Psychological Bulletin
1978, Vol. 85, No. 2, 410-416

Canonical Correlation Analysis: A General Parametric Significance-Testing System

Thomas R. Knapp
College of Education
University of Rochester

Significance tests for nine of the most common statistical procedures (simple correlation, t test for independent samples, multiple regression analysis, one-way analysis of variance, factorial analysis of variance, analysis of covariance, t test for correlated samples, discriminant analysis, and chi-square test of independence) can all be treated as special cases of the test of the null hypothesis in canonical correlation analysis for two sets of variables.

Ten years ago Cohen (1968) explained in language familiar to social scientists that the analysis of variance is a special case of multiple regression analysis and that both are subsumed under what the mathematical statisticians call the *general linear model*. The key concepts that link these two otherwise quite different techniques are the notion of a *dummy variable* (one for each of the "between" degrees of freedom) and the fact that differences between means and correlations between variables are analogous methodological concepts (a large difference between two sample means on a single variable conveys essentially the same information as a high correlation between that variable and the dummy variable of sample membership). Some authors of recent statistics textbooks for the social sciences, for example, Kerlinger and Pedhazur (1973), Cohen and Cohen (1975), and Roscoe (1975), have devoted full chapters or sections within chapters to similar explanations of this equivalence.

The purpose of the present article is to extend Cohen's arguments even further by showing how virtually all of the commonly encountered parametric tests of significance can be treated as special cases of canonical correlation analysis, which is the general procedure for investigating the relationships between two sets of variables. Much of what follows has been compiled from various scattered sources. Originality is claimed only for the sections on correlated-sample t tests and chi-square tests. Consideration of chi-square tests for contingency tables as a type of canonical correlation analysis was mentioned by Darlington, Weinberg, and Walberg (1973) but was not further elaborated upon in that article.

The following section contains a formulation of the basic canonical problem in matrix notation (familiarity with matrix algebra, including a knowledge of eigenvalues and eigenvectors, is assumed)[1] and a specification of the associated F test of the significance of the correlation between linear composites of two sets of variables for sample data. The next sections describe how canonical correlation analysis can be used for simple (two-variable) correlation, the t test for independent sample means, multiple regression analysis, one-way analysis of variance, factorial analysis of variance, analysis of covariance, the t test for correlated sample means, discriminant

This article is concerned solely with traditional hypothesis testing, which continues to dominate empirical methodology (rightly or wrongly). No attempt is made to compare the relative merits of tests of significance and other modes of inference.

Requests for reprints should be sent to Thomas R. Knapp, College of Education, University of Rochester, Rochester, New York 14627.

[1] For those who are not familiar with eigenvalues and eigenvectors, especially the former, there are excellent sections on these topics in Cooley and Lohnes (1971), Tatsuoka (1971), and Press (1972).

analysis, and the chi-square test of independence of two variables. Some of these techniques are illustrated with numerical examples, using the Project Talent Test Battery data contained in the appendix of the multivariate text by Cooley and Lohnes (1971). The final section consists of a brief summary and a few remarks regarding assumptions.

The General Problem

As stated in a number of standard textbooks in multivariate analysis, for example, Anderson (1958), Morrison (1976), Cooley and Lohnes (1971), and Tatsuoka (1971), if there is one set of p variables and another set of q variables (where q is usually taken to be less than or equal to p), the principal objective of canonical correlation analysis is to find a linear combination of the p variables that correlates maximally with a linear combination of the q variables and, for sample data, to test the statistical significance of that correlation. The weights for the q variables in the second set are obtained by finding the elements of the eigenvector $\mathbf{v}_1$ associated with the largest eigenvalue λ_1 of the matrix $M = R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$, where R_{YY}^{-1} is the inverse of the $q \times q$ matrix of intercorrelations among the q variables, R_{YX} is the $q \times p$ matrix of cross-correlations between the variables of the two sets, R_{XX}^{-1} is the inverse of the $p \times p$ matrix of intercorrelations among the p variables, and R_{XY} is the $p \times q$ transpose of R_{YX}. The weights for the p variables in the first set are obtained by finding the elements of the vector

$$\mathbf{v}_2 = \lambda_1^{-\frac{1}{2}}R_{XX}^{-1}R_{XY}\mathbf{v}_1.$$

The maximal canonical correlation r_c is the square root of λ_1. Its significance is tested by referring to a table of the F sampling distribution the following statistic for pq and $ms - pq/2 + 1$ degrees of freedom (the latter need not be an integer):

$$F = (1 - \Lambda^{1/s})/pq \div (\Lambda^{1/s})/(ms - pq/2 + 1),$$

where

$\Lambda = \prod_{i=1}^{q} (1 - \lambda_i)$, that is, the product of all the $1 - \lambda_i$s for $i = 1, 2, \ldots, q$ (Wilks, 1932),

$\lambda_i = i$th eigenvalue of M for $i = 1, 2, \ldots, q$,

$m = N - 3/2 - (p + q)/2$, where N is the sample size,

$s = [(p^2q^2 - 4) \div (p^2 + q^2 - 5)]^{\frac{1}{2}}$.

The test (Rao, 1952) is exact if either p or q is less than or equal to two and is approximate otherwise.

Table 1
Example of Standard Canonical Correlation Analysis

Data						
ID	X_1	X_2	X_3	Y_1	Y_2	Y_3
1	10	20	18	9	12	9
2	4	15	13	7	10	10
⋮	⋮	⋮	⋮	⋮	⋮	⋮
504	8	12	37	5	6	8
505	1	11	27	14	11	11

$p = q = 3$
$pq = 9$
$m = 500.5$
$s = (77/13)^{\frac{1}{2}} = 2.4337$
$ms - pq/2 + 1 = 1214.5668$[a]

Results

$\lambda_i = .3509, .0150, .0003$
$(r_c = \sqrt{.3509} = .59)$
$\Lambda = .6393$
$F(9, 1215) = 27.24$

Note. Data taken from Cooley and Lohnes (1971). Variables are from the Project Talent Test Battery: X_1 = Variable 17 (Sociability Inventory), X_2 = Variable 18 (Physical Science Interest Inventory), X_3 = Variable 19 (Office Work Interest Inventory), Y_1 = Variable 13 (Creativity Test), Y_2 = Variable 14 (Mechanical Reasoning Test), Y_3 = Variable 15 (Abstract Reasoning Test); ID = identification number.

[a] Rounded to 1215 for use with the F sampling distribution.

Table 1 contains a summary of the results of a typical canonical correlation analysis, with $p = q = 3$. The variables in one set are Variables 17 (Sociability Inventory), 18 (Physical Science Interest Inventory), and 19 (Office Work Interest Inventory) of the Project Talent Test Battery (Cooley & Lohnes, 1971). The variables in the other set are Variables 13 (Creativity Test), 14 (Mechanical Reasoning Test), and 15 (Abstract Reasoning Test) of the same battery.

Simple Correlation

For $p = 1$ and $q = 1$ (i.e., one variable in each set), $R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$ reduces to $R_{YX}R_{XY} = r^2$, where r is the correlation between the two variables. The largest eigenvalue of the scalar r^2 is r^2 itself.

The formula for F reduces to

$$(r^2/1) \div (1 - r^2)/(N - 2),$$

since $p = q = 1$, $s = (-3/-3)^{\frac{1}{2}} = \sqrt{1} = 1$, $\Lambda^{1/s} = \Lambda = 1 - \lambda_1 = 1 - r^2$, $1 - \Lambda^{1/s} = r^2$, and $m = N - 5/2$. This is, of course, the square of the well-known t ratio for testing the significance of a correlation coefficient. Since $F(1, N - 2) = t^2(N - 2)$, all is well.

t Test for Independent Sample Means

Several authors, for example, McNemar (1969), Welkowitz, Ewen, and Cohen (1976), and Roscoe (1975), have pointed out the equivalence of the test of the significance of the difference between two independent sample means and the test of the significance of a point-biserial correlation coefficient. The trick is to create a dichotomous dummy variable for which a score of one is indicative of membership in one of the samples and a score of zero is indicative of membership in the other sample (and therefore nonmembership in the first sample). The same trick is used when canonical correlation analysis is applied to an independent samples test on the means. The formula for F is identical to that for simple correlation in the previous section. One variable is the dummy dichotomy of group membership, and the other variable is the continuous criterion variable.

Multiple Regression Analysis

The major difference between multiple regression analysis and canonical analysis is that the former employs just one variable in the second set, that is, $q = 1$. Therefore, $R_{YY}^{-1}R_{YX}R_{XX}^{-1}R_{XY}$ reduces to $R_{YX}R_{XX}^{-1}R_{XY}$, which is recognizable as $R_{YX}\mathbf{b}$, where $\mathbf{b}$ is the column vector of beta weights (standardized partial regression coefficients), and which in turn reduces to the scalar $r_{Y \cdot X_1, X_2, \ldots, X_p}^2$ (since R_{YX} is a row vector of the correlations of each of the variables in the first set with the variable in the second set), that is, the square of the multiple correlation coefficient. The largest eigenvalue of r^2 is again r^2 itself. The formula for F reduces to $(r^2/p) \div (1 - r^2)/(N - p - 1)$, the traditional formula for testing the significance of a multiple r, since $q = 1$, $s = [(p^2 - 4) \div (p^2 - 4)]^{\frac{1}{2}} = 1$, $\Lambda^{1/s} = \Lambda = 1 - \lambda_1 = 1 - r^2$, $1 - \Lambda^{1/s} = r^2$, and $m = N - 3/2 - (p + 1)/2$.

The case of $p = 2$ and $q = 1$ presents a special difficulty, since $p^2 + q^2 - 5 = 0$, and s is undefined. (F is still the same multiple regression F, however.) There is a test of the significance of a canonical correlation coefficient, due to Bartlett (1941), that is not subject to this constraint. One calculates $\chi^2 = -[N - \frac{1}{2}(p + q + 1)] \log_e \Lambda$ and refers that value to a table of the chi-square sampling distribution for pq degrees of freedom.

One-Way Analysis of Variance

The equivalence of a one-way analysis of variance for k independent samples to a multiple regression analysis with $p = k - 1$ is well documented in Cohen's (1968) article, so a very brief treatment should suffice here. Membership in Sample 1 is denoted by variable X_1 (which equals one if an observation is in Sample 1 and equals zero otherwise), membership in Sample 2 is denoted by variable $X_2, \ldots$, membership in Sample $k - 1$ is denoted by X_{k-1}. Members of Sample k receive scores of zero on variables X_1 through X_{k-1}, so that a kth variable is not only unnecessary but would produce a linear dependency that would render R_{XX} noninvertible. The F ratio is exactly the same as the F ratio for multiple regression analysis, with the numerator equal to the mean square between and with the denominator equal to the mean square within. (There is actually a total sum of squares term in the numerator and in the denominator, but they cancel each other.)

Factorial Analysis of Variance

Cohen (1968) and others have shown how the equivalence of the analysis of variance and multiple regression analysis also extends to factorial analysis of variance and to the

analysis of covariance (see next section). For the simplest case of two independent variables, one with a categories and the other with b categories, one needs $a - 1$ dummy variables to represent group membership for the first variable, $b - 1$ dummy variables to represent group membership for the second variable, and $(a-1)(b-1) = ab-a-b+1$ dummy variables to represent group membership for their interaction. This produces a total of $p = (a - 1) + (b-1) + (ab-a-b+1) = ab-1$ new independent variables. Membership in the interaction groups can be defined by the following ordinary multiplication rules (1 and 0 are the codes for the main effect groups): $1 \times 1 = 1, 1 \times 0 = 0, 0 \times 1 = 0$, and $0 \times 0 = 0$. Three canonical analyses are required (one for each main effect and one for the effect of the entire set of variables; the magnitude of the interaction effect is determined by subtraction). Three F ratios are generated. The dummy dichotomies for the main effect of one independent variable are uncorrelated with the dummy dichotomies for the main effect of the other independent variable, but they are correlated with the dummy dichotomies for their interaction, even if the cell frequencies are proportional. There are other coding schemes, for example, orthogonal coding and effect coding, for which all main effect and interaction variables are uncorrelated, but all schemes produce the same desired F ratios (see Kerlinger & Pedhazur, 1973).

This procedure can be extended to take care of factorial designs for any number of independent variables, with the interaction coding determined in a similar manner, that is, by multiplying the main effect codes. For example, for four independent variables, the score on the first second-order interaction variable for an observation with main effect codes of 0, 1, and 0 would be $0 \times 1 \times 0 = 0$.

Analysis of Covariance

Using canonical correlation analysis (or multiple regression analysis, for that matter) to do an analysis of covariance is cumbersome though straightforward, so only the simplest case of one principal independent variable of k categories, one continuous dependent variable, and one continuous covariable is mentioned here.

It helps to start by stating the general research question that the analysis of covariance seeks to answer, namely, What is the effect (in the most liberal sense of the word) of the principal independent variable on the dependent variable that is over and above the effect of the covariable itself? This suggests (requires) that two analyses be performed: (a) the so-called *full-model* test that includes the principal independent variable and the covariable and (b) the *reduced-model* test that includes the covariable alone. An r^2 is obtained for each test, and a subsequent F ratio that is a function of the difference between the two r^2s is determined. Cohen's (1968) article provides the necessary details.

t Test for Correlated Sample Means

A canonical analysis of the difference between two correlated sample means also requires full-model and reduced-model considerations. Each member of each of the two samples gets a score on the principal treatment variable (1 if in Sample 1, 0 if in Sample 2) and a score on each of $N - 1$ pairing variables (1 if a member of the given pair, 0 if not), where N is the number of pairs. An r^2 is obtained for the full model, which includes all of these variables, and another r^2 is obtained for the reduced model, which includes only the pairing variables. The F test of the difference between the two r^2s provides an indication of whether or not the principal independent variable has an effect that is significantly greater than the effect of the pairing variables themselves.

Discriminant Analysis

This increasingly popular procedure is actually a multivariate analysis of variance in reverse; that is, there is one categorical dependent variable[2] of group membership and there are two or more continuous independent

[2] It is of course possible to have a factorial discriminant analysis with group membership defined along two or more dependent variables, but such designs are rarely encountered in social science research.

variables. The p independent variables are treated as though one were carrying out a multiple regression analysis, and the dependent variable of k categories is coded in the same way that an independent variable is treated in one-way analysis of variance, that is, by creating $q = k - 1$ dummy dichotomies. A standard $p \times q$ canonical analysis is then applied to the resulting system, and the F test of the largest canonical correlation coefficient determines whether or not the k groups are significantly separable on the p variables.

Table 2 contains a summary of the results of a canonical correlation analysis of the separability of Project Talent Test Battery Variables 11 (English Test), 12 (Reading Comprehension Test), and 16 (Mathematics Test) for five groups of students defined by Variable 8 (College Plans).

Table 2
Example of Application of Canonical Analysis to a Discriminant Problem

Data

ID	X_1	X_2	X_3	Y_1	Y_2	Y_3	Y_4
1	87	39	20	0	0	0	1
2	76	15	15	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
504	87	22	18	0	0	0	0
505	105	45	26	1	0	0	0

$p = 3$
$q = 4$
$pq = 12$
$m = 497.5$
$s = (140/20)^{\frac{1}{2}} = 2.6458$
$ms - pq/2 + 1 = 1311.2606$[a]

Results

$\lambda_i = .2706, .0163, .0033, .0000$
$\Lambda = .7152$
$F(12, 1311) = 14.83$

Note. Data taken from Cooley and Lohnes (1971). Variables are from the Project Talent Test Battery: X_1 = Variable 11 (English Test), X_2 = Variable 12 (Reading Comprehension Test), X_3 = Variable 16 (Mathematics Test); Y_1, Y_2, Y_3, Y_4 = dummy variables for five groups of students defined by Variable 8 (College Plans); ID = identification number.

[a] Rounded to 1311 for use with the F sampling distribution.

Chi-Square Test of Independence of Two Variables

An extreme case of canonical correlation analysis is the situation in which all p variables and all q variables are dummy dichotomies. This is exactly what happens in a $j \times k$ contingency table where X_1 (equal to one or zero) indicates membership or nonmembership in row 1, X_2 indicates membership or nonmembership in row 2,..., and X_{j-1} indicates membership or nonmembership in row $j - 1$; Y_1 indicates membership or nonmembership in column 1, Y_2 indicates membership or nonmembership in column 2,..., and Y_{k-1} indicates membership or nonmembership in column $k - 1$. Since each observation can be represented by a pattern of ones and zeros on one set of $p = j - 1$ variables and another set of $q = k - 1$ variables, a test of the independence of the two original categorical variables can be regarded as just one more special instance of a $p \times q$ canonical analysis. Since $\chi^2(pq) = pq \cdot F(pq, \infty)$, the actual χ^2 value can be explicitly calculated, if desired, by multiplying the canonical F by pq ($ms - pq/2 + 1$ is close enough to ∞ for most data).

Table 3 contains a summary of the canonical approach to a chi-square test of the independence of Project Talent Test Battery Variables 1 (School Size) and 8 (College Plans).[3]

So What?

The fact that each of several popular statistical techniques can be regarded as a special case of canonical correlation analysis raises two very interesting questions:

1. Should they be taught that way?
2. Should they be run that way?

The answer to both questions is, probably not. Students who study canonical correlation analysis should surely be told that a wide variety of statistical procedures is subsumed under the canonical model, and some time

[3] Since the number of observations for one of the school sizes was so small (9), it was eliminated from the analysis, and therefore the contingency table is 3×5 with 496 observations rather than 4×5 with 505 observations.

should be spent in showing them why this is, using arguments and examples similar to those contained in this article. This is not to say, however, that all students should start with canonical correlation analysis and then go on to study its special cases (although the idea is indeed tempting), since there is so much matrix algebra that one must know before one can study canonical analysis. Nor does it necessarily follow that students should study canonical correlation analysis only and forget about all of the special jargon and formulas that are associated with t tests, analyses of variance and covariance, chi-square, and so on. They will still have to read the research literature in which such things abound and will be at a distinct disadvantage if they have not been exposed to these matters in their statistics courses.

The same sorts of considerations are relevant regarding the actual carrying out of the analyses themselves. Although a suitably written canonical correlation program is capable of handling a simple correlation or t test problem, for example, it would be computationally inefficient to do so unless it was the only program one happened to have around. (In his article, Cohen, 1968, pointed out that virtually all computer installations have a good working multiple regression program but may not have a collection of analysis of variance programs.) The t test for correlated samples is a case in point. Setting up all of those pairing variables and doing full-model and reduced-model analyses may consume an inordinate amount of time. On the other hand it could be argued that a front-end routine that is part of the canonical package can do all of the dummy coding internally for the users, that is, they would submit jobs thinking that they were doing one thing and the computer would actually do another. (Some analyses of variance are actually run as regression analyses anyhow.) Another argument in favor of a single supercanonical program is that certain computers are so fast that the difference between running an analysis of variance as an analysis of variance, for example, and running it as a special case of a general canonical problem is only a matter of milliseconds, and the latter approach would therefore not be cost-ineffective.

Table 3
Example of Application of Canonical Analysis to Chi-Square Test

Data						
ID	X_1	X_2	Y_1	Y_2	Y_3	Y_4
1	0	1	0	0	0	1
2	1	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
495	0	0	0	0	0	0
496	0	0	1	0	0	0

$p = 2$
$q = 4$
$pq = 8$
$m = 491.5$
$s = (60/15)^{\frac{1}{2}} = 2$
$ms - pq/2 + 1 = 980$

Results

$\lambda_i = .0156, .0007, .0000, .0000$
$\Lambda = .9837$
$F(8, 980) = 1.01$
$\chi^2 = pq \cdot F = 8(1.01) = 8.08$

Note. Data taken from Cooley and Lohnes (1971). Variables are from the Project Talent Test Battery: X_1, X_2 = Variable 1 (School Size); Y_1, Y_2, Y_3, Y_4 = Variable 8 (College Plans); ID = identification number.

The pedagogical and computational implications of the generality of the canonical approach therefore depend on the kinds of students (and faculty!) one has (good or poor math backgrounds, how much statistics they will ultimately be studying, etc.) and the kinds of computing facilities (hardware and software) that are available. It is also a matter of philosophy. The deductive approach would argue for teaching the canonical approach first, then its subcases, and running all analyses as canonical analyses. The more popular and probably more psychologically defensible inductive approach would argue for working up toward canonical analyses (pointing out matters such as $F = t^2$ along the way) and having a variety of special-purpose computing routines.

One final note about the assumptions underlying Rao's F test, especially the question of what happens to continuity in general, and normality in particular, whenever one introduces all of those dummy dichotomies:

Contrary to popular misconception, the significance test does not assume multivariate normality for the total system of variables. What is assumed is a normal distribution of one set of variables for each combination of values for the other set of variables.[4] Also assumed is homoscedasticity, that is, equal dispersion for such distributions, and homogeneity of regression for covariance-type analyses. The technique that would seem to suffer most from assumption violation is the application of canonical analyses to chi-square, with no apparent hope for normality at all. But there is hidden normality whenever the expected frequencies for the cells are large enough to permit approximately normal distributions of obtained frequencies around them, which is why the authors of most statistics textbooks include so many admonitions against small expected frequencies.

[4] As Cohen (1968) pointed out in his article, the equivalence between the analysis of variance and regression analysis holds for fixed effects.

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,* 29–38.

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

Cohen, J., & Cohen, P. *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, N.J.: Erlbaum, 1975.

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

Darlington, R. B., Weinberg, S. L., & Walberg, H. J. Canonical variate analysis and related techniques. *Review of Educational Research,* 1973, *43,* 433–454.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research.* New York: Holt, Rinehart & Winston, 1973.

McNemar, Q. *Psychological statistics* (4th ed.). New York: Wiley, 1969.

Morrison, D. F. *Multivariate statistical methods* (2nd ed.). New York: McGraw–Hill, 1976.

Press, S. J. *Applied multivariate analysis.* New York: Holt, Rinehart & Winston, 1972.

Rao, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

Roscoe, J. T. *Fundamental research statistics for the behavioral sciences* (2nd ed.). New York: Holt, Rinehart & Winston, 1975.

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

Welkowitz, J., Ewen, R. B., & Cohen, J. *Introductory statistics for the behavioral sciences* (2nd ed.). New York: Academic Press, 1976.

Wilks, S. S. Certain generalizations in the analysis of variance. *Biometrika,* 1932, *24,* 471–474.

Received March 1, 1977 ■