THE ROLE OF MODELS IN PRINCIPAL COMPONENT AND FACTOR ANALYSIS

Jan De Leeuw
University of Leiden

Introduction

The paper by Caussinus is very useful, because it raises questions such as 'why models ?' and 'why statistics ?' in a rather direct way. In this sense the paper is different from the Takane and Shibayama paper, which accepts the traditional statistical framework and the corollary that maximum likelihood is better than least squares. The methodological problems are much more in the open here. This is due partly to the fact that Caussinus writes about principal component analysis (PCA), which is a technique used mostly in non-statistical contexts, and partly to the fact that Caussinus is French. Benzécri's first principle (1973, p. 3) is that data analysis is basically applied linear algebra and geometry, and not applied probability. And even if we consider probability theory as basic, as is done in at least some French data analysis, often not much attention is paid to the models of inferential statistics. Thus Caussinus' paper gives us a nice opportunity to compare the French data analysis with the Anglo-American inferential statistics approach.

The general question Caussinus discusses, briefly in this paper, and somewhat more extensively in Caussinus (1986), is the role of models in data analysis. He does this in the context of PCA and factor analysis (FA), which are, of course, the corner stones of psychometrics. In this discussion I first formulate my own opinion about the role of models, which does not differ greatly from that of Caussinus. After this I discuss a general class of models for PCA/FA, from which Caussinus has made a somewhat arbitrary choice.

Why models ?

Models are used for data reduction. Why is it useful to reduce the amount of data ? There are various reasons mentioned in the literature. In the first place *communication.*

Replacing a large amount of data by a smaller amount is useful for reporting. In the second place *interpretation*. Raw data can often be replaced by summary statistics which are easier to comprehend. In the third place *prediction*. If we want to predict, then we need a smooth versions of the data. Extrapolation from irregular data points is very risky. Thus we can also mention *smoothing* as a possible use of modelling. Closely related is the fact that modeling will often lead to a more *stable* representation of the data. Finally it is often said that models form the *link between theory and data*. This is perhaps true in some cases, but there are many social science situations in which there is not enough theory to make this plausible. Or in which the theory is not specific enough to define the model unambiguously.

Thus we can say that the models usually studied in data analysis give low-dimensional representations of the data, which are easier to communicate and interprete, and which are perhaps more stable and smooth than the original data. Models are, more or less by definition, never true, and always approximations.

Why statistics ?

Models have a *structural* or *systematic* component, and an *error* or *disturbance* component. **Data = Structure + Error**, or, in Tukey's terminology, **Data = Smooth + Rough**. Data analysis concentrates on modelling the structural component, statistics pays a lot of attention to modelling the errors. We have seen why modelling the structural components is useful. It is used, essentially, to get rid of the errors, which are non-informative anyway. Why do we also want to model the errors ? This is because an error model makes it possible to indicate how we can get maximum stability for a given smoothness.

An important complication in using the error models of statistics is that *statistical models are not about the actual data* .The statements of inferential statistics are statements about the *framework of replication* , about what will happen if we repeat our experiment. We do not actually have to repeat it, in fact the model takes over the burden from the experimenter. Thus we have to be satisfied with the fact that the error models discussed in statistics are all framed in terms of random variables, and are consequently all about the replication framework. They cannot be falsified without getting involved in an infinite regress (Hartigan, 1983, section 1.5, De Leeuw, 1984), and their link with actual data and actual data analysis is problematical.

Which models ?

Caussinus discusses three models for PCA and FA, a selection which is a bit arbitrary. Moreover it is not very obvious from his discussion that his models differ greatly in their degree of specificity, and that consequently he is perhaps using the word 'model' in at least two different meanings. His Model I is a simple data reduction technique, in which a function of two variables is approximated by a finite sum of products of functions of one variable. In formula: we want

$$\|x(\omega,t) - \Sigma_p \alpha_p(\omega)\beta_p(t)\|^2$$

to be as small as possible, with $\|.\|$ some Hilbert space norm. If it is the norm of the direct product of the space of individuals (indexed by ω) and the space of variables or time-points (indexed by t), then the duality diagram of Cailliez and Pagès (1976), discussed recently in Psychometrika by Tenenhaus and Young (1985), applies. If the number of individuals and variables is finite, then we can use matrices. It is clear that Model I only models the structure, not the errors, although of course choice of the norm implies a certain weighting of the errors.

In the case in which we deal with random variables it is often more natural to use the notation

$$E\{\|\underline{x}(t) - \Sigma_p \underline{\alpha}_p\beta_p(t)\|^2\}$$

for the loss function (we use the Dutch convention of underlining random variables). Again no model for errors is involved, we merely solve an approximation problem. The expectation is, of course, with respect to the probability defined on the carrier space. This probability can be defined theoretically. We can assume, for example, that it is Gaussian. No data are involved. The probability can also be 'empirical', however, by which we mean that it is the average of, say, n indicators. If we make assumptions about these indicators, for example that they are independent and identically distributed, then we are actually in the context of Model II of Caussinus. We have n observations (sample paths) $\underline{x}_i(t)$, which are iid. This could be called a statistical model, although a rather weak one. We can strengthen it by assuming normality, as in most of classical multivariate analysis.

PCA/FA models

The remaining models are perhaps best described as FA models, not as PCA models. This conforms with current practice. We limit ourselves to n observations on m variables, and we study the very general model

$$\underline{x}_{ij} = \tau + \underline{\lambda}_i + \underline{\mu}_j + \Sigma_p \underline{\alpha}_{ip}\underline{\beta}_{jp} + \underline{\varepsilon}_{ij}.$$

in which all quantities are random variables (underlined), defined on the same probability space. Compare De Leeuw (1973) for some results on this general model. There are at least two important special cases which are of interest, both because of Caussinus' paper and because of historical reasons. If τ and the $\underline{\lambda}_i$ are identically zero, if the $\underline{\mu}_j$ are constants (have zero variance), if the $\underline{\beta}_{jp}$ are constants, if the $\underline{\alpha}_{ip}$ and $\underline{\varepsilon}_{ij}$ all have zero expectation and are all independent of each other, then the model can be written in terms of m-vectors as

$$\underline{x}_i = \mu + B\underline{\alpha}_i + \underline{\varepsilon}_i.$$

Assuming unit variance of the $\underline{a}_{ip}$ gives the dispersion

$$E\{(\underline{x}_i - \mu)(\underline{x}_i - \mu)'\} = BB' + \Delta^2,$$

with $\Delta^2 = E(\underline{\varepsilon}_i\underline{\varepsilon}_i')$. This is the (random factor score) factor analysis model, associated with the names of Lawley, Bartlett, and Joreskøg. If we assume, in addtion, that $\Delta^2 = \delta^2 I$, then multinormal maximum likelihood estimates of the structural parameters β_{jp} and δ^2 can be computed from a PCA.

The random factor score model with equal uniquenesses is not mentioned by Caussinus, but another special case is. If we make the same assumptions as above, except for the α_{ip} which are now assumed to be constants, then we have the fixed factor score model. It can be written, in the special case in which also the μ_j are zero, as $\underline{X} = AB' + \underline{E}$. Elements of $\underline{E}$ have zero expectation, are all independent, and satisfy $E(\underline{\varepsilon}_{ij}^2) = \delta_j^2$. If all δ_j^2 are equal, then we have the fixed factor score model with equal uniquenesses, and this is Model III of Caussinus.

It seems to me that these models are best discussed in terms of the general model above, they are closer to FA than to PCA. The difference between the fixed and random score model must be discussed in terms of the replication framework they use. Of course this repliacation framework is almost always implicit. It seems to me (and to Caussinus) that the fixed score model is more realistic. This is despite the fact that in current statistical practice the random score model is far more popular.

Choice of metric

As Caussinus indicates the choice of metric (or metrics) in which we perform a PCA is mainly relevant for the descriptive versions of the technique (his Model I). For the statistical versions the choice of the metric is dictated by the model. Of course this is only an advantage if our confidence in the error part of the model is great. If we are merely imposing prior knowledge for mathematical convenience, it may very well be a disadvantage. In the 'early' or 'exploratory' stages of data analysis we need flexibility. It seems to me that the spirit of French data analysis, and also of Caussinus' paper, is that the statistical models must be regarded as instances of leading cases (Mallows and Tukey, 1982) or gauges (Gifi, 1981) for the general technique PCA, which is regarded as more fundamental than each of the models. Also compare Caussinus (1986).

Benzécri's second principle of data analysis (1973, p. 6) is that models must be derived from the data, not the other way around. This is also the point of view taken by mathematical physicists such as Kalman (1983) and Willems (1986). Imposing a great deal of prior structure, which cannot be refuted, is dangerous in areas in which there is little prior information.

References

Benzécri, J.-P., et al. (1973). **L'Analyse des Données , II. L'Analyse des Correspondances.** Paris: Dunod.

Cailliez, F., & Pagès, J.-P. (1976). **Introduction à l'Analyse des Données.** Paris: SMASH.

Caussinus, H. (1986). Quelques Réflexions sur la Part des Modèles Probabilistes en Analyse des Données. In E. Diday et al. (eds), **Data Analysis and Informatics IV.** Amsterdam: North Holland Publishing Co.

De Leeuw, J. (1973). **A Generalization of the Young-Whittle Model.** Report RB 006-73. Department of Data Theory, University of Leiden.

De Leeuw, J. (1984). Models of Data. **Kwantitatieve Methoden** , 5, 17-30.

Gifi, A. (1981). **Nonlinear Multivariate Analysis.** Department of Data Theory, University of Leiden.

Hartigan, J. (1983). **Bayes Methods.** Berlin: Springer.

Kalman, R.E. (1983). Identifiability and Modeling in Econometrics. In P.R. Krishnaiah (ed.), **Developments in Statistics.** New York: Academic Press.

Mallows, C.L., & Tukey, J.W. (1982). An Overview of Techniques of Data Analysis Emphasizing its Exploratory Aspects. In J. Tiago de Oliveira et al. (eds.), **Some Recent Advances in Statistics.** New York: Academic Press.

Tenenhaus, M, & Young, F.W. (1985). An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis, and Other Methods for Quantifying Multivariate Data. **Psychometrika,** 50, 91-119.

Willems, J.C. (1986). **From Time Series to Linear System.** Mathematics Institute, University of Groningen.