Series Editor's Introduction

Exploratory multivariate analysis (EMVA) is a very popular class of statistical techniques. Inventories of both mainframe and personal computing show that multiple regression (MR) and principal component analysis (PCA) are the techniques used most often, and that computing is usually done through packages such as *SAS, SPSS* or *BMDP*. The techniques discussed in this book extend standard EMVA methods in various ways. Starting point is the book by GIFI (1990), in fact Van de Geer's preface mentions explicitly that the book can be seen as a discussion of the *GIFI-system.* We must emphasize, however, that in many respects Van de Geer's book is more elegant than GIFI's. The GIFI book is huge and rambling, with large within-chapter variance, with many ideas that have not been properly worked out, a pillow filled with bricks. Van de Geer's book presents the basic GIFI techniques and approach at a uniform level, with a very strong emphasis on the linear geometry. This emphasis on the geometry is typical for Van de Geer's approach to EMVA. It can also be seen in his previous books, (such as Van de Geer, 1986). In the second volume the geometry is combined with discussion of the GIFI computational package.

Let us look a little bit more in detail into the question why this approach to EMVA is different from the standard treatment of, say, PCA and MR. It is most obviously different in its treatment of regression. Regression is

often presented in standard statistical texts as a method for fitting the so-called linear model. The linear model tells us that each of our observations is sampled from a normal distribution. The normal distributions for the various data points have the same variances, but different means, and the means are in a p-dimensional linear subspace of n-dimensional space. Making these assumptions leads to elegant algebra, and to elegant geometry, and the assumptions justify the use of MR in various ways. Nevertheless it should be emphasized that in all but a tiny number of situations, these assumptions are somewhere between far-fetched and preprosterous. Although regression is the workhorse of applied statistics, it has this status mostly because of its geometrical, computational and interpretational (causal) properties, not because of its statistical elegance. In this book multiple regression analysis disappears almost completely. It is a very special case of canonical analysis, in which we have two sets of variables, and one of the two sets consists of only one variable.

In the GIFI-system there is a much more central place for principal component analysis, although ultimately PCA is yet another special case of multi-set canonical analysis. PCA is usually presented somewhat differently from regression analysis. Basically, there are a number of algebraic optimality criteria that PCA satisfies. For instance, making a linear combination of the variables, with maximal variance. Or finding a low-rank approximation to the data matrix. These algebraic optimality criteria usually have straightforward geometrical interpretations, in terms of hyperplanes and ellipsoids in multidimensional space. The geometrical interpretations are emphasized throughout by Van de Geer, and in this book he extends his approach from linear MVA to nonlinear MVA.

In some multivariate analysis textbooks statisticians try to fit PCA into a model-fitting mould, in order to approach it with likelihood methods. This does not lead to anything useful, as far as I can see.

What is *nonlinear* MVA? There could be some misunderstandings here. The usual MVA techniques look for linear relationships between the variables, and because of this they work in linear spaces, and use the interpretational niceties that come with Euclidean geometry (hyperplanes, projections, ellipsoids). In modern computational statistics there are a number of techniques around that look for nonlinear relationships between variables. While this is interesting, it turns out that in practice this does not (yet) lead to useful approaches for routine analysis of large data sets. Too much has to be done by expert tuning, and too much of the computing is interactive. Between classical MVA and these advanced and rather delicate nonlinear techniques, we find MVA with *optimal scoring*. We continue to look for linear relationships, i.e., we continue to do linear

regression and component analysis, but we build transformations of the variables into the techniques. We optimize fit not only over the structural parameters of the linear part (the regression coefficients, the component loadings), but we also optimize over certain classes of transformations. This is the approach taken in the early optimal scaling work of De Leeuw, Young, and Takane (Young, 1981), and it is also the approach mostly taken by Van de Geer. One makes classical techniques more general by adding the optimal scaling options.

One of the key components of the GIFI system is consequently a system of optimal scaling methods. We need to define classes of transformations over which the loss functions are optimized. As a starting point we use the fact that variables are, in the last analysis, always discrete. This means that we can code them by using *indicator matrices,* elsewhere known as *dummies,* and impose various restrictions of an ordinal and numerical nature on the quantifications of the categories. This is explained in considerable detail in Volume 1 of this book.

The GIFI system, described in its final form by De Leeuw (1984) and De Leeuw and Van Rijckevorsel (1988), goes further than just generalizing classical techniques by giving them optimal scaling front-ends. It forces all classical multivariate techniques into a common framework by emphasizing a particular distance-based geometrical interpretation, and a corresponding distance-based loss function. The different MVA techniques are introduced as special cases, which impose particular sets of constraints on the basic optimization problem of minimizing the loss function. Van de Geer's book can be thought of as an introduction to this more general approach. His own personal emphasis on the linear geometry of MVA pervades the whole book, and makes it a special contribution to the literature.

JAN DE LEEUW
SERIES EDITOR