MULTIVARIATE AND MULTIDIMENSIONAL ANALYSIS

K. Van Steen
Harvard School of Public Health, Boston, MA, U.S.A

G. Molenberghs
Limburgs Universitair Centrum, Hasselt University, Belgium

Keywords: Multivariate regression, variables relationships, Dimensionality reduction, multivariate normal distribution, principal component, canonical correlation, factor analysis, discriminant analysis

Contents

Summary

An overview is given of methods for Gaussian and non-Gaussian multivariate outcomes. The methods for Gaussian outcomes are, of course, rooted in the multivariate normal distribution. Apart from multivariate regression and multivariate analysis of variance, emphasis is placed on classical multivariate methods, such a principal component analysis, canonical correlation analysis, factor analysis, and discriminant analysis. In addition, more modern methods such as structural equation models, path analysis, and projection pursuit are given attention. A brief treatment is given of methods for non-Gaussian outcomes.

1. Introduction

Surveys typically involve costly techniques for data collection on many parameters in an attempt to deal with non-response. In some cases, the cost of collecting data on several (dependent) variables may be relatively small compared to the expensive implementation of multiple treatments within a clinical trials setting, giving rise to a wealth of data to be explored. In either situation, measurements are collected over several variables, quantitative (i.e., numerical) or qualitative (i.e., categorical). Analyzing these data using one or two variables at a time implies ignoring potentially valuable information. Whether data is collected in science, business or engineering, it is worthwhile to process all of the data simultaneously and to investigate interrelationships between variables. This is reflected in the multitude of publications on multivariate data analysis.
Typically, multivariate data are displayed in the form of a data matrix and statistical methods for analyzing and describing the data are simplified using matrix algebra formulations. Graphical displays of multivariate data are a handy tool to check model assumptions, or to detect overall patterns or interactions of three or more variables in the exploratory phase of the analysis.

Roughly speaking, multivariate data may be viewed as a collection of n points in some high p-dimensional space, where the points correspond to individuals or subjects and the axes (or dimensions) correspond to the measured variables. What most multivariate analyses seek to find is a low q-dimensional approximation of the original p-space in such a way that the new configuration of points has retained as much "information" as possible, different interpretations of "information" leading to different analyses techniques.

In this chapter, multivariate analysis refers to a set of techniques that allow the presence of more than one outcome variable. In the most general setting, the investigator has both a set of dependent variables (some of which are continuous, discrete, categorical, binary, …) and a set of independent variables (some of which are continuous, discrete, categorical, binary, …) at his/her disposal. Since it is infeasible to classify the multitude of available tools according to a fixed structure, we opt to select particular methods for common sub-problems of the general problem.

2. Continuous Outcomes

The multivariate normality of the distribution from which data are drawn plays a key role in multivariate statistics. It not only it offers an elegant framework, but the sampling distribution of a lot of statistics is in fact approximately multivariate normal. Therefore, we will mainly focus on multivariate settings with continuously distributed outcomes.

2.1 Multivariate Linear Regression

The multivariate linear model

$$\underset{n\times p}{Y} = \underset{n\times(q+1)}{X}\ \underset{(q+1)\times p}{\beta} + \underset{n\times p}{\varepsilon}$$

with

$$\underset{n\times p}{E(\varepsilon)} = \underset{n\times p}{0}$$

$$\text{Cov}(\text{vector}\varepsilon) = \underset{p\times p}{\Sigma} \otimes \underset{n\times n}{I_n}$$

is a natural extension of the univariate linear model with n individuals to the case where there are p dependent variables. The vector β is an unknown vector of parameter estimates and needs to be estimated. The covariance matrix Σ allows for different responses in the same individual to be correlated. The identity matrix I_n in the Kronecker product $\Sigma \otimes I_n$ expresses the assumption that the residuals of each p-variate observation are independent.

As soon as the interest lies in performing significance tests and deriving confidence intervals, distributional assumptions have to be imposed. In many cases, the assumption of multivariate normally distributed error terms is plausible. In this case, overall significance tests are based on two variance expressions

$$W = (Y - \hat{Y})^T (Y - \hat{Y})$$

and

$$T = \left(Y - \frac{1}{n} J_n Y \right)^T \left(Y - \frac{1}{n} J_n Y \right),$$

with $\hat{Y}$ the $n \times p$ matrix of predicted values and J_n an $n \times n$ matrix with all entries equal to 1. It can be shown that under the null hypothesis of no relationship, $\beta = 0$, T and W are both Wishart matrices derived from multivariate normal random variables with the same covariance matrix. Tests are derived by examining the eigenvalues of $T^{-1}W$ and include Wilks' criterion, Lawley-Hotelling trace, Pillai's trace and the minimum eigenvalue. The most commonly used test is based on Wilks' criterion and is a likelihood ratio test against the general alternative $\beta \neq 0$. As for all aforementioned statistics, its distribution can be approximated by an F statistic. When the sample size is large, χ^2 approximations hold.

Depending on the organization of response and/or covariate information, this simple model is able to answer many pragmatic questions, such as:

(i) Is there a difference in mean response between groups? If the covariates are quantitative binary, separating the data into groups a (multivariate) analysis of variance approach can be adopted.

(ii) How can groups be compared with post-test (dependent variables) and pre-test (independent variables) data? This problem may be solved by relying on a multivariate analysis of covariance.

(iii) What is the relationship between one set of variables and another set of variables, thereby generalizing the concept of correlation (between two variables) and multiple correlation (between a variable and a set of variables)? This is typically the question of interest in a canonical correlation analysis.

(iv) How can we adequately explain observed correlations in a data set? Answering this question is the goal of path analysis or structural equation analysis.

(v) How can the covariates be used to discriminate individuals or subjects in terms of response levels? Discriminant analysis may provide a useful tool.

(vi) How can we explain the variance-covariance structure present in the data, yet reduce the dimensionality of the data? This is one of the concerns of a principal component analysis.

(vii) How can we reduce the dimensionality of the data without destroying the original proximities? An appropriate configuration of the data is found via multidimensional scaling techniques.

(viii) Is there a way to detect features in the data by developing and inspecting visual displays? In the exploratory stage of a data analysis, a graphical analysis is useful to gain more insight into the data at hand.

(ix) Is there a geographical effect present in the data? Here, the focus of the analysis is to grasp the correlation structure between measurements from individuals or subjects that are part of the same geographical neighborhood. This issue is addressed in spatial analysis.

(x) Is there a serial correlation effect? If multiple measurements are taken over time, time trends can be estimated by properly accounting for the correlation structure between within-individual recordings. This is one of the focuses of a repeated measures analysis.

Although all of the research questions above reside in a multivariate or multidimensional framework, there clearly are many fundamental differences, thoroughly affecting the mode of analysis. This will be commented upon in the remainder of this chapter.

2.2 Multivariate Analysis of Variance and Covariance

Multivariate analysis of variance (MANOVA) offers a framework to compare two or more groups of subjects on several dependent variables simultaneously. This type of analysis can be carried out within the regression model as specified in Section 2.1, by describing group memberships via dummy coding.

The assumptions for MANOVA are a natural extension from the univariate case: (i) the observations on the dependent variables follow a multivariate normal distribution, (ii) the observations are independent and (iii) the population covariance matrices for the dependent variables are equal. Several studies report on how type I and type II errors are affected when either of these assumptions is violated. Typically, one is interested in testing the multivariate null hypothesis of equal population mean vectors. Mimicking a univariate analysis of variance approach, multivariate generalizations of the univariate within sum of squares, W, and between sum of squares, B, are introduced. Note that in the multivariate case both W and B involve cross-products, apart from sum of squares. They can easily be converted to numbers via the concept of a 'generalized inverse'.

Probably the most popular multivariate test statistic is Wilk's Λ:

$$\Lambda = \frac{|W|}{|T|}, \quad 0 \le \Lambda \le 1.$$

Here, the generalized variance is the determinant of $T = B + W$, where in T the observations in each group are deviated about the overall mean, for each variable. The sampling distribution of Wilk's Λ can be approximated by Bartlett's χ^2 for moderate to large sample sizes and by Rao's F for small sample sizes. In particular cases, it follows an exact F-distribution.

Omnibus F-tests for making multivariate comparisons of population means are often based on Hotelling T^2, after which univariate *t*-tests can be performed or Tukey simultaneous confidence intervals can be calculated to identify pairs of groups that may be responsible for overall significance. Hotelling T^2 is also appropriate for testing multivariate contrasts. This is particularly useful when there exists a theoretical or empirical basis for specific alleged group differences.

A multivariate analysis of covariance (MANCOVA) is also able to detect mean group differences on a set of variables, but differs from MANOVA in the sense that it allows controlling for one or more covariates that may otherwise bias the analysis results.

-
-
-

Bibliography

Chambers, J.M., Cleveland, W.S., Kleiner, B. and Tukey, P.A. (1983), *Graphical Methods for Data Analysis*, Belmont, CA: Wadsworth [This is one of many books expanding on graphical techniques for multivariate data]

Cox, T.F. and Cox, M.A.A. (1994), *Multidimensional Scaling*, Chapman & Hall [This book covers a variety of techniques with applications]

Cox, D.R. and Wermuth, N. (1996), *Multivariate Dependencies: Models, analysis and interpretation*, Chapman & Hall [This book on chain graphs explains how to use and to interpret them]

Fahrmeir, L. and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag [This is a book concerned with the use of generalized linear models for univariate and multivariate regression analysis]

Johnson, R.A. and Wichern, D.W. (1992), *Applied Multivariate Statistical Analysis*, Prentice Hall [Presents a multitude of concepts and methods of multivariate analysis]

Krzanowski, W.J. (1988), *Principles of Multivariate Analysis* - A User's Perspective, Oxford Science Publications [A user-friendly introduction to modern multivariate statistical analysis]

Krzanowski, W.J. and Marriott, F.H.C. (1994), *Multivariate Analysis* - Part I, Edward Arnold [An extensive overview of basic multivariate analyses techniques, including graphical display forms and inferential procedures]

Lauritzen, S.L. (1996), *Graphical Models*, Oxford University Press [Mathematical background to the theory of graphical models]

Seber, G.A.F. (1984), *Multivariate Observations*, Wiley series in probability and mathematical statistics [One of the earliest works]

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics*, Wiley [The first published book on graphical modeling]

Biographical Sketches

Geert Molenberghs is Professor of Biostatistics at the Limburgs Universitair Centrum in Belgium. He received the B.S. degree in mathematics (1988) and a Ph.D. in biostatistics (1993) from the Universiteit Antwerpen. Geert Molenberghs published methodological work on the analysis of non-response in clinical and epidemiological studies. He serves as an associate editor for Biostatistics and is Joint Editor of Applied Statistics. He is an officer of the Belgian Statistical Society and the Belgian Region of the International Biometric Society. He serves on the Executive Committee of the International Biometric Society. He has held visiting positions at the Harvard School of Public Health (Boston, MA). With Geert Verbeke, he is a co-author of books on longitudinal data.

Kristel Van Steen holds B.S. and Ph.D. degrees in mathematics from the University of Ghent (Belgium) and a Master of Science in Biostatistics degree from the Limburgs Universitair Centrum (Belgium). Her research interests focus on sensitivity analysis in incomplete data, pseudo-likelihood methodology, statistical methods for quality of life data, statistical genetics, and model selection. She is currently a postdoctoral research fellow at the Harvard School of Public Health, Boston, MA.