Canonical Correlation Analysis in a Predictive System

Author(s): Randall B. Dunham and Dennis J. Kravetz

Source: *The Journal of Experimental Education*, Summer, 1975, Vol. 43, No. 4 (Summer, 1975), pp. 35-42

Published by: Taylor & Francis, Ltd.

Stable URL: https://www.jstor.org/stable/20151046

CANONICAL CORRELATION ANALYSIS IN A PREDICTIVE SYSTEM

RANDALL B. DUNHAM and DENNIS J. KRAVETZ

ABSTRACT

This discussion focuses on the use and interpretation of canonical correlation analysis in a predictive system using a numerical example. The concepts examined include: the canonical solution; canonical weights; significance tests; correlations of variables with variates; predictive effectiveness; and utility of individual predictor variables. The emphasis is on interpretation of the canonical analysis.

CANONICAL CORRELATION may be used in a predictive system when two or more predictor variables and two or more criteria variables are being considered. Canonical analysis determines a linear combination for each set of variables such that the two sets are maximally related to one another (5). In the present paper, one set of variables will be considered *predictor variables* and the other set *criteria variables*. More than one pair of linear combinations may be formed in canonical analysis. The maximum number of linear combinations which may be obtained is equal to the number of variables in the smaller set. The successive predictor linear combinations (variates) are each orthogonal (uncorrelated) to one another and the criteria variates also are orthogonal to one another. This paper does not intend to study the mathematical theory underlying canonical correlation analysis. Its purpose is to provide a guide for the actual use of this analytic tool in a predictive model. The computational procedures involved are presented by Cooley and Lohnes (3) and by Tatsuoka (9) and should be pursued independently by the reader.

The following numerical example will be used to describe the use of the canonical tool. The example uses data contrived for the prediction of a set of four criteria variables by a set of five predictor variables. Each of the criteria variables are measures of academic performance in introductory college courses: psychology, sociology, philosophy, and creative writing. The set of predictor variables consists of SAT Verbal, SAT Quantitative, high school junior GPA, high school senior GPA, and the number of extracurricular activities participated in. The use of canonical correlation analysis allows the simultaneous treatment of all four of the criteria variables. The alternative statistical treatment would involve a set of independent univariate analyses predicting each of the criteria variables one at a time. As will be demonstrated in this paper, the canonical analysis provides all of the information available from univariate analyses plus a great deal more. The intercorrelations of these standardized variables are presented in Table 1.

PROCEDURE

The first step in the use of the canonical analysis is to solve for the sets of linear combinations. This procedure is analogous to the simultaneous principal component analysis for each of variables, with rotation then being performed to maximally relate the two sets. The canonical weights and canonical correlations for the authors' sets of variables are presented in Table 2.

Table 1.–Intercorrelations of All Variables

	P1	P2	P3	P4	P5	C1	C2	C3	C4
P1 SAT Verbal	1.00								
P2 SAT Quantitative	60	1.00							
P3 Junior GPA	30	35	1.00						
P4 Senior GPA	40	40	60	1.00					
P5 Extracurricular	20	15	25	25	1.00				
C1 Psychology	30	40	20	30	15	1.00			
C2 Sociology	30	35	20	30	10	70	1.00		
C3 Philosophy	35	35	25	35	05	60	50	1.00	
C4 Writing	40	25	30	40	10	30	35	30	1.00

Table 2.–Canonical Solution

Predictor Variables	Standardized Predictor Weights PV1	PV2	PV3	PV4	Criteria Variables	Standardized Criteria Weights CV1	CV2	CV3	CV4
SAT Verbal	.50	-.70	-.19	.23	Psychology	.12	1.03	-.86	-.73
SAT Quantitative	.17	1.19	.44	-.15	Sociology	.14	.11	.45	1.36
Junior GPA	.10	-.35	.03	-1.22	Philosophy	.43	-.31	1.04	-.52
Senior GPA	.52	-.16	.03	.79	Writing	.60	-.64	-.63	-.02
Extracurricular	-.08	.39	-.97	-.01	R_{c_j} =	.57	.29	.10	.01
					$\lambda_j = R_{c_j}^2$ =	.33	.08	.01	.00

The canonical correlations express the degree of relationship between successive predictor-criteria pairs of linear combinations. Bartlett (1, 2) has given a method for obtaining a chi-square estimate for testing the significance of each canonical function. The appropriate test statistic for each canonical function is summarized in Table 3. Note that each eigenvalue (λ_j) is equal to the corresponding canonical correlation squared (R_{cj}^2). In the present example, two canonical functions are significant at the .005 level of confidence. The squares (R_{cj}^2) of these values represent the percent of overlapping variance from one set of variates to the other. The percent of shared variance for the respective pairs of predictor and criteria linear combinations in the present example are $R_{c_1}^2$ = 33 percent, $R_{c_2}^2$ = 8 percent, $R_{c_3}^2$ = 1 percent, and $R_{c_4}^2$ = 0 percent. This is *not* an indication of the percent of variance accounted for in the original set of variables.

Table 3.–Significance Tests

Significance Test Of:	χ^2	Degrees of Freedom	p Level
Variate 1	$\chi_1^2 = 293.4$	$(M_c)(M_p) = (5)(4) = 20$	$p < .005$
Variate 2	$\chi_2^2 = 51.06$	$(M_c-1)(M_p-1) = (4)(3) = 12$	$p < .005$
Variate 3	$\chi_3^2 = 5.94$	$(M_c-2)(M_p-2) = (3)(2) = 6$	$p < .50$
Variate 4	$\chi_4^2 = 0.00$	$(M_c-3)(M_p-3) = (2)(1) = 2$	$p \leq 1.0$

M_c = number of criteria variables

M_p = number of predictor variables

$\Lambda_i = \prod_i^{M_c} (1-\lambda_i)$

$\chi_i^2 = -[N - 1 - .5\ (M_p + M_c + 1)]\ \ln \Lambda_i$

Table 4.–Structure Matrix

	PVI	PVII	PVIII	PVIV	CVI	CVII	CVIII	CVIV
SAT Verbal	.82	-.07	-.10	.08	.47	-.02	-.01	.00
SAT Quantitative	.70	.65	.20	-.13	.40	.18	.02	.00
Junior GPA	.60	-.14	-.10	-.73	.34	-.04	-.01	-.01
Senior GPA	.82	-.07	-.10	.08	.47	-.02	-.01	.00
Extracurricular	.20	.30	-.93	-.09	.11	.09	-.09	.00
Psychology	.38	.21	.01	.00	.66	.73	-.11	-.09
Sociology	.38	.13	-.02	.01	.66	.46	.15	.58
Philosophy	.44	.05	-.06	.00	.76	.18	.56	-.28
Writing	.47	-.11	.04	.00	.82	-.38	-.42	.08

The nature of the relationships. In a method commonly used for interpreting a canonical relationship, the standardized canonical weights are examined (see Table 2). These weights are analogous to beta weights in multiple linear regression. Darlington (4) has reviewed the dangers of interpreting these weights because of their instability which is a function of the predictor variables intercorrelations. These problems are relevant in the present case where the variables within each set are intercorrelated.

A method for interpreting canonical relationships which is not severely affected by the intercorrelations of the variables in a set was proposed by Meredith (7). In this method, the correlations between the original variables and the linear combinations of these variables are calculated and examined. These correlations, which comprise the "structure matrix," are shown in Table 4. Examination of the columns in the upper left quadrant of the matrix helps to reveal which of the predictor variables are most highly related to each of the predictor variates. The columns in the lower right quadrant help to reveal which of the criteria variables are most highly related to each of the criteria variates. The first criterion variate, which is composed of creative writing, philosophy, psychology, and sociology (in that order), was predicted by the first predictor variate, composed primarily of SAT Verbal, senior GPA, SAT Quantitative, and junior GPA (in that order). The second criterion variate, which is composed primarily of psychology and, to a lesser degree, sociology and writing, was predicted by the second predictor variate, which is largely made up of SAT Quantitative and extracurricular activities. Examination of the other two quadrants of the structure matrix yields parallel information.

Structure matrices such as the one in Table 4 are helpful for interpretation purposes but can also be misleading. The relationship of a given variable with a canonical variate is not proportional to the correlation of the variable with that variate. Rather, the relationship is proportional to the square of the correlation between the variable and the variate. By squaring the coefficient, we can tell the percent of variance which may be accounted for by each variable. A matrix composed of such squared correlations, the *variance matrix*, is shown in Table 5. Use of this matrix is similar to that of the structure matrix. For example, the values in the lower left quadrant of the variance matrix equal the percent of variance in each criteria variable which may be accounted for by each predictor variate. In each quadrant of the variance matrix, the values which represent the proportion of variance accounted for by each variate are represented by the column sums divided by the number of rows. It is suggested that the variance matrix be used to facilitate the interpretation of canonical correlation analysis. This matrix will give the relative importance of each variable, and the structure matrix will give the directionality of the relationship. For example, in the first variate, the upper right quadrant of the variance matrix demonstrates that SAT Verbal ($L_{jk}^2 =$ .22) and senior GPA ($L_{jk}^2 = .22$) were the two most powerful predictor variables, while the extracurricular activities variable was least effective ($L_{jk}^2 = .01$). The structure matrix indicates positive directionality for all variables on the first variate.

The importance of interpreting canonical variates on the basis of the variance matrix rather than the weight matrix is easy to understand. When the variables in a set are intercorrelated, canonical weights will not express the overall relationship of a given variable with the linear combination. This is because variance which is shared by any of the variables in a set will only contribute to enlarging the weight of one variable. In our example, SAT Verbal and SAT Quantitative are correlated .60 with one another as are junior and senior GPA. On the basis of the predictor weights in Table 2, the first predictor variate would be interpreted as being composed of primarily SAT Verbal and senior GPA. Examination of the variance matrix in Table 5, however, shows that, in addition to SAT Verbal and senior GPA, SAT Quantitative and junior GPA also have sizable relationships with the first predictor variate.

Predictive effectiveness. Stewart and Love (8) provide a measure of "redundancy" which can be used to answer the question of predictive effectiveness and several related questions. Cooley and Lohnes (3) point out that only the squared structure loadings (variance matrix elements) are needed for computing this redundancy index. Following the notation used by Stewart and Love, the elements in the variance matrix are designated L_{jk}^2, the number of predictor variables M_p, and the number of criteria variables M_c. Looking at the upper left quadrant of the variance matrix, the proportion of the variance in the predictor set of variables extracted by each predictor variate can be computed by summing the column values and dividing by the number of variables

$$\left(\sum_{j=1}^{M_p} L_{jk}^2 \,/\, M_p\right).$$

In our example, the percent of variance in the predictor set of variables extracted by the predictor variates is shown in Table 5 to be equal to 45 percent for the first predictor variate and 11 percent, 19 percent, and 12 percent for the remaining three variates. The total percent of variance of

Table 5.—Variance Matrix

	Predictor Variate #1	PV #2	PV #3	PV #4	Unique Variance in Criteria	CV #1	CV #2	CV #3	CV #4	$R^2_{p.c}$
SAT Verbal	.68	.01	.01	.01	.027	.22	.00	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .22$
SAT Quantitative	.49	.42	.04	.02	.032	.16	.03	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .19$
Junior GPA	.36	.02	.01	.54	.006	.12	.00	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .12$
Senior GPA	.68	.01	.01	.01	.030	.22	.00	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .22$
Extracurricular	.04	.10	.86	.01	.007	.01	.01	.01	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .03$
$\sum_{j=1}^{M_p} L_{jk}^2 =$	2.25	.56	.93	.59		.73	.04	.01	.00	
$\sum_{j=1}^{M_p} L_{jk}^2 / M_p =$	.45	.11	.19	.12		.15	.01	.00	.00	

$$\overline{R^2_{p.p}} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_p} L_{jk}^2 / M_p = .87 \qquad \overline{R^2_{p.c}} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_p} L_{jk}^2 / M_p = .16$$

	PV #1	PV #2	PV #3	PV #4	$R^2_{c.p}$	CV #1	CV #2	CV #3	CV #4
Psychology	.15	.04	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .19$	.44	.54	.01	.01
Sociology	.14	.02	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .16$	.43	.21	.02	.34
Philosophy	.19	.00	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .19$	.58	.03	.31	.08
Writing	.22	.01	.00	.00	$\sum_{k=1}^{M_c} L_{jk}^2 = .23$	.67	.15	.17	.01
$\sum_{j=1}^{M_c} L_{jk}^2 =$	.70	.07	.00	.00		2.12	.93	.51	.44
$\sum_{j=1}^{M_c} L_{jk}^2 / M_c =$	.18	.02	.00	.00		.53	.23	.13	.11

$$\overline{R^2_{c.p}} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_c} L_{jk}^2 / M_c = .20 \qquad \overline{R^2_{c.c}} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_c} L_{jk}^2 / M_c = 1.00$$

the set of predictor variables extracted by the four predictor variates ($\overline{R_{p.p}^2}$) is equal to the sum of these four column sums–

$$(\overline{R_{p.p}^2} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_c} L_{jk}^2 \; / \; M_p)$$

–since the successive canonical variates are orthogonal. In the present example, the total percent of variance in the set of predictor variables extracted by the four predictor variates is equal to 87 percent.

Similar computations in the lower right quadrant of the variance matrix will yield information on the percent of variance in the set of criteria variables extracted by the criteria variates. The amounts computed from ($\sum_{j=1}^{M_c} L_{jk}^2 \; / \; M_c$) are equal to 53 percent, 23 percent, 13 percent, and 11 percent, respectively, for a total of 100 percent. (Whenever the number of components derived is equal to the number of variables, the total variance extracted will be equal to 100 percent.)

The most important information is contained in the other two quadrants of the variance matrix. In the lower left quadrant, the percent of variance in the set of criteria variables extracted by each of the predictor variates is equal to the column sums of that quadrant of the variance matrix divided by the number of criterion variables ($\sum_{j=1}^{M_c} = L_{jk}^2 \; / \; M_c$). This is the redundancy coefficient for each variate. In the present example, the column sums divided by M_c are equal to 18 percent, 2 percent, 0 percent, and 0 percent. The total redundancy, or total variance in the criteria set of variables extracted by the set of predictor variates, is equal to the sum of these four column sums ($\overline{R_{c.p}^2} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_c} L_{jk}^2 \; / \; M_c$). In the present example, the total redundancy is equal to 20 percent. This is the most critical value for evaluating the effectiveness of a predictor system. It tells us the total amount of variance extracted by the set of predictor variables–variance accounted for given the predictor set. Redundancy coefficients do not necessarily increase with the addition of another criterion variable to the criterion set. As pointed out by Stewart and Love (8), the redundancy measure is equal to the "mean squared multiple correlation" of each criterion variable given the set of predictor variables. Each of the row sums in the lower left quadrant of the variance matrix

$$(\sum_{k=1}^{M_c} L_{jk}^2)$$

can then be regarded as a multiple correlation squared. Consequently, the addition of a criterion variable with a squared multiple correlation smaller than the existing redundancy coefficient will reduce the new redundancy coefficient, while the addition of a criterion variable with a squared multiple correlation larger than the redundancy coefficient will increase the new redundancy coefficient. Similarly, the removal of a criterion variable with a squared multiple correlation smaller than the existing redundancy coefficient will increase the new redundancy coefficient, while the removal of a criterion variable with a squared multiple correlation larger than the existing redundancy coefficient will decrease the new redundancy coefficient. A more complete discussion of this issue may be found in a paper by Kravetz and Dunham (6). The redundancy in the predictor variables given the criteria variables may be found by referring to the upper right quadrant of the variance matrix. The computation

$$(\overline{R_{p.c}^2} = \sum_{k=1}^{M_c} \sum_{j=1}^{M_p} L_{jk}^2 \,/\, M_p)$$

yields a figure of 16 percent in the present example.

This value is not of primary interest in a purely predictive system. However, as indicated previously, the individual values of L_{jk}^2 in this quadrant may be used as indicators of the predictive effectiveness of individual predictor variables.

Which variable should be used in the predictive system? It may be desirable to eliminate certain predictor variables due to the inefficiency of those variables in the predictive system. Darlington (4) has discussed methods for determining the amount of unique variance accounted for by each predictor variable in multiple regression. A simple extension of these techniques to canonical analysis is made by Kravetz and Dunham (6). If the structure matrix in Table 4 consisted of part correlations (controlling for all other variables in the same set), the amount of unique variance in each of the variates accounted for by each variable would be equal to the square of that part correlation. When a given predictor variable is removed from the predictive system, the resultant drop in predictive efficiency ($R_{c.p}^2$) is an indication of the unique variance in the criteria variable set accounted for by that variable. Small drops will indicate that the variable eliminated added little to the predictive power of the system. Parallel analysis could be performed for the criteria variable relationships with the predictor variance accounted for. This could be helpful when choosing a criteria set in some situations. The drop in criteria variable redundancy (given the predictor set of variables) which results from dropping each of the predictor variables one at a time is shown in Table 5. The unique variance in the criteria set of variables accounted for by each of the predictor variables is 2.7 percent for SAT Verbal, 3.2 percent for SAT Quantitative, 0.6 percent for junior GPA, 3.0 percent for senior GPA, and 0.7 percent for extracurricular activities. On the basis of this information, decisions can be made for inclusion or exclusion of particular predictor variables. In our example, the decision might be made to not use junior GPA or extracurricular activities since the criteria variable redundancy (given the predictor set of variables) would drop only 0.6 percent if junior GPA were not used or 0.7 percent if extracurricular activities were not used. In practice, as in any predictive system, the cost of each predictor must be weighed against its predictive utility.

GENERAL COMMENTS

It is important that the user of canonical analysis read a presentation of the technique such as that by Cooley and Lohnes (3) since there are a number of standard restrictions and procedures which are not discussed in this paper. Cooley and Lohnes present a useful FORTRAN computer program for canonical correlation analysis which calculates the structure matrix, redundancy measures, and significance tests. Additionally, the paper by Kravetz and Dunham (6) deals more completely with the effects of various parameters (sample size, number of variables in each set, reliability of variables, etc.) on the canonical solutions and on the methods discussed in this paper.

Generally, it is suggested that only significant canonical variates be included for evaluation and interpretation. The techniques discussed in this paper may be used with only significant relationships by setting values for nonsignificant relationships in the weight matrix, structure matrix, and variance matrix equal to zero before beginning interpretation. As in the present case, nonsignificant variates seldom account for large portions of variance if retained. In any event, nonsignificant variates should never be interpreted since they are known to have a high probability of being due to chance alone.

Note that, in the case where there are more criteria variables than predictor variables, the same techniques may be applied as long as the distinction between the two sets is kept clear. No critical assumptions would be violated.

REFERENCES

1. Bartlett, M.S., "The Statistical Significance of Canonical Correlations," *Biometrika*, 32:29-38, 1941.
2. Bartlett, M.S., "Multivariate Analysis," Supplement to *The Journal of the Royal Statistical Society*, 9:176-197, 1947.
3. Cooley, W.W. and Lohnes, P.R., *Multivariate Data Analysis*, Wiley, New York, 1971.
4. Darlington, R.B., "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69: (No. 3):161-182. 1968.
5. Hotelling, H., "The Most Predictable Criterion," *Journal of Educational Psychology*: 26:139-142, 1935.
6. Kravetz, D.J. and Dunham, R.B., "Canonical Analysis: Some Important Parameters," University of Illinois, Champaign, 1974 (mimeo).
7. Meredith, W., "Canonical Correlation with Fallible Data," *Psychometrika*, 29:55-65, 1964.
8. Stewart, D.K. and Love, W.A., "A General Canonical Correlation Index," *Psychological Bulletin*, 70:160-163, 1968.
9. Tatsuoka, M.M., *Multivariate Analysis*, Wiley, New York, 1971.