325-75

INSTITUTE FOR RESEARCH ON POVERTY DISCUSSION PAPERS

CANONICAL CORRELATION AND THE RELATIONS BETWEEN SETS OF VARIABLES

Franklin D. Wilson

UNIVERSITY OF WISCONSIN - MADISON

CANONICAL CORRELATION AND THE RELATIONS BETWEEN SETS OF VARIABLES

Franklin D. Wilson

University of Wisconsin, Madison

December 1975

The research reported here was supported in part by funds granted to the Institute for Research on Poverty at the University of Wisconsin-Madison by the Department of Health, Education, and Welfare pursuant to the Economic Opportunity Act of 1964. The opinions expressed are solely those of the author.

ABSTRACT

This paper reviews and develops summary measures of associations between multiple sets of variables through the application of canonical correlation analysis. These measures are subsequently applied to a specific research problem. Some of the data analysis situations for which canonical correlation is appropriate are also discussed.

CANONICAL CORRELATION AND THE RELATIONS BETWEEN SETS OF VARIABLES

I. Introduction

Sociologists are becoming increasingly sophisticated in their use of multivariate statistical models, and a number of excellent sources are now available in the literature. (See Van de Geer, 1971; Blalock, 1971; Goldberger and Duncan, 1973.) We wish to direct sociologists' attention to a multivariate statistical technique, whose usefulness as a data analysis tool has a great deal of appeal when interest centers around the joint distribution of two or more sets of variables. The technique, canonical correlation, was developed by Hotelling (1935, 1936) more than three decades ago, and is used rather extensively in biometrics and psychometrics. In sociological literature, Klatzky and Hodge (1971) used the technique to analyze intergenerational occupational mobility, Van de Geer (1971) used the technique to estimate the parameters of unobservable variable models, and Hauser and Goldberger (1972) noted the similarities between canonical correlation and confirmatory factor analysis in the estimation of unobservable variable models. However, these applications do not begin to exhaust the potential usefulness of canonical correlation analysis. Some of the data analysis situations for which canonical correlation is appropriate are discussed in this paper.

Consider a situation in which a researcher is in a position to separate the variables of interest to him into sets, and he is in

a position to postulate "flows" of influences among the variable sets based upon information obtained from a theoretical model. The researcher's primary objective is in determining to what extent and at what point the distributions of these variable sets intersect. All multivariate statistical techniques are designed to provide answers to these types of questions, though perhaps from different data analytic points of view. Moreover, the researcher is interested in answering the following questions as a means of evaluating the plausibility of some specific hypotheses implied in his theoretical model: (1) What is the total relationship between the dependent and the independent variable sets? (2) In instances in which the independent set consists of several theoretically distinct subsets, one may ask what is the relative contribution of each subset to the total amount of variation explained in the dependent set? (3) Which variables in the dependent and independent set(s) respectively contributed most to the total amount of variation shared between the sets? Those familiar with univariate correlation and regression analysis will immediately recognize that questions one and two are practically identical to those that one would ask if the relations between individual variables are pursued. Indeed, it has been shown that certain aspects of canonical correlation analysis are simple extensions of univariate correlation theory (Rozeboom, 1965, 1968).

This paper presents a pedagogically oriented review of much of the technical literature that has been presented on canonical correlation (see Bartlett, 1941, 1947; Anderson, 1957; Morrison, 1967; Cooley and Lohnes, 1971; Van de Geer, 1971). We think that the

specific problems explored here, by way of an example, should stimulate a greater interest in the general usefulness of this multivariate statistical technique. Within this context, the current discussion focuses on two specific objectives.

First, as is known by most practicing methodologists, it is suggested that canonical correlation analysis offers a parsimonious way to reduce the complexities involved in relating several dependent variables to several independent variables, particularly when it is appropriate to conceptualize dependent and independent variables respectively as indicators of theoretical constructs. However, it should be noted that the approach employed here has neither the statistical precision nor the theoretical parsimony of simultaneous statistical models, particularly when the research problem calls for their use, and their assumptions can be met (see Hauser and Goldberger, 1972; Burt, 1973; Duncan and Goldberger, 1973). On the other hand, it can be argued that deficiencies in the data and/or in the theoretical model should not deter researchers from examining, at least in an exploratory manner, the fruitfulness of a theoretical approach to a subject that is defined as problematic. In this respect, canonical correlation, as it is applied in this paper, can provide the researcher with an alternative whose requirements are less stringent than those characteristic of simultaneous estimation procedures.

The second objective involves an attempt to resolve some of the problems frequently encountered in trying to interpret canonical solutions. It is probably the case that one of the main reasons why canonical correlation is so infrequently used by researchers has to do

with the difficulty associated with interpreting canonical roots and vectors. It is suspected that this problem of interpretation arises partly from a lack of appreciation of exactly what is being done when the relationship between sets of variables are subjected to a canonical correlation analysis. We take the position that the interpretation problem can be practically eliminated if it can be shown that canonical correlation is a parsimonious way of decomposing a set of multiple correlation coefficients. Thus, it will be shown that both the canonical coefficients and vectors can be given interpretations that are as meaningful as computing multiple and multiple-partial correlation coefficients.

II. Applications

In this section the particular approach taken toward canonical correlation is applied to a specific research problem addressed by Wilson's (1973) study of the determinants of housing status. The interest is in analyzing the determinants of the quality of housing occupied by primary families who owned their dwelling unit in 1960. The dependent set Y, housing quality, is composed of measures of whether the dwelling unit is in standard condition (Y_1), the age of the unit (Y_2), and a measure of the quality attributes of the unit (Y_3). The independent set Z consists of measures of marital duration (W_1), the total number of children present in the family (W_2), age of the youngest child (W_3), education (X_1), occupational prestige (X_2), and total family income (X_3). The first three Z variables are defined as measures of family status (W), and the latter

three are defined as measures of socioeconomic status (X).[2] The observed correlation among these measures is exhibited in Table 1.

Figure 1 summarizes the expected direction of the relationships among the variable sets. It should be noted that the model as diagrammed postulates relationships among theoretical constructs represented by the variable sets (cf. Sullivan, 1972). The reason for this relates to the fact that evaluating the full implication of the model may require the use of more than one canonical solution. Thus, for example, the correlation between Y and its indicators may require two different sets of estimates in order to determine the total effects of W and X.

In any event, the model hypothesizes that the effect of family status on housing quality is expected to be negative, largely because of the influence exerted by family size and age of the youngest child. Large families are least likely to be in a position to spend a great deal on housing consumption. Socioeconomic status should have a negative effect on family status, because size of family is inversely related to all three measures of socioeconomic status. Finally, socioeconomic status should have a positive effect on housing quality, because the quality of the housing environment should reflect social status considerations.

With respect to the research questions posed earlier, a full evaluation of the implications of the model diagrammed in Figure 1

Table 1. Correlations between Measures of Housing Quality, Family Status, and Socioeconomic Status for Whites Who Owned Their Home in 1960 (N = 8700)

Variables	Symbol	Y_1	Y_2	Y_3	W_1	W_2	W_3	X_1	X_2	X_3
Housing quality										
Condition of unit	Y_1	1.000	.158	.016	.031	-.049	-.010	.112	.105	.116
Age of unit	Y_2	.158	1.000	.181	.355	.146	.069	.220	.098	.106
Quality attributes of unit	Y_3	.016	.181	1.000	-.020	-.057	-.009	.266	.300	.278
Family status										
Marital duration	W_1	.031	.355	-.020	1.000	.479	.228	.357	.059	.096
Number of children	W_2	-.049	.146	-.057	.479	1.000	.661	.212	.041	.145
Age of youngest child	W_3	-.010	.069	-.009	.228	.661	1.000	.101	.016	.223
Socioeconomic status										
Education	X_1	.112	.220	.266	.357	.212	.101	1.000	.534	.296
Occupational prestige	X_2	.105	.098	.300	.059	.041	.016	.534	1.000	.310
Family income	X_3	.116	.106	.278	.096	.145	.223	.296	.310	1.000

Source: 1960 Census 1/1000 Public Use Sample Tapes.

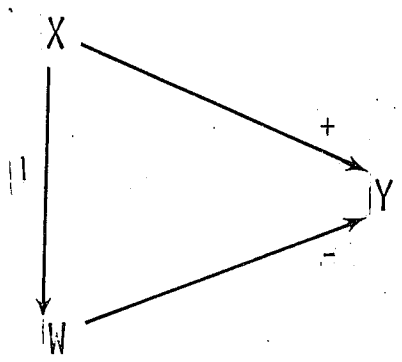

Figure 1. The Determinants of Housing Quality.

require (1) a measure, analogous to a multiple R^2, which can be used to summarize the overall predictive ability of the model; (2) measures, analogous to multiple-partial correlation coefficients, which can be used to determine the relative contribution made by the independent subsets W and X to the total variation explained in the dependent set Y; (3) measures that can be used to interpret the direction (positive versus negative) of the relationships between the variable sets; and (4) measures, when used in conjunction with those in (3), that can aid in determining which measured indicator variable(s) of the respective sets played significant role(s) in determining the overall relationships between the variable sets.

Matrix notation is employed throughout this exposition in order to clarify and enhance the derivations of specific measures. For illustrative purposes, let Y represent a P_1 x N matrix ($P_1 = 3$), W a P_w x N matrix ($P_w = 3$), X a P_x x N matrix ($P_x = 3$), and Z a P_2 x N matrix ($P_2 = P_w + P_x = 6$). Note further that N refers to sample size, and P_j refers to the number of variables in each set, respectively. Assuming all variables are expressed in standard form, the relationship between sets Y and Z can be expressed in terms of the following equations:

$$\begin{aligned} U &= A' Y \\ V &= B' Z \end{aligned} \qquad (1)$$

where

U and V are $K \times N$ matrices of canonical variates, and

A' and B' are $K \times P_j$ matrices of canonical weights.

The rows of U and V are linear combinations of the variables in sets Y and Z respectively. The relationship between the j^{th} linear combination in U and V can be expressed in terms of a canonical correlation coefficient. There are K such canonical coefficients possible. The problem addressed by canonical correlation reduces to finding: (1) the matrices A and B of canonical weights, and (2) a $K \times 1$ column vector C, with elements c_j $(j = 1,\ldots,k)$, which are the correlations between linear combinations of the variables in set Y with those in set Z. In order to find the vector C and the matrices A and B, we form the products[3]

$$\begin{bmatrix} Y' \\ Z' \end{bmatrix} \underline{\begin{bmatrix} Y & Z \end{bmatrix}} = \begin{bmatrix} Y'Y & Y'Z \\ Z'Y & Z'Z \end{bmatrix},$$

multiplying by $1/N$,

$$1/N \begin{bmatrix} Y'Y & Y'Z \\ Z'Y & Z'Z \end{bmatrix} = \begin{bmatrix} R_{YY} & R_{YZ} \\ R_{ZY} & R_{ZZ} \end{bmatrix},$$

and solve the following set of homogeneous equations:[4]

$$\left[(R_{YY})^{-1} R_{YZ} (R_{ZZ})^{-1} R_{ZY} - \lambda_j I \right] A_J = 0 \qquad (2)$$

$$\left[(R_{ZZ})^{-1} R_{ZY} (R_{YY})^{-1} R_{YZ} - \mu_j I \right] B_J = 0, \qquad (3)$$

where

λ_J and μ_J are characteristic roots $(J = 1,\ldots,K)$,

I is the identity matrix,

A_J and B_J are $P_j \times 1$ column vectors of canonical weights $(J = 1,\ldots,K)$ (These vectors are the transpose of the row vectors in A' and B'.)

and where the following constraints are imposed:

(1) R is of full rank, e.g., $(R_{YY})^{-1}$ and $(R_{ZZ})^{-1}$ exist.

(2) Y is a $P_1 \times N$ matrix.

Z is a $(P_w + P_x) \times N = P_2 \times N$ matrix.

(3) The first $k \leq \min(P_1, P_2)$ characteristic roots of $(R_{YY})^{-1} R_{YZ} (R_{ZZ})^{-1} R_{ZY}$ are distinct.

(If $P_2 > P_1$, then all of the roots extracted, $(R_{ZZ})^{-1} R_{ZY} (R_{YY})^{-1} R_{YZ}$, will not be distinct. The number of nondistinct roots will be equal to $P_2 - P_1$.)

(4) $A'_J R_{YY} A_J = 1$ and $B'_J R_{ZZ} B_J = 1$,

in order that the canonical variates in U and V are in standard form.

$$(5)\quad A_I' R_{YY} A_J = 0 \text{ and } B_I' R_{ZZ} B_J = 0.$$

Applying equations (2) and (3) to the observed correlation matrix displayed in Table 1, we have

$$\lambda = \mu = C^2 = \begin{bmatrix} .178 \\ .135 \\ .003 \end{bmatrix} \qquad C = \begin{bmatrix} .422 \\ .368 \\ .055 \end{bmatrix}$$

and

$$A = \begin{bmatrix} .365 & -.103 & .930 \\ .137 & .944 & -.103 \\ .921 & -.315 & -.353 \end{bmatrix} \qquad B = \begin{bmatrix} .022 & .912 & .527 \\ -.352 & .018 & -1.312 \\ .053 & -.020 & .486 \\ .448 & .050 & -.047 \\ .388 & -.094 & -.339 \\ .526 & -.070 & .125 \end{bmatrix}.$$

It can be observed that the characteristic roots of equations (2) and (3) are identical and are the squared canonical correlation coefficients. Since all of the canonical coefficients are significant beyond the (.01) level of rejection using Wilk's lambda (Barlett, 1941, 1974),[5] we are confronted with the problem of interpreting the substantive significance of at least the first two canonical coefficients.

A. Multiple Coefficients

The key to interpreting canonical coefficients is recognition of the fact that they are defined as the correlations between linear combinations of the original variables in sets Y and Z, and not the

correlations between the original variables themselves. Thus, each squared canonical coefficient is a measure of a certain amount of the total variation shared between two sets of variables. A measure of the total amount of variation shared between two sets of variables can also be obtained, which is analogous to but not identical with the squared product moment correlation coefficient, or with the squared multiple correlation coefficient (when the independent set is composed of two or more independent variable subsets). This coefficient has been termed the Squared Vector Multiple Correlation Coefficient (hereafter referred to as SVMC) (Srikantan, 1970). The coefficient SVMC is defined (Rozeboom, 1965, 1968; Srikantan, 1970) as

$$\text{SVMC} = R^2 = 1 - \prod_{j=1}^{k} (1 - c_j^2) , \qquad (4)$$

where $\prod$ indicates sequential multiplication.

Now,

$$\text{VCA} = \prod_{j=1}^{k} (1 - c_j^2)$$

is the Vector Coefficient of Alienation, or the vector correlation between Y and the residual of $Y - \hat{Y}$, where $\hat{Y}$ is the least squares estimates of the variables in Y. Thus, the correlation between Y and $Y - \hat{Y}$ is also a canonical relationship that conforms to equations (2) and (3). Therefore,

$$\text{SVMC} = 1 - \text{VCA} .$$

If the researcher is interested in estimating the total amount of variation shared between two sets of variables, SVMC is the appropriate measure. With respect to the first research question posed earlier,

$$\text{SVMC} = 1 - .709 = .291 ,$$

which suggests that 29 percent of the variation of the variates in set U can be explained by the variates in set V. Note particularly that the interpretation is applied to the variates and not the original set of variables.

Srikantan (1970) presents two other multiple canonical coefficients that may be appropriate for some research problems. However, we favor SVMC because it is a direct extension of the squared product moment correlation coefficient. The major disadvantage of all of these measures is that their interpretations are not necessarily equivalent to the proportion of variation in the variables of set Y that can be explained by the variables in set Z. Measures that permit this type of interpretation are available, and are our next topic of discussion. (See Stewart and Love, 1968; Miller and Farr, 1971; Alpert and Peterson, 1972; Wood, 1972.)

It was noted previously that the number of nonzero and positive c_j^2 values derived from equation (2) is determined by the rank of the variance-covariance matrix (the correlation matrix in the example) associated with the smallest variable set. For example, if the Y matrix contains three variables and the Z matrix six, the

maximum of nonzero and positive c_j^2 values is limited to three (although one, and perhaps all three, may not be statistically significant). Consequently, it is statistically possible to explain all the variation in the variables in set Y and only 50 percent of the variation in the variables of set Z (see Alpert and Peterson, 1972).[6] One aspect of the interpretation problem alluded to earlier with respect to canonical correlation is the symmetric character of the squared canonical correlation coefficients and its multiples. Thus, our immediate objective is to develop an asymmetric measure of explained variation, which is analogous to the squared multiple correlation coefficient. It will be recalled that the squared multiple correlation coefficient is a measure of the amount of variation in a given variable that can be explained by a linear combination of predicting variables. Stewart and Love (1968), and Miller and Farr (1971) have developed a measure for canonical analysis that is analogous to the squared multiple correlation coefficient and can be interpreted as the proportion of the variation in set Y which can be explained by set Z. We will denote these measures as $_dR_{y\cdot z}$ when the emphasis is on explaining the variation in set Y, and $_dR_{z\cdot y}$ when the emphasis is on explaining the variation in set Z. In general,

$$_dR_{y\cdot z} \neq {}_dR_{z\cdot y} \quad .$$

It is this asymmetric quality of this measure (hereafter referred to as total redundancy) that makes it a more useful measure than either c_j^2 or its multiples. As a measure of association, it has the

following desirable qualities: (1) $_{d}R_{y \cdot z}$ will be zero if and only if $R_{YZ} = 0$; and (2) it will achieve a value of 1 if and only if the variation in each of the y_i variables can be completely explained by the variations in Z, e.g., $R_{YZ} = 1$.

For illustrative purposes, we shall focus mainly on the derivation of $_{d}R_{y \cdot z}$, since $_{d}R_{z \cdot y}$ can be obtained in a similar manner. It can be shown that $_{d}R_{y \cdot z}$ is an arithmetic average of the squared multiple correlation coefficients obtained from predicting each Y_i variable from all of the variables in Z. First, we define the P_j x K matrices R^2_{YU} and R^2_{ZV}.

$$R_{YU} = R_{YY} A$$

$$= \begin{bmatrix} .387 & .044 & .921 \\ .347 & .938 & -.020 \\ .919 & -.155 & -.362 \end{bmatrix}$$

and

$$R^2_{YU} = \begin{bmatrix} .150 & .002 & .848 \\ .121 & .879 & .000 \\ .845 & .024 & .131 \end{bmatrix} .$$

Similarly,

$$R_{ZV} = \begin{bmatrix} .093 & .992 & -.014 \\ -.115 & .472 & -.750 \\ -.005 & .201 & -.241 \\ .724 & .331 & -.237 \\ .752 & -.035 & -.348 \\ .717 & .003 & -.017 \end{bmatrix} \quad \text{and,}$$

$$R^2_{ZV} = \begin{bmatrix} .008 & .985 & .000 \\ .013 & .225 & .563 \\ .000 & .040 & .058 \\ .524 & .109 & .056 \\ .565 & .001 & .121 \\ .514 & .000 & .000 \end{bmatrix} .$$

The $r^2_{yi,uj}$ and $r^2_{zi,vj}$ elements in R^2_{YU} and R^2_{ZV} respectively, are defined as the proportion of the variation in the i^{th} variable in Y or Z that can be explained by the j^{th} canonical variate in U or V, respectively. Postmultiplying R^2_{YU} and R^2_{ZV} by the C^2 $K \times 1$ column vector (the vector of squared canonical correlation coefficients), we have

$$R^2_{YU}\, C^2 = Q_Y$$

$$Qy = \begin{bmatrix} .150 & .002 & .848 \\ .121 & .879 & .000 \\ .845 & -.024 & .131 \end{bmatrix} \begin{bmatrix} .178 \\ .135 \\ .003 \end{bmatrix}$$

$$= \begin{bmatrix} .030 \\ .140 \\ .154 \end{bmatrix}$$

and

$$R^2_{ZV}\, C^2 = Q_Z$$

$$R^2_{ZV}\, C^2 = \begin{bmatrix} .008 & .985 & .000 \\ .013 & .225 & .563 \\ .000 & .040 & .058 \\ .524 & .109 & .056 \\ .565 & .001 & .121 \\ .514 & .000 & .000 \end{bmatrix} \begin{bmatrix} .178 \\ .135 \\ .003 \end{bmatrix}$$

$$Q_Z = \begin{bmatrix} .135 \\ .050 \\ .007 \\ .108 \\ .101 \\ .093 \end{bmatrix}$$

yields a $P \times 1$ column vector of squared multiple correlation coefficients. Postmultiplying Q_Y further by a $P \times 1$ unity vector τ yields

$$Q'\tau = R^2_{y \cdot z} = \sum_{i=1}^{P_1} R^2_{y_i \cdot z} \quad .$$

Inasmuch as $R^2_{y \cdot z}$ is simply the sum of the $R^2_{y_i \cdot z}$ values predicting each variable in Y given the variables in Z, it is possible that the former can achieve a value greater than one. The maximum value of $R^2_{y \cdot z}$ is equal to $Tr(R_{YY})$, or the number of variables in Y. Ideally, one would want to employ a measure to explain variation that conforms to the limits of $(0,1)$, which makes $R^2_{y \cdot z}$ less attractive as a measure of association. The asymmetric measure ${}_dR_{y \cdot z}$ corrects for this undesirable quality by dividing $R^2_{y\,z}$ by the number of variables in Y. Total redundancy can thus be defined as

$$
\begin{aligned}
{}_{d}R_{y\cdot z} &= \sum_{j=1}^{k} c_j^2 \left[\sum_{i=1}^{P_1} \sum_{j=1}^{k} r^2_{yi,uj} / P_1 \right] \\
&= \frac{1}{P_1} \left[\sum_{j=1}^{k} (R_J' R_J)\, c_j^2 \right] \\
&= \frac{1}{P_1} (R^2_{y\cdot z}) \\
&= \frac{1}{P_1} \sum_{i=1}^{P_1} R^2_{yi\cdot Z} \\
&= (.030 + .140 + .154)/3 \\
&= .108,
\end{aligned}
$$

where the $r_{yi,uj}$'s are elements of R_J and, the R_Js are the $P_1 \times 1$ column vectors of R_{YU}.

The size of the multiple redundancy measure indicates that socioeconomic status and family status combined explain 11 percent of the variation in the measures of housing quality. However, inasmuch as the theoretical model postulates asymmetric relationships among the variable sets, this measure is of little use in this respect. The measures most relevant for this task are the multiple-partial measures of redundancy, which are developed below.

B. Multiple-Partial Coefficients

In instances in which the independent variable set can be decomposed into subsets, we can define a set of multiple-partial coefficients. These coefficients can be used to determine the relative contribution made by each subset of Z to the total amount of

variation explained in set Y. In the example, the independent set Z is composed of two sets of independent variables, i.e., the subset W of family status variables, and the subset X of socioeconomic status variables. The first step in the computation of the multiple-partial coefficients involves computing the redundancy measures ${}_{d}R_{y\cdot w}$ and ${}_{d}R_{y\cdot x}$, which indicate the amount of variation in set Y that can be explained by sets W and X separately. Once this is accomplished, ${}_{d}R_{y\cdot z}$ can be decomposed into the following components:

(1) $$ {}_{d}R_{y\cdot w(x)} = ({}_{d}R_{y\cdot z} - {}_{d}R_{y\cdot x})/(1 - {}_{d}R_{y\cdot x}) $$

$$ = (.108 - .070)/(1 - .070) $$

$$ = .041, $$

which indicates that the relative contribution of family status to the total amount of variation explained in housing quality is (.041) or 38 percent [(.041/.108) 100].

(2) $$ {}_{d}R_{y\cdot x} = ({}_{d}R_{y\cdot z} - {}_{d}R_{y\cdot w})/(1 - {}_{d}R_{y\cdot w}) $$

$$ = (.108 - .047)/(1 - .047) $$

$$ = .064, $$

which indicates that the relative contribution of socioeconomic status to the total amount of variation explained in housing quality is (.064) or 59 percent [(.064/.108) 100].

(3) Finally, we should point out that components (1) and (2) define the "unique" contribution of sets W and X. It is statistically possible that some portion of the total variation explained in set Y by sets W and X might represent the combined effect of these

independent subsets. This can occur when the independent subsets are highly interrelated and therefore may exert common influence. (See Duncan, 1970; Coleman, 1970, for examples.) The third component can be derived as a residual,

$$
\begin{aligned}
{}_{d}R_{y\cdot wx(wx)} &= {}_{d}R_{y\cdot z} - ({}_{d}R_{y\cdot x} + {}_{d}R_{y\cdot w})/1 - ({}_{d}R_{y\cdot x} + {}_{d}R_{y\cdot w}) \\
&= {}_{d}R_{y\cdot z} - ({}_{d}R_{y\cdot x(w)} + {}_{d}R_{y\cdot w(x)}) \\
&= .108 - (.064 + .041) \\
&= .003,
\end{aligned}
$$

which in our case is very small. The reader should note, however, that while the application of the above decomposition to situations in which there are more than two independent subsets might appear straightforward, it may be more difficult to interpret component (3), because this component would then be equal to the sum of all possible nonredundant combinations of covariations existing between the subsets.

The multiple-partial measures of redundancy provide the answer to the second question posed earlier. Clearly, the relationships between socioeconomic and family status with housing quality, though small, are nonzero. But the theoretical model postulates not only that the observed relationships are nonzero but also that they should be in a specific direction. With the multiple-partials, we can only say that the relationships are of a certain size; we cannot say whether they imply positive or negative relationships. This applies as well to the other canonical coefficients discussed earlier and largely results from the way in which these coefficients are computed. The direction of the

variable-set relationships and the issue of which specific variables within the dependent and independent sets, respectively, are responsible for the total relations between variable sets can be determined by further manipulating the $r_{yi,uj}$ and $r_{zi,vj}$ elements of R_{YU} and R_{ZV}, respectively.

C. Canonical Variate-Observed Variable Relations

If we used all of the information obtained from the matrices R_{YU} and R_{ZV} and the vector C, a more precise description of the relationships between socioeconomic and family status and housing quality would be as depicted in Figure 2, where the relations between the variates (α_j) are determined by applying constraints (4) and (5), and the relations between variates and indicators are defined as $r_{ij} = r_{yi,uj}$ or $r_{ij} = r_{zi,vj}$.

A useful indicator of between-set relationships is the sign and size of the $r_{yi,uj}$ and $r_{zi,vj}$ values. If we wanted to relate a variable in set Y with a variable in set Z, the sign of the r_{ij} values are important, because they indicate the direction of the association between the two variables as measured by the product moment correlation coefficient. Indeed, the product moment correlation coefficient has simply been subjected to decomposition and can be estimated from the following equation:

$$r_{yi,zi} = r_{yi,uj}\, c_j\, r_{zi,vj}. \qquad (5)$$

Applying this equation to the relations between Y_2 and Z_1 $(Z_1 = W_1)$, we have

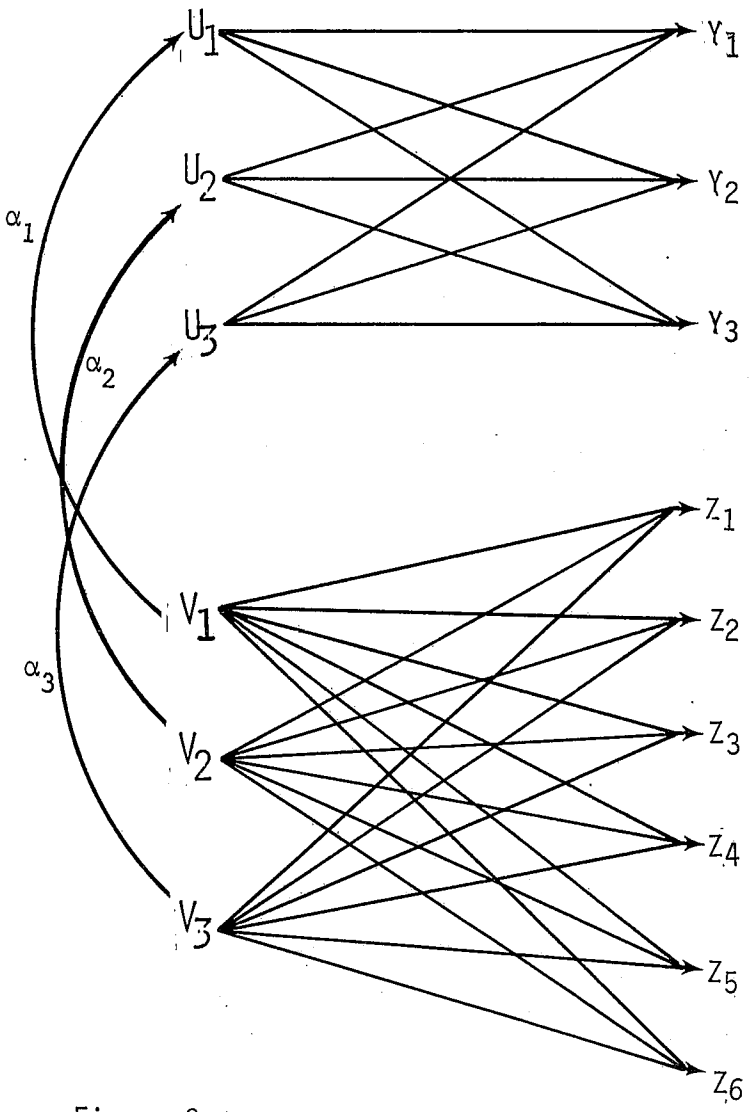

Figure 2.

$$r_{y2,z1} = .347\ (.422)\ .093 + .938\ (.368)\ .992 + (-.020)\ (.055)\ .014$$
$$= .013 + .342 + .000$$
$$+ .356.$$

More generally, the matrix R_{12} can be reproduced by applying the following equation:

$$\hat{R}_{12} = R_{YU}\, S\, R'_{ZV}, \qquad (6)$$

where R_{YU} and R_{ZV} are $P_1 \times K$ and $P_2 \times K$ matrices respectively, and S is a diagonal matrix with the J^{th} canonical correlation in the $K \times 1$ column vector C as elements.

Thus, the signs of the $r_{ji,uj}$ and $r_{zi,vj}$ values can be used to determine the general direction of the relationships between the variable sets. On the other hand, the sizes of these values are poor indicators of between-set relationships by themselves because they only indicate the contribution made by the i^{th} variable in sets Y or Z to the total amount of variation extracted by the j^{th} canonical variate from all the variables in each set, respectively. If they are weighted by the squared canonical correlation coefficients, they provide some indication of the amount of variation explained in the i^{th} variable of one set given all the variables in the other set via the k^{th} canonical relationship. The sum of these values for each variable across the j^{th} canonical relationship is equal to the squared multiple correlation coefficient for that variable given the variables in the other set. The reader will recall that the $P \times 1$ column vector Q of squared multiple correlation coefficients was defined as

$$R^2_{YU}\, c^2 = Q .$$

Now we wish to decompose each of the squared multiple correlation coefficients into an additive set of values that can be associated with each canonical variate extracted from set Y. Thus, if we multiply each $r^2_{yi,uj}$ value in R^2_{YU} by the c^2_j value it is associated with, we have

$$r^2_{yi,uj}\, c^2_j = l_{yi,uj} ,$$

which is a measure of the amount of variation explained in the i^{th} variable of set Y by the variables in set Z via the j^{th} canonical variate. (In the language of factor analysis, l_{ij} is simply the square of the loading of the i^{th} variable on the j^{th} factor.) It should be obvious that, by definition,

$$\sum_{j=1}^{k} r_{yi,uj}\, c^2_j = \sum_{j=1}^{k} l_{yi,uj}$$

$$= R^2_{yi\cdot Z}$$

and

$$\sum_{i=1}^{P_1} \sum_{j=1}^{k} r_{yi,uj}\, c^2_j = \sum_{i=1}^{P_1} \sum_{j=1}^{k} l_{yi,uj}$$

$$= \sum_{i=1}^{P_1} R^2_{yi\cdot Z}$$

$$= R^2_{Y\cdot Z} .$$

The $l_{yi,uj}$ (or $l_{zi,vj}$) values, then, not only provide us with a means of determining which variable in each set made the largest contribution to the j^{th} canonical relationship, but it also indicates what

proportion of the total variation explained in a given variable can be associated with the j^{th} canonical relationship. Thus, the total redundancy measure ${}_dR_{y\cdot z}$ can be used to estimate the total amount of variation in Y that can be explained by Z, and its decomposition into an additive set of values permit the determination of which variable in Y is actually being explained.

Table 2 reports the empirical estimates derived from most of the measures we have discussed in this chapter. The last column in the table reports the multiple and multiple-partial redundancy measures, whose relative sizes suggest that both socioeconomic and family status are related to the quality of the housing environment inhabited by owner households. As was noted earlier, our objectives are to determine not only whether socioeconomic status and family status are related to housing quality, but we want to determine whether the hypothesized directions of these relationships are confirmed by the data. We noted that the overall direction of the relationships between sets can only be determined by analyzing the signs and relative sizes of the relationships between the observed measures and the canonical variates. For each canonical solution extracted from equation (2), Table 2 reports r_{ij} and l_{ij} values for each of the measured variables. As a further aid to interpretation, the third column under each canonical solution reports the l_{ij} values as proportions of the total variation explained in each of the variables (as represented by multiple R^2 coefficients).

From Table 2 it can be observed that socioeconomic status appears to be related to housing quality because of the positive relationships between measures of the former and condition and quality attributes of dwellings. This observation is supported by the values reported under

Table 2. Canonical Relationships between Housing Quality, and Family Status and Socioeconomic Status for Whites Who Owned Their Home in 1960 (N = 8700)

Variable Sets	1st Canonical			2nd Canonical			3rd Canonical			$R^2_{J_i \cdot K}$	$d^R J \cdot K(L)$
	r_{ij}	l_{ij}	$\frac{l_{ij}}{R^2_{J_i \cdot K}}$	r_{ij}	l_{ij}	$\frac{l_{ij}}{R^2_{J_i \cdot K}}$	r_{ij}	l_{ij}	$\frac{l_{ij}}{R^2_{J_i \cdot K}}$		
Housing quality											.108[a]
Condition of unit	.387	.027	.90	.044	.000	.00	.921	.003	.10	.030	
Age of unit	.347	.021	.15	.938	.119	.85	-.020	.000	.00	.140	
Quality attributes of unit	.919	.150	.97	-.155	.003	.03	-.362	.000	.00	.154	
Family status											.0412[b]
Marital duration	.093	.002	.01	.992	.133	.99	-.014	.000	.00	.135	
Number of children	-.115	.002	.04	.472	.030	.60	-.750	.180	.33	.050	
Age of youngest child	-.005	.000	.00	.201	.001	1.00	-.241	.000	.00	.001	
Socioeconomic status											.064[c]
Education	.724	.093	.86	.331	.015	.14	-.237	.000	.00	.108	
Occupational prestige	.752	.100	1.00	-.035	.000	.00	-.348	.000	.00	.101	
Family income	.717	.091	1.00	.003	.000	.00	-.017	.000	.00	.091	

[a] The amount of variation in set Y which can be explained by sets W and X.

[b] Proportion of the total variation explained in set Y, which can be attributed to the effect of family status.

[c] Proportion of the variation explained in set Y, which can be attributed to the effect of socioeconomic status.

the first canonical solution in which the signs of the coefficients are all positive and the $l_{ij}/R^2_{J_i.K}$ values are at least (.86). The first canonical solution captures practically all of the covariation that exists between socioeconomic status and housing quality. Thus, with respect to this relationship, our expectations are confirmed.

The relationship between family status and housing quality emerges in the second canonical solution. Again, using the $l_{ij}/R^2_{J_i.K}$ values as the basis for evaluation, it is evident that age of dwelling is being explained by marital duration and number of children. Clearly, the basis of the relationships that housing quality haye with socio-economic status and family status are not the same. Moreover, it should be equally clear that our expectations in regards to the underlying reasons for the relationship between housing quality and family status are not confirmed. We postulated a negative relationship because it was suggested that large families are more likely to live in poorer quality housing. We find, on the other hand, that the relationship is positive and it is marital duration, not age and number of children, that is the basis for this relationship. These results are consistent with the argument that families age with their units.

It was predicted that socioeconomic status would be negatively related to family status because of the inverse relationship between size of family and the three measures of socioeconomic status. These relationships are reported in Table 3. Socioeconomic status explained an average of 9 percent of the variation in family status. Moreover, it is clearly evident that the positive relationship between marital duration and education is responsible for the overall relationship

Table 3. Canonical Relationship between Family Status and Socioeconomic Status for Whites Who Owned Their Home in 1960 (N = 8700)

Variable Sets	1st Canonical			2nd Canonical			$R^2_{J_i\ K}$	$d^R_{J\cdot K}$
	r_{ij}	l_{ij}	$l_{ij}/R^2_{J_i\cdot K}$	r_{ij}	l_{ij}	$l_{ij}/R^2_{J_i\cdot K}$		
Family status								
Marital status	-.989	.151	1.00	-.090	.000	.00	.152	.091
Number of children	.604	.057	.90	-.371	.007	.10	.063	
Age of youngest child	.335	.017	.29	.928	.041	.71	.058	
Socioeconomic status								
Education	.913	.129	1.00	-.097	.000	.00	.130	
Occupational prestige	.153	.004	1.00	-.027	.000	.00	.004	
Family income	.294	.013	.25	.910	.039	.75	.053	
Canonical correlations[a]	.393			.217				

[a]All canonical coefficients are significant beyond the .01 level of rejection.

between these variable sets. The relationship between education and number of children, though small, is positive, while income and occupational prestige seem to bear no relationship to this variable. Finally, family income appears to be positively related to age of youngest child with respect to both the first and second canonical solution, a relationship which our theoretical model did not predict.

III. Discussion

One of the main reasons why these particular sets of variables were chosen in order to demonstrate the utility of canonical correlation analysis relates to the structure of the observed correlation matrix. First, the within-set and between-set correlations are rather small, which is due in part to the particular manner in which these variables (particularly the measures of housing quality) were operationalized via the census. Even given these low values and the exploratory nature of the theoretical model under review, it would still be of some interest to determine the reasonableness of the model in terms of whether it warrants further investigation. The conceptualization of the observed variables as indicators of specific theoretical constructs would appear to this writer to be a reasonable approach to take toward these data. This is the primary reason why the model as depicted in Figure 1 is defined in terms of the relationships between sets of variables, although we were also interested in the issue of which variables within each set were responsible for the between-set relationships. Moreover, it should be apparent that a canonical solution is derived mainly from the between-set correlation matrix R_{YZ} and the correlation between

variates and indicator variables are largely a function of the structure of this matrix. Thus, the attempt here was not to find the optimal correlation between a theoretical construct and its indicators, but rather to simply summarize the relationships between variable sets without implying that an optimal set of relations were obtained. Admittedly, this goal is less ambitious and less parsimonious than what would be obtained using a simultaneous estimation procedure.

However, viewed from another angle, the technique employed presents a clear picture of the complexity of the relationships between the dependent set and each of the independent sets. We were able to detect the fact that measures of socioeconomic status and family status are differentially related to measures of housing quality. What this means essentially is that if housing quality were related separately to socioeconomic and family status, different variables in the former set would have emerged as being largely responsible for the total relationship between the variable sets. In other words, the correlations between indicator variables and canonical variates would vary depending on the nature of the variables in each set. This is an undesirable state of affairs, because unless we can assume that the effects of indicator variables within each independent set are the same with respect to each indicator in the dependent set, there is no single "best" estimate of the unobserved-unobserved correlations and the unobserved-indicator correlations. For example, if the first canonical solution is taken as the best overall estimate of the relationship of housing quality with socioeconomic status and family status, then we would have virtually eliminated the relationship between housing quality and family status, since that relationship emerged in the second canonical solution, not the first.

The problem of differential association between dependent and independent sets is likely to increase in complexity as the number of independent sets are increased which, in some instances, necessitates the application of less restrictive and less precise statistical models in order to evaluate the implications of the researcher's theoretical model. Thus, our main argument is simply that the measures we have proposed here can be used to partially overcome this problem when more sophisticated and restrictive statistical models should not be applied.

FOOTNOTES

1. The measure "quality attributes of the dwelling unit" is defined as that proportion of value of property which remains after eliminating from it the effects of its measured determinants. (See Wilson, 1973.)

2. Age of dwelling unit, age of youngest child, total number of persons in the family, education and occupational prestige (Duncan scale) are expressed in logarithms. The generalized least squares estimate of units in standard condition is employed. This estimate takes the form:

$$Y_i \left[\; 1/P \; (1 - P)\right]^{\frac{1}{2}}$$

where

Y_i is the observed (0, 1) value of the variable.

P is the OLS estimate of the probability of living in a standard unit.

The data for this analysis are derived from the 1960 Census 1/1,000 Public Use Sample tapes.

3. The interested reader can find an extensive discussion of derivations in the technical literature cited earlier.

4. If one solves equation (2), then the vector B_J can be obtained as follows:

$$B_J = \frac{(Z'Z)^{-1} Z'Y \; A_J}{[\; \lambda_j \;]^{\frac{1}{2}}} \quad \text{OR} \quad B_J = \frac{(R_{ZZ})^{-1} R_{ZY} \; A_J}{[\; \lambda \;]^{\frac{1}{2}}} \; .$$

5. Wilk's lambda conforms approximately to the chi square distribution with $(P_1)(P_2)$ degrees of freedom.

6. The latter is true if and only if the matrix is of full rank, otherwise more variation can be explained. This is the primary reason why it is

frequently suggested that the number of variables in the dependent set should be equal to or less than the number of variables in the independent set.

REFERENCES

Alpert, Mark I., and Peterson, Robert A.

"On the Interpretation of Canonical Analysis." Journal of Marketing Research 9 (May 1972): 187-192.

Anderson, T. W.

Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons, 1958.

Bartlett, M. S.

"The Statistical Signficance of Canonical Correlations." Biometrika 32 (January 1941): 29-38.

"Multivariate Analysis." Journal of the Royal Statistical Society Supplement 9: (1947): 176-190.

Blalock, H. M., ed.

Causal Models in the Social Sciences. Chicago: Aldine-Atherton, 1971.

Burt, Ronald S.

"Confirmatory Factor-Analytic Structures and The Theory Construction Process." Sociological Methods and Research 2 (November 1973): 131-190.

Coleman, J. S.

"Reply to Cain and Watts." American Sociological Review 35 (April 1970): 242-248.

Cooley, William A., and Lohnes, Paul R.

Multivariate Procedures for the Behavioral Sciences. New York: John Wiley and Sons, 1971.

Duncan, Otis Dudley.

"Partials, Partitions, and Paths." In Sociological Methodology, edited by Edgar F. Borgatta and George W. Bohrnstedt. San Francisco: Jossey-Bass, 1970.

Goldberger, Arthur S., and Duncan, Otis D., eds.

Structural Equation Models in the Social Sciences. New York: Seminar Press, Inc., 1973.

Hauser, Robert M., and Goldberger, Arthur S.

"The Treatment of Unobservable Variables in Path Analysis." In Sociological Methodology, edited by Herbert L. Costner. San Francisco: Jossey-Bass, 1971.

Hotelling, Harold.

"The Most Predictable Criterion." Journal of Educational Psychology 26 (February 1935): 139-142.

Klatzky, Sheila, and Hodge, Robert W.

"A Canonical Correlation Analysis of Occupational Mobility." Journal of the American Statistical Association 66 (March 1971): 16-22.

Morrison, Donald F.

Multivariate Statistical Methods. New York: McGraw-Hill, 1967.

Rozeboom, William.

"Linear Correlations Between Sets of Variables." Psychometrika 30 (March 1965): 57-71.

"The Theory of Abstract Partials: An Introduction." Psychometrika 33 (June 1968): 133-167.

Srikantan, K. S.

"Canonical Association Between Nominal Measurements." Journal of the American Statistical Association 65 (March 1970): 284-292.

Stewart, Douglas, and Love, William.

"A General Canonical Correlation Index." Psychological Bulletin 70 (September 1968): 160-163.

Sullivan, John L.

"Multiple Indicators and Complex Casual Models." In Casual Models in the Social Sciences, edited by H. M. Blalock. New York: Aldine-Atherton, 1971.

Van de Geer, John P.

Introduction to Multivariate Analysis for the Social Sciences. San Francisco: W. H. Freeman and Company, 1971.

Wilson, Franklin D.

"Dimensions of Housing Status." Unpublished Doctoral Dissertation, Washington State University, Pullman Washington, 1973.

Wood, Donald A.

"Toward the Interpretation of Canonical Dimensions." Multivariate Behavioral Research 7 (October 1972): 477-482.