Smooth Test for Equality of Distributions*

Anil K. Bera Aurobindo Ghosh Zhijie Xiao

August 03, 2010.

Keywords: Goodness-of-Fit test, Probability integral transform, Score test, Two-sample test, Empirical distribution.

*The authors like to thank the Co-Editor, two referees, Peter Phillips, Yanqing Fan and seminar participants at the Tinbergen Institute, University of Amsterdam, University of Maryland, the Far Eastern Meetings of the Econometric Society at Seoul, and European Meeting of the Econometric Society, Madrid, for helpful comments. Usual disclaimers apply. Addresses: Anil K. Bera, Dept. of Economics, University of Illinois at Urbana-Champaign, 1407 W. Gregory Drive, Urbana IL 61801 (email: abera@uiuc.edu, tel: 217-333-4596); Aurobindo Ghosh, School of Economics, Singapore Management University, 90 Stamford Road, Singapore-178903 (email: aurobindo@smu.edu.sg, tel:+65-6828-0863); and Zhijie Xiao, Dept. of Economics, Boston College, Chestnut Hill, MA 02467 (email: zhijie.xiao@bc.edu, tel: 617-552-1709).

Abstract

The two-sample version of the celebrated Pearson goodness-of-fit problem has been a topic of extensive research, and several tests like the Kolmogorov-Smirnov and Cramér-von Mises have been suggested. Although these tests perform fairly well as omnibus tests for comparing two probability density functions (PDFs), they may have poor power against specific departures like in location, scale, skewness and kurtosis. We propose a new test for the equality of two PDFs based on a modified version of the Neyman smooth test using empirical distribution functions minimizing size distortion in finite samples. The suggest test can detect the specific directions of departure from the null hypothesis. Specifically, it can identify deviations in the directions of mean, variance, skewness or tail behavior. We derive the smooth test using Rao's score principle. In finite sample, the actual probability of type-I error depends on the relative sizes of the two samples. We proposed two different approaches to deal with this problem and show that, under appropriate conditions, the proposed tests are asymptotically distributed as chi-squared. We also study the finite sample size and power properties of our proposed test. As an application of our procedure, we compare the age distributions of employees with small employers in New York and Pennsylvania with group insurance before and after the enactment of the "Community Rating" legislation in New York. It has been a conventional wisdom that if community rating is enforced (where group health insurance premium does not depend on age or any other physical characteristics of the insured), then the insurance market will collapse since only older or less healthy patients would prefer group insurance. We find that there are significant changes in the age distribution in the population in New York owing mainly to a shift in location and scale.

1 INTRODUCTION

One of the old, celebrated problems in statistics is the *two-sample* version of Pearson (1900) goodness-of-fit problem [Lehmann (1953) and Darling (1957)]. Suppose we have two samples $X_1, X_2, ..., X_n$ and $Y_1, Y_2, ..., Y_m$ from two *unspecified* absolutely continuous distributions with cumulative distribution functions (CDFs), $F(x)$ and $G(x)$, respectively. The problem is to test the hypothesis $H_0 : F = G$. Most of the tests in the literature are based on some distance measures between the two empirical distribution functions (EDFs), $F_n(x)$ and $G_m(x)$. For example, the Kolmogorov-Smirnov criterion uses [see, for instance, Darling (1957, p. 828)]

$$D_{nm} = \sqrt{\frac{nm}{n+m}} \sup_{-\infty<x<\infty} |F_n(x) - G_m(x)| . \tag{1}$$

The Cramér- von Mises statistic is based on the measure [see, for example, Anderson (1962, p. 1148)]

$$W^2_{nm} = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 \, dH_{n+m}(x) , \tag{2}$$

where $H_{n+m}(x)$ is the EDF of the two samples together, i.e., $H_{n+m}(x) = [nF_n(x) + mG_m(x)]/(m+n)$. Anderson and Darling (1952) modification of (2) is given by [see Darling (1957, p. 827)]

$$A^2_{nm} = \frac{nm}{n+m} \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 \, \psi(H_{n+m}(x)) \, dH_{n+m}. \tag{3}$$

Here $\psi(.)$ is some non-negative weight function chosen to accentuate the distance between $F_n(x)$ and $G_m(x)$ where the test is desired to have sensitivity. A statistically appealing weight function $\psi(u) = [u(1-u)]^{-1}$ has the effect of weighing the tails heavily since the function is large near $u = 0$ and $u = 1$ [Anderson and Darling (1954, p. 767)]. However, the computation and implementation of these tests are not trivial, and being *omnibus* tests, they lack power in *specific* directions (Janssen, 2000). Moreover, they do not have good finite sample power properties.

We propose a test for H_0 using Neyman's (1937) smooth test principle. To motivate the proposed test, let us first consider the *one-sample* Pearson goodness-

of-fit test of $H_0^{'}: F(x) = F_0(x)$ where $F_0(x)$ is a *specified* CDF with $f_0(x)$ as the corresponding probability density function (PDF). Define the probability integral transform (PIT)

$$Z_i = F_0(X_i) = \int_{-\infty}^{X_i} f_0(\omega)\, d\omega, i = 1, 2, ..., n. \tag{4}$$

If $H_0^{'}$ is true, then $Z_1, Z_2, ..., Z_n$ are independently and identically distributed (IID) as uniform random variates $U(0,1)$ irrespective of F_0. Test of $H_0^{'}$ is identical to the test of uniformity of Z in $(0,1)$. Therefore, in some sense, "all" testing problems can be converted into testing only *one kind of hypothesis* [see Neyman (1937, pp. 160-162) and Bera and Ghosh (2002, p. 178)]. Neyman (1937) considered the following *smooth* alternative to the uniform density

$$h(z) = c(\theta) \exp\left[\sum_{j=1}^{k} \theta_j \pi_j(z)\right],\ 0 < z < 1, \tag{5}$$

where $c(\theta)$ is the constant of integration depending on $\theta_1, \theta_2, ..., \theta_k$, and $\pi_j(z)$ are orthogonal polynomials of order j satisfying

$$\int_0^1 \pi_i(z)\, \pi_j(z)\, dz = \delta_{ij} \text{ where } \delta_{ij} = \begin{cases} = 1, \text{ if } i = j \\ = 0, \text{ if } i \neq j. \end{cases} \tag{6}$$

Neyman called (5) a smooth alternative since with small θs, $h(z)$ is close to, and has few intersections with the density under the null hypothesis $U(0,1)$. Neyman (1937, pp. 163-164) used $\pi_j(z)$ as the normalized Legendre polynomials $\pi_j(z) = a_{j0} + a_{j1}z + ... + a_{jj}z^j, a_{jj} \neq 0$, satisfying the orthogonality conditions in (6). Explicitly for $k = 1, 2, 3, 4$, these are: $\pi_1(z) = \sqrt{12}\left(z - \frac{1}{2}\right)$, $\pi_2(z) = \sqrt{5}\left(6\left(z - \frac{1}{2}\right)^2 - \frac{1}{2}\right)$, $\pi_3(z) = \sqrt{7}\left(20\left(z - \frac{1}{2}\right)^3 - 3\left(z - \frac{1}{2}\right)\right)$, $\pi_4(z) = 210\left(z - \frac{1}{2}\right)^4 - 45\left(z - \frac{1}{2}\right)^2 + \frac{9}{8}$. We can, therefore, test $H_0^{'}$ by testing $H_0^{''}: \theta_1 = \theta_2 = ... = \theta_k = 0$ in (5). Using the generalized Neyman-Pearson (N-P) lemma, Neyman (1937) derived the locally most powerful symmetric test for $H_0^{''}$ against the alternative H_1 : At least one $\theta_j \neq 0$, for

some j. Under H_0, asymptotically the test statistic

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2 \ \sim \chi_k^2, \text{ where } u_j = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \pi_j \left(Z_i\right), j = 1, ..., k. \tag{7}$$

Neyman suggested this test to rectify some of the drawbacks of Pearson (1900) goodness-of-fit statistic [see Bera and Ghosh (2002) for more on this and for a historical perspective].

Now turning to the problem of testing $H_0 : F = G$ in the two sample case, let us start assuming that $F(.)$ is known. We construct a new random variable $Z = F(Y)$. The CDF of Z is given by

$$H(z) = \Pr(Z \leq z) = \Pr\left(Y \leq F^{-1}(z)\right) = G\left(Q(z)\right), \tag{8}$$

where $Q(z) = F^{-1}(z)$ is the quantile function of Z. Therefore, the PDF of Z can be written as [see Neyman (1937, p. 161), Pearson (1938, p. 138)]

$$h(z) = \frac{d}{dz} H(z) = \frac{g\left(Q(z)\right)}{f\left(Q(z)\right)}, \ 0 < z < 1. \tag{9}$$

Although this is a ratio of two PDFs, $h(z)$ is a proper density function in the sense that $h(z) \geq 0$, $z \in (0,1)$ and $\int_0^1 h(z)\, dz = 1$, if we assume that F and G are also strictly increasing functions. In the literature $h(.)$ is known under different names. Ćwik and Mielniczuk (1989) termed it the *relative density,* while Parzen (1992, p. 7) and Handcock and Morris (1999, p. 22) called it the *comparison density function* and the *density ratio,* respectively. We will call it *ratio density function* (RDF) since it is both a ratio of two densities and a proper density itself. Under $H_0 : F = G$, $h(z) = 1$, i.e., $Z \sim U(0,1)$. Under the alternative hypothesis $H_1 : F \neq G$, $h(z)$ will differ from 1, providing a basis for the Neyman smooth test. Under the alternative, we take $h(z)$ as given in (5) and test $\theta_1 = \theta_2 = ... = \theta_k = 0$. Therefore, the test utilizes (9) which looks more like a "likelihood (density) ratio". To see the exact form of $h(z)$, let us consider some particular cases. When the two distributions differ only in location; for example, $f(.) \equiv \mathcal{N}(0,1)$ and $g(.) \equiv N(\mu, 1)$ $(\mu \neq 0)$, $\ln(h(z)) = \mu z - \frac{1}{2}\sigma^2$ which is *linear* in z. Similarly, if the distributions differ only in scale parameter, such as,

$f(.) \equiv \mathcal{N}(0,1)$ and $g(.) \equiv \mathcal{N}(0,\sigma^2)$, $\sigma^2 \neq 1$, $\ln(h(z)) = \frac{z^2}{2}\left[1 - \frac{1}{\sigma^2}\right] - \frac{1}{2}\ln\sigma^2$, a *quadratic* function of z. As plotted in Figure 1, the first and second-order normalized polynomials $\pi_1(z)$ and $\pi_2(z)$ can capture this type of differences in location and scale aspects of the distributions. In Figure 2, we provide plots of $h(z)$ when $f(.)$ and $g(.)$, differ in skewness and kurtosis terms, respectively. These plots resemble closely with the plots of the third and the fourth normalized Legendre polynomials $\pi_3(z)$ and $\pi_4(z)$ in Figure 3. Therefore, we believe that the test will not only be powerful but also will be informative in identifying particular source(s) of departure(s) from H_0.

Our approach of testing $H_0 : F = G$ is related to those based on the distance functions between F and G mentioned earlier. Under Neyman's smooth test formulation, we consider the distance between $h(z)$ and 1, i.e., $\frac{d}{dz}\left[G\left(F^{-1}(z)\right) - z\right]$, $0 < z < 1$. $G\left(F^{-1}(z)\right) - z$ is a familiar quantity in the literature of tests based on differences of EDFs, such as those in (1)-(3). Due to the equality [see Serfling (1980, pp. 110-111)]

$$\sup_{-\infty<x<\infty} |F(x) - G(x)| = \sup_{0<z<1} \left|G\left(F^{-1}(z)\right) - z\right|,$$

we can transform the domain of our distance calculation from $(-\infty, \infty)$ to $(0,1)$. Moreover, $\sqrt{m}\left[G_m\left(F^{-1}(z)\right) - z\right]$ converges to the well-known Brownian bridge process in distribution [see, for instance, Billingsley (1968, p. 104)]. Neyman's approach also makes the test procedure more informative by looking at the distance of $h(z)$ from 1 towards particular direction(s) specified by (5).

The rest of the paper is organized as follows. In Section 2, we derive Neyman's smooth test for the one-sample problem using Rao's (1948) score principle for testing $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$. In Section 3, we propose a test statistic for equality of distributions in two samples and derive its asymptotic distribution. Optimality criteria for the finite sample implementation of the test is discussed in Section 3.1. The general case with relaxed condition on the relative sample sizes is studied in Section 4. In Section 5, we apply the proposed test for comparing two age distributions in New York and Pennsylvania small group insurance markets. Finite sample behavior of the test and sample size selection are investigated in Section 6 through Monte

Carlo experiments. Finally, the paper is concluded in Section 7 with a summary of our results and indications of future research.

2 NEYMAN SMOOTH TEST

For simplicity we first consider the test when $F(x)$ is known, and we use the PIT as defined by $Z_i = F(Y_i)$. Neyman (1937, pp. 166-180) derived a locally most powerful symmetric unbiased test for $H_0 : \theta_1 = \theta_2 = ... = \theta_k = 0$ in (5) against $H_1 : \theta_j \neq 0$ at least for one j, and he called it an unbiased critical region of type-C. This type-C critical region is an extension of the locally most powerful unbiased (LMPU) test (type-A region) of Neyman and Pearson (1936) from a single parameter case to a multiparameter situation. We denote the power function as $\beta(\theta_1, \theta_2, ..., \theta_k) = \beta(\theta) \equiv \beta$. Assuming that the power power function $\beta(\theta)$ is twice differentiable in the neighborhood of $H_0 : \theta = 0$, an unbiased critical region of type-C of size α is obtained by maximizing $\frac{\partial^2 \beta}{\partial \theta_j^2}\Big|_{\theta=0}$ subject to

$$\beta(\mathbf{0}) = \beta(0, 0, ...0) = \alpha. \tag{10}$$

$$\beta_j = \frac{\partial \beta}{\partial \theta_j}\bigg|_{\theta=0} = 0, \ j = 1, 2, ..., k. \tag{11}$$

$$\beta_{jl} = \frac{\partial^2 \beta}{\partial \theta_j \partial \theta_l}\bigg|_{\theta=0} = 0, \ j, l = 1, 2, ..., k, j \neq l. \tag{12}$$

$$\beta_{jj} = \frac{\partial^2 \beta}{\partial \theta_j^2}\bigg|_{\theta=0} = \frac{\partial^2 \beta}{\partial \theta_1^2}\bigg|_{\theta=0} = \beta_{11}, \ j = 2, 3, ..., k. \tag{13}$$

It is clear that conditions (10) and (11) are, respectively, for the size and unbiasedness. Conditions (12) and (13) ensures that equal departures from $\theta = 0$, in *all* directions should lead to the same power. Therefore, for this type-C critical region, the approximate power function is $\beta(\theta) = \alpha + \frac{1}{2}\beta_{11} \sum_{j=1}^{k} \theta_j^2$. After some algebra and using multiparameter version of the generalized Neyman-Pearson lemma, Neyman obtained his optimal test statistic. The resulting statistic, however, takes a very simple form as given in Proposition 1.

Proposition 1 (Neyman, 1937) *The type-C critical region is given by*

$$\Psi_k^2 = \sum_{j=1}^{k} u_j^2 \geq C_\alpha, \tag{14}$$

where $u_j = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j (z_i)$, *and for large* m *the critical point* C_α *is determined from* $\Pr[\chi_k^2 \geq C_\alpha] = \alpha$.

We now show that the test statistic Ψ_k^2 can simply be obtained using the Rao (1948) score (RS) test principle. Taking (5) as the PDF under the alternative hypothesis, the log-likelihood function $l(\theta)$ can be written as

$$l(\theta) = m \ln c(\theta) + \sum_{j=1}^{k} \theta_j \sum_{i=1}^{m} \pi_j (z_i). \tag{15}$$

The RS test for testing the null $H_0 : \theta = \theta_0$ is given by

$$RS = s(\theta_0)' \mathcal{I}(\theta_0)^{-1} s(\theta_0), \tag{16}$$

where $s(\theta)$ is the score vector $\partial l(\theta) / \partial \theta$, and $\mathcal{I}(\theta)$ is the information matrix $E\left[-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'}\right]$. It is straightforward to see that

$$s(\theta_j) = \frac{\partial l(\theta)}{\partial \theta_j} = m \frac{\partial \ln c(\theta)}{\partial \theta_j} + \sum_{i=1}^{m} \pi_j (z_i) = m \frac{\partial \ln c(\theta)}{\partial \theta_j} + \sqrt{m} u_j, \ j = 1, .., k. \tag{17}$$

Differentiating the identity $\int_0^1 h(z)\, dz = 1$ with respect to θ_j, we have

$$\frac{\partial c(\theta)}{\partial \theta_j} \int_0^1 \exp\left[\sum_{j=1}^{k} \theta_j \pi_j (z)\right] dz + c(\theta) \int_0^1 \exp\left[\sum_{j=1}^{k} \theta_j \pi_j (z)\right] \pi_j (z)\, dz = 0. \tag{18}$$

Evaluating (18) at $\theta = \mathbf{0}$, we have $\left.\frac{\partial \ln c(\theta)}{\partial \theta_j}\right|_{\theta=0} = 0$. Thus, under the null hypothesis

$$s(\theta_j) = \sqrt{m} u_j. \tag{19}$$

To get the information matrix, note from (17) that

$$\frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_l} = m\frac{\partial^2 \ln c(\theta)}{\partial\theta_j\partial\theta_l}, \tag{20}$$

is a constant. Therefore, under H_0 the (j,l)th element of the information matrix $\mathcal{I}(\theta)$ is simply $-m\partial^2 c(\theta)/\partial\theta_j\partial_l$ evaluated at $\theta = 0$. Differentiating (18) with respect to θ_l and evaluating it at $\theta = 0$, after some simplification, we have

$$\left.\frac{\partial^2 c(\theta)}{\partial\theta_j\partial\theta_l}\right|_{\theta=0} + \int_0^1 \pi_j(z)\,\pi_l(z)\,dz = 0. \tag{21}$$

Using (6)

$$\left.\frac{\partial^2 c(\theta)}{\partial\theta_j\partial\theta_l}\right|_{\theta=0} = -\delta_{jl}, \tag{22}$$

and

$$\mathcal{I}(\theta_0) = mI_k, \tag{23}$$

where I_k is a $k \times k$ identity matrix. Combining (16), (20) and (23) the RS test statistic has the simple form

$$RS = \sum_{j=1}^{k} u_j^2. \tag{24}$$

If we write $c(\theta) = \exp[\psi(\theta)]$ in (5), the density function can be expressed as

$$h(z) = \exp\left[\psi(\theta) + \sum_{j=1}^{k}\theta_j\pi_j(z)\right]. \tag{25}$$

This choice of the alternative can be recast into the well-known *maximum entropy (ME)* characterization. Suppose, we have k moment restrictions on the random variable Z given by $E[\pi_j(z)] = c_j$, $j = 1, 2, ..., k$. Then, among all possible distributions, the density function that maximizes the entropy $-E[\ln h(z)]$ has the form (25), where θ_j is the Lagrange multiplier corresponding to the j^{th} moment restriction in the ME optimization, $j = 1, 2, ..., k$ [see, for example, Kagan, Linnik, Rao and Ramachandran (1973, p. 409)]. In this sense, we seek power towards the most "plausible" density function that satisfies the above k moment conditions and maximizes the entropy measure.

The components that Neyman considered in his test is connected to an ANOVA like decomposition of the distribution. If we consider the classical Cramér- von Mises test for $H_0 : F_X(.)$ is uniformly distributed against $H_1 : F_X(.)$ is not uniformly distributed, the test statistic is $CvM = n\int_0^1 \left[\hat{F}_X(y) - y\right]^2 dy$, or equivalently, if we test that $H_0 : F_X(.) = F_0(.)$ against $H_1 : F_X(.) \neq F_0(.)$, we consider $n\int_0^1 \left[\hat{F}_X(.) - F_0(.)\right]^2 dF_0(.)$, where $\hat{F}_X(.)$ is the empirical distribution function of X. If we consider the Fourier transforms: $\theta_{2j-1} = \int_0^1 \cos(2\pi jx)\, dF(x)$, $\theta_{2j} = \int_0^1 \sin(2\pi jx)\, dF(x)$, $j = 1, 2, ...$, then the testing problem is equivalent to testing $H_0 : \theta_j = 0$, $j = 1, 2, ...$ against H_1 : at least one of the $\theta_j \neq 0$. If we let $\hat{\theta}_j$ be the empirical Fourier coefficients, then the CvM test can be expressed as

$$CvM = \frac{n}{2\pi^2} \sum_{j=1}^{\infty} j^{-2} \left(\hat{\theta}_{2j-1}^2 + \hat{\theta}_{2j}^2\right). \tag{26}$$

From the representation in (26), we can see that the classical CvM test (and also K-S) put increasingly smaller weights (j^{-2}) to high frequency components (i.e., put a weight j^{-2} to $e^{i2\pi jx}$ with large j). These weights (j^{-2}) make the CvM type test effectively use the first few components of θ_j. In Neyman's smooth test, instead of putting a decreasing weight, the selected components are equally weighted. This allows us to focus on the selected directions of departure from the null distribution, and this can make the smooth test superior particularly when we are trying to identify specific directions of departure and not just an overall departure [see, for example, Durbin and Knott (1972), Eubank and LaRiccia (1992) for studies on this issue]. Durbin (1972) and Durbin and Knott (1972) also suggested that only the first few $\theta_j s$ should be examined which is essentially the spirit of Neyman's smooth test with a small value of k in the specification of the density in (5) under the alternative hypothesis.

3 TESTING EQUALITY OF DISTRIBUTIONS

In this section we consider two-samples of n and m observations $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ from *unspecified* absolutely continuous distributions with CDFs $F(x)$ and $G(x)$, and test the hypothesis $H_0 : F = G$. Without loss of generality, we assume that $m \leq n$.

Without knowledge about the distribution function F, $Z_i = F(Y_i)$ is unknown and thus Ψ_k^2 in (14) is not feasible in practice. We, therefore, consider using the empirical distribution function (EDF)

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x), \tag{27}$$

in place of F, to construct

$$\hat{Z}_i = F_n(Y_i) = \frac{1}{n}\sum_{l=1}^{n} I(X_l \leq Y_i), \ i = 1, ..., m, \tag{28}$$

where $I(.)$ is an indicator function. Substituting Z_i by $\hat{Z}_i$ in (14), we obtain the following generalized version of the statistic

$$\hat{\Psi}_k^2 = \sum_{j=1}^{k}\left[\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)\right]^2. \tag{29}$$

As noted earlier, the infeasible Ψ_k^2 converges to a χ^2-distribution with degree of freedom k and has all the local optimality properties of Rao's score test. We now show that under certain conditions, $\hat{\Psi}_k^2$ has the same limiting distribution. To achieve that goal, we need to demonstrate that the departure due to replacing $F(.)$ by the EDF $F_n(.)$ is asymptotically negligible. For convenience of asymptotic analysis, we make the following assumptions.

ASSUMPTION 1: $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$ are *independent and identically distributed* with CDFs $F(x)$ and $G(x)$ respectively, and are independent of each other.

ASSUMPTION 2: $\mathrm{E}\left(\pi_j(Z_i)^2\right) < \infty$, for $j = 1, \cdots, k$.

We state the result in Theorem 1, proof of which is given in the Appendix.

Theorem 1 *Under Assumptions 1 and 2 and the null hypothesis that $F = G$, if $\frac{(\log\log n)m}{n} \to 0$ as $m, n \to \infty$, $\hat{\Psi}_k^2 \Rightarrow \chi_k^2$.*

From Theorem 1 we can see that the size of the sample used in estimating the distribution function F should be larger in magnitude than the test sample size. If we estimate the CDF of X based on observations $\{X_i\}_{i=1}^n$, F_n converges to F at

the rate $\sqrt{n}$ (pointwise), i.e., the estimation error in $\hat{Z}_i$ is of the order $n^{-\frac{1}{2}}$. Thus, as $\pi_j(.)'s$ can be shown to be first-order Lipschitz continuous, the accumulated estimation error in $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)$ has the order $\sqrt{m}/\sqrt{n}$, which goes to zero if n increases to ∞ at a rate faster than m. Thus to obtain consistent tests, we require that the sample size n used in estimating the empirical distribution should be larger than m, the test sample size. Incidentally, similar problem arises in various statistical procedures. For example, in simulation based inference [e.g., Gourieroux and Monfort (1996)] the conditional moments are estimated based on simulations. If the number of simulations increases fast enough relative to the sample size, the resulting estimator has the same asymptotic distribution as the estimator based on known conditional moments.

3.1 Relative Magnitude of Two Sample Sizes

Theorem 1 provides an upper bound for m given n so that $\hat{\Psi}_k^2$ is asymptotically equivalent to Ψ_k^2. Under this condition, a wide range of sample sizes can be chosen and all provide asymptotically equivalent tests, although the finite sample performance of these tests may differ substantially. A natural question is: what is the optimal rate of m relative to n that minimizes the order of size distortion?

Theorem 2 *Under the null hypothesis,*

$$\hat{\Psi}_k^2 = \chi_k^2 + O_p\left(\frac{1}{\sqrt{m}}\right) + O_p\left(\sqrt{\frac{m}{n}}\right),$$

and thus the optimal relative magnitude of m and n that minimizes the size distortion is $m = O\left(n^{\frac{1}{2}}\right)$.

Heuristically, we can decompose $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)$ into

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\pi_j\left(\hat{Z}_i\right) - \pi_j\left(Z_i\right)\right]. \tag{30}$$

The first part of (30), $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right)$ converges to a standard normal variate. The larger the m, the faster it goes to the limiting normal distribution. The second part of

(30), $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left[\pi_j\left(\hat{Z}_i\right)-\pi_j\left(Z_i\right)\right]$, which is the error term coming from estimating $F\left(.\right)$, converges to zero under the conditions in Theorem 1. The larger the n relative to m, the smaller the term. To optimize the sampling properties of the test, a trade-off has to be made to balance these two components, giving an optimal relative magnitude between m and n.

An exact analytical formula of the optimal relative sample sizes will be dependent on the exact formulation of the higher order terms, and to the best of our knowledge, cannot be obtained. Theorem 2 gives the optimal relative magnitude between m and n, which substantially narrows down the range of choices for the sample sizes. To minimize the distortion coming from estimating the distribution function, we prefer n to be large relative to m. On the other hand, to obtain fast convergence to χ^2 in the limit, we want a large m. Thus, a trade-off has to be made to minimize size distortion. We balance these two terms so that they are of the same order of magnitude, giving the optimal relative magnitude as $m = O(\sqrt{n})$. Monte Carlo experiment results reported in Section 6 indicate that a simple rule of thumb, such as, $m = \sqrt{n}$ based on Theorem 2 provides satisfactory results.

Alternatively, since we know that $\{X_l\}_{l=1}^{n}$ are from the same distribution with CDF F. We may divide the index set $\mathcal{N} = \{1, \ldots, n\}$ into two mutually exclusive and exhaustive sets $\mathcal{N}_1$ and $\mathcal{N}_2$ with cardinalities n_1 and n_2, with $n_1 + n_2 = n$, and define the **training set**

$$\mathcal{X}_1 = \{(X_j), j \in \mathcal{N}_1\}$$

and the testing set

$$\mathcal{X}_2 = \{(X_j), j \in \mathcal{N}_2\}.$$

Then estimate $F(\cdot)$ using data $\mathcal{X}_1$ and construct

$$F_{n_1}\left(X_i\right) = \frac{1}{n_1}\sum_{j\in\mathcal{N}_1} I\left(X_j \leq X_i\right), \text{ for } i \in \mathcal{N}_2.$$

Notice that $\mathcal{X}_1$ and $\mathcal{X}_2$ are from the same distribution with CDF F, $F\left(X_i\right)$ $(i \in \mathcal{N}_2)$ are uniformly distributed, and $F_{n_1}\left(X_i\right)$ provides an estimator for the uniform distribution. Hence, we can calculate the estimated $\hat{\Psi}_k^2$ based on the EDF $F_{n_1}\left(X_i\right)$,

$i = 1, 2, ..., n_2$, that should have an asymptotic χ^2 distribution with k degrees of freedom. Our objective is to minimize some distance between the empirical distribution function of $\hat{\Psi}_k^2$ and the χ_k^2 CDF in finite samples, i.e., it is equivalent to the solution for either x or t

$$\arg\min_{n_2} d_{n_2}^r(\hat{F}_{\hat{\Psi}_k^2}(x), F_{\chi_k^2}(x)) \equiv \arg\min_{n_2} d_{n_2}^r(F_{\chi_k^2}\left(\hat{F}_{\hat{\Psi}_k^2}^{-1}(t)\right), t).$$

If we take the criterion function $d_{n_2}^r(.)$ to be the Anderson-Darling statistic over r replications, the optimal sample size for the testing sample n_2 is a solution to

$$\arg\min_{n_2} -r + \frac{1}{r}\sum_{i=1}^{r}(2i-1)\left[\log\left(p_{(i)}\right) + \log\left(1 - p_{(r-i+1)}\right)\right], \tag{31}$$

where $p_{(i)} = F_{\chi_k^2}\left(\hat{F}_{\hat{\Psi}_k^2}^{-1}\left(\hat{\Psi}_{(i)}^2\right)\right)$ and $\hat{\Psi}_{(1)}^2 \leq \hat{\Psi}_{(2)}^2 \leq ... \leq \hat{\Psi}_{(r)}^2$.

For each value of n_2, we can calculate the above criterion function. We choose n_2 that minimizes the above criterion. Finally we choose the test sample size $m = (n_2/n_1) \times n$.

4 THE GENERAL CASE

The test $\hat{\Psi}_k^2$ that we proposed in the previous section is an asymptotic test and, by appropriately choosing the relative sample sizes (so that one sample is smaller than the other), has an asymptotic χ_k^2 distribution under the null hypothesis. The test is asymptotically distribution-free (ADF) and thus remains asymptotically equivalent to the original one-sample version smooth test Ψ_k^2. In some applications, we may have two samples with similar sample sizes, i.e. the case where m and n are of the same magnitude, say $m = \lambda n$, where $\lambda \in (0, 1]$. In this case, the distribution of $\hat{\Psi}_k^2$ will be affected by the estimation error of $F(\cdot)$, and it will no longer be asymptotically χ_k^2 distributed. We now discuss the appropriate modification needed to restore the original ADF (χ_k^2) property.

For convenience of asymptotic analysis, we slightly modify Assumption 2 as follows:

ASSUMPTION 2′: $\mathrm{E}\left(\pi_j\left(Z_i\right)^{2+\epsilon}\right) < \infty$, for some $\epsilon > 0$, $j = 1, \cdots, k$.

Since the basic ingredients in constructing $\hat{\Psi}_k^2$ are $\frac{1}{\sqrt{m}} \sum_{i=1}^m \pi_j\left(\hat{Z}_i\right)$, $j = 1, \cdots, k$, we first derive the joint distribution of $\frac{1}{\sqrt{m}} \sum_{i=1}^m \pi_j\left(\hat{Z}_i\right)$, $j = 1, \cdots, k$. Let

$$\Pi_k = \left(\frac{1}{\sqrt{m}} \sum_{i=1}^m \pi_1\left(\hat{Z}_i\right), \cdots\cdots, \frac{1}{\sqrt{m}} \sum_{i=1}^m \pi_k\left(\hat{Z}_i\right) \right)',$$

The limiting behavior of Π_k is summarized in the following Theorem.

.

Theorem 3 *Under Assumptions 1 and 2′ and H_0:$F = G$, if $m = \lambda n$, Π_k converges to a k-dimensional normal variate with covariance matrix*

$$\Omega_k = I_k + \lambda \Delta_k, \tag{32}$$

where I_k is the k-dimensional identity matrix and Δ_k is a $k \times k$ matrix whose (j, r)-th element is

$$\delta_{jr} = E\left[\dot{\pi}_j\left(Z_1\right) \dot{\pi}_r\left(Z_2\right) \left\{F\left(Y_1 \wedge Y_2\right) - F\left(Y_1\right) F\left(Y_2\right)\right\}\right],\ j = 1, \cdots, k;\ r = 1, \cdots, k,$$

with $\dot{\pi}_j\left(.\right)$ denoting the first derivative of $\pi_j\left(.\right)$ and $Y_i \wedge Y_j = \min\left(Y_i, Y_j\right)$.

Result in Theorem 3 shows that when the two sample sizes are of the same order of magnitude, the limiting distribution of Π_k is different from its infeasible counterpart when the true CDF is used. Consequently, the limiting distribution of $\hat{\Psi}_k^2$ is no longer distribution free, and thus can not be directly used as a test statistic in our inference problem. However, Theorem 3 indicates that the limiting variate of Π_k is still multivariate normal and an asymptotic χ^2-test can be constructed after appropriate standardization.

It is also instructive to note that the change in the asymptotic covariance of Π_k; from I_k [see, equation (23)], it changes to $I_k + \lambda \Delta_k$ when estimated CDF is used. However, given the result in Theorem 3, we can construct a valid test by replacing

Δ_k by its consistent estimator. Letting

$$\widehat{\delta}_{jr} = \frac{1}{nm^2} \sum_{l=1}^{n} \sum_{i=1}^{m} \sum_{t=1}^{m} \left[\dot{\pi}_j \left(\widehat{Z}_i \right) \dot{\pi}_r \left(\widehat{Z}_t \right) \left[I \left(X_l \leq Y_i \wedge Y_t \right) - \widehat{Z}_i \widehat{Z}_t \right] \right]$$

and

$$\widehat{\Delta}_k = \begin{bmatrix} \widehat{\delta}_{11} & \cdots & \widehat{\delta}_{1k} \\ \cdots & \ddots & \\ \widehat{\delta}_{k1} & \cdots & \widehat{\delta}_{kk} \end{bmatrix},$$

we may estimate the covariance matrix Ω_k by

$$\widehat{\Omega}_k = I_k + \frac{m}{n} \widehat{\Delta}_k.$$

The following modified testing statistic

$$\widetilde{\Psi}_k^2 = \Pi_k' \widehat{\Omega}_k^{-1} \Pi_k \tag{33}$$

is asymptotically distributed as χ_k^2 and can be used when m and n are of the same order.

Under the sample size condition of Theorem 1, $\widehat{\Omega}_k \to I_k$, the modified testing statistic $\widetilde{\Psi}_k^2$ is asymptotically equivalent to $\hat{\Psi}_k^2$ and the limiting result still holds in that case. Thus we may treat $\widetilde{\Psi}_k^2$ as a generalized version of the smooth test in Section 2, which includes the previous test as a special case.

Remark: An alternative approach to deal with this preliminary estimation error is to use bootstrap or sub-sampling method to construct a valid test. To do so, we need to construct a resampled test under the restriction of null hypothesis. Let $n^* + m^* = n$, $m^*/n^* = m/n$, and consider resampling n^* and m^* observations from $\{X_l\}_{l=1}^n$ and denote them, respectively, as $\{X_i^*\}_{i=1}^{n^*}$, and $\left\{Y_j^*\right\}_{j=1}^{m^*}$. Let us define

$$F_{n^*}^* (x) = \frac{1}{n^*} \sum_{i=1}^{n^*} I \left(X_i^* \leq x \right), \hat{Z}_i^* = F_{n^*}^* (Y_i^*), \ i = 1, ..., m^*.$$

Then we can construct

$$\hat{\Psi}_k^{*2} = \sum_{j=1}^{k} \left[\frac{1}{\sqrt{m^*}} \sum_{i=1}^{m^*} \pi_j \left(\hat{Z}_i^* \right) \right]^2 . \tag{34}$$

The limiting null distribution of the test statistic can then be approximated by repeating the above steps many times. The above resampling-based approximation is asymptotically valid if the limit of the conditional distribution of $\hat{\Psi}_k^{*2}$ is the same as that of $\hat{\Psi}_k^2$. We will study the empirical distribution of $\hat{\Psi}_k^{*2}$ in Section 6. This bootstrapped version of the smooth test is in similar spirit as developed in Mora and Neumeyer (2005) for the two-sample test for regression errors based on empirical processes. However, we are looking at the unconditional distributions of the two-samples, and are not assuming any models generating the regression residuals, hence we are avoiding complications due to model selection and parameter estimation.

5 AN APPLICATION TO COMPARING TWO AGE DISTRIBUTIONS

We consider an insurance market where there are several insurance companies providing health insurance competing for clients. These clients can be grouped into different risk categories depending on their "proneness" or "propensity" of having bad health. However, the main problem that the insurance companies face is one of *adverse selection* since they cannot see beforehand what type of client they are insuring, high risk or low risk (Akerlof, 1970). Rothschild and Stiglitz (1976) considered an insurance market setup where there are two types of clients, high and low risk. They claimed that a health insurance contract based purely on risk categories will ensure that the high risk client chooses to pay higher premium. Moreover, both high and low risk clients will buy full (or complete) insurance coverage. However, if it is not possible to write a health insurance contract based on the risk categories due to either legislation or other restrictions then there could be two possible scenarios. First, a healthier (low risk) individual will chose to buy less than complete coverage while the less healthy (high risk) individuals will buy full insurance, so the

insurance market will still function although this will not be as efficient as risk based contracts. The second scenario happens if further regulations restrict or prohibit the selling of less than full insurance. In this case, the healthy or low risk individuals will stop buying coverage which means that gradually the insured population will be made up of more high risk individuals. Thus, the insurance company has to pay out more often. This will cause the premium to go up, so healthier individuals will drop their coverage even further and this cycle will continue. Finally, this will result in a total collapse of the insurance market. This scenario is referred to as the "*Adverse Selection Death Spiral*" [Buchmueller and DiNardo (2002)].

To find the evidence of any existence of "Adverse Selection Death Spiral", the state of New York where legislation for enforcing "community rating" (premium fixed by community and not by risk category) was enacted in 1993 can be compared with Pennsylvania, where no such legislations was enacted. We would like to test for the difference between the age distributions of the adult civilian population between 18 and 64, before and after 1993. The data is from 1987-1996 March Current Population Survey covering questions on whether an individual have insurance coverage, and if so whether the coverage is through a small employer and other similar questions. New York and Pennsylvania were selected due to their similarities both geographically and demographically.

Our objective is to use the smooth test to determine if there is a difference between the age distributions in a state before and after 1993. The population selected was individuals who were covered by group insurance policies sponsored by their employers who had 100 or less employees. This was divided into two parts, one for the individuals before 1993 and the other for those after 1993. Figures 4 and 5 give the estimated PDFs using kernel density estimator. We estimated the PDF $f(x)$ of the sample $x_1, x_2, ..., x_n$ using kernel density estimator $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x-x_i}{b}\right)$, where the bandwidth is *quadruple* of $b = 1.06\min(\hat{\sigma}_x, IQR/1.34)n^{-\frac{1}{5}}$, $\hat{\sigma}_x$ is the estimated standard deviation of $x_i's$ and IQR is the interquartile range for the sample and $K(.)$ is the kernel suggested by Parzen (1962) [also see, Silverman (1986) pp. 45-47].

Table 1 presents the values of some of the commonly used test statistics discussed in Section 1 based on the EDF for the age distributions of group insured by small

employers before and after 1993 in New York and Pennsylvania along with the 0.1% critical values of all the modified statistics [Stephens (1970)]. Except for the test on D^+ [which considers the positive part of (1)], all the tests clearly indicate that the two distributions are different. However, these test statistics are not informative about the nature of the deviation from the equality of two distributions.

Suppose now, that the sample $y_1, y_2, ..., y_m$ comes from a population after 1993. The probability integral transforms (PIT) of each y_i is based on

$$\hat{z}_i = F_n(y_i), \ i = 1, 2, ..., m. \tag{35}$$

If the age distributions before and after 1993 are the same then $\hat{z}_1, \hat{z}_2, ..., \hat{z}_m$ should be approximately distributed as $U(0,1)$. In see Figure 6 we present the histogram of the New York data and note the departure from uniformity.

To elucidate more information from the data, we use the smooth test and its components based on the statistic $\hat{\Psi}_k^2$ in (29) with $k = 4$. Table 2 provides the smooth test results. For New York, we have $\hat{\Psi}_k^2 = 117.5011$ that is highly significant at any standard significance level. Therefore, we may conclude that the age distributions before and after 1993 are quite different. Now considering the four (asymptotically independent) components of $\hat{\Psi}_k^2$, each of which $\left(\hat{u}_j^2\right)$ is approximately distributed as χ_1^2, we note that the hypothesis of equality of two distributions are most possibly rejected due to the differences in the first (location) and second (scale) order moments. The constributions of $\hat{u}_3^2$ and $\hat{u}_4^2$ are indeed very small. As we discussed in Section 3.1 there could be size distortion in "uncritical" use of $\hat{\Psi}_4^2$ with asymptotic χ_4^2 critical values. In Table 5 (discussed in details in the next section 6) we provide "exact" critical values using $\hat{\Psi}_k^{*2}$ in (34). It is observed that the value of $\hat{\Psi}_k^2 = 117.5011$ is still highly significant. Another way to rectify the size distortion of $\hat{\Psi}_4^2$ is to exploit the result in Theorem 2, and use a "smaller" test sample size m. Again as we will discuss in Section 6, our simulation study shows that (see Figure 7) optimal value of m is around 500. Therefore, we repeat the calculations of $\hat{\Psi}_4^2$ and its components with $n = 4598$ but $m = 500$ (randomly drawn). The test statistics and the corresponding p-values are reported in Table 3. The qualitative result, i.e., the significance of $\hat{\Psi}_4^2$ and the two components $\hat{u}_1^2$ and $\hat{u}_2^2$, remain unchanged. Yet another way to to

handle size distortion when m and n are of the same order of magnitude (without subsampling) is to use the modified statistic $\widetilde{\Psi}_4^2$ in (33). Its value for New York is 83.37 which is also highly significant. Therefore, we come to the same conclusion using various approaches. The results for the Pennsylvania data are also similar (see Table 2). Thus the population age distributions before and after 1993 in both the states are indeed different irrespective of the "Community Rating" legislation and the sources of the differences are mainly through location and scale.

6 MONTE CARLO RESULT

To study the finite sample properties of the smooth test we generate the data under the null hypothesis of equality of distributions by subsampling from the population of group insured civilian adults before 1993 in New York as used in our application in Section 5. In Figure 8, we plot the empirical size of the test as a function of m when the nominal level is 5%. The estimation sample size n is fixed at 5000, also drawn at random from the same population. We evaluate the size of the test for $m = 50$ to 3500 with an increment of 10 based on 2000 replications. There is a high "relative" variation in the empirical size; however, we observe an increasing trend of the test size with m, though it never crosses the 10% level. Furthermore, there are several values of m for which the actual sizes are close to the nominal size of 5%.

In order to fully explore the finite sample performance of the suggested tests, in addition to studying closeness of the actual and nominal sizes, we also evaluated the goodness of the overall χ^2 approximations (χ_1^2 or χ_4^2, depending on the test statistics) of the distributions under the null hypothesis [see David (1939, 1947)]. Using the values of the test statistics from 10000 replications with $n = 2500$ and $m = 500$ in Figures 9 and 10, we plot the estimated kernel densities and the corresponding theoretical χ^2 densities. In the panels of Figure 9, it is hard to distinguish between the theoretical χ_1^2 and the empirical densities of $\hat{u}_j^2$ ($j = 1, 2, 3, 4$), the components of $\hat{\Psi}_4^2$. From Figure 10, we note that the estimated kernel density of $\hat{\Psi}_4^2$ is not too far off from that of the theoretical χ_4^2. This is a worst-case scenario, since here $m = 500$ is substantially high for $n = 2500$.

Although a good overall match of the empirical distribution of a test statistic to

theoretical χ^2 is desirable, having the true size of the test close to the nominal level (say, 5%) is very important. In Table 4, we present the actual size when $n = 2500$ and $m = 500$ and 50, with 10000 replications. Theorem 2 suggests that for minimum size distortion m need to be in the order of $\sqrt{n}$. From that point of view, the actual sizes when $m = 500$ are pretty good (only for $\hat{u}_4^2$, the size is 0.083 is much higher than 5%); qualitatively, we could have inferred this by looking at the tail parts of the empirical distributions in Figures 9 and 10. For $m = 50$, the results are indeed very good, the actual sizes never cross the nominal 5% level.

Subsampling from the New York data used in the empirical application is convenient to study the finite sample size properties of the test. However, to study finite sample power-size trade-off in a systematic way we need a different framework. We generate the data using the following mixture of Gaussian and log-normal densities:

$$X = B * X_1 + (1 - B) * X_2. \tag{36}$$

Observations under the null hypothesis are generated taking $B \sim Bernoulli(0.3)$, $\ln X_1 \sim N(-1, 4)$ and $X_2 \sim N(1.2, 1.21)$, and under the alternative $B \sim Bernoulli(0.5)$, $\ln X_1 \sim N(-0.1, 1)$ and $X_2 \sim N(1.75, 0.81)$. To consider a small sample size take $n = 625$ and increase the value of m in increments of 5 upto $m = 250$, and perform 200 replications. The estimated size and power of the test are plotted in Figure 11. Upto $m = 50$, the size is around 0.05 and after that it tends to increase. The estimated power of the test using $\hat{\Psi}_4^2$ increases dramatically with m; for example, when $m = 50$, the power is around 0.85 (with empirical size close to 0.05).

Next we take $n = 2000$, and increase the value of m with increments of 10, upto $m = 2000$ (so that m and n are of the same order of magnitude), and perform 2000 replications. Figure 12 provides the size, (raw) power and sizer-adjusted power of $\hat{\Psi}_4^2$. The test has very good power properties even at small sample size, say $m = 100$. The size adjusted power, though decreasing with m, is also quite good. Of course, given these results our recommendation would be not to use $\hat{\Psi}_4^2$ mechanically when m and n are of similar order due to significant size distortion. When $n = 2000$ and $m = 1000$ the empirical size of $\hat{\Psi}_4^2$ is close to 20%; however, for this case if we use our modified test $\widetilde{\Psi}_4^2$ the size reduces to close to 5%. To get an idea of the finite

sample distribution of $\widetilde{\Psi}_4^2$ (relative to $\hat{\Psi}_4^2$ and χ_4^2) under the null hypothesis we plot these three densities in Figure 13. It is clear that the kernel density function of $\widetilde{\Psi}_4^2$ almost overlaps with that of χ_4^2, and while the density of $\hat{\Psi}_4^2$ has a much thicker tail requiring much larger critical values for valid inference. To see the impact of m on the modified test statistic $\widetilde{\Psi}_4^2$, we repeat the above experiment (of Figure 13) but with $n = m = 2000$, and plot the three densities in Figure 14. Now although the estimated kernel density of $\widetilde{\Psi}_4^2$ does not overlap with that of χ_4^2 completely, the deviation is not all that substantial. The departure of the empirical density of $\hat{\Psi}_4^2$ from χ_4^2 is more pronounced compared to that in Figure 13.

The empirical densities of $\hat{\Psi}_4^2$ and $\widetilde{\Psi}_4^2$ under the alternative hypothesis are plotted in Figures 15 and 16, respectively for $n = 2000, m = 1000$ (i.e., $\lambda = 0.5$) and $n = m = 1000$.(i.e. $\lambda = 1.0$) along with the χ_4^2 density. The distributions of the unmodified $\hat{\Psi}_4^2$ lie further to the right to those of the modified counterparts, reflecting the higher rejection rates for $\hat{\Psi}_4^2$ due to its much larger size.

We also implement the resample based statistic $\hat{\Psi}_4^{*2}$ in (34) when n and m are of similar order of magnitude, taking $n = m = 2000$ and performing 5000 replications. The simulated density under the null hypothesis alomg with that of χ_4^2 are plotted in Figure 17.These two densities are quite different. Based on numerical values of $\hat{\Psi}_4^{*2}$ we find the empirical critical values and report those in Table 5. These values are almost twice those for the values of the standard χ_4^2.

7 SUMMARY AND FUTURE RESEARCH

We propose a smooth test for comparing the distributions in a two sample setup. Unlike traditional *omnibus* goodness-of-fit tests such as the Kolmogorov-Smirnov or Cramér-von Mises procedures, the smooth test helps us identify the sources of departure from the null hypothesis of equality of two densities. We derived the smooth test from Rao's score principle, and this it enjoys the optimality properties of the score test (Bera and Bilias, 2001). We have also investigated the choices of the relative sizes of the estimation and test samples to minimize size distortion. Our Monte Carlo results reveal that the smooth test (with proper choice of the sample size) and its modified form have good finite sample properties in terms of size, power

and closeness to the theoretical χ^2 distributions.

There are several directions of future research that we would like to pursue.

Following Neyman (1937) we have kept the value of k fixed at 4. In some applications, however, it may be desirable to choose k as an increasing function of the sample size using some model selection criterion, such as the Schwarz's criterion (see for example, Ledwina, 1994; Kallenberg, Oosterhoff, Schriever, 1985, Inglot, Kallenberg and Ledwina, 1994). If k is too high, the effectiveness of the test in each direction could be diluted as pointed out by Neyman (1937) (also see the discussion in Bera and Ghosh, 2002, p.205). Furthermore, the accumulated stochastic errors (discussed in Section 6) in each direction would be large and that can deteriorate the performance of the test. Janic-Wróblewska and Ledwina (2000) proposed a data-driven version of the smooth test, where k can take a maximum value $d(m) = o\left(\{m/\log m\}^{1/9}\right)$, which increases with m at a very slow rate. It would be quite interesting to combine our methodology with that of Janic-Wróblewska and Ledwina (2000).

As suggested by one of the referees, another very useful extension would be to consider non-orthogonal moment functions such as, $E\left(X^j - Y^j\right),\ j = 1, 2, ...$ and follow the GMM-type formulation of Newey (1985). With this approach, though we lose local optimality property of Rao score principle (see, for example, Bera and Ghosh (2002), and references therein) and identification of the directions of departure from the null hypothesis, we could avoid the problem of choosing of m and use the complete samples.

We can extend our test similar to what Fan (1996) proposed adaptively using both Neyman's technique and a wavelet based procedure for comparing global and local departures from the null hypothesis (also see Fan and Huang, 2001). Finally, the test can be modified to allow for possible dependence in the data, particularly in the context of time series or panel data.

APPENDIX: PROOFS

Proof of Theorem 1

We first give two Lemmas that will be used in our proof.

Lemma 1 *The normalzed Legendre polynomials are given by*

$$\pi_j(z) = \frac{\sqrt{2j+1}}{j!}\frac{d^j}{dz^j}\left(z^2 - z\right)^j,\ 0 \le z \le 1, j = 1, 2, 3... \tag{A.1}$$

Proof. The standard Legendre polynomials [see, for instance, Kendall and Stuart (1973, p. 460)] are

$$P_j(x) = \frac{1}{2^j j!}\frac{d^j}{dx^j}\left(x^2 - 1\right)^j, -1 \le x \le 1, j = 1, 2, 3... \tag{A.2}$$

where

$$\int_{-1}^{1} P_j(x) P_l(x)\, dx = \begin{cases} 0, & \text{if } j \neq l, \\ \frac{2}{2j+1}, & \text{if } j = l. \end{cases} \tag{A.3}$$

Now, changing variable to $z = \frac{x+1}{2}$,or $x = 2z - 1$,

$$\begin{aligned} \tilde{P}_j(z) &= \frac{1}{2^j j!} \times \frac{1}{2^j}\frac{d^j}{dz^j}\left\{2z-1)^2 - 1\right\}^j \\ &= \frac{1}{2^{2j} j!}\frac{d^j}{dz^j}\left\{4z^2 - 4z\right\}^j \\ &= \frac{1}{j!}\frac{d^j}{dz^j}\left\{z^2 - z\right\}^j. \end{aligned}$$

Since $\int_0^1 \left(\tilde{P}_j(z)\right)^2 dz = \int_{-1}^1 2^{-1}\left(P_j(x)\right)^2 dx = (2j+1)^{-1}$ (see equation A.3), normalizing we have

$$\pi_j(z) = \frac{\sqrt{2j+1}}{j!}\frac{d^j}{dz^j}\left(z^2 - z\right)^j,\ 0 \le z \le 1, j = 1, 2, ... \tag{A.4}$$

■

Lemma 2 *If $\pi_j(.)$ is the normalized Legendre polynomial of degree j defined on $[0,1]$, then the first order Lipschitz condition holds, that is*

$$\left|\pi_j(\hat{z}) - \pi_j(z)\right| \le M\left|\hat{z} - z\right|, \tag{A.5}$$

where M is a positive constant, and z and $\hat{z}$ are any two points between 0 and 1.

Proof. Using the mean value thoerem on $\pi_j(z)$ in (A.1), we have

$$\begin{aligned}\pi_j(\hat{z}) - \pi_j(z) &= \frac{\sqrt{2j+1}}{j!}(\hat{z}-z)\frac{d}{dz}\pi_j(z^*) \\ &= \frac{\sqrt{2j+1}}{j!}(\hat{z}-z)\frac{d^{j+1}}{dz^{j+1}}\left(z^2-z\right)^j\Big|_{z=z^*}, \end{aligned} \tag{A.6}$$

where z^* is such that $|z^*-z| < |\hat{z}-z|$.

Expanding the polynomial $(z^2-z)^j$, and differentiating with respect to z upto order $j+1$,

$$\begin{aligned}\frac{d}{dz}\pi_j(z) &= \frac{\sqrt{2j+1}}{j!}\frac{d^{j+1}}{dz^{j+1}}\left(z^2-z\right)^j \\ &= \frac{d^{j+1}}{dz^{j+1}}\sum_{l=0}^{j}\left(z^2\right)^l(z)^{j-l}(-1)^{j-l} \\ &= \frac{\sqrt{2j+1}}{j!}\sum_{l=1}^{j}\frac{(l+j)!}{(l-1)!}\frac{j!}{l!\,(l-j)!}(-1)^{j-l}z^{l-1} \\ &\leq \frac{\sqrt{2j+1}}{j!}\sum_{l=1}^{j}l\frac{(l+j)!}{(l)!}\frac{j!}{l!\,(l-j)!}\left|(-1)^{j-l}\right|\left|z^{l-1}\right| \\ &\leq \frac{j(j+1)}{2}\sqrt{2j+1}\sum_{l=1}^{j}\frac{(l+j)!}{(l)!l!\,(l-j)!} \\ &= \frac{j(j+1)}{2}\sqrt{2j+1}\,{}_2F_1(1-j,2+j;2;-1), \end{aligned} \tag{A.7}$$

where the finite Hypergeometric function ${}_2F_1(.)$ is defined by

$${}_2F_1(a,b;c;x) = 1 + \frac{ab}{c}x + \frac{a(a+1)b(b+1)}{c(c+1)2!}x^2 + ... \tag{A.8}$$

Hence, using (A.3) and (A.7),

$$\begin{aligned}&|\pi_l(\hat{z}) - \pi_l(z)| \\ &= |\hat{z}-z|\left|\frac{d}{dz}\pi_j(z)\Big|_{z=z^*}\right| \\ &\leq |\hat{z}-z|\frac{j(j+1)}{2}\sqrt{2j+1}\,|{}_2F_1(1-j,2+j;2;-1)| \\ &= M_j|\hat{z}-z|, \end{aligned} \tag{A.9}$$

where M_j is a finite positive number for any finite positive interger j. ■

Proof of Theorem 1

Proof. We need to show that $\widehat{\Psi}_k^2 - \Psi_k^2 = o_p(1)$. Notice that

$$\begin{aligned}
\widehat{\Psi}_k^2 &= \sum_{j=1}^{k} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(\widehat{Z}_i) \right]^2 \\
&= \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \pi_j(Z_i) + \sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2 \\
&= \Psi_k^2 + R_{1,m,n} + R_{2,m,n},
\end{aligned}$$

where

$$\Psi_k^2 = \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \pi_j(Z_i) \right]^2, \; R_{1,m,n} = \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2,$$
$$R_{2,m,n} = 2 \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \pi_j(Z_i) \right] \left[\sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right].$$

We show that, as $m, n \to \infty$, and $(\log \log n)m/n \to 0$, $R_{1,m,n} = o_p(1)$, $R_{2,m,n} = o_p(1)$. We first consider $R_{1,m,n}$,

$$R_{1,m,n} = \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2 \leq \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right]^2.$$

By Lemma 2, $|\pi_j(u) - \pi_j(v)| \leq M |u - v|$, for some finite number $M > 0$. Thus

$$\begin{aligned}
R_{1,m,n} &\leq \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right]^2 \leq M^2 \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left| \widehat{Z}_i - Z_i \right| \right]^2 \\
&\leq M^2 \sum_{j=1}^{k} \frac{1}{m} \left[m \max_{1 \leq i \leq m} \left| \widehat{Z}_i - Z_i \right| \right]^2.
\end{aligned}$$

Recalling $\hat{Z}_i = F_n(Y_i)$ and $Z_i = F(Y_i)$, and noting that $\sup_u |F_n(u) - F(u)|$ is maximally of order $\sqrt{(\log \log n)/n}$ as $n \to \infty$ [see, e.g., Shorack and Wellner (1986)], we have

$$\max_{1 \leq i \leq m} \left| \widehat{Z}_i - Z_i \right| = O_p(\sqrt{(\log \log n)/n}),$$

and

$$R_{1,m,n} \leq mM^2 \sum_{j=1}^{k} \left[\max_{1 \leq i \leq m} \left| \widehat{Z}_i - Z_i \right| \right]^2 = O_p((\log \log n)m/n),$$

which converges to 0 if $(\log \log n)m/n \to 0$ as $m, n \to \infty$.

Similarly,

$$\begin{aligned} |R_{2,m,n}| &= 2 \sum_{j=1}^{k} \frac{1}{m} \left| \sum_{i=1}^{m} \pi_j(Z_i) \right| \left| \sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right| \\ &\leq 2 \sum_{j=1}^{k} \frac{1}{m} \left| \sum_{i=1}^{m} \pi_j(Z_i) \right| \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right]. \end{aligned}$$

Note that $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(Z_i)$ converges to a standard normal variable and

$$\sum_{i=1}^{m} \pi_j(Z_i) = O_p(m^{1/2}).$$

Again using Lemma 2,

$$\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \leq M \sum_{i=1}^{m} \left| \widehat{Z}_i - Z_i \right| \leq Mm \max_{1 \leq i \leq m} \left| \widehat{Z}_i - Z_i \right| = O_p(m\sqrt{(\log \log n)/n}).$$

Thus

$$\begin{aligned} |R_{2,m,n}| &\leq 2 \sum_{j=1}^{k} \frac{1}{m} \left| \sum_{i=1}^{m} \pi_j(Z_i) \right| \left[\sum_{i=1}^{m} \left| \pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right| \right] \\ &= O_p\left(\frac{1}{m} \times m^{1/2} \times m\sqrt{(\log \log n)/n} \right) \\ &= O_p(\sqrt{(\log \log n)m/n}), \end{aligned}$$

converges to 0 when $(\log \log n)m/n \to 0$ as $m, n \to \infty$. ■

Proof of Theorem 2

Proof. From the proof of Theorem 1 we know that the asymptotic χ^2 test statistic $\widehat{\Psi}_k^2$ can be decomposed into $\Psi_k^2 + R_{1,m,n} + R_{2,m,n}$. We analyze each of these terms to show how the relative magnitudes of m and n affect the size of the test.

The leading term is

$$\Psi_k^2 = \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \pi_j(Z_i) \right]^2 .$$

By construction, conditional on X, for each j, $V_{ji} = \pi_j(Z_i)$ $(i = 1, ..., m)$ are m independent and identically distributed random variables with mean zero and unit variance. Thus

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} V_{ji} \Rightarrow N(0,1) \equiv \xi_j,$$

where ξ_j $(j = 1, ..., k)$ are k independent standard normal variates. In addition, assuming that V_{ji} possesses moments up to the forth order, by standard result of Edgeworth expansion [see, e.g., Rothenberg (1984) and references therein], the probability density function of $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} V_{ji}$ has the following expansion

$$f(x) \approx \varphi(x) \left[1 + \frac{k_3 H_3(x)}{6\sqrt{m}} + \frac{3k_4 H_4(x) + k_3^2 H_6(x)}{72m} \right],$$

where $\varphi(.)$ is the density of standard normal, k_r is the r-th cumulant, and H_r is the Hermite polynomial of degree r defined as $H_r(x) = (-1)^r \varphi^{(r)}(x)/\varphi(x)$. Thus, $m^{-1/2} \sum_{i=1}^{m} V_{ji}$ can be expanded into a leading term of standard normal variable ξ_j plus a second order term, say $\frac{1}{\sqrt{m}} A_j$, of $O_p(m^{-1/2})$ and a third term, $\frac{1}{m} B_j$, of order $O_p(m^{-1})$, i.e.,

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} V_{ji} \approx \xi_j + \frac{1}{\sqrt{m}} A_j + \frac{1}{m} B_j.$$

Therefore, Ψ_k^2 can be expanded as

$$\sum_{j=1}^{k} \xi_j^2 + \frac{1}{\sqrt{m}} A + \frac{1}{m} B + o_p(\frac{1}{m}),$$

where the leading term $\sum_{j=1}^{k} \xi_j^2$ is the χ^2 random variable with k degree of freedom and the second term $\frac{1}{\sqrt{m}} A$ is of order $O_p(m^{-1/2})$. To obtain a good approximation of the χ^2 distribution, a large m is preferred.

Now we turn to the estimation of the distribution function. We have

$$R_{1,m,n} = \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2$$
$$= \frac{m}{n} \sum_{j=1}^{k} \left[\frac{1}{m} \sum_{i=1}^{m} \sqrt{n} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2 = \frac{m}{n} C.$$

Notice that $U_{ji} = \sqrt{n} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] = O_p(1)$, and $\frac{1}{m} \sum_{i=1}^{m} U_{ji} = O_p(1)$, and

$$C = \sum_{j=1}^{k} \left[\frac{1}{m} \sum_{i=1}^{m} \sqrt{n} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]^2 = O_p(1),$$

and thus $R_{1,m,n} = O_p\left(\frac{m}{n}\right)$. Similarly,

$$R_{2,m,n} = 2 \sum_{j=1}^{k} \frac{1}{m} \left[\sum_{i=1}^{m} \pi_j(Z_i) \right] \left[\sum_{i=1}^{m} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]$$
$$= 2 \sqrt{\frac{m}{n}} \sum_{j=1}^{k} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(Z_i) \right] \left[\frac{1}{m} \sum_{i=1}^{m} \sqrt{n} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right]$$
$$= O_p\left(\sqrt{\frac{m}{n}} \right)$$

since

$$D = 2 \sum_{j=1}^{k} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j(Z_i) \right] \left[\frac{1}{m} \sum_{i=1}^{m} \sqrt{n} \left[\pi_j(\widehat{Z}_i) - \pi_j(Z_i) \right] \right] = O_p(1).$$

Therefore,

$$\widehat{\Psi}_k^2 = \Psi_k^2 + R_{1,m,n} + R_{2,m,n}$$
$$= \sum_{j=1}^{k} \xi_j^2 + \frac{1}{\sqrt{m}} A + \sqrt{\frac{m}{n}} D + o_p(\frac{1}{\sqrt{m}} + \sqrt{\frac{m}{n}}),$$

where $\frac{1}{\sqrt{m}} A$, $\sqrt{\frac{m}{n}} D$, etc., are higher order terms that brings size distortion in finite sample, but are $o_p(1)$. In particular, the leading terms are of order $O_p\left(\frac{1}{\sqrt{m}}\right)$ and $O_p\left(\sqrt{\frac{m}{n}}\right)$, respectively. ■

Proof of Theorem 3

Proof. We first analyze the limiting behavior of $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)$, $j=1,\cdots k$. Notice that

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(\hat{Z}_i\right)=\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right)+\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\dot{\pi}_j\left(Z_i\right)\left[\hat{Z}_i-Z_i\right]+\underbrace{\mathcal{R}_{m,n}}_{\sqrt{m}/n}$$

where the remainder term $\mathcal{R}_{m,n}=O_p(\sqrt{m}/n)=o_p(1)$, and the leading term $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\pi_j\left(Z_i\right)$ follows a simple CLT, thus we focus on the behavior of $\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\dot{\pi}_j\left(Z_i\right)\left[\hat{Z}_i-Z_i\right]$. Notice that when $m=\lambda n=m_n$, $\lambda>0$, we may re-write

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\dot{\pi}_j\left(Z_i\right)\left[\hat{Z}_i-Z_i\right]=\sum_{l=1}^{n}\frac{1}{n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l\leq Y_i\right)-F(Y_i)\right)\right)=\sum_{l=1}^{n}\xi_{nl},$$

where $\xi_{nl}=\frac{1}{n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l\leq Y_i\right)-F(Y_i)\right)\right)$, $l=1,\cdots,n$, are uncorrelated over l. Let $\mathcal{F}_l=\sigma\left\{X_\ell,\ell\leq l;\,Y_i,i=1,\cdot\cdot,m_n\right\}$, then ξ_{nl} is $\mathcal{F}_l$ measurable and $\mathrm{E}(\xi_{nl}|\mathcal{F}_{l-1})=0$. Thus $\sum_{l=1}^{n}\xi_{nl}$ is a summation of uncorrelated arrays, and its limiting behavior can be analyzed using the central limit theorem method developed for triangular arrays.

Notice that

$$s_n^2=\mathrm{E}\left(\sum_{l=1}^{n}\xi_{nl}\right)^2=\sum_{l=1}^{n}\mathrm{E}\xi_{nl}^2\to\lambda\delta_{jj},\text{ as }n\to\infty,$$

where

$$\delta_{jj}=\mathrm{E}\left[\dot{\pi}_j\left(Z_i\right)\dot{\pi}_j\left(Z_s\right)\left\{F\left(Y_i\wedge Y_s\right)-F\left(Y_i\right)F\left(Y_s\right)\right\}\right].$$

We re-standardize ξ_{nl} by s_n, and verify that

$$\max_{1\leq l\leq n}\left|s_n^{-1}\xi_{nl}\right|\xrightarrow{P}0.$$

Notice that

$$\Pr\left(\max_{1\leq l\leq n}\left|s_n^{-1}\xi_{nl}\right|>\epsilon\right)\leq\frac{\sum_{1\leq l\leq n}\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^2 1(|s_n^{-1}\xi_{nl}|>\epsilon)}{\epsilon^2},$$

and by Hölder's inequality,

$$\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^2 1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon) \leq \left(\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^{2p}\right)^{1/p} \left(\mathrm{E}\left|1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon)\right|^q\right)^{1/q},$$

where $\frac{1}{p} + \frac{1}{q} = 1$. We look at $\left(\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^{2p}\right)^{1/p}$ and $\left(\mathrm{E}\left|1(|s_n^{-1}\xi_{nl}| > \epsilon)\right|^q\right)^{1/q}$. Notice that

$$s_n^{-1}\xi_{nl} = \frac{1}{ns_n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right).$$

By the Minkowski's inequality,

$$\begin{aligned}
&\left[\mathrm{E}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right)^{2p}\right]^{1/2p} \\
&\leq \sum_{i=1}^{m_n}\left[\mathrm{E}\left(\frac{1}{\sqrt{m_n}}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right)^{2p}\right]^{1/2p} \\
&= m_n^{1/2}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right\|_{2p}.
\end{aligned}$$

Thus

$$\begin{aligned}
\left(\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^{2p}\right)^{1/p} &= \left(\mathrm{E}\left|\frac{1}{ns_n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right)\right|^{2p}\right)^{1/p} \\
&\leq \frac{1}{\left(ns_n\right)^2}\left(m^{1/2}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right\|_{2p}\right)^2 \\
&= \frac{\lambda}{ns_n^2}\left(\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right\|_{2p}\right)^2 \\
&\approx \frac{1}{n\delta_{jj}}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \leq Y_i\right) - F(Y_i)\right)\right\|_{2p}^2
\end{aligned}$$

and, by the Minkowski's inequality again,

$$
\begin{aligned}
&\mathrm{E}\left|1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon)\right|^q \\
&= \mathrm{E}\left|1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon)\right| \\
&= \Pr\left(\left|\frac{1}{ns_n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right)\right| > \epsilon\right) \\
&\le \frac{1}{\epsilon^b}\mathrm{E}\left|\frac{1}{ns_n}\left(\frac{1}{\sqrt{m_n}}\sum_{i=1}^{m_n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right)\right|^b \\
&\le \frac{1}{\epsilon^b}\frac{1}{n^b s_n^b m_n^{b/2}}\left(\sum_{i=1}^{m_n}\left(\mathrm{E}\left|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right|^b\right)^{1/b}\right)^b \\
&= \frac{\lambda^{b/2}}{n^{b/2}s_n^b\epsilon^b}\left(\mathrm{E}\left\|\dot{\pi}_j\left(Z_1\right)\left(I\left(X_l \le Y_1\right) - F(Y_1)\right)\right\|_b\right)^b \\
&\approx \frac{1}{n^{b/2}\delta_{jj}^{b/2}\epsilon^b}\left(\mathrm{E}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_b\right)^b.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^2 1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon) \\
&\le \left(\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^{2p}\right)^{1/p}\left(\mathrm{E}\left|1(\left|s_n^{-1}\xi_{nl}\right| > \epsilon)\right|^q\right)^{1/q} \\
&\le \frac{\lambda}{ns_n^2}\left(\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_{2p}\right)^2 \frac{\lambda^{b/2q}}{n^{b/2q}s_n^{b/q}\epsilon^{b/q}}\left(\mathrm{E}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_b\right)^{b/q} \\
&= \frac{\lambda^{1+b/2q}}{n^{1+b/2q}s_n^{2+b/q}\epsilon^{b/q}}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_{2p}^2\left(\mathrm{E}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_b\right)^{b/q}.
\end{aligned}
$$

Under Assumptions 1 and $2'$, by appropriately choosen p, q, and b,

$$
\begin{aligned}
&\Pr\left(\max_{1\le l\le n}\left|s_n^{-1}\xi_{nl}\right| > \epsilon\right) \\
&\le \frac{\sum_{1\le l\le n}\mathrm{E}\left|s_n^{-1}\xi_{nl}\right|^2 1(|s_n^{-1}\xi_{nl}| > \epsilon)}{\epsilon^2} \\
&\le \frac{1}{\epsilon^{2+b/q}}\frac{\lambda^{1+b/2q}}{n^{b/2q}s_n^{2+b/q}}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_{2p}^2\left(\mathrm{E}\left\|\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l \le Y_i\right) - F(Y_i)\right)\right\|_b\right)^{b/q}
\end{aligned}
$$

$\to \infty$, as $n \to \infty$.

By analysis of McLeish (1974),

$$\sum_{l=1}^{n} s_n^{-1} \xi_{nl} \rightarrow N(0,1).$$

Thus,

$$\left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \dot{\pi}_j \left(Z_i \right) \left[\hat{Z}_i - Z_i \right] \right) \rightarrow N(0, \lambda \delta_{jj}).$$

Notice that $\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j \left(Z_i \right)$ and $\frac{1}{n} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \sum_{l=1}^{n} \dot{\pi}_j \left(Z_i \right) \left(I \left(X_l \leq Y_i \right) - Z_i \right)$ are uncorrelated, we have

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j \left(\hat{Z}_i \right) \Rightarrow N(0, 1 + \lambda \delta_{jj}). \tag{A.10}$$

Thus, by result of (A.10) and application of the Cramer-Wold device,

$$\Pi_k = \left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_1 \left(\hat{Z}_i \right), \cdots\cdots, \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_k \left(\hat{Z}_i \right) \right)$$

converges to a k-diemsional normal variate with covariance matrix $\Omega_k = I_k + \lambda \Delta_k$, where

$$\Delta_k = \begin{bmatrix} \delta_{11} & \cdots & \delta_{1k} \\ \cdots & \ddots & \cdots \\ \delta_{k1} & \cdots & \delta_{kk} \end{bmatrix}$$

since, when $j \neq r$,

$$E \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j \left(Z_i \right) \right] \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_r \left(Z_i \right) \right] = 0 \text{ because } \pi_j \left(Z_i \right) \text{ and } \pi_r \left(Z_i \right) \text{ are orthogonal ,}$$

$$\mathrm{E} \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_j \left(Z_i \right) \right] \left[\frac{1}{n} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \sum_{l=1}^{n} \dot{\pi}_r \left(Z_i \right) \left(I \left(X_l \leq Y_i \right) - Z_i \right) \right] = 0,$$

$$\mathrm{E} \left[\frac{1}{n} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \sum_{l=1}^{n} \dot{\pi}_j \left(Z_i \right) \left(I \left(X_l \leq Y_i \right) - Z_i \right) \right] \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} \pi_r \left(Z_i \right) \right] = 0,$$

and

$$\mathrm{E}\left[\frac{1}{n}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\sum_{l=1}^{n}\dot{\pi}_j\left(Z_i\right)\left(I\left(X_l\leq Y_i\right)-Z_i\right)\right]\left[\frac{1}{n}\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\sum_{l=1}^{n}\dot{\pi}_r\left(Z_i\right)\left(I\left(X_l\leq Y_i\right)-Z_i\right)\right]\rightarrow\lambda\delta_{jr}.$$

■

REFERENCES

Anderson, T. W. (1962), "On the Distribution of the two-sample Cramer-von Mises Criterion," *Annals of Mathematical Statistics*, 33, 1148-1159.

Anderson, T. W. and Darling, D.A. (1952), "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, 23, 193-212.

— (1954), "A test of goodness of fit,". *Journal of the American Statistical Association*, 49, 765-769.

Bera, A. K. and Ghosh, A. (2002), "Neyman's smooth test and its applications in Econometrics." In: *Handbook of Applied Econometrics and Statistical Inference*, Eds. A. Ullah, A. Wan and A. Chaturvedi, Marcel Dekker, pp. 177-230.

Billingsley, P. (1968), *Convergence in Probability Measures*, John Wiley: New York.

Buchmueller, T. and DiNardo, J. (2002), "Did community rating induce an adverse selection death spiral? Evidence from New York, Pennsylvania, and Connecticut," *American Economic Review*, 92, 280-294.

Ćwik, J and Mielniczuk, J. (1989), "Estimating density ratio with application to discriminant analysis," *Communications in Statistics: Theory and Methods*, 18, 3057-3069.

Darling, D.A. (1955), "The Cramér-Smirnov test in the parametric case," *Annals of Mathematical Statistics*, 26, 1-20.

— (1957), "The Kolmogorov-Smirnov, Cramér-von Mises tests," *Annals of Mathematical Statistics*, 28, 823-838.

David, F. N. (1939), "On Neyman's "Smooth" Test for Goodness of Fit: I. Distribution of the Criterion #2 when Hypothesis Tested is True," *Biometrika,* 31, 191-199.

— (1947), "A 'Smooth' test for goodness of fit,"*Biometrika,* 4, 299-310.

Durbin, J. (1973), "Distribution theory for tests based on the sample distribution function," In *Regional Conference Series in Applied Mathematics*, *Vol. 9*, SIAM: Philadelphia, PA.

Durbin, J. and Knott, M. (1972), "Components of Cramer-von Mises statistics. I," *Journal of the Royal Statistical Society. Series B (Methodological)*, 34, 290-307.

Eubank, R.L. and LaRiccia, V.N. (1992), "Asymptotic comparison of Cramér-von Mises and nonparametric function estimation techniques for testing goodness-of-fit," *Annals of Statistics*, 20, 1412-1425.

Fan, J. (1996), "Test of significance based on wavelet thresholding and Neyman's truncation," *Journal of the American Statistical Association,* 91, 674-688.

Fan, J. and Huang, L-S. (2001), "Goodness-of-fit tests for parametric regression models," *Journal of the American Statistical Association*, 96, 640-652.

Gourieroux, C. and Monfort, A. (1996), *Simulation-based Econometric Methods*. Oxford University Press: Oxford.

Handcock, M.S. and Morris, M. (1999), *Relative Distribution Methods in Social Sciences*. Springer: New York.

Horowitz, J. L (2002), "The bootstrap in econometrics," *Statistical Science*, 18, Silver Anniversary of the Bootstrap, 18, 211-218.

Inglot, T., Kallenberg, W.C.M., Ledwina, T., (1994), "Power approximations to and power comparison of smooth goodness-of- fit tests," *Scandinavian Journal of Statistics* 21, 131-145.

Janic-Wróblewska, A. and Ledwina, T. (2000), "Data driven rank test for two-sample problem," *Scandinavian Journal of Statistics*, 27, 281-297.

Janssen, A. (2000), "Global power functions of goodness of fit tests," *The Annals of Statistics*, 28, 239-253.

Kendall, M.G. and Stuart, A. (1973), *The Advanced Theory of Statistics, 3rd Edition, Vol. 2,* Hafner: New York.

Kagan, A.M., Linnik,Yu. V. and Rao, C.R. (Translated from Russian text by B. Ramachandran), (1973), *Characterization Problems in Mathematical Statistics.* Wiley: New York.

Kallenberg, W.C.M., Oosterhoff, J., Schriever, B.F., (1985), "The number of

classes in 2 goodness of fit test," *Journal of the American Statistical Association* 80, pp. 959-968.

Ledwina, T. (1994), "Data-driven version of Neyman's smooth test of fit," *Journal of the American Statistical Association,* 89, 1000-1005.

Lehmann, E. L. (1953), "The power of rank tests," *Annals of Mathematical Statistics*, 24, 23-43.

McLeish, D. L. (1974), "Dependent Central Limit Theorems and Invariance Principles," *Annals of Probability*, 2, 620-628.

Mora, J. and Neumeyer, N., (2005), "The two-sample problem with regression errors: An empirical process approach," Working Paper, Departamento de Fundametos del Análisis Económico, Universidad de Alicante.

Newey, W.K., (1985), "Generalized method of moments specification testing," *Journal of Econometrics* 29, 229-256.

Neyman, J. (1937), "'Smooth test' for goodness of fit," *Skandinaviske Aktuarietidskrift,* 20, 150-199.

Neyman, J. and Pearson, E.S. (1936), "Contributions to the theory of testing statistical hypothesis I: Unbiased critical regions of Type A and A_1," *Statistical Research Memoirs,* 1, 1-37.

Parzen, E. (1992), "Comparison change analysis," in *Nonparametric Statistics and Related Topics*, Ed. A. Saleh. Elsevier: Holland, pp. 3-15.

Rao, C. R. (1948), "Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation," *Proceedings of the Cambridge Philosophical Society,* 44, 50-57.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons: New York.

Shorack, G. R. and Wellner, J. (1986), *Empirical Processes with Applications to Statistics*. John Wiley & Sons: New York.

Stock, J.H., J.H. Wright and M. Yogo (2002), "A Survey of weak instruments and weak identification in generalized method of moments," *The Journal of Business and Economic Statistics*, 20, 518-529

List of Tables

Test Statistic	New York	Pennsylvania	Critical Values Upper 0.1%
D^+	1.128	0.8915	1.859
D^-	4.3492	4.4053	1.859
KS	4.3492	4.4053	1.95
Kuiper	5.4809	5.3015	2.303
CvM	5.3503	5.1944	1.167
A-D	28.4875	25.2846	6.0
W	2.0858	1.9058	0.385

Table 1. Goodness-of-Fit Statistics based on EDF.

Source	$\hat{\Psi}_4^2$	$\hat{u}_1^2$	$\hat{u}_2^2$	$\hat{u}_3^2$	$\hat{u}_4^2$	$\widetilde{\Psi}_4^2$
New York $(n = 4548, m = 2517)$	117.5011	69.0543	46.0559	0.819	1.572	83.3719
p-value (χ^2)	0.0000	0.0000	0.0000	0.3655	0.2099	0.0000
Pennsylvania $(n = 3113, m = 1875)$	96.3788	64.9253	21.5569	8.7567	1.13	63.0314
p-value (χ^2)	0.0000	0.0000	0.0000	0.0031	0.2878	0.0000

Table 2. Smooth statistic and its four components.

Source	$\hat{\Psi}_4^2$	$\hat{u}_1^2$	$\hat{u}_2^2$	$\hat{u}_3^2$	$\hat{u}_4^2$
Test Statistic	27.6095	7.4413	13.7468	3.8773	2.544
p-value	0.0000	0.0064	0.0002	0.0489	0.1107

Table 3. Neyman's smooth statistic for NY and components $(n = 4548, m = 500)$.

Source	$\hat{\Psi}_4^2$	$\hat{u}_1^2$	$\hat{u}_2^2$	$\hat{u}_3^2$	$\hat{u}_4^2$
Asymptotic Null Distribution	χ_4^2	χ_1^2	χ_1^2	χ_1^2	χ_1^2
actual size ($n = 2500, m = 500$)	0.0577	0.0423	0.0529	0.0497	0.083
actual size ($n = 2500, m = 50$)	0.0425	0.0473	0.0474	0.0461	0.0473

Table 4. Actual sizes for 5% nominal size tests.

Level (α)	0.10	0.05	0.025	0.01	0.005
Critical Value for $\widehat{\Psi}_4^{*2}$	15.29	18.45	22.10	26.24	27.60
Critical Value for χ_4^2	7.78	9.49	11.14	13.28	14.86

Table 5. Critical Values for the resampled smooth test and χ_4^2

List of Figure Titles and Legends

Figure 1. 1^{st} and 2^{nd} normalized Legendre polynomials, $\pi_1(z)$ and $\pi_2(z)$.

Figure 2. RDF with different skewness and kurtosis. $f(x)$: Standard Normal; $g(x)$: Standardized χ^2_3 (left) or t_3 (right).

Figure 3. 3^{rd} and 4^{th} normailzed Legendre polynomials, $\pi_3(z)$ and $\pi_4(z)$.

Figure 4. Density estimates for New York.

Figure 5. Density estimates for Pennsylvania.

Figure 6. Histogram of Probability Integral Transform.

Figure 7. Criteria function for optimal sample size for closeness under H_0 ($n = 4548$; repl.$= 200$; $m_{opt} = 455$).

Figure 8. Test sample size and test size for $n = 5000$ and increasing m ($r = 2000$).

Figure 9. Density Plots for individual U_i^2 under H_0 and theoretical χ^2_1.

Figure 10. Density Plot for $\hat{\Psi}^2_4$ under H_0 and theoretical χ^2_4.

Figure 11. Size and Power of the smooth test for simulated data ($n = 625$).

Figure 12. Size and Power of the smooth test for simulated data ($n = 2000$).

Figure 13. Density of the modified $\widehat{\Psi}^2_k$ under H_0 when $\gamma = 0.5$ ($n = 2000, m = 1000, r = 2000$)

Figure 14. Density of the modified $\widehat{\Psi}^2_k$ under H_0 when $\gamma = 1$. ($n = 1000, m = 1000, r = 1000$)

Figure 15. Density of the modified $\widehat{\Psi}^2_k$ under H_1 when $\gamma = 0.5$ ($n = 2000, m = 1000, r = 1000$)

Figure 16. Density of the modified $\widehat{\Psi}^2_k$ under H_1 when $\gamma = 1$. ($n = 1000, m = 1000, r = 1000$)

Figure 17. Density of the resampled $\hat{\Psi}^{*2}_k$ under H_0 ($n = 2000, m = 2000, r = 5000$)

List of Figure Artwork

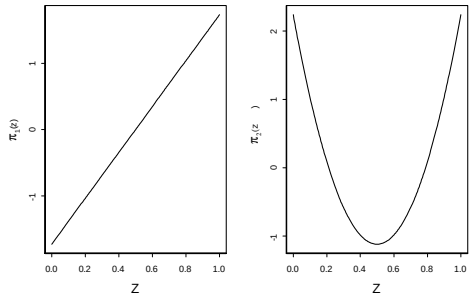

Figure 1.

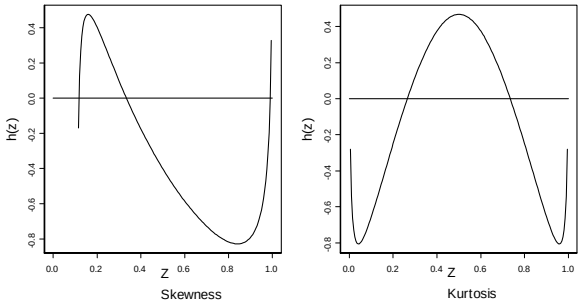

Figure 2.

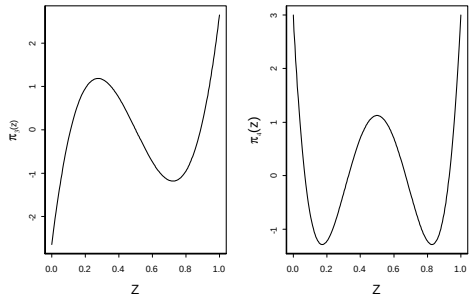

Figure 3.

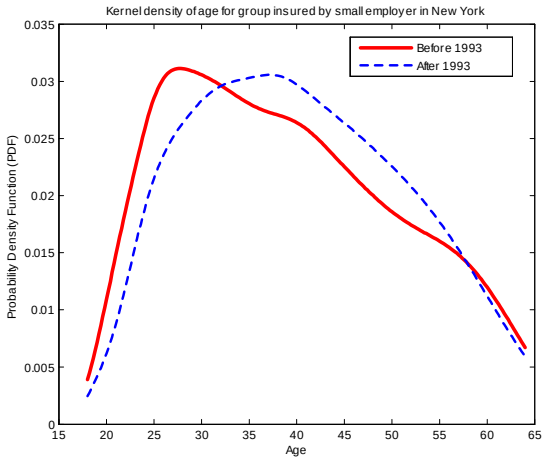

Figure 4.

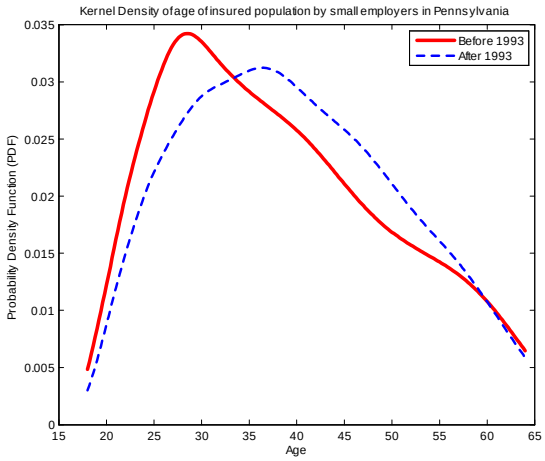

Figure 5.

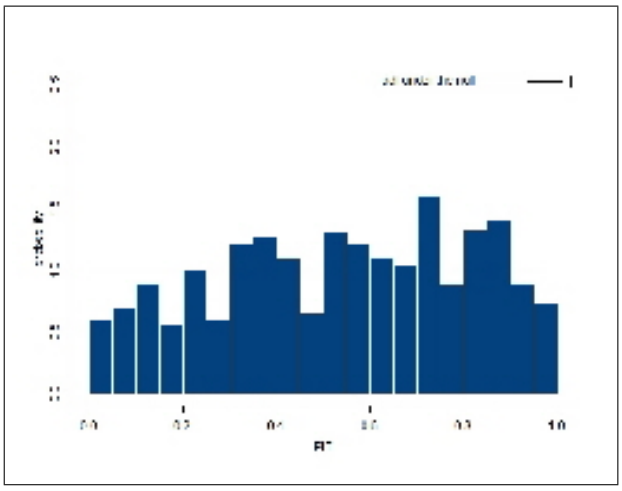

Figure 6.

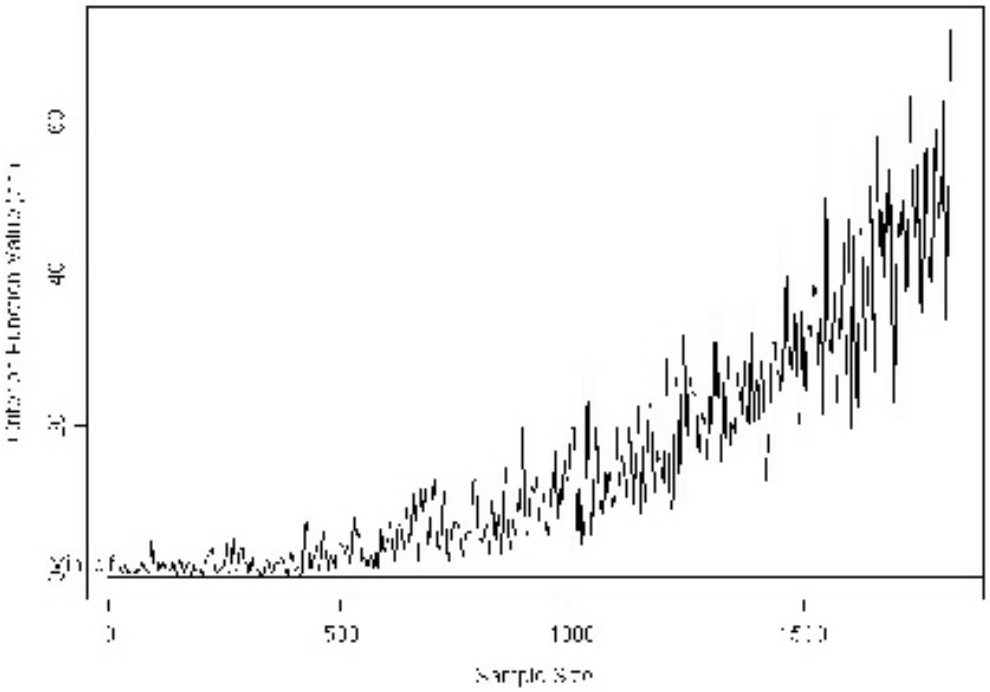

Figure 7.

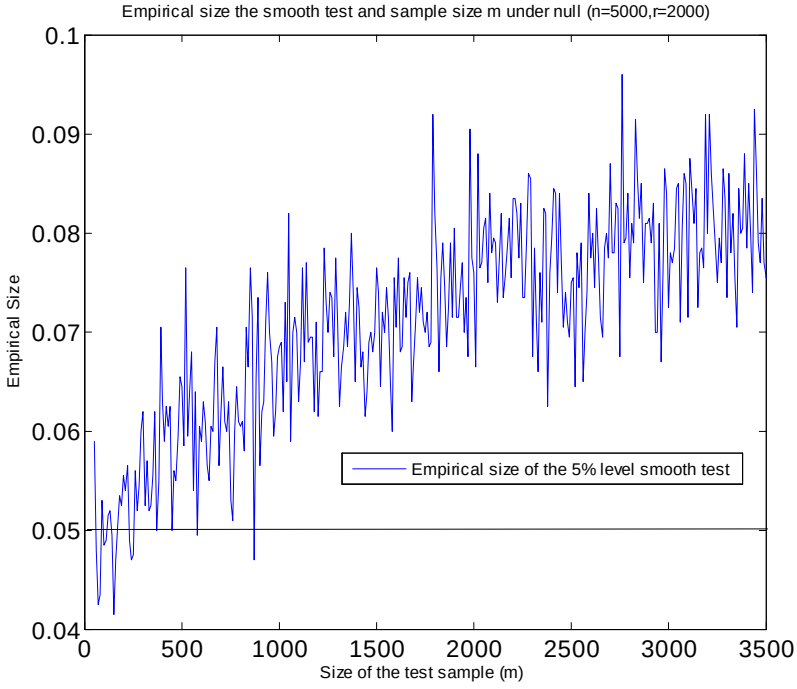

Figure 8.

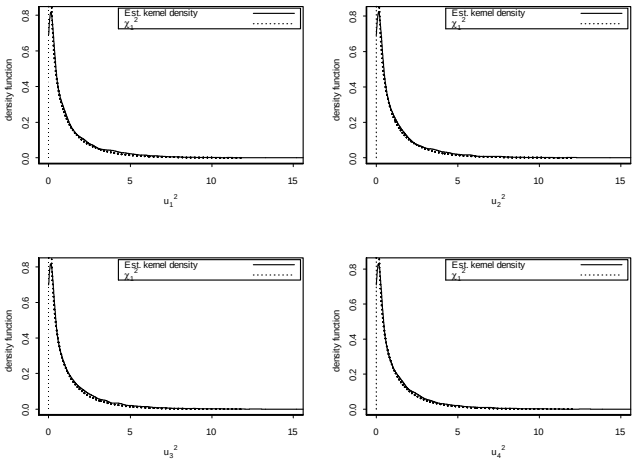

Figure 9.

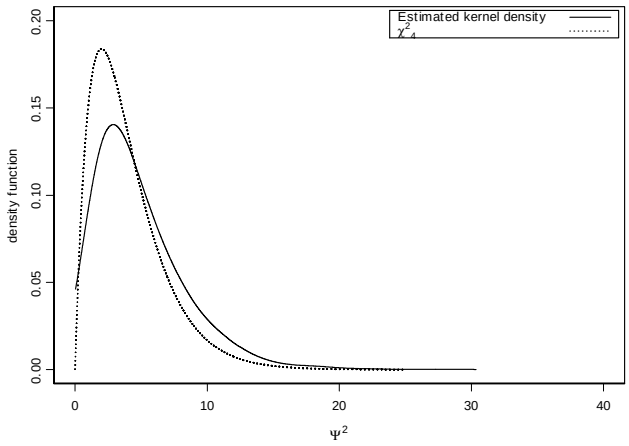

Figure 10.

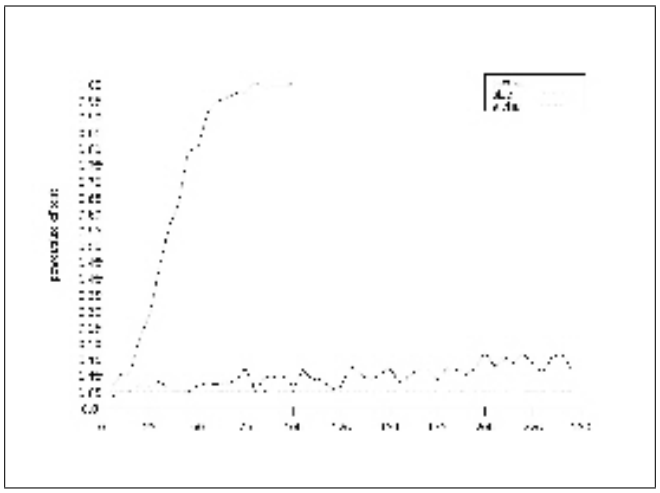

Figure 11.

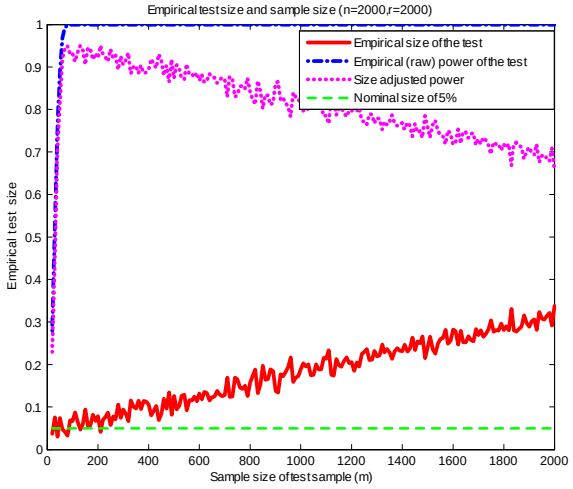

Figure 12.

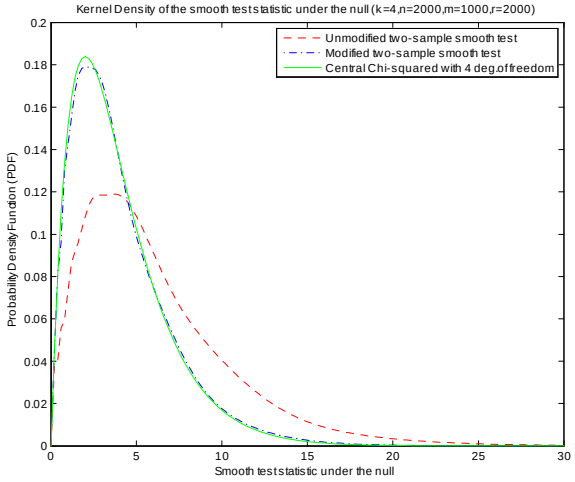

Figure 13.

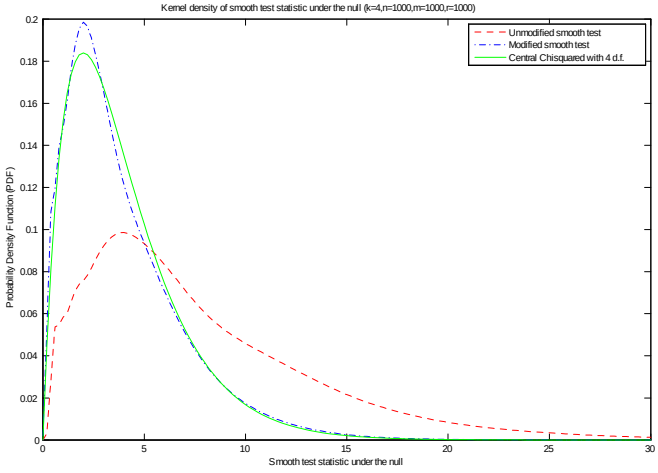

Figure 14.

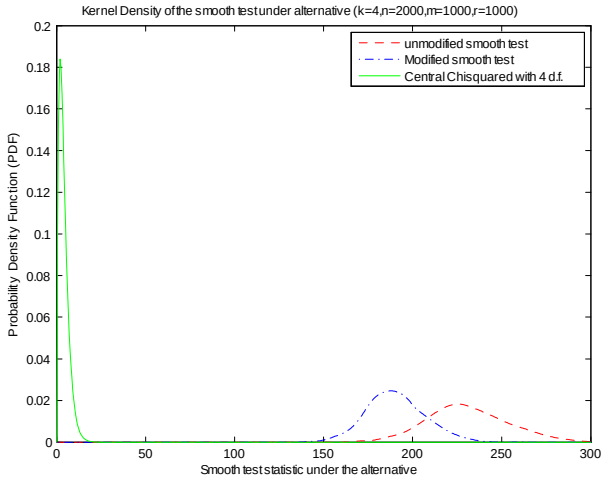

Figure 15.

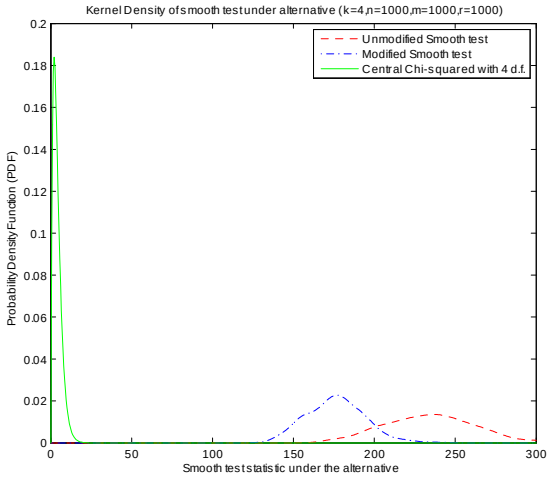

Figure 16.

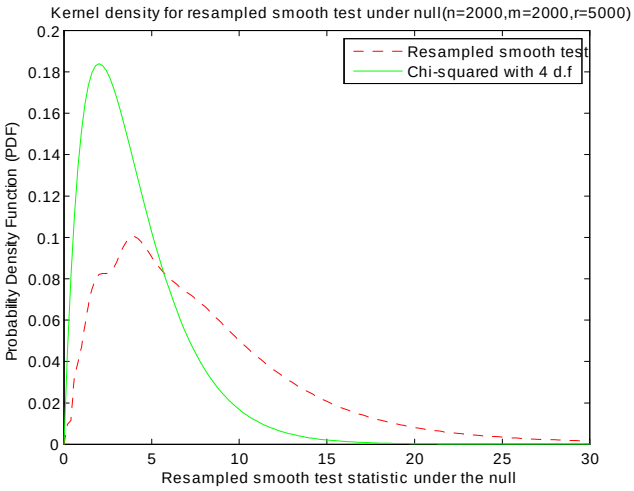

Figure 17.