The Annals of Statistics
1987, Vol. 15, No. 3, 1215–1228

ON MEASURING INTERNAL DEPENDENCE IN A SET OF RANDOM VARIABLES[1]

BY ROBERT A. KOYAK

The Johns Hopkins University

To measure dependence in a set of random variables, a multivariate analog of maximal correlation is considered. This consists of transforming each of the variables so that the largest partial sums of the eigenvalues of the resulting correlation matrix is maximized. A "maximalized" measure of association obtained in this manner permits statements to be made about the strength of internal dependence exhibited by the random variables. It is shown, under a weak regularity condition, that optimizing transformations exist and that they satisfy a geometrically interpretable fixed point property. If the variables are jointly Gaussian, then the identity transformation is shown to be optimal, which extends Kolmogorov's result for canonical correlation to the principal components setting.

1. Introduction. Consider two random vectors $\mathbf{Y} = (Y_1, \ldots, Y_p)'$ and $\mathbf{Z} = (Z_1, \ldots, Z_q)'$, each defined relative to a common probability space, with ranges $R(Y) \subset \mathbf{R}^p$ and $R(Z) \subset \mathbf{R}^q$. An often used measure of association between $\mathbf{Y}$ and $\mathbf{Z}$ is the *canonical correlation coefficient*

$$\rho(\mathbf{Y}, \mathbf{Z}) = \sup_{a, b} \operatorname{corr}(a'\mathbf{Y}, b'\mathbf{Z}), \tag{1.1}$$

where the supremum is over $a \in \mathbf{R}^p$ and $b \in \mathbf{R}^q$. As a measure of association, the usefulness of $\rho(\mathbf{Y}, \mathbf{Z})$ depends on the extent to which the relationship between $\mathbf{Y}$ and $\mathbf{Z}$ is linear, as it is if $\mathbf{Y}$ and $\mathbf{Z}$ are jointly Gaussian. This restriction can be overcome by considering instead the *maximal* canonical correlation coefficient

$$\rho^*(\mathbf{Y}, \mathbf{Z}) = \sup_{\theta, \phi} \operatorname{corr}(\theta(\mathbf{Y}), \phi(\mathbf{Z})), \tag{1.2}$$

where the supremum is taken over all Borel measurable mappings θ: $R(\mathbf{Y}) \to \mathbf{R}^1$ and ϕ: $R(\mathbf{Z}) \to \mathbf{R}^1$ having finite second moments.

In the special case where $\mathbf{Y}$ and $\mathbf{Z}$ are univariate, $\rho^*(\mathbf{Y}, \mathbf{Z})$ is simply the *maximal correlation coefficient*, which has received considerable attention in the statistical literature. Gebelein (1941) is credited with first formulating the concept of maximal correlation and for giving it its present name. The mathematical properties of maximal correlation have been explored by several authors, among whom the contributions of Sarmanov (1958a, b), Rényi (1959) and Csáki and Fischer (1963) are prominent. We note the following properties of the maximal

Received March 1986; revised October 1986.

[1]Research supported by Office of Naval Research Contract No. N00014-79-C-0801, Air Force Office of Scientific Research Grant No. 82-0029C and National Science Foundation Grant MCS 80-02698.

AMS 1980 *subject classification*. 62H20.

Key words and phrases. Maximal correlation, dimensionality reduction, transformation.

correlation coefficient [see, for example, Csáki and Fischer (1963)]:

PROPERTY 1. (a) $0 \leq \rho^*(\mathbf{Y}, \mathbf{Z}) \leq 1$.

(b) $\rho^*(\mathbf{Y}, \mathbf{Z}) = 0$ if and only if $\mathbf{Y}$ and $\mathbf{Z}$ are stochastically independent.

(c) $\rho^*(\mathbf{Y}, \mathbf{Z}) = 1$ if there exist functions θ^* and ϕ^* such that $\theta^*(\mathbf{Y}) = \phi^*(\mathbf{Z})$, a.e.

(d) If $\mathbf{Y}$ and $\mathbf{Z}$ are jointly Gaussian, then $\rho^*(\mathbf{Y}, \mathbf{Z}) = \rho(\mathbf{Y}, \mathbf{Z})$, i.e., the maximal correlation and the usual product correlation coincide.

It is interesting to note that a converse to Property 1(c), in general, does not hold: The maximal correlation need not be attained by any transformations of $\mathbf{Y}$ and $\mathbf{Z}$. An example in which this phenomenon is exhibited can be found in Rényi (1959). However, pathologies of this type can be ruled out if the bivariate distribution satisfies the mild regularity condition

$$\iint \Omega^2(y, z)\, dF_y\, dF_z < \infty, \tag{1.3}$$

where $\Omega(y, z) = dF_{yz}(y, z)/dF_y(y)\, dF_z(y)$ is the Radon–Nikodym derivative of the joint distribution of Y and Z with respect to the product distribution of its marginals.

Lancaster (1958, 1969) used the term ϕ^2-finite to describe bivariate distributions satisfying (1.3). Since these distributions have $\Omega(y, z) \in L_2(F_y \times F_z)$, one can expand $\Omega(y, z)$ as a Fourier series with respect to an orthonormal basis in the product distribution

$$\Omega(y, z) = 1 + \sum_{n=1}^{\infty} \rho_n \theta_n(y) \phi_n(z), \tag{1.4}$$

where $\sum_{n=1}^{\infty} \rho_n^2 < \infty$. Buja (1985) pointed out that (1.4) can be interpreted as a singular value decomposition with this choice of basis, which in turn gives the canonical form of Lancaster. The constant term on the right in (1.4) is the largest, but trivial, singular value corresponding to constant singular vectors. The remaining singular vectors have zero means and unit variances; furthermore, the largest of the remaining singular values coincides with $\rho^*(\mathbf{Y}, \mathbf{Z})$, and the corresponding singular vectors are the optimizing transformations $\theta^*(\mathbf{Y})$ and $\phi^*(\mathbf{Z})$.

It is readily seen that Properties 1(a)–1(c) also hold for general random vectors $\mathbf{Y}$ and $\mathbf{Z}$ in the canonical correlation context. However, the extension of Property 1(d) is not obvious and is the subject of what is known as Kolmogorov's canonical correlation problem. The solution to this problem, due to Kolmogorov, states that the linear functions of $\mathbf{Y}$ and $\mathbf{Z}$ obtained in a classical canonical correlation analysis also achieve the maximal canonical correlation. A proof of this can be found in Lancaster (1969).

Canonical correlation measures the cohesion between two sets of random variables and is suitable as a measure of association in a multivariate context when a partitioning of the variables is appropriate to the analysis. However, a view in which the variables are considered symmetrically is needed if an

objective of the analysis is to obtain information of a more general nature about the multivariate distribution. For an $\mathbf{R}^M$-valued random vector $\mathbf{X} = (X_1, \ldots, X_M)'$, the smallest-dimensional subsurface of $\mathbf{R}^M$ that does an "adequate job" of containing its variability gives a heuristic measure of the internal association exhibited by $\mathbf{X}$. If strong linear dependence is exhibited, then observations on $\mathbf{X}$ will tend to concentrate in a lower-dimensional linear manifold of $\mathbf{R}^M$. A measure of this concentration is given by the terms $g_k(\mathbf{X})$, which denote the sums of the k largest eigenvalues of the correlation matrix of $\mathbf{X}$, for $1 \leq k \leq M - 1$. We note some of its properties:

PROPERTY 2. (a) $k \leq g_k(\mathbf{X}) \leq M$ for all $1 \leq k \leq M - 1$.

(b) If $g_{k_1}(\mathbf{X}) = k_1$ for some k_1, then necessarily $g_k(\mathbf{X}) = k$ for all $1 \leq k \leq M - 1$, implying that the coordinate variables of $\mathbf{X}$ are uncorrelated. If, in addition, $\mathbf{X}$ is Gaussian, then the coordinate variables of $\mathbf{X}$ are mutually independent.

(c) If $g_k(\mathbf{X}) = M$ for some k, then $g_{k_1}(\mathbf{X}) = M$ for all $k_1 > k$, $1 \leq k_1, k \leq M - 1$.

The inequality on the left in Property 2(a) simply states that the eigenvalues of a correlation matrix *majorize* the diagonal elements. This is obtained as a special case of a theorem due to Schur [see Marshall and Olkin (1979)]. From this majorization property, Property 2(b), also follows. Property 2(c) refers to the condition where $M - k$ of the eigenvalues of the correlation matrix of $\mathbf{X}$ equal 0, or equivalently, where $R(\mathbf{X})$ is contained in a k-dimensional affine space of $\mathbf{R}^M$. For, if $\Gamma(M, k)$ denotes the space of $M \times M$ projection matrices having rank no greater than k, we have, letting $\tilde{\mathbf{X}}$ represent standardized $\mathbf{X}$,

$$g_k(\mathbf{X}) = M - \inf_{A \in \Gamma(M,k)} \sum_{j=1}^{M} E\left[\tilde{X}_j - A_j'\tilde{X}\right]^2, \tag{1.5}$$

with A_j denoting the jth column of A. Hence, $g_k(\mathbf{X}) = M$ implies $\tilde{\mathbf{X}} = A\tilde{\mathbf{X}}$ a.e. for some matrix $A \in \Gamma(M, k)$.

The multivariate data analytic method associated with g_k is principal components. The matrix A achieving the infimum in (1.5) is the eigenprojection of the correlation matrix of $\mathbf{X}$ corresponding to the k largest eigenvalues. If $z_1, \ldots, z_k$ are the normalized eigenvectors corresponding to the k largest eigenvalues (taken in decreasing order), then

$$A_j'\tilde{\mathbf{X}} = \sum_{i=1}^{k} z_{ji} z_i'\tilde{\mathbf{X}} = \sum_{i=1}^{k} z_{ji} F_i \tag{1.6}$$

and F_i is called the ith principal component. The principal components are uncorrelated random variables (independent if $\mathbf{X}$ is Gaussian), with variances equal to the corresponding eigenvalues of the correlation matrix. In geometric terms, the first k principal components provide an orthogonal coordinate system for the k-dimensional linear manifold which best fits the data.

The effectiveness of a principal component analysis can be severely limited in the presence of nonlinear relationships, as is true of any of the classical

multivariate techniques derived from normal theory. Hence, $g_k(\mathbf{X})$ is inadequate as a general measure of the internal dependencies exhibited in $\mathbf{X}$. In analogy with maximal correlation, we derive a "maximalized" version of $g_k(\mathbf{X})$,

$$g_k^*(\mathbf{X}) = \sup_{\phi} g_k(\phi(\mathbf{X})), \tag{1.7}$$

where the supremum is taken over all Borel measurable transformations ϕ: $R(\mathbf{X}) \to \mathbf{R}^M$ taking the form $\phi(\mathbf{X}) = (\phi_1(X_1), \ldots, \phi_M(X_M))'$ having finite second moments. We will assume, without loss of generality, that $E[\phi_j(X_j)] = 0$ and $E[\phi_j^2(X_j)] = 1$ for every j. The relationship between $g_k^*(\mathbf{X})$ and $g_k(\mathbf{X})$ parallels that between the correlation and maximal correlation coefficients:

PROPERTY 3. (a) $k \le g_k^*(\mathbf{X}) \le M$ for all $1 \le k \le M - 1$.

(b) If $g_{k_1}^*(\mathbf{X}) = k_1$ for some k_1, then $g_k^*(\mathbf{X}) = k$ for all $1 \le k \le M - 1$, and the coordinate variables of $\mathbf{X}$ are pairwise independent.

(c) If $g_k^*(\mathbf{X}) = M$ for some k, then $g_{k_1}^* = M$ for all $k_1 > k$, $1 \le k_1, k \le M - 1$. If, in addition, there exists a transformation ϕ^* such that $g_k^*(\mathbf{X}) = g_k(\phi^*(\mathbf{X})) = M$, then there exists a matrix $A \in \Gamma(M, k)$ such that $\phi^*(\mathbf{X}) = A\phi^*(\mathbf{X})$ a.e.

For the independence assertion in Property 3(b), suppose X_1 and X_2, say, are dependent. Then, by Property 1(b), $\rho^*(X_1, X_2) > 0$ and $g_1^*(\mathbf{X})$ is therefore greater than 1. It is also clear from this argument that if the X_j are pairwise, but not *mutually* independent, $g_k^* = k$ would still obtain. In Property 3(c), we again note that the existence of an "optimal transformation" ϕ^* is not automatically ensured. However, we show in Section 2 that optimal transformations exist if each of the bivariate marginal distributions of $\mathbf{X}$ satisfies (1.3). In analogy with Kolmogorov's result for canonical correlation, it is of interest to ask whether $g_k^*(\mathbf{X}) = g_k(\mathbf{X})$ for all $1 \le k \le M - 1$ if $\mathbf{X}$ is Gaussian. Since Property 1(d) ensures that nonlinear transformations of the X_j cannot increase the individual correlations, it seems both plausible and intuitive that Kolmogorov's result extends to principal components. We give a proof of this in Section 3.

2. Mathematical framework. We begin by considering the space of all Borel-measurable function of the form $\phi(\mathbf{X}) = (\phi_1(\mathbf{X}), \ldots, \phi_M(\mathbf{X}))'$, which satisfy

$$\begin{aligned} &E\left[\phi_j(\mathbf{X})\right] = 0, \\ &E\left[\phi_j^2(\mathbf{X})\right] < \infty, \qquad j = 1, \ldots, M. \end{aligned} \tag{2.1}$$

The inner product of any two members ϕ and ψ of this space is defined as

$$\langle \phi, \psi \rangle_M = \sum_{j=1}^{M} \langle \phi_j, \psi_j \rangle = \sum_{j=1}^{M} E\left[\phi_j(\mathbf{X})\psi_j(\mathbf{X})\right], \tag{2.2}$$

and the pseudonorm derived from (2.2) is denoted by

$$\|\phi\|_M = \langle \phi, \phi \rangle^{1/2}. \tag{2.3}$$

This function space, which we will denote $\mathbf{H}^M$, can be viewed as the M-fold Cartesian product of Hilbert space $\mathbf{H}$ with itself, where $\mathbf{H}$ is the space of all

scalar-valued functions of $\mathbf{X}$ having zero mean and finite variance. Denote by $\tilde{\mathbf{H}}^M$ the subspace of $\mathbf{H}^M$ for which an element takes the form

$$\phi(\mathbf{X}) = (\phi_1(X_1), \ldots, \phi_M(X_M))', \tag{2.4}$$

which, in turn, can be expressed as the Cartesian product of Hilbert spaces $\mathbf{H}_1, \ldots, \mathbf{H}_M$, where $\mathbf{H}_j$ is the subspace of $\mathbf{H}$ consisting of scalar-valued functions depending on $\mathbf{X}$ only through X_j. We restrict our attention to transformations satisfying

$$\|\phi_j\| = 1, \qquad j = 1, \ldots, M, \tag{2.5}$$

which entails no loss in generality due to the scale invariance property of the correlation matrix. The set of transformations satisfying (2.5) will be denoted Φ, which is a proper subset of the $\sqrt{M}$-sphere in $\tilde{\mathbf{H}}^M$.

A transformation $\phi^* \in \Phi$ shall be called *optimal for a k-factor multivariate dimensionality reduction analysis* (MDRA) if $g_k(\phi^*(\mathbf{X})) = g_k^*(\mathbf{X})$. An optimal transformation ϕ^* will typically depend on k, although this dependence has been notationally suppressed. We will, subsequently, also drop the notational dependence on $\mathbf{X}$ and simply denote the corresponding quantities by $g_k(\phi)$ and g_k^*, respectively. Our, perhaps cumbersome, choice of nomenclature for this problem is motivated by our desire to avoid confusion with a different nonlinear generalization of principal components, which is discussed in Koyak (1985).

The k-factor MDRA problem can be given a distance interpretation in $\mathbf{H}^M$. Let $\mathbf{S}_k$ denote the set of of elements of $\mathbf{H}^M$, which take the form $A\phi$, where A is an element of $\Gamma(M, k)$ and ϕ is an element of $\tilde{\mathbf{H}}^M$. In analogy with (1.5), we write

$$\begin{aligned} g_k(\phi) &= \sup_{A \in \Gamma(M,k)} E[\operatorname{trace} A\phi\phi'] \\ &= M - \inf_{A \in \Gamma(M,k)} \|\phi - A\phi\|_m^2, \end{aligned} \tag{2.6}$$

so that

$$M - g_k(\phi) = \inf_{\zeta \in \mathbf{S}_k} \|\phi - \zeta\|_M^2. \tag{2.7}$$

An optimal transformation ϕ^* therefore attains the minimum distance between an element in Φ and its closest point in $\mathbf{S}_k$. Deriving an optimal transformation (even under complete stochastic knowledge) is a substantial problem, as neither $\mathbf{S}_k$ nor Φ are linear or even convex subsets of $\tilde{\mathbf{H}}^M$. An algorithm for constructing optimal transformations and a discussion of its mathematical properties can be found in Koyak (1985). At a point in Φ for which the distance between the two sets is minimized, projecting into $\mathbf{S}_k$ and then projecting back into Φ reproduces the point of minimization. This is equivalently expressed as a fixed point property:

$$T_k\phi^* = \phi^*, \tag{2.8}$$

where T_k denotes the compound of the two nonlinear projection operators.

3. Existence of optimal transformations. Let g_k^* denote the supremum of $g_k(\phi)$ over Φ and let $\phi^{(n)}$ denote a sequence in Φ for which $g_k(\phi^{(n)}) \to g_k^*$ as $n \to \infty$. Because Φ is not strongly compact in $\tilde{\mathbf{H}}^M$, it does not directly follow that there exists an element of Φ for which this supremum is attained. For a bivariate random variable, a one-factor MDRA problem is equivalent to finding a transformation to achieve the maximal correlation, so the force of Rényi's example mentioned earlier is applicable to the present framework.

A mild condition on the distribution of $\mathbf{X}$ is required to ensure the existence of optimal transformations. Let P_j denote the conditional expectation operator mapping $\mathbf{H}$ into $\mathbf{H}_j$,

$$P_j\psi = E\big[\psi(\mathbf{X})|X_j\big], \tag{3.1}$$

and let $P_j|\mathbf{H}_i$ denote the restriction of P_j to $\mathbf{H}_i$.

ASSUMPTION 3.1. The restricted conditional expectation operators $P_j|\mathbf{H}_i$ are compact for every pair (i, j) with $i \neq j$.

PROPOSITION 3.1. *A sufficient condition for Assumption* 3.1 *to hold is that each of the bivariate marginal distributions of* $\mathbf{X}$ *satisfies* (1.4).

PROOF. This is a result of long standing in the theory of integral operators and can be found, for example, in Edwards (1965). The canonical form (1.4), however, permits a straightforward proof, which we now proceed to give.

Let i and j be any two positive integers less than or equal to M, where $i \neq j$. In accordance with (1.4), write

$$\Omega(x_i, x_j) = 1 + \sum_{N=1}^{\infty} \rho_n \phi_{j,n}(x_j)\phi_{i,n}(x_i), \tag{3.2}$$

where the ρ_n are square summable and $\{\phi_{r,n}\}$ is an orthonormal basis for $\mathbf{H}_r$, $r = i, j$. Let $\{\psi_j^{(m)}\}$ denote a weakly convergent sequence in $\mathbf{H}_j$ and let ψ_j denote the weak limit. The proposition is proved by showing that $P_i\psi_j^{(m)} \to_s P_i\psi_j$ as $m \to \infty$.

From (3.2) we have

$$P_i\psi_j^{(m)} = \sum_{n=1}^{\infty} \rho_n \langle \psi_j^{(m)}, \phi_{j,n}\rangle \phi_{i,n} \tag{3.3}$$

and therefore

$$\big\| P_i\big(\psi_j^{(m)} - \psi_j\big)\big\|^2 = \sum_{n=1}^{\infty} \rho_n^2 \langle \psi_j^{(m)} - \psi_j, \phi_{j,n}\rangle^2. \tag{3.4}$$

Since every weakly convergent sequence is bounded in norm, the dominated convergence theorem can be applied to take the limit as $m \to \infty$ inside the summation in (3.4) and doing so gives a limiting value of 0 for the term on the left. Therefore, $P_i\psi_j^{(m)} \to_s P_i\psi_j$ as claimed. □

PROPOSITION 3.2. *Under Assumption* 3.1, *for every integer* k *between* 1 *and* $M - 1$, *there exists an element* $\phi^* \in \Phi$, *depending on* k, *such that* $g_k(\phi^*) = g_k^*$.

PROOF. Let $\{\phi^{(n)}\}$ denote a sequence of Φ-transformations for which $g_k(\phi^{(n)}) \to g_k^*$ and let $\{\phi^{(n')}\}$ denote a subsequence converging weakly to an element $\tilde{\phi}$. The weak convergence of each of the coordinate functions $\phi_j^{(n')}$ to $\tilde{\phi}_j$ in $\mathbf{H}_j$ is thereby implied and $\|\tilde{\phi}_j\| \leq 1$ for all j (examples abound where these inequalities are strict). Assume initially that none of these norms is equal to zero. Let $d_j = \|\tilde{\phi}_j\|^{-1}$ and let $D = \operatorname{diag}(d_1, \ldots, d_M)$. Take $\phi^* = D\tilde{\phi}$, which is an element of Φ. If it is shown that

$$g_k(\phi^*) \geq \lim_{n' \to \infty} g_k\big(\phi^{(n')}\big) = g_k^*, \tag{3.5}$$

it will directly follow that ϕ^* is an optimal transformation.

Let $U(\phi)$ have elements $u_{ij}(\phi)$ equal to $\langle \phi_i, \phi_j \rangle$. For $\phi \in \Phi$, we have $u_{jj}(\phi) = 1$ for all j. For $i \neq j$ we have

$$\begin{aligned} u_{ij}\big(\phi^{(n')}\big) &= \langle \phi_i^{(n')}, \phi_j^{(n')} \rangle \\ &= \Big\langle P_j\big(\phi_i^{(n')} - \tilde{\phi}_i\big), \phi_j^{(n')} \Big\rangle + \langle \tilde{\phi}_i, \phi_j^{(n')} \rangle. \end{aligned} \tag{3.6}$$

The first inner product on the second line of (3.6) goes to zero as $n' \to \infty$, since

$$\Big|\Big\langle P_j\big(\phi_i^{(n')} - \tilde{\phi}_i\big), \phi_j^{(n')} \Big\rangle\Big| \leq \Big\| P_j\big(\phi_i^{(n')} - \tilde{\phi}_i\big) \Big\|,$$

which by Assumption 3.1 goes to zero as $n' \to \infty$. Using the weak convergence of $\{\phi_j^{(n')}\}$ on the second inner product on the second line of (3.6), it is seen that $\lim_{n' \to \infty} u_{ij}(\phi^{(n')}) = u_{ij}(\tilde{\phi})$ for all pairs $i \neq j$.

For any symmetric $M \times M$ matrix U, let $G_k(U)$ denote the sum of the k largest eigenvalues. Then $G_k(\hat{U}) = g_k^*$, where $\hat{U} = \lim_{n' \to \infty} U(\phi^{(n')})$. The covariance matrix of ϕ^* satisfies

$$\begin{aligned} \mathbf{U}(\phi^*) &= D\mathbf{U}\big(\tilde{\phi}\big)D', \\ \mathbf{U}(\phi^*) - I &= D(\hat{\mathbf{U}} - I)D'. \end{aligned} \tag{3.7}$$

It will be shown that $G_k(D(\hat{\mathbf{U}} - I)D') \geq G_k(\hat{\mathbf{U}} - I)$. Since every $M \times M$ rotation matrix diagonalizes the identity matrix, it will follow from (3.7) that ϕ^* is optimal, since

$$\begin{aligned} g_k(\phi^*) &= G_k\big(D(\hat{\mathbf{U}} - I)D'\big) + k, \\ g_k^* &= G_k(\hat{\mathbf{U}} - I) + k. \end{aligned} \tag{3.8}$$

The idea, then, is that if the convergence of $\{\phi^{(n')}\}$ is not strong, by concentrating on the off-diagonal elements of the covariance matrix, one can find an element of Φ which is optimal. Of course, if the convergence is strong, the limit is optimal a fortiori.

To facilitate this argument, the following lemma is introduced:

LEMMA 3.1. *Let A be an $M \times M$ symmetric matrix, and W an $M \times M$ symmetric matrix with each diagonal element equal to zero. Let D denote a diagonal matrix with each diagonal element $d_{jj} \geq 1$. There exists a diagonal matric Δ with each diagonal element δ_{jj} equal to $+1$ or -1 such that*

$$\operatorname{trace} AW \leq \operatorname{trace} A\Delta DWD'\Delta'.$$

PROOF. Let D_j and Δ_j denote diagonal matrices where

$$(D_j)_{ii} = \begin{cases} 1, & i \neq j, \\ d_{jj}, & i = j, \end{cases}$$

$$(\Delta_j)_{ii} = \begin{cases} 1, & i \neq j, \\ \delta_{jj}, & i = j. \end{cases} \tag{3.9}$$

Obtain a recursive system of matrices as follows:

$$\begin{aligned} W_0 &= W, \\ W_r &= \Delta_r D_r W_{r-1} D_r' \Delta_r' \\ &= \left[\prod_{j=1}^{r} \Delta_j D_j\right] W \left[\prod_{j=1}^{r} \Delta_j D_j\right]', \end{aligned} \tag{3.10}$$

for $r = 1, \ldots, M$. At the end of this recursion, $W_M = \Delta DWD'\Delta'$. The lemma will be proved by establishing the chain of inequalities

$$\operatorname{trace} AW_{r-1} \leq \operatorname{trace} AW_r, \tag{3.11}$$

through a choice of Δ.

Start with $r = 1$. Using $w_{11} = 0$, one can write

$$\begin{aligned} \operatorname{trace} AW_1 &= \sum_{i=1}^{M} \sum_{j=1}^{M} a_{ij} w_{ij}^{(1)} \\ &= \sum_{i=2}^{M} \sum_{j=2}^{M} a_{ij} w_{ij}^{(0)} + 2\delta_{11} d_{11} \sum_{i=1}^{M} a_{i1} w_{i1}^{(0)} \\ &\geq \sum_{i=2}^{M} \sum_{j=2}^{M} a_{ij} w_{ij}^{(0)} + 2 \sum_{i=1}^{M} a_{i1} w_{i1}^{(0)} \\ &= \operatorname{trace} AW_0, \end{aligned} \tag{3.12}$$

by choosing

$$\delta_{11} = \operatorname{sign}\left[\sum_{i=1}^{M} a_{i1} w_{i1}^{(0)}\right].$$

We proceed recursively in this manner. At the rth stage, use $w_{rr} = 0$ and take

$$\delta_{rr} = \operatorname{sign}\left[\sum_{i=1}^{M} a_{ir} w_{ir}^{(r-1)}\right],$$

to obtain $\operatorname{trace} AW_r \geq \operatorname{trace} AW_{r-1}$ by the same argument used in (3.12), thus finishing the proof of the lemma. □

We now complete our proof of Proposition 3.2. Let $A \in \Gamma(M, k)$ denote the eigenprojection matrix corresponding to the k largest eigenvalues of $\hat{U}$ (and hence also of $\hat{U} - I$). For any matrix Δ satisfying the conditions of Lemma 3.1, we have

$$\begin{aligned} g_k(\phi^*) &= g_k(\Delta\phi^*) \\ &= \sup_{B \in \Gamma(M,k)} \operatorname{trace} B\Delta D(\hat{U} - I)D'\Delta' + k \\ &\geq \operatorname{trace} A\Delta D(\hat{U} - I)D'\Delta' + k. \end{aligned} \tag{3.13}$$

However, by Lemma 3.1, Δ can be chosen so that

$$\operatorname{trace} A\Delta D(\hat{U} - I)D'\Delta' + k \geq \operatorname{trace} A(\hat{U} - I) + k = g_k^*. \tag{3.14}$$

From (3.13) and (3.14) together, we get $g_k(\phi^*) \geq g_k^*$, implying $g_k(\phi^*) = g_k^*$ since g_k^* is the supremum of $g_k(\phi)$ over $\phi \in \Phi$, which completes the proof in the case that none of the $\|\tilde{\phi}_j\|$ is equal to zero.

Now, if any of the terms $\|\tilde{\phi}_j\|$ should equal zero, the limiting correlation matrix $\hat{\mathbf{U}}$ will contain zeros on the off-diagonal for every such term. Replacing $\tilde{\phi}_j$ by any ϕ_j in $\mathbf{H}_j$ having norm 1 will produce a correlation matrix whose sum of the k largest eigenvalues is no smaller than that of $\hat{\mathbf{U}}$. Lemma 3.1 can be employed to produce an element ϕ^* in Φ attaining the supremum of $g_k(\phi)$, finishing the proof. □

4. Measuring association: The Gaussian case. Now suppose that $\mathbf{X}$ is Gaussian, with each X_j marginally $N(0,1)$. A pairwise stochastic representation of $\mathbf{X}$ is given by

$$X_i = \rho_{ij}X_j + \left(1 - \rho_{ij}^2\right)^{1/2}\varepsilon_{ij} \quad \text{a.e.}, \tag{4.1}$$

where ε_{ij} is distributed as $N(0,1)$ independent of X_j and ρ_{ij} is the correlation coefficient for X_i and X_j. For a positive integer r, let h_r^* denote the rth standardized Hermite–Chebyshev polynomial under $N(0,1)$. Note that $\{h_r^*(X_j)\}$ can be taken as an orthonormal system over $\mathbf{H}_j$ for each j, and an arbitrary polynomial transformation of order L in $\mathbf{H}_j$ can be expressed as

$$\phi_j(X_j) = \sum_{n=1}^{L} \tau_{jn} h_n^*(X_j), \tag{4.2}$$

for some constants τ_{jn}. We require

$$\sum_{n=1}^{L} \tau_{jn}^2 = 1, \tag{4.3}$$

so that $\|\phi_j\| = 1$. Let $\mathbf{C}_{L,j}$ denote the subset of $\mathbf{H}_j$ consisting of polynomials having order no greater than L and let $\mathbf{C}_L$ denote their M-fold Cartesian product over j. Using the fact that $\langle h_r^*(X_i), h_s^*(X_j)\rangle = \delta_{rs}\rho_{ij}^r$ [Lancaster (1969)], a typical correlation for an element $\phi \in \mathbf{C}_L \cap \Phi$ takes the form

$$\langle \phi_i, \phi_j \rangle = \sum_{n=1}^{L} \tau_{in}\tau_{jn}\rho_{ij}^n. \tag{4.4}$$

Take $\mathbf{C} = \bigcup_{L=1}^{\infty} \mathbf{C}_L$. It is well known that $\mathbf{C}$ is a complete class in $\tilde{\mathbf{H}}^M$.

PROPOSITION 4.1. *If $\phi^* \in \Phi$ satisfies $g_k(\phi^*) \geq g_k(\phi)$ for every $\phi \in \mathbf{C}_L \cap \Phi$, for every positive integer L, then ϕ^* is an optimal transformation for a k-factor MDRA problem.*

PROOF. This is a straightforward consequence of the completeness of $\mathbf{C}$ and the continuity of $g_k(\phi)$ with respect to the norm topology. □

Let ξ denote the identity map: $\xi(\mathbf{X}) = \mathbf{X}$. Observe that ξ_j coincides with h_1^* for every j, from which it is clear that ξ is an element of $\mathbf{C}_L \cap \Phi$ for every positive integer $\mathbf{L}$. The assertion of optimality of ξ will be proven if it is shown that ξ is optimal among all $\phi \in \mathbf{C}_L \cap \Phi$, regardless of L and k.

In order to make use of the representation afforded by (4.4) we introduce

DEFINITION 4.1. Let A and B denote matrices having common dimension. Their Schur (or Hadamard) product, denoted $A \circ B$, is the matrix of the same dimension composed of the elementwise products of A and B:

$$[A \circ B]_{ij} = a_{ij} b_{ij}.$$

An exposition on the range of uses of Schur products in multivariate statistics can be found in Styan (1973).

Let $\mathbf{U}_0 = \mathbf{U}(\xi)$ denote the $M \times M$ correlation matrix of the identity, with (i, j)th element ρ_{ij}. For a transformation ϕ satisfying (4.2) and (4.3), let $T_n = \operatorname{diag}(\tau_{1n}, \ldots, \tau_{Mn})$ for $n = 1, \ldots, L$, and let the nth Schur power of $\mathbf{U}_0$ be denoted $\mathbf{U}_0^{(n)}$, with (i, j)th element ρ_{ij}^n. The covariance matrix of ϕ then has the representation

$$\mathbf{U}(\phi) = \sum_{n=1}^{L} T_n' \mathbf{U}_0^{(n)} T_n. \tag{4.5}$$

Take $\mathbf{U}_0^{(0)}$ to be the $M \times M$ matrix of 1's, and reexpress (4.5) as

$$\begin{aligned} \mathbf{U}(\phi) &= \left[\sum_{n=1}^{L} T_n' \mathbf{U}_0^{(n-1)} T_n \right] \circ \mathbf{U}_0 \\ &= D(\phi) \circ \mathbf{U}(\xi). \end{aligned} \tag{4.6}$$

Each of the matrices $\mathbf{U}_0^{(n-1)}$ is positive semidefinite: $\mathbf{U}_0^{(0)}$ has M as its only nonzero eigenvalue and for $n \geq 1$, $\mathbf{U}_0^{(n)}$ is the correlation matrix obtained by taking h_n^* as the transformation on each coordinate. Since each diagonal element of $D(\phi)$ is equal to one, $D(\phi)$ is itself a (possibly degenerate) correlation matrix. Therefore, (4.6) suggests a Schur factorization of the covariance matrix of an element of $\mathbf{C}_L \cap \Phi$ into two correlation matrices, one of which corresponds to the identity transformation.

For the case $k = 1$, (4.6) is sufficient to establish the optimality of ξ. Let $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_M$ be the eigenvalues of $\mathbf{U}_0$, with $z_1, \ldots, z_m$ the associated eigenvectors. Collect these eigenvectors into a matrix Z, where the jth column of Z is z_j, and let $Z_j = \operatorname{diag}(z_j)$. Then

$$D(\phi) \circ \mathbf{U}(\xi) = \sum_{j=1}^{M} \gamma_j Z_j' D(\phi) Z_j. \tag{4.7}$$

If w is the eivenvector of $\mathbf{U}(\phi)$ corresponding to its largest eigenvalue, then

$$
\begin{aligned}
g_1(\phi) &= w'[D(\phi)\circ\mathbf{U}(\xi)]w = \sum_{j=1}^{M}\gamma_j w'Z_j'D(\phi)Z_j w \\
&\le \gamma_1 w'\left[\sum_{j=1}^{M}Z_j'D(\phi)Z_j\right]w = \gamma_1 w'[I\circ D(\phi)]w \\
&= \gamma_1 = g_1(\xi).
\end{aligned} \tag{4.8}
$$

This argument does not depend on L and it follows from Proposition 4.1 that ξ is therefore an optimal transformation in Φ.

The derivation of (4.8) can be found in Gekeler (1981); it was originally established by Schur (1911). However, it does not extend in a simple way for k greater than one. We extend this result to the multifactors case with the introduction of an additional concept:

DEFINITION 4.2 [Marshall and Olkin (1979)]. Let $\mathbf{x}$ and $\mathbf{y}$ denote M-vectors, with $x_{[i]}$ and $y_{[i]}$ representing the ith largest of their respective elements. We say that $\mathbf{x}$ is *majorized* by $\mathbf{y}$, denoted $\mathbf{x} \prec \mathbf{y}$, if

$$\sum_{i=1}^{m} x_{[i]} \le \sum_{i=1}^{m} y_{[i]},$$

for every $m = 1,\ldots, M-1$, with equality holding for $m = M$.

For an arbitrary symmetric $M \times M$ matrix U, let $\lambda(U)$ be the eigenvalues of U written as a vector, with elements ordered from largest to smallest, and let $\Lambda(U) = \operatorname{diag}\lambda(U)$. The following theorem extends the ideas expressed in (4.8).

LEMMA 4.1 [Bapat and Sunder (1985)]. *Let A and B be $M \times M$ matrices, with A self-adjoint and B a correlation matrix. Then*

$$\lambda(A\circ B) \prec \lambda(A). \tag{4.9}$$

The reader is referred to the paper by Bapat and Sunder for the proof.

THEOREM 4.1. *Let $\mathbf{X}$ be distributed as Gaussian with mean zero and marginal variances each equal to 1. Then, for every $k = 1,\ldots, M-1$, the identity transformation $\xi \in \Phi$ is optimal for the k-factor MDRA problem.*

PROOF. Fix integer $L > 0$ and consider the problem restricted to $\mathbf{C}_L \cap \Phi$. An element $\phi \in \mathbf{C}_L \cap \Phi$ takes the form suggested by (4.2) with covariance matrix $\mathbf{U}(\phi)$ having the Schur factorization (4.6). As noted earlier, $D(\phi)$ is itself a correlation matrix, and since $\mathbf{U}(\xi)$ is a correlation matrix, it is self-adjoint. It therefore follows from Lemma 4.1 that

$$\lambda(\mathbf{U}(\phi)) \prec \lambda(\mathbf{U}(\xi)). \tag{4.10}$$

Majorization, however, implies

$$g_k(\phi) \le g_k(\xi), \qquad k = 1,\ldots, M-1. \tag{4.11}$$

Since equality is obtained by choosing $\phi = \xi$, ξ is an optimal transformation in

$\mathbf{C}_L \cap \Phi$. Since this is true regardless of the value of L, Proposition 4.1 implies that ξ is optimal in Φ and the proof is complete. □

It is interesting to consider whether the force of this result applies to other spherically symmetric distributions having finite second moments. This can be answered in the negative by considering a r.v. $\mathbf{X}$ having a spherically symmetric distribution with the identity as its correlation matrix. Since the coordinate variables of $\mathbf{X}$ are independent in this case if and only if $\mathbf{X}$ is Gaussian [Kelker (1970)], Property 3(b) asserts that the identity is not optimal. I am grateful to a referee from a different journal for bringing this to my attention.

5. Discussion. Maximalized measures of association are informative in two ways:

(i) They quantify association/dependence under minimal assumptions on the variables.

(ii) The corresponding optimal transformations prescribe actions that can be taken to enhance a linear analysis.

By rephrasing the problem in a function space context, linear data analysis techniques such as canonical correlation and principal components can be given nonlinear generalizations related to deriving a maximalized measure of association. These techniques can be made operative on data by establishing a fixed point property such as (2.8), which forms the rationale for an iterative method of constructing both the maximalized measure of association and its corresponding optimal transformations. This idea has been successfully implemented by Breiman and Friedman (1985) in the context of least-squares multiple regression, Breiman and Ihaka (1984) in discriminant analysis, Owen (1983) in autoregressive time series and Koyak (1985) in principal components.

We emphasize that taking this approach to data analysis is not a new idea: It is a research area of long standing, especially in the psychometric community. The technique known as optimal scoring, often attributed to Fisher (1940) but with antecedents due to Hirschfeld (1935) and others, is an algorithm for finding the maximal correlation of two categorical variables. The ideas of optimal scoring form the basis for the alternating least-squares techniques of de Leeuw, Young and Takane (1976) and the "French School" method of correspondence analysis of Benzécri (1969) and his colleagues. The nonlinear principal components framework developed by de Leeuw (1982) is a finite-dimensional predecessor to the work represented in this paper.

Our treatment of $g_k(\phi)$ as the natural multivariate analog of maximal correlation may seem unduly restrictive in that only marginal transformations are considered. In the framework we have adopted, it is not clear how general joint transformations can be accommodated. A different framework might allow general nonlinear transformations in lieu of the linear functions comprising the principal components, taking the least-squares approach embodied in (1.5) to derive a measure of association. This is similar to the nonlinear factor analysis of McDonald (1967) and Hastie (1984) has developed an interesting methodology in this spirit. A drawback to using nonlinear functions of more than one variable is

the lack of computationally efficient smoothing or estimation techniques that operate in high dimensions.

The canonical form, given by (1.4) for bivariate distributions, points to the need for caution in interpreting the outcome of a maximal correlation analysis since the solution is only the lead nontrivial term in an expression relating the distribution to that obtained under independence. That this is not an idle concern can be demonstrated with an example, reported by several of the discussants following the paper by Breiman and Friedman [Pregibon and Vardi (1985) and Buja and Kass (1985)] and attributed by one of them to Charles Stone. Consider a bivariate r.v. (X, Y) for which $R(X) = A \cup B$ and $R(Y) = C \cup D$, with A and B and C and D disjoint Borel sets on the real line and where $P(Y \in C|X \in A) = P(Y \in D|X \in B) = 1$. Then $\rho^*(X, Y) = 1$ and it is achieved by the transformations $\phi^*(X) = 1$, $X \in A$, and $\theta^*(Y) = 1$, $Y \in C$.

From a practical point of view, using bivalent transformations to enhance a linear analysis is very unappealing, since other aspects of the dependence are completely eliminated. In the above example, the canonical form (1.4) tells the story: The largest nontrivial singular value is equal to 1 and all others are 0. Hence, an analysis which relies solely on the largest nontrivial singular value, which is essentially a maximal correlation analysis, can be highly misleading. Finding transformations corresponding to these lesser singular values, easily accommodated in ACE or optimal scoring routines, is a recommendable complement to the analysis as Buja (1985) pointed out.

The lesson of this example also bears on the general multivariate problem, of which maximal correlation analysis is a special case. It may be considered a drawback to using maximalized measures of association that an effective interpretation cannot be made in the absence of the corresponding optimal transformations. It should be noted, however, that a "pathology" like that demonstrated in the preceding example can be useful when the difficulty of detecting inhomogeneity in multivariate data is considered. Unlike the bivariate case, there does not seem to be a convenient or tractable canonical form for general multivariate distributions in which g_k^* plays the role of a lead nontrivial singular value, so an obvious analog of a lower eigensolution does not present itself.

Acknowledgments. I would like to express by deepest gratitude to my thesis advisor, Leo Breiman, for suggesting the subject of this paper to me and for his guidance in bringing it to its present form. Also, I would like to thank Kenneth Wachter for his assistance and Ingram Olkin for alerting me to important references.

REFERENCES

BAPAT, R. B. and SUNDER, V. S. (1985). On majorization and Schur products. *Linear Algebra Appl.* **72** 107–117.

BENZÉCRI, J. P. (1969). Statistical analysis as a tool to make patterns emerge from data. In *Methodologies of Pattern Recognition* (S. Watanabe, ed.) 35–60. Academic, New York.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BREIMAN, L. and IHAKA, R. (1984). Nonlinear discriminant analysis via scaling and ACE. Technical Report 40, Dept. Statistics, Univ. of California, Berkeley.

BUJA, A. (1985). Theory of bivariate ACE. Technical Report 74, Dept. Statistics, Univ. of Washington, Seattle.

BUJA, A. and KASS, R. E. (1985). Some observations on ACE methodology. *J. Amer. Statist. Assoc.* **80** 602–607.

CSÁKI, P. and FISCHER, J. (1963). On the general notion of maximal correlation. *Magyar Tud. Akad. Mat. Kutato Int. Közl.* **8** 27–51.

DE LEEUW, J. (1982). Nonlinear principal component analysis. *COMPSTAT 1982* 72–85. North-Holland, Amsterdam.

DE LEEUW, J., YOUNG, F. W. and TAKANE, Y. (1976). Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika* **41** 471–503.

EDWARDS, R. E. (1965). *Functional Analysis: Theory and Practice*. Holt, Rinehart and Winston, New York.

FISHER, R. A. (1940). The precision of discriminant functions. *Ann. Eugenics* **10** 422–429.

FRANKLIN, J. (1968). *Matrix Theory*. Prentice-Hall, Englewood Cliffs, N.J.

GEBELEIN, H. (1941). Das statistische Problem der Korrelation als Variations—und Eigenwertproblem und sein Zusammenhang mit der Ausgleichungsrechnung. *Z. Angew. Math. Mech.* **21** 364–379.

GEKELER, E. (1981). On the pointwise product and the mean value theorem. *Linear Algebra Appl.* **35** 183–191.

HARDY, G. H., LITTLEWOOD, J. E. and PÓLYA, G. (1929). Some simple inequalities satisfied by convex functions. *Messenger Math.* **58** 145–152.

HASTIE, T. (1984). Principal curves and surfaces. Technical Report 11, Dept. Statistics, Stanford Univ.

HIRSCHFELD, H. O. (1935). A connection between correlation and contingency. *Proc. Cambridge Philos. Soc.* **31** 520–524.

KELKER, D. (1970). Distribution theory for spherical distributions and a location-scale parameter generalization. *Sankhyā Ser. A* **32** 419–430.

KOYAK, R. (1985). Optimal transformations for multivariate linear reduction analysis. Ph.D. thesis, Dept. Statistics, Univ. of California, Berkeley.

LANCASTER, H. O. (1958). The structure of bivariate distributions. *Ann. Math Statist.* **29** 719–736.

LANCASTER, H. O. (1969). *The Chi-Squared Distribution*. Wiley, New York.

MARSHALL, A. W. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic, New York.

MCDONALD, R. P. (1967). *Nonlinear Factor Analysis*. Psychometric Monograph 15, Psychometric Soc. William Byrd Press, Richmond, Va.

OWEN, A. (1983). Optimal transformations for autoregressive time series models. Technical Report ORION020, Dept. Statistics, Stanford Univ.

PREGIBON, D. and VARDI, Y. (1985). Comment on "Estimating optimal transformations for multiple regression and correlation," by L. Breiman and J. H. Friedman. *J. Amer. Statist. Assoc.* **80** 598–601.

RÉNYI, A. (1959). On measures of dependence. *Acta. Math. Acad. Sci. Hungar.* **10** 441–451.

SARMANOV, O. V. (1958a). The maximal correlation coefficient (symmetric case). *Dokl. Akad. Nauk SSSR* **120** 715–718. (In Russian.)

SARMANOV, O. V. (1958b). The maximal correlation coefficient (non-symmetrical case). *Dokl. Akad. Nauk SSSR* **121** 52–55. (In Russian.)

SCHUR, I. (1911). Bemerkungen zur Theorie der beschränken Bilinearformen mit unendlich vielen Veränderlichen. *J. Reine Angew. Math.* **140** 1–28.

STYAN, G. (1973). Hadamard products and multivariate statistical analysis. *Linear Algebra Appl.* **16** 217–240.

DEPARTMENT OF
MATHEMATICAL SCIENCES
THE JOHNS HOPKINS UNIVERSITY
BALTIMORE, MARYLAND 21218