The Annals of Statistics
1980, Vol. 8, No. 4, 865–869

CANONICAL VARIABLES AS OPTIMAL PREDICTORS

BY V. J. YOHAI AND M. S. GARCIA BEN
Universidad de Buenos Aires

Let $\mathbf{X} = (X_1, \cdots, X_m)'$ and $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ be two random vectors. Given any random vector $\mathbf{Z}$, let $\mathbf{Y}_{\mathbf{Z}}^*$ be the best linear predictor of $\mathbf{Y}$ based on $\mathbf{Z}$. Let p be any natural number smaller than m. We consider the problem of finding the p-dimensional random vector $\mathbf{Z} = (Z_1, \cdots, Z_p)'$ where each component Z_i is a linear function of $\mathbf{X}$, which minimizes the determinant of $E(\mathbf{Y} - \mathbf{Y}_{\mathbf{Z}}^*)(\mathbf{Y} - \mathbf{Y}_{\mathbf{Z}}^*)'$. We show that $Z_1, \cdots, Z_p$ coincide with the first p canonical variables (except for a nonsingular linear transformation). We also show that the square of the $(p + 1)$th canonical correlation coefficient measures the relative improvement in the prediction of $\mathbf{Y}$ when $p + 1$ Z_i's are used instead of p.

1. Introduction. Let $\mathbf{X} = (X_1, \cdots, X_m)'$ and $\mathbf{Y} = (Y_1, \cdots, Y_n)'$ be two random vectors and assume $m \leqslant n$. Assume also that $E(\mathbf{X}) = E(\mathbf{Y}) = \mathbf{0}$ and let

$$\begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}$$

be the covariance matrix of $(\mathbf{X}', \mathbf{Y}')'$, where $\mathbf{\Sigma}_{11}$ and $\mathbf{\Sigma}_{22}$ are nonsingular matrices.

Classically, the problem of canonical correlation consists of finding vectors $\mathbf{b}_1, \cdots, \mathbf{b}_m$ in R^m and $\mathbf{c}_1, \cdots, \mathbf{c}_n$ in R^n such that if $V_i = \mathbf{b}_i'\mathbf{X}$ and $W_i = \mathbf{c}_i'\mathbf{Y}$ then

(i) V_1, W_1 are the two linear functions of $\mathbf{X}$ and $\mathbf{Y}$ respectively, with variance 1, which have correlation coefficient with maximum absolute value.

(ii) for $i \leqslant m$, V_i is the linear function of $\mathbf{X}$ with variance 1, uncorrelated with $V_1, \cdots, V_{i-1}$, and W_i is the linear function of $\mathbf{Y}$ with variance 1, uncorrelated with $W_1, \cdots, W_{i-1}$, such that the pair (V_i, W_i) has a correlation coefficient with maximum absolute value.

(iii) for $i > m$, W_i has variance 1 and is uncorrelated with $W_1, \cdots, W_{i-1}$.

For $1 \leqslant i \leqslant m$ the pair of variables (V_i, W_i) is called the ith pair of canonical variables and the absolute value of its correlation coefficient ρ_i is called the ith canonical correlation. Clearly $\rho_1^2 \geqslant \rho_2^2 \geqslant \cdots \rho_m^2$ is satisfied.

It is well known [Rao (1973, Section 8f)] that to solve this problem it suffices to find a $m \times m$ matrix $\mathbf{B}$ and a $n \times n$ matrix $\mathbf{C}$ such that

$$\mathbf{B}'\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{B} = \mathbf{R}_1 \tag{1.1}$$

$$\mathbf{B}'\mathbf{\Sigma}_{11}\mathbf{B} = \mathbf{I}_m \tag{1.2}$$

$$\mathbf{C}'\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{C} = \mathbf{R}_2 \tag{1.3}$$

$$\mathbf{C}'\mathbf{\Sigma}_{22}\mathbf{C} = \mathbf{I}_n \tag{1.4}$$

Received August 1978; revised March 1979.
AMS 1970 *subject classifications.* Primary 62H20.
Key words and phrases. Canonical variables, canonical correlations, linear predictors.

where $\mathbf{I}_m$ is the $m \times m$ identity matrix and where $\mathbf{R}_1$ and $\mathbf{R}_2$ are two diagonal matrices with decreasing elements in their diagonals. Then the vectors $\mathbf{b}_i (1 \leq i \leq m)$ and $\mathbf{c}_i (1 \leq i \leq n)$ which solve the problem of canonical correlation are given by the columns of $\mathbf{B}$ and $\mathbf{C}$ respectively. The ith diagonal element of $\mathbf{R}_1$ and $\mathbf{R}_2$ is ρ_i^2 if $1 \leq i \leq m$ and if $i > m$ the ith diagonal element of $\mathbf{R}_2$ is 0. The treatment of the case where $m > n$ is analogous.

The solution to the canonical correlation problem is unique (except for a change of sign in the $\mathbf{b}_i$'s or the $\mathbf{c}_i$'s) if and only if the numbers ρ_i^2 are all different.

In this approach to the canonical correlation problem, the vectors $\mathbf{X}$ and $\mathbf{Y}$ play symmetrical roles, but in many practical problems their roles differ. This happens for example when the components of $\mathbf{X}$ are observable variables correlated to the components of $\mathbf{Y}$, while the components of $\mathbf{Y}$ are not observable or have high cost of observation. In this case the researcher may be interested in using $\mathbf{X}$ to predict $\mathbf{Y}$. If m is very large it would be useful to summarize the information contained in $\mathbf{X}$ in a few variables $Z_1, \cdots, Z_p$, linear functions of $\mathbf{X}$:

$$Z_i = \mathbf{a}_i'\mathbf{X},$$

choosing $\mathbf{a}_i$, $1 \leq i \leq p$, such that the vector $\mathbf{Z} = (Z_1, \cdots, Z_p)'$ be the best for linearly predicting the vector $\mathbf{Y}$. This may be formalized as follows: Let $\mathbf{Y}_Z^*$ be the least square predictor of $\mathbf{Y}$ based on $\mathbf{Z}$. Then $\mathbf{Y}_Z^*$ is given by [Rao (1973, Section 4g)]:

$$\mathbf{Y}_Z^* = E(\mathbf{Y}\mathbf{Z}')\, E(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}. \tag{1.5}$$

It is well known that $\mathbf{Y}_Z^*$ is the best linear predictor of $\mathbf{Y}$ based on $\mathbf{Z}$ using either of the following criteria:

(i) it minimizes $E(\|\mathbf{Y} - \mathbf{Y}_Z^*\|^2)$,

and

(ii) it minimizes $|E(\mathbf{Y} - \mathbf{Y}_Z^*)(\mathbf{Y} - \mathbf{Y}_Z^*)'|$,

among all predictors of the forms $\mathbf{Y}_Z^* = \mathbf{D}\mathbf{Z}$, where $\mathbf{D}$ is any $n \times p$ matrix. ($|\ |$ indicates the matrix determinant and $\|\ \|$ the vector Euclidean norm).

Then we may define the best p-vector $\mathbf{Z}$ for predicting $\mathbf{Y}$ using two different criteria:

(a) the vector $\mathbf{Z}$ which minimizes

$$E(\|\mathbf{Y} - \mathbf{Y}_Z^*\|^2) \tag{1.6}$$

or

(b) the vector $\mathbf{Z}$ which minimizes

$$|E(\mathbf{Y} - \mathbf{Y}_Z^*)(\mathbf{Y} - \mathbf{Y}_Z^*)'|. \tag{1.7}$$

The problem of finding $\mathbf{Z}$ which minimizes (1.6) is treated in Rao (1973, Chapter 8, Problem 2). The variables $Z_1, \cdots, Z_p$ which solve this problem are in general different from the first p canonical variables, being the same in the particular case that $\mathbf{\Sigma}_{22}$ is of the form $\lambda\, \mathbf{I}_n$ where λ is a scalar.

On the other hand, if the criterium for choosing $\mathbf{Z}$ is to minimize (1.7), we will show that a solution is to choose $\mathbf{Z} = (V_1, \cdots, V_p)'$ where $V_1, \cdots, V_p$ are the first canonical variables.

2. Proofs. We will prove the following theorem:

THEOREM. *Consider the problem of choosing a $m \times p$ matrix $\mathbf{A}^*$ such that $\mathbf{Z}^* = \mathbf{A}^{*\prime}\mathbf{X}$ minimizes* (1.7), *among all the p-dimensional vectors $\mathbf{Z} = \mathbf{A}'\mathbf{X}$. Then*

(i) *The $m \times p$ matrix $\mathbf{A}_0$ given by the first p columns of the matrix $\mathbf{B}$ satisfying* (1.1) *and* (1.2) *is a solution to this problem.*

(ii) *If $\rho_p^2 > \rho_{p+1}^2$ then every other solution $\mathbf{A}^*$ is of the form $\mathbf{A}^* = \mathbf{A}_0\mathbf{G}$ where $\mathbf{G}$ is any nonsingular $p \times p$ matrix.*

On the other hand with k equal eigenvalues $\rho_{q+1}^2 = \rho_{q+2}^2 = \cdots = \rho_{q+k}^2$ and $q < p < q + k$ the $p - q$ last columns of $\mathbf{A}_0$ can be chosen as any set of $p - q$ orthogonal eigenvectors associated with the common eigenvalue, and in this framework every solution can be written in the form $\mathbf{A}^* = \mathbf{A}_0\mathbf{G}$.

(iii) *The minimum value of* (1.7) *is given by*

$$|\mathbf{\Sigma}_{22}| \prod_{i=1}^{p} (1 - \rho_i^2).$$

PROOF. Replacing in (1.7) $\mathbf{Y}_Z^*$ by its expression (1.5), it turns out that (1.7) is equivalent to

$$|E(\mathbf{YY}') - E(\mathbf{YZ}')E(\mathbf{ZZ}')^{-1}E(\mathbf{ZY}')|. \tag{2.1}$$

Let us note that the best linear predictor of $\mathbf{Y}$ based on $\mathbf{Z}$ is the same as the best linear predictor of $\mathbf{Y}$ based on $\mathbf{DZ}$ for any $p \times p$ nonsingular matrix $\mathbf{D}$. $\mathbf{D}$ may be always chosen such that the covariance matrix of $\mathbf{DZ}$ be the identity. Then without loss of generality we may choose $\mathbf{A}^*$ among the matrices $\mathbf{A}$ such that:

$$E(\mathbf{ZZ}') = \mathbf{A}'\mathbf{\Sigma}_{11}\mathbf{A} = \mathbf{I}_p. \tag{2.2}$$

Replacing $\mathbf{Z}$ by $\mathbf{A}'\mathbf{X}$ in (2.1) and using (2.2) the expression (1.7) to be minimized may be written

$$|\mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{AA}'\mathbf{\Sigma}_{12}|$$

and this is equal to [Press (1972, Formula 2.4.2)]:

$$|\mathbf{\Sigma}_{22}|\,|\mathbf{I}_p - \mathbf{A}'\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{A}|. \tag{2.3}$$

Since the first factor does not depend on $\mathbf{A}$, the problem of minimizing (1.7) is reduced to finding a $m \times p$ matrix $\mathbf{A}$ such that

$$|\mathbf{I}_p - \mathbf{A}'\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{21}\mathbf{A}| \tag{2.4}$$

is minimized, subject to the restriction (2.2).

Let $\mathbf{B}$ be a matrix satisfying (1.1) and (1.2) and put $\mathbf{H} = \mathbf{B}^{-1}\mathbf{A}$. Then replacing $\mathbf{A} = \mathbf{BH}$ and using (1.1), (2.4) is equivalent to

$$|\mathbf{I}_p - \mathbf{H}'\mathbf{R}_1\mathbf{H}|. \tag{2.5}$$

Moreover by (2.2) and (1.2) the matrix $\mathbf{H}$ satisfies

$$\mathbf{H}'\mathbf{H} = \mathbf{I}_p. \tag{2.6}$$

Given any $p \times p$ symmetric matrix $\mathbf{A}$ we denote by $\lambda_i(\mathbf{A})$ the ith largest eigenvalue of $\mathbf{A}$. Then (2.5) is equivalent to

$$\prod_{i=1}^{p} (1 - \lambda_i(\mathbf{H}'\mathbf{R}_1\mathbf{H})). \tag{2.7}$$

According to Lemma 2.6 of Okamoto (1969) a $p \times m$ matrix $\mathbf{H}^*$ minimizes (2.7) if and only if

$$\mathbf{H}^* = \mathbf{S}\mathbf{Q}, \tag{2.8}$$

where $\mathbf{Q}$ is any nonsingular $p \times p$ matrix, in particular we may take

$$\mathbf{Q} = \mathbf{I}_p$$

and $\mathbf{S}$ is any $m \times p$ matrix whose columns are eigenvectors of $\mathbf{R}_1$ corresponding to the first p largest eigenvalues. Since $\mathbf{R}_1$ is diagonal with nondecreasing elements in its diagonal, the first p vectors of the canonical base on $\mathbf{R}^P$ satisfy this property. Therefore $\mathbf{H}^*$ may be taken equal to

$$\mathbf{H}_0 = \begin{pmatrix} \mathbf{I}_p \\ \mathbf{O} \end{pmatrix},$$

where $\mathbf{O}$ denotes the $(m - p) \times p$ matrix with all its elements 0. Then $\mathbf{A}_0 = \mathbf{B}\mathbf{H}_0$ is a solution to the problem of minimizing (1.7), where the matrix $\mathbf{A}_0$ is formed by the first p columns of $\mathbf{B}$.

The proof of (iii) follows immediately replacing $\mathbf{A}$ by $\mathbf{A}_0$ in (2.3) and using (1.1).

To prove (ii) it is enough to observe that given any other matrix $\mathbf{A}^*$ such that $\mathbf{Z}^* = \mathbf{A}^{*\prime}\,\mathbf{X}$ minimizes (1.7), we may obtain a nonsingular matrix $\mathbf{D}$ such that $\tilde{\mathbf{A}} = \mathbf{A}^*\mathbf{D}'$ satisfies (2.2). Denote $\mathbf{H}^* = \mathbf{B}^{-1}\tilde{\mathbf{A}}$. Then from (2.8) and the fact that $\rho_p^2 > \rho_{p+1}^2$ we have

$$\mathbf{H}^* = \begin{pmatrix} \mathbf{I}_p \\ \mathbf{O} \end{pmatrix}\mathbf{Q}$$

where $\mathbf{Q}$ is a $p \times p$ nonsingular matrix. Then

$$\mathbf{A}^* = \mathbf{B}\begin{pmatrix} \mathbf{I}_p \\ \mathbf{O} \end{pmatrix}\mathbf{Q}\mathbf{D}'^{-1} = \mathbf{A}_0\mathbf{Q}\mathbf{D}'^{-1}$$

and denoting $\mathbf{G} = \mathbf{Q}\mathbf{D}'^{-1}$ we obtain (ii).

In the case where $\rho_{q+1}^2 = \rho_{q+2}^2 = \cdots = \rho_{q+k}^2$ and $q < p < q + k$ the matrix of the $p - q$ last columns of $\mathbf{A}_0$ can be replaced by

$$(\mathbf{b}_{q+1}, \mathbf{b}_{q+2}, \cdots, \mathbf{b}_{q+k})\mathbf{F}$$

where F is a $k \times (p - q)$ matrix such that $\mathbf{F}'\mathbf{F} = \mathbf{I}_{p-q}$. After this change the solution $\mathbf{A}^* = \mathbf{A}_0\mathbf{G}$ depends on both $\mathbf{F}$ and $\mathbf{G}$, and every solution which minimizes (2.4) can be written in this way.

REMARK. Point (iii) of the theorem yields an interpretation of the square of the $p + 1$-canonical correlation ρ_{p+1}: it measures the relative improvement in the prediction of **Y** when a $(p + 1)$-dimensional vector **Z** is used instead of a p-dimensional one. In effect from point (iii) of the above theorem we have that the determinant of the covariance matrix of the residual vector $\mathbf{Y} - \mathbf{Y}_{\mathbf{Z}}^*$ when an optimal p-dimensional vector **Z** is used is

$$|\Sigma_{22}| \prod_{i=1}^{p} (1 - \rho_i^2).$$

If a $(p + 1)$-dimensional optimal vector **Z** is used the determinant will be reduced to

$$|\Sigma_{22}| \prod_{i=1}^{p+1} (1 - \rho_i^2)$$

and then the relative reduction of the determinant is ρ_{p+1}^2.

Acknowledgment. We are grateful to a referee for suggesting extensions of our original results to the case where some of the eigenvalues are equal.

REFERENCES

[1] OKAMOTO, M. (1969). Optimality of principal components. In *Proc. Second Internat. Symp. Multivariate Anal.* (P. R. Krishnaiah, ed.). 673–685. Academic Press, New York.

[2] PRESS, S. J. (1972). *Applied Multivariate Analysis*. Holt, Rinehart and Winston, U.S.A.

[3] RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

DEPARTAMENTO DE MATEMATICAS
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
CIUDAD UNIVERSITARIA
PABELLON 1
1428 BUENOS AIRES
ARGENTINA