Abstracts of the Annual Conference of Finnish Statisticians 2007, "Multivariate Methods".
Eds. J. Niemelä, S. Puntanen, and E. P. Liski. Dept. of Mathematics, Statistics and Philosophy,
University of Tampere, 2007, ISBN 978-951-44-6957-2, pages 12–16

40 years of multivariate analysis

Lauri Tarkkonen
Department of Mathematics and Statistics
P.O. Box 54
FI-00014 University of Helsinki, FINLAND
Email: lauri.tarkkonen@helsinki.fi

1 Introduction

The basic theories of multivariate statistical analysis has been available for more than sixty years but the applications have been possible for some 40 years when the computing power of the mainframe computers become available to most scientist in behavioral sciences, medicine and economics.

In Finland the pioneers of multivariate analysis are Toivo Vahervuo and Yrjö Ahmavaara, whose book Johdatus faktorianalyysiin (1958) (Introduction to Factor Analysis), was preceding the actual computerized era and was dealing with factor analysis only. Their influence was more in the field of psychology, and first in the early sixties Seppo Mustonen brought the whole repertoire of multivariate analysis to the statisticians in Finland. Seppo Mustonen's contribution to the field did not restrict to teaching the theory for most of the future professors in Statistics, but he had a strong influence in the computational development of the multivariate analysis, starting with the Survo packages for mainframes in the sixties, through some very innovative programs for the Wang minicomputers and today the whole statistical problem solving and publishing system on today's PCs. I would like to characterize today's SURVO as the only publishing program that can perform a factor analysis.

2 Multivariate or not?

In the late sixties when I started my career at the university, the multivariate analysis was something that could separate the real experts from some ordinary statisticians. I became a statistical consultant and got the best view of what happened on the applications field. Because the psychology had a head start, they were using factor analysis for about everything. If you did not use factor analysis in your work, you were just ordinary.

The analysis was performed in the computing center by mathematicians or statisticians and they knew very little about the substance, but they knew most of

what was necessary for squeezing the data into the computers and out came the printout, where the most important text was: Normal termination.

The statisticians learned their multivariate analysis reading the classic T.W. Anderson: Introduction to multivariate analysis. There the main method was Principal Components analysis, the factor analysis was somewhere in the appendix. Even though the two analyses were very different in the theoretical sense, factor analysis model (1) and the principal components model (2):

$$x = Af + e, \tag{1}$$

$$u_i = b_i' x \text{ and } \sigma_u^2 \text{ maximized by } b_i, \text{ and } u_i \text{ orthogonal to } u_j. \tag{2}$$

They lead to very similar computational formulas:

$$((\Sigma - \Psi) - \lambda I)a = 0, \tag{1}$$

$$(\Sigma - \lambda I)a = 0. \tag{2}$$

The only difference being the estimation of communalities, by subtracting estimates Ψ from the estimated Σ. The scaling by the diagonal of Σ came much later with the development of the Maximum Likelihood Factor Analysis.

One of the key issues was the rotation of the factors. The direction of the principal components was given by the maximization of their variances, but the factors did not have this kind of restriction, the idea was to rotate them to a direction giving them a meaningful interpretation.

Here was one of the crucial problems in the early days. The ones who performed the analysis did not know what the meaningful direction was and the scientist having that information was not allowed near the computers.

In the early days, with manual calculation the rotation was done manually as well. The variables were drawn on millimeter plotting paper as vectors and the factors were drawn on some transparent paper and the factors were rotated manually and visually. Because the know how was not available for the ones performing the computations, it was necessary to develop analytical rotation methods, the principles were taken from Thurstone's simple structure criteria. The most common analytic rotation method was the "Varimax", maximizing the variances of the factor loadings, thus having mainly zeros and ones, thus serving well the Thurstone criteria. Here one has to remind everybody about Touko Markkanen's cosine rotation, there some variables as orthogonal as possible were given as the target for the rotation. This gave some means to the pursuit of meaningful factors. Of course this sound method did not get strong international approval.

3 Exploratory vs. confirmatory

The use of almost any multivariate analysis was purely exploratory, all sorts of variables were collected and joined in the same data and then the magic wand (the factor analysis) transformed this into a theory. The problem was that almost any set of variables having correlations of different magnitude gave some kind of factor structure. So suddenly finding a factor structure was not anymore enough.

In the mid seventies Lawley and Jöreskog made some fundamental development in the theory of factor analysis. They developed the Maximum Likelihood estimators for the factor loadings. Suddenly the ugly duckling became a swan. The method that most real statisticians regarded just as a gimmick used by psychologists that actually had no theoretical base, became a bona fide statistical model and the previous king, principal components analysis became just a mathematical way of transforming a number of correlating variables to a minimum set of orthogonal variables, useful for example as predictors in some regression models.

In many fields of application this change went unnoticed, as they made reference to previous work in their field and they did not bother to learn the new theory of multivariate analysis. Some prominent statisticians paid attention, for example the previously referred T.W. Anderson rewrote his book and factor analysis climbed from the appendix to the company of real multivariate models.

Some software writers have not seen this change. SPSS has a module called Factor Analysis, and the default values perform you Principal Components. So a big number of innocent users do not know the difference. They trust the programmers know. In the same time when the library programs were developed bigger and more comprehensive the programmers tried to help the users, as the only acceptable slogan was: Easy to use. You do not have to learn anything. Even though research was supposed to be a handicraft, where every problem deserved specific solutions tailor-made for the research problem in hand. When you try to discover new facts, you sometimes need new way of using your tools.

K.G. Jöreskog made full use of the development of the Maximum Likelihood estimation of the Factor Analysis. Because now it was possible to get some tests for the goodness of fit of the statistical multivariate models, he expanded the idea of multiple equation models, known in econometrics, by bringing latent factors into the picture and came up with the Structural Equation Models with Likelihood estimation. His first program for estimation of the parameters of these models was called LISREL and for years that was a synonym for SEM, Structural equation models. Jöreskog's model was actually only a factor analysis model, where the direction of the factors was more or less fixed by the user.

Now the factor analysis had shaken the worst insults, that it is not possible to falsify the model, so it could not be used as any kind of a proof for a theory. In the late seventies and early eighties the LISREL models were greeted with real enthusiasm, they were supposed to solve all the problems in behavioral science research. The biggest thing was that now the scientist were exploring whole structures instead of single parameters.

As always, big enthusiasm lasted some years and then the reality struck back. This was too complicated for the scientists without very good training in statistics and use of computers and the LISREL models lost it's appeal. Then some seven to ten years later they came back, now the users had done more of their homework and new programs like Amos made the application easier.

4 Linear vs. nonlinear

The early years of multivariate analysis was based on covariance structures and correlations, meaning that only linear relationships between variables were relevant. Of course some relationships could be made linear by suitable transformations of the data. But the life is not linear and all variables are not quantitative, but there are plenty of qualitative variables and even in some cases when the variable seems to be perfectly quantitative the changes of variable values from one to the next is more qualitative than quantitative. The effect of the new child to a family life and consumption patterns is much bigger when the amount of children is changing from 0 to 1 than from 1 to 2 or from 5 to 6. Still in most cases a variable "number of children in the family" is used as a bona fide quantitative variable.

To be able to deal with more complicated relationships the Euclidean metric defined by the correlations or covariances must be replaced by some more complicated metrics. One way to deal with this problem is to use a metric based on χ^2.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \left(\frac{f_{ij} - f_{i.} f_{.j}/n}{f_{i.} f_{.j}/n} \right)$$

The term in the brackets $\left(\frac{f_{ij} - f_{i.} f_{.j}/n}{f_{i.} f_{.j}/n} \right)$ is called the contribution to χ^2. High values tell about strong tendency to reject the independence hypothesis. A singular value decomposition of the matrix of contributions, is decomposing the dependencies expressed by the contributions.

Sir R.A. Fisher has brought up in the thirties a method of displaying multidimensional structure of a two-dimensional frequency table and then it was found out that the Burt table, where one could say, that the variances were replaced on the main diagonal by the frequency distribution and the covariances of two variables was replaced by the two dimensional frequency table, thus one number the correlation was replaced by a two dimensional frequency distribution, giving much more information about the relationships that ever could be stored in one number only.

The French developed the correspondence analysis, the main contributor being Ludovic Lebart. In the late eighties the correspondence analysis returned from exile back to the English speaking statistical world.

5 Future

The future of the multivariate analysis is depending on the other hand the development of the programs and the development of the users professional skill and ambitions. The relationships will be more and more complicated, but the users must know more of the models they want. At the moment the users are given standard tools, it is like trying to feed everybody with the same factory made meatballs, even though some customers would like to have gourmet meals.

A good example in this is the development of the factor analysis is the big statistical program systems, SAS and SPSS. You get actually principal components,

with some measurement error variances swept under the rug. On the other hand Survo offers you a graphical rotation, allowing you to get exactly the factors you want or if you do not want to rotate by hand, you can give auxiliary variables, to guide the direction of the factors, so that they have exactly the taste you want to have. Not just something a programmer feels you should have.