

FOREWORD

This book by Gifi is a nonstandard treatment of some of the more important parts of multivariate data analysis. A heavy emphasis has been placed on plotting observations and/or variables in low-dimensional Euclidean space. This clearly shows its psychometric origins, because of the obvious relationships with factor analysis and component analysis. Much less emphasis than usual is paid to regression and analysis of variance, and the various multivariate generalizations of the linear model.

There are other ways in which the book is nonstandard. It is organized, as many other multivariate data analysis books are, around various key techniques, such as principal component analysis, canonical analysis, and so on, but in this book these techniques are identified, much more than is usually done, with computer programs and the type of output they reproduce. A technique is defined, more or less, as a tool into which you feed data of a particular type and format, and which then reproduces output of a particular type and format. This very instrumental interpretation of data analysis is contrasted, in many places, with the model-oriented approach of classical statistics. Models do have a place in Gifi's philosophy: not as tentative approximations to the truth, but as devices to sharpen and standardize the data analytic techniques. At the same time the inferential methods of statistics are explicitly introduced as one of various ways in which the stability of data analysis tools can be investigated.

Of course we have to realize that there is, according to many statisticians, one appropriate way to analyse data: first formulate a model on the basis of prior knowledge, then compute the likelihood of the data given the model, and then estimate and test the appropriateness of the model by maximizing the likelihood function. There are some variations on this scheme, depending on what exactly this prior knowledge includes, but these variations are minor compared to the prescription seemingly advocated by Gifi. First choose a technique (implemented in a computer program) on the basis of the format of your data, then apply this technique, and study the output. The prior knowledge is not mentioned explicitly, but more importantly it is not isolated in any one particular time and place in the data analysis process. The classical recipe suggests that prior knowledge (theory, prejudice, experience, expecta-

tions) enters in the formulation of the model, but then it should not be used at all during the other stages of the process. These are supposed to be mechanical, almost as applying the gigantic computer program MAXLIK, which handles data of any type, or the even more gigantic program TBAYES, which takes data and prior information as input and transforms this to posterior information. In the book by Gifi many other places where prior information, and expertise, play a role in data analysis are indicated. Not all thought and creativity of the investigator, and the history of his science, is concentrated in the mysterious phase in which the model is formulated.

The techniques actually discussed in the book have a very strong emphasis on geometry, in the sense that they make pictures of data. As we said, this places the techniques rather firmly in the tradition of factor analysis, or perhaps even more clearly in the tradition of multidimensional scaling. Another relationship with multidimensional scaling is that variables are transformed optimally, using monotone transformations or splines, in order to improve the fit of the low-dimensional representation. This is different from traditional factor analysis, in which the fit is improved by computing additional factors and then rotating them to simple structures. Gifi incorporates the additional factors in the nonlinear transformations, and keeps the pictures simple and preferably in two or at most three dimensions. The pictures are, of course, descriptive devices. In most cases they have to represent an enormous amount of information, and they can approximate that by showing the dominant relations in the data, its most coarse features. This may be useless, or even offensive, to people who want specific answers to specific questions. Again the psychometric (or social science) origins of Gifi's system become apparent here, because in the large scale survey type of investigations in which multivariate methods are likely to be used the questions are usually not very specific.

Gifi was one of the first investigators to use the bootstrap and jackknife to study the stability of multivariate representations. The reasons for this are clear: the amount of prior knowledge one needs to use these techniques is minimal, especially if one uses the perturbation interpretation discussed in Chapter 12 of the book. If we use the idea that a statistical technique is a tool, that can be applied in many different situations, although it is not necessarily optimal for any one specific situation, then this fits in rather nicely with the ideas behind bootstrapping, jackknifing, cross validating, and using randomization tests. The tools have their own quality control devices built in, as it were, and because of that they become more useful and more informative at relatively little extra cost.

The application of a computer program, or a set of related computer programs, as a statistical toolbox around which various applications can be

organized, is much more common now than it was during the Gifi project. We now have books built around GLIM, BMDP, the DataDesk, SPSS, and so on. The emphasis on computations to replace assumptions, and the use of resampling and Monte Carlo in that context, is very important in modern statistics. Optimal transformations occur with great regularity in the statistical journals these days, in combination with semi-parametric models, splines, smoothing, and Box-Cox families. Correspondence analysis (sometimes disguised as ACE) can now be discussed among consenting statisticians, without causing frothing at the mouth and the throwing of stones. It was not always like this! The Gifi project started as a one-person effort in 1968. It became a more-than-one-person effort in 1974 and a quite-a-lot-of-persons effort in 1978. It was generously supported by ZWO, the Netherlands Organization for the Advancement of Pure Research, but was frowned upon by official statisticians, and led to the types of semi-religious debates that typically occur in the foundations of statistics area. That this resistance was eventually largely overcome was not in the least due to internal developments in the discipline of statistics itself.

There are a few persons I would like to thank personally, on behalf of Albert Gifi, of course: in the first place John Van de Geer, who taught us the geometric approach to multivariate analysis that pervades the book, who provided us with a Department to work in, and with the freedom we needed; secondly, Doug Carroll and Joe Kruskal at Bell Laboratories in Murray Hill, who invited several of us to come over there and who provided access to many resources, tools and organizations; thirdly, Forrest Young and Yoshio Takane, who provided invaluable assistance at a very critical point in the development of these techniques; fourthly, Chris Haveman, of SPSS, who remembered from his days at the Computing Center of the University of Leiden that people there were doing useful things in multivariate data analysis; and finally, Richard Gill, formerly of the Center for Mathematics and Informatics in Amsterdam, who provided valuable support and connections to the world of academic statistics. There are innumerable other persons who provided various services at various points in time, and I would like to thank them collectively as well.

As explained elsewhere in the book, Gifi is not quite dead yet. Many of the topics mentioned in the book, but not worked out in detail there, have been taken up in dissertations and research monographs. The basic approach has been extended to path models and dynamic systems, and variations based on multinormal likelihood theory are also available. Links to multidimensional scaling have been strengthened, as have the links to classical multivariate statistical theory. However, the book, this book, is the only comprehensive,

programmatic, encompassing statement of the general Gifi philosophy, and the way in which it is implemented on the digital computer.

Los Angeles, 22 April, 1989

Jan de Leeuw