0} + O_p(n^{-1})$, where $J^b_{2n,0} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} H^b_{n,ij,0}$. Note that $E(H^b_{n,ij,0} | \mathcal{Z}_i) = 0$, i.e., $J^b_{2n,0}$ is a degenerate U-statistic. Also note that $H^b_{n,ij,0}$ is unrelated to the smoothing parameters $(h_1, \ldots, h_q)$. Then it is easy to show that $E[(J^b_{2n,0})^2] = O(n^{-2})$, which implies that $J^b_{2n,0} = O_p(n^{-1})$. This together with (A.24) lead to

$$J^b_{2n} = O_p\left(n^{-1}\right) = o_p\left(\left(n^2 h_1 \ldots h_q\right)^{-1/2}\right). \tag{A.25}$$

Combining (A.17), (A.20), (A.23) and (A.25), and the fact that $J_{1n} = O_p(n^{-1})$, this completes the proof of Theorem 3.1. □

References

Ahmad, I., van Belle, G., 1974. Measuring affinity of distributions. In: Proschan, Serfling, R. (Eds.), Reliability and Biometry, Statistical Analysis of Life Testing. SIAM.

Ahmad, I.A., Li, Q., 1997. Testing independence by nonparametric kernel method. Statistics and Probability Letters 34, 201–210.

Aitchison, J., Aitken, C.G.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63 (3), 413–420.

Anderson, G., 2001. The power and size of nonparametric tests for common distributional characteristics. Econometric Reviews 20 (1), 1–30.

Anderson, N.H., Hall, P., Titterington, D.M., 1994. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. Journal of Multivariate Analysis 50, 41–54.

Bhattacharya, R.N., Rao, R.R., 1986. Normal Approximations and Asymptotic Expansions. R.E. Krieger Publishing Company.

Billingsley, P., 1968. Convergence of Probability Measures. Wiley.

Davidson, R., MacKinnon, J.G., 2000. Bootstrap tests: How many bootstraps? Econometric Reviews 19, 55–68.

de Jong, P., 1987. A central limit theorem for generalized quadratic forms. Probability Theory and Related Fields 75, 261–277.

Fan, Y., 1998. Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. Econometric Theory 14, 604–621.

Fan, Y., Gencay, R., 1993. Hypothesis testing based on modified nonparametric estimation of an affinity measure between two distributions. Journal of Nonparametric Statistics 4, 389–403.

Fan, Y., Li, Q., 1999. Central limit theorem for degenerate u-statistics of absolute regular processes with application to model specification testing. Journal of Nonparametric Statistics 10, 245–271.

Fan, Y., Li, Q., 2000. Consistent model specification tests: Kernel-based tests versus Bierens's icm tests. Econometric Theory 16, 1016–1041.

Fan, Y., Ullah, A., 1999. On goodness-of-fit tests for weakly dependent processes using kernel method. Journal of Nonparametric Statistics 11, 337–360.

Ghosh, B.K., Huang, W.M., 1991. The power and optimal kernel of the Bickel–Rosenblatt test for goodness-of-fit. Annals of Statistics 19, 999–1009.

Grund, B., Hall, P., 1993. On the performance of kernel estimators for high-dimensional sparse binary data. Journal of Multivariate Analysis 44, 321–344.

Hall, P., 1981. On nonparametric multivariate binary discrimination. Biometrika 68 (1), 287–294.

Hall, P., 1984. Central limit theorem for integrated square error of multivariate nonparametric density estimators. Journal of Multivariate Analysis 14, 1–16.

Hall, P., Li, Q., Racine, J.S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. The Review of Economics and Statistics 89, 784–789.

Hall, P., Racine, J.S., Li, Q., 2004. Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association 99 (468), 1015–1026.

Hong, Y., White, H., 2005. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Econometrica 73 (3), 837–901.

Kiefer, N.M., Racine, J.S., 2008. The smooth colonel meets the reverend. Manuscript. Cornell University.

Lee, J., 1990. U-statistics: Theory and Practice. Marcel Dekker, New York.

Li, Q., 1996. Nonparametric testing of closeness between two unknown distributions. Econometric Reviews 15, 261–274.

Li, Q., Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. Journal of Multivariate Analysis 86, 266–292.

Mammen, E., 1992. When does bootstrap work? In: Asymptotic Results and Simulations. Springer-Verlag, New York.

Ossiander, M., 1987. A central limit theorem under metric entropy with L_2 bracketing. The Annals of Probability 15 (3), 897–919.

Ouyang, D., Li, Q., Racine, J.S., 2006. Cross-validation and the estimation of probability distributions with categorical data. Journal of Nonparametric Statistics 18 (1), 69–100.

Racine, J., Maasoumi, E., 2007. A versatile and robust metric entropy test of time reversibility and dependence. Journal of Econometrics 138, 547–567.

Robinson, P.M., 1991. Consistent nonparametric entropy-based testing. Review of Economic Studies 58, 437–453.

Rosenblatt, M., 1975. A quadratic measure of deviation of two-dimensional density estimates and a test of independence. Annals of Statistics 3, 1–14.

Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer.

Stone, C.J., 1984. An asymptotically optimal window selection rule for kernel density estimates. Annals of Statistics 12, 1285–1297.