Two-Sample Test Statistics for Measuring Discrepancies between Two Multivariate Probability Density Functions Using Kernel-Based Density Estimates

NIALL H. ANDERSON

University of Glasgow, Scotland

PETER HALL

Australian National University, Canberra, Australia, and CSIRO Division of Mathematics and Statistics, Sydney, Australia

AND

D. M. TITTERINGTON

University of Glasgow, Scotland

Test statistics are proposed for testing equality of two p-variate probability density functions. The statistics are based on the integrated square distance between two kernel-based density estimates and are two-sample versions of the statistic studied by Hall (1984, *J. Multivariate Anal.* **14** 1-16). Particular emphasis is laid on the case where the two bandwidths are fixed and equal. Asymptotic distributional results and power calculations are supplemented by an empirical study based on univariate examples. © 1994 Academic Press, Inc.

1. INTRODUCTION

Hall (1984) analyses the integrated squared error between a kernel-based density estimate of a multivariate probability density function (pdf) and the true pdf itself. In an obvious, brief notation, the quantity of interest is

$$I=\int (\hat{f}_h-f)^2,$$

where $\hat{f}_h$ denotes the density estimate, h denotes the associated bandwidth, and f is the true pdf. In particular, Hall (1984) derives a central limit

Received June 22, 1992; revised October 18, 1993.

AMS 1985 *subject classifications*: primary 62G10; secondary 62E40.

Key words and phrases: bootstrap, density estimation, integrated square distance, kernel method, nonparametric two-sample test.

0047-259X/94 $6.00

theorem for I which relies, among other things, on conditions imposed on the bandwidth used within $\hat{f}_h$. From a practical point of view, I is a natural test statistic for a significance test against the hypothesis that f is indeed the correct pdf.

The objective of the present paper is to investigate two-sample versions of I; the most natural statistic is

$$T_{h_1h_2} = \int (\hat{f}_1 - \hat{f}_2)^2, \tag{1.1}$$

in which, for $j = 1, 2$, $\hat{f}_j$ is a density estimate based on the jth sample and using smoothing parameter h_j.

In the main section, Section 2, we concentrate on the case where the kernel function underlying the density estimates is a spherically symmetric p-variate density and our point of departure is the (unweighted) statistic $T_{h_1h_2}$. Some generalizations are described in Section 3.

2. Theory and Applications of The two-Sample Statistics Based on a Spherically Symmetric Kernel Function

2.1. *Introduction*

As indicated at the end of Section 1, we concentrate here on the case where, given two independent random samples $\{X_{j1}, ..., X_{jn_j}\}$, $j = 1, 2$, from p-variate distributions with densities f_j, $j = 1, 2$, we estimate f_j by

$$\hat{f}_j(x) = (n_j h_j^p)^{-1} \sum_{i=1}^{n_j} K\{(x - X_{ji})/h_j\}, \qquad j = 1, 2,$$

in which h_j is a bandwidth and K is a spherically symmetric p-variate density.

In Section 2.2 we examine $T_{h_1h_2}$, given in (1.1), as a statistic for testing the hypothesis $f_1 = f_2$. It turns out to be convenient to concentrate on the case $h_1 = h_2 = h$ (with $T_{h_1h_2}$ then written as T_h), and to show that the minimum distance at which the statistic can discriminate between f_1 and f_2 is $\delta = n^{-1/2}h^{-p/2}$, where $n = n_1 + n_2$, and n_1 and n_2 are assumed to be of the same order of magnitude. The results suggest that the bandwidth h should be fixed if close alternatives are to be distinguished. Section 2.3 derives the theory related to the case $h = 1$, with the notation $T_1 \equiv T$. It turns out that T is not now asymptotically Normal. Section 2.3 also suggests an alternative test statistic, U, defined in (2.9), the practical implementation of which is easier than it is for T. Section 2.4 investigates the power of tests based on U, Section 2.5 discusses the implementation of the bootstrap, and

Section 2.6 illustrates the relative performance of T and U by carrying out an empirical bootstrap study for simple univariate examples.

2.2. *Basic Asymptotic Results*

We begin our discussion by examining

$$T_{h_1h_2} = \int (\hat{f}_1 - \hat{f}_2)^2,$$

where $\hat{f}_1$ and $\hat{f}_2$ are as defined at the beginning of Section 2.1.

Under the hypothesis that $f_1 = f_2$, and assuming that f_1 has two continuous, square-integrable derivatives, we may show that

$$E(T_{h_1h_2}) = \kappa_1\{(n_1h_1)^{-1} + (n_2h_2)^{-1}\} + \kappa_2(h_1^2 - h_2^2)^2 \int (\nabla^2 f_1)^2 + O(n_1^{-1} + n_2^{-1}) + o(|h_1^2 - h_2^2|^2), \tag{2.1}$$

where $\kappa_1 = \int K^2$, $\kappa_2 = \frac{1}{2} \int (z^{(1)})^2 K(z)\, dz$, and ∇^2 is the Laplacian operator. In the event that h_1 and h_2 are different, say $h_1/h_2 \to c \neq 1$ as $n_1, n_2 \to \infty$, it is generally necessary to estimate $\int (\nabla^2 f_1)^2$, and even the smaller order terms represented by $o(|h_1^2 - h_2^2|^2)$ in (2.1), in order properly to centre the test statistic $T_{h_1h_2}$. While this can be done root-n consistently, subject to appropriate smoothness conditions on f, it is nevertheless cumbersome. Therefore we assume that $h_1 = h_2 = h$, say, and base our test directly on $T_{h_1h_2} = T_h$, without any empirical centring.

In order to assess the power of a test based on T_h we should ascertain its performance against a local alternative hypothesis. To this end, let $f_1 = f$ denote a fixed density, let g be a function such that $f_2 = f + \delta g$ is a density for all sufficiently small $|\delta|$, and let t_h denote the α-level critical point of the distribution of T_h under the null hypothesis H_0 that $\delta = 0$:

$$P_{H_0}(T_h > t_h) = \alpha.$$

Our test consists of rejecting H_0 if $T_h > t_h$. We claim that, if h is chosen to converge to zero as $n_1, n_2 \to \infty$ (which is necessary if $\hat{f}_j$ is consistently to estimate f_j), then the minimum distance at which the test can discriminate between f_1 and f_2 is $\delta = n^{-1/2}h^{-p/2}$.

This claim may be formalized as follows. Let $H_1 = H_1(c)$ denote the alternative hypothesis that $\delta = n^{-1/2}h^{-p/2}c$, where $c \neq 0$, and define

$$\pi(c) = \lim_{n \to \infty} P_{H_1}(T_h > t_h).$$

We shall show below that this limit is well-defined, that $\alpha < \pi(c) < 1$ for $0 < |c| < \infty$, and that $\pi(c) \to 1$ as $|c| \to \infty$.

To verify the claims above, observe first that

$$\begin{aligned} T_h = & \int \{\hat{f}_1 - \hat{f}_2 - E_{H_1}(\hat{f}_1 - \hat{f}_2)\}^2 \\ & + 2 \int \{\hat{f}_1 - \hat{f}_2 - E_{H_1}(\hat{f}_1 - \hat{f}_2)\}\, E_{H_1}(\hat{f}_1 - \hat{f}_2) \\ & + \int \{E_{H_1}(\hat{f}_1 - \hat{f}_2)\}^2, \end{aligned}$$

and $\int \{E_{H_1}(\hat{f}_1 - \hat{f}_2)\}^2 \sim \delta^2 \int g^2$. Arguing as in Hall (1984), we may show that if n_1, $n_2 \to \infty$ such that n_1/n_2 is bounded away from zero and infinity, and if $h \to 0$ and $nh^p \to \infty$, then under H_1,

$$nh^{p/2} \left[\int \{\hat{f}_1 - \hat{f}_2 - E_{H_1}(\hat{f}_1 - \hat{f}_2)\}^2 - \kappa_1(n_1^{-1} + n_2^{-1})\, h^{-1} \right],$$

$$n^{1/2}\delta^{-1} \int \{\hat{f}_1 - \hat{f}_2 - E_{H_1}(\hat{f}_1 - \hat{f}_2)\}\, E_{H_1}(\hat{f}_1 - \hat{f}_2)$$

are asymptotically independent and normally distributed with zero means and finite, nonzero variances, the latter not depending on c. Therefore, if $\delta = n^{-1/2} h^{-p/2} c$, then under H_1,

$$nh^{p/2} \left[T_h - \kappa_1(n_1^{-1} + n_2^{-1}) h^{-1} - \{1 + o(1)\}\, \delta^2 \int g^2 \right]$$

is asymptotically normally distributed with zero mean and finite, nonzero variance $\sigma^2(c)$, the latter being an increasing function of c. The claims made about $\pi(c)$ in the previous paragraph follow directly from this result.

2.3. *The Case of Fixed Bandwidth*

The results in the previous subsection suggest that, in order to discriminate between distributions distant only $n^{-1/2}$ apart, we should fix the bandwidth h. Without loss of generality we take $h = 1$.

Define

$$a_j(x) = \int K(x - y)\, f_j(y)\, dy, \qquad j = 1, 2.$$

Then the squared distance between a_1 and a_2, $I(f_1, f_2) = \int (a_1 - a_2)^2$, is a measure of the distance of f_1 from f_2. By Parseval's identity,

$$I = (2\pi)^{-p} \int |\check{K}(\check{f}_1 - \check{f}_2)|^2,$$

where $\check{f}_1$, $\check{f}_2$, and $\check{K}$ denote the Fourier transforms of f_1, f_2, and K, respectively. We may deduce from this result that, provided $\check{K}$ does not vanish on an interval, $I(\cdot, \cdot)^{1/2}$ is a metric on the class of all densities.

This observation motivates the following regularity condition on K:

$$K \text{ is bounded, absolutely integrable, and has a Fourier transform which does not vanish on any interval.} \tag{2.2}$$

The kernel $K(x) = (x^{-1} \sin x)^k$, for any integer $k \geqslant 2$, is bounded and integrable but has a Fourier transform which vanishes outside $(-\frac{1}{2}k, \frac{1}{2}k)$; it would be unsuitable for our purposes. On the other hand, functions such as the p-variate uniform and standard normal densities, and a p-variate form of Epanechnikov's kernel (Silverman, 1986, p. 76) all satisfy (2.2). For most purposes the condition of integrability may be dropped from (2.2); for example, if $K(x)$ denotes the indicator function of the semi-infinite rectangle $\pi(-\infty, x^{(j)}]$ then the test which we shall propose below is a p-variate version of the familiar Kolmogorov–Smirnov test. However, without the assumption of integrability the Fourier transform of K is not well-defined, and then it is awkward to characterize the class of K's for which $I(\cdot, \cdot)$ separates densities (e.g., such that $I(\cdot, \cdot)^{1/2}$ is a metric).

As in Subsection 2.1, let $\{X_{j1}, ..., X_{jn_j}\}$, $j = 1, 2$, represent random samples from distributions with densities f_j, $j = 1, 2$. An unbiased estimator of $a_j(x)$ is given by

$$\hat{a}_j(x) = n_j^{-1} \sum_{i=1}^{n_j} K(x - X_{ji}),$$

motivating the test statistic

$$T = \int (\hat{a}_1 - \hat{a}_2)^2. \tag{2.3}$$

We could include a weight function in the integrand. Theory and practice for that case are virtually identical to those for the basic statistic defined in (2.3).

Under the null hypothesis H_0 that $f_1 = f_2$, the expected value and asymptotic variance of T are given by

$$E_{H_0}(T) = (n_1^{-1} + n_2^{-1}) J_1, \qquad \operatorname{var}_{H_0}(T) \sim (n_1^{-1} + n_2^{-1})^2 J_2,$$

where it is assumed that $n_1, n_2 \to \infty$ such that n_1/n_2 is bounded away from zero and infinity, and where

$$J_1 = \kappa_1 - \int a^2, \qquad J_2 = \iint M(x_1, x_2)^2 f(x_1) f(x_2)\, dx_1\, dx_2, \qquad a = a_1,$$
$$M(x_1, x_2) = \int \{K(x - x_1) - a(x)\}\{K(x - x_2) - a(x)\}\, dx. \tag{2.4}$$

If K is bounded and integrable then M is bounded, and so $J_2 < \infty$.

However, T is not asymptotically normally distributed. As a prelude to describing the asymptotic distribution of T we first note that we may represent M as an orthogonal expansion in its eigenfunctions with respect to the weight f,

$$M(x_1, x_2) = \sum_{k=1}^{\infty} \lambda_k \omega_k(x_1)\, \omega_k(x_2), \tag{2.5}$$

where

$$\int M(x_1, x_2)\, \omega_k(x_1)\, f(x_1)\, dx_1 = \lambda_k \omega_k(x_2), \tag{2.6}$$

$$\int \omega_k(x)\, \omega_l(x)\, f(x)\, dx = \delta_{kl}, \tag{2.7}$$

(the Kronecker delta). The expansion in (2.5) converges in L^2, in the sense that

$$\iint \left\{M(x_1, x_2) - \sum_{k=1}^{m} \lambda_k \omega_k(x_1)\, \omega_k(x_2)\right\}^2 f(x_1)\, f(x_2)\, dx_1\, dx_2 \to 0$$

as $m \to \infty$. See Indritz (1963, p. 209ff), and also Neuhaus (1977) and Hall (1979). Note too that in the notation of (2.5), $J_2 = \sum \lambda_k^2$.

Let $Z_{11}, Z_{12}, ..., Z_{21}, Z_{22}, ...$ denote independent standard normal random variables. Assume that for a parameter n diverging to $+\infty$ we have $n_j = n_j(n) \sim \rho_j n$, where $0 < \rho_j < \infty$. We shall show in the Appendix that

$$\begin{aligned} n\{T - (n_1^{-1} + n_2^{-1})\, J_1\} &\to S \\ &= \sum_{k=1}^{\infty} \lambda_k \{(\rho_1^{-1/2} Z_{1k} - \rho_2^{-1/2} Z_{2k})^2 - (\rho_1^{-1} + \rho_2^{-1})\} \end{aligned} \tag{2.8}$$

in distribution.

For practical implementation of the test based on T it is necessary to estimate the centring constant, $(n_1^{-1} + n_2^{-1})\, J_1$, and hence to estimate J_1. It is also of interest to estimate the asymptotic variance under H_0, i.e., to estimate J_2. Observe that $J_0 = \int a^2$ and J_2 are estimated root-n consistently by

$$\hat{J}_{0j} = \{n_j(n_j - 1)\}^{-1} \sum_{i_1 \neq i_2}\sum L(X_{ji_1} - X_{ji_2}),$$

$$\begin{aligned} \hat{J}_{2j} = {} & \{n_j(n_j - 1)\}^{-1} \sum_{i_1 \neq i_2}\sum L(X_{ji_1} - X_{ji_2})^2 - 2\{n_j(n_j - 1)(n_j - 2)\}^{-1} \\ & \times \sum_{i_1 \neq i_2 \neq i_3 \neq i_1}\sum\sum L(X_{ji_1} - X_{ji_2})\, L(X_{ji_2} - X_{ji_3}) + \hat{J}_{0j}^2, \end{aligned}$$

respectively. Alternatively, J_0 and J_2 may be estimated from the pooled sample, $\mathscr{X} = \{X_{11}, ..., X_{1n_1}, X_{21}, ..., X_{2n_2}\}$. Let $\hat{J}_k$ denote any one of the three possible etimators of J_k (based on either one of the individual samples, or on the pooled sample), for $k=0$, 2. Put $\hat{J}_1 = \kappa_1 - \hat{J}_0$. Then $\hat{J}_k = J_k + O_p(n^{-1/2})$, and so under H_0, noting (2.8),

$$U \equiv n\{T - (n_1^{-1} + n_2^{-1})\hat{J}_1\} \to S \tag{2.9}$$

in distribution. An asymptotic statistical test may be based on the value of U, by rejecting the null hypothesis if U exceeds the appropriate critical point. However, the distribution of S depends on the unknowns $\lambda_1, \lambda_2, ...,$ and the larger values of these quantities would have to be estimated. Therefore it seems that a more practical approach is to consider bootstrap methods, discussed in subsection 2.5.

In many instances it is not absolutely essential to centre the statistic T, as done in (2.8) and (2.9). If we were to work directly with T then of course the limit result (2.8) would change to

$$nT \to \sum_{k=1}^{\infty} \lambda_k(\rho_1^{-1/2} Z_{1k} - \rho_2^{-1/2} Z_{2k})^2 \tag{2.10}$$

in distribution. Then we would not need to compute $\hat{J}_1$. However, for this approach to be valid the right-hand side of (2.10) must be well-defined, which virtually requires us to assume that $\sum |\lambda_k| < \infty$. Our present regularity conditions only guarantee the weaker condition $\sum \lambda_k^2 < \infty$, which is sufficient for the series defining S to converge in L^2.

2.4. *Power of Tests Based on U*

We assume that for a parameter n diverging to $+\infty$, we have $n_j = n_j(n) \sim \rho_j n$ where $0 < \rho_j < \infty$. Let H_1 denote the alternative hypothesis that $f_1 = f$ (fixed) and $f_2 = f + \delta g$, where $\delta = n^{-1/2} c$ and $c \neq 0$. Here, g is a fixed, bounded, integrable function with the property that $f + \delta g$ is a density for all sufficiently small $|\delta|$. Define

$$J_3 = \int \left[\int \{K(x-y) - a(x)\}\, g(x)\, dx\right]^2 dy,$$

$$r_k = \iint \{K(x-y) - a(x)\}\, g(x)\, \omega_k(y)\, f(y)\, dx\, dy$$

$$= \iint K(x-y)\, g(x)\, \omega_k(y)\, f(y)\, dx\, dy,$$

the last identity holding if $\lambda_k \neq 0$. On the probability space supporting the Z_{jk}'s, define random variables Y_1, Y_2 such that $E(Y_1) = E(Y_2) = 0$, $E(Y_1^2) = E(Y_2^2) = 4J_3$, $E(Y_1 Y_2) = 0$, $E(Y_j Z_{jk}) = r_k$, $E(Y_1 Z_{2k}) = E(Y_2 Z_{1k}) = 0$, and the Y's and Z's have joint normal distributions. We shall prove in the Appendix that under H_1,

$$U \to S + 2c(\rho_1^{-1/2} Y_1 - \rho_2^{-1/2} Y_2) + c^2 J_4$$

in distribution, where S was defined at (2.3) and $J_4 = \int (K * g)^2$. Therefore, the asymptotic value of the probability that the version of our test with nominal level α correctly rejects H_0 when H_1 is true equals

$$\pi(c) = P\{S + 2c(\rho_1^{-1/2} Y_1 - \rho_2^{-1/2} Y_2) > s_\alpha - c^2 J_4\},$$

where s_α denotes the upper α-level point of the distribution of S. Our assumption at (2.2) that the Fourier transform of K does not vanish on any interval guarantees that $J_4 \neq 0$, and so $\pi(c) \to 1$ as $|c| \to \infty$.

2.5. *Bootstrap*

Since the distribution of S is unknown, bootstrap methods provide an attractive approach to determining a critical point for the test. Resampling may be from either the pooled sample $\mathscr{X} = \{X_{11}, ..., X_{1n_1}, X_{21}, ..., X_{2n_2}\}$ or from one of the individual samples. Using the pooled sample will provide somewhat greater level accuracy when H_0 is true, and obviously there are disadvantages in resampling from one of the individual samples if it is much smaller in size than the other. However, neither of the two approaches to bootstrap resampling appears to have any general advantages from the viewpoint of power.

Since the distribution of S depends on the unknowns λ_k in a manner which does not allow the effect to be removed by simply Studentizing, there seems little point in using sophisticated bootstrap methods such as percentile-t. The Studentized statistic is not even asymptotically pivotal on this occasion. Therefore we suggest a simpler, percentile method. Let $\{X_{j1}^*, ..., X_{jn_j}^*\}$, $j = 1, 2$, denote independent resamples drawn randomly, with replacement, using one of the two approaches. Compute

$$\hat{a}_j^*(x) = n_j^{-1} \sum_{i=1}^{n_j} K(x - X_{ji}^*), \qquad T^* = \int (\hat{a}_1^* - \hat{a}_2^*)^2.$$

Let $\hat{J}_0^*$ denote the version of $\hat{J}_0$ calculated from the resamples rather than the samples, and define $\hat{J}_1^* = \kappa_1 - \hat{J}_0^*$ and

$$U^* = n\{T^* - (n_1^{-1} + n_2^{-1}) \hat{J}_1^*\}.$$

Given $0 < \alpha < 1$, let $\hat{u}_\alpha$ denote the solution of the equation

$$P(U^* > \hat{u}_\alpha \mid \mathscr{X}) = \alpha.$$

A nominal α-level test of H_0 is to reject that hypothesis if $U > \hat{u}_\alpha$.

2.6. *Some Numerical Results*

In this section we report a simulation study that compares the performance of T (from (2.3)) and U (from (2.9)). The study was based on Normal mixture distributions and two cases were considered. In each case, we took

$$f_2(x) = \phi(x; 0, 1),$$

where $\phi(x; \mu, \sigma^2)$ denotes the univariate Normal probability density function corresponding to $X \sim N(\mu, \sigma^2)$, and

$$f_1(x) = (1-p)\,\phi(x; 0, 1) + p\phi(x; 0, \sigma^2),$$

with $p = cn^{-1/2}$. Of the two cases, $\sigma^2 = 2$ in case (a), and $\sigma^2 = 4$ in case (b), so that both cases, and case (a) in particular, are demanding so far as testing is concerned. The Epanechnikov kernel function was used.

Table I displays the empirical powers obtained from bootstrap tests based on T and U, carried out as described in Section 2.5, for various values of n_1, n_2, and c. In each bootstrap test 199 bootstrap resamples were generated, and each quoted value for power was based on 1000 replications. In general and not surprisingly, values of power increase with c and

TABLE I

Empirical Powers for T and U with $f_1(x) = (1-p)\,\phi(x; 0, 1) + p\phi(x; 0, \sigma^2)$, and $p = cn^{-1/2}$; 1000 Replications with 199 Bootstrap Resamples

			c			
1	2	Statistic	1.0	2.0	4.0	6.0
			(a) $\sigma^2 = 2$			
20	60	T	0.076	0.053	0.079	0.131
		U	0.066	0.082	0.107	0.172
50	150	T	0.066	0.072	0.063	0.130
		U	0.054	0.075	0.087	0.132
100	300	T	0.047	0.064	0.071	0.079
		U	0.048	0.054	0.104	0.136
			(b) $\sigma^2 = 4$			
20	60	T	0.068	0.109	0.166	0.334
		U	0.076	0.115	0.259	0.498
50	150	T	0.056	0.064	0.162	0.279
		U	0.058	0.087	0.235	0.480
100	300	T	0.065	0.066	0.107	0.233
		U	0.048	0.094	0.241	0.451

the values for case (b) are typically larger than those for case (a). Also, typically, U turns out to be more powerful than T.

Note that in this paper we are not trying optimally to estimate

$$\int (f_1 - f_2)^2,$$

the population-based measure of which $T_{h_1 h_2}$ might be regarded as an estimate. As a referee has remarked, optimal choices for h_1 and h_2 in $T_{h_1 h_2}$ would then correspond to undersmoothing, relative to the values that would be best for density estimation, in order to compensate for the smoothing effect of the integral. Here, in contrast, relative oversmoothing is appropriate; $T_{h_1 h_2}$ is being used for a totally different purpose.

3. Generalizations

We begin by generalizing our definition of T given in Section 2; see (2.3). First, in our definition of $\hat{a}_j$ we generalize $K(x_1 - x_2)$ to $K(x_1, x_2)$, where K is now a bivariate function of the p-vectors x_1 and x_2. In this notation, $\hat{a}_j$ becomes

$$\hat{a}_j(x) = n_j^{-1} \sum_{i=1}^{n_j} K(x, X_{ji}).$$

Secondly, we incorporate a nonnegative weight function w into our definition of T, as discussed just subsequent to (2.3):

$$T = \int (\hat{a}_1 - \hat{a}_2)^2 \, w. \tag{3.1}$$

If we take K to be symmetric and square-integrable with respect to the weight function, i.e., to satisfy

$$\int K^2(x_1, x_2) \, w(x_1) \, w(x_2) \, dx_1 \, dx_2 < \infty, \tag{3.2}$$

then we may represent K as an orthogonal expansion with respect to w,

$$K(x_1, x_2) = \sum_{k=1}^{\infty} v_k \omega_k(x_1) \, \omega_k(x_2),$$

where

$$\int K(x_1, x_2) \, \omega_k(x_1) \, w(x_1) \, dx_1 = v_k \omega_k(x_2), \qquad \int \omega_k \omega_l w = \delta_{kl}.$$

In this notation,

$$T = \sum_{k=1}^{\infty} v_k^2 \left\{ n_1^{-1} \sum_{i=1}^{n_1} \omega_k(X_{1i}) - n_2^{-1} \sum_{i=1}^{n} \omega_k(X_{1i}) \right\}^2. \tag{3.3}$$

Formula (3.3) makes it explicitly clear that T is based on comparing the means of the eigenfunctions ω_k, i.e., comparing $E\omega_k(X_{11})$ and $E\omega_k(X_{21})$. The weights v_k^2 apportion the amount of emphasis we give to individual eigenfunctions. The analogue of v_k^2 in Section 2 is λ_k. This is perhaps clearer from an asymptotic analysis, as follows. If n_1, $n_2 \to \infty$ in such a way that $n_j \sim \rho_j n$ then, under the null hypothesis that $f_1 = f_2 = f$,

$$nT \to S' = \sum_{k=1}^{\infty} v_k^2 (\rho_1^{-1/2} Z_{1k} - \rho_2^{-1/2} Z_{2k})^2$$

in distribution, where $\{Z_{jk},\ j = 1, 2,\ k \geqslant 1\}$ denote jointly normally distributed random variables with zero means and covariances $E(Z_{jk} Z_{jl}) = \int \omega_k \omega_l f$, $E(Z_{1k} X_{2l}) = 0$. Note that condition (3.2) guarantees $\sum v_k^2 < \infty$, and so S' is well-defined.

We may test H_0: $f_1 = f_2$ by rejecting this hypothesis if the value of T is too large. Again, the distribution of T may be approximated using bootstrap methods, much as discussed in Subsection 2.5. Let $\mathscr{X}$ and $\{X_{j1}^*, ..., X_{jn_j}^*\}$, $j = 1, 2$, be as described there, and re-define

$$\hat{a}_j^*(x) = n_j^{-1} \sum_{i=1}^{n_j} K(x, X_{ji}^*), \qquad T^* = \int (\hat{a}_1^* - \hat{a}_2^*).$$

A test with asymptotic level α may be obtained by rejecting H_0 if $T > \hat{t}_\alpha$, where $\hat{t}_\alpha$ is the solution of the equation

$$P(T^* > \hat{t}_\alpha \mid \mathscr{X}) = \alpha.$$

To appreciate that this test has good power properties, let us consider local alternatives of the type discussed in Subsection 2.4, where $f_1 = f$ (fixed) and $f_2 = f + n^{-1/2} cg$. Assume that each $v_k \neq 0$ and $\{\omega_k\}$ is a complete orthonormal sequence in the space of all functions sharing the same support as f and g. Then the asymptotic value, $\pi(c)$, of the probability that the test rejects H_0 when H_1 is true, satisfies $\pi(c) \to 1$ as $|c| \to \infty$. Without the assumptions that the sequence $\{\omega_k\}$ is complete, the test may not be able to discriminate against certain alternatives. (Consider, for example, the case where g is orthogonal to $\{\omega_k\}$.) However, in many practical problems we are interested in detecting departures which, if they exist at all, are most likely reflected in location or scale difference. We could confine attention to such departures, and so increase the power of the test for discriminating against them. Of course, the penalty paid is that such a test has no power for discriminating against departures orthogonal to location and scale.

Consider the problem of testing for location and scale differences among two-dimensional distributions. Let $x = (x^{(1)}, x^{(2)})^{\mathrm{T}}$. Given two bivariate functions ψ_1 and ψ_2, and bivariate data $\mathscr{X} = \{X_{11}, ..., X_{1n_1}, X_{21}, ..., X_{2n_2}\}$, put

$$\langle \psi_1, \psi_2 \rangle = (n_1 + n_2)^{-1} \sum_{j=1}^{2} \sum_{k=1}^{n_j} \psi_1(X_{ji})\, \psi_2(X_{ji}), \qquad \|\psi_1\|^2 = \langle \psi_1, \psi_1 \rangle,$$

$$\psi_{11}(x) = x^{(1)}, \qquad \psi_{12}(x) = x^{(2)}, \qquad \psi_{13}(x) = x^{(1)^2}, \qquad \psi_{14}(x) = x^{(2)^2},$$

$$\psi_{15} = x^{(1)} x^{(2)}, \qquad \psi_{21} = \psi_{11},$$

$$\psi_{2i} = \psi_{1i} - \langle \psi_{1i}, \psi_{2,i-1} \rangle \, \|\psi_{2,i-1}\|^{-2} \psi_{2,i-1} - \cdots$$

$$- \langle \psi_{1i}, \psi_{2i} \rangle \, \|\psi_{21}\|^{-2} \psi_{21},$$

$$\omega_i = \psi_{2i} \|\psi_{2i}\|^{-1}, \qquad 1 \leqslant i \leqslant 5.$$

Then the functions ω_i are orthonormal with respect to the weight $d\hat{F}$, where $\hat{F}$ is the empirical distribution function of $\mathscr{X}$. In effect, the first two ω_i's represent location, and the next three denote principally scale.

Appendix

Mean, Variance, and Asymptotic Distribution of the Test Statistic, T, Defined in (2.3)

Under the hypothesis that $f_1 = f$ and $f_2 = f + \delta g$ we have

$$E(T) = \int \left[E(\hat{a}_1 - E\hat{a}_1)^2 + E(\hat{a}_2 - E\hat{a}_2)^2 + \{E(\hat{a}_1 - \hat{a}_2)\}^2 \right]$$

$$= n_1^{-1} \left\{ K_1 - \int (K * f)^2 \right\}$$

$$+ n_2^{-1} \left[\kappa_1 - \int \{K * (f + \delta g)\}^2 \right] + \delta^2 \int (K * g)^2$$

$$= \begin{cases} (n_1^{-1} + n_2^{-1})\, J_1 & \text{if } \delta = 0 \\ (n_1^{-1} + n_2^{-1})\{1 + o(1)\}\, J_1 + \delta^2 \int (K * g)^2 & \text{if } \delta \to 0. \end{cases}$$

Further, defining

$$S_{1j} = \sum_{i=1}^{n_j} \int \{K(x - X_{ji}) - a_j(x)\}^2 \, dx,$$

$$S_{2j} = \sum_{i_1 \neq i_2} \sum \int \{K(x - X_{ji_1}) - a_j(x)\} \{K(x - X_{ji_2}) - a_j(x)\} \, dx,$$

$$S_3 = \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \int \{K(x - X_{1i_1}) - a_1(x)\}\{K(x - X_{2i_2}) - a_2(x)\}\, dx,$$

$$S_{4j} = \sum_{i=1}^{n_j} \int \{K(x - X_{ji}) - a_j(x)\}\, g(x)\, dx,$$

we have

$$T - E(T) = \sum_{j=1}^{2} n_j^{-2}\{(S_{1j} - ES_{1j}) + S_{2j}\}$$
$$- 2n_1^{-1}n_2^{-1}S_3 + 2\delta(n_1^{-1}S_{41} - n_2^{-1}S_{42}).$$

Now, $S_{1j} - ES_{1j} = O_p(n_j^{1/2})$, which means that the contribution of $S_{1j} - ES_{1j}$ to the formula above is of order $n^{-3/2}$. This means that it is negligible relative to the contributions of the other terms, which are all of size n^{-1}. Indeed, the asymptotic variances of S_{21}, S_{22}, S_3, S_{41} and S_{42} are $n_1^2 J_2$, $n_2^2 J_2$, $n_1 n_2 J_2$, $n_1 J_4$ and $n_2 J_4$, respectively. These quantities all have zero means and are uncorrelated with one another. It follows that $T - E(T)$ has asymptotic variance $\{n_1^{-2} + n_2^{-2} + 4(n_1 n_2)^{-1}\}\, J_2 + 4\delta^2(n_1^{-1} + n_2^{-1})\, J_4$.

With M, λ_k and ω_k defined by (2.4)–(2.7), we have

$$S_{2j} = \sum_{k=1}^{\infty} \lambda_k \left[\left\{\sum_{i=1}^{n_j} \omega_k(X_{ji})\right\}^2 - \sum_{i=1}^{n_j} \omega_k(X_{ji})^2\right].$$

This representation is correct for $j = 1, 2$ under H_0, and for $j = 1$ under H_1. Under H_1, when $j = 2$, the λ_k's and ω_k's should be replaced by versions which depend on δ and converge to their counterparts (under H_0, i.e., with $\delta = 0$) as $\delta \to 0$. Since $\int M(x_1, x_2)\, f(x_2)\, dx_2 = 0$ then we may deduce from (2.6) that if $\lambda_k \neq 0$ then $E\{\omega_k(X_{ji})\} = 0$. By (2.7), $E\{\omega_k(X_{ji})\, \omega_l(X_{ji})\} = \delta_{kl}$, whence it follows that under H_0 or H_1,

$$n_j^{-1} S_{2j} \to \sum_{k=1}^{\infty} \lambda_k (Z_{jk}^2 - 1) \qquad (j = 1, 2), \qquad (n_1 n_2)^{-1/2} S_3 \to \sum_{k=1}^{\infty} \lambda_k Z_{1k} Z_{2k}$$

jointly in distribution. Therefore,

$$n\left(\sum_{j=1}^{2} n_j^{-2} S_{2j} - 2n_1^{-1} n_2^{-1} S_3\right)$$
$$\to \sum_{j=1}^{2} \rho_j^{-1} \sum_{k=1}^{\infty} \lambda_k (Z_{jk}^2 - 1) - 2\rho_1^{-1/2}\rho_2^{-1/2} \sum_{k=1}^{\infty} \lambda_k Z_{1k} Z_{2k}$$
$$= \sum_{k=1}^{\infty} \lambda_k \{(\rho_1^{-1/2} Z_{1k} - \rho_2^{-1/2} Z_{2k})^2 - (\rho_1^{-1} + \rho_2^{-1})\}.$$

More simply, S_{4j} is asymptotically normally distributed with zero mean and variance $n_j J_4$. Its asymptotic correlation with $\sum_i \omega_k(X_{ji})$ is $n\rho_j r_k$. Joint asymptotic normality may be proved using the Cramér–Wold device.

Acknowledgments

The support of the U. K. Science and Engineering Research Council is gratefully acknowledged, in the form of a Visiting Fellowship Research Grant (for P.H.) and a research studentship (for N.H.A.).

References

[1] Hall, P. (1979). On the invariance principle for *U*-statistics, *Stochastic Process. Appl.* **9** 163–174.

[2] Hall, P. (1984). Central limit theorem for integrated squared error for multivariate nonparametric density estimator. *J. Multivariate Anal.* **14** 1–16.

[3] Indritz, J. (1963). "*Methods in Analysis.*" Macmillan, New York.

[4] Neuhaus, G. (1977). Functional limit theorems for *U*-statistics in the degenerate case. *J. Multivariate Anal.* **7** 424–439.

[5] Silverman, B. W. (1986). "*Density Estimation for Statistics and Data Analysis*". Chapman and Hall, London.