Ann. Inst. Statist. Math.
Vol. 44, No. 4, 755–768 (1992)

A PROCEDURE FOR ASSESSING VECTOR CORRELATIONS

JÉRÔME ALLAIRE AND YVES LEPAGE*

Département de mathématiques et de statistique,
Université de Montréal, Montréal, Québec, Canada H3C 3J7

(Received November 21, 1990; revised November 20, 1991)

Abstract. Three known measures of multivariate relationship are presented. Under the null hypothesis of lack of multivariate relationship between K random vectors, the asymptotic joint distributions of the $\binom{K}{2}$ values taken by these measures for all possible pairs $(X^{(i)}, X^{(j)})$, $1 \leq i < j \leq K$, is used to construct tests of the null hypothesis based on the maximum and more generally, on the greatest values of the measures. The asymptotic power of the tests is also obtained under a sequence of alternatives.

Key words and phrases: Multivariate relationship, matrix correlation, asymptotic distributions, elliptical distributions, hypothesis testing.

1. Introduction

Suppose we have K sets of measurements on n individuals where each set of measurements represents a specific characteristic. For example, we measure a vector of biological variables such as age, height and weight, a vector of biochemical variables like cholesterol, albumin and calcium levels in the blood and a vector of variables representing mental traits of individuals. We are then interested in determining the existence of relations between these vectors of characteristics and if there is one, to find which vectors are significantly related. Similar work has been done by Cameron and Eagleson (1985) when considering only random variables.

To measure the relation between sets of variables we use three known measures of multivariate relationship. The first measure proposed by Stewart and Love (1968) is classified as a redundancy measure while the two others proposed by Escoufier (1973) and Cramer and Nicewander (1979) are known as measures of multivariate association. We can calculate the $\binom{K}{2}$ different values taken by these measures to compare the K vectors two at a time. We will suppose throughout that the parent distribution is in the class of elliptical distributions. This class of

* The author was partially supported by a grant of the Natural Sciences and Engineering Research Council of Canada.

distributions is more general than the multivariate normal model and yet it is quite simple (see, for example, Devlin *et al.* (1976)). Under this class of distributions, the asymptotic (as $n \to \infty$) joint distributions of the $\binom{K}{2}$ measures derived in Allaire and Lepage (1990) are used to construct statistical tests to decide when large values are significant. The procedure is approximate since it is based on asymptotic null distributions but it can give an indication when assessing the significance of large measures of multivariate relationship.

In Section 2, we present the three known measures of multivariate relationship, some properties and their asymptotic joint distributions under the null hypothesis. We propose tests of lack of relationship based on the maximum of the measures in Section 3 and obtain the asymptotic non-null distributions of the tests statistics under a sequence of alternatives in Section 4. In Section 5, we construct tests based on the s largest measures of relationship, $1 \leq s \leq K(K-1)/2$. Finally, an application of the procedure is presented in Section 6.

2. Measures of multivariate relationship

Let

$$X = \begin{pmatrix} X^{(1)} \\ \vdots \\ X^{(K)} \end{pmatrix}$$

be a $p \times 1$ random vector where $X^{(i)}$ is $p_i \times 1$, $i = 1, \ldots, K$ and $\sum_{i=1}^{K} p_i = p$. Define for $i = 1, \ldots, K$, $\mu^{(i)} = E(X^{(i)})$ and for $i, j = 1, \ldots, K$,

$$\Sigma_{ij} = \operatorname{Cov}(X^{(i)}, X^{(j)}) = E\{(X^{(i)} - \mu^{(i)})(X^{(j)} - \mu^{(j)})'\}.$$

Write

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \vdots \\ \mu^{(K)} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1K} \\ \vdots & & \vdots \\ \Sigma_{K1} & \cdots & \Sigma_{KK} \end{pmatrix}$$

and assume Σ positive definite. Consider a random sample $X_1, \ldots, X_n$ $(n > p)$ drawn from X where

$$X_\alpha = \begin{pmatrix} X_\alpha^{(1)} \\ \vdots \\ X_\alpha^{(K)} \end{pmatrix}$$

for $\alpha = 1, \ldots, n$. The usual unbiased estimators of μ and Σ are

$$\bar{X} = \begin{pmatrix} \bar{X}^{(1)} \\ \vdots \\ \bar{X}^{(K)} \end{pmatrix} \quad \text{and} \quad S = \begin{pmatrix} S_{11} & \cdots & S_{1K} \\ \vdots & & \vdots \\ S_{K1} & \cdots & S_{KK} \end{pmatrix}$$

where for $i = 1, \ldots, K$, $\bar{X}^{(i)} = (1/n) \sum_{\alpha=1}^{n} X_\alpha^{(i)}$ and for $i, j = 1, \ldots, K$,

$$S_{ij} = \frac{1}{n-1} \sum_{\alpha=1}^{n} (X_\alpha^{(i)} - \bar{X}^{(i)})(X_\alpha^{(j)} - \bar{X}^{(j)})'.$$

We consider first, the measure of multivariate relationship proposed by Stewart and Love (1968) and defined between $X^{(i)}$ and $X^{(j)}$ for $i, j = 1, \ldots, K$, $i \neq j$, by

$$RV_{ij}^{(1)} = \frac{\operatorname{tr}(S_{ij} S_{jj}^{-1} S_{ji})}{\operatorname{tr}(S_{ii})}$$

where $\operatorname{tr}(\cdot)$ is the trace operator. It is classified as a redundancy measure since it is based on the prediction of $X^{(i)}$ by $X^{(j)}$ (see Cramer and Nicewander (1979)). We consider also two measures of multivariate association. These measures are generalizations of the correlation coefficient. A first measure is proposed by Escoufier (1973) and is defined by

$$RV_{ij}^{(2)} = \frac{\operatorname{tr}(S_{ij} S_{ji})}{\sqrt{\operatorname{tr}(S_{ii}^2)\operatorname{tr}(S_{jj}^2)}};$$

a second measure is presented by Cramer and Nicewander (1979) and is defined by

$$RV_{ij}^{(3)} = \frac{\operatorname{tr}(S_{ii}^{-1} S_{ij} S_{jj}^{-1} S_{ji})}{p_i}$$

for $i, j = 1, \ldots, K$, $i \neq j$. These measures possess the following properties: (a) $0 \leq RV_{ij}^{(h)} \leq 1$ for $h = 1, 2, 3$; (b) $RV_{ij}^{(h)} = RV_{ji}^{(h)}$ for $h = 2$ and for $h = 3$ if we replace p_i by $\min\{p_i, p_j\}$; (c) when $p_i = p_j = 1$, $RV_{ij}^{(h)}$ reduces to the squared correlation coefficient between variables $X^{(i)}$ and $X^{(j)}$, for $h = 1, 2, 3$; (d) when $p_i = 1$, $RV_{ij}^{(h)}$ reduces to the squared multiple correlation coefficient between the variable $X^{(i)}$ and the vector $X^{(j)}$ for $h = 1, 3$; (e) $RV_{ij}^{(1)}$ and $RV_{ij}^{(3)}$ are functions of canonical correlations and (f) $RV_{ij}^{(3)}$ is invariant under nonsingular linear transformations of either $X^{(i)}$ or $X^{(j)}$ while $RV_{ij}^{(1)}$ and $RV_{ij}^{(2)}$ are invariant under orthogonal transformations of either set of variables. For the proof of these properties and other results on measures of multivariate relationship, the reader is referred to Cramer and Nicewander (1979), Ramsay *et al.* (1984) and Lazraq and Cléroux (1988).

Suppose that the distribution of X is in the class of elliptical distributions $E_p(\mu, V)$ with mean vector μ, covariance matrix $\Sigma = \alpha V$ for some α and kurtosis parameter κ (see Muirhead (1982)). This class of distributions generalizes the multivariate normal distribution ($\kappa = 0$) and contains a large number of alternatives to the normal model like the ϵ-contaminated multivariate normal distribution and the multivariate t distribution. It includes long-tailed and short-tailed distributions. We are interested in the null hypothesis

$$H_0 : \Sigma_{ij} = 0, \quad 1 \leq i < j \leq K \tag{2.1}$$

which does not imply mutual independence of the vectors $X^{(i)}$ (except when $\kappa = 0$) but only non-correlation between the components of $X^{(i)}$ and $X^{(j)}$ for all $1 \leq i < j \leq K$. In Allaire and Lepage (1990), it is shown that H_0 is equivalent to the hypothesis of lack of multivariate relationship between $X^{(i)}$ and $X^{(j)}$ that is, for fixed h, $h \in \{1,2,3\}$, $\rho V_{ij}^{(h)} = 0$, $1 \le i < j \le K$ where $\rho V_{ij}^{(h)}$ is the population measure of multivariate relationship defined in the same manner as $RV_{ij}^{(h)}$ replacing S by Σ for $1 \le i < j \le K$ and $h = 1,2,3$. In Allaire and Lepage (1990) it is also shown that for fixed h, $h \in \{1,2,3\}$, the $RV_{ij}^{(h)}$'s, $1 \le i < j \le K$, under H_0, are asymptotically (as $n \to \infty$) independent and distributed as

$$(2.2) \qquad \frac{(1+\kappa)}{\operatorname{tr}(\Sigma_{ii})} \sum_{k=1}^{p_i} \sum_{l=1}^{p_j} \lambda_k^{(i)} W_{ijkl}^2 \quad \text{if} \quad h = 1,$$

$$(2.3) \qquad \frac{(1+\kappa)}{\sqrt{\operatorname{tr}(\Sigma_{ii}^2)\operatorname{tr}(\Sigma_{jj}^2)}} \sum_{k=1}^{p_i} \sum_{l=1}^{p_j} \lambda_k^{(i)} \lambda_l^{(j)} W_{ijkl}^2 \quad \text{if} \quad h = 2,$$

$$(2.4) \qquad \frac{(1+\kappa)}{p_i} \chi_{p_i p_j}^2 \quad \text{if} \quad h = 3$$

where $\lambda_1^{(i)}, \ldots, \lambda_{p_i}^{(i)}$ are the eigenvalues of Σ_{ii}, $i = 1, \ldots, K$, and the W_{ijkl} are independent $N(0,1)$ random variables. Distributions (2.2) and (2.3) were actually derived for the case $K = 2$ in Lazraq and Cléroux (1989) and Cléroux and Ducharme (1989) respectively.

3. Tests of lack of relationship based on the maximum $RV_{ij}^{(h)}$, $1 \le i < j \le K$

When K variables are measured on n individuals and the measurements are assumed to be normally distributed, the likelihood ratio test for the independence of the K variables against the alternative that there is only one non-zero correlation is based on the maximum correlation coefficient (see Moran (1980)). However we often measure on each individual K sets of variables where each set corresponds to a specific characteristic and we are interested in the relationship that may exist between these characteristics.

If it is suspected that there is only one non-zero measure of multivariate relationship at the population level or one non-zero Σ_{ij}, $i < j$, tests of the null hypothesis given by (2.1), based on the maximum of the $RV_{ij}^{(h)}$, $1 \le i < j \le K$, for fixed h, $h \in \{1,2,3\}$, seem adequate in this situation. For normally distributed random vectors, the likelihood ratio test for H_0 against this alternative is based (see Allaire and Lepage (1991)) on the maximum of the $\binom{K}{2}$ possible values of a measure of multivariate relationship called the Hotelling-Rozeboom measure (see Cramer and Nicewander (1979)) defined between $X^{(i)}$ and $X^{(j)}$, $i < j$, by one minus the vector alienation coefficient which is the determinant of the matrix $I_{p_j} - S_{jj}^{-1} S_{ji} S_{ii}^{-1} S_{ij}$. With the results of the preceding section, asymptotic tests can be constructed. The tests consist of rejecting the null hypothesis if

$$n \max_{1 \le i < j \le K} RV_{ij}^{(h)} > c_\alpha^{(h)}$$

where h is fixed, $h \in \{1,2,3\}$ and $c_\alpha^{(h)}$ is the $(1-\alpha)$-th quantile of the asymptotic distribution of $n \max_{1 \le i < j \le K} RV_{ij}^{(h)}$. To obtain the critical values $c_\alpha^{(h)}$, $h = 1,2,3$, set $Y_{ij}^{(h)}$, $1 \le i < j \le K$ the random variables such that under H_0,

$$n \begin{pmatrix} RV_{12}^{(h)} \\ \vdots \\ RV_{K-1\ K}^{(h)} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} Y_{12}^{(h)} \\ \vdots \\ Y_{K-1\ K}^{(h)} \end{pmatrix} \tag{3.1}$$

where $\xrightarrow{\mathcal{L}}$ stands for convergence in distribution as $n \to \infty$. For fixed h, $h \in \{1,2,3\}$, the random variables $Y_{ij}^{(h)}$, $1 \le i < j \le K$, are independent and their distributions are given by (2.2), (2.3) and (2.4). Therefore, we have for fixed h and positive x,

$$\begin{aligned} P_0^{(h)}(x) &= \lim_{n\to\infty} P\left(n \max_{1 \le i < j \le K} RV_{ij}^{(h)} > x\right) \\ &= 1 - \lim_{n\to\infty} P(nRV_{12}^{(h)} \le x, \ldots, nRV_{K-1\ K}^{(h)} \le x) \\ &= 1 - P(Y_{12}^{(h)} \le x, \ldots, Y_{K-1\ K}^{(h)} \le x) \\ &= 1 - \prod_{1 \le i < j \le k}\prod P(Y_{ij}^{(h)} \le x). \end{aligned}$$

Hence, the critical values $c_\alpha^{(h)}$ can be obtained by iteration on x so that $P_0^{(h)}(x) = \alpha$. In practice, it is much simpler to evaluate the critical levels of the tests from

$$P_0^{(h)}\left(n \max_{1 \le i < j \le K} rv_{ij}^{(h)}\right) = 1 - \prod_{1 \le i < j \le k}\prod P\left(Y_{ij}^{(h)} \le n \max_{1 \le i < j \le K} rv_{ij}^{(h)}\right) \tag{3.2}$$

where, for $h = 1,2,3$, $rv_{ij}^{(h)}$ are the observed values of $RV_{ij}^{(h)}$, $1 \le i < j \le K$. For $h = 3$, the chi-square distribution function with $p_i p_j$ degrees of freedom is needed but for $h = 1$ and 2, since $Y_{ij}^{(h)}$ is distributed as a sum of eigenvalues times chi-square random variables, the Imhof algorithm (1961) can be used. The unknown parameters are replaced by consistent estimates: Σ is replaced by S; the eigenvalues of Σ_{ii} are replaced by the eigenvalues of S_{ii} and κ is replaced by a consistent estimate (see, for example, Cléroux and Ducharme (1989)), $\hat{\kappa}$ equal to a third of the average of the sample kurtosis coefficients of the p variables.

4. Asymptotic power of proposed tests

If the null hypothesis is false, there is at least one pair, say (i_0, j_0), for which $\Sigma_{i_0 j_0} \ne 0$ or alternatively $\rho V_{i_0 j_0}^{(h)} > 0$, then we have for $h = 1,2,3$,

$$\begin{aligned} \lim_{n\to\infty} P\left(n \max_{1 \le i < j \le K} RV_{ij}^{(h)} > c_\alpha^{(h)}\right) &= \lim_{n\to\infty} P\left(\bigcup_{1 \le i < j \le K} \{nRV_{ij}^{(h)} > c_\alpha^{(h)}\}\right) \\ &\ge \lim_{n\to\infty} P(nRV_{i_0 j_0}^{(h)} > c_\alpha^{(h)}) = 1 \end{aligned}$$

since for $h = 1, 2, 3$, $RV_{i_0 j_0}^{(h)}$ is a consistent estimate of $\rho V_{i_0 j_0}^{(h)} > 0$. Therefore, the tests proposed in the preceding section are consistent. In the following, we consider the sequence of alternative hypotheses

$$H_{1n} : \begin{cases} \Sigma_{ij} = 0, & (i, j) \neq (i_0, j_0) \\ \Sigma_{j_0 i_0} = \Sigma'_{i_0 j_0} = \dfrac{A}{\sqrt{n}} & \end{cases}$$

where $i_0 < j_0$ and A is a fixed $p_{j_0} \times p_{i_0}$ matrix. We derive, for fixed h, $h \in \{1, 2, 3\}$, the asymptotic (as $n \to \infty$) joint distribution of $nRV_{ij}^{(h)}$, $1 \leq i < j \leq K$, under H_{1n}. First, we prove the following lemma.

LEMMA 4.1. *Let S be the sample covariance matrix obtained from a sample of size n drawn from an elliptical distribution with covariance matrix Σ and kurtosis parameter κ. Then under H_{1n}, we have*

$$Z_n = \sqrt{n} \begin{pmatrix} \mathrm{vec}(S_{21}) \\ \vdots \\ \mathrm{vec}(S_{K\ K-1}) \end{pmatrix} \xrightarrow{\mathcal{L}} Z$$

where Z is distributed as $N_f(\mu_Z, \Sigma_Z)$, $f = \sum\sum_{1 \leq i < j \leq K} p_i p_j$.

$$\mu_Z = \begin{pmatrix} 0 \\ \vdots \\ \mathrm{vec}(A) \\ \vdots \\ 0 \end{pmatrix},$$

$$\Sigma_Z = (1 + \kappa) \begin{pmatrix} \Sigma_{11} \otimes \Sigma_{22} & & 0 \\ & \ddots & \\ 0 & & \Sigma_{K-1\ K-1} \otimes \Sigma_{KK} \end{pmatrix}.$$

$\mathrm{vec}(S_{ji})$ is the $p_j p_i \times 1$ vector formed by stacking the columns of the $p_j \times p_i$ matrix S_{ji} for $1 \leq i < j \leq K$ and $\otimes$ denotes the Kronecker product of matrices.

PROOF. We have to show that for any $f \times 1$ vector

$$\lambda = \begin{pmatrix} \lambda_{12} \\ \vdots \\ \lambda_{K-1\ K} \end{pmatrix} \in \mathbb{R}^f,$$

$\lambda' Z_n \xrightarrow{\mathcal{L}} \lambda' Z$. We can write (see Allaire and Lepage (1990))

$$\sqrt{n} \left(\begin{pmatrix} \mathrm{vec}(S_{21}) \\ \vdots \\ \mathrm{vec}(S_{K\ K-1}) \end{pmatrix} - \begin{pmatrix} \mathrm{vec}(\Sigma_{21}) \\ \vdots \\ \mathrm{vec}(\Sigma_{K\ K-1}) \end{pmatrix} \right) \xrightarrow{\mathcal{L}} N_f(0, \Omega)$$

where the $f \times f$ matrix Ω may be partitioned into $p_ip_j \times p_kp_l$ matrices

$$\Omega_{ji}^{lk} = (1+\kappa)((\Sigma_{ik} \otimes \Sigma_{jl}) + K_{p_ip_j}(\Sigma_{jk} \otimes \Sigma_{il})) + \kappa \operatorname{vec}(\Sigma_{ji})(\operatorname{vec}(\Sigma_{lk}))'$$

for $1 \leq i < j \leq K$ and $1 \leq k < l \leq K$, $K_{p_ip_j}$ is a $p_ip_j \times p_ip_j$ commutation matrix defined by

$$K_{p_ip_j} = \sum_{k=1}^{p_i}\sum_{l=1}^{p_j}(\Delta_{kl} \otimes \Delta'_{kl})$$

and Δ_{kl} is a $p_i \times p_j$ matrix with all its elements being zero except the entry (k,l) which is unity. Since Z_n is asymptotically multivariate normal, $\lambda' Z_n$ is asymptotically normal, hence

$$\frac{\lambda' Z_n - \sqrt{n}\sum\sum_{1\leq i<j\leq K}\lambda'_{ij}\operatorname{vec}(\Sigma_{ji})}{\left(\sum\sum_{1\leq i<j\leq K}\sum\sum_{1\leq k<l\leq K}\lambda'_{ij}\Omega_{ji}^{lk}\lambda_{kl}\right)^{1/2}} \xrightarrow{\mathcal{L}} N(0,1).$$

Replacing $\Sigma_{j_0i_0}$ by $A/\sqrt{n}$ and using the fact that $\Sigma_{ji} = 0$ for $(i,j) \neq (i_0, j_0)$, we get

$$\sqrt{n}\sum\sum_{1\leq i<j\leq K}\lambda'_{ij}\operatorname{vec}(\Sigma_{ji}) = \lambda'_{i_0j_0}\operatorname{vec}(A) = \lambda'\mu_Z$$

and

$$\begin{aligned}
&\sum\sum_{1\leq i<j\leq K}\sum\sum_{1\leq k<l\leq K}\lambda'_{ij}\Omega_{ji}^{lk}\lambda_{kl} \\
&= \sum\sum_{1\leq i<j\leq K}\lambda'_{ij}\Omega_{ji}^{ji}\lambda_{ij} + \sum\sum_{1\leq i<j\leq K}\sum\sum_{\substack{1\leq k<l\leq K \\ (i,j)\neq(k,l)}}\lambda'_{ij}\Omega_{ji}^{lk}\lambda_{kl} \\
&= \sum\sum_{\substack{1\leq i<j\leq K \\ (i,j)\neq(i_0,j_0)}}\lambda'_{ij}\Omega_{ji}^{ji}\lambda_{ij} + \lambda'_{i_0j_0}\Omega_{j_0i_0}^{j_0i_0}\lambda_{i_0j_0} + O(n^{-1/2}) \\
&= \sum\sum_{1\leq i<j\leq K}\lambda'_{ij}(1+\kappa)(\ \Sigma_{ii} \otimes \Sigma_{jj})\lambda_{ij} + O(n^{-1/2}) \\
&= \lambda'\Sigma_Z\lambda + O(n^{-1/2})
\end{aligned}$$

since

$$\Omega_{j_0i_0}^{j_0i_0} = (1+\kappa)(\Sigma_{i_0i_0} \otimes \Sigma_{j_0j_0}) + O(n^{-1})$$

under H_{1n}. The result is obtained from Serfling (1980). □

It follows from this lemma that under H_{1n}, the $RV_{ij}^{(h)}$'s, $1 \leq i < j \leq K$, remain asymptotically independent for fixed h, $h \in \{1,2,3\}$, but a non-centrality parameter is introduced in the asymptotic distribution of $nRV_{i_0j_0}^{(h)}$. The proof is identical as in Allaire and Lepage (1990).

THEOREM 4.1. *Under the assumptions of Lemma* 4.1 *and under* H_{1n}, *we have for fixed* h, $h \in \{1,2,3\}$,

$$n\begin{pmatrix} RV_{12}^{(h)} \\ \vdots \\ RV_{K-1\ K}^{(h)} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} Y_{12}^{(h)} \\ \vdots \\ Y_{K-1\ K}^{(h)} \end{pmatrix}$$

where for $h = 1,2$ *the* $Y_{ij}^{(h)}$*'s*, $1 \leq i < j \leq K$, *are independent and distributed as in* (2.2) *and* (2.3) *except that for* $Y_{i_0j_0}^{(h)}$ *the random variables* $W_{i_0j_0kl}$ *are replaced by* $W_{i_0j_0kl}^{(h)}$ *which are independent and distributed as* $N(\delta_{kl}^{(h)},1)$ *for* $k = 1,\ldots,p_{i_0}$, $l = 1,\ldots,p_{j_0}$, *where*

$$\delta_{kl}^{(h)} = \begin{cases} C_{kl}^{(1)'}\dfrac{(\Sigma_{i_0i_0}^{-1/2}\otimes I_{p_{j_0}})}{\sqrt{1+\kappa}}\operatorname{vec}(A) & \text{if } h=1, \\ C_{kl}^{(2)'}\dfrac{(\Sigma_{i_0i_0}^{-1/2}\otimes \Sigma_{j_0j_0}^{-1/2})}{\sqrt{1+\kappa}}\operatorname{vec}(A) & \text{if } h=2, \end{cases}$$

$C_{kl}^{(h)}$ *is the normalized eigenvector corresponding to the eigenvalue*

$$\begin{cases} \lambda_k^{(i_0)} \text{ of multiplicity } p_{j_0} \text{ of } (\Sigma_{i_0i_0}\otimes I_{p_{j_0}}) & \text{if } h=1, \\ \lambda_k^{(i_0)}\lambda_l^{(j_0)} \text{ of } (\Sigma_{i_0i_0}\otimes\Sigma_{j_0j_0}) & \text{if } h=2 \end{cases}$$

and $Y_{i_0j_0}^{(3)}$ *is distributed as* $((1+\kappa)/p_{i_0})\chi^2_{p_{i_0}p_{j_0}}(\delta^2)$ *where*

$$\delta^2 = \frac{(\operatorname{vec}(A))'(\Sigma_{i_0i_0}\otimes\Sigma_{j_0j_0})^{-1}\operatorname{vec}(A)}{1+\kappa}.$$

As a consequence, we can calculate the asymptotic power of the tests of Section 3. It is given, for $h = 1,2,3$, by

$$\lim_{n\to\infty} P_{H_{1n}}\left(n \max_{1\leq i<j\leq K} RV_{ij}^{(h)} > c_\alpha^{(h)}\right) = 1 - \prod\prod_{1\leq i<j\leq K} P(Y_{ij}^{(h)} \leq c_\alpha^{(h)})$$

where the distribution of $Y_{ij}^{(h)}$, $1 \leq i < j \leq K$, is given above.

More generally, when $\Sigma'_{ij} = A_{ij}/\sqrt{n}$ for the s pairs (i,j), $1 \leq s \leq K(K-1)/2$, belonging to the set $I = \{(i_1,j_1),\ldots,(i_s,j_s)\}$ and $\Sigma_{ij} = 0$ for $(i,j) \notin I$, $1 \leq i < j \leq K$, we can prove that the random variables $Y_{ij}^{(h)}$ of Theorem 4.1 are independent and distributed as in (2.2), (2.3) and (2.4) for $(i,j) \notin I$ and distributed as $Y_{i_0j_0}^{(h)}$ for $(i_0,j_0) \in I$ replacing (i_0,j_0) by (i_r,j_r) and A by $A_{j_ri_r}$ for $r = 1,\ldots,s$.

5. Generalization to the s largest $RV_{ij}^{(h)}$, $1 \leq s \leq K(K-1)/2$

When it is suspected that there is not only one, but a few non-zero measures of multivariate relationship between the $X^{(i)}$'s, we can then utilize also the information contained in the second largest, third largest, ... measures of multivariate relationship, not only the maximum. Let $P_{ij}^{(h)} = 1 - F_{ij}^{(h)}(nRV_{ij}^{(h)})$ for $1 \leq i < j \leq K$ and fixed h, $h \in \{1, 2, 3\}$ where $F_{ij}^{(h)}$ are the distribution functions of the random variables $Y_{ij}^{(h)}$ given by (3.1) when $\Sigma_{ij} = 0$. Since the $Y_{ij}^{(h)}$'s are continuous, under H_0 the $P_{ij}^{(h)}$'s are asymptotically independent and uniformly distributed over the unit interval $(0, 1)$. Keeping in mind that, under H_0, the $RV_{ij}^{(h)}$'s are not asymptotically identically distributed, we shall base our tests on the quantities $P_{ij}^{(h)}$ which represent the individual asymptotic critical levels or p-values of the observed $RV_{ij}^{(h)}$'s. Therefore, instead of using the largest $RV_{ij}^{(h)}$'s we will consider the smallest $P_{ij}^{(h)}$'s. First, we prove the following lemma.

LEMMA 5.1. *Let $T = (T_1 \cdots T_r)'$ be a vector statistic based on a random sample of size n drawn from a continuous population and $U = (U_1 \cdots U_r)'$ a random vector with continuous distribution function such that $P(U_i \neq U_j) = 1$ for all $i \neq j$. If*

$$T = \begin{pmatrix} T_1 \\ \vdots \\ T_r \end{pmatrix} \xrightarrow{\mathcal{L}} U = \begin{pmatrix} U_1 \\ \vdots \\ U_r \end{pmatrix},$$

then for $1 \leq s \leq r$, we have

$$\begin{pmatrix} T_{(1)} \\ \vdots \\ T_{(s)} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} U_{(1)} \\ \vdots \\ U_{(s)} \end{pmatrix}$$

where $T_{(i)}$ and $U_{(i)}$ are the i-th order statistics of the T_i's and U_i's respectively.

PROOF. Using the notation of Lindgren (1976), let $\nu = (\nu_1, \ldots, \nu_s)$ be a permutation of the integers $1, \ldots, r$ taken s at a time. Each of these $r!/(r-s)!$ permutations defines a region in $\mathbb{R}^r$ given by

$$R_\nu = \{x = (x_1 \cdots x_r) \mid x_{\nu_1} < \cdots < x_{\nu_s} < x_{\nu_{s+1}}, \ldots, x_{\nu_r}\}$$

where $\nu_1, \ldots, \nu_s, \nu_{s+1}, \ldots, \nu_r$ is a permutation of the integers $1, \ldots, r$. Note that each R_ν is the union over the $(r-s)!$ possible permutations of $\nu_{s+1}, \ldots, \nu_r$ of the mutually exclusive sets $\{x \mid x_{\nu_1} < \cdots < x_{\nu_r}\}$. The set of regions $\{R_\nu\}_\nu$ constitutes the sample space except for the boundaries which have probability zero under the assumption of continuity. Let $R \subseteq \mathbb{R}^s$ be the projection of R_ν on $\mathbb{R}^s$ corresponding to the specific permutation $\nu = (1, \ldots, s)$. Let A be any subset of R and A_ν be the set of points in R_ν whose s smallest coordinates when ordered yield a point in A. Thus we have

$$\begin{aligned}
\lim_{n\to\infty} P((T_{(1)}\cdots T_{(s)}) \in A) &= \lim_{n\to\infty} P\left((T_1\cdots T_r) \in \bigcup_\nu A_\nu\right) \\
&= \lim_{n\to\infty} \sum_\nu P((T_1\cdots T_r) \in A_\nu) \\
&= \sum_\nu \lim_{n\to\infty} P((T_1\cdots T_r) \in A_\nu) \\
&= \sum_\nu P((U_1\cdots U_r) \in A_\nu) \\
&= P\left((U_1\cdots U_r) \in \bigcup_\nu A_\nu\right) \\
&= P((U_{(1)}\cdots U_{(s)}) \in A). \qquad \square
\end{aligned}$$

From this lemma, we deduce, for fixed h, $h \in \{1,2,3\}$ and for $1 \le s \le K(K-1)/2$, that

$$\begin{pmatrix} P_{(1)}^{(h)} \\ \vdots \\ P_{(s)}^{(h)} \end{pmatrix} \xrightarrow{\mathcal{L}} \begin{pmatrix} U_{(1)} \\ \vdots \\ U_{(s)} \end{pmatrix}$$

where $P_{(i)}^{(h)}$, for $i = 1,\ldots,s$ is the i-th order statistic of the $P_{ij}^{(h)}$'s $1 \le i < j \le K$ and $U_{(i)}$ is the i-th order statistic of a random sample of size $K(K-1)/2$ uniformly distributed over the unit interval. The joint distribution of $U_{(1)} \le \cdots \le U_{(s)}$ is given by the density function (see David (1981))

$$f(u_1,\ldots,u_s) = \begin{cases} \dfrac{r!(1-u_s)^{r-s}}{(r-s)!} & \text{if } 0 < u_1 < \cdots < u_s < 1, \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

with $r = K(K-1)/2$. We propose to reject the null hypothesis if at least one of the ordered critical levels $P_{(i)}^{(h)}$ is small. If the desired asymptotic significance level of the test is α then s critical values $c_1,\ldots,c_s$ are required such that

$$P\left(\bigcup_{i=1}^{s}\{U_{(i)} \le c_i\}\right) = \alpha. \tag{5.2}$$

The choice of the c_i's can be done in many ways. Since in general we have no information on the type of alternative to consider and we wish to give equal weight to each $P_{(i)}^{(h)}$, we suggest choosing the c_i's such that the individual probabilities $P(U_{(i)} \le c_i)$ be equal for $i = 1,\ldots,s$. We can compute these probabilities by using the fact that $U_{(i)}$ has a beta$(i, r-i+1)$ distribution. The c_i's are obtained by iteration so that equation (5.2) is satisfied where the left-hand side is computed using the density function (5.1) which gives

$$1-P\left(\bigcap_{i=1}^{s}\{U_{(i)}>c_i\}\right)=1-\int_{c_s}^{1}\int_{c_{s-1}}^{u_s}\cdots\int_{c_1}^{u_2}f(u_1,\ldots,u_s)du_1\cdots du_s.$$

The critical values c_i, $i=1,\ldots,s$, decrease as s increases so it is possible that $P_{(1)}^{(h)}$ considered by itself is significant but considered together with $P_{(2)}^{(h)}$ is not significant. More generally, the $s-1$ largest measures of multivariate relationship can yield significance while the s largest may not. One should then look more closely at the $s-1$ pairs $(X^{(i)},X^{(j)})$ corresponding to the significant $P_{ij}^{(h)}$'s.

6. An application

Consider the national track records for men of 55 countries given in Dawkins (1989). This data set is part of a larger set collected by Belcham and Hymans (1984) for the 1984 Los Angeles olympic games. There are 8 variables corresponding to 8 different track events ranging from the 100 meters race to the marathon. The variables are grouped into three subsets of variables representing three types of races. First, there are three sprints: the 100, 200 and 400 meters events; then there are two middle-distance races: the 800 and 1500 meters events; finally there are three long-distance races: the 5 and 10 kilometers events and the marathon. The sizes of the subvectors are respectively $p_1=3$, $p_2=2$ and $p_3=3$.

It is of course suspected that there exist relations between the three types of races. However, we do not know which type of races are significantly related when simultaneously comparing them two at a time. It is known that performance in running reflects the energetic and mechanic capacities of the athletes (see, for example, Péronnet and Thibault (1989)) and that these capacities have different relative importances which are specific for different race distances (Svendenhag and Sjödin (1984)). For example, the maximal aerobic power (MAP) is very important for middle-distance races while the percentage of MAP sustained and the anaerobic capacity are the primary physiological factors of marathon runners and sprinters respectively. However, since the statistical unit used is a country, we will find out whether the nations that produce world class runners succeed in doing so for all types of races or only for certain pairs of race types. This would reflect for example, the genetic background of the runners or the training programs held in each country.

As noted by Dawkins (1989), if the raw data were analysed in the same units, too much weight would be put on the long-distance races, particularly on the marathon. Therefore, each variable is rescaled to give mean 0 and standard deviation 1 so that the transformed variables represent the relative performances of a country in different events.

A goodness of fit test for normality is performed on each variable. The Kolmogorov statistic modified by Stephens (1974) is used. For every event, the normality assumption is rejected at the level 0.05 except for the 200 meters race.

Also a consistent estimate of κ, the kurtosis parameter which is zero for a multivariate normal population, is obtained through the use of an algorithm found in Cléroux and Ducharme (1989). The data set yields $\hat{\kappa} = 1.201$ thus the underlying distribution can not be assumed to be multivariate normal.

Table 1 gives the results of the tests of Section 3. The critical levels of the tests for $H_0 : \rho V_{ij} = 0$, $1 \leq i < j \leq 3$. based on the maximum $RV_{ij}^{(h)}$ are computed from formula (3.2) with Imhof's algorithm (1961) using the subroutine FQUAD found in Chapter 9 of Koerts and Abrahamse (1969). The computations are performed on a CYBER 170 series, model 835/855 computer using the FORTRAN 5 programming language. The significant results given in Table 1, agree with the one obtained when we apply the asymptotic test proposed by Muirhead and Waternaux (1980) which yields a critical level smaller than 0.01.

Table 1. Tests statistics and critical levels for $H_0 : \Sigma_{ij} = 0$, $1 \leq i < j \leq 3$.

Measure of multivariate relationship	$\max_{1\leq i<j\leq 3} rv_{ij}$	Pair (i,j) corresponding to $\max_{1\leq i<j\leq 3} rv_{ij}$	Critical level
Stewart & Love	0.8235	$(2,3)$	0.00016
Escoufier	0.8323	$(2,3)$	0.00002
Cramer & Nicewander	0.4422	$(2,3)$	0.00245

Although we strongly reject the null hypothesis, so that there are significantly large measures of multivariate relationship in the data, we do not want to stop here. We are interested in assessing which measures can be thought to be significantly large. Table 2 gives the ordered critical levels for each pair $(X^{(i)}, X^{(j)})$, $1 \leq i < j \leq 3$, for the individual tests of $\rho V_{ij}^{(h)} = 0$, based on the three measures of multivariate relationship. It gives also the estimate $\hat{\kappa}$ computed from the components of $X^{(i)}$ and $X^{(j)}$. It is interesting to note that for all three measures, the relationship is the highest between middle-distance and long-distance races while it is the lowest between long-distance races and sprints.

To use the tests of the preceding section with all $s = 3$ ordered critical levels, we need to determine the critical values c_1, c_2 and c_3. As discussed in Section 5, this can be done so that equal weights are given to the $P_{(i)}^{(h)}$'s. $i = 1,2,3$. We find at the significance level 0.01, $c_1 = 0.0012$, $c_2 = 0.0350$ and $c_3 = 0.1531$. Not only the null hypothesis is rejected at the level 0.01 since at least one ordered $P_{ij}^{(h)}$ is smaller than it's corresponding critical value, but we see that all three critical levels are smaller than their corresponding critical values. Therefore, we can believe all three measures of multivariate relationship to be significantly large at the 0.01 significance level.

Table 2. Ordered critical levels for the individual tests of $H_0 : \rho V_{ij}^{(h)} = 0$, $h = 1, 2, 3$.

Measure of multivariate relationship	Ordered critical levels	Corresponding measure rv_{ij}	Pair (i, j)	$\hat{\kappa}$
Stewart & Love	0.000016	0.8235	(2,3)	0.8836
	0.000376	0.6652	(1,2)	1.5062
	0.002369	0.5413	(1,3)	1.2112
Escoufier	0.000001	0.8323	(2,3)	0.8836
	0.000080	0.7136	(1,2)	1.5062
	0.000368	0.5140	(1,3)	1.2112
Cramer & Nicewander	0.000240	0.4422	(2,3)	0.8836
	0.006215	0.4103	(1,2)	1.5062
	0.030323	0.2472	(1,3)	1.2112

Acknowledgements

The authors are indebted to Professor Luc Léger of the Department of Physical Education of the Université de Montréal for helpful discussions concerning the application. The authors are also grateful to the referees for suggestions which led to the improvement of the paper.

References

Allaire, J. and Lepage, Y. (1990). Tests de l'absence de liaison entre plusieurs vecteurs aléatoires pour les distributions elliptiques, *Statist. Anal. Données*, **15**, 21–46.

Allaire, J. and Lepage, Y. (1991). On a likelihood ratio test for independence, *Statist. Probab. Lett.*, **11**, 449–452.

Belcham, P. and Hymans, R. (1984). *IAAF/ATFS Track and Field Statistics Handbook for the 1984 Los Angeles Olympic Games*, International Amateur Athletic Federation, London.

Cameron, M. A. and Eagleson, G. K. (1985). A new procedure for assessing large sets of correlations, *Austral. J. Statist.*, **27**, 84–95.

Cléroux, R. and Ducharme, G. R. (1989). Vector correlation for elliptical distributions, *Comm. Statist. Theory Methods*, **18**, 1441–1454.

Cramer, E. M. and Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association, *Psychometrika*, **49**, 403–423.

David, H. A. (1981). *Order Statistics*, 2nd ed., Wiley, New York.

Dawkins, B. (1989). Multivariate analysis of national track records, *Amer. Statist.*, **43**, 110–115.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R. (1976). Some multivariate applications of elliptical distribution, *Essays in Probability and Statistics* (ed. S. Ikeda), 365–395, Shinko Tsusho, Tokyo.

Escoufier, Y. (1973). Le traitement des variables vectorielles, *Biometrics*, **29**, 751–760.

Imhof, P. (1961). Computing the distribution of quadratic forms in normal variates, *Biometrika*, **48**, 419–426.

Koerts, J. and Abrahamse, A. P. J. (1969). *On the Theory and Application of the General Linear Model*, Rotterdam University Press, Rotterdam.

Lazraq, A. and Cléroux, R. (1988). Étude comparative de différentes mesures de liaison entre deux vecteurs aléatoires et tests d'indépendance, *Statist. Anal. Données*, **13**, 15–18.

Lazraq, A. and Cléroux, R. (1989). Tests of homogeneity between coefficient of multivariate association in elliptical distributions, Thèse de Doctorat, Faculté des Études Supérieures, Université de Montréal.

Lindgren, B. W. (1976). *Statistical Theory*, 3rd ed., Macmillan, New York.

Moran, P. A. P. (1980). Testing the largest of a set of correlation coefficients, *Austral. J. Statist.*, **22**, 289–297.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, New York.

Muirhead, R. J. and Waternaux, C. M. (1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for non-normal populations, *Biometrika*, **67**, 31–43.

Péronnet, F. and Thibault, G. (1989). Mathematical analysis of running performance and world running records, *Journal of Applied Physiology*, **67**, 453–465.

Ramsay, J. O., ten Berge, J. M. F. and Styan, G. P. H. (1984). Matrix correlation, *Psychometrika*, **49**, 403–423.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

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

Stewart, D. and Love, W. (1968). A general canonical correlation index, *Psychological Bulletin*, **70**, 160–163.

Svendenhag, J. and Sjödin, B. (1984). Maximal and submaximal oxygen uptakes and blood lactate levels in elite male middle- and long-distance runners, *International Journal of Sports Medicine*, **5**, 225–261.