Chapter 6
Introduction to Multivariate Procedures

Chapter Table of Contents

Chapter 6
Introduction to Multivariate Procedures

Overview

The procedures discussed in this chapter investigate relationships among variables without designating some as independent and others as dependent. Principal component analysis and common factor analysis examine relationships within a single set of variables, whereas canonical correlation looks at the relationship between two sets of variables. The following is a brief description of SAS/STAT multivariate procedures:

CORRESP	performs simple and multiple correspondence analyses, using a contingency table, Burt table, binary table, or raw categorical data as input. Correspondence analysis is a weighted form of principal component analysis that is appropriate for frequency data.
PRINCOMP	performs a principal component analysis and outputs standardized or unstandardized principal component scores.
PRINQUAL	performs a principal component analysis of qualitative data and multidimensional preference analysis.
FACTOR	performs principal component and common factor analyses with rotations and outputs component scores or estimates of common factor scores.
CANCORR	performs a canonical correlation analysis and outputs canonical variable scores.

Many other SAS/STAT procedures can also analyze multivariate data, for example, the CATMOD, GLM, REG, CALIS, and TRANSREG procedures as well as the procedures for clustering and discriminant analysis.

The purpose of *principal component analysis* (Rao 1964) is to derive a small number of linear combinations (principal components) of a set of variables that retain as much of the information in the original variables as possible. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering, and so on. Principal component analysis can also be viewed as an attempt to uncover approximate linear dependencies among variables.

The purpose of *common factor analysis* (Mulaik 1972) is to explain the correlations or covariances among a set of variables in terms of a limited number of unobservable, latent variables. The latent variables are not generally computable as linear combinations of the original variables. In common factor analysis, it is assumed that the variables are linearly related if not for uncorrelated random error or *unique variation* in each variable; both the linear relations and the amount of unique variation can be estimated.

Principal component and common factor analysis are often followed by rotation of the components or factors. *Rotation* is the application of a nonsingular linear transformation to components or common factors to aid interpretation.

The purpose of *canonical correlation analysis* (Mardia, Kent, and Bibby 1979) is to explain or summarize the relationship between two sets of variables by finding a small number of linear combinations from each set of variables that have the highest possible between-set correlations. Plots of the canonical variables can be useful in examining multivariate dependencies. If one of the two sets of variables consists of dummy variables generated from a classification variable, the canonical correlation is equivalent to canonical discriminant analysis (see Chapter 19, "The CANDISC Procedure"). If both sets of variables are dummy variables, canonical correlation is equivalent to simple correspondence analysis.

The purpose of *correspondence analysis* (Lebart, Morineau, and Warwick 1984; Greenacre 1984; Nishisato 1980) is to summarize the associations between a set of categorical variables in a small number of dimensions. Correspondence analysis computes scores on each dimension for each row and column category in a contingency table. Plots of these scores show the relationships among the categories.

The PRINQUAL procedure obtains linear and nonlinear transformations of variables using the method of alternating least squares (Young 1981) to optimize properties of the transformed variables' covariance or correlation matrix. PROC PRINQUAL nonlinearly transforms variables, improving their fit to a principal component model. The name, PRINQUAL, for principal components of qualitative data, comes from the special case analysis of fitting a principal component model to nominal and ordinal scale of measurement variables (Young, Takane, and de Leeuw 1978). However, PROC PRINQUAL also has facilities for smoothly transforming continuous variables. All of PROC PRINQUAL's transformations are also available in the TRANSREG procedure, which fits regression models with nonlinear transformations. PROC PRINQUAL can also perform metric and nonmetric multidimensional preference (MDPREF) analyses (Carroll 1972). The PRINQUAL procedure produces very little displayed output; the results are available in an output data set.

Comparison of the PRINCOMP and FACTOR Procedures

Although PROC FACTOR can be used for common factor analysis, the default method is principal components. PROC FACTOR produces the same results as PROC PRINCOMP except that scoring coefficients from PROC FACTOR are normalized to give principal component scores with unit variance, whereas PROC PRINCOMP by default produces principal component scores with variance equal to the corresponding eigenvalue. PROC PRINCOMP can also compute scores standardized to unit variance.

PROC PRINCOMP has the following advantages over PROC FACTOR:

- PROC PRINCOMP is slightly faster if a small number of components is requested.
- PROC PRINCOMP can analyze somewhat larger problems in a fixed amount of memory.
- PROC PRINCOMP can output scores from an analysis of a partial correlation or covariance matrix.
- PROC PRINCOMP is simpler to use.

PROC FACTOR has the following advantages over PROC PRINCOMP for principal component analysis:

- PROC FACTOR produces more output, including the scree (eigenvalue) plot, pattern matrix, and residual correlations.
- PROC FACTOR does rotations.

If you want to perform a common factor analysis, you must use PROC FACTOR instead of PROC PRINCOMP. Principal component analysis should never be used if a common factor solution is desired (Dziuban and Harris 1973; Lee and Comrey 1979).

Comparison of the PRINCOMP and PRINQUAL Procedures

The PRINCOMP procedure performs principal component analysis. The PRINQUAL procedure finds linear and nonlinear transformations of variables to optimize properties of the transformed variables' covariance or correlation matrix. One property is the sum of the first n eigenvalues, which is a measure of the fit of a principal component model with n components. Use PROC PRINQUAL to find nonlinear transformations of your variables or to perform a multidimensional preference analysis. Use PROC PRINCOMP to fit a principal component model to your data or to PROC PRINQUAL's output data set. PROC PRINCOMP produces a report of the principal component analysis and output data sets. PROC PRINQUAL produces only an output data set and an iteration history table.

Comparison of the PRINCOMP and CORRESP Procedures

As summarized previously, PROC PRINCOMP performs a principal component analysis of interval-scaled data. PROC CORRESP performs correspondence analysis, which is a weighted form of principal component analysis that is appropriate for frequency data. If your data are categorical, use PROC CORRESP instead of PROC

PRINCOMP. Both procedures produce an output data set that can be used with the %PLOTIT macro. The plots produced from the PROC CORRESP output data set graphically show relationships among the categories of the categorical variables.

Comparison of the PRINQUAL and CORRESP Procedures

Both PROC PRINQUAL and PROC CORRESP can be used to summarize associations among variables measured on a nominal scale. PROC PRINQUAL searches for a single nonlinear transformation of the original scoring of each nominal variable that optimizes some aspect of the covariance matrix of the transformed variables. For example, PROC PRINQUAL could be used to find scorings that maximize the fit of a principal component model with one component. PROC CORRESP uses the crosstabulations of nominal variables, not covariances, and produces multiple scores for each category of each nominal variable. The main conceptual difference between PROC PRINQUAL and PROC CORRESP is that PROC PRINQUAL assumes that the categories of a nominal variable correspond to values of a single underlying interval variable, whereas PROC CORRESP assumes that there are multiple underlying interval variables and therefore uses different category scores for each dimension of the correspondence analysis. PROC CORRESP scores on the first dimension match the single set of PROC PRINQUAL scores (with appropriate standardizations for both analyses).

Comparison of the TRANSREG and PRINQUAL Procedures

Both the TRANSREG and PRINQUAL procedures are data transformation procedures that have many of the same transformations. These procedures can either directly perform the specified transformation (such as taking the logarithm of the variable) or search for an optimal transformation (such as a spline with a specified number of knots). Both procedures can use an iterative, alternating-least-squares analysis. Both procedures create an output data set that can be used as input to other procedures. PROC PRINQUAL displays very little output, whereas PROC TRANSREG displays many results. PROC TRANSREG has two sets of variables, usually dependent and independent, and it fits linear models such as ordinary regression and ANOVA, multiple and multivariate regression, metric and nonmetric conjoint analysis, metric and nonmetric vector and ideal point preference mapping, redundancy analysis, canonical correlation, and response surface regression. In contrast, PROC PRINQUAL has one set of variables, fits a principal component model or multidimensional preference analysis, and can also optimize other properties of a correlation or covariance matrix. PROC TRANSREG performs hypothesis testing and can be used to code experimental designs prior to their use in other analyses.

See Chapter 3, “Introduction to Regression Procedures,” for more comparisons of the TRANSREG and REG procedures.

References

Carroll J.D. (1972), "Individual Differences and Multidimensional Scaling," in R.N. Shepard, A.K. Romney, and S.B. Nerlove (eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Volume 1)*, New York: Seminar Press.

Dziuban, C.D. and Harris, C.W. (1973), "On the Extraction of Components and the Applicability of the Factor Model," *American Educational Research Journal*, 10, 93–99.

Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

Hanson, R.J. and Norris, M.J. (1981), "Analysis of Measurements Based on the Singular Value Decomposition," *SIAM Journal on Scientific and Statistical Computing*, 2, 363–373.

Kshirsagar, A.M. (1972), *Multivariate Analysis*, New York: Marcel Dekker, Inc.

Lebart, L., Morineau, A., and Warwick, K.M. (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, New York: John Wiley & Sons, Inc.

Lee, H.B. and Comrey, A.L. (1979), "Distortions in a Commonly Used Factor Analytic Procedure," *Multivariate Behavioral Research*, 14, 301–321.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press.

Mulaik, S.A. (1972), *The Foundations of Factor Analysis*, New York: McGraw-Hill Book Co.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*, Toronto: University of Toronto Press.

Rao, C.R. (1964), "The Use and Interpretation of Principal Component Analysis in Applied Research," *Sankhya A*, 26, 329–358.

Van den Wollenberg, A.L. (1977), "Redundancy Analysis—An Alternative to Canonical Correlation Analysis," *Psychometrika*, 42, 207–219.

Young, F.W. (1981), "Quantitative Analysis of Qualitative Data," *Psychometrika*, 46, 357–388.

Young, F.W., Takane, Y., and de Leeuw, J. (1978), "The Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method with Optimal Scaling Features," *Psychometrika*, 43, 279–281.