ED 109 237 TM 004 699

AUTHOR Rim, Eui-Do
TITLE A Stepwise Canonical Procedure and the Shrinkage of Canonical Correlations.
PUB DATE [Apr 75]
NOTE 13p.; Paper presented at the Annual Meeting of the American Educational Pesearch Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Analysis of Variance; *Comparative Analysis; *Correlation; *Factor Analysis; Matrices; Prediction; *Predictor Variables; *Statistical Analysis; Validity

ABSTRACT

A stepwise canonical procedure, including two selection indices for variable deletion and a rule for stopping the iterative procedure, was derived as a method of selecting core variables from predictors and criteria. The procedure was applied to simulated data varying in the degree of built in structures in population correlation matrices, number of variables, and number of cases. A double cross-validation method was used to test the stability of the canonical correlations. The magnitudes and shrinkage of the largest and the mean significant canonical correlations were compared by means of MANOVA across the different decision rules, built in correlation structures. number of starting variables, and number of cases. (Author)

A STEPWISE CANONICAL PROCEDURE AND THE SHRINKAGE OF CANONICAL CORRELATIONS

Eui-Do Rim
Research for Better Schools, Inc.

A E R A
Washington, D. C., 1975

A STEPWISE CANONICAL PROCEDURE AND - THE SHRINKAGE OF CANONTICAL CORRELATIONS[1]

Eui-Do Rim

Research for Better Schools, Inc.

INTRODUCTION

Educational research in natural settings is inherently multivariate in structure and as such involves numerous related variables and correlated criteria. Selection and placement problems of a school, a firm, or an industry are cases in point, as is the evaluation of educational programs and policy. However, in most instances it is not practicable to take all possible predictors and criteria into consideration. The researcher usually has to select restricted numbers of variables from each of predictor and vriterion variable sets, even when he/she uses multivariate research designs.

The primary concern of the present study was to develop an analytical stepwise canonical procedure that could be used to systematize the selection of core variables from two sets of variables, namely, predictors and criteria. The procedure was designed to select those core variables which could explain, in the most parsimonious way, the relationships between the two sets of variables. It was also hoped that the core variables selected by the procedure would provide users with more stable statistics, as tested by the cross-validation method, than the use of whole sets of variables.

[1] This study is a part of the dissertation which was submitted in partial fulfillment of the requirements for the degree of Ph.D. in Education in the Graduate College of the University of Illinois at Urbana-Champaign, 1972. The author would like to extend his gratitude to Dr. M. M. Tatsuoka, academic advisor, and his dissertation committee members for all the support they rendered for this study.

At the same time, the present study aimed to compare the magnitudes of canonical correlations and shrinkage in relation to the proportion of variables retained, across different built-in population correlation structure groups, the number of starting variables, and the number of cases (sample size).

ANALYTICAL PROCEDURE

The problem of deriving an analytical procedure involved the development of an optimum selection procedure, a selection index, and a stopping rule. A backward elimination procedure was derived for this study. Four procedures were examined as potential sources for deriving a selection index: the correlation of each variable with its canonical-variate (similar to Huberty's study, 1971); the principal components of both predictor, and criterion correlation matrices; the multiple correlation of each variable with the variables in the other variable set (Hall, 1971); and eigenvectors obtained from canonical analysis. The two selection indices chosen for use in this study were derived via the eigenvector approach.

Selection Index I

As Roy (1957) observed, the largest eigenvalue and its associated eigenvector play important roles in most multivariate analyses. The weight vectors ($\underline{u}_{k1}$ and $\underline{v}_{m1}$) associated with the largest eigenvalue (the square of the largest canonical correlation, ρ_1^2) were used as criteria for selecting a variable to be deleted. In other words, the variable that had the smallest absolute value was deleted at each stage of the variable elimination procedure.

Selection Index II

The second index was derived from a weighted sum of the absolute values of vectors that were associated with significant canonical correlations. The values $\sum_{i=1}^{r} |\rho_i^2 u_{ki}|$ and $\sum_{i=1}^{r} |\rho_i^2 v_{mi}|$ were computed for each of the predictors and criteria, respectively, where r represented the number of significant canonical correlations as tested by Bartlett's method. The variable that had the smallest value was deleted at each stage of the stepwise analysis.

A combined criterion of rational and arbitrary criteria was established as a rule for deciding when to stop the process of variable deletion. The stepwise procedure for variable elimination was terminated when the deletion of a variable led to the satisfying of either one of the stopping rules described below. At that point, the last variable deleted was restored into its variable set.

Stopping Rule 1

For practical purposes, the first stopping rule was derived from Bartlett's (1941) χ^2 approximation. In the significance testing of canonical correlations of p predictors and q criteria,

$$- \{ \nu - \frac{1}{2} (p + q + 1) \} \ln \Lambda \qquad (1)$$

is distributed approximately as χ^2 with pq degrees of freedom. Then one variable is removed either from predictors or from criteria,

$$- \{ \nu - \frac{1}{2} (p' + q' + 1) \} \ln \Lambda' \qquad (2)$$

is now distributed as χ^2 with p'q' degrees of freedom, where p'q' is either (p - 1)q or p(q -1), so that

$$\frac{(\chi^2 pq - \chi^2 p'q')/(pq - p'q')}{\chi^2\ p'q'/p'q'} \tag{3}$$

is at least approximately distributed as an F-variable with (pq - p'q') and p'q' degrees of freedom for the numerator and denominator, respectively. If the F-value exceeded a predetermined value such as $p<.05$, the stepwise procedure of variable deletion was terminated.

In the same way the significance of the decrease of significant canonical correlations, resulting from the deletion of one variable, was tested by computing the difference of two χ^2 differences and their degrees of freedom;

$$\frac{\{\chi^2_{pq} - \chi^2_{(p-r)(q-r)}\} - \{\chi^2_{p'q'} - \chi^2_{(p'-r)(q'-r)}\}}{\{\chi^2_{p'q'} - \chi^2_{(p'-r)(q'-r)}\} / (p' + q' - r)} = F\ , \tag{4}$$

with 1 and (p' + q' - r) degrees of freedom for the numerator and denominator, respectively, and where r represented the number of significant canonical correlations.

Stopping Rule 2

Hotelling (1936) has indicated that Wilks' lambda, defined by

$$\Lambda = \prod_{i=1}^{q} (1 - \rho_i^2)\ , \tag{5}$$

is an index of alienation. Then $(1-\Lambda)$ represents the variance accounted for by canonical correlations. The ratio,

$$\frac{1 - \Lambda_{p'q'}}{1 - \Lambda_{pq}}\ , \tag{6}$$

shows the proportion of variance accounted for by the reduced set of variables as compared to that accounted for by the whole set of variables.

Hence, the stepwise variable deletion procedure was terminated when

$$\frac{1 - \Lambda_{p'q'}}{1 - \Lambda_{pq}} < (1 - \alpha) \text{ , say, } .90 \text{ .} \qquad (7)$$

Cross-Validation

The stability of sample canonical correlations was tested by a cross-valication method. The cross-validated canonical correlations were computed by elementwise division of the diagonal elements of $\underline{U}_1'$ $\underline{R}_{xy}2$ $\underline{V}1$ -- the numerator -- by the corresponding square root of the elementwise product of a diagonal element of $\underline{U}_1'$ $R_{xx}2$ $\underline{U}_1$ and one of $\underline{V}_1'$ $\underline{R}_{yy}2$ $\underline{V}_1$ -- the denominator;

$$r_{ci} = \frac{\underline{u1}_i' \; \underline{R}_{xy}2 \; \underline{v1}_i}{\sqrt{(\underline{u1}_j' \; \underline{R}_{xx}2 \; u1_i) \; (\underline{v1}_i' \; \underline{R}_{yy}2 \; \underline{v1}_i)}} \qquad (8)$$

where the subscript "1" represents the derivation sample, and the subscript "2," the validation sample. A double cross-validation method was used; that is, the weights derived from one sample were cross-validated on the other sample, and vice versa.

DATA AND METHODS OF COMPARISON

The analytical procedure was applied to a series of simulated data varying in degree of built-in structures in the form of population correlation matrices, number of starting variables, and number of cases (sample size). Two varying factors were built in the structures of population correlation matrices: the number of clusters in the predictor and criterion sets (0-0, 0-1, 1-1, 1-2, and 2-2 cluster groups); and, high or low predictor-criterion intercorrelations. For convenience, the same number of variables were used as predictors and criteria at the beginning. Cases with 6, 9, and 12 variables for each of the predictor and criterion sets were investigated in this study. Thus, thirty-six different population correlation matrices were generated for this study.

From each population correlation matrix, two small and two large sample correlation matrices were generated by using Kaiser-Dickman's method (1962) in order to contrast small and large sample cases, and for cross-validation purposes (144 sample correlation matrices in total).

The magnitudes of resulting canonical correlations and skrinkage, and the proportion of variables retained were compared by using 3 X 3 X 6 X 2 X 2 fixed factorial design multivariate analysis of variance across: variable deletion methods (3 levels: Method I = use of whole set of variables; Method II = use of Selection Index I; and Method III = use of Selection Index II), the initial numbers of variables (3 levels: 12, 18, and 24 variable groups), the six built-in population correlation structure groups, the high and low predictor-criterion intercorrelation groups, and small and large sample groups. The MANOVA was performed for the largest canonical correlation and its shrinkage and for mean significant canonical correlations and mean shrinkage, separately. Canonical correlation coefficients were transformed into Fisher's Z's and proportions of variables retained were converted into radians by an arcsine transformation before the MANOVA was performed (Edwards, 1968).

Data generation, canonical analyses, cross-validation, data transformation and data analysis were all processed by IBM 360/50-75 at the University of Illinois by using the SOUPAC and UMAVAC programs.[2]

FINDINGS AND DISCUSSIONS

About 36 percent of the variables were eliminated by using Selection Index I, and approximately 38 percent by Selection Index II. The deletion of variables brought about considerable decrease in the magnitudes of the largest and mean significant canonical correlations. However, the magnitudes of shrinkage were significantly smaller for the reduced sets of variables than for the whole sets of variables. (The difference between the shrinkage of the largest canonical correlation of the reduced sets of variables by Selection Index I and that of the whole sets of variables, however, was not statistically significant.)

Therefore, it is concluded that the elimination of lesser contributing variables by either Selection Index I or Selection Index II results in more stable canonical correlation. It still remains to be determined if the sample canonical correlations obtained by the present stepwise procedure are more stable than those obtained by any random selection method.

When the canonical correlations resulting from the use of Selection Index I were compared with those resulting from the use of Selection Index II, the former were slightly larger than the latter. A few more variables were deleted in the latter case. The magnitudes of shrinkage of the largest and mean significant canonical correlations of the Selection Index I group

[2] Department of Computer Science, University of Illinois, SOUPAC program descriptions: Statistically Oriented Users Programming and Consulting, 1972. The UMAVAC was originally developed by J. D. Finn and adapted for use at the University of Illinois by J. L. Wardrop and T. J. Blish.

were also larger than those of the Selection Index II group, although the differences were not statistically significant. Most other comparisons between Selection Index I and Selection Index II showed similar results. Hence, it seems that the strategy of choosing a selection index should depend upon the user's intention. Selection Index I caused premature termination when the variable to be deleted had large weights on significant canonical-variates other than the first one. This was especially true when the deleted variable had large weights on the second and the third variates. Selection Index I is recommended when the user is interested only in the largest canonical correlation and when he can be satisfied with a rough sieving of variables. Selection Index II quires more complicated computations, but gives more stable sample statistics with fewer variables.

One of the most important sources of variation in both univariate and multivariate analyses of variance is the sample size factor. In the present study, it was the dominating source of variation in the analyses of mean significant canonical correlations and shrinkage. Apparently more variables were deleted from the small sample group (43%) than from the large sample group (32%). However, the magnitudes of the resulting largest and mean significant canonical correlations of the small sample group were still larger than those of the large sample group. The difference between the means of mean significant canonical correlations of the small and large sample groups was prominent because the number of significant canonical correlations increased as the number of cases increased. The amounts of decrease in canonical correlations caused by the deletion of variables of the small sample group were evidently larger than those of the large sample group, mostly because more variables were eliminated from the former group.

The shrinkage in the small sample group was also larger than that in the large sample group. When the decremental tendencies of shrinkage were compared for each sample size group, the decreases in shrinkage due to deletion of variables were conspicuous for the small sample group, but there was no noticeable change for the large sample group. This implies that when sample size is small, the proposed variable elimination procedures will be more effective in achieving more stable pairs of canonical-variates.

The predictor-criterion intercorrelation factor was another important source of variation. Noticeably more variables were removed from the high intercorrelation group (48%) than from the low intercorrelation group (28%). The magnitudes of the largest and mean significant canonical correlations of the high intercorrelation group were still larger than those of the low intercorrelation group. The amounts of decrease in the largest and mean significant canonical correlations due to the variable elimination were significantly larger for the high intercorrelation group than for the low intercorrelation group, because more variables were deleted from the former group.

On the other hand, the magnitudes of shrinkage in the high intercorrelation group were smaller than those in the low intercorrelation group. The same tendency remained after deletion of variables. Consequently, it can be said that the present variable-deletion procedures are applicable to both high and low intercorrelation cases, and they are more efficient in the sense that they give more stable sample statistics with fewer variables when the predictor-criterion intercorrelations are large.

The number of starting variables was the next most important factor in the present study. Average percentages of the eliminated variables were 31, 35, and 47 percent for the groups starting with 12, 18, and 24 variables, respectively. When the proportions of the eliminated variables are converted into the actual number of variables, four variables on the average were deleted from the group starting with 12 variables, six from the group starting with 18 variables, and eleven from the group starting with 24 variables. The largest and mean significant canonical correlations of the whole set of variables of the group starting with 18 variables (.982 and .840) were significantly larger than those of the groups starting with 12 (.925 and .834) and 24 variables (.967 and .817). The mean of the largest canonical correlation was originally the smallest for the group starting with 12 variables, and while the mean of significant canonical correlation means was the smallest for the group starting with 24 variables. The variable deletion did not affect the relative standings of the largest and mean significant canonical correlations of these groups. However, the shrinkage for the original whole set of variables of the group starting with 12 variables was significantly larger than those for the groups starting with 18 and 24 variables. Generally, deletion of variables caused decrease in amounts of shrinkage. The declining rate was the sharpest for the group starting with 24 variables and the mildest for the group starting with 12 variables. Accordinglv, it is concluded that the proposed variable-elimination methods are more useful when the number of

REFERENCES

Anderson, T. W. *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1958.

Bartlett, M. S. The statistical significance of canonical correlations. *Biometrika*, 1941, 32, 29 - 37.

Cooley, W. W. and Lohnes, P. R. *Multivariate data analysis.* New York: Wiley, 1971.

Edwards, A. L. *Experimental design in psychological research.* (3rd ed.) New York: Holt, 1968.

Efrovmson, M. A. Multiple regression analysis. Chapter 17 in A. Ralston and H. S. Wilf, *Mathematical methods for digital computers.* New York: Wiley, 1960, 191 - 203.

Hall, C. E. Generalizing the Wherry-Doolittle battery reducing procedure to canonical correlation and MANOVA. *The Journal of Experimental Education*, 1971, 39, 47 - 51.

Hotelling, H. Relations between two sets of variables. *Biometrika*, 1936, 28, 321 - 377.

Hubertv, C. J. On the variable selection problem in multiple group discriminant analysis. Paper presented at the 1971 AERA annual meeting. (mimeo.)

Kaiser, H. F. and Dickman, K. Sample and population score matrices from an arbitrary population correlation matrix. *Psychometrika*, 1962, 27, 179 - 182.

Roy, S. N. *Some aspects of multivariate analysis.* New York: Wiley, 1957.

Tatsuoka, M. M. *Multivariate analysis: techniques for educational and psychological research.* New York: Wiley, 1971.