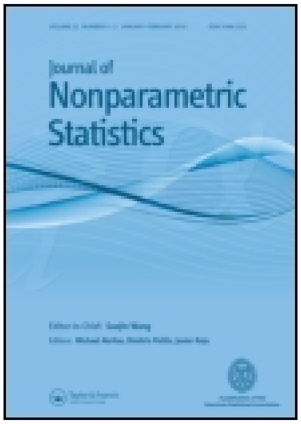

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

Goodness of fit in multidimensions based on nearest neighbour distances

Sean Zhou [a] & S. Rao Jammalamadaka [b]

[a] Department of Mathematics, University of Western Australia, Nedlands, WA, 6009

[b] Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, 93106

Available online: 12 Apr 2007

Nonparametric Statistics, Vol. 2, pp. 271–284
Reprints available directly from the publisher
Photocopying permitted by license only

GOODNESS OF FIT IN MULTIDIMENSIONS BASED ON NEAREST NEIGHBOUR DISTANCES

SEAN ZHOU

Department of Mathematics, University of Western Australia, Nedlands, WA, 6009

S. RAO JAMMALAMADAKA[1]

Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106

(Received March 2, 1992)

This paper is concerned with tests based on nearest neighbour distances for the goodness of fit problem in multidimensions a la Bickel and Brieman (1983). We argue that the nearest neighbour distances provide a natural extension to multidimensions, of the idea of "spacings", which have been extensively used on the real line. The asymptotic distribution theory for a general class of these test statistics is studied both under the null hypothesis as well as under an appropriately converging sequence of alternatives. The results are used to obtain the Pitman asymptotic relative efficiencies of such statistics and to discuss optimal tests in this class.

KEYWORDS: Goodness of fit in multi-dimensions, nearest neighbour distances, Pitman asymptotic relative efficiencies.

1. INTRODUCTION

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with a common density function $f(x)$ on R^d, i.e., each X_i is a d-dimensional vector ($d \geq 1$). The basic goodness of fit problem is to test

$$H_0: F = F_0$$

where F_0 is a specified distribution function (d.f.) on R^d. Often a preliminary test of this type on model-checking precedes all the rest of statistical inference.

On the real line, i.e. for $d = 1$, broadly speaking there are three general approaches to testing the goodness of fit hypothesis.

(a) χ^2 *methods:* Fix cells or class intervals and compare the observed frequencies in each cell with what is expected under H_0. This classical procedure goes back to Pearson (1900). For a recent review, see Moore and Spurill (1976).

(b) *Empirical d.f. methods:* Compute the empirical d.f.

$$F_n(x) = \frac{1}{n}(\text{number of } X_i \leq x)$$

and check how far this is from the postulated one by using a distance $d(\cdot, \cdot)$.

[1] Research supported in part by ONR Grant number N00014-93-1-0174.

Reject H_0 if $d(F_n, F_0) > c_\alpha$. For example, Kolmogorov–Smirnov test:

$$d(F_n, F_0) = \sqrt{n} \sup_n |F_n(x) - F_0(x)|.$$

See for instance, Shorack and Wellner (1986).

(c) *Spacings methods:* This can be considered as the "dual" approach to χ^2. We fix a frequency, say m, (in m-step spacings) and consider the length of the interval formed by X_i's which contains m successive observations. More precisely, we order

$$-\infty = X_{(0)} \leq X(1) \leqq X_{(2)} \leq \cdots \leq X_{(n)} \leq X_{(n+1)} = \infty$$

and define 1-step spacings:

$$D_i = F_0(X_{(i+1)}) - F_0(X_{(i)}), \qquad i = 0, 1, \ldots, n$$

m-step spacings (overlapping):

$$D_i^{(m)} = F_0(X_{(i+m)}) - F_0(X_{(i)}), \qquad i = 0, 1, \ldots, n - m,$$

or m-step spacings (disjoint or non-overlapping):

$$D_{i,m}^{(m)} = F_0(X_{((i+1)m)}) - F_0(X_{(im)}), \qquad i = 0, 1, \ldots, \left[\frac{n}{m}\right] - 1.$$

There is a vast literature on goodness of fit tests based on spacings. Pyke (1965) is a good review. Among others are Kuo and Rao (1981), Hall (1986), Jammalamadaka, Zhou and Tiwari (1989), etc. The general idea is to compare the null probability measure of the coverages, i.e. $F_0((X_{(i)}, X_{(i+1)}])$, with its expected value under H_0, namely, $1/n$.

There are, of course, many other procedures such as probability plots, moment techniques etc besides several adhoc methods. An important reference in this connection is the handbook by D'Agostino and Stephens (1986). See also Andrews *et al.* (1973) and Koziol (1986) in connection with testing multivariate normality.

In multidimensional spaces, the search for goodness of fit tests that are general and practical, similar to those mentioned above for one dimension, remains an open problem. Although the χ^2 methods are still available in theory, there are difficulties associated with the arbitrariness of choosing the classes or cells. The distribution-free nature of the procedures of Kolmogorov–Smirnov type tests does not extend to higher dimensions. Since there is no unique ordering (and order statistics) in multi-dimensional spades, we cannot define spacings, as in one dimension. However, noting that the general idea of spacings is to compare $F_0((X_{(i)}, X_{(i+1)}])$, with their expected value $1/n$ under H_0, we may extend this concept to multidimensions by choosing appropriate coverages to replace the interval $(X_{(i)}, X_{(i+1)}]$. One immediate choice is the "nearest-neighbour" ball $B(X_i, R_i)$, which has been studied by Bickel and Breiman (1983), where $B(x, r) = \{y : \|y - x\| < r\}$ is the ball of radius r around x with the usual Euclidean norm $\|\cdot\|$ on R^d, and R_i is the nearest-neighbour distance from X_i defined by

$$R_i = \min_{j \neq i} \|X_j - X_i\|.$$

In one dimension, $B(X_i, R_i)$ reduces to the smaller of the two 1-step spacing intervals surrounding X_i. Let

$$F_i = F_0(B(X_i, R_i)) = \int_{B(X_i,R_i)} f(y)\,\mathrm{d}y = \int_{\|y\|<R_i} f(X_i+y)\,\mathrm{d}y,$$

denote the coverage probability of $B(X_i, R_i)$, where $f(x)$ is the density under H_0. It is interesting to note that for any fixed x, $F_0(B(x, r_j))$ where $r_j = \|X_j - x\|$, are i.i.d. with a uniform distribution on $(0, 1)$. Moreover, $P(F_i > a) = (1-a)^{n-1}$, independent of the null distribution F_0, and $E[F_i] = 1/n$ under H_0, same as the expected value of 1-step spacings in one dimension. We consider the class of test statistics based on F_i of the form

$$T_n = T_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nF_i) - E_0 h(nF_i)],$$

where h is a real-valued function defined on $[0, \infty)$ and E_0 denotes the expectation under H_0. Since the computation of the coverage probabilities F_i poses a numerical problem, one may consider the following approximation for F_i, proposed by Bickel and Breiman (1983):

$$D_i = f(X_i)V(R_i)$$

where $V(r)$ denotes the volume of $B(0, r)$. Note that R_i is small for large n and so $f(x) \approx f(X_i)$ for $x \in B(X_i, R_i)$, assuming $f(x)$ is continuous. Thus for large n,

$$F_i = F_0(B(X_i, R_i)) = \int_{B(X_i,R_i)} f(x)\,\mathrm{d}x \approx f(X_i)V(R_i) = D_i.$$

We will also consider tests based on $\{D_i\}$, of the form

$$T_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nD_i) - E_0 h(nD_i)].$$

In the following sections, we study the limiting behaviour of both T_n and T_n^* under the null hypothesis as well as under a sequence of alternatives converging to H_0. Although T_n^* is simpler in computation than T_n, we focus primarily on T_n in our study since it is of independent theoretical interest as the more legitimate extension of the concept of spacings. Also, one can numerically evaluate using a computer, the $\{F_i\}$ and hence T_n.

Our results show that the limiting distribution of T_n (and T_n^*) is independent of the null distribution F_0 and thus provide asymptotically distribution-free tests (which of course is not surprising in view of the results of Bickel and Breiman (1983) about D_i). More specifically, T_n converges in distribution to $N(0, \sigma^2)$ under H_0 and to $N(\mu, \sigma^2)$ under a sequence of alternatives converging to H_0 at a rate of $n^{-1/4}$, for $d < 8$. The test statistic here, does not depend on the alternatives. Thus an "optimal" test among the class of tests $T_n(h)$ provides an omnibus test of uniformity, irrespective of the alternatives.

Bickel and Breiman (1983) study tests based on $\{D_i\}$. It should be remarked that while they derive the asymptotic behaviour of the empirical process based on $\{D_i\}$ under the null hypothesis, they do not consider the limiting distribution under a sequence of close alternatives. Schilling (1983) studies the limiting distribution of a weighted version of the empirical process based on $\{D_i\}$ under a sequence of alternatives converging to H_0 at a rate of $n^{-1/2}$. His results show that the unweighted process $\sum_{i=1}^{n} I(nD_i \leq t)$ has no power against these alternatives. This does not however pinpoint that the correct rate at which the alternatives should converge is $n^{-1/4}$ for the unweighted case, nor study tests and their relative efficiencies for these alternatives. Schilling (1983) shows that by choosing an appropriate weight function $w(x)$ based on the alternatives, the weighted version $\sum_{i=1}^{n} w(nD_i)I(nD_i \leq t)$ can detect alternatives converging at the rate of $n^{-1/2}$. In comparison with this, the test statistic T_n is less powerful *when the alternatives are specified* (on which the appropriate weight function $w(x)$ depends). But most frequently, the alternatives are not given in a goodness of fit problem so that the correct weighting cannot be determined. In contrast, our tests T_n and T_n^* are independent of the alternatives and they are uniformly powerful against the general alternatives which are at a distance of $n^{-1/4}$ from H_0. In fact, when the alternatives are known, one can always use the likelihood ratio test to achieve the maximum power. It is interesting to note that these results parallel those for spacings tests (see Holst and Rao (1981)).

The spacings tests as well as the current procedures based on $\{F_i\}$ have the same local power properties as *comparable* chi-square procedures and in fact, better in many instances (See Jammalamadaka and Tiwari (1987)). By *comparable* chi-square procedures we mean the following: because of the duality we alluded to earlier (i.e. fixing cells and comparing the frequencies as in chi-square, versus, fixing frequencies and comparing the lengths/volumes of the cells as in spacings/nearest neighbor methods) the present spacings/nearest neighbour methods should be compared to chi-square procedures with expected frequency of one. Such chi-square tests are asymptotically normal and are not as efficient as the spacings tests. See Jammalamadaka and Tiwari (1987). Similar comparisons are also possible between the usual chi-square test with a finite number of cells and the spacings tests (specifically the Greenwood statistic which is the sum of squares of spacings) where the length of the step m is a fixed finite fraction of the sample size. In this latter case, spacings tests also have an asymptotic chi-square distribution and one can handle the estimated parameter problems also as was done in Wells *et al.* (1992). The results of this investigation will appear elsewhere. To summarize, spacings type tests can meet or beat the Chi-square procedures, when appropriate comparisons are made.

In Section 2, we will state the results about the limiting distributions of T_n and T_n^*. Since the details of the proofs are quite technical and lengthy, although they are either basically routine or similar to those of Bickel and Breiman (1983), we provide only an outline of the proofs in Section 3, to keep the paper short.

2. RESULTS

Proposition 1 gives the limiting process of the empirical process based on F_i.

Let $\{\xi_n(t): 0 \leq t \leq \infty\}$ be the normalized empirical process of $\{nF_i : i =$

$1, \ldots, n\}$, i.e.

$$\xi_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I(nF_i \leq t) - P(nF_i \leq t)].$$

PROPOSITION 1. If the following assumption on the density $f(x)$ under H_0 holds:

(A1)
(i) $\{f > 0\} = \{x \in R^d \ ; f(x) > 0\}$ is open in R^d, and
(ii) f is uniformly bounded and continuous on $\{f > 0\}$,

then under H_0, the process $\{\xi_n(t) : 0 \leq t \leq \infty\}$ converges weakly to a Gaussian process $\{\xi(t) : 0 \leq t \leq \infty\}$ with mean zero and covariance function

$$K(s, t) = e^{-t} - e^{-s-t}\left[1 - s + st - \int_{W(s,t)} (e^{\beta(s,t,w)} - 1)\, \mathrm{d}w\right], \qquad 0 \leq s \leq t \leq \infty,$$

where

$$W(s, t) = \{w \in R^d : r_1 \leq \|w\| \leq r_1 + r_2\},$$
$$\beta(s, t, w) = \int_{B(0,r_1) \cap B(w,r_2)} \mathrm{d}z$$

and r_1, r_2 are given by $V(r_1) = t$, $V(r_2) = s$.

The proof of Proposition 1 is based on the results of Bickel and Breiman (1983) on D_i. The main argument is that the empirical process of $\{nF_i\}$ is close to that of $\{nD_i\}$ so that they have the same weak limit. As a result of Proposition 1, we can obtain the limiting distribution of T_n under the null hypothesis based on the fact that T_n can be represented as a functional of the process $\{\xi_n(t)\}$. The results are stated in the following Proposition 2 and its corollary which provides simpler sufficient conditions:

PROPOSITION 2. Assume that (A1) holds and a real-valued function h on $[0, \infty)$ (possibly infinity at 0,) satisfies assumption:

(A2)
(i) $h(t)$ is of bounded variation on any closed interval in $(0, \infty)$;
(ii) $|\int_0^\infty (te^{-t})^{1/2}\, \mathrm{d}h(t)| < \infty$.

Then

$$T_n = T_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nF_i) - Eh(nF_i)] \xrightarrow{d} N(0, \sigma^2(h)),$$

under H_0 where $\sigma^2(h) = \int_0^\infty \int_0^\infty K(s, t)\, \mathrm{d}h(s)\, \mathrm{d}h(t)$ and $K(s, t)$ is as in Proposition 1.

COROLLARY . Suppose the following assumption on $h(\cdot)$ holds:

(A3)
(i) h is absolutely continuous in $(0, \infty)$ and its derivative h' is bounded on any closed interval in $(0, \infty)$,
(ii) $t^{\alpha} h^2(t)$ is bounded for some $\alpha < 1$,
(iii) $e^{-\beta t} h^2(t)$ is bounded for some $\beta < 1$.

Then $T_n \xrightarrow{d} N(0, \sigma^2(h))$ under H_0.

Some typical examples of h satisfying (A3) include $h(t) = t^\gamma (\gamma > -1/2)$, $h(t) = \log t$ and $h(t) = |t - c|$, and cover most statistics in the literature. The following two theorems will show that the empirical process $\{\xi_n(t)\}$ and the test statistic T_n will converge to a nondegenerate limit under a sequence of alternatives converging to the null hypothesis at a rate of $n^{-1/4}$, which will enable us to obtain the asymptotic relative efficiencies for T_n. We consider the limiting distribution under the following sequence of alternatives:

$$H_{1n} : X_i \sim f(x) + n^{-1/4} l(x). \tag{2.1}$$

where $\int l(x)\,\mathrm{d}x = 0$. Such a sequence of alternatives was studied in one dimension by a number of authors. See for example, Jammalamadaka and Tiwari (1987).

THEOREM 3. If (A1) and the following (A4) and (A5) hold:

(A4) f is twice continuously differentiable on $\{f > 0\}$;
(A5) l is supported in a compact subset of $\{f > 0\}$ and is twice continuously differentiable on $\{f > 0\}$,

then under H_{1n},

$$\xi_n(t) \to \xi(t) + \left(\frac{t^2}{2} - t\right) e^{-t} \int \frac{l^2(x)}{f(x)}\,\mathrm{d}x$$

weakly, for dimension $d < 8$, where $\xi(t)$ is as defined in Proposition 1.

THEOREM 4. Assume (A1) and (A3)–(A5) hold, $d < 8$. Then

$$T_n = T_n(h) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [h(nF_i) - Eh(nF_i)] \xrightarrow{d} N(\mu(h), \sigma^2(h))$$

under $\{H_{1n}\}$, where

$$\mu(h) = \int_0^\infty \left(\frac{t^2}{2} - t\right) e^{-t}\,\mathrm{d}h(t) \int \frac{l^2(x)}{f(x)}\,\mathrm{d}x$$

$$\sigma^2(h) = \int_0^\infty \int_0^\infty K(s, t)\,\mathrm{d}h(s)\,\mathrm{d}h(t)$$

and $K(s, t)$ is as in Proposition 1.

The basic idea for the proof of Theorem 3 is to use a Taylor expansion on the mean function of the process, while Theorem 4 follows from Theorem 3 just as Proposition 2 follows from Proposition 1, although the details are a bit complicated.

¿From Theorem 4, it can be shown that the Pitman Asymptotic Relative Efficiency (ARE) of $T_n(h_1)$ relative to $T_n(h_2)$ for two different functions $h_1(\cdot)$ and $h_2(\cdot)$ is

$$\mathrm{ARE}(T_n(h_1), T_n(h_2)) = \frac{\mathrm{Eff}(T_n(h_1))}{\mathrm{Eff}(T_n(h_2))}$$

where

$$\mathrm{Eff}(T_n(h)) = \frac{\mu^2(h)}{\sigma^2(h)} = \frac{\left[\int_0^\infty (t - t^2/2) e^{-t}\,\mathrm{d}h(t)\right]^2}{\int_0^\infty \int_0^\infty K(s, t)\,\mathrm{d}h(s)\,\mathrm{d}h(t)} \int \frac{l^2(x)}{f(x)}\,\mathrm{d}x$$

is called the "efficacy" of $T_n(h)$. As a result, the optimal statistic (in Pitman sense) in the class of $T_n(h)$ is the one with the function h that maximizes

$$e(h)=\frac{\left[\int_0^\infty (t-t^2/2)e^{-t}h'(t)\,\mathrm{d}t\right]^2}{\int_0^\infty\int_0^\infty K(s,t)h'(s)h'(t)\,\mathrm{d}s\,\mathrm{d}t}.$$

This optimization problem, we find, is quite nontrivial but can be reduced to a standard procedure of solving integral equation. To see this, suppose that $g(t)$ is a solution of the following integral equation:

$$\int_0^\infty K(s,t)g(s)\,\mathrm{d}s=(t-t^2/2)e^{-t}. \tag{2.2}$$

Then, by considering $\int_0^\infty\int_0^\infty K(s,t)g(s)h'(t)\,\mathrm{d}s\,\mathrm{d}t$ as the inner product of g and h' and applying the Cauchy-Schwartz inequality we obtain

$$e(h)=\frac{\left[\int_0^\infty\int_0^\infty K(s,t)g(s)h'(t)\,\mathrm{d}t\right]^2}{\int_0^\infty\int_0^\infty K(s,t)h'(s)h'(t)\,\mathrm{d}s\,\mathrm{d}t}\le\int_0^\infty\int_0^\infty K(s,t)g(s)g(t)\,\mathrm{d}s\,\mathrm{d}t$$

with equality if $h'(t)=g(t)$. Thus the optimal function $h(\cdot)$ is given by $h(t)=\int_0^t g(s)\,\mathrm{d}s$ where $g(s)$ is the solution of the integral equation (2.2). Note that $e(h)$ depends only on h, independent of the alternatives as well as the null hypothesis.

The corresponding results for T_n^* are summarized in the following theorem:

THEOREM 5. (a) If $f(x)$ satisfies (A1) and h is a function of bounded variation on $[0,\infty)$, then

$$T_n^*=T_n^*(h)=\frac{1}{\sqrt{n}}\sum_{i=1}^n[h(nD_i)-Eh(nD_i)]\xrightarrow{d}N(0,\sigma^2(h)),$$

under H_0, where $\sigma^2(h)$ is the same as in Proposition 2.

(b) If the conditions in part (a), as well as (A4) and (A5) hold, then

$$T_n^*=T_n^*(h)\xrightarrow{d}N(\mu(h),\sigma^2(h))$$

under $\{H_{1n}\}$ given by (2.1), where $\mu(h)$ and $\sigma^2(h)$ are as in Theorem 4.

Remark. The condition on function h in Theorem 5 is stronger than (A2) and many useful functions such as polynomials and logarithms do not have bounded variation on $[0,\infty)$. However, if a function satisfies part (i) of (A2) (i.e., of bounded variation on closed intervals in $(0,\infty)$), a truncation can turn into a function of bounded variation on $[0,\infty)$ and still keep T_n^* little changed. For example, if we are interested in using function $h(t)=t^2$, say, then its truncated version

$$\bar{h}(t)=t^2I(t\le A)+A^2I(t>A),\qquad 0<A<\infty,$$

would have bounded variation on $[0,\infty)$. Since $h(t)=\bar{h}(t)$ for $t\in[0,A]$, and A can be chosen large enough so that $T_n^*(\bar{h})$ is virtually unchanged from $T_n^*(h)$. This should provide an approximation good enough for all practical purposes.

3. OUTLINE OF THE PROOFS

Proof of Proposition 1. The proof is based on the results of Bickel and Breiman (1983) which show that the normalized empirical process of $\{nD_i : i = 1, \ldots, n\}$, say $\eta_n(t)$, converges weakly to $\xi(t)$. Since the tightness of $\{\xi_n(t)\}$ can be proved in a way very similar to that of $\{\eta_n(t)\}$, we need only to verify that $\mathrm{Var}(\xi_n(t) - \eta_n(t) \to 0$, which establishes that these two processes converge to the same limit, namely $\xi(t)$.

First we cite an inequality provided by Bickel and Breiman (1983): Let $g_n(x, r)$ and $h_n(x, r)$ be two bounded functions defined on $R^d \times [0, \infty)$. Put $h_1 = h_n(X_1, n^{1/d}R_1)$ and $g_2 = g_n(X_2, n^{1/d}R_2)$. Then $\exists$ constant $M < \infty$ such that

$$|\mathrm{Cov}(h_1, g_2)| \le M \, \|g_n\| \, (n^{-1}E\,|h_1| + E\,|h_1F_1|) \qquad \forall n > 4. \tag{3.1}$$

Let $\epsilon > 0$. Note that $nF_0(B(x, n^{-1/d}r)) \to f(x)V(r)$ as $n \to \infty$, hence by Lebesgue's Dominated Convergence Theorem (LDCT)

$$P(n^{1/d}R_1 > M) = \int [1 - F_0(B(x, n^{-1/d}M))]^{n-1} f(x)\,\mathrm{d}x \to \int e^{-f(x)V(M)} f(x)\,\mathrm{d}x.$$

The last integral tends to zero as $M \to \infty$. it follows that $\exists M$ and N such that

$$P(n^{1/d}R_1 > M) < \epsilon \qquad \forall n > N. \tag{3.2}$$

Next, since $P(nF_1 > t) = (1 - t/n)^{n-1}$,

$$\begin{aligned} &P(nF_1 \le t,\, nD_1 > t,\, n^{1/d}R_1 \le M) \\ &= \int\int_0^t P(nD_1 > t,\, n^{1/d}R_1 \le M \mid X_1 = x,\, nF_1 = u) \frac{n-1}{n}\left(1 - \frac{u}{n}\right)^{n-2} f(x)\,\mathrm{d}u\,\mathrm{d}x. \end{aligned} \tag{3.3}$$

Given $X_1 = x$ and $nF_1 = u < t$, put $\delta = (t - u)/2V(M) > 0$. Then by (A1), for $z \in \{f > 0\}$, $|f(x + n^{-1/d}z) - f(x)| < \delta$ $\forall z \in B(0, M)$ and large n. Hence when $n^{1/d}R_1 \le M$,

$$u = nF_1 = n\int_{\|y\|<R_1} f(x + y)\,\mathrm{d}y \ge (f(x) - \delta)nV(R_1) \ge nD_1 - \delta V(M)$$

so that $nD_1 \le u + \delta V(M) = (u + t)/2 < t$. Consequently the conditional probability in (3.3) equals zero for large n and so $P(nF_1 \le t, nD_1 > t, n^{1/d}R_1 \le M) \to 0$ as $n \to \infty$ by (3.3) and LDCT. This together with (3.2) proves that $P(nF_1 \le t, nD_1 > t) \to 0$. It follows that

$$\begin{aligned} &E[I(nF_1 > t) - I(nD_1 > t)]^2 \\ &\quad = P(nF_1 > t) + P(nD_1 > t) - 2P(nF_1 > t, nD_1 > t) \\ &\quad = P(nF_1 > t) + P(nD_1 > t) - 2[P(nD_1 > t) - P(nF_1 \le t, nD_1 > t)] \\ &\quad\ \to e^{-t} + e^{-t} - 2e^{-t} = 0. \end{aligned} \tag{3.4}$$

Now take $g_n(x, r) = h_x(x, r) = I\{nF_0(B(x, n^{-1/d}r)) \le t\} - I\{f(x)V(r) \le t\}$ and use (3.1) and (3.4), we obtain

$$\begin{aligned} &n\,\mathrm{Cov}[I(nF_1 \le t) - I(nD_1 \le t),\, I(nF_2 \le t) - I(nD_2 \le t)] \\ &\le nM_1\left\{\frac{1}{n}E\,|I(nF_1 \le t) - I(nD_1 \le t)| + E\,|[I(nF_1 \le t) - I(nD_1 \le t)]F_1|\right\} \\ &\le M_1\{E\,|I(nF_1 \le t) - I(nD_1 \le t)| + n[E(I(nF_1 \le t) - I(nD_1 \le t))^2E(F_1^2)]^{1/2}\} \to 0 \end{aligned}$$

where M_1 is a constant. Here we used the fact that $n[E(F_1^2)]^{1/2} = n[2/n(n+1)]^{1/2}$ is bounded. Finally

$$\operatorname{Var}(\eta_n(t) - \xi_n(t)) \le E[I(nF_1 \le t) - I(nD_1 \le t)]^2 \\ + (n-1)\operatorname{Cov}[I(nF_1 \le t) - I(nD_1 \le t),\ I(nF_2 \le t) - I(nD_2 \le t)] \to 0.$$

Proof of Proposition 2: The proof uses a method described in Shorack and Wellner (1986, pp. 737–739). First note that due to Proposition 1, we can assume that $\|\xi_n - \xi\|_\infty = \sup_t |\xi_n(t) - \xi(t)| \xrightarrow{P} 0$. We will show that $\operatorname{Var}(\xi_n(t)) \le Mte^{-t}$ for some constant M. Let $\epsilon > 0$. Then

$$\operatorname{Var}\left[\int_0^\delta \xi_n(t)\,\mathrm{d}h(t)\right] = \int_0^\delta \int_0^\delta E[\xi_n(s)\xi_n(t)]\,\mathrm{d}h(s)\,\mathrm{d}h(t) \\ \le \left[\int_0^\delta \{\operatorname{Var}(\xi_n(t))\}^{1/2}\,\mathrm{d}h(t)\right]^2 \le M\left[\int_0^\delta (te^{-t})^{1/2}\,\mathrm{d}h(t)\right]^2 < \epsilon^3$$

for some $\delta > 0$ by (A2). Thus

$$P\left(\left|\int_0^\delta \xi_n(t)\,\mathrm{d}h(t)\right| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\operatorname{Var}\left[\int_0^\delta \xi_n(t)\,\mathrm{d}h(t)\right] < \frac{\epsilon^3}{\epsilon^2} = \epsilon. \tag{3.5}$$

Similarly

$$P\left(\left|\int_A^\infty \xi_n(t)\,\mathrm{d}h(t)\right| \ge \epsilon\right) < \epsilon \quad \text{for some large } A. \tag{3.6}$$

The same argument will also give

$$P\left(\left|\int_0^\delta \xi(t)\,\mathrm{d}h(t)\right| \ge \epsilon\right) < \epsilon, \qquad P\left(\left|\int_A^\infty \xi(t)\,\mathrm{d}h(t)\right| \ge \epsilon\right) < \epsilon \tag{3.7}$$

for some $\delta > 0$ and $A < \infty$. Finally, because

$$\int_\delta^A |[\xi_n(t) - \xi(t)]\,\mathrm{d}h(t)| \le \|\xi_n - \xi\|_\infty \int_\delta^A \mathrm{d}|h|(t) \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty,$$

(3.5)–(3.7) imply

$$T_n = \int_0^\infty -\xi_n(t)\,\mathrm{d}h(t) \xrightarrow{d} \int_0^\infty -\xi(t)\,\mathrm{d}h(t).$$

The last stochastic integral has a $N(0, \sigma^2(h))$ distribution by a standard argument on a Gaussian process (c.f. Sharack and Wellner (1986), pp. 42–43). It remains to show that $\operatorname{Var}(\xi_n(t)) \le Mte^{-t}\ \forall t \ge 0$ and $n \ge 4$. Since $\operatorname{Var}(\xi_n(t)) = 0$ for $t \ge n$, we need only to consider $t < n$. By (3.1) $\exists M_1$ such that $\forall n \ge 4$,

$$n\operatorname{Cov}[I(nF_1 \le t),\ I(nF_2 \le t)] \le M_1\{P(nF_1 \le t) + E[nF_1 I(nF_1 \le t)]\}.$$

Because

$$P(nF_1 \le t) = 1 - \left(1 - \frac{t}{n}\right)^{n-1} = \frac{t}{n}\sum_{k=0}^{n-2}\left(1 - \frac{t}{n}\right)^k \le t \qquad \forall t \in [0, n]$$

and

$$E[nF_1 I(nF_1 \le t)] \le tP(nF_1 \le t) \le t,$$

we have

$$\begin{aligned}\operatorname{Var}(\xi_n(t)) &= \operatorname{Var}[I(nF_1 \le t)] + (n-1)\operatorname{Cov}[I(nF_1 \le t), I(nF_2 \le t)]\\ &\le P(nF_1 \le t) + M_1\{P(nF_1 \le t) + E[nF_1 I(nF_1 \le t)]\}\\ &\le (1+2M_1)t \le (1+2M_1)ete^{-1} \qquad \forall t \in [0, 1].\end{aligned} \tag{3.8}$$

Similarly

$$\begin{aligned}\operatorname{Var}(\xi_n(t)) &= \operatorname{Var}[I(nF_1 > t)] + (n-1)\operatorname{Cov}[I(nF_1 > t), I(nF_2 > t)]\\ &\le P(nF_1 > t) + M_2\{P(nF_1 > t) + E[nF_1 I(nF_1 > t)]\}\end{aligned} \tag{3.9}$$

for some constant M_2. For $t \in (1, n]$,

$$P(nF_1 > t) = \left(1 - \frac{t}{n}\right)^{n-1} = e^{(n-1)\log(1-t/n)} < e^{-(n-1)(t/n)} \le ete^{-t} \tag{3.10}$$

and

$$\begin{aligned}E[nF_1 I(nF_1 > t)] &= \frac{n-1}{n}\int_t^n u\left(1-\frac{u}{n}\right)^{n-2} du = t\left(1-\frac{t}{n}\right)^{n-1} + \left(1-\frac{t}{n}\right)^n\\ &\le (te+1)e^{-t} \le (e+1)te^{-t}.\end{aligned} \tag{3.11}$$

Combining (3.9) through (3.11) yields

$$\operatorname{Var}(\xi_n(t)) \le [e + M_2(2e+1)]te^{-t} \qquad \forall t \in [1, n)$$

which together with (3.8) completes the proof.

Proof of the corollary of Proposition 2: Clearly (A3)-(i) implies (A2)-(i). So we need only to show that (ii)–(iii) of (A3) imply (ii) of (A2). Now

$$\begin{aligned}\text{(A3)-(ii)} &\Rightarrow t^{-1}h^2(t) = t^{-(1+\alpha)}t^{\alpha}h^2(t) \le Ct^{-(1+\alpha)} \quad \text{(for some constant } C)\\ &\Rightarrow t^{-1/2}|h(t)| \le \sqrt{C}\,t^{-1/2(1+\alpha)} \quad (\alpha < 1)\\ &\Rightarrow \int_0^1 t^{-1/2}|h(t)|\,dt < \infty \quad \text{and} \quad t^{1/2}|h(t)| \le \sqrt{C}\,t^{1/2(1-\alpha)} \to 0 \quad \text{as} \quad t \downarrow 0\\ &\Rightarrow \left|\int_0^1 t^{1/2}\,dh(t)\right| = \left|h(1) - \frac{1}{2}\int_0^1 t^{-1/2}h(t)\,dt\right| < \infty\\ &\Rightarrow \left|\int_0^1 t^{1/2}\,dh(t)\right| = \left|h(1) - \frac{1}{2}\int_0^1 (te^{-1})^{1/2}\,dh(t)\right| < \infty\end{aligned}$$

and

$$\begin{aligned}\text{(A3)-(iii)} &\Rightarrow (te^{-t})^{-1}h^2(t) \le Cte^{-(1-\beta)t} \quad \text{(for some constant } C)\\ &\Rightarrow (te^{-t})^{-1/2}|h(t)| \le \sqrt{C}\,t^{1/2}e^{-1/2(1-\beta)t} \quad (\beta < 1)\\ &\Rightarrow \int_1^\infty (te^{-t})^{1/2}|h(t)|\,dt < \infty \quad \text{and} \quad (te^{-t})^{1/2}h(t) \to 0 \quad \text{as} \quad t \to \infty\\ &\Rightarrow \left|\int_1^\infty (te^{-t})^{1/2}\,dh(t)\right| = \left|-e^{-1/2}h(1) - \frac{1}{2}\int_1^\infty (t^{-1/2} - t^{1/2})e^{-1/2t}h(t)\,dt\right|\\ &\le |h(1)| + \frac{1}{2}\int_1^\infty (1+t^{1/2})e^{-1/2t}|h(t)|\,dt < \infty.\end{aligned}$$

Thus (A2)-(ii) holds.

To prove Theorem 3 and Theorem 4, we define

$$S_n = S_n(x) = \inf\{r : nF_0(B(x, r)) = t\} \tag{3.12}$$

for $n > t$ and $x \in \{f > 0\}$. Clearly S_n is well-defined because $nF(B(x, r))$ is continuous in r. It is easy to see that

(a) $nF(B(x, S_n)) = t \, \forall x \in \{f > 0\}$ and $n > t$;
(b) $S_n = O(n^{-1/d})$ and $V(S_n) = O(n^{-1})$.

Proof of Theorem 3: Let P_n and P_0 denote the probability measures under H_{1n} and H_0 respectively. Write

$$\xi_n(t) = \xi_n^{(n)}(t) + [\bar{\xi}_n(t) - \xi_n^{(n)}(t)] + m_n(t)$$

where

$$\xi_n^{(n)}(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I\{nF_i^{(n)} \le t\} - E_n I\{nF_i^{(n)} \le t\}]$$

$$F_i^{(n)} = F_n(B(X_i, R_i))$$

$$\bar{\xi}_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [I\{nF_i \le t\} - E_n I\{nF_i \le t\}]$$

and

$$m_n(t) = \sqrt{n}\,[P_n(nF_1 > t) - P_0(nF_1 > t)]$$

It is clear that $\xi_n^{(n)}(t) \to \xi(t)$ weakly under H_{1n}. Moreover, $E_n[\bar{\xi}_n(t) - \xi_n^{(n)}(t)] = 0$ and it can be shown that

$$\operatorname{Var}(\bar{\xi}_n(t) - \xi_n^{(n)}(t)) \to 0$$

in a way similar to the proof of $\operatorname{Var}(\xi_n(t) - \eta_n(t)) \to 0$ in Theorem 1 with $F_i^{(n)}$ in place of D_i (the only difference is in the proof of $P(nF_1 \le t, nF_1^{(n)} > t) \to 0$, but this is actually easier because $|F_1^{(n)} - F_1| \le n^{-1/4} C F_1$ for some constant C by (A5)). It follows that $\bar{\xi}_n(t) - \xi_n^{(n)}(t) \xrightarrow{P} 0$ and so it suffices to show that

$$m_n(t) \to \left(\frac{t^2}{2} - t\right) e^{-t} \int \frac{l^2(x)}{f(x)}\, dx \tag{3.13}$$

as $n \to \infty$. Let S_n be as in (3.12). Then

$$\begin{aligned}
P_n(nF_1 > t) &= \int P_n(nF_1 > t \mid X_1 = x) f(x)\, dx \\
&= \int P_n(R_1 > S_n \mid X_1 = x) f(x)\, dx \\
&= \int \left\{1 - \int_{\|y\| < S_n} [f(x+y) + n^{-1/4} l(x+y)]\, dy\right\}^{n-1} f(x)\, dx \\
&= \int \{1 - F_0(B(x, S_n)) - n^{-1/4} L\}^{n-1} f(x)\, dx \\
&= \int \left\{1 - \frac{t}{n} - n^{-1/4} L\right\}^{n-1} f(x)\, dx
\end{aligned}$$

where $L = \int_{\|y\|<S_n} l(x+y)\,dy$. We also have $P_0(nF_1 > t) = \left(1 - \frac{t}{n}\right)^{n-1}$. Thus

$$\begin{aligned} m_n(t) &= \sqrt{n}\,\{P_n(nF_1 > t) - P_0(nF_1 > t)\} \\ &= \sqrt{n} \int \left[\left(1 - \frac{t}{n} - n^{-1/4}L\right)^{n-1} - \left(1 - \frac{t}{n}\right)^{n-1}\right] f(x)\,dx \\ &\quad + n^{-1/4} \int \left(1 - \frac{t}{n} - n^{-1/4}L\right)^{n-1} l(x)\,dx \\ &= I_1 + I_2, \quad \text{say.} \end{aligned} \tag{3.14}$$

A Taylor expansion shows that if $a_n = O(1)$, then

$$\begin{aligned} (n-1)\log(1 - n^{-5/4}a_n) &= (n-1)\{-n^{-5/4}a_n + O(n^{-5/2})\} \\ &= n^{-1/4}a_n - n^{-5/4}a_n + O(n^{-3/2}) \\ &= n^{-1/4}a_n + O(n^{-5/4}) \end{aligned}$$

and so

$$\begin{aligned} (1 - n^{-5/4}a_n)^{n-1} &= \exp\{-n^{-1/4}a_n + O(n^{-5/4})\} \\ &= 1 - n^{-1/4}a_n + O(n^{-5/4}) + \tfrac{1}{2}[-n^{-1/4}a_n + O(n^{-5/4})]^2 + O(n^{-3/4}) \\ &= 1 - n^{-1/4}a_n + \tfrac{1}{2}n^{-1/2}a_n^2 + o(n^{-1/2}). \end{aligned} \tag{3.15}$$

By (A5)

$$\begin{aligned} nL &= \int_{\|y\|<S_n} nl(x+y)\,dy = \int_{\|y\|<S_n} n\left[l(x) + \sum_{i=1}^{d} l_i'(x)y_i + O(\|y\|^2)\right] dy \\ &= nl(x)V(S_n) + nO(S_n^2)V(S_n) \end{aligned} \tag{3.16}$$

where $(y_1, \ldots, y_d) = y$ and $l_i' = \partial l / \partial x_i$, which shows $nL = O(1)$ and so by (3.15),

$$\begin{aligned} \left(1 - n^{-1/4}\frac{L}{1 - t/n}\right)^{n-1} &= \left(1 - n^{-5/4}\frac{nL}{1 - t/n}\right)^{n-1} \\ &= 1 - n^{-1/4}\frac{nL}{1 - t/n} + \frac{1}{2}n^{-1/2}\left(\frac{nL}{1 - t/n}\right)^2 + o(n^{-1/2}). \end{aligned} \tag{3.17}$$

Thus by (3.14),

$$\begin{aligned} I_1 &= \sqrt{n} \int \left(1 - \frac{t}{n}\right)^{n-1} \left[\left(1 - n^{-1/4}\frac{L}{1 - t/n}\right)^{n-1} - 1\right] f(x)\,dx \\ &= -n^{1/4} \int \left(1 - \frac{t}{n}\right)^{n-2} (nL) f(x)\,dx + \frac{1}{2} \int \left(1 - \frac{t}{n}\right)^{n-3} (nL)^2 f(x)\,dx + o(1). \end{aligned} \tag{3.18}$$

Similar to (3.16) we can obtain

$$t = nF_0(B(x, S_n)) = \int_{\|y\|<S_n} nf(x+y)\,dy = nf(x)V(S_n) + nO(S_n^2)V(S_n). \tag{3.19}$$

Combining (3.16) and (3.19) yields $nL = tl(x)/f(x) + O(n^{-2/d})$ and so by (3.18)

$$I_1 = -n^{1/4}\left(1 - \frac{t}{n}\right)^{n-2} t \int l(x)\,dx + O(n^{1/4-2/d}) + \frac{1}{2}\left(1 - \frac{t}{n}\right)^{n-3} t^2 \int \frac{l^2(x)}{f(x)}\,dx + o(1).$$

Note that $\int l(x)\,\mathrm{d}x = 0$. Thus for $d < 8$, as $n \to \infty$,

$$I_1 \to \tfrac{1}{2}t^2 e^{-t} \int \frac{l^2(x)}{f(x)}\,dx. \tag{3.20}$$

Similarly for $d < 8$,

$$\begin{aligned} I_2 &= n^{1/4} \int \left(1 - \frac{t}{n}\right)^{n-1} l(x)\,\mathrm{d}x - \int \left(1 - \frac{t}{n}\right)^{n-2} (nL)l(x)\,\mathrm{d}x + o(1) \\ &\to -te^{-t} \int \frac{l^2(x)}{f(x)}\,\mathrm{d}x \quad \text{as} \quad n \to \infty. \end{aligned} \tag{3.21}$$

Finally (3.13) follows from (3.14) together with (3.20) and (3.21).

Proof of Theorem 4: The proof of Theorem 4 is similar to that of Proposition 2. The idea is to find a function $M(t)$ such that $\mathrm{Var}_n(\xi_n(t)) \le M(t)$ and $\int [M(t)]^{1/2}\,\mathrm{d}h(t) < \infty$, where Var_n denotes the variance under H_{1n}. Again by using (3.1) we obtain

$$\mathrm{Var}_n(\xi_n(t)) \le P_n(nF_1 \le t) + M_1\{P_n(nF_1 \le t) + E_n[nF_1 I(nF_1 \le t)]\} \tag{3.22}$$

and

$$\mathrm{Var}_n(\xi_n(t)) \le P_n(nF_1 > t) + M_1\{P_n(nF_1 > t) + E_n[nF_1 I(nF_1 > t)]\} \tag{3.23}$$

for some constant M_1, where P_n and E_n denote the probability measure and the expectation under H_{1n} respectively. Define $F^{(n)}(A) = \int_A f_n(x)\,\mathrm{d}x$ for $A \subset R^d$. By (A5), $|l| \le Cf$ for some constant C, hence $|F^{(n)}(A) - F_0(A)| \le n^{-1/4}CF_0(A)$ and

$$|nF^{(n)}(B(x, S_n)) - t| = |nF^{(n)}(B(x, S_n)) - nF_0(B(x, S_n))| \le n^{-1/4}Ct.$$

Consequently

$$\begin{aligned} P_n(nF_1 > t) &= \int P_n(R_1 > S_n \mid X_1 = x) f_n(x)\,\mathrm{d}x \\ &= \int [1 - F^{(n)}(B(x, S_n))]^{n-1} f_n(x)\,\mathrm{d}x \\ &\le \int \left[1 - (1 - n^{-1/4}C)\frac{t}{n}\right]^{n-1} f_n(x)\,\mathrm{d}x. \end{aligned} \tag{3.24}$$

Let $\beta < 1$ satisfy (A3) and $\beta_1 \in (\beta, 1)$. Then by (3.24), $\exists N_1$ such that $\forall n > N_1$

$$P_n(nF_1 > t) \le \left(1 - \beta_1 \frac{t}{n}\right)^{n-1} \le e e^{-\beta_1 t} \qquad \forall t \ge 0 \tag{3.25}$$

and

$$\begin{aligned} E_n[nF_1 I(nF_1 > t)] &= -\int_{\{u > t\}} u\,\mathrm{d}P_n(nF_1 > u) \\ &= tP_n(nF_1 > t) + \int_t^n P_n(nF_1 > u)\,\mathrm{d}u \\ &\le ete^{-\beta_1 t} + \frac{e}{\beta_1} e^{-\beta_1 t}. \end{aligned} \tag{3.26}$$

Combining (3.25) and (3.26) with (3.23) we can see that $\exists M_2$ and N_2 such that

$$\mathrm{Var}_n(\xi_n(t)) \leq M_2 t e^{-\beta_1 t} \qquad \forall t \geq 1 \quad \text{and} \quad n > N_2. \tag{3.27}$$

In a similar way we can argue that $\exists N_3$ such that $\forall n > N_3$,

$$P_n(nF_1 \leq t) \leq 1 - \int \left[1 - (1 + n^{-1/4}C)\frac{t}{n}\right]^{n-1} f_n(x)\,\mathrm{d}x \leq 1 - \left(1 - \frac{2t}{n}\right)^{n-1}$$

$$= \frac{2t}{n}\sum_{j=0}^{n-2}\left(1 - \frac{2t}{n}\right)^j \leq 2t \qquad \forall t \in [0, 1]$$

and $E_n[nF_1 I(nF_1 \leq t)] \leq tP_n(nF_1 \leq t) \leq t$. Hence by (3.22), there is a constant M_3 such that $\mathrm{Var}_n(\xi_n(t)) \leq M_3 t \ \forall t \in [0, 1]$. This together with (3.27) shows the existence of the constants M_4 and N such that $\mathrm{Var}_n(\xi_n(t)) \leq M_4 t e^{-\beta_1 t}\, \forall t \geq 0$ and $n > N$. Finally taking $M(t) = M_4 t e^{-\beta_1 t}$ completes the proof.

Proof of Theorem 5: Due to the condition on h, Part (a) follows immediately from the weak convergence of $\eta_n(t)$ (the empirical process of $\{nD_i\}$ to $\xi(t)$ under H_0. As for Part (b), it is sufficient to show that the process $\eta_n(t)$ converges weakly to the same limit as that of $\xi_n(t)$ under $\{H_{1n}\}$, which can be proved by using arguments similar to those in the proof of Theorem 3, except that the S_n in Theorem 3 should be replaced by $[t/nf(x)]^{1/d}$.

References

1. D. F. Andrews, R. Gnanadesikan and J. L. Warner (1973). Methods forcasting multivariate normality, *Multivariate Analysis III,* (P. R. Krishnaiah ed.) Academic press, New York, 95–116.
2. P. J. Bickel and L. Breiman (1983) Sum of functions of nearest neighbour distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probability* **11,** no. 1, 185–214.
3. M. Kuo and J. S. Rao (1981). Limit theory and efficiencies for tests based on higher order spacings. *Statistics: Applications and New Directions,* Proceedings of Golden Jubilee Conference of Indian Statistical Institute, Statistical Publishing Society, Calcutta, 333–352.
4. R. B. D'Agostino and M. A. Stephens (1986). *Goodness of fit techniques* Marcel Dekker, New York.
5. P. Hall (1986). On powerful distributional tests based on sample spacings. *J. Multivariate Analysis,* **19,** 201–224.
6. L. Holst and J. S. Rao (1981). Asymptotic spacings theory with applications to the two-sample problem. *Canadian J. Statist.,* **9,** 79–89.
7. S. R. Jammalamadala and R. C. Tiwari (1987). Efficiencies of some disjoint spacings tests relative to a chi-square test. *New perspectives in Theoretical and Applied statistics* (Eds. M. L. Puri, J. Vilaplana and W. Wertz) John Wiley, 311–317.
8. S. R. Jammalamadaka, X. Zhou and R. C. Tiwari (1989). Asymptotic efficiencies of spacings tests for goodness of fit. *Metrika,* **36,** 355–377.
9. J. A. Koziol (1986). Assessing multivariate normality- a compendium. *Commun. Statist.- Theory and Methods,* **15,** 2763–2783.
10. D. S. Moore and M. C. Spurill (1975). Unified large sample theory of general chi-squared statistics for tests of fit. *Ann. Statist.,* **3,** 599–616.
11. K. Pearson (1990). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from a random sample. *Phil. Mag. Sers.,* **50,** 157–172.
12. R. Pyke (1965). Spacings. *J. R. Statist. Soc. B,* **27,** 395–449.
13. M. F. Schilling (1983). Goodness of fit testing in R^m based on the weighted empirical distribution of certain nearest neighbour distances. *Ann. Statist.,* **11,** No. 1, 1–12.
14. G. R. Shorack and J. A. Wellner (1986). *Empirical processes with applications to statistics.* Wiley, New York.
15. M. Wells, S. R. Jammalamadaka and R. C. Tiwari (1992). Large sample theory of spacings statistics for tests of fit for composite hypothesis, To appear in *Jour. Roy. Statist. Soc.* (*Ser. B*).