Biometrika (2012), **99**, 1, *pp.* 1–17

Printed in Great Britain

Advance Access publication on 31 July 2012

Testing equality of a large number of densities

BY D. ZHAN

Department of Human Resources, Texas A&M University, College Station, Texas 77843, U.S.A.
dzhan@tamu.edu

AND J. D. HART

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
hart@stat.tamu.edu

SUMMARY

The problem of testing equality of a large number of densities is considered. The classical k-sample problem compares a small, fixed number of distributions and allows the sample size from each distribution to increase without bound. In our asymptotic analysis the number of distributions tends to infinity but the size of individual samples remains fixed. The proposed test statistic is motivated by the simple idea of comparing kernel density estimators from the various samples to the average of all density estimators. However, a novel interpretation of this familiar type of statistic arises upon centering it. The asymptotic distribution of the statistic under the null hypothesis of equal densities is derived, and power against local alternatives is considered. It is shown that a consistent test is attainable in many situations where all but a vanishingly small proportion of densities are equal to each other. The test is studied via simulations, and an illustration involving microarray data is provided.

Some key words: Kernel density estimation; k-sample problem; Local alternatives; Omnibus test; U-statistics.

1. INTRODUCTION

A recurring theme in modern statistics is dealing with a large number of rather small data sets. For example, microarrays produce data that can be framed in this way. The use of clustering in such problems is common. To wit, one assumes that all data sets come from a relatively small number of distributions, and the goal is to cluster the data sets so that those in a given cluster have a common distribution. In such problems it is advisable to verify that clustering is indeed necessary. A formal test of the null hypothesis that all data sets come from a single distribution could prevent a spurious clustering. Such a test is the subject of this paper. We propose a test that has good power when the number of small data sets tends to infinity, but the sample sizes of individual data sets are fixed.

The k-sample problem, i.e., testing whether k data sets come from the same population, is a classical one in statistics. In this setting k is assumed to be fixed, and asymptotic analysis proceeds by letting each of the k sample sizes tend to infinity. In contrast, we consider a setting where the number of data sets tends to infinity, but all sample sizes are bounded by a common value. To distinguish our situation from the classical one, we use the notation p for the number of data sets. This is also in keeping with the modern terminology of large p, small n problems. We are mainly interested in cases where n is quite small, small enough that a nonparametric test of the equality of distributions would have essentially no power if p were small, as in the classical

setting. Our test is based on the idea of comparing kernel density estimates computed from all the small data sets. When the null hypothesis is true, the differences between these density estimates will be relatively small, whereas if the alternative is true, the differences will be larger.

For the sake of simplicity, the main treatment of our proposed test assumes that all data sets are of the same, fixed size n. However, we also propose a test for the setting where sample sizes are different, and provide conditions on the sample sizes that ensure asymptotic normality of the test statistic. A proof of this result is given in our Supplementary Material.

In settings with n small, the power of a test derives from having a sufficiently large number of data sets whose distributions differ from the norm. We investigate the power of our test by assuming that the distributions for the various data sets are independently drawn from a countable collection of distributions. It is shown that, as $p \to \infty$, our test is consistent against such alternatives. We also consider local alternatives for which the proportion of data sets having the same distribution, call it g, tends to 1 as $p \to \infty$. Here, if the number of distributions different from g is just larger than $p^{1/2}$, then our test is still consistent.

The two-sample Kolmogorov–Smirnov test (Stephens, 1974) is a useful and popular nonparametric method for testing whether two samples are from the same distribution. Other traditional two-sample goodness-of-fit tests based on empirical distribution functions include the Cramér–von Mises and Anderson–Darling tests (Anderson & Darling, 1954; Anderson, 1962; Stephens, 1974, 1986). Recently, Jimenez-Gamero et al. (2009) proposed a test based on empirical characteristic functions. Anderson, Hall and Titterington (1994), Louani (2000), and Cao and Van Keilegom (2006) considered tests for the two-sample problem based on kernel density estimates.

As for testing equality of more than two distributions, Kiefer (1959) proposed extensions of the Kolmogorov–Smirnov and Cramér–von Mises tests to the k-sample setting, while Scholz and Stephens (1987) extended the Anderson–Darling test to that case. Martínez-Camblor, de Uña Álvarez and Corral (2008) proposed a test for comparing k samples that is based on kernel density estimators. The authors suggest that density-based tests may be more powerful than ones based on the empirical distribution function. The simulations of Martínez-Camblor and de Uña Álvarez (2009) provide evidence that tests based on L_1 distances between kernel density estimators are generally more powerful than either tests based on empirical distribution functions or the test of Martínez-Camblor, de Uña Álvarez and Corral (2008).

Aside from the large literature on microarray analysis (for example, Efron, 2004), we are aware of only a few articles that deal with inference for a large number of small data sets. Cox and Solomon (1986) investigated methods for checking the fit of a model that assumes many small samples are normally distributed with a common variance. Cox and Solomon (1988) proposed a test for detecting within-samples serial correlation when the data consist of many small samples. Park and Park (2012) consider the classical analysis of variance problem when the number of data sets is large.

2. THE DATA AND PROPOSED TEST

The observed data are X_{ij} ($j = 1, \ldots, n$; $i = 1, \ldots, p$). It is assumed that $X_{i1}, \ldots, X_{in}$ is a random sample from density f_i ($i = 1, \ldots, p$), and the samples are taken independently of each other. Our interest is in testing the null hypothesis $H_0 : f_1 = \cdots = f_p$ against the alternative that not all the densities are the same. We let $N(a, b)$ denote a normal distribution with mean a and variance b.

Our test of H_0 is based on comparing kernel density estimates from each of the p samples with a single kernel estimate using all the data pooled together. Define

$$\hat{f}_i(x) = \frac{1}{nh} \sum_{j=1}^{n} \phi\left(\frac{x - X_{ij}}{h}\right), \quad i = 1, \ldots, p,$$

where ϕ is the standard normal density and $h > 0$ is the bandwidth of the kernel estimate. The pooled estimator is

$$\hat{f}(x) = \frac{1}{p} \sum_{i=1}^{p} \hat{f}_i(x),$$

which is equivalent to a kernel estimator based on all np data values pooled together. We have chosen the bandwidths of these $p + 1$ kernel estimates to be the same, in order to ensure that, when H_0 is true, all estimates are estimating the same function, in the sense that $E\{\hat{f}_i(x) - \hat{f}(x)\} = 0$. The same principle has been used in related settings, such as that in Young and Bowman (1995).

Thoughout this paper we use a normal kernel. It would be interesting to consider the effect of using a different kernel, although doing so is beyond the scope of the current paper. Our intuition is that other symmetric, unimodal kernels will not yield substantially different results than those obtained with a normal kernel. As in the setting of density estimation, the bandwidth h probably has a larger effect than kernel choice, and we will give some results on the effect of h.

Our ultimate test statistic is derived by starting from a statistic analogous to that of Young and Bowman (1995) in the regression setting. Define

$$\begin{aligned} T_p &= \frac{1}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} \left\{ \hat{f}_i(x) - \hat{f}(x) \right\}^2 dx \\ &= \int_{-\infty}^{\infty} \frac{1}{p} \sum_{i=1}^{p} \left\{ \hat{f}_i(x) - \hat{f}(x) \right\}^2 dx \\ &= \frac{1}{p} \sum_{i=1}^{p} \int_{-\infty}^{\infty} \hat{f}_i^2(x)\, dx - \int_{-\infty}^{\infty} \hat{f}^2(x)\, dx. \end{aligned} \tag{1}$$

The statistic T_p is an obvious one to use for testing equality of the densities. It is simply a kernel estimate analog of an analysis of variance sum of squares statistic. However, centering T_p so that its limiting distribution under H_0 has mean zero is a nontrivial operation. It turns out that one must subtract from T_p an estimator of $E(T_p)$ whose variance is of the same order in p as that of T_p. Rather than taking this approach, evaluation of the integrals in (1) reveals an effortless way of doing the centering, and leads to a test statistic with an alternative motivation.

The following well-known property is used a number of times subsequently.

Property 1. The convolution of $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ densities is $N(0, \sigma_1^2 + \sigma_2^2)$.

Using Property 1 we have

$$\int_{-\infty}^{\infty} \hat{f}_i^2(x)\,dx = \frac{1}{n^2 2^{1/2} h} \sum_{j=1}^{n}\sum_{l=1}^{n} \phi\left(\frac{X_{ij}-X_{il}}{2^{1/2}h}\right),$$

$$\int_{-\infty}^{\infty} \hat{f}^2(x)dx = \frac{1}{p^2 n^2 2^{1/2} h} \sum_{i=1}^{p}\sum_{k=1}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n} \phi\left(\frac{X_{ij}-X_{kl}}{2^{1/2}h}\right).$$

We may therefore write

$$T_p = \frac{(p-1)}{p}\frac{1}{n2^{1/2}h}\phi(0) + \frac{p-1}{p^2}\frac{1}{n^2 2^{1/2} h} \sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n} \phi\left(\frac{X_{ij}-X_{il}}{2^{1/2}h}\right)$$
$$-\frac{1}{p^2 n^2 2^{1/2} h} \sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n} \phi\left(\frac{X_{ij}-X_{kl}}{2^{1/2}h}\right).$$

Suppose that H_0 is true and let the common density of each observation be f. Then, again using Property 1, for $j \neq l$, $i \neq k$ and arbitrary $1 \leq r, s \leq n$,

$$E\left\{\frac{1}{2^{1/2}h}\phi\left(\frac{X_{ij}-X_{il}}{2^{1/2}h}\right)\right\} = E\left\{\frac{1}{2^{1/2}h}\phi\left(\frac{X_{ir}-X_{ks}}{2^{1/2}h}\right)\right\}$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2^{1/2}h}\phi\left(\frac{x-y}{2^{1/2}h}\right) f(x)f(y)\,dxdy$$
$$= \int_{-\infty}^{\infty} f_h^2(x)\,dx,$$

where

$$f_h(x) = E\{\hat{f}(x)\} = \int_{-\infty}^{\infty} \frac{1}{h}\phi\left(\frac{x-y}{h}\right) f(y)\,dy.$$

To center the statistic under H_0, we first drop the term depending on $\phi(0)$, since it contains no information about the underlying densities. Secondly, up to known multipliers, the two sums in T_p have the same expectations under H_0. Therefore, defining

$$S_W = \frac{1}{pn(n-1)2^{1/2}h} \sum_{i=1}^{p}\sum_{j=1}^{n}\sum_{l=1,l\neq j}^{n} \phi\left(\frac{X_{ij}-X_{il}}{2^{1/2}h}\right),$$

$$S_B = \frac{1}{p(p-1)n^2 2^{1/2} h} \sum_{i=1}^{p}\sum_{k=1,k\neq i}^{p}\sum_{j=1}^{n}\sum_{l=1}^{n} \phi\left(\frac{X_{ij}-X_{kl}}{2^{1/2}h}\right),$$

the unstudentized version of our test statistic is $S = S_W - S_B$, which has mean zero under H_0. As we will subsequently show, $E(S)$ is larger than zero when the alternative hypothesis is true. The subscripts W and B stand for within and between, respectively. The statistic S_W is an intra-samples parameter estimate, since X_{ij} and X_{il} come from the same small data set, whereas S_B is an inter-samples estimate, since X_{ij} and X_{kl} come from different data sets. These two statistics estimate the same parameter, $\int_{-\infty}^{\infty} f_h^2(x)dx$, under H_0, but different parameters under the alternative.

Using U-statistic technology, we will show in Section 3 that $S/\mathrm{var}(S)^{1/2}$ converges in distribution to a standard normal random variable as $p \to \infty$ with n and h fixed. Not surprisingly,

$\mathrm{var}(S)$ depends upon the unknown density f under H_0. It is therefore necessary to construct a consistent estimator of $\mathrm{var}(S)$ in order to obtain the final version of our test statistic that is asymptotically normal under H_0. We will propose such an estimator in Section 3.

3. ASYMPTOTIC NULL DISTRIBUTION OF TEST STATISTIC

Theorem 1 establishes the asymptotic normality of the statistic S. Its proof is given in the Appendix and uses the theory of U-statistics. The following notation is useful in the proof and in defining a consistent estimator of the variance of S. Let $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ and define

$$h_1(x) = \frac{1}{n(n-1)2^{1/2}h} \sum_{j=1}^{n} \sum_{l=1, j\neq l}^{n} \phi\left(\frac{x_j - x_l}{2^{1/2}h}\right), \tag{2}$$

$$h_2(x, y) = \frac{1}{n^2 2^{1/2}h} \sum_{j=1}^{n} \sum_{l=1}^{n} \phi\left(\frac{x_j - y_l}{2^{1/2}h}\right). \tag{3}$$

THEOREM 1. *Let n and h be fixed for all p, and suppose that X_{ij} ($j = 1, \ldots, n$; $i = 1, \ldots, p$) are independent and identically distributed with common density f. Then $S/\mathrm{var}(S)^{1/2}$ converges in distribution to a random variable having the standard normal distribution as p tends to ∞.*

Remark 1. An implicit assumption in Theorem 1 is that the limiting variance of $p^{1/2}S$ is larger than zero. In order for this variance to be zero, h_1 would have to be a linear function of h_3, as defined by (A2). If this is possible at all, it would be a pathological case, as one can see upon comparing definitions (2) and (A2).

To make Theorem 1 practically useful, we need a consistent estimator of the null value of $p\mathrm{var}(S)$. Using U-statistic technology, it suffices to replace $\mathrm{var}(S)$ by the variance of the following projection of $S_W - S_B$:

$$\frac{1}{p}\sum_{i=1}^{p} h_1(X_i) - \frac{2}{p}\sum_{i=1}^{p} h_3(X_i) + E(S_B).$$

If h_3 were known, we could simply use the sample variance of $h_1(X_1) - 2h_3(X_1), \ldots, h_1(X_p) - 2h_3(X_p)$ to estimate $\sigma^2 = \mathrm{var}\{h_1(X_1) - 2h_3(X_1)\}$. However, unlike h_1, h_3 depends upon the unknown null density f. We may deal with this problem by estimating h_3, as in Sen (1960) and Schucany and Bankson (1989). Define

$$\hat{h}_3(x; i) = \frac{1}{(p-1)} \sum_{k=1, k\neq i}^{p} h_2(X_k, x), \quad i = 1, \ldots, p.$$

The quantity σ^2 is estimated by the sample variance, call it $\hat{\sigma}^2$, of $h_1(X_i) - 2\hat{h}_3(X_i; i)$ ($i = 1, \ldots, p$). The statistic $\hat{\sigma}^2$ is essentially a jackknife variance estimator, as studied by Arvesen (1969) in the case of U-statistics for univariate random variables. We now have the following corollary to Theorem 1.

COROLLARY 1. *Let the conditions of Theorem* 1 *be satisfied, and suppose that $\hat{\sigma}^2$ is the estimator defined immediately above. Then $p^{1/2}S/\hat{\sigma}$ converges in distribution to a random variable having the standard normal distribution as p tends to ∞.*

Corollary 1 is proven using Theorem 1, the consistency of $\hat{\sigma}^2$, as shown in the Supplementary Material, and Slutsky's theorem. Formally, our test of the null hypothesis that all data have a common distribution is rejected at nominal level α when $S_p = p^{1/2}S/\hat{\sigma}$ is larger than the $(1-\alpha)$ quantile of the standard normal distribution. The test is one-sided since, as shown in the next section, $E(S) > 0$ under the alternative hypothesis.

4. ASYMPTOTIC POWER PROPERTIES

4·1. *Fixed alternatives*

It is not reasonable to expect that our test can consistently detect every fixed alternative to $H_0 : f_1 = f_2 = \cdots = f_p$ as $p \to \infty$ with n fixed, as the following argument illustrates. Suppose we have p data sets each of size n, where the np observations are mutually independent. We know that the first m data sets come from $N(\mu_1, 1)$ and the other $p - m$ come from $N(\mu_2, 1)$. If we want to detect a difference in the two distributions we cannot expect better power than that of the standard two-sample test of $H_0 : \mu_1 = \mu_2$. The limiting power of this test as $p \to \infty$ with m and n fixed is equal to the probability that the random variable $(nm)^{1/2}|\bar{X} - \mu_2|$ exceeds a standard normal critical value, where $\bar{X}$ is the sample mean of the nm observations in the first m data sets. Since n and m are fixed this probability is less than 1, and hence the test is inconsistent.

This leads to the question of what constitutes a reasonable fixed alternative in our setting. In Condition 1 below, we define alternatives in which densities are randomly selected from a countable collection of densities. This is tantamount to a clustering model in which it is assumed that all data within a cluster have a common distribution and there are countably many clusters.

Condition 1. For each p, densities $f_1, \ldots, f_p$ are drawn independently from a countable collection $\{g_1, g_2, \ldots\}$ of densities. For each i, f_i is identical to g_r with probability ρ_r ($r = 1, 2, \ldots$). Once a density is selected, a random sample of size n is drawn from that density. We assume that each $\rho_r < 1$, which ensures that the alternative hypothesis is true, almost surely, for all p sufficiently large.

Condition 1 constitutes a fixed alternative in the sense that $f_1, \ldots, f_p$ are drawn from the same collection of densities for every p.

Let z_α be the $1 - \alpha$ quantile of the standard normal distribution, and let $\mu = E(S)$. Then the power of our nominal size α test is

$$\mathrm{pr}\left(\frac{p^{1/2}S}{\hat{\sigma}} > z_\alpha\right) = \mathrm{pr}\left\{\frac{p^{1/2}(S-\mu)}{\hat{\sigma}} > z_\alpha - \frac{p^{1/2}\mu}{\hat{\sigma}}\right\}. \tag{4}$$

The key to obtaining good power is that, except in a trivial case, the parameter μ is larger than 0 under Condition 1, as established in the following lemma. All lemmas and theorems in this section are proven in the Appendix, unless otherwise specified.

LEMMA 1. *Let $g_r(\,\cdot\,; h)$ be the convolution of g_r with the $N(0, h^2)$ density. Under Condition* 1*,*

$$E(S) = \mu = \sum_{r=1}^{\infty}\sum_{s>r}\rho_r\rho_s\int\{g_r(x;h) - g_s(x;h)\}^2\,dx. \tag{5}$$

If in addition there are two distinct integers r and s such that $\rho_r > 0$, $\rho_s > 0$ and g_r and g_s are different in the sense that their characteristic functions are different, then $\mu > 0$.

Under the model entailed by Condition 1, the unconditional density of each observation X_{ij} is the same, and equal to the mixture $\sum_{r=1}^{\infty} \rho_r g_r$. For this reason it may seem odd that $\mu > 0$. The reason that μ is larger than 0 under our alternative model is that in this case observations in the same data set are correlated. One may regard the model entailed by Condition 1 as a random effects model. Even though observations within a small data set of size n are conditionally independent, they are dependent with respect to their unconditional joint distribution. This, along with the fact that observations in different small data sets are unconditionally independent, leads to $\mu > 0$.

We may now state a main result about the power of our test.

THEOREM 2. *Assume that Condition* 1 *holds. Then with* n *and* h *fixed for all* p, $p^{1/2}(S - \mu)/\hat{\sigma}$ *converges in distribution to a random variable with the standard normal distribution as* $p \to \infty$. *Under the further condition that* $\mu > 0$, *the test that rejects* H_0 *for large values of* $p^{1/2}S/\hat{\sigma}$ *is consistent, i.e., its power tends to 1 as* $p \to \infty$.

Remark 2. In order for a kernel density estimator to be consistent for the underlying density, it is necessary, in general, for its bandwidth to tend to 0 as sample size tends to infinity. In the current setting it is not required that $h \to 0$ for test consistency. This is because consistency results from $g_r(\cdot; h)$ being different from $g_s(\cdot; h)$ for some r and s, which only requires that g_r and g_s be different.

4·2. *Local alternatives*

We continue to assume that the alternative model of the previous section holds, but now we assume that $\rho_r \to 0$ for $r = 2, 3, \ldots$, which implies that $\rho_1 \to 1$. Therefore, the alternative becomes very close to the null in that the proportion of densities equal to g_1 is very close to 1. The question is, how quickly may the proportions tend to 0 and still allow our test to have substantial power?

Our local alternative is defined by the following condition

Condition 2. When the number of data sets is p, the probability of selecting density g_r is $\rho_{rp} = a_r \epsilon_p$, $r = 2, 3, \ldots$, where $\{a_2, a_3, \ldots\}$ is a positive summable sequence, $\{\epsilon_p\}$ is a positive sequence that tends to 0 as $p \to \infty$, and $\epsilon_p \sum_{r=2}^{\infty} a_r < 1$ for $p = 1, 2, \ldots$. Note that $\rho_{1p} = 1 - \epsilon_p \sum_{r=2}^{\infty} a_r$.

The power of our test against local alternatives is provided by the following theorem.

THEOREM 3. *Under Condition* 2, *the limiting power of the test based on* $p^{1/2}S/\hat{\sigma}$ *is*

$$1 - \Phi\left[z_\alpha - \frac{p^{1/2}\epsilon_p}{\sigma} \sum_{s=2}^{\infty} a_s \int \{g_1(x; h) - g_s(x; h)\}^2 \, dx\right], \tag{6}$$

where σ^2 *is the limiting variance of* $p^{1/2}S$ *when all data have density* g_1, *and* Φ *is the cumulative distribution function of a* $N(0, 1)$ *distribution.*

If at least one of $g_2, g_3, \ldots$ is different from g_1, and $p^{1/2}\epsilon_p \to \infty$ as $p \to \infty$, then Theorem 3 entails that the power of our test tends to 1 as $p \to \infty$. Furthermore, our test can detect $p^{1/2}$ alternatives in that, if $\epsilon_p = p^{-1/2}$, the limiting power is greater than α, and given explicitly by (6). It may seem that expression (6) does not depend on n, but indeed it does. The parameter σ depends on n, and as n increases, σ decreases, which increases power.

5. BANDWIDTH SELECTION

To this point we have assumed that the bandwidth is fixed as $p \to \infty$. Expression (6) shows that, in general, the power of the test will depend on which fixed bandwidth one chooses. Furthermore, the choice of h that optimizes power depends on information that is difficult to estimate. Therefore, the first method we discuss for choosing h sidesteps the question of power, and simply provides a bandwidth whose scale is commensurate with the scale of the observations.

Our method is motivated by the maximal smoothing principle of Terrell and Scott (1985) and Terrell (1990). For a Gaussian kernel their bandwidth is $h_{OS} = 1.144\sigma_{\text{pop}}n^{-1/5}$, which is an upper bound on an asymptotically optimal bandwidth when the population standard deviation is σ_{pop}. A data-driven bandwidth is obtained by substituting either the sample standard deviation or a more robust estimate of scale for σ_{pop} in h_{OS}.

In our setting, we have a different estimate of scale for each data set. We thus suggest using s_{pool}, where s^2_{pool} is the average of all p sample variances. Our choice of h is thus

$$\hat{h} = 1.144 s_{\text{pool}} n^{-1/5}. \tag{7}$$

Another possibility is to use a more robust scale estimate, such as the median of all p standard deviations. In general the choice (7) does not produce the best power. However, it has the virtue of simplicity and stability, inasmuch as s_{pool} is based on a large number, np, of observations.

Define, for each $h > 0$,

$$S_p(h) = \frac{S_W - S_B}{\hat{\sigma}/p^{1/2}}, \tag{8}$$

where the quantities on the right hand side of (8) are defined in Section 3. It is important to point out that, so long as the variance of the null distribution is finite, the limiting distribution of $S_p(\hat{h})$ as $p \to \infty$ is the same as that of $S_p(h_{OS})$, where $h_{OS} = 1.144\{\text{var}(X_{ij})\}^{1/2}n^{-1/5}$. This is so because, as shown in the Supplementary Material, $\hat{h}$ differs from h_{OS} by $O(p^{-1/2})$, and as a result $S_p(\hat{h}) = S_p(h_{OS}) + O(p^{-1/2})$.

There are certainly other possibilities for choosing h, the most obvious of which is estimating an h that maximizes power. This approach has been taken in other testing problems based on smoothing methods; see, for example, Kulasekera and Wang (1997), Doksum and Schafer (2006) and Martínez-Camblor and de Uña Álvarez (2013). Expression (4) suggests that a value of h maximizing $S_p(h)$, i.e., one producing the smallest P-value, would be close to the a priori h that maximizes power. Of course, if one rejected H_0 when the smallest P-value, call it $\widehat{P}$, was less than α, then the size of the test would be larger than α. If the actual null distribution of $\widehat{P}$ were determined, then one could perform a valid test based on $\widehat{P}$. It should be clear though, that this test will have smaller power than a test based on $S_p(h)$ with a power-optimal fixed h. Nonetheless, the $\widehat{P}$-based test might be a reasonable compromise between an optimal statistic and one based on bandwidth (7). The requisite adjustment to the critical value of $\widehat{P}$ could be made using the bootstrap. To do so, one may draw samples of size np randomly and with replacement from X_{ij} ($j = 1, \ldots, n$; $i = 1, \ldots, p$). For each sample drawn, p small data sets each of size n are constructed, and $\hat{P}^*$ is computed from these p data sets in exactly the same way $\hat{P}$ was computed from the original data. The null hypothesis would be rejected if $\hat{P}$ is smaller than the α quantile of all the values of $\hat{P}^*$.

6. NUMERICAL RESULTS

6·1. *Simulations using $S_p(\hat{h})$*

The simulations in this section employ the test statistic $S_p(\hat{h})$ with $\hat{h}$ defined as in (7). For each case in which the the null hypothesis is true the number of replications is 2000 and otherwise it is 1000. Initially we consider the level properties of our large sample test. Four different possibilities were considered for the common distribution under the null hypothesis. These are a standard normal distribution, a t distribution with 3 degrees of freedom, a bimodal mixture of two normal distributions with means -2 and 2 and common standard deviation 1, and a gamma density with shape parameter 3 and scale 1. Since our test statistic is invariant to location and scale under the null hypothesis, the results reported are not affected by the particular location and scale chosen for a density.

Table 6·1 gives results for the gamma density. Results for the other three densities are not shown here as they are similar to the gamma case. The estimated level of our large sample test was consistently close to the nominal level. On the basis of a binomial test with level of significance 0.05, none of the empirical rejection rates in Table 1 is significantly larger than the corresponding nominal level. Considering all four densities, only three of 192 cases had an empirical rate significantly larger than the nominal level, and all three of these occurred at nominal level 0.10. If anything, the rates tended to be slightly too small.

We now consider three types of alternatives to the null hypothesis. One type, Case 1, is such that there are only two densities, standard normal and $N(1,1)$, with $\rho 100\%$, $\rho = 0.1, 0.2$, of the p data sets being drawn from the latter. Another type, Case 2, also has only two densities, standard normal and $N(0,4)$, with $\rho 100\%$, $\rho = 0.1, 0.2$, of the p data sets being drawn from $N(0,4)$. In Case 3 densities are drawn from a continuous scale mixture: $f_i|\beta_i \sim$ gamma with shape 3 and scale β_i, and $\beta_i \sim$ gamma with shape 50 and scale 1/50.

For each of the three settings we considered a sort of oracle test to see how well our nonparametric test fares in comparison to a good parametric test. The oracle was a likelihood ratio test. For Case 1 this was the classical F-test from analysis of variance, for Case 2 a Gaussian-based likelihood ratio test of the null hypothesis of equal variances, and for Case 3 a likelihood ratio test of equal scale parameters assuming the data are gamma distributed with shape parameter known to be 3.

All results are for a nominal α of 0.05 and are given in Table 6·1. An interesting aspect of the results is the substantial increase in power resulting from a small increase in n. This knowledge could be quite useful in deciding how many experimental units would be needed to ensure detection of differences among distributions. With the possible exception of Case 2, our nonparametric test performed reasonably well in comparison to the parametric tests. In all three cases the nonparametric test had empirical power that was at least 69% that of the likelihood ratio test when $n > 3$ and $p \geq 500$.

6·2. *Simulation investigating bandwidth effect*

We also did a small study to investigate the effect of bandwidth on the power of the test. We considered the Case 2 scale alternative with $\rho = 0.1$, $n = 3$ and $p = 500$. Let $\mathrm{Pow}(h)$ denote the power of a size 0.05 test based on test statistic $S_p(h)$. For each h in a grid of 100 values between 0.25 and 5, $\mathrm{Pow}(h)$ was estimated by generating 2000 data sets from the alternative and determining the proportion of cases in which $S_p(h)$ exceeded 1.645. A local linear estimate $\widehat{\mathrm{Pow}}(h)$ of $\mathrm{Pow}(h)$ based on these results was unimodal, with $\widehat{\mathrm{Pow}}(0.25) = 0.14$ and $\widehat{\mathrm{Pow}}(5) =$

Table 1. *Empirical rejection rates (%) of large sample test when the null hypothesis is true and the common density is gamma with shape 3 and scale 1. Each value is the percentage of rejections in 2000 replications.*

		Nominal α (%)		
n	p	1	5	10
2	100	0.95	4.35	9.20
	500	0.40	3.85	9.30
	1000	0.95	4.85	9.35
	5000	1.30	4.75	9.95
3	100	0.40	3.35	7.55
	500	0.75	4.05	9.05
	1000	0.90	4.85	8.95
	5000	0.95	5.15	10.10
5	100	0.40	3.65	8.25
	500	0.70	3.35	8.05
	1000	0.45	4.20	9.05
	5000	0.80	4.65	9.35
10	100	0.30	2.85	7.90
	500	0.60	4.15	9.60
	1000	0.45	4.50	9.65
	5000	0.75	4.35	9.20

0.06. The maximum estimated power was 0.33, which occurred at $h = 1.1$. An estimate of the standard error of each estimated power is no more than 0.01.

We note that the Table 6·1 entry for the normal scale case with $n = 3$, $p = 500$ and $\rho = 0.1$ is 32, and therefore our data-driven bandwidth produced power that differs insubstantially from what is possible using the best fixed bandwidth. Although this result is in just a single setting, it is nonetheless encouraging.

Finally, we investigated an idea equivalent to that mentioned in the last paragraph of Section 5. The test statistic was $M_p = \max_{0.25 \le h \le 5} S_p(h)$. The null distribution of M_p was approximated in the case $n = 3$ and $p = 500$ by generating data from the the standard normal distribution. On the basis of 2000 replications the 95th percentile of M_p was estimated to be 1.915. We then generated 2000 replications from the alternative (10% $N(0, 4)$ and 90% $N(0, 1)$), and found that the proportion of cases in which M_p exceeded 1.915 was 0.31. Again this is encouraging in that the power is quite close to the best empirical power obtained with a fixed bandwidth test.

6·3. *Microarray data*

Here we present an analysis of data collected by Professor Robert Chapkin and coworkers at Texas A&M University; see Davidson et al. (2004). The data analyzed are part of a much larger data set, but provide a good example of our methodology. The data are from five rats, all of which were subjected to the same treatment. There are 8038 logged gene expression levels from each

Table 2. *Empirical power (%) of nonparametric ($S_p(\hat{h})$) and parametric (LR) tests. The nominal α is 5% and each table value is the percentage of rejections in 1000 replications. In the normal location case, $(1-\rho)100\%$ of the data sets are drawn from $N(0,1)$ and $\rho 100\%$ are drawn from $N(1,1)$. In the normal scale case, $(1-\rho)100\%$ of the data sets are drawn from $N(0,1)$ and $\rho 100\%$ are drawn from $N(0,4)$. In the gamma case, data set i is gamma with shape 3 and scale β_i, $i = 1, \ldots, p$, where $\beta_1, \ldots, \beta_p$ are independent and identically distributed gamma with shape 50 and scale 1/50.*

		Normal location				Normal scale				Gamma	
		$\rho = 0.1$		$\rho = 0.2$		$\rho = 0.1$		$\rho = 0.2$			
n	p	$S_p(\hat{h})$	LR	$S_p(\hat{h})$	LR	$S_p(\hat{h})$	LR	$S_p(\hat{h})$	LR	$S_p(\hat{h})$	LR
2	100	13	20	29	39	9	19	13	31	11	19
	500	43	57	84	93	17	57	35	76	26	53
	1000	66	81	98	100	26	81	52	96	43	78
	5000	100	100	100	100	76	100	98	100	95	100
3	100	22	39	55	73	10	43	20	65	17	32
	500	76	93	99	100	32	91	71	100	48	83
	1000	96	100	100	100	56	99	92	100	76	98
	5000	100	100	100	100	99	100	100	100	100	100
5	100	45	77	89	99	22	82	47	98	33	58
	500	99	100	100	100	69	100	99	100	88	99
	1000	100	100	100	100	94	100	100	100	99	100
	5000	100	100	100	100	100	100	100	100	100	100
10	100	86	100	100	100	51	100	94	100	72	94
	500	100	100	100	100	100	100	100	100	100	100
	1000	100	100	100	100	100	100	100	100	100	100
	5000	100	100	100	100	100	100	100	100	100	100

rat, and so $n = 5$ and $p = 8038$. Denoting the original data Y_{ij} ($i = 1, \ldots, 8038$; $j = 1, \ldots, 5$), the data analyzed were $X_{ij} = Y_{ij} - \bar{Y}_j$, where $\bar{Y}_j = \sum_{i=1}^{8038} Y_{ij}/8038$. This transformation was done to effectively eliminate additive rat effects.

Before conducting the main analysis, we investigated the possibility of correlation between log(expression levels) of different genes on the same rat. A common form of correlation in microarray data is autocorrelation with respect to gene proximity (Koren, Tirosh and Barkai, 2007). Treating the data of each rat as a time series of length 8038, we computed the sample autocorrelation function for each rat at lags 1 to 1000. The first lag autocorrelation for each rat was between 0.09 and 0.10, and no other autocorrelation for any rat exceeded 0.09. On this basis, the assumption of independence across genes seems like a reasonable working hypothesis.

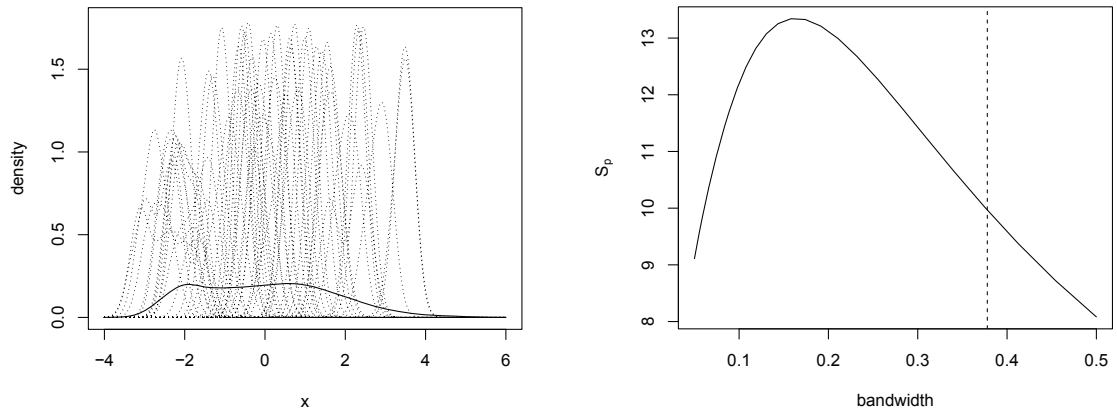

Fig. 1. Plots arising from the gene expression data. On the left, the dashed lines are kernel estimates for 50 randomly selected genes, and the solid line is a kernel estimate using all 5(8038) data values. On the right is the value of test statistic S_p as a function of bandwidth for the differenced rat data. The vertical line indicates the bandwidth selected by the rule (7).

For a given gene, the log(expression level) varies from rat to rat, and potentially this distribution could differ from gene to gene. Initially we test for equality of these 8038 distributions. Figure 1 gives an impression of how kernel density estimates differ from gene to gene. The most obvious differences are in terms of location, which seem substantial enough that any reasonable test should reject equality of distributions. Indeed our test does find significance, with the value of S_p equal to 280.07, leading to a P-value of essentially zero.

What is not so clear from Figure 1 is whether there are differences between distributions other than ones due to location. For example, it would be of interest to know if there are differences in scale. Hart and Cañette (2011) devised a rank-based test specifically for detecting scale differences in the current setting, applied it to the rat data, and found strong evidence in favor of scale differences. So, there are significant non-location differences among distributions, and it is of interest to see if some version of our kernel-based test can detect them.

We have $X_{ij} = \mu_i + \epsilon_{ij}$, where μ_i is some convenient location parameter of f_i, and f_i is the density from which $X_{i1}, \ldots, X_{in}$ are drawn. For the sake of concreteness we take μ_i to be the mean of f_i. So, $E(\epsilon_{ij}|f_i) = 0$, and we wish to test the null hypothesis that the distribution of ϵ_{ij} is the same for all i. We can eliminate the effect of location by computing differences. For example,

$$\delta_{i12} = X_{i1} - X_{i2} = \epsilon_{i1} - \epsilon_{i2}.$$

When $n \geq 4$, there are at least two independent differences in each small data set. We may thus apply our method to test the null hypothesis that the distribution of δ_{ijk} is the same for all i. Unfortunately, this hypothesis is not equivalent to the null hypothesis that the distribution of ϵ_{ij} is the same for all i. However, if the distribution of, say, δ_{1jk} is different from that of δ_{2jk}, then it must be true that the distributions of ϵ_{1j} and ϵ_{2j} are different. The only deficit of the procedure

is that there exist some exceptional cases where the distributions of ϵ_{1j} and ϵ_{2j} are different, but those of δ_{1jk} and δ_{2jk} are the same. In those cases the power of the test would equal its level.

For each of the 8038 small data sets, we randomly selected one of the 15 ways in which two independent differences can be formed, and applied our test to the resulting set of differences. In this case $n = 2$, and the test statistic S_p was 9.967, yielding a P-value of essentially zero. We have thus found differences in distributions other than ones of location type.

Both tests in this section used the bandwidth (7), which in the case of the differences-based test was 0.378. The effect of h on the test was considered by computing the test statistic S_p for a grid of bandwidths between 0.05 and 0.5; the results are seen in Figure 1. This plot is similar to the significance trace proposed by Young and Bowman (1995). They suggested that a P-value be computed as a function of h. This function, termed the significance trace, is definitive if it lies completely above or below the nominal α. We have plotted the test statistic rather than P-value since the latter quantity is less than machine precision for each h. There is overwhelming evidence to reject H_0 regardless of which bandwidth is used.

7. DEALING WITH UNEQUAL SAMPLE SIZES

In practice sample sizes will often differ. The form of our test statistic is readily modified to account for this situation. Let the sample sizes for the p data sets be $n_1, \ldots, n_p$, each of which is assumed to be at least 2, and define

$$\tilde{S}_W = \frac{1}{(M-N)2^{1/2}h} \sum_{i=1}^{p} \sum_{j=1}^{n_i} \sum_{l=1, l\neq j}^{n_i} \phi\left(\frac{X_{ij} - X_{il}}{2^{1/2}h}\right),$$

$$\tilde{S}_B = \frac{1}{(N^2-M)2^{1/2}h} \sum_{i=1}^{p} \sum_{k=1, k\neq i}^{p} \sum_{j=1}^{n_i} \sum_{l=1}^{n_k} \phi\left(\frac{X_{ij} - X_{kl}}{2^{1/2}h}\right),$$

where $N = \sum_{i=1}^{p} n_i$ and $M = \sum_{i=1}^{p} n_i^2$. It is still true that $E(\tilde{S}_W - \tilde{S}_B) = 0$ under the null hypothesis. We may thus use as test statistic $p^{1/2}(\tilde{S}_W - \tilde{S}_B)/\tilde{\sigma}$, where, as described in our Supplementary Material, $\tilde{\sigma}$ is a suitably modified version of $\hat{\sigma}$.

In accordance with the notion that individual sample sizes are small, we assume that the n_is are bounded for all i and p. This means that there exist distinct sample sizes $m_1 < \cdots < m_J$ that comprise all the sample sizes that will occur as $p \to \infty$. The following condition is used in establishing an asymptotic normality result for our statistic when sample sizes are unequal.

Condition 3. The only possible samples sizes are $m_1 < \cdots < m_J$. Let p_i be the number of data sets having sample size m_i ($i = 1, \ldots, J$). Then, for $i = 1, \ldots, J$, p_i/p tends to π_i as $p \to \infty$, where each $\pi_i > 0$.

The following theorem is proven in the Supplementary Material.

THEOREM 4. *Let h be fixed for all p, suppose that the sample sizes satisfy Condition* 3*, and let X_{ij} ($j = 1, \ldots, n_i$; $i = 1, \ldots, p$) be independent and identically distributed with common density f. Then $p^{1/2}(\tilde{S}_W - \tilde{S}_B)/\tilde{\sigma}$ converges in distribution to a random variable having the standard normal distribution as p tends to ∞.*

8. CONCLUDING REMARKS

A number of different avenues for building on our methodology have presented themselves during the course of our study. One that should be reasonably straightforward is to multivariate data. We have also thought of alternative definitions of the test statistic. One possibility is to use a Kullback–Leibler discrepancy, or log-likelihood ratio:

$$\sum_{i=1}^{p}\sum_{j=1}^{n}\left\{\log \hat{f}_i(X_{ij}) - \log \hat{f}(X_{ij})\right\}. \tag{9}$$

The main reason we have not pursued such a statistic is that it is less amenable to asymptotic analysis than the one we considered. Nonetheless, it would be interesting to investigate how (9) compares with S_p in terms of power.

Further study on the effect of bandwidth is also of interest. Simulations have shown that in some cases fairly substantial gains in power can be obtained by using the right fixed bandwidth instead of (7). The simulation of Section 6.2 indicates promise for the maximizer of $S_p(h)$ with respect to h, and so this test also deserves further study.

Finally, it is of interest to develop a procedure that indicates the main reason for significance of our test. In our microarray example, it seemed plausible from an examination of the data that location differences were the main reason for significance. Ideally, though, one would like to be able to quantify this notion. A way in which this has been done previously for an omnibus statistic is to write it as a sum of components; see, for example, Durbin and Knott (1972) and Parr and Schucany (1982). Since our statistic is of L_2 form, it is possible, using Parseval's formula, to write it as a sum of components corresponding to different orthogonal functions. The most important part of such an approach is finding a basis for which the components have meaningful interpretations.

ACKNOWLEDGEMENT

The authors thank an associate editor and three referees for many helpful comments and Professors Raymond J. Carroll and Robert Chapkin for allowing use of their data. The work of Professor Hart was supported by the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional technical details.

APPENDIX

Proof of Theorem 1

With $X_i = (X_{i1}, \ldots, X_{in})$ for $i = 1, \ldots, p$, we may write $S_W = \sum_{i=1}^{p} h_1(X_i)/p$, which is just a mean of independent and identically distributed random variables. We may also write

$$S_B = \frac{1}{p(p-1)}\sum_{i=1}^{p}\sum_{k\neq i} h_2(X_i, X_k),$$

which is a multivariate U-statistic. When the bandwidth h is fixed, S_W and S_B have all moments finite since each is a bounded random variable.

From Serfling (1980), Section 5.3.1, the projection of S_B is defined as

$$\hat{S}_B = \sum_{r=1}^{p} E(S_B|X_r) - (p-1)E(S_B), \tag{A1}$$

where $E(S_B) = E[h_2(X_i, X_k)] = \theta$. An important fact (Serfling, 1980) is that $p^{1/2}(\hat{S}_B - S_B) = o_p(1)$, which implies that $S_W - S_B$ and $S_W - \hat{S}_B$ have the same asymptotic distribution. So it suffices to find $\mathrm{var}(S_W - \hat{S}_B)$, which is easier to find than $\mathrm{var}(S_W - S_B)$, since $\hat{S}_B$ is a sum of independent and identically distributed random variables.

Define, for $i \neq k$ and each x,

$$h_3(x) = E\left\{h_2(X_i, X_k)|X_k = x\right\} = E\{h_2(X_i, x)\}. \tag{A2}$$

Using equation (2), p. 188 of Serfling (1980),

$$\hat{S}_B = \frac{2}{p}\sum_{i=1}^{p} h_3(X_i) - \theta, \tag{A3}$$

which, as promised, is a sum of independent and identically distributed random variables.

Defining $\hat{S} = S_W - \hat{S}_B$, we have

$$\mathrm{var}(\hat{S}) = \frac{1}{p}\mathrm{var}\left\{h_1(X_i) - 2h_3(X_i)\right\} \equiv \frac{1}{p}\sigma^2. \tag{A4}$$

From the central limit theorem it follows that $p^{1/2}\hat{S}/\sigma$ converges in distribution to a standard normal random variable.

Proof of Lemma 1

For $l \neq j$ and $i \neq k$, we have

$$\begin{aligned}
\mu &= \frac{1}{2^{1/2}h}\left\{E\phi\left(\frac{X_{ij} - X_{il}}{2^{1/2}h}\right) - E\phi\left(\frac{X_{ij} - X_{kl}}{2^{1/2}h}\right)\right\} \\
&= \frac{1}{2^{1/2}h}\left\{\sum_{r=1}^{\infty}\rho_r \int\int \phi\left(\frac{x-y}{2^{1/2}h}\right) g_r(x)g_r(y)\,dxdy - \int\int \phi\left(\frac{x-y}{2^{1/2}h}\right) m(x)m(y)\,dxdy\right\} \\
&= \frac{1}{2^{1/2}h}\int\int \phi\left(\frac{x-y}{2^{1/2}h}\right)\left\{\sum_{r=1}^{\infty}\rho_r g_r(x)g_r(y)\sum_{s=1}^{\infty}\rho_s - \sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\rho_r\rho_s g_r(x)g_s(y)\right\} dxdy \\
&= \frac{1}{2^{1/2}h}\int\int \phi\left(\frac{x-y}{2^{1/2}h}\right)\sum_{r=1}^{\infty}\sum_{s=1}^{\infty}\rho_r\rho_s g_r(x)\left\{g_r(y) - g_s(y)\right\} dxdy \\
&= \frac{1}{2^{1/2}h}\int\int \phi\left(\frac{x-y}{2^{1/2}h}\right)\sum_{r=1}^{\infty}\sum_{s>r}\rho_r\rho_s\{g_r(x) - g_s(x)\}\{g_r(y) - g_s(y)\}\,dxdy.
\end{aligned} \tag{A5}$$

It is straightforward to show that

$$\frac{1}{2^{1/2}h}\int\int \phi\left(\frac{x-y}{2^{1/2}h}\right)\{g_r(x) - g_s(x)\}\{g_r(y) - g_s(y)\}\,dxdy = \int\{g_r(x;h) - g_s(x;h)\}^2\,dx,$$

thus establishing the first part of Lemma 1. To establish that $\mu > 0$, it is enough to show that one of the summands in μ is positive. By assumption, there exist r and s such that $\rho_r\rho_s > 0$ and g_r and g_s have different characteristic functions, call them ψ_r and ψ_s. Plancherel's formula entails that

$$\int\{g_r(x;h) - g_s(x;h)\}^2\,dx = \frac{1}{2\pi}\int e^{-h^2t^2/2}|\psi_r(t) - \psi_s(t)|^2\,dt.$$

By the continuity of characteristic functions, there exists an interval throughout which $|\psi_r(t) - \psi_s(t)| > 0$, and so the last integral is positive.

Proof of Theorem 2

The vectors $X_1, \ldots, X_p$ are independent and identically distributed with common density $m(x_1, \ldots, x_n) = \sum_{r=1}^{\infty} \rho_r \prod_{j=1}^{n} g_r(x_j)$. Therefore, the same proof as that for Theorem 1 implies that $p^{1/2}(S - \mu)/\sigma$ converges in distribution to a standard normal random variable, where $\sigma^2 > 0$ is the variance of $h_1(X_1) - 2h_3(X_1)$. In the Supplementary Material it is shown that $\hat{\sigma}$ converges in probability to σ. Using (4) and the assumption $\mu > 0$, consistency of the test follows.

Proof of Theorem 3

Because the alternative hypothesis is no longer fixed, one must use a central limit theorem that allows a triangular array structure. For this reason we denote the observed n-vectors by $X_{p1}, \ldots, X_{pp}$. These vectors are independent and identically distributed for any given p, but their common distribution changes with p.

The quantity $S_W - S_B$ may still be approximated by its U-statistic projection, since our local alternatives do not affect the property that the difference between $S_W - S_B$ and its projection is $o_p(p^{-1/2})$. For each p, define $\delta_p = h_1(X_{p1}) - 2h_3(X_{p1})$. Using Liapounov's central limit theorem (Chung 1974, pp. 196-200), our asymptotic normality result is established by verifying that

$$\lim_{p \to \infty} \frac{E\{|\delta_p - E(\delta_p)|^3\}}{p^{1/2} \mathrm{var}(\delta_p)^{3/2}} = 0. \tag{A6}$$

Using the fact that each of h_1 and h_3 is bounded by $\phi(0)/(2^{1/2}h)$, it follows that $E\{|\delta_p - E(\delta_p)|^3\}$ is bounded by some constant C for all p. In the Supplementary Material it is shown that $\mathrm{var}(\delta_p)$ converges to $\sigma^2 = \mathrm{var}\{h_1(X) - 2\tilde{h}_3(X)\}$ as $p \to \infty$, where the n components of X are independent and identically distributed as g_1, and, for each x, $\tilde{h}_3(x) = E\{h_2(X, x)\}$. It is assumed that $\sigma^2 > 0$, and hence (A6) is verified.

For the rest of the proof, first use (5) to obtain

$$\begin{aligned}
\mu &= \sum_{r=1}^{\infty} \sum_{s>r} \rho_{rp}\rho_{sp} \int \{g_r(x;h) - g_s(x;h)\}^2 \, dx \\
&= \rho_{1p} \sum_{s=2}^{\infty} \rho_{sp} \int \{g_1(x;h) - g_s(x;h)\}^2 \, dx + \sum_{r=2}^{\infty} \sum_{s>r} \rho_{rp}\rho_{sp} \int \{g_r(x;h) - g_s(x;h)\}^2 \, dx \\
&= \epsilon_p \sum_{s=2}^{\infty} a_s \int \{g_1(x;h) - g_s(x;h)\}^2 \, dx + O_p(\epsilon_p^2).
\end{aligned} \tag{A7}$$

Expression (6) now follows from (4), the asymptotic normality of S, (A7), and the fact, as shown in the Supplementary Material, that $\hat{\sigma}$ converges in probability to σ.

REFERENCES

ANDERSON, T. W. (1962). On the distribution of the two-sample Cramér–von Mises criterion. *Ann. Math. Statist.* **33**, 1148–1159.

ANDERSON, T. W. & DARLING, D. A. (1954). A test of goodness-of-fit. *J. Am. Statist. Assoc.* **49**, 765–769.

ANDERSON, N. H., HALL, P. & TITTERINGTON, D. M. (1994). Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Mult. Anal.* **50**, 41–54.

ARVESEN, J. N. (1969). Jackknifing U-statistics. *Ann. Math. Statist.* **40**, 2076–2100.

CAO, R. & VAN KEILEGOM, I. (2006). Empirical likelihood tests for two-sample problems via nonparametric density estimation. *Can. J. Statist.* **34**, 61–67.

CHUNG, K. L. (1974). *A Course in Probability Theory*. New York: Academic Press, Inc.

COX, D. R. & SOLOMON, P. J. (1986). Analysis of variability with large numbers of small samples. *Biometrika* **73**, 543–554.

COX, D. R. & SOLOMON, P. J. (1988). On testing for serial correlation in large numbers of small samples. *Biometrika* **75**, 145–148.

DAVIDSON, L. A., NGUYEN, D. V., HOKANSON, R. M., CALLAWAY, E. S., ISETT, R. B., TURNER, N. D., DOUGHERTY, E. R., WANG, N., LUPTON, J. R., CARROLL, R. J. & CHAPKIN, R. S.(2004). Chemopreventive n-3 polyunsaturated fatty acids reprogram genetic signatures during colon cancer initiation and progression in the rat. *Cancer Research* **64**, 6797–6804.

DOKSUM, K. & SCHAFER, C. (2006). Powerful choices: tuning parameter selection based on power. In *Frontiers in Statistics*, H. L. Koul & J. Fan, eds. London: Imperial College Press.

DURBIN, J. & KNOTT, M. (1972). Components of Cramér-von Mises statistics. *J. R. Statist. Soc. B* **34**, 290–307.

EFRON, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Am. Statist. Assoc.* **99**, 96–104.

HART, J. D. & CAÑETTE, I. (2011). Nonparametric estimation of distributions in random effects models. *J. Comp. Graph. Statist.* **20**, 461–478.

JIMENEZ-GAMERO, M. D., ALBA-FERNANDEZ, V., MUÑOZ-GARCIA, J. & CHALCO-CANO, Y. (2009). Goodness-of-fit tests based on empirical characteristic functions. *Comp. Statist. Data Anal.* **53**, 3957–3971.

KIEFER, J. (1959). k-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises test. *Ann. Math. Statist.* **30**, 420–447.

KOREN, A., TIROSH, I. & BARKAI, N. (2007). Autocorrelation analysis reveals widespread spatial biases in microarray experiments. *BMC Genomics* **8**, 164.

KULASEKERA, K. B. & WANG, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *J. Am. Statist. Assoc.* **92**, 500–511.

LOUANI, D. (2000). Large deviation for L_1-distance in kernel density estimation. *J. Statist. Plan. Infer.* **90**, 177–182.

MARTÍNEZ-CAMBLOR, P., DE UÑA ÁLVAREZ, J. & CORRAL, N. (2008). k-Sample test based on the common area of kernel density estimator. *J. Statist. Plan. Infer.* **138**, 4006–4020.

MARTÍNEZ-CAMBLOR, P. & DE UÑA ÁLVAREZ, J. (2009). Non-parametric k-sample tests: density functions vs distribution functions. *Comp. Statist. Data Anal.* **53**, 3344–3357.

MARTÍNEZ-CAMBLOR, P. & DE UÑA ÁLVAREZ, J. (2013). Studying the bandwidth in k-sample smooth tests. *Comp. Statist.* **28**, 875-892.

PARK, J. & PARK, D. (2012). Testing the equality of a large number of normal population means. *Comp. Statist. Data Anal.* **56**, 1131–1149.

PARR, W. C. & SCHUCANY, W. R. (1982). Minimum distance estimation and components of goodness-of-fit statistics. *J. R. Statist. Soc. B* **44**, 178–189.

SCHOLZ, F. W. & STEPHENS, M. A. (1987). k-sample Anderson–Darling test. *J. Am. Statist. Assoc.* **82**, 919–924.

SCHUCANY, W. R. & BANKSON, D. M. (1989). Small sample variance estimators for U-statistics. *Austral. J. Statist.* **31**, 417–426.

SEN, P. K. (1960). On some convergence properties of U-statistics. *Calcutta Statist. Assoc. Bull.* **10**, 1–18.

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

STEPHENS, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *J. Am. Statist. Assoc.* **69**, 730–737.

STEPHENS, M. A. (1986). Tests based on EDF statistics. In *Goodness of Fit Techniques*, R. B. D'Agostino & M. A. Stephens, eds. New York: Marcel Dekker, Inc.

TERRELL, G. R. (1990). The maximal smoothing principle in density estimation. *J. Am. Statist. Assoc.* **85**, 470–477.

TERRELL, G. R. & SCOTT, D. W. (1985). Oversmoothed nonparametric density estimates. *J. Am. Statist. Assoc.* **85**, 209–214.

YOUNG, S. G. & BOWMAN, A. W. (1995). Non-parametric analysis of covariance. *Biometrics* **51**, 920–931.

[*Received September* 2012. *Revised December* 2012]