PSYCHOMETRIKA–VOL. 46, NO. 2.
JUNE, 1981

RELATIONSHIPS BETWEEN REDUNDANCY ANALYSIS, CANONICAL CORRELATION, AND MULTIVARIATE REGRESSION

KEITH E. MULLER

UNIVERSITY OF NORTH CAROLINA

This paper attempts to clarify the nature of redundancy analysis and its relationships to canonical correlation and multivariate multiple linear regression. Stewart and Love introduced redundancy analysis to provide non-symmetric measures of the dependence of one set of variables on the other, as channeled through the canonical variates. Van den Wollenberg derived sets of variates which directly maximize the between set redundancy. Multivariate multiple linear regression on component scores (such as principal components) is considered. The problem is extended to include an orthogonal rotation of the components. The solution is shown to be identical to van den Wollenberg's maximum redundancy solution.

Introduction

Hotelling [1936] derived canonical correlation as a method for finding linear combinations of two sets of variables which are maximally correlated. Stewart and Love [1968] introduced redundancy analysis as an interpretational aid to canonical correlation. Controversy arose immediately as to what it actually measures and whether what it measures is of any utility. [See Dawson, Note 1, for a detailed review.] Van den Wollenberg [1977] derived sets of variates which directly maximize the between set redundancy (whereas the canonical solution maximizes the between set canonical correlation).

This paper considers a number of aspects of redundancy analysis in an attempt to clarify its nature. First, a seemingly unrelated problem is defined. Its solution is shown to be the same as van den Wollenberg's. Consequently, the model equation studied gives some insight into the nature of redundancy analysis.

Some Standard Results

This section summarizes a number of standard results in regression. This provides a convenient way to introduce notation and context needed in the sequel. First, consider the usual multivariate multiple regression model equation for q criterion variables and p predictors, with both sets standardized (zero mean and unit variance):

$$\underset{n \times q}{Z_y} = \underset{(n \times p)(p \times q)}{Z_x B} + \underset{n \times q}{E} \quad . \tag{1}$$

Assume that the number of observations, $n \gg p + q$, and that the usual least squares assumptions of independence, linearity and homoscedasticy hold. It may be the case that the rank of Z_x is p^*, strictly less than p. Then, of course, the usual estimator of B does not exist. One standard response is to replace the original p variables with p^* (full rank)

This research was supported in part by U.S. Environmental Protection Agency contract 68-02-3402. The author gratefully acknowledges the stimulation of Maurice Tatsuoka and Beth Dawson-Saunders in first interesting him in redundancy analysis, as well as a useful change suggested by Warren Sarle.

Requests for reprints should be sent to Keith E. Muller, Department of Biostatistics, Rosenau Hall 201H, University of North Carolina, Chapel Hill, North Carolina, 27514.

0033-3123/81/0600-3017$00.75/0

variables, such as the principal components, or other "factor" variables. The procedure identifies a new (full rank of p^*) model:

$$\underset{n \times q}{Z_y} = \underset{(n \times p^*)(p^* \times q)}{Z_{x^*} B_1} + \underset{n \times q}{E}, \tag{2}$$

$$\underset{n \times p^*}{Z_{x^*}} = \underset{(n \times p)(p \times p^*)}{Z_x T} \tag{3}$$

The value of B_1 is uniquely defined, once a particular transformation T has been chosen, and the usual estimator of B_1 exists.

It is often convenient to require orthogonality for the new variables:

$$\frac{1}{n} Z'_{x^*} Z_{x^*} = R_{x^*x^*} = I_{p^*} = T' R_{xx} T. \tag{4}$$

Hence, any choice of F, $p \times p^*$, rank p^*, such that

$$R_{xx} = FF' \tag{5}$$

implies

$$T = F(F'F)^{-1}. \tag{6}$$

These results imply a choice for $\hat{B}$, namely

$$\hat{B}_0 = T\hat{B}_1 \tag{7}$$

$$\hat{B}_0 = T(R^{-1}_{x^*x^*} R_{x^*y}) \tag{8}$$

$$= TT' R_{xy}. \tag{9}$$

Since TT' is a generalized inverse of R_{xx} [see Khatri, 1976, for a related discussion], the factor score approach and generalized inverse approach to the less then full rank regression problem are equivalent in many important ways.

A Related Problem

Assume that an orthogonal (rank $r \leq p^*$) rotation, A_*, of the factor scores is desired. The model then becomes

$$\underset{n \times q}{Z_y} = \underset{(n \times p)(p \times p^*)(p^* \times r)(r \times q)}{Z_x T A_* B_*} + \underset{n \times q}{E} \tag{10}$$

Taking the expectation of (10) and replacing the expectation of Z_y with Z_y gives an equation to be solved for estimators of the parameters:

$$Z_y = Z_x T A_* B_*. \tag{11}$$

Premultiplying by $(1/n)A'_* T' Z'_x$ produces

$$A'_* T' R_{xy} = A'_* T' R_{xx} T A_* B_* \tag{12}$$

$$= A'_* A_* B_* \tag{13}$$

$$= B_*. \tag{14}$$

Using this result in (11) and premultiplying by $(1/n)T' Z'_x$ gives

$$T' R_{xy} = T' R_{xx} T A_* A'_* T' R_{xy} \tag{15}$$

$$= A_* A'_* T' R_{xy}. \tag{16}$$

Postmultiplying by $R_{yx}TA_*$, the equation then becomes

$$T'R_{xy}R_{yx}TA_* = A_*A_*'T'R_{xy}R_{yx}TA_*. \tag{17}$$

Upon reflection, it can be seen that this equation is satisfied by choosing A_* as a matrix whose columns are eigenvectors of the symmetric matrix

$$M = T'R_{xy}R_{yx}T. \tag{18}$$

Equivalence to Maximizing Redundancy

The purpose behind the above development is quite simple. It is easy to show that the above solution for choosing linear combinations of Z_x is equivalent to van den Wollenberg's solution for choosing linear combinations of Z_x which maximize the redundancy of Z_y given Z_x. Stewart and Love [1968] defined the redundancy statistic as the mean variance of one set explained by a canonical variate of the other set. The total redundancy is the sum over the canonical variates. The k^{th} redundancy is equal to a constant times the inner product of the canonical factor loadings:

$$\frac{\rho^2}{q}\mathbf{b}_k'R_{yy}R_{yy}\mathbf{b}_k = Rd_{y|x}(k). \tag{19}$$

Here k indicates a particular canonical variate, $\mathbf{b}_k$ the canonical weights for Z_y.

Van den Wollenberg [1977] expressed the redundancy in an alternate form. First, one of the two solution equations from canonical correlation is

$$\mathbf{0} = R_{yx}\mathbf{a}_k - \rho_k R_{yy}\mathbf{b}_k. \tag{20}$$

Here $\mathbf{a}_k$ is the k^{th} set of canonical weights for Z_x. Hence

$$\rho_k R_{yy}\mathbf{b}_k = R_{yx}\mathbf{a}_k. \tag{21}$$

Taking the inner product of each side separately gives

$$\rho_k^2\mathbf{b}_k'R_{yy}R_{yy}\mathbf{b}_k = \mathbf{a}_k'R_{xy}R_{yx}\mathbf{a}_k. \tag{22}$$

It follows immediately that

$$Rd_{y|x}(k) = \frac{1}{q}\mathbf{a}_k'R_{xy}R_{yx}\mathbf{a}_k. \tag{23}$$

Therefore the redundancy statistic may also be characterized as the mean squared loading of one set on a canonical variate of the other set. Some of the controversy centers on the fact that redundancy is not symmetric, while canonical correlations are (with respect to the labeling of one set as Z_x and the other as Z_y). Of course, the usual multiple R is also not symmetric in that sense.

To see the claimed equivalence, consider van den Wollenberg's solution equation to maximize redundancy:

$$(R_{xy}R_{yx} - \lambda R_{xx})\mathbf{a} = \mathbf{0}. \tag{24}$$

The linear combination of Z_x sought is $\mathbf{a}$, $p \times 1$. This two-matrix eigenvalue problem may be solved in various ways. The most common technique used in canonical correlation uses a factor, F, of R_{xx}. It may be applied here as follows:

$$(R_{xy}R_{yx} - \lambda FF')\mathbf{a} = \mathbf{0}. \tag{25}$$

Premultiplying by $(F'F)^{-1}F'$ gives

$$((F'F)^{-1}F'R_{xy}R_{yx} - \lambda F')\mathbf{a} = \mathbf{0}. \tag{26}$$

Letting

$$\mathbf{a} = F(F'F)^{-1}\mathbf{a}_* = T\mathbf{a}_* \tag{27}$$

the expression (26) becomes

$$((F'F)^{-1}F'R_{xy}R_{yx}F(F'F)^{-1} - \lambda I)\mathbf{a}_* = \mathbf{0}. \tag{28}$$

With T defined as in (6), (28) becomes

$$(T'R_{xy}R_{yx}T - \lambda I)\mathbf{a}_* = \mathbf{0}. \tag{29}$$

This again is the eigenvector solution equation for the matrix M defined in (18). Note that throughout, either Z_x or Z_y or both may be less than full rank.

Conclusions

The results of the last two sections may be summarized as follows: (i) equation (10) is the model equation for maximizing redundancy, (ii) which is equivalent to finding a method of moments estimate of a choice of orthogonal rotation of factor scores used for regression. Redundancy analysis stands between canonical correlation and multivariate multiple regression. Canonical correlation may be thought of as a process of orthogonalizing the Z_x correlation matrix and the Z_y correlation matrix, then providing an orthogonal transformation of each (orthogonal) set to maximize and orthogonalize the between set correlations. Multivariate multiple regression provides a single transformation from one original space to the other original space. Redundancy analysis orthogonalizes one set of variables (at a time) and then provides an orthogonal transformation to use in predicting into the original space of the other set. Redundancy analysis should be treated as evaluating adequacy of regression (prediction) and not association. It shifts the usual emphasis in canonical correlation toward multivariate multiple regression. In the former, the two sets of variables stand in a strongly symmetric relationship. In the latter the two sets are usually treated and discussed quite differently.

REFERENCES

Dawson, B. *The sampling distribution of the canonical redundancy statistic.* Unpublished doctoral dissertation, University of Illinois, 1977.

Hotelling, H. Relations between two sets of variates. *Biometrika*, 1936, *28*, 321–377.

Khatri, C. G. A note on multiple and canonical correlation for a singular covariance matrix. *Psychometrika*, 1976, *41*, 465–470.

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

van den Wollenberg, A. L. Redundancy analysis, an alternative for canonical correlation analysis. *Psychometrika*, 1977, *42*, 207–219.

Manuscript received 9/4/80
Final version received 12/8/80