Reports (1–4 double-spaced typewritten pages) should describe tools, products, and/or services found useful by practicing evaluators. Reports should also include the context of use and sufficient information for interested readers to contact the user directly for more detailed information, that is, name, address, and telephone number. Books, computer software and hardware, and new techniques — or old ones used in new ways — are just a few of the items that may be described.

Nonlinear Multivariate Analysis

JAN DE LEEUW

[**Editors' note:** Dr. de Leeuw was invited to provide a short overview of the methodology that he and his colleagues have developed for analyzing large-scale data bases.]

In evaluation research and survey analysis, multivariate analysis procedures applied to large data sets are very important. In the physical and biological sciences we are often interested in relating a very small number of variables, with relatively low measurement errors, using specific models based on explicit prior knowledge about the functional form of the relationship. This makes it possible to fit models to small amounts of data and still find stable relationships. If we do not have prior knowledge and if the level of measurement error is high, we have to increase the amounts of data by taking larger samples and/or by analyzing more variables. Increasing the number of observations decreases the sampling errors and increasing the number of variables hopefully decreases the measurement errors.

There are two basic problems with applying standard multivariate analysis techniques to large surveys or panel studies. In the first place, the amount of computation involved can be prohibitive, at least with the resources that are often available. Not everybody has access to a Cray supercomputer — sometimes a PC or a Mac is all that is available. However, this problem is admittedly less important than it used to be. The second problem is that the assumptions needed to interpret classical multivariate analysis techniques are usually quite inappropriate. Often the techniques are designed for numerical variables, usually linearity of regression is required, and for

Evaluation Practice, Vol. 11, No. 2, 1990, pp. 155-157
ISSN: 0191-8036

most of the inferential statements we even need the assumptions of multivariate normality. All these assumptions are routinely violated in survey and evaluation research, and often they do not even make sense.

There are two standard solutions to the second problem. The first one is to use ad hoc procedures. Replace "disagree completely" by -2, "disagree" by $-1, \ldots$, "agree completely" by $+2$, and apply the techniques for numerical variables such as principal component analysis and regression. This is highly arbitrary, and it gives (or should give) the investigator a bad conscience. The second solution is to apply nonparametric techniques, such as log-linear modelling, to small subsets of the variables. This, however, can throw away the baby with the bath water, because we give up our idea to get measurement stability by using multiple indicators for latent constructs. There is a third solution available, however, which tries to avoid both the inflicting of a bad conscience and the murder of the baby.

In 1968 I started the Nonlinear Multivariate Analysis project at the University of Leiden. About 30 people have been working on this project, about 15 doctoral dissertations have been written, and about 25 large computer programs have been produced. This year the book *Nonlinear Multivariate Analysis*, written under the collective pseudonym Albert Gifi, has appeared in the Wiley Series on Applied Probability and Statistics. In 1987, the book *Progress in Component and Correspondence Analysis*, edited by my colleague Jan van Rijckevorsel and myself, appeared in the same series.

Four of our main computer programs (ANACOR for correspondence analysis, HOMALS for multiple correspondence analysis, PRINCALS for nonlinear principal components analysis, and OVERALS for nonlinear, multi-set, canonical correlation analysis) were brought out by SPSS, in the module CATEGORIES. Somewhat earlier the programs PRINQUAL, TRANSREG, ADDALS, and CORAN, which were developed by SAS and the University of North Carolina, Chapel Hill, in close cooperation with the Department of Data Theory at the University of Leiden, were brought out as additional user programs by SAS. Related software has appeared in BMDP (the CA module for correspondence analysis) and in a public domain package from Jerome Freidman, Department of Statistics, Stanford University. Programs for nonlinear multiple regression and nonlinear principal component analysis are also available in the BLSS statistics package from the University of California, Berkeley.

The basic ideas behind the nonlinear multivariate analysis techniques can be briefly explained. What the investigator needs, besides data, is a question, often based on a theory which can be fairly vague, e.g., How well can we predict income (or some transformation of income) from these background variables? Can these 45 variables (perhaps after transformation) be interpreted as measurement of one latent construct? Does this path model describe the relationships between my variables (or their transformations) reasonably well? In all these questions the notion of transformation (or, in the case of nominal and ordinal variables, of quantification) is important. We do not merely fit the model over the structural parameters (regression coefficients, path coefficients, component loadings), but we also find *optimal transformations* which make the fit of the model as good as possible. These optimal transformations are often restricted in that we require them to be monotone, or smooth, or a polynomial, or a spline. But the basic idea is to maximize the multiple correlation, the largest eigenvalues, the canonical correlations, and so on, over both parameters and transformations. This is done by optimized, alternating, least- squares algorithms so that the programs can deal with really large data sets.

This note is to give an idea of what these new techniques are about, what problems they are designed for, and what is available in terms of software and books. Not all of the programs are in SPSS or SAS; there are programs for nonlinear dynamic systems, path analysis, discriminant analysis and factor analysis which are available in FORTRAN source code or, with minimal interfaces, for the PC. The most important programs have been bundled, with a graphics user interface, for the Mac. There are other variations including programs which use nonlinear maximum likelihood instead of least squares.

For more information, contact Jan de Leeuw, Department of Mathematics, UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024-1555; or Ita G.G. Kreft, Chrystal Clear Statistical Consulting, 2874 Nicada Drive, Los Angeles, CA 90077.