Multivariate Analysis.

Hervé Abdi[1]
The University of Texas at Dallas

Introduction

As the name indicates, multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than one variable. Several of these techniques were developed recently in part because they require the computational capabilities of modern computers. Also, because most of them are recent, these techniques are not always unified in their presentation, and the choice of the proper technique for a given problem is often difficult.

This article provides a (non-exhaustive) catalog in order to help decide when to use a given statistical technique for a given type of data or statistical question and gives a brief description of each technique. This paper is organized according to the number of data sets to analyze: one or two (or more). With two data sets we consider two cases: in the first case, one set of data plays the role of predictors (or independent) variables (IV's) and the second set of data corresponds to measurements or dependent variables (DV's); in the second case, the different sets of data correspond to different sets of DV's.

One data set

Typically, the data tables to be analyzed are made of several measurements collected on a set of units (e.g., subjects). In general, the units are rows and the variables columns.

Interval or ratio level of measurement: principal component analysis (PCA)

This is the oldest and most versatile method. The goal of PCA is to decompose a data table with correlated measurements into a new set of uncorrelated (i.e., orthogonal) variables. These variables are called, depending upon the context, principal components, factors, eigenvectors, singular vectors, or loadings. Each unit is also assigned a set of scores which correspond to its projection on the components.

The results of the analysis are often presented with graphs plotting the projections of the units onto the components, and the loadings of the variables

[1] In: Lewis-Beck M., Bryman, A., Futing T. (Eds.) (2003). *Encyclopedia of Social Sciences Research Methods.* Thousand Oaks (CA): Sage.
Address correspondence to
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utdallas.edu/∼herve

(the so-called "circle of correlations"). The importance of each component is expressed by the variance (i.e., eigenvalue) of its projections or by the proportion of the variance explained. In this context, PCA is interpreted as an orthogonal decomposition of the variance (also called inertia) of a data table.

Nominal or ordinal level of measurement: correspondence analysis (CA), multiple correspondence analysis (MCA)

CA is a generalization of PCA to contingency tables. The factors of CA give an orthogonal decomposition of the Chi-square associated to the table. In CA, rows and columns of the table play a symmetric role and can be represented in the same plot. When several nominal variables are analyzed, CA is generalized as MCA. CA is also known as dual or optimal scaling or reciprocal averaging.

Similarity or distance: multidimensional scaling (MDS), additive tree, cluster analysis

These techniques are applied when the rows and the columns of the data table represent the same units and when the measure is a distance or a similarity. The goal of the analysis is to represent graphically these distances or similarities. MDS is used to represent the units as points on a map such that their Euclidean distances on the map approximate the original similarities (classic MDS, which is equivalent to PCA, is used for distances, nonmetric MDS for similarities). Additive tree analysis and cluster analysis are used to represent the units as "leaves" of a tree with the distance "on the tree" approximating the original distance or similiarity.

Two data sets, Case one: one independent variable set and one dependent variable set

Multiple linear regression analysis (MLR)

In MLR, several IV's (which are supposed to be fixed or equivalently are measured without error) are used to predict with a least square approach one DV. If the IV's are orthogonal, the problem reduces to a set of univariate regressions. When the IV's are correlated, their importance is estimated from the partial coefficient of correlation. An important problem arises when one of the IV's can be predicted from the other variables because the computations required by MLR can no longer be performed: This is called multicolinearity. Some possible solutions to this problem are described in the following section.

Regression with too many predictors and/or several dependent variables

Partial least square (PLS) regression (PLSR)

PLSR addresses the multicolinearity problem by computing latent vectors (akin to the components of PCA) which explains both the IV's and the DV's. This very versatile technique is used when the goal is to predict more than one DV. It combines features from PCA and MLR: The score of the units as well as the loadings of the variables can be plotted as in PCA, and the DV's can be estimated (with a confidence interval) as in MLR.

Principal component regression (PCR)

In PCR, the IV's are first submitted to a PCA and the scores of the units are then used as predictors in a standard MLR.

Ridge regression (RR)

RR accommodates the multicolinearity problem by adding a small constant (the ridge) to the diagonal of the correlation matrix. This makes the computation of the estimates for MLR possible.

Reduced rank regression (RRR) or redundancy analysis

In RRR, the DV's are first submitted to a PCA and the scores of the units are then used as DV's in a series of standard MLR's where the original IV's are used as predictors (a procedure akin to an inverse PCR).

Multivariate analysis of variance (MANOVA)

In MANOVA the IV's have the same structure as in a standard ANOVA, and are used to predict a set of DV's. MANOVA computes a series of ordered orthogonal linear combinations of the DV's (i.e., factors) with the constraint that the first factor generates the largest F if used in an ANOVA. The sampling distribution of this F is adjusted to take into account its construction.

Predicting a nominal variable: discriminant analysis (DA)

DA, which is mathematically equivalent to MANOVA, is used when a set of IV's are used to predict the group to which a given unit belongs (which is a nominal DV). It combines the IV's in order to create the largest F when the groups are used as a fixed factor in an ANOVA.

Fitting a model: confirmatory factor analysis (CFA)

In CFA, the researcher first generates one (or a few) model(s) of an underlying explanatory structure (i.e., a construct) which is often expressed as a graph. Then the correlations between the DV's are fitted to this structure. Models are evaluated by comparing how well they fit the data. Variations over CFA are called structural equation modelling (SEM), LISREL, or EQS.

Two (or more) data sets, Case two: two (or more) dependent variable sets

Canonical correlation analysis (CC)

CC combines the DV's to find pairs of new variables (called canonical variables, CV, one for each data table) which have the highest correlation. However, the CV's, even when highly correlated, do not necessarily explain a large portion of the variance of the original tables. This make the interpretation of the CV sometimes difficult, but CC is nonetheless an important theoretical tool because most multivariate techniques can be interpreted as a specific case of CC.

Multiple factor analysis (MFA)

MFA combines several data tables into one single analysis. The first step is to perform a PCA of each table. Then each data table is normalized by dividing all the entries of the table by the first eigenvalue of its PCA. This transformation—akin to the univariate Z-score—equalizes the weight of each table in the final solution and therefore makes possible the simultaneous analysis of several heterogenous data tables.

Multiple correspondence analysis (MCA)

MCA can be used to analyze several contingency tables; it generalizes CA.

Parafac and Tucker3

These techniques handle three-way data matrices by generalizing the PCA decomposition into scores and loadings in order to generates three matrices of loading (one for each dimension of the data). They differ by the constraints they impose on the decomposition (Tucker3 generates orthogonal loadings, Parafac does not).

Indscal

Indscal is used when each of several subjects generates a data matrix with the same units and the same variables for all the subjects. Indscal generates a common Euclidean solution (with dimensions) and expresses the differences between subjects as differences in the importance given to the common dimensions.

Statis

Statis is used when at least one dimension of the three-way table is common to all tables (e.g., same units measured on several occasions with different variables). The first step of the method performs a PCA of each table and generates a similarity table (i.e., cross-product) between the units for each table. The similarity tables are then combined by computing a cross-product matrix and performing its PCA (without centering). The loadings on the first component of this analysis are then used as weights to compute the *compromise data table* which is the weighted average of all the tables. The original table (and their units) are projected into the compromise space in order to explore their communalities and differences.

Procustean analysis (PA)

PA is used to compare distance tables obtained on the same objects. The first step is to represent the tables by MDS maps. Then PA finds a set of transformations that will make the position of the objects in both maps as close as possible (in the least square sense).

*

References

[1] Borg I., & Groenen P. (1997). *Modern multidimensional scaling.* New York: Springer-Verlag.

[2] Escofier, B., & Pagès, J. (1988). *Analyses factorielles multiples.* Paris: Dunod.

[3] Johnson R.A., & Wichern D.W. (2002). *Applied multivariate statistical analysis.* Upper Saddle River (NJ): Prentice-Hall.

[4] Naes T., & Risvik E. (Eds.) (1996). *Multivariate analysis of data in sensory science.* New York: Elsevier.

[5] Weller S.C., & Romney A.K., (1990). *Metric scaling: Correspondence analysis.* Thousand Oaks (CA): Sage.