OBTAINING PART AND BIPARTIAL CANONICAL CORRELATION ESTIMATES FROM PROC CANCORR: A BIOMEDICAL APPLICATION

David A. Ludwig and Steven N. Blair
University of South Carolina

Introduction

Large numbers of variables are often collected and examined during observational investigation. Statistical techniques that can uncover possible empirical relationships based on all the variables under study would be beneficial in determining the dimensionality of large data sets. Recent advances in statistical computing have made available once near impossible multivariate manipulations. Theoretical models used to define empirical multivariate structure require additional consideration due to the added complexity associated with modelling the interdependence of more than two variables. Timm and Carlson (1976) have noted that the choice of a correct multivariate model should be based on pragmatic reasoning and not statistical outcome. Incorrect modelling may result in spurious statistical conclusions which do not honestly reflect the underlying structure of the data. Therefore, the ability to manipulate key parameters of a multivariate model is essential for correct interpretation.

The multivariate general linear model provides numerous avenues for simultaneously analysing large numbers of intercorrelated variables. Specifically, canonical correlation is becoming increasingly popular as a data analytic method. Two reasons for the increased interest in this technique may be (a) part, partial, and bipartial adjustment, which adjusts estimates based on comitant varition, and (b) redundancy analysis, which provides a more meaningful interpretation of shared variance between two sets of variables than the overall squared canonical correlation.

Although the most recent version of PROC CANCORR provides a statement to perform partial canonical correlation, no specifications or examples are given to aid the researcher in accomplishing part or bipartial analysis. By utilizing a large biomedical data base, this paper illustrates how part and bipartial canonical redundancy analyses can be obtained from PROC CANCORR via PROC GLM.

Statistical Concepts

The specifics of redundancy analysis, part, partial, and bipartial adjustment have been discussed in numerous articles and books and are much too lengthy to reiterate. Although a general description of these methods will be given to justify their use, more detailed accounts can be found in Cooley and Lohnes (1971) and Thorndike (1978).

The concept of part, partial, and bipartial simple correlation was developed from about the turn of the century to the early 1940's. More recently, these concepts have been generalized to sets of variates and the notion of part, partial, and bipartial canonical correlation. Timm and Carlson (1976) provide one of the more recent detailed expositions on the subject of part and bipartial canonical correlation. The multivariate extension and matrix analog of part correlation (sometimes referred to as semi-partial correlation), part canonical correlation, estimates the relationship between two sets of variates (predictor and criterion set). After a third set of variates (control set), is removed from the criterion set. Unlike partial adjustment, which adjusts both predictor and criterion sets, part correlation would provide a more meaningful model when the control set influences the criterion set but not the predictor set. Similarly, bipartial canonical correlation corrects both criterion and predictor sets using unique control sets. These control sets may or may not contain subsets of common variables depending on the relationship that is believed to underlie the data. Adjustment is accomplished with ordinary least squares and the residualized forms of the variables are then used in an unadjusted canonical correlation analysis.

Since traditional canonical correlation analysis is typically difficult to interpret, at least in terms of explained variance, it has not been a popular statistical methodology. Steward and Love (1968) improved on the interpretability of canonical correlation with the introduction of the redundancy index. Composite wise, and in terms of dependent and independent variables, the redundancy index is the mean variance of the criterion set that is explained by a particular canonical variate of the predictor set. Summing the redundancy indices yields an overall redundancy index which is interpreted as the variance of the criterion variables which is explained by the predictor variables (i.e., redundant variance).

Studies utilizing higher order canonical partialling in combination with redundancy analysis are virtually nonexistant. Thorndike (1978) has speculated on the great descriptive and exploratory power these two techniques may have when used together. The following example is designed to stimulate interest in these techniques, along with demonstrating how they can be accomplished in SAS.

Example Problem and SAS Program

In an attempt to explore the association between different patterns of sub-cutaneous body fat distribution and coronary heart disease (CHD) risk factors, data on 2063 adult males between

the ages of 18 and 65 were obtained from the archives of the Institute for Aerobics Research in Dallas, Texas. The analysis was purely exporatory with the premise that the distribution of body fat (how a person is fat) may be equally important as total body composition when examining the association with CHD. Seven skin folds (predictor set) were compared with six blood components (criterion set) in which elevated levels had previously been linked to an increased risk of CHD. The blood components included:

1. total cholesterol
2. triglycerices
3. uric acid
4. glucose
5. systolic blood pressure
6. diastolic blood pressure.

Since variability associated with location was of prime importance, total body composition, as measured by hydrostatic weight, was partialed from the blood components but not the seven skin folds (part adjustment). It was felt that partialling hydrostatic weight from the skin folds would not be appropriate since hydrostatic weight is itself a measure of body fat. This was the first step in the analysis. In the second step, and in addition to the part adjustment of hydrostatic weight, age was partialed from both the criterion and predictor sets (bipartial adjustment). The latter model could now be interpreted in terms of variance explained after hydrostatic weight while holding age constant. Due to the exploratory nature of the investigation, tests of significance were not performed.

Obtaining the correct statistics based on the bipartial model is relatively easy. Once the data has been inputted and prior to executing PROC CANCORR, GLM is used to residualize the criterion variables. The model statement in GLM lists the six blood components as dependent effects and hydrostatic weight as the independent variable. An output statement creates a new data set containing the six residualized variables along with the seven skin folds. This data set is now ready for PROC CANCORR. If PROC CANCORR is used without a partial statement the resulting output would be based on part adjustment. By listing age in the partial statement, a bipartial analysis would result. This approach works when one of the control sets is a subset of the other. If the control sets are unique, GLM must be used twice. Once to residualize the criterion set and once to residualize the predictor set. The syntax is as follows:

```
PROC GLM;
MODEL CHOL TRIG URIC GLUC SYST DIAS = HYDRO/
NOUNI;
OUTPUT OUT = ADJUST RESID = RCHOL RTRIG
RURIC RGLUC RSYST RDIAS;
PROC CANCORR DATA = ADJUST REDENDANCY;
PARTIAL AGE;
VAR RCHOL RTRIG RURIC RGLUC RSYST RDIAS;
WITH CHEST AXILLA TRICEPS BACK ABDOMEN HIP
THIGH;
*COMMENT* OUTPUT BASED ON BIPARTIAL MODEL;
```

If the control sets were unique, the syntax would contain two GLM statements and the partial statement would not be used. Therefore, the same results (within rounding error and provided the data set has no missing values) could be obtained with the following:

```
PROC GLM;
MODEL CHOL TRIG URIC GLUC SYST DIAS = HYDRO
AGE/NOUNI;
OUTPUT OUT = ADJUST RESID = RCHOL RTRIG
RURIC RGLUC RSYST RDIAS;
PROC GLM DATA = ADJUST;
MODEL CHEST AXILLA TRICEPS BACK ABDOMEN HIP
THIGH = AGE/NOUNI;
OUTPUT OUT = ADJUST2 RESID = RCHEST RAXILLA
RTRICEPS RBACK RABDOMEN RHIP RTHIGH;
PROC CANCORR DATA = ADJUST2 REDUNDANCY;
VAR RCHOL RTRIG RURIC RGLUC RSYST RDIAS;
WITH RCHEST RAXILLA RTRICEPS RBACK RABDOMEN
RHIP RTHIGH;
```

By manipulating the independent effects in the GLM model statements any form of partialling can be accomplished.

A scree test performed on the eigenvalues associated with the six canonical correlations from the unadjusted analysis indicated that only the first canonical composite was needed to represent the relationship between the two sets of data. Ten percent of the total criterion set variance was redundant with (explained by) the predictor set, with most of this coming from the first composite. The canonical loadings for the first canonical variables indicated a moderate relationship between the skin folds of the upper and lower torso and triglycerides. All of the loadings were in the positive direction. After bipartial adjustment only four percent of the total criterion variance was redundant with the predictor set. All of the canonical loadings were reduced substantially except for triglycerides. However, peripheral skin folds (triceps and thigh) now had negative loadings. The results suggest that patterns of fat deposition may prove to be a moderating factor in the prediction of coronary heart disease.

Recommendations

During the course of this investigation, pertinent issues in computing and statistical inference arose. They are presented here as suggestions and recommendations for possible improvements in the CANCORR procedure.

1. Provide for part and bipartial adjustment within the CANCORR PROC.
2. Provide a scree test in PROC CANCORR.
3. Investigate and possibly implement component rotation like that suggested by Wollenberg (1977).

References

Colley, W.W., & Lohnes P.R. Multivariate data analysis,. New York: Wilwy, 1971

Stewart, D., & Love, W. A general canonical correlation index. Psychological Bulletin, 1968, 70(3), 160-163.

Throndike, R.M. Correlation procedures for research. New York: Gardner, 1978.

Timm, N.H., & Carlson, J.E. Part and bipartial canonical correlation analysis. Psychometrika, 1976, 41(2), 159-176.

Wollenberg, A.L. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika, 1977, 42(2), 207-219.

Author's Address

David A. Ludwig
Department of Epidemiology and Biostatistics
College of Health
University of South Carolina
Columbia, SC 29208