AN ALIGNED GOODNESS OF FIT TEST FOR THE MULTIVARIATE TWO-SAMPLE MODEL: LOCATIONS UNKNOWN

by

Pranab Kumar Sen

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1463

June 1984

AN ALIGNED GOODNESS OF FIT TEST FOR THE MULTIVARIATE TWO-SAMPLE MODEL : LOCATIONS UNKNOWN *

Pranab Kumar SEN

1. INTRODUCTION

Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_m$ be m independent and identically distributed (i.i.d.) random vectors (r.v.) with a continuous distribution function (d.f.) F, defined on the Euclidean space E^p, for some $p \geq 1$. Also, let $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_n$ be n i.i.d.r.v. with a continuous d.f. G, defined on E^p. We consider the model :

$$(1.1) \quad F(\underset{\sim}{x}) = F_o(\underset{\sim}{x} - \underset{\sim}{\theta}_1) \quad \text{and} \quad G(\underset{\sim}{x}) = G_o(\underset{\sim}{x} - \underset{\sim}{\theta}_2), \; \underset{\sim}{x} \in E^p,$$

where $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$ are unknown (location) parameters and the forms of F_o and G_o are also not known. The problem is to test for the hypotheses :

$$(1.2) \quad H_0 : F_o = G_o \quad \text{against} \quad H_1 : F_o \neq G_o,$$

treating $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$ as nuisance parameters. For $\underset{\sim}{\theta}_1 = \underset{\sim}{\theta}_2$, under H_0, F and G are the same, so that one may use the usual two-sample Kolmogorov - Smirnov test statistic for testing H_o against H_1. If $F_m(x)$ and $G_n(x)$ be respectively the first and second sample empirical d.f., then, this statistic can conveneiently be written as

$$(1.3) \quad K^+_{mn} = \sup\{ (mn/N)^{\frac{1}{2}} [F_m(\underset{\sim}{x}) - G_n(\underset{\sim}{x})] : \underset{\sim}{x} \in E^p \} \quad \text{(one-sided case)}$$

$$(1.4) \quad K_{mn} = \sup\{ (mn/N)^{\frac{1}{2}} | F_m(\underset{\sim}{x}) - G_n(\underset{\sim}{x})| : \underset{\sim}{x} \in E^p \} \quad \text{(two-sided case)}$$

where N = m+n is the total sample size. For the univariate case (i.e., p = 1), under H_0, the distribution of K^+_{mn} or K_{mn} does not depend on F (continuous), and hence , they are genuinely distribution-free when F and G are the same.

* Work partially supported by the (U.S.) Office of Naval Research, Contract N00014-83-K-0387.

For small values of m and n , the exact distribution of K_{mn}^{+} or K_{mn} (when F=G) can be obtained by direct enumeration ; we may refer to Hodges(1957) for a survey of computational methods and some numerical studies. Vincze(1960) has also presented a survey of some of these results. For large sample sizes, we have for every $t \geq 0$, under $F = G$,

$$(1.5) \quad \lim_{m,n\to\infty} P\{ K_{mn}^{+} \geq t \} = \exp(-2t^2) ,$$

$$(1.6) \quad \lim_{m,n\to\infty} P\{ K_{mn} \geq t \} = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2t^2) ,$$

and, actually, these tails dominate the exact distributions for the finite sample case. In the multivariate case (i.e., $p \geq 2$), this exact distribution-freeness of K_{mn}^{+} or K_{mn} is not generally true (excepting when all the p coordinates of the $\underset{\sim}{X}_i$ (and $\underset{\sim}{Y}_j$) are all mutually independent. However, when $F = G$, even in the multivariate case, all the N vectors $\underset{\sim}{X}_1,\ldots,\underset{\sim}{X}_m, \underset{\sim}{Y}_1,\ldots,\underset{\sim}{Y}_n$ are i.i.d.r.v. (with the d.f. F), so that the Chatterjee-Sen(1964) rank-permutation principle can be adapted to generate the permutational distribution of K_{mn}^{+} or K_{mn} over the set of N! (conditionally) equally likely perumtations of these N vectors, and this yields a conditionally (permutationally) distribution-free test for $F = G$. This task becomes prohibitively labourious when N increases, so that one is inclined to use suitable approximations to this permutation distribution. This has been studied in detail in Bickel (1969), and that reveals the asymptotic distribution-freeness of K_{mn}^{+} and K_{mn} when $F = G$.

With respect to the model (1.1), when $\underset{\sim}{\theta}_1 \neq \underset{\sim}{\theta}_2$, under H_0 in (1.2), F and G need not be the same, and hence, the test based on K_{mn}^{+} or K_{mn} may not be valid (may even be inconsistent). One possibility to overcome this problem is to consider suitable estimators $\hat{\underset{\sim}{\theta}}_{1,m}$ and $\hat{\underset{\sim}{\theta}}_{2,n}$ of $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$, respectively, form the residuals(vectors) $\underset{\sim}{X}_i - \hat{\underset{\sim}{\theta}}_{1,m}$, i= 1,...,m and $\underset{\sim}{Y}_j - \hat{\underset{\sim}{\theta}}_{2,n}$, j= 1,..., n, and to base a test on these aligned observations. Within each sample, the residual vectors are generally not independent (they are symmetrically

dependent (or exchangeable) with their marginal distribution and the dependence pattern both depending on the underlying d.f. and the sample size m or n), but, residuals from the other samples are independent of each other. This distorts the exchangeability of all the N residual vectors (even when $F_0 = G_0$ and /or $m = n$), so that the rank-permutation principle may not be applicable . From the asymptotic distributional point of view, for the two-sample Kolmogorov-Smirnov statistics based on such aligned observations, even if one chooses some efficient estimators of $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$, for the related (aligned) empirical distributional processes, the covariance functions may become quite involved and no simple distributional theory may therefore be available. This problem has been studied in the univariate case in Sen(1984) where a variant form of the Kolmogorov-Smirnov statistic has been used (under an additional assumption of symmetry of F_0 and G_0) , and it has been shown that this alternative one is asymptotically distribution-free under quite general regularity conditions.

The primary objective of the current study is to consider some multivariate generalizations of aligned Kolmogorov-Smirnov type tests , to validly apply them for the testing problem in (1.2), and to present their asymptotic propoerties too. Along with the preliminary notions, the proposed tests are presented in Section 2. Asymptotic properties of the tests are then considered in Section 3. Some general comments are made in the concluding section.

2. THE PROPOSED TESTS

To eliminate the nuisance parameters from the testing situation, first, we consider some suitable estimators of the locations $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$. Let $\hat{\underset{\sim}{\theta}}_{1,m}$ and $\hat{\underset{\sim}{\theta}}_{2,n}$ be two arbitrary *translation-invariant* estimators of $\underset{\sim}{\theta}_1$ and $\underset{\sim}{\theta}_2$, respectively. We assume them to be'square-root n' consistent , that is ,

$$(2.1) \qquad m^{\frac{1}{2}}||\hat{\underset{\sim}{\theta}}_{1,m} - \underset{\sim}{\theta}_1|| = O_p(1) \quad \text{and} \quad n^{\frac{1}{2}}||\hat{\underset{\sim}{\theta}}_{2,n} - \underset{\sim}{\theta}_2|| = O_p(1) ,$$

where $||\ .||$ stands for the Euclidean norm, and we assume further that there exists a positive λ_0 $(0 < \lambda_0 < \frac{1}{2})$, such that

(2.2) $\quad 0 < \lambda_o \leq \lambda_N = m/N \leq 1 - \lambda_o < 1, \quad N \geq N_o .$

Consider then the residuals (vectors)

(2.3) $\quad \hat{\underset{\sim}{X}}_i = \underset{\sim}{X}_i - \hat{\underset{\sim}{\theta}}_{1,m}$, i=1,...,m; $\hat{\underset{\sim}{Y}}_j = \underset{\sim}{Y}_j - \hat{\underset{\sim}{\theta}}_{2,n}$, j=1,...,n.

For these aligned observations, we consider the empirical d.f.'s

(2.4) $\quad \hat{F}_m(\underset{\sim}{x}) = m^{-1}\Sigma_{i=1}^m I(\hat{\underset{\sim}{X}}_i \leq \underset{\sim}{x})$, $\hat{G}_n(\underset{\sim}{x}) = n^{-1}\Sigma_{j=1}^n I(\hat{\underset{\sim}{Y}}_j \leq \underset{\sim}{x})$, $\underset{\sim}{x} \in E^p$,

where $\underset{\sim}{a} \leq \underset{\sim}{b}$ means the coordinatewise inequality $a_i \leq b_i$, i=1,...,p. Further, let us define

(2.5) $\quad \hat{\underset{\sim}{X}}_i^* = (\,|\hat{X}_{i1}|,\ldots,|\hat{X}_{ip}|\,)'$ and $\hat{\underset{\sim}{Y}}_j^* = (|\hat{Y}_{j1}|,\ldots,|\hat{Y}_{jp}|)'$,

and denote the empirical d.f.'s for these observations by

(2.6) $\quad \hat{F}_m^*(\underset{\sim}{x}) = m^{-1}\Sigma_{i=1}^m I(\hat{\underset{\sim}{X}}_i^* \leq \underset{\sim}{x})$ and $\hat{G}_n^*(\underset{\sim}{x}) = n^{-1}\Sigma_{j=1}^n I(\hat{\underset{\sim}{Y}}_j^* \leq \underset{\sim}{x})$, $\underset{\sim}{x} \in E_p^*$,

where E_p^* , the positive quadrant of E^p, is defined by

(2.7) $\quad E_p^* = \{(x_1,\ldots,x_p) : 0 \leq x_j < \infty,\ j=1,\ldots,p\}$ $(\subset E^p)$.

Our proposed tests are based on the empirical d.f.'s in (2.6). Specifically, we consider Kolmogorov-Smirnov Type test statistics , though others may be used in the same manner. Define

(2.8) $\quad \hat{K}_{mn}^{*+} = \sup\{\,(mn/N)^{\frac{1}{2}}\,[\,\hat{F}_m^*(\underset{\sim}{x}) - \hat{G}_n^*(\underset{\sim}{x})] : \underset{\sim}{x} \in E_p^*\,\}$

(2.9) $\quad \hat{K}_{mn}^{*} = \sup\{\,(mn/N)^{\frac{1}{2}}\,|\,\hat{F}_m^*(\underset{\sim}{x}) - \hat{G}_n^*(\underset{\sim}{x})\,| : \underset{\sim}{x} \in E_p^*\,\}.$

Our primary concern is to study the properties of the test statistics $\hat{K}_{mn}^{*+}$ and $\hat{K}_{mn}^{*}$ with a view to testing for the hypotheses in (1.2). In this context, we define the true residuals by

(2.10) $\quad \underset{\sim}{X}_i^o = \underset{\sim}{X}_i - \underset{\sim}{\theta}_1$, i=1,...,m and $\underset{\sim}{Y}_j^o = \underset{\sim}{Y}_j - \underset{\sim}{\theta}_2$, j=1,...,n ,

Then, parallel to (2.4)-(2.6), we define

(2.11) $\quad F_m^o(\underset{\sim}{x}) = m^{-1}\Sigma_{i=1}^m I(\underset{\sim}{X}_i^o \leq \underset{\sim}{x})$, $G_n^o(\underset{\sim}{x}) = n^{-1}\Sigma_{j=1}^n I(\underset{\sim}{Y}_j^o \leq \underset{\sim}{x})$, $\underset{\sim}{x} \in E^p$,

(2.12) $\quad \underset{\sim}{X}_i^{o*} = (\,|X_{i1}^o|,\ldots,|X_{ip}^o|\,)'$, $\underset{\sim}{Y}_j^{o*} = (\,|Y_{j1}^o|,\ldots,|Y_{jp}^o|)'$,

(2.13) $\quad F_m^{o*}(\underset{\sim}{x}) = m^{-1}\Sigma_{i=1}^m I(\underset{\sim}{X}_i^{o*} \leq \underset{\sim}{x})$ and $G_n^{o*}(\underset{\sim}{x}) = n^{-1}\Sigma_{i=1}^n I(\underset{\sim}{Y}_j^{o*} \leq \underset{\sim}{x})$, $\underset{\sim}{x} \in E_p^*$.

Note that under H_0 in (1.2), both $\underset{\sim}{X}_i^o$ and $\underset{\sim}{Y}_j^o$ have the common (unknown) d.f. F_o, so that $\underset{\sim}{X}_i^{o*}$ and $\underset{\sim}{Y}_j^{o*}$ have also a common d.f., which we denote by F_o^* . Parallel

to (2.8) and (2.9), we define

$$(2.14)\qquad K^{*+}_{mn} = \sup\{\ (mn/N)^{\frac{1}{2}}\ [\ F^{o*}_m(\underset{\sim}{x}) - G^{o*}_n(\underset{\sim}{x})]\ :\ \underset{\sim}{x} \in E^*_p\ \},$$

$$(2.15)\qquad K^{*}_{mn} = \sup\{\ (mn/N)^{\frac{1}{2}}\ |\ F^{o*}_m(\underset{\sim}{x}) - G^{o*}_n(\underset{\sim}{x})|\ :\ \underset{\sim}{x} \in E^*_p\ \}\ .$$

Our basic goal is to establish a general asymptotic (stochastic) equivalence result for $\hat{K}^{*+}_{mn}$ and K^{*+}_{mn} (as well as $\hat{K}^{*}_{mn}$ and K^{*}_{mn}), and this will enable us to study the asymptotic properties of the proposed test statistics thgrough those of K^{*+}_{mn} and K^{*}_{mn}.

We may recall that the estimators $\hat{\underset{\sim}{\theta}}_{1,m}$ and $\hat{\underset{\sim}{\theta}}_{2,n}$ are assumed to be translation-invariant, so that the residuals in (2.3) are translation-invariant too. Thus, without any loss of generality, we may assume that

$$(2.16)\qquad \underset{\sim}{\theta}_1 = \underset{\sim}{\theta}_2 = \underset{\sim}{0}\ .$$

We assume further that the d.f. F_o and G_o are both diagonally symmetric about the origin, that is, $\underset{\sim}{X}^o_i$ and $(-1)\underset{\sim}{X}^o_i$ (and $\underset{\sim}{Y}^o_j$ and $(-1)\underset{\sim}{Y}^o_j$) both have the common d.f. F_o (and G_o). We also assume that the d.f. F_o and G_o both have uniformly continuous probability density functions (p.d.f.), f_o and g_o, respectively, almost everywhere (a.e.). With these preliminary notions and basic assumptions, we proceed on to study the general properties of the proposed tests.

3. ASYMPTOTIC PROPERTIES OF THE PROPOSED TESTS

First, we consider the following asymptotic equivalence result which provides the access to our subsequent analysis.

Theorem 1. Under the regularity conditions of Section 2, as $N \to \infty$,

$$(3.1)\qquad \sup\{\ m^{\frac{1}{2}}\ |\hat{F}^{*}_m(\underset{\sim}{x}) - F^{o*}_m(\underset{\sim}{x})|\ :\ \underset{\sim}{x} \in E^*_p\ \} \xrightarrow{p} 0\ ,$$

$$(3.2)\qquad \sup\{\ n^{\frac{1}{2}}\ |\hat{G}^{*}_n(\underset{\sim}{x}) - G^{o*}_n(\underset{\sim}{x})|\ :\ \underset{\sim}{x} \in E^*_p\ \} \xrightarrow{p} 0\ .$$

Proof. We shall prove the first assertion in (3.1); the proof of (3.2) follows precisely on the same line. By virtue of (2.1), we may write

$$(3.3)\qquad \underset{\sim}{t}_m = m^{\frac{1}{2}}(\hat{\underset{\sim}{\theta}}_{1,m} - \underset{\sim}{\theta}_1)\ ,\ \text{so that}\ \ ||\underset{\sim}{t}_m|| = O_p(1)\ .$$

Moreover, note that by definition,

$$(3.4)\quad \hat{F}^{*}_{m}(\underset{\sim}{x}) = \hat{F}^{*}_{m}(x_1,\ldots,x_p)$$
$$= \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ \hat{F}_m(\,i_jx_j+(1-i_j)(-x_j-),1\le j\le p\,),$$

$$(3.5)\quad F^{O*}_{m}(x) = F^{O*}_{m}(x_1,\ldots,x_p)$$
$$= \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-),\ 1\le j\le p)\ ,$$

for every $\underset{\sim}{x}\ \varepsilon\ E^{*}_{p}$. Therefore, for every $\underset{\sim}{x}\ \varepsilon\ E^{*}_{p}$,

$$(3.6)\quad m^{\frac{1}{2}}\ [\ \hat{F}^{*}_{m}(\underset{\sim}{x}) - F^{O*}_{m}(\underset{\sim}{x})\ ]$$
$$= \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ m^{\frac{1}{2}}\ [\ \hat{F}_m(i_jx_j+(1-i_j)(-x_j-),1\le j\le p)$$
$$- F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-),1\le j\le p)\ ]\ .$$

Recall further that

$$(3.7)\quad \hat{F}_m(\underset{\sim}{x}) = F^{O}_{m}(\ \underset{\sim}{x} + m^{-\frac{1}{2}}\ \underset{\sim m}{t}\ )\ ,\ \text{for every}\ \underset{\sim}{x}\ \varepsilon\ E^{p}.$$

Therefore, by (3.6) and (3.7), we obtain that

$$(3.8)\quad m^{\frac{1}{2}}[\ \hat{F}^{*}_{m}(\underset{\sim}{x}) - F^{O*}_{m}(\underset{\sim}{x})\ ]$$
$$= \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ m^{\frac{1}{2}}\{[\ F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-)+m^{-\frac{1}{2}}t_j,1\le j\le p)$$
$$- F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-),\ 1\le j\le p)\ ]\}$$
$$= \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ m^{\frac{1}{2}}\{[\ F_{o}(i_jx_j+(1-i_j)(-x_j-)+m^{-\frac{1}{2}}t_j,1\le j\le p)$$
$$- F_{o}(i_jx_j+(1-i_j)(-x_j-),1\le j\le p)]$$
$$+[\ F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-)+m^{-\frac{1}{2}}t_j,\ 1\le j\le p) - F_{o}(i_jx_j+(1-i_j)(-x_j-)+m^{-\frac{1}{2}}t_j,1\le j\le p)$$
$$-F^{O}_{m}(i_jx_j+(1-i_j)(-x_j-),1\le j\le p) + F_{o}(i_jx_j+(1-i_j)(-x_j-),1\le j\le p)\ ]\ \}\ .$$

Now, by virtue of the assumed diagonal symmetry of F_o, $f_o(\underset{\sim}{x}) = f_o((-1)\underset{\sim}{x})$, for every $\underset{\sim}{x}\ \varepsilon\ E^{p}$, so that using (3.3) along with the a.e. uniform continuity of f_o , we obtain by the Taylor expansion that

$$\Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ m^{\frac{1}{2}}[F_{o}(i_jx_j+(1-i_j)(-x_j-)+\ m^{-\frac{1}{2}}t_j\ ,\ 1\le j\le p)$$
$$(3.9)\qquad - F_{o}(i_jx_j+(1-i_j)(-x_j-),\ 1\le j\le p\ )\ ]$$
$$= \Sigma^{p}_{k=1}t_k\ \ \Sigma_{i_1=0,1}\cdots\Sigma_{i_p=0,1}\ (-1)^{p-\Sigma i_j}\ (\partial/\partial y_k)F_{o}(i_jx_j+(1-i_j)(-x_j)+y_j,1\le j\le p)\Big|_{\underset{\sim}{y}=\underset{\sim}{0}}$$
$$+\ o(||\underset{\sim}{t}||).o(1)\ .$$

We like to show that the first term on the right hand side of (3.9) is identically equal to 0. Towards this, for simplicity, we take $p = 2$; a very similar treatment holds for general $p \geq 2$. Note that for every $x_1 \geq 0$ and $x_2 \geq 0$,

$$(\partial/\partial y_1)[\, F_o(x_1+y_1,x_2+y_2) - F_o(x_1+y_1,-x_2+y_2) - F_o(-x_1+y_1,x_2+y_2) + F_o(-x_1+y_1,-x_2+y_2)\,]\Big|_{y=0}$$

$$= \int_{-\infty}^{x_2} f_o(x_1,u)du - \int_{-\infty}^{-x_2} f_o(x_1,u)du - \int_{-\infty}^{x_2} f_o(-x_1,u)du + \int_{-\infty}^{-x_2} f_o(-x_1,u)du$$

$$= \int_{-\infty}^{x_2} f_o(x_1,u)du - \int_{-\infty}^{-x_2} f_o(x_1,u)du - \int_{-x_2}^{\infty} f_o(x_1,u)du + \int_{x_2}^{\infty} f_o(x_1,u)du$$

(3.10) (by the diagonal symmetry of f_o)

$$= \int_{-\infty}^{\infty} f_o(x_1,u)du - \int_{-\infty}^{\infty} f_o(x_1,u)du = 0 .$$

The same identity holds with the partial derivative with respect to y_2 at $\underset{\sim}{y} = \underset{\sim}{0}$. Also, by (3.3), uniformly in $\underset{\sim}{x} \in E_p^*$, the second term on the right hand side of (3.9) is $o_p(1)$. Hence, uniformly in $\underset{\sim}{x} \in E_p^*$, (3.9) is $o_p(1)$. For the treatment of the last term on the right hand side of (3.8), we make use of the weak convergence of the sample distributional process in the multi-dimensional case. For every $m \geq 1$, we define $W_m^o = \{ W_m^o(\underset{\sim}{x}) , \underset{\sim}{x} \in E^p \}$, by letting

$$(3.11) \qquad W_m^o(\underset{\sim}{x}) = m^{\frac{1}{2}} [\, F_m^o(\underset{\sim}{x}) - F_o(\underset{\sim}{x}) \,] , \underset{\sim}{x} \in E^p.$$

Then the weak convergence of the empirical process W_m^o to a tied-down Gaussian function follows by an appeal to the standard results on weak convergence in the multi-parameter case [see for example, Neuhaus (1971)]. If we define the modulus of continuity $\omega_\delta(g)$ as $\sup\{ |g(\underset{\sim}{x}) - g(\underset{\sim}{y})| : |\underset{\sim}{x} - \underset{\sim}{y}| < \delta\}, \delta > 0$, then by the results in Section 5 of Neuhaus(1971), we conclude that

$$(3.12) \quad \lim_{\delta \to 0} \limsup_{m \to \infty} P\{ \omega_\delta(W_m^o) > \varepsilon \} = 0 , \ \forall \varepsilon > 0.$$

[Actually, we may rewrite W_m^o as a process on $D[0,1]^p$ (by using the marginal probability integral transformations) and then verify (3.1_2) by reference to Neuhaus(1971). However, by the uniform continuity of f_o (as has been assumed), we may bypass these details and state the result as in (3.12)]. Since $||\underset{\sim}{t}|| = o_p(1)$ (as assumed in (3.3)), $m^{-\frac{1}{2}}||\underset{\sim}{t}|| \xrightarrow{p} 0$ as $m \to \infty$, so that by (3.1_2), we conclude that the last term on the right hand side of (3.8) being bounded

by $2^p \omega_{\delta_m}(W_m)$, where $\delta_m = m^{-\frac{1}{2}}||\underset{\sim}{t}|| \xrightarrow{p} 0$, converges in probability to 0 (a.e.), as $m \to \infty$. Consequently, we obtain that

$$(3.13) \quad \sup\{ m^{\frac{1}{2}} |\hat{F}^*_m(\underset{\sim}{x}) - F^{o*}_m(\underset{\sim}{x})| : \underset{\sim}{x} \in E^*_p \} \xrightarrow{p} 0, \text{ as } m \to \infty.$$

This prove (3.1). Q.E.D.

Now, by virtue of Theorem 1, we conclude that under the regularity conditions of Section 2, as $N \to \infty$,

$$(3.14) \quad \sup\{ N^{\frac{1}{2}} | \hat{F}^*_m(\underset{\sim}{x}) - F^{o*}_m(\underset{\sim}{x}) - \hat{G}^*_n(\underset{\sim}{x}) + G^{o*}_n(\underset{\sim}{x})| : \underset{\sim}{x} \in E^*_p \} \xrightarrow{p} 0.$$

By (2.8), (2.9), (2.14), (2.15) and (3.14), we conclude that under the assumed regularity conditions, as $N \to \infty$,

$$(3.15) \quad |\hat{K}^{*+}_{mn} - K^{*+}_{mn}| \xrightarrow{p} 0 \quad \text{and} \quad |\hat{K}^*_{mn} - K^*_{mn}| \xrightarrow{p} 0.$$

Note that the stochastic equivalence result in (3.15) is of quite general nature, and it holds irrespective of the null hypothesis H_0 in (1.2) being true or not. For the univariate case, under H_0, K^{*+}_{mn} and K^*_{mn} are genuinely distribution-free and the asymptotic distributions in (1.5) and (1.6) pertain to these statistics. Thus, $\hat{K}^{*+}_{mn}$ and $\hat{K}^*_{mn}$ are asymptotically distrinution-free and their asymptotic distributions are given by (1.5) and (1.6). The situation is different for the multi-variate case. K^{*+}_{mn} and K^*_{mn} are generally not distribution-free under H_0 (unless the p coordinates of $\underset{\sim}{X}^o_i$ (or $\underset{\sim}{Y}^o_i$) are stochatically independent). Keeping the asymptotic equivalence result in (3.15) in mind, we may proceed as follows:

(i) Permutational distributional approach: As the Chatterjee-Sen(1964) rank-permutation principle can be adapted to generate the permutational (conditional) distribution of K^{*+}_{mn} or K^*_{mn} (if the $\underset{\sim}{X}^{o*}_i$ and $\underset{\sim}{Y}^{o*}_i$ were observable), we use (3.15) and consider the following approximation to this permutation distribution. Define the $\hat{\underset{\sim}{X}}^*_i$ and $\hat{\underset{\sim}{Y}}^*_j$ as in (2.5). Pool these N aligned vectors into a combined set and then consider all possible $\binom{N}{m}$ partitioning of these into two subsets of sizes m and n respectively. For each such partitioning, consider the

two-sample Kolmogorov-Smirnov type statistics K_{mn}^{*+} and K_{mn}^{*}, defined as in (2.8)-(2.9). Use the (discrete) distribution over the $\binom{N}{m}$ realizations (with equal probability assigned to each of these mass points), and by the usual ordering of these $\binom{N}{m}$ realizations of K_{mn}^{*+} or K_{mn}^{*}, the cut-off point (critical value) can be obtained as in an usual permutation test. The essential difference here is that the critical value determined from this permutation distribution is not (generally) the exact (conditional) one (i.e., may not correspond to the exact (conditional) significance level). But, these critical values are good approximations to the exact ones. This procedure, though deterministic, may become quite computationally laborious as N increases, and hence, for large sample sizes, one may like to use the following weak convergence approach which is more adaptable in practice.

(ii) Weak convergence approach: As in Section 2, we denote the d.f. of $\underset{\sim}{X}_i^{o*}$ by F_o^*, and let F_{oj}^* be the jth marginal d.f., for j=1,...,p. Let $I_p = [0,1]^p$ be the unit p-cube, and for every $\underset{\sim}{t} = (t_1,\ldots,t_p) \in I_p$ and $\underset{\sim}{s} = (s_1,\ldots,s_p) \in I_p$, let

$$\gamma(\underset{\sim}{s},\underset{\sim}{t}) = F_o^*\big(F_{o1}^{*-1}(s_1 \wedge t_1),\ldots,F_{op}^{*-1}(s_p \wedge t_p)\big) - F_o^*\big(F_{o1}^{*-1}(s_1),\ldots,F_{op}^{*-1}(s_p)\big) F_o^*\big(F_{o1}^{*-1}(t_1),\ldots,F_{op}^{*-1}(t_p)\big). \tag{3.16}$$

Consider then a Gaussian process $W^o = \{ W^o(\underset{\sim}{t}), \underset{\sim}{t} \in I_p \}$ where $EW^o(\underset{\sim}{t}) = 0$ for every $\underset{\sim}{t} \in I_p$, and

$$EW^o(\underset{\sim}{s})W^o(\underset{\sim}{t}) = \gamma(\underset{\sim}{s},\underset{\sim}{t}), \text{ for every } \underset{\sim}{s}, \underset{\sim}{t} \in I_p. \tag{3.17}$$

Let

$$\kappa^+ = \sup_{\underset{\sim}{t} \in I_p} W^o(\underset{\sim}{t}) \quad \text{and} \quad \kappa = \sup_{\underset{\sim}{t} \in I_p} |W^o(\underset{\sim}{t})|. \tag{3.18}$$

Then, by reference to the general weak convergence results for the multi-dimensional empirical processes [viz., Neuhaus(1971)], we conclude that under H_0 in (1.2),

$$K_{mn}^{*+} \xrightarrow{\mathcal{D}} \kappa^+ \quad \text{and} \quad K_{mn}^{*} \xrightarrow{\mathcal{D}} \kappa, \text{ as } N \to \infty. \tag{3.19}$$

Thus, by virtue of (3.15) and (3.19), we may use the distribution of κ^+ or κ to provide asymptotic approximations to the null distribution of K_{mn}^{*+} or K_{mn}^{*}.

In the univariate case, the covariance function $\gamma(s,t)$ reduces to $s \wedge t - st$, so that W^o is a standard Brownian bridge on $[0,1]$, and hence, the distributions of κ^+ and κ are given by (1.5)-(1.6). In the multivariate case, i.e., for $p \geq 2$, the distribution of κ^+ or κ depends on the covariance function $\gamma(\underset{\sim}{s},\underset{\sim}{t})$. For some particular form of this covariance function, Dugue (1969) has studied the characteristic functions of certain functionals of W^o, and his results may be extended to our case as well. However, for our case, the covariance function $\gamma(\underset{\sim}{s},\underset{\sim}{t})$ is not specified (even under H_o), and hence, the Fourier coefficients appearing in the characteristic function are not known. This makes it difficult to use the inversion of the characteristic function to obtain the distribution function of κ^+ or κ in a series form. Actually, using the basic results of Kiefer(1961), it can be shown that for every $p(\geq 1)$ and $\varepsilon > 0$, there exist positive constants $c_1(p,\varepsilon)$ and $c_2(p,\varepsilon)$, such that for every $t \geq 0$,

$$(3.20) \quad P\{ \kappa^+ \geq t \} \leq c_1(p,\varepsilon).\exp\{ -(2-\varepsilon)t^2 \} ,$$

$$(3.21) \quad P\{ \kappa \geq t \} \leq c_2(p,\varepsilon).\exp\{ -(2-\varepsilon)t^2 \} ,$$

and this could have been used to provide suitable bounds for the critical values of κ^+ and κ. However, the constants $c_1(p,\varepsilon)$ and $c_2(p,\varepsilon)$ depend very much on the underlying covariance function $\gamma(\underset{\sim}{s},\underset{\sim}{t})$ (besides depending on p and ε), and hence, even for a given ε, we may not be able to have their values computed when $\gamma(\underset{\sim}{s},\underset{\sim}{t})$ is not known. For this reason, we may take recourse to a modified bootstrap method to estimate the critical levels of κ^+ and κ.

Define the aligned vectors $\hat{X}_i^*$ and $\hat{Y}_j^*$ as in (2.5) and pool all these N vectors into a combined sample. Let $\hat{H}_N^*$ be the empirical d.f. of these N aligned vectors. From the distribution $\hat{H}_N^*$, we draw K independent samples, each of size N. We denote the observations in the rth sample by $\underset{\sim}{Z}_{r1},\ldots,\underset{\sim}{Z}_{rN}$, for $r = 1,\ldots,K$. Note that the $\underset{\sim}{Z}_{rk}$ are drawn with replacement from the finite population of the N units (related to the vectors $\hat{X}_1^*,\ldots,\hat{X}_m^*,\hat{Y}_1^*,\ldots,\hat{Y}_n^*$). For the rth sample, we partition the observations into two subsets $(\underset{\sim}{Z}_{r1},\ldots,\underset{\sim}{Z}_{rm})$

and $(Z_{m+1},\ldots,Z_N)$, and then compute the statistics K^{*+}_{mn} and K^{*}_{mn} as in Section 2, [see (2.14)-(2.15)]; we denote these statistics by T^{+}_{r} and T_r, respectively, for $r = 1,\ldots,K$. Let $T^{+}_{K,1} \leq \ldots \leq T^{+}_{K,K}$ (and $T_{K,1} \leq \ldots \leq T_{K,K}$) be the ordered values of the T^{+}_{r} (and T_r) from the r pseudo-samples, and we choose K (large) in such a way that corresponding to the chosen level of significance α ($0 < \alpha < 1$) , we have

(3.22) $\quad M = K(1 - \alpha)$ a positive integer.

Then, our estimate of the α-level critical values of κ^{+} and κ are

(3.23) $\quad T^{+}_{K,M}$ and $T_{K,M}$, respectively.

Note that the above procedure is conceptually very similar to the permutational approach outlined in (i). The basic difference is here we are sampling with replacement, while in the earlier case, it was without replacement ; but both are based on the aligned observations ($\hat{X}^{*}_{i}$ and $\hat{Y}^{*}_{j}$). While in the earlier case, we do not have to choose K (which is rather arbitrary), we have in all $\binom{N}{m}$ possible partitioning , and as N increases this may become prohibitively large. For example, when $m = n = 25$, $\binom{50}{25}$ is an enormously large number, while for the modified bootstrap procedure, we may choose K as 500 or more to have good estimators of the critical values. Actually, with respect to the permutational distributional approach too, one may choose a subset of K permutations (at random) from the set of $\binom{N}{m}$ possible ones, and based on this subset, one may consider the estimates of the critical values. These will have essentially the same properties , but, sampling without replacement may make the variability even somewhat smaller. On this ground, we would recommend the use of the modified permutational approach where a subset of K (random) permutations are used in the estimation of the critical levels.

We may comment briefly on the consistency of the tests based on the statistics in (2.8)-(2.9). By virtue of (3.15), (3.19) and (3.20)-(3.21), the tests are consistent against the same class of alternatives for which the tests

based on K^{*+}_{mn} and K^{*}_{mn} are consistent. Note that if F^{*}_{o} and G^{*}_{o} stand respectively for the d.f. of $\underset{\sim}{X}^{o*}_{i}$ and $\underset{\sim}{Y}^{o*}_{j}$, i.e.,they are the population counterparts of F^{o*}_{m} and G^{o*}_{n} defined in (2.13), then the test based on K^{*+}_{mn} is consistent against the class of alternatives that $\sup\{ F^{*}_{o}(\underset{\sim}{x}) - G^{*}_{o}(\underset{\sim}{x}) : \underset{\sim}{x} \in E^{*}_{p} \} > 0$, and the test based on K^{*}_{mn} is consistent against a larger class for which $\sup\{ | F^{*}_{o}(\underset{\sim}{x}) - G^{*}_{o}(\underset{\sim}{x}) | : \underset{\sim}{x} \in E^{*}_{p}\} > 0$. Note that ideally we would have expected our tests to be consistent against the class of alternatives for which $\sup\{ F_{o}(\underset{\sim}{x}) - G_{o}(\underset{\sim}{x}) : \underset{\sim}{x} \in E^{p} \} > 0$ and $\sup\{| F_{o}(\underset{\sim}{x}) - G_{o}(\underset{\sim}{x})| : \underset{\sim}{x} \in E^{p} \} > 0$. This is easy to verify that these two sets of classes are not necessarily the same. In the univariate case, $F^{*}_{o}(x) = 2F_{o}(x)-1$, $\forall\ x \geq 0$, so that these two classes are the same. However, in the bivariate or general multivariate case, they need not be the same. As an example, consider the case of $p = 2$. Let $\underset{\sim}{X}^{o}$ have the diagonally symmetric d.f. F_{o} . Also, let $\underset{\sim}{Y}^{o}$ have the diagonally symmetric d.f. G_{o} . We denote $\underset{\sim}{Y}^{o} = (Y^{o}_{1},Y^{o}_{2})$ and assume that $(Y^{o}_{1},-Y^{o}_{2})$ has the same d.f. F_{o} .Then , in general F_{o} and G_{o} are not the same (exception , when Y^{o}_{1} and Y^{o}_{2} are independent). On the other hand, if we define $\underset{\sim}{X}^{o*}$ and $\underset{\sim}{Y}^{o*}$ as in before, then, for them, the corresponding d.f. F^{*}_{o} and G^{*}_{o} will be the same. This apparent difference is mainly due to the fact that in the multivariate case, the symmetry of the marginal distributions fails to ensure the same for the joint distributions, and hence, using particular constructions for these joint distributions, one may obtain the same joint distributions for the absolute values of the coordinates which may not have the unique inversions. In view of this fact, we have to keep in mind that the class of alternatives for which the proposed tests are consistent. The pathological example cited above should not be over-emphasized;in practice , this works out well.

For local alternatives, the weak convergence of the two empirical distribution processes can be employed to obtain the asymptotic distributions of $\hat{K}^{*+}_{mn}$ and $\hat{K}^{*}_{mn}$ in terms of the boundary crossing probabilities of some drifted Gaussian process (in the multi-parameter case), and local asymptotic efficiency results may then be obtained as in Hájek and Šidák (1967, p.272).

4. SOME GENERAL REMARKS

The primary reason for working with the absolute values of the variables in (2.5) is to eliminate the effect of the location estimators (possible under the assumption of diagonal symmetry of the underlying F_o and G_o) .If we would have worked with the residuals in (2.3), even under the assumed diagonal symmetry of F_o and G_o , the equivalence result in (3.1) and (3.2) may not hold when $\hat{F}_m^*$ and $\hat{G}_n^*$ (and F_m^{o*} and G_n^{o*}) are replaced by $\hat{F}_m$ and $\hat{G}_n$ (and F_m and G_n) , respectively. The weak convergence of the empirical distributional processes when some parameters are estimated has been studied by a host of workers [viz., Durbin (1973)], and this may be employed to find out the asymptotic distribution of the Kolmogorov-Smirnov type statistics based on the residuals in (2.3). However, this involves a mutidimensional Gaussian process whose covariance function depends on some other functionals of the density functions f_o and g_o , and thereby, are more difficult to adopt in practice(even when f_o and g_o are diagonally symmetric). The proposed tests are operationally more simple and share good asymptotic properties too.

In the multivariate nonparametric problems, often, dimension-reduction is employed , and on the reduced data, some simpler tests are used. For example, Anderson (1966) considered some 'cutting function' (a scalar valued function of the vector $\underset{\sim}{X}$ or $\underset{\sim}{Y}$), by which one can transform the p-variate data to the univariate case, and then use the classical univariate nonparametric tests to test for the multivariate situation. We may also refer to Friedman and Rafsky (1979) where this technique has been employed at a more sophisticated level. In either case, these authors considered the basic hypothesis of equality of the two multivariate distributions, and the choice of the cutting function has been justified on some natural grounds. In general, apart from the arbitrariness of this choice of cutting functions, there is also a loss of efficiency due to this dimension-reduction. If one is interested in testing for a hypothesis

similar to (1.2), one may use a cutting function, estimate its location in a convenient way, work with the residuals, and under an assumption of symmetry of the distribution of this cutting function (around its median), one can use the aligned test based on $\hat{K}^{*+}_{mn}$ or $\hat{K}^{*}_{mn}$ (defined on the residual cutting functions). For this univariate case, one may even use the asymptotic results in (1.5)-(1.6). However, in general, this will entail some loss of efficiency due to dimension-reduction, and also the class of alternatives for which such a test will be consistent will be a subset of the class considered in Section 3.

In this paper, we have considered an aligned Kolmogorov-Smirnov type test (for both one-sided and two-sided alternatives). It is also possible to consider some related test statistics. For example, we may define the aligned empirical d.f.'s $\hat{F}^{*}_{m}$ and $\hat{G}^{*}_{n}$ as in (2.6), let $\hat{H}^{*}_{N} = N^{-1}\{ m\hat{F}^{*}_{m} + n\hat{G}^{*}_{n} \}$, and consider an aligned Cramér-von Mises' type test statistic

$$(4.1) \qquad \mathcal{L}_N = \frac{mn}{N} \int_{E_p^*} \int [\hat{F}^{*}_{m}(\underset{\sim}{x}) - \hat{G}^{*}_{n}(\underset{\sim}{x})]^2 d\hat{H}^{*}_{N}(\underset{\sim}{x}) .$$

Again, by virtue of Theorem 1, we may establish the asymptotic equivalence of the statistic in (4.1) and the parallel version based on the d.f.'s F^{o*}_{m}, G^{o*}_{n} and H^{o*}_{N} ($= N^{-1}\{ mF^{o*}_{m} + nG^{o*}_{n}\}$). For the distribution theory, we may proceed as in Section 3 . One advantage of the statistic $\mathcal{L}_N$ in (4.1) is that its permutational moments can be computed with relatively more ease, though its asymptotic distribution is no less simpler than that of K^{*+}_{mn} or K^{*}_{mn} . In a different context [viz. Majumdar and Sen(1978)], the study of the local (approximate) Bahadur-efficiency of Kolmogorov-Smirnov and Cramér-von Mises tests reveals that generally the Kolmogorov-Smirnov test may perform better than the Cramér-von Mises test. A very similar conclusion holds here.

REFERENCES

[1] T.W. Anderson, Some nonparametric procedures based on statistically equivalent blocks, Multivariate Analysis (ed.P.R. Krishnaiah), pp.5-27(1966).

[2] P.J. Bickel, A distribution-free version of the Smirnov two-sample test in the p-variate case, Ann. Math. Statist. 40, 1-23 (1969).

[3] S.K. Chatterjee - P.K. Sen, Nonparametric tests for the bivariate two-sample location problem, Calcutta Statist. Assoc. Bull. 13, 18-58 (1964).

[4] D. Dugue, Characteristic functions of random variables connected with Brownian motion and of the von Mises' multidimensional ω_n^2 ,Multivariate Analysis,II (ed. P.R. Krishnaiah), pp.289-301 (1969).

[5] J. Durbin, Some results for the bivariate goodness of fit problem, in Nonparametric Tech. in Statist. Infer. (ed. M.L. Puri), pp. 435-449 (1970)

[6] J. Durbin, Weak convergence of sample distribution function when parameters are estimated, Ann. Statist. 1, 279-290 (1973)

[7] J.H. Friedman - L.C. Rafsky, Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests, Ann. Statist. 7, 697-717 (1979)

[8] J. Hajek - Z. Sidak, Theory of Rank Tests, Academia, Prague (1967).

[9] J.L. Hodges,Jr., The significance probability of the Smirnov two-sample test, Arkiv for Math. 3, 469-486 (1957).

[10] J. Kiefer, On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, Pacific J. Math. 11, 649-660 (1961).

[11] H. Majumdar - P.K. Sen, Nonparametric tests for multiple regression under progressive censoring, J. Multivar. Anal. 8, 73-95 (1978).

[12] G. Neuhaus, On weak convergence of stochastic processes with multidimensional time parameter, Ann. Math. Statist. 42, 1285-1295 (1971).

[13] P.K. Sen, On a Kolmogorov-Smirnov type aligned test , Statist. Probability Lett. 2 , (1984).

[14] I. Vincze, On two-sample tests based on order statistics, Publication. Math. 10, 82-87.