

A NONPARAMETRIC TEST FOR EQUALITY OF DISTRIBUTIONS WITH MIXED CATEGORICAL AND CONTINUOUS DATA

QI LI
DEPARTMENT OF ECONOMICS
TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-4228

ESFANDIAR MAASOUMI
DEPARTMENT OF ECONOMICS
SOUTHERN METHODIST UNIVERSITY
DALLAS, TX 75275-0496

JEFF RACINE
DEPARTMENT OF ECONOMICS
SYRACUSE UNIVERSITY
SYRACUSE, NY 13244-1020

ABSTRACT. In this paper we consider the problem of testing for equality of two 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 statistic is shown to have an (asymptotic) normal null distribution. However, we advocate the use of bootstrap methods in order to better approximate its null distribution in finite-sample settings. Simulations show that the proposed test enjoys substantial power gains relative to both a conventional frequency-based test and a smoothing test based on ad hoc smoothing parameter selection, while a demonstrative empirical application to the wage income ‘differential’ between men and women underscores the utility of the proposed approach in mixed data settings.

Key words: Mixed discrete and continuous variables; Density testing; Nonparametric smoothing; Cross-validation.

Date: June 3, 2004.

Li’s research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine would like to thank the Center for Policy Research at Syracuse University for their generous support, and Jose Galdo for his data efforts.

E. Maasoumi is the corresponding author. His contact information is Department of Economics, Southern Methodist University, Dallas, TX 75275-0496, Email: maasoumi@mail.smu.edu, Tel: (214) 768-4298.

1. INTRODUCTION

It is difficult to think of a more ubiquitous test in applied statistics than the test for equality of distributions. The most popular variants involve simply testing whether moments of two distributions, such as their means and/or variances, differ, or perhaps whether their quantiles differ. Comparing distributions, or reconstructing indirectly observed distributions (such as the counterfactuals in program evaluation) is implicit and ever present in almost all statistical/econometric work. It is not difficult to conjure up cases for which moment-based tests lack power, however, while the same can be said for parametric tests requiring specification of the null distribution. Generally, interest truly lies in detecting *any* potential difference between two distributions, 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 & van Belle (1974), Mammen (1992), Fan & Gencay (1993), Li (1996), Fan & 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. One might instead, however, consider kernel “smoothing” the discrete variables as well, and there is a rich literature in statistics on smoothing discrete variables and its potential benefits; see Aitchison & Aitken (1976), Hall (1981), Grund & Hall (1993), Scott (1992), Simonoff (1996), and Li & Racine (2003), among others. Though smoothing discrete variables 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 highly desirable. The test

developed herein is an extension of existing frequency-based ‘smooth’ kernel tests, while ‘non-smooth’ (i.e. empirical CDF) tests of distributional differences have recently been examined and reviewed in Anderson (2001).

In this paper we propose a kernel-based test for equality of distributions mounted on a square integral metric defined over mixed continuous/discrete variables. Similar 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 & White (2000), Racine & Maasoumi (Under Revision), and Ahmad & Li (1997). We use data-driven bandwidth selection methods, smooth both the continuous and discrete variables, 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 size and 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 also we 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 for the continuous and discrete variables involved. Section 2 presents the test statistics and their properties, Section 3 presents two simulation experiments designed to assess the finite-sample performance of the estimator, while Section 4 presents a demonstrative empirical application to the wage income ‘differential’ between men and women. Section 5 concludes, and all proofs are relegated to the appendix.

2. THE TEST STATISTIC

We consider the case where we are faced with a mixture of discrete and continuous data. Let $X = (X^c, X^d) \in R^q \times \mathcal{S}^r$, where X^c is the continuous variable having dimension q , and X^d is the discrete variable having dimension r and assuming values in $\mathcal{S}^r = \prod_{s=1}^r \{0, 1, \dots, c_s - 1\}$. 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. We are interested in testing the null hypothesis that

$$H_0 : f(x) = g(x) \text{ for almost all } x \in R^q \times \mathcal{S}^r$$

against the alternative hypothesis H_1 that $f(x) \neq g(x)$ on a set with positive measure. We first discuss how to estimate $f(\cdot)$ and $g(\cdot)$ and then outline the test statistic.

Let x_s^d and X_{is}^d denote the s th components of x^d and X_i^d respectively. Following Aitchison & Aitken (1976), for $x_s, X_{is}^d \in \mathcal{S}_s^r = \{0, 1, \dots, c_s - 1\}$ (x_s^d takes c_s different values), we define a univariate kernel function

$$(1.1) \quad l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 1 - \lambda_s & \text{if } X_{is}^d = x_s^d, \\ \lambda_s / (c_s - 1) & \text{if } X_{is}^d \neq x_s^d, \end{cases}$$

where the range of 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. Here we use $I(\cdot)$ to denote an indicator function, $I(A) = 1$ if the event A holds true, zero otherwise. If $\lambda_s = (c_s - 1)/c_s$, $l(X_{is}^d, x_s^d, \frac{c_s - 1}{c_s}) = 1/c_s$ 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

$$(1.2) \quad L_{\lambda, x, x_i} = \prod_{t=1}^r l(X_{it}^d, x_t^d, \lambda_t) = \prod_{s=1}^r \{\lambda_s / (c_s - 1)\}^{I_{x_{is}^d \neq x_s^d}} (1 - \lambda_s)^{I_{x_{is}^d = x_s^d}},$$

where $I_{x_{is}^d \neq x_s^d} = I(X_{is}^d \neq x_s^d)$, and $I_{x_{is}^d = x_s^d} = I(X_{is}^d = x_s^d)$.

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

$$(1.3) \quad W_{h,x,x_i} = \prod_{s=1}^q \frac{1}{h_s} w\left(\frac{X_{is}^c - x_s^c}{h_s}\right).$$

The final product kernel for all components, discrete and continuous, is given by

$$(1.4) \quad K_{\gamma,x,x_i} = W_{h,x,x_i} L_{\lambda,x,x_i},$$

where $\gamma = (h, \lambda)$, L_{λ,x,x_i} and W_{h,x,x_i} are defined in (1.2) and (1.3), respectively.

We estimate the joint density of $f(x)$ by

$$(1.5) \quad \hat{f}(x) = \frac{1}{n_1} \sum_{i=1}^{n_1} K_{\gamma,x,x_i},$$

where $K_{\gamma,x,x_i} = W_{h,x,x_i} L_{\lambda,x,x_i}$, $\gamma = (h, \lambda)$.

Similarly, we estimate the joint density of $g(x)$ by

$$(1.6) \quad \hat{g}(x) = \frac{1}{n_2} \sum_{i=1}^{n_2} K_{\gamma,x,y_i},$$

where $K_{\gamma,x,y_i} = W_{h,x,y_i} L_{\lambda,x,y_i}$.

A test statistic can be constructed based on the integrated squared density difference given by $I = \int [f(x) - g(x)]^2 dx = \int [f(x)dF(x) + g(x)dG(x) - 2f(x)dG(x)]$. $F(\cdot)$ and $G(\cdot)$ are the cumulative distribution functions for X and Y , respectively, and $\int dx = \sum_{x^d \in \mathcal{S}^d} \int dx^c$. 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,

$$(1.7) \quad \begin{aligned} I_n &= \frac{1}{n_1} \sum_{i=1}^{n_1} \hat{f}(X_i) + \frac{1}{n_2} \sum_{i=1}^{n_2} \hat{g}(Y_i) - \frac{2}{n_2} \sum_{i=1}^{n_2} \hat{f}(Y_i) \\ &= \frac{1}{n_1^2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K_{\gamma,x_i,x_j} + \frac{1}{n_2^2} \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} K_{\gamma,y_i,y_j} - \frac{2}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{\gamma,x_i,y_j}. \end{aligned}$$

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 $x^d \in \mathcal{S}^r$, both $f(\cdot, x^d)$ and $g(\cdot, x^d)$ are bounded and continuous functions (continuous with respect to x^c). The kernel function $w(\cdot)$ is a bounded, non-negative second order kernel.

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

Note that in (C1) we assume X_i (Y_i) is independent of X_j (Y_j) for $j \neq i$. When $n_1 = n_2 = n$, however, we do allow for the possibility that X_i and Y_i are correlated, as in panel-type cases where data are collected from n individuals for two different time periods. 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 for degenerate U-statistics with weakly dependent data as given in Fan & 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 below) will also need to be modified; block or stationary bootstrapping or subsampling methods will be more 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 hold 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, it requires that the two sample sizes have the same order of magnitude.

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

Theorem 2.1. *Under conditions (C1) to (C3), we have, under H_0 , that*

$$T_n = (n_1 n_2 h_1 \dots h_q)^{1/2} (I_n - c_n) / \sigma_n \rightarrow N(0, 1) \text{ in distribution,}$$

where $c_n = \frac{w(0)^q}{h_1 \dots h_q} [\prod_{s=1}^r (1 - \lambda_s)] \left[\frac{1}{n_1} + \frac{1}{n_2} \right]$, and where

$$\sigma_n^2 = 2n_1 n_2 h_1 \dots h_q \left[\sum_{i=1}^{n_1} \sum_{j=1}^{n_1} \frac{(K_{\gamma,ij}^x)^2}{n_1^4} + \sum_{i=1}^{n_2} \sum_{j=1}^{n_2} \frac{(K_{\gamma,ij}^y)^2}{n_2^4} + 2 \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{(K_{\gamma,ij}^{x,y})^2}{n_1^2 n_2^2} \right].$$

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 $I_n \rightarrow \int [f(x) - g(x)]^2 dx \equiv C > 0$ (in probability), $c_n = o(1)$, $\sigma_n = O_p(1)$. Hence, T_n will have the order 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 size and 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 the following cross-validation method for selecting (h, λ) .

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 $z_i = y_i$ for $n_1 + 1 \leq i \leq n_1 + n_2$. Let $\tilde{f}(z_i) = (N - 1)^{-1} \sum_{j \neq i}^N K_{\gamma, z_i, z_j}$ be the leave-one-out estimate of $f(z_i)$. We choose (h, λ) to minimize the following cross-validation function:

$$(1.8) \quad CV(h, \lambda) = \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \bar{K}_{\gamma, z_i, z_j} - \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i}^N K_{\gamma, z_i, z_j},$$

where $\bar{K}_{\gamma, ij}^z = \bar{W}_{h, ij} \bar{L}_{\lambda, ij}$, $\bar{W}_{h, ij} = \prod_{s=1}^q h_s^{-1} \bar{w} \left(\frac{z_i^c - z_j^c}{h_s} \right)$, $\bar{w}(v) = \int w(u) w(v - u) du$ is the two-fold convolution kernel, $\bar{L}_{\lambda, ij} = \sum_{z \in \mathcal{S}^r} L_{\lambda, x, x_i} L_{\lambda, x, x_j}$, and $K_{\gamma, ij}^z = W_{h, z_i, z_j} L_{\lambda, z_i, z_j}$.

Letting $(\hat{h}_1, \dots, \hat{h}_q)$ and $(\hat{\lambda}_1, \dots, \hat{\lambda}_r)$ denote the cross-validated values of (h_1, \dots, h_q) and $(\lambda_1, \dots, \lambda_r)$, Li & Racine (2003)¹ have shown that $\hat{h}_s/h_s^0 - 1 \rightarrow 0$ in probability, and $\hat{\lambda}_s/\lambda_s^0 -$

¹Li & Racine (2003) only consider the case for which $h_1 = \dots = h_q = h$ and $\lambda_1 = \dots = \lambda_r = \lambda$. It is straightforward to generalize the result of Li & Racine (2003) to the vector h and λ case, and the result should be modified as given here.

$1 \rightarrow 0$ in probability, where $h_s^0 = a_s^0 n^{-1/(q+4)}$, and $\lambda_s^0 = b_s^0 n^{-2/(q+4)}$, a_s^0 and b_s^0 are some finite constants, while h_s^0 and λ_s^0 are the optimal smoothing parameters that minimize the integrated squared difference $E[\int(\hat{f}(z) - f(z))^2 dz]$.

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. *Under conditions (C1) to (C3), under H_0 we have*

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

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

The proof of Theorem 2.2 is given in the appendix.

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 , we advocate the use of the following bootstrap procedure in applied settings.

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 them $\{y_i^*\}_{i=1}^{n_2}$. Compute a test statistic \hat{T}_n^* in the same way as \hat{T}_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 $B = 1,000$), 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 .

Note that we use the same $(\hat{h}, \hat{\lambda})$ when computing \hat{T}_n^* , i.e., we do not 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 shows that the bootstrap method works.

Theorem 2.3. *Under conditions (C1) to (C3), we have*

$$\hat{T}_n^* = (n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} (\hat{T}_n^* - \hat{c}_n^*) / \hat{\sigma}_n^* \rightarrow N(0, 1) \text{ in distribution in probability,}$$

where \hat{c}_n^* and $\hat{\sigma}_n^*$ are defined the same way as in \hat{c}_n and $\hat{\sigma}_n$ but with (x_i, y_i) replaced by (x_i^*, y_i^*) .

The proof of Theorem 2.3 is given in the appendix.

In the bootstrap hypothesis testing literature, the notion of ‘convergence in distribution with probability one’ is often used to describe the asymptotic behavior of bootstrap tests. ‘Convergence in distribution in probability’ is much easier to establish than ‘convergence in distribution with probability one’, and runs parallel to that of ‘convergence in probability’ and ‘convergence with probability one’; see Li, Hsiao & Zinn (2003) for a detailed definition of ‘convergence in distribution in probability’.

3. MONTE CARLO SIMULATIONS

We consider the finite-sample performance of the proposed test. In particular, we consider the behavior of the test relative to the conventional frequency approach for mixed data.

3.1. Testing Equality of Density Functions with Mixed Data. We consider two mixed data DGPs. The first allows us to examine the test’s size, while the second permits an assessment of power. For DGP0, we have

$$g(x, z) \equiv f(x, z) \sim f(x)p(z), f(x) \sim N(0, 1),$$

$$z \in \{0, 1, 2, 3\}, Pr(Z = j) = (0.20, 0.30, 0.15, 0.35), \text{ for } j = 0, \dots, 3,$$

while for DGP1, we have

$$\begin{aligned}
 f(x, z) &\sim f(x)p(z), f(x) \sim N(0, 1), \\
 Pr(Z = j) &= (0.20, 0.30, 0.15, 0.35), \text{ for } j = 0, \dots, 3, \\
 g(x, z) &\sim g(x)p(z), g(x) \sim N(0.5, 1), \\
 Pr(Z = j) &= (0.20, 0.30, 0.15, 0.35), \text{ for } j = 0, \dots, 3.
 \end{aligned}$$

That is, the continuous components of $f(\cdot)$ and $g(\cdot)$ differ in their means under the alternative. Evidently, by looking at cases in which conventional tests may perform well we are being conservative relative to the power performance of our proposed tests in general.

We consider three tests of the hypothesis $H_0 : g(x, z) = f(x, z)$ a.e: i) the proposed test with cross-validated h and λ , ii) the conventional frequency test with cross-validated h and $\lambda = 0$, and iii) the conventional ad hoc test with $h = 1.06\sigma n^{-1/5}$ and $\lambda = 0$. Empirical size and power are summarized in tables 1 through 3.

Tables 1 through 3 suggest the following; i) our test is very correctly sized, while the other test sizes are reasonable as well, (ii) the proposed method enjoys substantial power gains, especially in small sample situations relative to the conventional frequency test ($\lambda = 0$), (iii) cross-validation works quite well in this setting, yielding results for even the conventional frequency test that are comparable to those based on the optimal bandwidth $h = 1.06\sigma n^{-1/5}$, and (iv) the consistency of the tests is evident in the large sample experiments with power approaching one.

4. AN APPLICATION TO PANEL DATA

We consider a data panel for 1980-2000 constructed from the Current Population Survey (CPS) March supplement on real incomes for white non-Hispanic workers having a high school education ages 25 to 55 years who were full-time workers working at least 30 hours a week and at least 40 weeks a year. Self-employed, farmers, unpaid family workers, and

TABLE 1. Mixed Data, CV h, λ

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
DGP0			
50	0.007	0.050	0.102
100	0.007	0.051	0.102
200	0.011	0.050	0.106
400	0.006	0.048	0.103
DGP1			
50	0.116	0.288	0.406
100	0.253	0.491	0.620
200	0.511	0.756	0.856
400	0.921	0.981	0.993

TABLE 2. Mixed Data, CV $h, \lambda = 0$

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
DGP0			
50	0.008	0.064	0.119
100	0.007	0.052	0.106
200	0.014	0.053	0.110
400	0.007	0.048	0.101
DGP1			
50	0.044	0.174	0.294
100	0.146	0.354	0.511
200	0.400	0.659	0.779
400	0.877	0.971	0.989

TABLE 3. Mixed Data, Ad Hoc $h = 1.06\sigma n^{-1/5}, \lambda = 0$

n	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
DGP0			
50	0.009	0.059	0.125
100	0.007	0.047	0.103
200	0.011	0.052	0.106
400	0.008	0.054	0.094
DGP1			
50	0.045	0.177	0.291
100	0.139	0.343	0.486
200	0.369	0.631	0.748
400	0.840	0.956	0.981

members of the Armed Forces are excluded. Wage income is the income category considered. Since CPS is not a “panel” of repeated observations on the same subjects, “dependence” over time is thought to be less of an issue here.

4.1. Testing Male Versus Female Income Equality. We first randomly sample 100 males and females per year (from the frame described earlier) to construct a male and female panel consisting of time (year treated as an ordered categorical variable) and real annual income (treated as continuous), with each panel having 2,100 observations. In this set up, dependence between the two panels is also a non-issue.

Figures 1 and 2 plot the estimated joint distribution of earnings and time. Bandwidths were selected via cross-validation and were $\hat{h} = 1.81\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.17$ for females ($\hat{\sigma}$ is the sample standard deviation of income), and $\hat{h} = 1.36\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.15$ for males. It can be seen that the distribution of female income appears to be more concentrated at lower incomes than for males.

We apply the proposed test using 399 bootstrap replications, resulting in $\hat{T}_n = 55.2$ with the 90th, 95th, and 99th percentiles under the null being 0.21, 0.67, and 1.31 respectively. The null of equality of male and female income distributions for the period 1980-2000 is soundly rejected, while the resampled percentiles indicate that the limiting normal distribution provides a poor approximation to the finite-sample null distribution even for a fairly large pooled sample of size 4,200. Our sample frame is sufficiently narrow and allows only age (and perhaps marital status) as a further explanation of this gender wage differential.

4.2. Testing Income Equality Over Time. Next we consider testing whether the joint distribution of incomes for males and females in a given year changes significantly over time. We randomly select 250 males and 250 females (from the original frame described earlier) for a given year to construct our joint sample, then apply the test for equality of the joint income/sex (continuous/discrete) distribution for two different time spans. We consider 1980 versus 2000 and 1990 versus 1995. The estimated joint densities are plotted in figures 3 and 4. Bandwidths were selected via cross-validation and were $\hat{h} = 1.91\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.000$ for 1980, $\hat{h} = 1.26\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.001$ for 1990, $\hat{h} = 0.80\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.076$ for 1995, and $\hat{h} = 0.80\hat{\sigma}n^{-1/5}$ and $\hat{\lambda} = 0.097$ for 2000, where $\hat{\sigma}$ is the sample standard deviation of income.

FIGURE 1. PDF of male real income

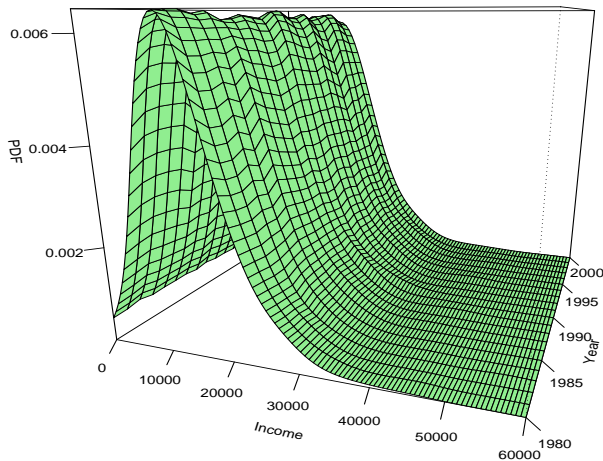
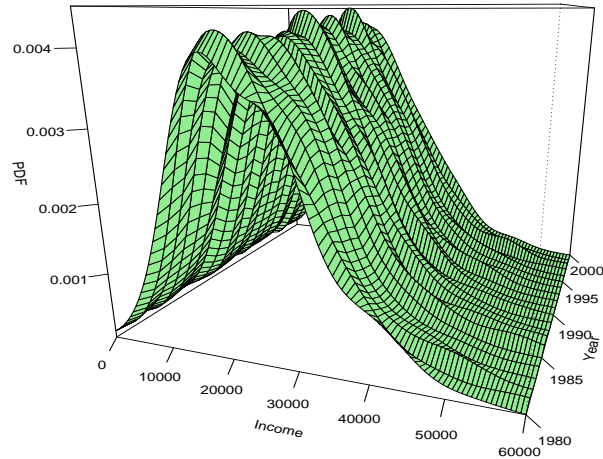


FIGURE 2. PDF of female real income

The density plots having the lowest modes in figures 3 and 4 represent female incomes (those having the highest represent male incomes).

Summarizing, we reject the null of equality in 1980 versus 2000 ($\hat{T}_n = 3.819$, $p = 0.005$), but fail to reject for 1990 versus 1995 ($\hat{T}_n = 0.326$, $p = 0.191$). Figure 4 suggests that the reason for the rejection of equality of income distributions in 1980 versus 2000 lies with a

FIGURE 3. PDF of real income, 1990 versus 1995

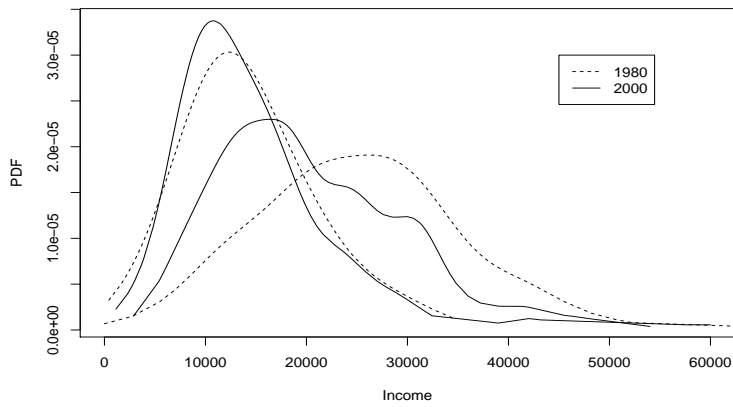
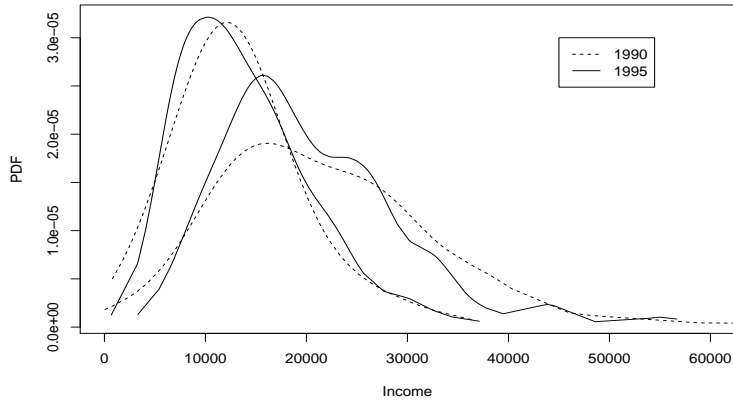


FIGURE 4. PDF of real income, 1980 versus 2000

leftward shift in the distribution of male real incomes over time. One possible explanation for this may be that we are looking at individuals who have only a high school education, and therefore tend to be employed in nonsupervisory manufacturing positions where real wages have been either constant or in decline during this time period.

5. CONCLUSION

We consider the problem of testing for equality of two 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. We advocate the use of bootstrap methods for obtaining the statistic's null distribution in finite-sample settings. Simulations show that the proposed test enjoys 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 male and female income based upon a 20-year panel underscores the novelty and flexibility of the proposed approach in mixed data settings.

Our approach can be extended to testing the equality of two unknown conditional densities, or testing the equality of two residual distributions. Hall, Racine & Li (forthcoming) have shown that the cross-validation method has the remarkable ability of potentially removing *irrelevant* conditional variables. In the testing framework this will lead to a more powerful test than a counterpart test that does not have this ability. We leave the exploration of these topics for future research.

APPENDIX A. APPENDIX

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

$$\begin{aligned}
 I_{1n} &= n_1^{-2} \sum_i K_{\gamma, x_i, x_i} + n_2^{-2} \sum_i K_{\gamma, y_i, y_i} - 2(n_1 n_2)^{-1} \sum_{i=1}^n K_{\gamma, x_i, y_i} \\
 &= (h_1 \dots h_q)^{-1} w(0)^q \left[\prod_{s=1}^r (1 - \lambda_s) \right] [n_1^{-1} + n_2^{-1}] - 2(n_1 n_2)^{-1} \sum_{i=1}^n K_{\gamma, x_i, y_i} \\
 &= c_n - 2(n_1 n_2)^{-1} \sum_{i=1}^n K_{\gamma, x_i, y_i},
 \end{aligned}$$

with $c_n = w(0)^q (h_1 \dots h_q)^{-1} \left[\prod_{s=1}^r (1 - \lambda_s) \right] [n_1^{-1} + n_2^{-1}]$, $n = \min\{n_1, n_2\}$, and

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

where $\sum_i = \sum_{i=1}^{n_1}$ if the summand has x_i , and $\sum_i = \sum_{i=1}^{n_2}$ if the summand has y_i , $\sum_{j \neq i}$ is similarly defined.

It is easy to show that $E[|I_{1n} - c_n|] = (n_1 n_2)^{-1} O(n E|K_{\gamma, x_i, y_i}|) = O(n^{-1})$. Therefore,

$$(1.9) \quad I_{1n} = c_n + O_p(n^{-1}).$$

Let $z_i = (x_i, y_i)$ and define $H_n(z_i, z_j) = K_{\gamma, x_i, x_j} + K_{\gamma, y_i, y_j} - 2K_{\gamma, x_i, y_j}$. For $i \neq j$, we have $E[H_n(z_i, z_j) | z_i] = 0$ under H_0 of $f = g$. Therefore, I_{2n} is a degenerate U-statistic. Defining $H = h_1 \dots h_q$, then it is easy to show that ($\delta_n = n_1/n_2$)

$$\begin{aligned}
 \text{var}(I_{2n}) &= E[(I_{2n})^2] \\
 &= 2 \sum_i \sum_{j \neq i} \{ n_1^{-4} E[(K_{\gamma, x_i, x_j})^2] + n_2^{-4} E[(K_{\gamma, y_i, y_j})^2] + 4(n_1 n_2)^{-2} E[(K_{\gamma, x_i, y_j})^2] + O(n^{-4}) \} \\
 &= \frac{2}{n_1 n_2} \{ \delta_n^{-1} E[(K_{\gamma, x_i, x_j})^2] + \delta_n E[(K_{\gamma, y_i, y_j})^2] + 4E[(K_{\gamma, x_i, y_j})^2] + o(1) \} \\
 &\equiv (n_1 n_2 H)^{-1} \{ \sigma_{n,0}^2 + o(1) \},
 \end{aligned}$$

where $\sigma_{n,0}^2 = H\{\delta_n^{-1}E[(K_{\gamma,x_i,x_j})^2] + \delta_n E[(K_{\gamma,y_i,y_j})^2] + 4E[(K_{\gamma,x_i,y_j})^2]\}$. It is straightforward, though tedious, to check that the conditions of Hall's (1984) central limit theorem for degenerate U-statistic holds. Thus, we have under H_0

$$(1.10) \quad (n_1 n_2 H)^{1/2} I_{2n} / \sigma_{n,0} \rightarrow N(0, 1) \text{ in distribution.}$$

It is easy to show that $\sigma_n^2 = E[\sigma_n^2] + o(1)$, and by the U-statistic H-decomposition, it follows that $\sigma_n^2 = \sigma_{n,0}^2 + o_p(1)$. Therefore, from (1.10) we obtain

$$(1.11) \quad (n_1 n_2 H)^{1/2} I_{2n} / \sigma_n \rightarrow N(0, 1) \text{ in distribution.}$$

(1.9) and (1.11) complete the proof of Theorem 2.1.

A.2. Proof of Theorem 2.2. Theorem 2.1 implies that when $h_s = h_s^0 = a_s^0 n^{-1/(q+4)}$ and $\lambda_s = \lambda_s^0 = b_s^0 n^{-2/(q+4)}$, the test statistic $\hat{T}_n(h^0, \lambda^0) \rightarrow N(0, 1)$ in distribution. Therefore, it is sufficient to prove that $\hat{T}_n(\hat{h}, \hat{\lambda}) - \hat{T}_n(h_0, \lambda_0) = o_p(1)$. For this, it suffices to show the following:

- (i) $(n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}_{2n} = (n_1 n_2 h_1^0 \dots h_q^0)^{1/2} I_{2n} + o_p(1)$,
- (ii) $(n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} [\hat{I}_{1n} - \hat{c}_n] = (n_1 n_2 h_1^0 \dots h_q^0)^{1/2} [I_{1n} - c_n] + o_p(1)$, and
- (iii) $\hat{\sigma}_n^2 = \sigma_n^2 + o_p(1)$.

Below we will only prove (i) since (ii) and (iii) are much easier to establish than (i) (and can be similarly proved).

Write $\hat{h}_s = \hat{a}_s n^{-1/(q+4)}$ and $\hat{\lambda}_s = \hat{b}_s n^{-2/(q+4)}$. From Theorem 3.1 of Li & Racine (2003), we know that $\hat{h}_s/h_s^0 - 1 \rightarrow 0$ and $\hat{\lambda}_s/\lambda_s^0 - 1 \rightarrow 0$ (in probability). This implies that $\hat{a}_s \rightarrow a_s^0$ and $\hat{b}_s \rightarrow b_s^0$ 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 $a_{1s} < a_s^0 < a_{2s}$ ($s = 1, \dots, q$) and $b_{1t} < b_t^0 < b_{2t}$ ($t = 1, \dots, r$). Let $c = (a_1, \dots, a_q, b_1, \dots, b_r)$, $c_0 = (a_1^0, \dots, a_q^0, b_1^0, \dots, b_r^0)$, and $\hat{c} = (\hat{a}_1, \dots, \hat{a}_q, \hat{b}_1, \dots, \hat{b}_r)$.

Then Lemma 1.1 shows that $A_n(c) \equiv (n_1 n_2 h_1 \dots h_q)^{1/2} I_{2n}(h, \lambda)$ (with $h_s = a_s n^{-1/(q+4)}$ and $\lambda_s = b_s n^{-2/(q+4)}$) 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 $\epsilon > 0$,

$$(1.12) \quad \lim_{n \rightarrow \infty} Pr [|B_n(\hat{c})| < \epsilon] = 1.$$

For any $\delta > 0$, denote the δ -ball centered at c_0 by $C_\delta = \{c : \|c - c_0\| \leq \delta\}$, where $\|\cdot\|$ denotes the Euclidean norm of a vector. By Lemma 1.1 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 exist a δ ($0 < \delta < 1$) and an N_1 , such that

$$(1.13) \quad Pr \left[\sup_{\|c' - c\| < \delta} |A_n(c') - A_n(c)| > \epsilon \right] < \eta_1$$

for all $n \geq N_1$

(1.13) implies that

$$(1.14) \quad Pr [|B_n(\hat{c})| > \epsilon, \hat{c} \in C_\delta] \leq Pr \left[\sup_{c \in C_\delta} |B_n(c)| > \epsilon \right] < \eta_1$$

for all $n \geq N_1$.

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

$$(1.15) \quad Pr [\hat{c} \notin C_\delta] \equiv Pr [\|\hat{c} - c_0\| > \delta] < \eta_2$$

for all $n \geq N_2$.

Therefore,

$$\begin{aligned}
Pr[|B_n(\hat{c})| > \epsilon] &= Pr[|B_n(\hat{c})| > \epsilon, \hat{c} \in C_\delta] + Pr[|B_n(\hat{c})| > \epsilon, \hat{c} \notin C_\delta] \\
(1.16) \qquad \qquad &< \eta_1 + \eta_2
\end{aligned}$$

for all $n \geq \max\{N_1, N_2\}$ by (1.14) and (1.15), 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 $P(A) \leq P(B)$).

(1.16) is equivalent to (1.12). This completes the proof of (i).

A.3. Proof of Theorem 2.3. First we can write $\hat{I}_n^* = \hat{I}_{1n}^* + \hat{I}_{2n}^*$, where \hat{I}_{jn}^* is the same as in \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 $E(\cdot | \{x_i\}_{i=1}^{n_1}, \{y_i\}_{i=1}^{n_2})$. By exactly the same arguments as we used in the proof of Theorem 2.1, one can show that $\hat{I}_{1n}^* - \hat{c}_n + O_p(n^{-1})$ by showing that $E^*|\hat{I}_{1n}^* - \hat{c}_n| = O_p(n^{-1})$ (note that $\hat{c}_n^* \equiv \hat{c}_n$). Also, one can show that $\hat{I}_{2n}^* - \hat{I}_{2n} = o_p((n^2H)^{-1/2})$ (by showing that $E^*[\hat{I}_{2n}^* - \hat{I}_{2n}]^2 = o_p((n^2H)^{-1})$ ($H = h_1 \dots h_q$), and that $\hat{\sigma}_n^{*2} - \hat{\sigma}_n^2 = o_p(1)$). Therefore, we have that,

$$(n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}_n^* / \hat{\sigma}_n^* - (n_1 n_2 \hat{h}_1 \dots \hat{h}_q)^{1/2} \hat{I}_n / \hat{\sigma}_n = o_p(1).$$

Thus, Theorem 2.3 follows from Theorem 2.1.

Lemma 1.1. Let $A_n(c) = (n_1 n_2 h_1 \dots h_q)^{1/2} I_{2n}(h, \lambda)$, where $h_s = a_s n^{-1/(q+4)}$, $\lambda_s = b_s n^{-2/(q+4)}$, $c = (a_1, \dots, a_q, b_1, \dots, b_r)$, $c_s \in [C_{1s}, C_{2s}]$ 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 $h_s = a_s n^{1/(q+4)}$ and $\lambda_s = b_s n^{-2/(q+4)}$, where $K_{c, ij} = W\left(\frac{X_j - X_i}{h}\right) L(X_j^d, X_i^d, \lambda)$, and letting $\delta = q/(4+q)$, $H^{-1/2} = (h_1 \dots h_q)^{-1/2}$, $C_1 = (a_1, \dots, a_q)'$, $C_2 = (b_1, \dots, b_r)'$, $\bar{C}_1 = \prod_{s=1}^q a_s$, and $\bar{C}_2 = \prod_{s=1}^r b_s$, then we have

$H^{-1/2}K_{c,ij} = \bar{C}_1 n^{\delta/2} W_{C_1,ij} L_{C_2,ij}$. Also, noting that $|L_{C'_2,ij} - L_{C_2,ij}| \leq \sum_{s=1}^r |b_s - b'_s| \leq r \|C_2 - C'_2\|$, we have

$$\begin{aligned}
& |(H')^{-1/2}K_{C',ij} - H^{-1/2}K_{C,ij}| = \left| n^{\delta/2} \left\{ (\bar{C}'_1)^{-1/2} W_{C'_1,ij} L_{C'_2,ij} - \bar{C}_1^{-1/2} W_{C_1,ij} L_{C_2,ij} \right\} \right| \\
& = \left| n^{\delta/2} \left\{ (\bar{C}'_1)^{-1/2} W_{C'_1,ij} [L_{C'_2,ij} - L_{C_2,ij}] + [(C'_1)^{-1/2} W_{C'_1,ij} - C_1^{-q/2} W_{C_1,ij}] L_{C_2,ij} \right\} \right| \\
(1.17) \quad & D_1 \left\{ (H')^{-1/2} W_{C'_1,ij} \|C'_2 - C_2\| + H^{-1/2} G \left(\frac{x_j - x_i}{h} \right) \|C'_1 - C_1\| \right\},
\end{aligned}$$

where $D_1 > 0$ is a finite constant. In the last equality we used $|L_{C_2,ij}| \leq 1$ and assumption (C3): also, we 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 (1.17), we have

$$\begin{aligned}
& E \{ A_n(c') - A_n(c) \}^2 \\
& = E \{ [(H')^{-1/2}K_{c',ij} - H^{-1/2}K_{c,ij}]^2 \} \\
& \leq 4E \left\{ [(H')^{-1}W^2 \left(\frac{x_j - x_i}{h'} \right) \|C'_2 - C_2\|^2 + H^{-1}G \left(\frac{x_j - x_i}{h} \right) \|C'_1 - C_1\|^2] \right\} \\
& \leq 4D_1 \left\{ \left[\int \int f(x_i) f(x_i + hv) W^2(v) dx_i dv \right] \|C'_2 - C_2\|^2 \right. \\
& \quad \left. + \left[\int \int f(x_i) f(x_i + w) G(w)^2 dx_i dw \right] \|C'_1 - C_1\|^2 \right\} \\
& \leq 4D_1 \sup_x f(x) \left\{ \left[\int W^2(v) dv \right] \|C'_2 - C_2\|^2 + \left[\int G(w)^2 dw \right] \|C'_1 - C_1\|^2 \right\} \\
(1.18) \quad & \leq D \|C' - C\|^2,
\end{aligned}$$

where D is a finite positive constant. 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).

REFERENCES

- Ahmad, I. A. & Li, Q. (1997), ‘Testing independence by nonparametric kernel method’, *Statistics and Probability Letters* **34**, 201–210.
- Ahmad, I. & van Belle, G. (1974), Measuring affinity of distributions, *in* Proschan & R. Serfling, eds, ‘Reliability and Biometry, Statistical Analysis of Life Testing’, SIAM.
- 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.
- Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley.
- 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. & Ullah, A. (1999), ‘On goodness-of-fit tests for weakly dependent processes using kernel method’, *Journal of Nonparametric Statistics* **11**, 337–360.
- 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., Racine, J. & Li, Q. (forthcoming), ‘Cross-validation and the estimation of conditional probability densities’, *Journal of the American Statistical Association* .
- Hong, Y. & White, H. (2000), ‘Asymptotic distribution theory for nonparametric entropy measures of serial dependence’, *Mimeo, Department of Economics, Cornell University, and UCSD* .
- Li, Q. (1996), ‘Nonparametric testing of closeness between two unknown distributions’, *Econometric Reviews* **15**, 261–274.
- Li, Q., Hsiao, C. & Zinn, J. (2003), ‘Consistent specification tests for semiparametric/nonparametric models based on series estimation methods’, *Journal of Econometrics* **112**, 295–325.
- 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? 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.
- Racine, J. & Maasoumi, E. (Under Revision), ‘A versatile and robust metric entropy test of time reversibility and dependence’, *Journal of Econometrics* .
- Robinson, P. M. (1991), ‘Consistent nonparametric entropy-based testing’, *Review of Economic Studies* **58**, 437–453.
- Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley.
- Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer.