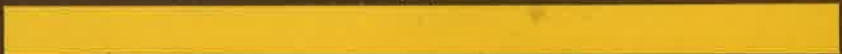


Probability Statistics and Design of Experiments



EDITED BY

R.R. BAHADUR

SOME GOODNESS OF FIT TESTS IN HIGHER DIMENSIONS BASED ON INTERPOINT DISTANCES

S. RAO JAMMALAMADAKA AND XIAN ZHOU
*Dept. of Statistics and Applied Probability,
University of California at Santa Barbara*

The goodness of fit problem for multidimensional data is considered based on "near-neighbor distances", i.e., all interpoint distances closer than a specified value r_n . The asymptotic distribution theory for two general classes of statistics based on these quantities is considered as well as their asymptotic relative efficiencies. Surprisingly, in the class of tests based only on small interpoint distances, the locally most powerful such statistic is merely the number of pairs that are closer than r_n . In another class of tests considered here, the optimal test performs as well as the likelihood ratio test.

1. INTRODUCTION

Goodness of fit tests for multivariate data have been studied by various authors, with special attention paid to multivariate normality. See for instance, Andrews et al (1973) and Koziol (1986). We study here tests for this general problem, based on the interpoint distances and derive their Pitman relative efficiencies. This enables us to pinpoint asymptotically optimal tests, within the classes of test statistics studied.

Let X_1, \dots, X_n be i.i.d. random vectors on \mathbf{R}^d with a common density $p(x)$, and $\|\cdot\|$ denote the usual Euclidean norm on \mathbf{R}^d . Jammalamadaka and Janson (1986) study the limiting distribution of the statistic

$$\sum_{1 \leq i < j \leq n} I(\|X_i - X_j\| < r_n) \quad (1.1)$$

where $I(\cdot)$ is the indicator function, as a special case of a generalized U -statistic. By an appropriate choice of r_n , we may call such pairs of (X_i, X_j) such that $\|X_i - X_j\| < r_n$, *near-neighbors* or r_n -close neighbors (which is

AMS (1980) subject classifications: Primary 62E20, 62G10, Secondary 62G20

Key words: Goodness of fit, multivariate data, interpoint distances, asymptotic relative efficiencies

somewhat analogous to the idea of higher order spacings in \mathbf{R}^1 as opposed to the *nearest-neighbors*, which are analogous in \mathbf{R}^1 , to the 1-step spacings. See for instance Kuo and Rao (1981) for tests based on spacings. In a parallel paper, we consider tests based on the nearest-neighbor distances, and derive results along the lines of Bickel and Breiman (1983) and Schilling (1986).

Here we will consider the following two families of test statistics based on "small" interpoint or r_n -close neighbor distances:

(i) F1: all statistics of the form

$$U_{n1} = U_{n1}(h) = \sum_{1 \leq i < j \leq n} h(r_n^{-1} \|X_i - X_j\|) \cdot I(\|X_i - X_j\| < r_n),$$

where h is a real-valued measurable bounded function defined on $[0, 1]$, and h is bounded and not almost everywhere zero on $[0, 1]$; $\{r_n\}$ is a sequence of numbers such that $r_n \rightarrow 0$ and $n^2 r_n^d \rightarrow \infty$ as $n \rightarrow \infty$. This family of test statistics is based on near-neighbor distances. The rate at which r_n goes to zero, is dictated by considerations of asymptotic normality of these statistics.

(ii) F2: all statistics of the form

$$U_{n2} = U_{n2}(h) = \sum_{1 \leq i < j \leq n} [h(X_i) + h(X_j)] I(\|X_i - X_j\| < r_n),$$

where h is a real-valued, measurable and bounded function defined on \mathbf{R}^d , not almost everywhere zero, and $\{r_n\}$ is as in (i).

In Section 2, the limiting distribution of the class of statistics U_{n1} and their asymptotic relative efficiencies (ARE's) are studied. It is found that in F1, the one with $h \equiv \text{constant}$ is optimal in the sense that it has the maximum efficacy. That is, the statistic given by (1.1) is the most efficient in the family F1. In Section 3, we study the AREs for the class of statistics F2. It is shown that the optimal test in this family F2 is asymptotically as efficient as the likelihood ratio test. Because of this fact, we saw no need to consider the larger class of statistics of the form:

$$\sum_{i < j} \psi(X_i, X_j) I(\|X_i - X_j\| < r_n).$$

One can study this class, but this requires more stringent assumptions on the function $\psi(\cdot, \cdot)$ as compared to those we make on $h(\cdot)$ in the definition of U_{n2} . The different assumptions and the results are illustrated through an application in Section 4.

A word about the notation: " $a_n \sim b_n$ " will mean that $a_n/b_n \rightarrow 1$ as $n \rightarrow \infty$ for two sequences of real numbers $\{a_n\}$ and $\{b_n\}$. We write \xrightarrow{d} for convergence in distribution.

2. THE ARE FOR THE STATISTICS IN FAMILY F1

Let H_0 be the null hypothesis given by $H_0: p(x) = p_0(x)$ where $p_0(x)$ is a specified density, and H_{1n} be a sequence of alternatives given by

$$H_{1n}: p(x) = p_n(x) = p_0(x) + v_n l_n(x), \quad (2.1)$$

where $v_n \rightarrow 0$ as $n \rightarrow \infty$ and l_n converges in L^1 and L^2 norms to some function l with

$$\int l(x) dx = 0.$$

Let E_0 and E_n denote the expectation under H_0 and H_{1n} respectively. Let f_n be a bounded measurable function defined on $\mathbf{R}^d \times \mathbf{R}^d$ which is symmetric (i.e. $f_n(x, y) = f_n(y, x)$ for $x, y \in \mathbf{R}^d$). Define

$$U_n = \sum_{1 \leq i < j \leq n} f_n(X_i, X_j),$$

$$g_n(x) = E_0 f_n(x, Y)$$

and

$$\sigma_n^2 = \frac{1}{2} n^2 [E_0 f_n^2 - (E_0 f_n)^2] + n^3 E_0 (g_n - E_0 f_n)^2. \quad (2.2)$$

From a slight generalization of Theorem 2.1 of Jammalamadaka and Janson (1986), we obtain

THEOREM 2.1 Suppose that as $n \rightarrow \infty$,

- (i) $\sup_{x, y} |f_n(x, y)| = o(\sigma_n)$.
 (ii) $\sup_x E_0 |f_n(x, Y)| = o(\sigma_n/n)$.

Then

$$\frac{U_n - \binom{n}{2} E_0 f_n}{\sigma_n} \xrightarrow{d} N(0, 1) \text{ under } H_0$$

and

$$\frac{U_n - \binom{n}{2} E_n f_n}{\sigma_n} \xrightarrow{d} N(0, 1) \text{ under } H_{1n}. \quad \square$$

Now take

$$f_n(x, y) = h(r_n^{-1} \|x - y\|) I(\|x - y\| < r_n) \quad (2.3)$$

as a special case. Then

$$E_0 f_n = \iint_{\|x-y\| < r_n} h(r_n^{-1} \|x - y\|) p_0(x) p_0(y) dx dy$$

$$\begin{aligned}
&= \int_{\|u\| < 1} h(\|u\|) \left[\int p_0(x) p_0(x + r_n u) dx \right] r_n^d du \\
&= r_n^d \int_{\|u\| < 1} h(\|u\|) (p_0 * p_0)(-r_n u) du, \quad (2.4)
\end{aligned}$$

where "*" denotes convolution.

Assume $p_0 \in L^2$. Then $p_0 * p_0$ is continuous. Note also that since h is bounded, the Lebesgue Dominated Convergence Theorem (LDCT) yields

$$\begin{aligned}
\int_{\|u\| < 1} h(\|u\|) (p_0 * p_0)(-r_n u) du &\rightarrow (p_0 * p_0)(0) \int_{\|u\| < 1} h(\|u\|) du \\
&= \int p_0^2(x) dx \int_{\|u\| < 1} h(\|u\|) du.
\end{aligned}$$

Thus by (2.4)

$$E_0 f_n \sim r_n^d \int p_0^2(x) dx \int_{\|u\| < 1} h(\|u\|) du. \quad (2.5)$$

Similarly

$$E_0 f_n^2 \sim r_n^d \int p_0^2(x) dx \int_{\|u\| < 1} h^2(\|u\|) du \quad (2.6)$$

and if $p_0 \in L^3$, then

$$\begin{aligned}
E_0 (g_n - E_0 f_n)^2 &\sim r_n^{2d} \left[\int p_0^3(x) dx - \left(\int p_0^2(x) dx \right)^2 \right] \\
&\quad \left[\int_{\|u\| < 1} h(\|u\|) du \right]^2 \\
&= r_n^{2d} \left[\int_{\|u\| < 1} h(\|u\|) du \right]^2 \text{Var}_0 [p_0(X)], \quad (2.7)
\end{aligned}$$

where Var_0 denotes the variance under H_0 . Substituting (2.5), (2.6) and (2.7) into (2.2) yields

$$\begin{aligned}
\sigma_n^2 &\sim \frac{1}{2} n^3 r_n^d \int_{\|u\| < 1} h^2(\|u\|) du \int p_0^2(x) dx \\
&\quad + n^3 r_n^{2d} \left[\int_{\|u\| < 1} h(\|u\|) du \right]^2 \text{Var}_0 [p_0(X)]. \quad (2.8)
\end{aligned}$$

Now we are ready to prove

THEOREM 2.2 Suppose $p_0 \in L^3, l \in L^3$ and let

$$\lim_{n \rightarrow \infty} \binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} = \mu, \tag{2.9}$$

where f_n is as given by (2.3). Then

$$\frac{U_{n1} - \binom{n}{2} E_0 f_n}{\sigma_n} \xrightarrow{d} \begin{cases} N(0, 1) \text{ under } H_0 \\ N(\mu, 1) \text{ under } H_{1n} \end{cases}$$

PROOF. Since $n^2 r_n^d \rightarrow \infty$ and h is not equal to zero almost everywhere on $[0, 1]$, (2.8) shows that

$$\sigma_n^2 \geq \frac{1}{2} n^2 r_n^d \int_{\|u\| < 1} h^2(\|u\|) du \int p_0^2(x) dx \rightarrow \infty.$$

Hence $\sup |f_n| \leq \|h\|_\infty = o(\sigma_n)$. Moreover, by the Cauchy-Schwarz inequality

$$\begin{aligned} E_0 |f_n(x, Y)| &= \int_{\|x-y\| < r_n} |h(r_n^{-1} \|x-y\|)| p_0(y) dy \\ &\leq \|h\|_\infty \left(\int_{\|x-y\| < r_n} p_0^2(y) dy \int_{\|x-y\| < r_n} dy \right)^{1/2} \\ &\leq \|h\|_\infty \left(\int_{\|x-y\| < r_n} p_0^2(y) dy \right)^{1/2} c_d^{1/2} r_n^{d/2}, \end{aligned}$$

where $c_d = \int_{\|u\| < 1} du = \pi^{d/2} / \Gamma(d/2 + 1)$. Thus

$$\sup_x E_0 |f_n(x, Y)| = o(r_n^{d/2}) = o(\sigma_n/n).$$

By Theorem 2.1 we have

$$\frac{U_{n1} - \binom{n}{2} E_0 f_n}{\sigma_n} \xrightarrow{d} N(0, 1)$$

under H_0 , and

$$\frac{U_{n1} - \binom{n}{2} E_0 f_n}{\sigma_n} = \frac{U_{n1} - \binom{n}{2} E_n f_n}{\sigma_n} + \binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \xrightarrow{d} N(\mu, 1)$$

under H_{1n} . □

In order to find the ARE of the statistic U_{n1} , we need to choose appropriate alternative sequence, i.e., the rate v_n in (2.1) as well as fine-tune the

choice of r_n , such that the limit μ in (2.9) exists and is neither zero nor infinity. This involves various cases, which we discuss after equation (2.11).

As in (2.4) and (2.5) and by the assumptions on p_0 and p_n we can obtain

$$E_n f_n \sim r_n^d \int_{\|u\| < 1} h(\|u\|) du \int [p_0(x) + v_n l(x)]^2 dx.$$

Case A: Assume $\int p_0(x) l(x) dx \neq 0$. Then

$$E_n f_n - E_0 f_n \sim 2r_n^d v_n \int_{\|u\| < 1} h(\|u\|) du \int p_0(x) l(x) dx. \quad (2.10)$$

For notational convenience, write

$$a(h) = \int_{\|u\| < 1} h(\|u\|) du.$$

Then by (2.8) and (2.10), the Pitman "efficacy"

$$\begin{aligned} & \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 \\ & \sim \frac{n^4}{4} \frac{4r_n^{2d} v_n^2 a^2(h) \left(\int p_0(x) l(x) dx \right)^2}{\frac{1}{2} n^2 r_n^d a(h^2) \int p_0^2(x) dx + n^3 r_n^{2d} a^2(h) \text{Var}_0(p_0(X))} \\ & = \frac{(n v_n^2) (n r_n^d) \left(\int p_0(x) l(x) dx \right)^2}{\frac{1}{2} a(h^2)/a^2(h) \int p_0^2(x) dx + n r_n^d \text{Var}_0(p_0(X))}. \end{aligned} \quad (2.11)$$

The limit of this expression in (2.11) will depend on the rate of r_n .

A(i) If $r_n^d \rightarrow 0$ such that $n r_n^d \rightarrow \infty$, then from (2.11) we can see that for $v_n = n^{-1/2}$,

$$\lim_{n \rightarrow \infty} \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 = \frac{\left[\int p_0(x) l(x) dx \right]^2}{\text{Var}_0(p_0(X))}.$$

Thus by Theorem 2.1,

$$\text{Eff}(U_{n1}) = \lim_{n \rightarrow \infty} \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 = \frac{\left[\int p_0(x) l(x) dx \right]^2}{\text{Var}_0(p_0(X))},$$

which does not depend on h . Therefore in this case all members in $F1$ are equally efficient.

A(ii) If $r_n^d \rightarrow 0$ such that $nr_n^d \rightarrow \alpha \in (0, \infty)$, then with $v_n = n^{-1/2}$,

$$\begin{aligned} \text{Eff}(U_{n1}) &= \lim_{n \rightarrow \infty} \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 \\ &= \frac{\alpha \left[\int p_0(x) l(x) dx \right]^2}{\frac{1}{2} a(h^2)/a^2(h) \int p_0^2(x) dx + \alpha \text{Var}_0(p_0(X))}. \end{aligned}$$

Thus maximizing $\text{Eff}(U_{n1})$ in this case is equivalent to maximizing

$$\frac{a^2(h)}{a(h^2)} = \frac{\left(\int_{\|u\| < 1} h(\|u\|) du \right)^2}{\int_{\|u\| < 1} h^2(\|u\|) du}. \quad (2.12)$$

By the Cauchy-Schwarz Inequality, the quantity in (2.12) is at most equal to

$$\frac{\int_{\|u\| < 1} h^2(\|u\|) du \int_{\|u\| < 1} du}{\int_{\|u\| < 1} h^2(\|u\|) du} = \int_{\|u\| < 1} du = c_d,$$

which happens at $h \equiv \text{constant}$. When $h \equiv 1$, $U_{n1}(1)$ is the same as that given in (1.1), which is the number of pairs of points within a distance r_n .

A(iii) If $nr_n^d \rightarrow 0$ then v_n should be taken as $(n^2 r_n^d)^{-1/2}$. Let $r_n^d = O(n^{-p})$, $1 < p < 2$. Then $v_n = (n^2 r_n^d)^{-1/2} = O(n^{-(2-p)/2})$. By Theorem 2.2, we get

$$\text{Eff}(U_{n1}) = \left[2 \frac{a^2(h)}{a(h^2)} \frac{\left[\int p_0(x) l(x) dx \right]^2}{\int p_0^2(x) dx} \right]^{1/(2-p)}$$

Thus again the U_{n1} with $h \equiv \text{constant}$ has the maximum efficacy in F1 which results in the test statistic (1.1).

Case B: Consider the case where $\int p_0(x) l(x) dx = 0$. Then by (2.10)

$$E_n f_n - E_0 f_n \sim r_n^d v_n^2 a(h) \int l^2(x) dx.$$

Hence

$$\begin{aligned} & \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 \\ & \sim \frac{n^4}{4} \frac{r_n^{2d} v_n^4 a^2(h) \left(\int I^2(x) dx \right)^2}{\frac{1}{2} n^2 r_n^d a(h^2) \int p_0^2(x) dx + n^3 r_n^{2d} a^2(h) \text{Var}_0(p_0(X))} \\ & = \frac{(n v_n^4) (n r_n^d) \left(\int I^2(x) dx \right)^2}{\frac{1}{2} a(h^2)/a^2(h) \int p_0^2(x) dx + n r_n^d \text{Var}_0(p_0(X))}. \end{aligned}$$

Thus when $\text{Var}_0(p_0(X)) > 0$, we have

B(i) If $n r_n^d \rightarrow \infty$, then for $v_n = n^{-1/4}$,

$$\text{Eff}(U_{n1}) = \frac{\left(\int I^2(x) dx \right)^2}{\text{Var}(p_0(X))},$$

which is again independent of h .

B(ii) If $n r_n^d \rightarrow \alpha \in (0, \infty)$, for $v_n = n^{-1/4}$,

$$\text{Eff}(U_{n1}) = \frac{\alpha \left[\int I^2(x) dx \right]^2}{\frac{1}{2} a(h^2)/a^2(h) \int p_0^2(x) dx + \alpha \text{Var}_0(p_0(X))}.$$

B(iii) If $n r_n^d \rightarrow 0$, $r_n^d = O(n^{-p})$ ($1 < p < 2$), then for $v_n = n^{-(2-p)/4}$,

$$\text{Eff}(U_{n1}) = \left[2 \frac{a^2(h)}{a(h^2)} \left[\int I^2(x) dx \right]^2 \right]^{1/(2-p)}.$$

When $\text{Var}_0(p_0(X)) = 0$, v_n and $\text{Eff}(U_{n1})$ are the same as in case (iii) above. In all these cases, we see that the optimal $h(\cdot)$ is obtained by maximizing $[a^2(h)/a(h^2)]$. Since this optimal choice is to pick $h \equiv 1$, the locally most powerful statistic in all these cases is still $U_{n1}(1)$ given in (1.1). In other words,

$$\text{ARE}(U_{n1}(1), U_{n1}(h)) \geq 1$$

for all $U_{n1}(h) \in F1$.

3. THE ARE FOR THE STATISTICS IN FAMILY F2

For U_{n2} , take $f_n(x, y) = [h(x) + h(y)] I(\|x - y\| < r_n)$ so that

$$U_{n2} = \sum_{1 \leq i < j \leq n} f_n(X_i, X_j).$$

As in the proof of Theorem 2.1, we can verify that f_n as above, still satisfies the conditions of Theorem 2.1. Therefore, for the class of statistics U_{n2} , we have the following result, which parallels Theorem 2.2 for the class $F1$.

THEOREM 3.1 Suppose $p_0 \in L^3$, $l \in L^3$ and

$$\lim_{n \rightarrow \infty} \binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} = \mu.$$

Then

$$\frac{U_{n2} - \binom{n}{2} E_0 f_n}{\sigma_n} \xrightarrow{d} \begin{cases} N(0, 1) \text{ under } H_0 \\ N(\mu, 1) \text{ under } H_{1n} \end{cases} \quad \square$$

With the assumption on p_0 , p_n and $l \in L^3$, we now evaluate μ and σ_n^2 .

$$\begin{aligned} E_0 f_n &= \iint_{\|x-y\| < r_n} [h(x) + h(y)] p_0(x) p_0(y) dx dy \\ &= r_n^d \iint_{\|u\| < 1} [h(x) + h(x + r_n u)] p_0(x) p_0(x + r_n u) du dx \quad (3.1) \end{aligned}$$

and

$$\begin{aligned} E_0 f_n &= r_n^d \iint_{\|u\| < 1} [h(x) + h(x + r_n u)] [p_0(x) + v_n l_n(x)] [p_0(x + r_n u) \\ &\quad + v_n l_n(x + r_n u)] du dx \end{aligned}$$

Hence

$$\begin{aligned} E_n f_n - E_0 f_n &\sim r_n^d \iint_{\|u\| < 1} 2h(x) 2v_n p_0(x) l_n(x) du dx \\ &\sim 4c_d r_n^d v_n \int h(x) p_0(x) l(x) dx. \quad (3.2) \end{aligned}$$

Similarly

$$E_0 f_n^2 \sim 4c_d r_n^d \int h^2(x) p_0^2(x) dx \quad \text{and}$$

$$E_0 (g_0 - E_0 f_n)^2 \sim 4r_n^{2d} c_d^2 \text{Var}_0 [h(X) p_0(X)].$$

Thus, from (2.2)

$$\sigma_n^2 \sim 2n^2 c_d r_n^d \int h^2(x) p_0^2(x) dx + 4n^3 r_n^{2d} c_d^2 \text{Var}_0 [h(X) p_0(X)] \quad (3.3)$$

and consequently,

$$\begin{aligned} & \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 \\ & \sim \frac{c_d (n v_n^2) (n r_n^d) \left(\int h(x) p_0(x) l(x) dx \right)^2}{\frac{1}{2} \int h^2(x) p_0^2(x) dx + c_d n r_n^d \text{Var}_0(h(X) p_0(X))} \end{aligned} \quad (3.4)$$

We have two cases now:

Case (i) Let $n r_n^d \rightarrow \infty$ and choose $v_n = n^{-1/2}$. Assume that $l(x)/p_0^2(x)$ is bounded. Then,

$$\text{Eff}(U_{n2}) = \lim_{n \rightarrow \infty} \left[\binom{n}{2} \frac{E_n f_n - E_0 f_n}{\sigma_n} \right]^2 = \frac{\left(\int h(x) p_0(x) l(x) dx \right)^2}{\text{Var}(h(X) p_0(X))}.$$

If we define

$$r(x) = l(x)/p_0(x),$$

then $E_0 r(X) = \int l(x) dx = 0$ and

$$\begin{aligned} \text{Eff}(U_{n2}) &= \frac{E_0^2[h(X) p_0(X) r(X)]}{\text{Var}(h(X) p_0(X))} = \frac{\text{Cov}_0^2[h(X) p_0(X), r(X)]}{\text{Var}_0(h(X) p_0(X))} \\ &\leq \text{Var}(r(X)) = \int l^2(x)/p_0(x) dx \end{aligned} \quad (3.5)$$

for all $U_{n2} \in \mathbf{F2}$, with the equality holding when $h(x) p_0(x) = r(x)$. Thus the statistic $U_{n2}(h)$ with $h(x) = r(x)/p_0(x) = l(x)/p_0^2(x)$ has the maximum efficacy, which is given by the last term in (3.5).

Consider the statistic

$$V_n = \sum_{i=1}^n r(X_i) = \sum_{i=1}^n l(X_i)/p_0(X_i),$$

which is asymptotically equivalent to the likelihood ratio test for testing the simple H_0 versus the simple alternative H_{1n} . By the Central Limit Theorem we can obtain that for $v_n = n^{-1/2}$,

$$\begin{aligned} \text{Eff}(V_n) &= \frac{\left(\int r(x) l(x) dx \right)^2}{\text{Var}_0(r(X))} = \frac{E_0^2(r^2(X))}{\text{Var}_0(r(X))} = E_0(r^2(X)) \\ &= \int l^2(x)/p_0(x) dx. \end{aligned}$$

This is the same as the right side of (3.5). Thus the U_{n2} with $h = l/p_0^2$ has the same efficacy as that of the likelihood ratio test if $n r_n^d \rightarrow \infty$.

Case (ii) Let $nr_n^d \rightarrow 0$ and choose $r_n^d = O(n^{-\nu})$, $\nu_n = (n^2 r_n^d)^{-1/2} = O(n^{-(2-\nu)/2})$. Assume that $l(x)/p_0(x)$ is bounded. Then, from (3.4)

$$\text{Eff}(U_{n_2}) = 2c_d \frac{\left(\int h(x) p_0(x) l(x) dx \right)^2}{\int h^2(x) p_0^2(x) dx} \leq 2c_d \int l^2(x) dx$$

for all $U_{n_2} \in \mathbf{F2}$, with the equality holding when $h(x) p_0(x) = l(x)$.

Thus the statistic U_{n_2} with $h(x) = l(x)/p_0(x)$ has the maximum efficacy, which is equal to $2c_d \int l^2(x) dx$. The resulting optimal statistic is

$$\sum_{1 \leq i < j \leq n} \left[\frac{l(X_i)}{p_0(X_i)} + \frac{l(X_j)}{p_0(X_j)} \right] I(\|X_i - X_j\| < r_n).$$

4. EXAMPLE

Now let us consider the null hypothesis

$$H_0: p_0(x, y) = K e^{-(x^2+y^2)/2} I(x^2+y^2 \leq R)$$

against a sequence of alternatives

$$H_{1n} \quad p_n(x, y) = K_n \exp \left(-(x^2 + 2\rho_n xy + y^2)/2(1 - \rho_n^2) \right) I(x^2 + y^2 \leq R)$$

where K and K_n are the constants such that $\int p_0 = \int p_n = 1$ and $\{\rho_n\}$ is a sequence of numbers in $[0, 1]$ converging to zero.

This is a truncated version of the problem of testing the bivariate normal distribution $N(0, 0, 1, 1, 0)$ versus $N(0, 0, 1, 1, \rho_n)$. The use of truncation is to satisfy the assumption that l/p_0^2 is bounded. It may not be as natural as the untruncated case, but should be good enough for all practical purposes because R can be chosen as large as we want.

In order to apply the results of the previous sections, we need to rewrite p_n in the form of $p_0 + \nu_n l_n$. This can be done simply by putting

$$l_n = \nu_n^{-1} (p_n - p_0).$$

In both families $\mathbf{F1}$ and $\mathbf{F2}$, we see that in order to get maximum efficacy, r_n should be chosen such that $nr_n^d \rightarrow \infty$. In this example, $d = 2$, hence we choose $r_n = n^{-\delta}$ with $0 < \delta < 1/2$. Then with $\rho_n = \nu_n = n^{-1/2}$, it can be shown that

$$l_n(x, y) \rightarrow l(x, y) = xy p_0(x, y) \quad (4.1)$$

in L^1 and L^2 norms. The proof of this convergence is not difficult to show but rather lengthy and therefore the details are omitted here. The other conditions on p_0 and l are easily seen to be satisfied and due to the trunca-

tion, l/p_0^2 is indeed bounded. Thus the best U_{n2} test statistic based on the bivariate sample $\{(X_i, Y_i), i = 1, \dots, n\}$ is

$$U_{n2}^* = \sum_{1 \leq i < j \leq n} [h(X_i, Y_i) + h(X_j, Y_j)] I[(X_i - X_j)^2 + (Y_i - Y_j)^2 < n^{-2\delta}] \quad (4.2)$$

where $0 < \delta < 1/2$ and

$$h(x, y) = \frac{l(x, y)}{p_0^2(x, y)} = \frac{1}{K} xy \exp((x^2 + y^2)) I(x^2 + y^2 \leq R^2).$$

The efficacy of U_{n2} in (4.2) is

$$\text{Eff}(U_{n2}^*) = \int \int l^2/p_0 = \iint_{x^2 + y^2 \leq R^2} Kx^2y^2 \exp(-(x^2 + y^2)/2) dx dy.$$

If R is large, then

$$\text{Eff}(U_{n2}^*) \cong \frac{1}{2\pi} \iint x^2y^2 \exp(-(x^2 + y^2)/2) dx dy = 1.$$

To obtain a critical region, we can use the result of Theorem 3.1 that under H_0 , viz.

$$\frac{U_{n2}^* - \binom{n}{2} E_0 f_n}{\sigma_n} \xrightarrow{d} N(0, 1).$$

By (3.1) and (3.3), it can be shown that

$$\begin{aligned} E_0 f_n &= 2r_n^d \left[c_d \int hp_0^2 + O(r_n^2) \right] = 2n^{-2\delta} \left[c_d \int l + O(r_n^2) \right] \\ &= 2n^{-2\delta} O(n^{-2\delta}) = O(n^{-4\delta}) \end{aligned}$$

and

$$\sigma_n^2 \sim 4n^3 r_n^{2d} c_d^2 \left[\int h^2 p_0^3 - \left(\int hp_0^2 \right)^2 \right] = 4n^3 n^{-4\delta} c_d^2 \int l^2/p = 4n^{3-4\delta} \pi^2.$$

Thus

$$\frac{\binom{n}{2} E_0 f_n}{\sigma_n} \sim \frac{n^2 O(n^{-4\delta})}{2n^{3/2-2\delta} \pi} = O(n^{1/2-2\delta}).$$

If, for instance, we take $\delta = 3/8$ so that

$$\binom{n}{2} E_0 f_n / \sigma_n = O(n^{-1/4}) \rightarrow 0,$$

then

$$U_{n2}/2\pi n^{3/4} \xrightarrow{d} N(0, 1).$$

Thus an asymptotic critical region with significance level α in this case, is

given by

$$U_{n2}/2\pi n^{3/4} > z_{1-\alpha}$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$ th quantile of the standard normal distribution function.

One can also construct tests using the U_{n1} statistic. However, because

$$\iint l(x, y)p_0(x, y)dxdy = K^2 \iint_{x^2+y^2 \leq R^2} xy \exp(-x^2 - y^2) dxdy \cong 0,$$

v_n has to be chosen as $n^{-1/4}$. Thus any U_{n1} test can only detect the alternatives converging to the null hypothesis at a rate of $n^{-1/4}$. Compare it with the convergence rate of $n^{-1/2}$ for U_{n2}^* , we see that $ARE(U_{n1}, U_{n2}^*) = 0$. Hence with large sample size, tests in the family **F1** are far less efficient than U_{n2}^* in this example.

5. CONCLUSIONS

The maximum efficacy that a U_{n1} test could possibly achieve under $v_n = n^{-1/2}$ is

$$\frac{\left(\int p_0(x) l(x) dx\right)^2}{\text{Var}_0(p_0(X))} = \frac{\text{Cov}_0^2(p_0(X), l(X)/p_0(X))}{\text{Var}_0(p_0(X))} < \text{Var}_0(l(X)/p_0(X)) = \int \frac{l^2(x)}{p_0(x)} dx \tag{5.1}$$

where the inequality must be strict. Otherwise $l(x)$ would be proportional to $p_0^2(x)$, which is impossible because l has to satisfy $\int l(x)dx = 0$. The last expression in (5.1) is just the efficacy of the best U_{n2} statistic. Thus it is clear that the class of tests in the family **F1** are not as efficient as the likelihood ratio test as well as the best test in **F2**. Indeed this is to be expected from the fact that the optimal test in **F1** does not take into account the $p_0(x)$ and the $l(x)$. But on the other hand, the optimal test in **F1**, which is independent of p_0 and l and uses only the number of pairs that are closer than r_n , can give a quick test.

REFERENCES

Andrews, D.F., Gnanadesikan, R. and Warner, J.L. (1973). Methods for assessing multivariate normality, *Multivariate Analysis III*, (P.R. Krishnaiah Ed.) Academic Press, New York, 95-116.

Bickel, P.J. and Breiman, L. (1983). Sum of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test *Ann. Probability* 11, No. 1, 185-214.

Jammalamadaka, S.R. and Janson, S. (1986). Limit theorems for a triangular scheme of U -statistics with applications to inter-point distances. *Ann. Probability* 14, 1347-1359.

- Koziol, J.A. (1986). Assessing multivariate normality—a compendium, *Commun. Statist.—Theory and Methods*, 15, 2763–2783.
- Kuo, M. and Rao, J.S. (1981). Limit theory and efficiencies for tests based on higher order spacings. *Statistics—Applications and New Directions*, Proceedings of the Golden Jubilee Conference of the Indian Statistical Institute, Statistical Publishing Society, Calcutta, 333–352.
- Schilling, M.F. (1983). Goodness of fit testing in R^m based on the weighted empirical distribution of certain nearest neighbor distances. *Ann. Statist.* 11, 1–12.