ED 317 616 TM 014 735

AUTHOR	Eason, Sandra
TITLE	Backward Variable Elimination Canonical Correlation and Canonical Cross-Validation.
PUB DATE	19 Jan 90
NOTE	19p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, January 25-27, 1990).
AVAILABLE FROM	Sandra Eason, University of New Orleans, Lakefront, New Orleans, LA 70148.
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Educational Research; *Heuristics; *Multivariate Analysis; Research Methodology
IDENTIFIERS	*Cross Validation

ABSTRACT

This paper suggests that multivariate analysis techniques are very important in educational research, and that one multivariate technique--canonical correlation analysis--may be particularly useful. The logic of canonical analysis is explained. It is suggested that a backward variable elimination strategy can make the method even more powerful, by yielding more parsimonious results and greater power against Type II error. Backward variable elimination canonical correlation analysis uses communality coefficients to eliminate variables with relatively small explanatory power. It is also suggested that cross-validation procedures should be implemented to augment interpretation. These analyses are illustrated using a small heuristic data set. Five data tables are included. (Author/TJH)

stepwis3.wp 1/19/90

BACKWARD VARIABLE ELIMINATION CANONICAL CORRELATION AND CANONICAL CROSS-VALIDATION

Sandra Eason

University of New Orleans 70148

Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX, January 25, 1990.

ABSTRACT

The present paper suggests that multivariate techniques are very important in educational research, and that one multivariate technique, i.e., canonical correlation analysis, may be particularly useful. The logic of canonical analysis is explained. It is suggested that a backward variable elimination strategy can make the method even more powerful, by yielding more parsimonious results and greater power against Type II error. It is also suggested that cross-validation procedures should be implemented to augment interpretation. These analyses are illustrated using a small heuristic data set.

BACKWARD VARIABLE ELIMINATION CANONICAL CORRELATION AND CANONICAL CROSS-VALIDATION

The knowledge explosion combined with rapid technological advances in computer hardware and software have created both a need and the means for asking questions that require multivariate data analysis (Fish, 1988). Various statistical techniques available for analysis of multivariate data include multivariate analysis of variance (MANOVA), factor analysis, discriminant analysis, and canonical correlation analysis. These analyses are appropriate for use in studies involving more than one dependent variable because **only** multivariate statistical procedures **simultaneously** consider all the relationships among all the variables being investigated. Multivariate techniques may be employed for purposes of detecting treatment effects (MANOVA or discriminant analysis; cf. Maxwell, in press), data reduction (factor analysis), and group discrimination or classification of subjects (discriminant analysis; Huberty & Barton, 1989; Huberty & Wisenbaker, in press). However, perhaps the most powerful and potentially useful of the multivariate designs is canonical correlation analysis, a technique which analyzes complex relations of variable sets.

Canonical Correlation Analysis

Canonical correlation analysis is a statistical method that examines the relation between two or more sets of variables. The degree of relation, labeled the canonical correlation coefficient, is determined by a correlation between scores on linear predictor composite variables and scores on linear criterion composite variables. The discussion of how these composite scores are created is beyond the scope of the current presentation, but the process is described in readable detail by Eason, Daniel and Thompson, (1990) and by Thompson (1984, 1988).

The canonical score composites maximize the correlation between variable sets. Canonical correlation analysis is such a powerful analytical technique because the method considers all the relationships among all the variables and does not require that any variables be converted to the nominal level of scale. The premise that canonical analysis is powerful is supported in a study by Chastain and Joe (1987). In a canonical correlation analysis utilizing variables of intelligence, the researchers suggest that their relatively strong relations imply that earlier findings concerning intelligence have been "ignored or masked by previous univariate methods" (p. 323).

The logic of canonical correlation analysis is similar to the logic of the univariate technique of multiple regression. Cohen (1968) notes that multiple regression is a general analytic procedure encompassing all univariate procedures. Conversely, Knapp (1978) suggests that a similar relationship exists between canonical correlation analysis and virtually all univariate and multivariate tests of significance. In multiple regression the optimum linear combination of predictor variables to estimate a criterion variable is derived. Similarly, in canonical correlation analysis maximum linear relationships between sets of variables are isolated.

Canonical correlation analysis can be employed when at least two predictor variables and two criterion variables are present. The number of linear relationships obtained, i.e., canonical functions, is equal to the number of variables in the smaller of the two variable sets (Thompson, 1984). Canonical functions are derived by the extraction of principal components from a matrix, **A**, derived from the bivariate correlation matrix (**R**) involving all the variables in the analysis (Thompson, 1984, p. 13).

The first canonical function explains the largest amount of common variance across variable sets. Additional canonical functions are orthogonal to (i.e., perfectly uncorrelated with) previous functions, just as principal components are always orthogonal when they are first extracted. By analyzing predictor variables as a set and criterion variables as a set, the interrelationships shared by the variables are fully considered. Complex relations between the variable sets are accounted for in determining the variance explained. A statistical method that considers complex relations is important to social scientists, since reality is complex, and it is important to use statistical methods that mirror the reality about which the researcher wishes to generalize.

Within canonical correlation analysis several types of coefficients are computed which provide greater insight for the researcher in interpretation of results. Two of the more prominent coefficients are standardized function coefficients and canonical structure coefficients. Standardized function coefficients are analogous to beta weights in multiple regression, or to factor pattern coefficients in factor analysis. These are the weights used to create the canonical composite scores actually correlated in the analysis (Eason et al., 1990).

However, standardized function coefficients are affected by multicollinearity, and can provide incomplete information as to a variable's contribution to a given set of results (Thompson & Borrello, 1985). The computation of a second coefficient, i.e., canonical structure coefficients, is necessary to fully explore the explanatory power of a given variable. Canonical structure coefficients are correlations between observed variables and the synthetic or latent canonical score composites. By squaring a structure coefficient the amount of variance a variable shares with a function is indicated. Thus, the sum of the squared structure coefficients for a variable across each function yields the amount of total variance a variable contributes to the overall solution.

The variance a variable contributes to the overall solution is labeled a communality coefficient (h^2) and is important to the conceptualization of backward variable elimination canonical correlation analysis (Thompson, 1984), a method originating in the work of Rim (1972). In the past researchers too frequently concluded their analyses after interpreting the standardized function coefficients but prior to investigating the canonical structure coefficients (Thompson, 1988). Such a practice fails to take into consideration the amount of variance explained by the variables. When only function coefficients are consulted, a true picture of a variable's importance is unavailable (Thompson & Borrello, 1985).

The purpose of the present paper is to describe an extension of canonical correlation analysis which utilizes canonical communality coefficients. The technique is called backward variable elimination canonical correlation analysis (Thompson, 1982a). CANBAK (Thompson, 1982c), a computer program specifically designed to implement this canonical analysis, was utilized to provide the concrete heuristic example of the procedure.

Backward Variable Elimination Canonical Correlation Analysis

Backward variable elimination canonical correlation is a technique that sequentially deletes variables with low communality coefficients. A canonical communality coefficient is the sum of squared structure coefficients for a variable across all canonical functions. The communality coefficient represents the explanatory power of a variable after all possible linear relations are maximized. Since the variance explained by the linear combination of variables is of primary importance to analytical findings, variables contributing small amounts of variance to the results offer relatively unimportant contributions to solutions. Such variables can be eliminated in the interest of greater parsimony and to conserve degrees of freedom so that statistical power against Type II error is improved (Stevens, 1986; Thompson, 1982a).

The backward variable elimination procedure analyzes data in the following manner. First, a full model canonical correlation analysis involving all possible canonical functions is computed.

Communality coefficients are consulted and the variable with the lowest coefficient is deleted from the analysis, provided the communality coefficients are not homogeneous. A second canonical correlation analysis is conducted disregarding the discarded variable. Following the analysis, communality coefficients are once again consulted. The procedure of variable deletion is complete when the communality coefficients are reasonably homogeneous. Inference can be made that the variables remaining in the analysis explain most of the variance in the overall solution.

Heuristic Example of the Method

To illustrate the technique a hypothetical data set involving a university/school study will be presented. The study involved 49 subjects participating in a project called "At Risk Kids." The objectives of the project were to positively affect middle school adolescents in two primary areas. Eight predictor variables assessed pretreatment attitudes while the two criterion variables assessed locus of control and reading aptitude. The computer program, CANBAK, facilitates the variable elimination analysis by automatically consulting the communality coefficients, deleting variables when appropriate, and recomputing the canonical solution. An additional feature of CANBAK is the computation of several canonical coefficients usually obtained manually. CANBAK also computes variate adequacy, redundancy, pooled redundancy, and index coefficients.

The four quadrants of the bivariate correlation matrix (**R**) for these data are presented in Table 1. The quadruple product matrix (Thompson, 1984, p. 13) for the canonical analysis of these data was derived from the following equation:

$$\mathbf{R}_{2,2}^{-1} \times \mathbf{R}_{2,1} \times \mathbf{R}_{1,1}^{-1} \times \mathbf{R}_{1,2} = \mathbf{A}_{2X2}$$

From the 10 x 10 correlation matrix presented in Table 1, the analysis extracted two canonical functions. The extraction is consistent with the indication that the number of functions corresponds to the size of the smaller variable set, a function of the fact that this set determines the rank of matrix **A**, the matrix from which the functions are actually derived. The squared canonical correlation for the first function was .90 (chi square = 102.74, $\underline{df} = 16$, $\underline{p} < .05$); the squared canonical correlation coefficient associated with the second function was .11 (chi square = 4.9, $\underline{df} = 7$, $\underline{p} > .05$).

Following the canonical correlation analysis, the procedure of deleting a variable with the lowest communality coefficient (h^2) was initiated. For the present analysis only members of the larger variable set were considered for deletion; such a procedure is appropriate if the smaller set is a true criterion set and one does not wish to exclude any criterion variables from the analysis. Table 2 indicates that at the first step in the analysis the second predictor variable, Program, had the lowest communality coefficient (.015). Thus, the variable was eliminated from further analysis.

At the second step of the analysis, as reported in Table 3, two canonical functions were extracted from the reduced nine x nine correlation matrix. Deletion of one variable reduced the degrees of freedom for the first function from 16 to 14. In studies with larger variable sets, i.e., eight predictor variables and four criterion variables, four degrees of freedom would be conserved with deletion of one variable. Thus, the more variables there are in each set, through a multiplicative, the greater will be the conservation of degrees of freedom used in statistical significance testing. Thompson (1982b) suggests, "The conservation of degrees of freedom can be sizeable, and tends to reduce the likelihood of Type II errors occurring as a function of variable set sizes" (p. 4). The squared canonical correlation for the first function in the second step was .90 (chi square = 102.72, df = 14, p < .05); the squared canonical correlation associated with the second function was .11 (chi square = 4.92, df = 6, p > .05). The analysis indicated that the fifth predictor variable, High School Diploma, had the smallest canonical communality coefficient (.22) and was therefore deleted.

At the third step the correlation matrix was reduced to an eight x eight matrix. The squared canonical correlations for the functions were .90 (chi square 103.60, df = 12, p < .05) and .11 (chi square = 4.90. df = 5, p > .05), respectively. Table 4 indicates variable six, Self Esteem, had the smallest canonical communality coefficient (.25) and was deleted. The analysis was concluded as the remaining variables had relatively homogeneous communality coefficients. Thus, the variables Program, High School Diploma, and Self Esteem minimally contributed to the prediction of outcomes for these "At Risk Kids." Table 5 presents the final solution in the canonical correlation analysis.

In summary, backward variable elimination canonical correlation analysis yields parsimonious results thereby conserving degrees of freedom and reducing the likelihood of a Type II error. The procedure utilizes communality coefficients to eliminate variables with relatively small explanatory power. CANBAK, a computer program designed specifically for the analysis, facilitates the computational process, although the same analysis could be conducted with SPSS-X or SAS, albeit with somewhat more difficulty.

Cross-Validation

Findings of a study can yield large effect sizes and statistically significant coefficients but remain of little importance to researchers, if results are not replicable. For example, a prediction equation derived from one sample may not be accurate in a different sample. Invariance or cross-validation procedures can provide the researcher with an estimate of the stability of results across samples (Fish, 1986).

Cross-validation, recommended as an appropriate invariance procedure for canonical correlation analysis (Fish, 1986; Thompson, 1984), involves splitting a sample randomly into two subgroups (usually of unequal size) and performing separate canonical correlation analyses on each subgroup. In addition, new predictor and criterion composite scores for one group are derived from standardized function coefficients of the second group. Likewise, predictor and criterion composite scores for the second group are derived from standardized function coefficients of the first group. The new composite scores are correlated and compared for an invariance estimate.

The invariance estimates obtained from the "At Risk Kids" analysis indicated that the results were invariant and therefore replicable. The squared canonical correlation coefficient for group one was .89 as compared to the invariance check of a squared canonical correlation coefficient of .78. For group two the squared canonical correlation coefficient was .98 as compared to the invariance check of a squared canonical coefficient of .90. Only the replicability of the statistically significant canonical function was cross-validated here, for illustrative purposes. However, cross-validation procedures can be performed to establish the invariance of the other functions.

Summary

The present paper has suggested that multivariate techniques are very important in educational research, and that one multivariate technique, i.e., canonical correlation analysis, may be particularly useful. The logic of canonical analysis is explained. It is suggested that a backward variable elimination strategy can make the method even more powerful, by yielding more parsimonious results and greater power against Type II error. It is also suggested that cross-validation procedures should be implemented to augment interpretation. These analyses are illustrated using a small heuristic data set.

References

Chastain, R. L., & Joe, G. W. (1987). Multidimensional relations between intellectual abilities and demographic variables. Journal of Educational Psychology, 79(3), 323-325.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Eason, S., Daniel, L., & Thompson, B. (1990, January). A review of practice in a decade's worth of canonical correlation analysis studies. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX.

Fish, L. (1986, November). The importance of invariance procedures as against tests of statistical significance. Paper presented at the annual meeting of the Mid-South Educational Research Association, Memphis. (ERIC Document Reproduction Service No. ED 278 707)

Fish, L. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Huberty, C.J., & Barton, R.M. (1989). An introduction to discriminant analysis. Measurement and Evaluation in Counseling and Development, 22, 158-168.

Huberty, C.J., & Wisenbaker, J.M. (in press). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), Advances in social science methodology (Vol. 2). Greenwich, CT: JAI Press.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance-testing system. *Psychological Bulletin*, *25*, 410-416.

Maxwell, S.E. (in press). Recent developments in MANOVA applications. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2). Greenwich, CT: JAI Press.

Rim, E. (1972). A stepwise canonical approach to the selection of "kernel" variables from two sets of variables. (Doctoral dissertation, University of Illinois at Urbana, 1972). *Dissertation Abstracts International*, *34*, 623A. (University Microfilms No. 73-17,386)

Stevens, J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thompson, B. (1982a). A logic for stepwise canonical correlation analysis. *Perceptual and Motor Skills*, *54*, 879-882.

Thompson, B. (1982b, February). *Stepwise canonical correlation analysis: A new research technique*. Paper presented at the annual meeting of the Southwest Educational Research Association, Austin, TX. (ERIC Document Reproduction Service No. ED 222 546)

Thompson, B. (1982c). CANBAK: A program which performs stepwise canonical correlation analysis. *Educational and Psychological Measurement*, *42*, 849-851.

Thompson, B. (1984). *Canonical correlation analysis*. Newbury Park, CA: SAGE.

Thompson, B. (1988, April). *Canonical correlation analysis: An explanation with comments on correct practice*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 295 957)

Thompson, B., & Borrello, G. M. (1985). The importance of structure coefficients in regression research. *Educational and Psychological Measurement*, *45*, 203-209.

Table 1
Correlation Matrix

Camp	1.000	.162	.597	.556	.445	.395	.400	.241	-.167	.501
Program	.162	1.000	.117	.108	.339	.554	.206	.207	.003	.112
Teacher	.597	.117	1.000	.605	.665	.486	.451	.458	-.102	.538
School W	.556	.108	.605	1.000	.529	.512	.506	.502	-.029	.937
High Sch	.445	.339	.665	.529	1.000	.672	.680	.483	-.135	.469
HS Diplo	.395	.554	.486	.512	.672	1.000	.467	.387	-.045	.441
College	.400	.206	.451	.506	.680	.467	1.000	.593	-.244	.522
Self Est	.241	.207	.458	.502	.483	.387	.593	1.000	-.012	.471
					$R_{1,1}$				$R_{1,2}$	
					$R_{2,1}$				$R_{2,2}$	
Locus C	-.167	.003	-.102	-.029	-.135	-.045	-.244	-.012	1.000	-.142
Read Apt	.501	.112	.538	.937	.469	.441	.522	.471	-.142	1.000

Table 2
Canonical Solution for Step One

	Function I			Function II			
	F	S	SSQ	F	S	SSQ	h^2
Camp	-.057	.514	.264	-.428	-.453	.206	.470
Program	.071	.120	.014	.102	.022	.001	.015
Teacher	.004	.561	.315	-.166	-.253	.064	.379
School Work	1.060	.994	.987	.582	.011	.000	.987
High School	-.087	.484	.234	.067	-.360	.130	.364
HS Diploma	-.080	.464	.216	.106	-.089	.008	.223
College	.111	.528	.278	-1.138	-.687	.472	.750
Self Esteem	-.028	.500	.250	.483	.015	.000	.250
Locus Control	.108	-.035	.001	1.004	.999	.999	1.000
Read Aptitude	1.010	.994	.988	.035	-.107	.012	1.000

Note. F =canonical function coefficient; S = canonical structure coefficient; SSQ = squared canonical structure coefficient; h^2= canonical communality coefficient.

Table 3
Canonical Solution for Step Two

	Function I			Function II			
	F	S	SSQ	F	S	SSQ	h^2
Camp	-.047	.515	.265	-.416	-.452	.205	.470
Teacher	-.012	.562	.316	-.190	-.251	.063	.379
School Work	1.044	.995	.990	.564	.016	.000	.990
High School	-.078	.485	.235	.079	-.359	.129	.364
HS Diploma	-.035	.465	.216	.172	-.087	.008	.224
College	.105	.529	.280	-1.150	-.686	.471	.751
Self Esteem	-.018	.501	.251	.500	.018	.000	.251
Locus Control	.107	-.037	.001	1.005	.999	.999	1.000
Read Aptitude	1.010	.994	.989	.037	-.106	.011	1.000

Note. F = canonical function coefficient; S = canonical structure coefficient; SSQ = squared canonical structure coefficient; h^2= canonical communality coefficient.

Table 4
Canonical Solution for Step Three

	Function I			Function II			
	F	S	SSQ	F	S	SSQ	h^2
Camp	-.050	.515	.265	-.407	-.458	.209	.475
Teacher	-.010	.562	.316	-.205	-.255	.065	.381
School Work	1.037	.995	.991	.601	.013	.000	.991
High School	-.099	.485	.235	.185	-.363	.132	.367
College	.107	.529	.280	-1.172	-.693	.481	.760
Self Esteem	-.019	.501	.251	.512	.016	.000	.251
Locus Control	.108	-.036	.001	1.004	.999	.999	1.000
Read Aptitude	1.010	.994	.989	.036	-.107	.011	1.000

Note. F = canonical function coefficient; S = canonical structure coefficient; SSQ = squared canonical structure coefficient; h^2= canonical communality coefficient.

Table 5
Canonical Solution for Final Step

	Function I			Function II			
	F	S	SSQ	F	S	SSQ	h^2
Camp	-.046	.515	.265	-.561	-.498	.248	.513
Teacher	-.015	.562	.316	-.085	-.279	.078	.394
School Work	1.032	.996	.991	.787	.007	.000	.991
High School	-.098	.484	.235	.182	-.396	.157	.392
College	.098	.528	.279	-1.013	-.753	.568	.847
Locus Control	.110	-.034	.001	1.004	.999	.999	1.000
Read Aptitude	1.010	.994	.988	.034	-.109	.012	1.000

Note. F = canonical function coefficient; S = canonical structure coefficient; SSQ = squared canonical structure coefficient; h^2= canonical communality coefficient.