132

MODELS OF DATA[*]

Jan de Leeuw[**]

Summary

We point out that models based on probability theory, and the statistical techniques derived from them, have limited applicability, at least in exploratory multivariate situations. Prior knowledge, if available, must be incorporated into the analysis to yield greater stability. If prior knowledge is not available, however, it must not be invented. This applies to both the structural and the replication-framework aspects of a model. The methods of gauging and stability analysis are introduced as alternatives. They use the notion of a technique as the pivot of data analysis, not that of a model. Homogeneity analysis is used as an example.

[*] Paper presented at the symposium 'Foundations of Statistics', organized by SMS/VVS, Utrecht, october 25, 1983.

[**] Department of Data Theory FSW/RUL, Middelstegracht 4, 2312 TW Leiden, tel. 071-148333-2255.

1 A common statistical model

Multivariate statistical analysis is usually based on the following model. A sequence $\underline{x}_i$ of independent, identically distributed, discrete random vectors is studied. The $\underline{x}_i$ assume values x_j with probabilities π_j, where j=l,...,m. The vector π, containing the π_j, lies in S^{m-1}, the unit simplex of R^m. In fact in most cases there is prior knowledge of the form $\pi \varepsilon \Omega$, with Ω a known subset of S^{m-1}.

Classical multivariate analysis is based on the multinormal distribution, which entails that it assumes the the $\underline{x}_i$ are continuous random vectors. Observed random variables will however necessarily be discrete (due to rounding), and continious models will always be approximations to discrete models. Thus we effectively assume without any loss of generality that the $\underline{x}_i$ are discrete (De Leeuw, 1983). In multinormal analysis the $\underline{x}_i$ are discretizations of multinormal variables. In multinomial analysis (often log-linear analysis) the set Ω is often defined in terms of (conditional) independence and (order of) interaction. Both the assumptions of independence and of identical distributions are serious restrictions of generality, but they are used because they define the simplest case. Also observe that Ω serves as prior knowledge. Thus our prior knowledge is in the form of set membership and not in the form of a prior distribution on S^{m-1}.

2 Usual statistical analysis

The techniques of classical multinormal analysis and of log-linear analysis are all based on the theory of best asymptotically normal estimation. This is a very restricted, but at the same time very convenient form of general (local asymptotically minimax) large sample theory. We give a brief outline. More information can be found in the recent paper by De Leeuw (1983). We estimate the vector π by using estimates of the form $\underline{\pi}_n = \Phi(p_n)$. Here $\underline{p}_n$ is the observed m-vector of proportions in the first n trials, the discrete version of the empirical distribution function. The function Φ maps S^{m-1} into Ω, is assumed to be differentiable, and is assumed to satisfy Fisher-consistency. In this context Fisher-consistency is $\Phi(p) = p$ for all $p \varepsilon \Omega$. This implies that $\underline{\pi}_n$ converges in probability to π. Differentiability implies that $\underline{\pi}_n$ is asymptotically normal. Fisher-consistency also implies a lower bound on the dispersion matrix of the asymptotic distribution. If this bound is attained, the estimate is called best asymptotically normal. Best asymptotically normal estimates can be computed by minimizing statistics such as $n(p_n - \pi)'\underline{P}_n^{-1}(\underline{p}_n \geq \pi)$ over $\pi \varepsilon \Omega$. Here $\underline{P}_n$ is any consistent estimate of Π, which is the diagonal matrix with the elements of π on the diagonal. If we use $\underline{P}_n = \Pi$ and minimize $n(\underline{p}_n - \pi)'\Pi^{-1}(\underline{p}_n - \pi)$ over π then we use minimum chi-quared estimation. If $\underline{P}_n$ has the observed frequencies on the diagonal, then we use modified minimum chi-squared estimation. Many variations are possible. But in all cases the minimum value of the statistic is asymptotically chi-squared. This is central chi-squared if the model is true, it is non-central chi-squared under local (contiguous) alternatives. All the results given above depend on the assumption that Ω is a smooth differentiable manifold, but generalizations are possible if Ω has 'corners'.

In fact most of the statistical techniques used in multivariate analysis are based on a very specific form of best asymptotically normal theory: maximum likelihood estimation and likelihood ratio testing. In many cases, however, more convenient alternatives are available with the same (first order) asymptotic properties.

3 Correspondence between model and data

The data in multivariate analysis, i.e. the recorded observations, are n vectors $z_1, \ldots, z_n$. Each one of the z_i is equal to one of the x_j, which are vectors in, say, R^t. Thus we can form the vector p, with p_j giving the proportion of z_i equal to x_j. Of course $p \varepsilon S^{m-1}$.

The model in section 1 is supposed to model these data. But how precisely? In many books and papers it seems that the z_i are identified without further ado with the $\underline{x}_i$, which makes p identical with $\underline{p}_n$. Thus it is assumed that the z_i are identically distributed independent random variables. But this is surely a very strange identification. A random variable is a function on a probability space, in this case a function taking the values x_j. And z_i is just a single vector, certainly not a function. Thus the implied identification of z_i with $\underline{x}_i$ is nonsense. A more careful analysis identifies z_i with a realization of the random variable $\underline{x}_i$, thus making $z_i = \underline{x}_i(\xi)$ for some ξ in the probability space on which $\underline{x}_i$ is defined. Equivalently the $n \times t$ matrix Z is a realization of the $n \times t$ random matrix $\underline{X}$, defined on the n fold product of the basic probability space.

The question which component of the model corresponds with the data Z is then answered. But the answer immediately gives rise to a new question. What is this probability space on which $\underline{X}$ is defined? Or, closely related to this, how can we investigate if our model is a satisfactory representation of reality? The only possible answer in many cases seems to be that the probability space is one of replications under identical conditions of the experiment that produced the result $\underline{X} = Z$. To interpret the model we have to imbed it in a framework of replications. In our case the framework implies that $\text{prob}(\underline{X} = Z)$ is the multinomial probability with parameters π, and observed frequencies np_j. Again observe that p is not a random variable, but a realization of $\underline{p}_n$. Statistics makes statements about the random variables, in our case about the operations performed with these random variables than are indicated by the theory of best asymptotically normal estimation. Thus statistics makes statements about the hypothetical framework of replications, and not about the data themselves.

Now if the replications can actually be carried out, then the framework can be tested as to its appropriateness. We can study empirically if the replications give relative frequencies close to the probabilities dictated by the multinomial model. But we cannot use the statistical methods that are available within the framework to investigate this. This would involve modelling the replications explicitly as realizations of independent and identically distributed random variables, which simply amounts to introducing a new and much wider framework of hypothetical replications. The model for n throws with a fair coin is $\text{prob}(\underline{X} = Z) = 2^{-n}$. If we want to verify or falsify this model by using k replications of the n throws, we cannot do this by modelling the replications. This simply would give rise to the model $\text{prob}(\underline{X}_1 = Z_1, \ldots, \underline{X}_k = Z_k) = 2^{-nk}$, which must be investigated in a similar way, and so on.

Thus we cannot falsify the model completely by statistical methods, at least not the basic framework of replication. We can merely find out if this framework is plausible or tenable by ordinary scientific reasoning. If we accept the model, with its corresponding framework of replications, then we can assert that $4n(\underline{p}_n - \frac{1}{2})^2$, for example, has a χ^2_1-distribution, with $\underline{p}_n$ the proportion of heads in n throws. This statistic can be used to test the fairness of the coin within the model of independent identically distributed replications. In a similar way the model of independent identically distributed observations could be tested within, for instance, the more general model of a first order stationary Markov chain. And so on. But in any case only components of the model are testable within the model. The replications can only be carried out outside the model, they cannot be fitted into the model because that would lead to an infinite regress.

Thus the usual chi-squared statistics, which we use as tests for components of the model, make statements about the hypothetical framework of replications. If we do not use this framework, then they are simply numbers, which indicate how well the model (without framework) fits the data. They measure distance between data and model in a weighted Euclidean metric, with the weights derived from the interpretation in terms of realizations of independent and identically distributed random variables. From ordinary scientific practice, outside the science of statistics, it is clear that these distance measures can be very useful indices even without accepting the framework on which their derivation depends.

4 Some replication frameworks

If we are sampling with replacement from a finite population with identifiable units, then the replication framework simply consists of all N^n samples, each having the same probability. Here N is the population-size, and $\pi_j = N_j/N$, with N_j the number of individuals with value x_j. If A is any subset of S^{m-1}, then we can define $p_n(A)$ as the proportion of samples which have their p in A. If B is a neighbourhood of the origin, then the law of large numbers asserts that $p_n(\pi + B) \rightarrow 1$ if $n \rightarrow \infty$. The central limit theorem says that

$p_n(\pi + n^{-\frac{1}{2}}B) \rightarrow N_\Pi(B)$, with N_Π the appropriate Gaussian measure. These results make it possible to apply the theory of best asymptotically normal estimation in this essentially combinatorial context. Roughly said: we transform p to $\Phi(p)$, where Φ is chosen in such a way that as many samples as possible have their $\Phi(p)$ near π. This purely combinatorial framework is quite satisfactory. There is nothing hypothetical about it, we merely count samples. Of course in most survey situations samples are drawn without replacement, using stratification and clustering. This introduces complications, but none of them seems very essential.

In case of a finite population the π_j are real, they are simply the population proportions. In principle they can be measured exactly. In case of geometrical probabilities or idealized physical models such as a fair coin, a perfect die, an ideal gas, the π_j can be calculated exactly from considerations of symmetry. But what is the probability that a given coin will produce heads if thrown by a given machine? Does such a probability exist, and in what sense? We think it exists in the same sense as true length exists. Although it is clearly an idealization, it can be defined with sufficient precision by empirical measurement procedures. Of course true length does not exist, but in most circumstances proceeding as if it existed does not entail large errors. In the case of true length we can design apparatus with sufficient precision, we can eliminate systematic errors, we can average over independent measurements. In the case of true probability we can continue to throw the coin until the proportion of heads seems stable to the precision desired. Thus in these cases we can act as if true probabilities exist, and we can approximate them as precisely as we wish. This also implies that our stochastic models can in principle be verified or falsified by empirical operations. The framework of replications is not only there as a model, it can actually be filled with empirical observations and it can be checked.

5 Social sciences

In de social sciences we often want to proceed in the same way as in the physical sciences. We have idealized models, such as simple learning theories, within which idealized probabilities can be computed exactly. We also have experiments, which can be embedded in a framework of replications and thus in a probability model. Unfortunately replications are seldomly carried out, which means that the framework is hardly ever tested. This has some obvious reasons. If replications are carried out, they often lead to very different and unexpected results. If attempts at replication fail, it is often very easy to find reasons why they fail. Circumstances have changed, subjects have aged, money has inflated, morals are different, laboratories have moved, governments have changed colour, and so on. In the social sciences replications in the classical sense of the word can be imagined, but they are often practically very difficult or impossible to carry out. Briefly we can say that the replication framework is imaginary in such cases.

What is the probability that human beings prefer beer to wine? Or even: what is the probability that individual A prefers beer to wine? These probabilities do not seem to exist in any real sense of the word. They are also imaginary. Perhaps we can merely speak about the probability that individual A indicates a preference for beer if asked at time B in country C in laboratory D, and so on. By making the replication framework more narrow the probability becomes more real, but the possibility of ever measuring it by actual replication disappears. This unfortunate tradeoff has prevented the social sciences from building up a body of empirically verified stable theoretical knowledge. It also makes the value of probability models very limited. We can easily build a model, but its basic assumptions cannot be tested, and its basic quantities are more imaginary than real. Thus a proper replication framework is missing in most social science situations.

Even if there is an acceptable framework of replications, the choice of the model is often problematic. In finite population surveys, for example, we have seen that the framework gives all samples of size n equal probability. If we sample with replacement, we can use the interpretation in terms of independent identically distributed random variables. But here the part of the model that specifies $\pi \varepsilon \Omega$ is often not very plausible. We can hardly expect it to be exactly true. In hypothetical populations (fair coins, all possible human beings) such models can be true exactly, but not in finite populations. We expect the four major blood groups to be approximately in the proportions dictated by the genetic model, but we do not expect these proportions to be exact for the population of the Netherlands at january 1, 1984. Even rational models, such as the Medelian model or simple learning theory models, can only be approximately true for finite populations.

Again the situation is worse in social sciences multivariate analysis. The path models (partial or conditional independence models) used there are only superficially rational in the same sense as Mendelian models. It is true, of course, that path analysis was developed in genetics to simplify the Mendelian calculations. But in the social sciences path analysis is used mainly in an exploratory way, for data reduction purposes, or for formalization of the investigators favourite prejudices. Nobody in his right mind will insist, when prompted, that in these situations models of the form $\pi \varepsilon \Omega$ are true, or even approximately true. The only thing we can say is that it would be nice if they were true. Or that in our idealized model of the social process they are true.

Let us briefly summarize the situation again, as we see it. Most papers in soaial science methodology journals that use probability models simply deal with calculations involving random variables. The same thing is true for biometric, econometric, and, not surprisingly, statistical journals. Als long as calculations with random variables are carried out we remain entirely within probability theory, i.e. within mathematics. The relevance of these results for data analysis, i.e. for science, must be demonstrated by linking the probability theory results to empirical data. The model must be interpreted, in the same way as differential equations describing motion or force must be interpreted. Probability models are interpreted by providing a suitable framework of replications. If such a framework cannot be found, is far-fetched, is untestable, then the model is irrelevant. If such a framework can be found and is testable, but is obviously false, then the model is not irrelevant, but obviously false.

We have seen, however, that even if the framework of replication is missing, it is still possible that the statistics computed in the random variable calculations can be valuable. It is common scientific practice to find out how false a model is by using either graphical aids or by computing some measure of fit. The chi-squared statistics of the best asymptotically normal theory are nice measures of fit. They have interesting properties under suitable frameworks of replications, and even if these frameworks do not apply they may behave in statisfactory ways. It seems to us that analysis of variance, chi-squared theory, log-linear theory, structural co-variance theory as in LISREL, do not enjoy their great popularity because they are optimal in restricted and largely irrelevant frameworks of replications. They are popular because they lead to nice representations, graphics, decompositions, arrows diagrams, algorithms, and so on. It is misleading to use irrelevant optimality criteria as a sales argument, good performance under a relatively large number of different (mostly non-probabilistic) conditions seems much more important for the social sciences. But this inevitably means that we must use more general performance criteria than optimality in the classical statistical sense.

6 Techniques

The first sections of this paper were mainly negative. We pointed out things that are misleading and do not make sense. This section is meant to be constructive. We discuss alternative practices that we prefer. For this we need the concept of a technique, which in a sense replaces the earlier pivot-concept of a model.

In our context a technique Ψ is a mapping of S^{m-1} into some representation space. We have seen that the theory of best asymptotically normal estimation makes it possible to define optimal techniques, given the model $\pi \varepsilon \Omega$ and given the replication framework. Thus the model dictates the technique, which is optimal given that the model is true. But models are never true, certainly not in social science contexts, and optimality is thus not relevant formulated in this way. We expect the technique to be 'quite good' if the model is 'approximately true', and this is the relevant property.

With a model we can try to associate an optimal technique, but with a given technique we can also try to associate a model for which the technique is optimal. In fact this was one of the ways in which Gauss derived the normal error theory model. This inverse procedure of pairing models and techniques is also valuable, but it is again limited by its focus on optimality. In fact optimality itself is too restricted a concept for a proper correspondence between models and techniques. For social science data analysis the whole idea of pairing models and techniques is too restrictive. We need a many-to-many correspondence between models and techniques. This has been pointed out earlier by Tukey (1962), Benzécri (1973), and by Gifi (1984), whose arguments have had a major influence on our point of view.

It is clear that our aims are related to the ideas behind the development of robust estimates. People involved in that field also study a many-to-many correspondence between, for example, univariate parent populations and location estimates. In the theory of robust estimation, however, the replication framework of ordinary statistics is copied. The notion of a model is generalized to that of a super-model, and the notion of minimum loss to that of minimax loss. These generalizations are still very much within classical statistics, or, as we prefer to say, within probability theory. For more complicated multivariate situations, in which there may be no replication framework, and no obvious (super)model we need different tools.

Thus we have the following situation. We start with a technique, in our case a mapping of S^{m-1} into Ω, and we want to know how good a tool this technique is. The question is, of course, vague. But this is a necessary consequence of the vague situation we are in. There is very little prior knowledge, and the prior knowlegde we have is almost completely negative. There is nothing we can safely assume. Thus there is little room for rationalism and deduction, and much room for empiricism and induction. Nevertheless we shall discuss two methods for studying the performance of our technique, and for contributing to the evaluation of its quality.

The first method is gauging. We apply our technique to data generated by a model. This can be an algebraic model, a probability model, a geometric model, or whatever. The idea is that we know all there is to know about this model (or this gauge) from a priori considerations, and that we want to find out how our technique represents this information in R^p. If it gives a good representation of the essential information, then we are satisfied. And we continue to apply our technique to another gauge. Among the gauges that are tried out there may be one for which the technique is optimal in a given sense, there may be another one for which the technique is optimal in another sense.

The second quality-control method is stability analysis. We study how the representation changes if we change the data, i.e. the vector p, in various ways. The basic idea is that an unimportant change in the data should cause only an unimportant change in the representation. This is related to the ideas behind robustness, but a little reflection shows that much of statistical analysis can be interpreted as stability analysis. But stability analysis is much more general, because there are many kinds of perturbations. We can leave out an individual, or a variable. We can study the effect of rounding error, of stochastic perturbations, of sampling, and so on. Thus in the same way as gauging generalizes the classical one-one pairing of models and techniques that dominates classical statistics, stability analysis generalizes the usual computation of (asymptotic) sampling distributions.

7 Example

To show that a program based on gauging and stability analysis can be carried out, we use a technique called homogeneity analysis or multiple correspondence analysis as an example. There are n objects, measured on T variables. Variable t has k_t possible values. Thus the observed multivariable has $k_1 \times \ldots \times k_T = m$ possible values, which are often called profiles or cells. The technique maps each of the n objects into R^q, Euclidean q-space. The representation is thus an n x q matrix H. For every possible H we can compute, for each variable t, the familiar partition $S_t = W_t + B_t$, which splits up the total dispersion of H into within-category and between-category dispersion. The q x q matrix S_t, the total dispersion, is, by construction, the same for all t. Homogeneity analysis maximizes the sum of the B_t, under the restriction that $S_t = I$. This means that objects with similar profiles tend to be close in the representation, profiles with high frequency tend to be near the centroid, unique profiles tend to be far from the centroid.

Our starting point is that homogeneity analysis is a technique which gives interesting representations of data from many different sources and types. For the details of the technique, and for many examples, we refer to the books by Gifi (1984), Greenacre (in press), Benzécri (1973, 1980), Nishisato (1980). We now proceed to answer the question: how good is homogeneity analysis?

Homogeneity analysis has been applied to many different gauges. We do not present a full list, but we mention the Guttman-scale, Thurstonean models, Coombsian models, Rasch models, Spearman models. In all these cases homogeneity analysis appears to represent the essential information in the model in a recognizable way, although in some cases care must be exercised in coding the variables. Properties of homogeneity analysis in models with total positivity, and in models in which all bivariate regressions can be linearized have been studied as well. The gauging results are again quite satisfactory. The multinormal distribution (for which we need an hypothetically infinite number of objects, of course) has been studied in considerable detail. The same thing is true for stationary processes, and Markov chains in particular. Homogeneity analysis can give good estimates of correlation matrices, even if we first discreticize and transform random variables. Most of these results van be found in Gifi (1984), Benzécri (1973, 1980), Heiser (1981), Schriever (1983). There are interesting statistical models in which homogeneity analysis gives consistent estimates of the parameters (De Leeuw, 1983).

Stability analysis for homogeneity analysis has also progressed considerably. Sampling stability has been studied using the classical delta method. This gives the usual confidence region information. The effects of deleting individuals and/or variables has been studied using both classical perturbation theory and algebraic perturbation theory. These results are in Gifi, but also in many publications of the French school. Resampling methods such as the Bootstrap and Jackknife have also been used in connection with homogeneity analysis. The effect of the discretization of continuous variables has been studied in some detail. Influence of missing information has been studied (Meulman, 1982). Numerically stable implementations have been developed.

8 Conclusions

Convential probability models cannot be applied in exploratory, multivariate, survey, social science investigations. This is not because there is something wrong with the models, but because there is something wrong with the reasoning linking models to data. There is no proper replication framework. The probabilities are imaginary. It seems to us that this problem also occurs in other contexts, but it is especially serious here. A second problem, which is perhaps more typical for social science investigations, is that there is no firmly established theoretical knowledge on which realistic restrictive models can be based. Thus model testing, which may be appropriate in other contexts, is out of place here.

An inevitable consequence is that much of data analysis is not probabilistic. There are many forms of stability, and replication stability is only one of these. There are many ways to indicate the size of a derivative, and the one chosen by the delta method is only one possibility.

For the 'foundations of statistical inference' debate, the main outcome of our analysis is that, at least in most practical situations we are aware of, the replication framework required by classical statistics is simply not available. Although we do not understand much of the 'foundations'-literature, it seems to us that most schools are firmly committed to probability models. The real differences seem to be confined to small-sample situations, in which the properties of the models become all-important. We think that statistical small-sample theory is of very limited relevance for data analysis. We think that Bayesian statistics not only idealizes reality beyond recognition, but it also tirannizes the data analist by comparing him with the immaculate perfection of the Coherent Person. It is bad enough to replace reality by models which are obviously false, one should not magnify the error by also replacing the scientific process by a model of it which is equally false.

9 References

Benzécri, J.P. e.a.	Analyse des Données. (2 volumes). Paris, Dunod, 1973.
Benzécri, J.P. e.a.	Practique de l'analyse des Donnes. (3 volumes). Paris, Dunod, 1980.

Gifi, A. Nonlinear multivariate analysis.
Leiden, DSWO-press, 1984.

Greenacre, M. Correspondence Analysis.
New York, Wiley, In Press.

Heiser, W.J. Unfolding analysis of proximity data.
Leiden, Data Theory, 1981 (dissertation).

Leeuw, J. de Models and methods for the analysis of correlation coefficients, Journal of Econometrics, 22, 1983, 113-138.

Meulman, J. Homogeneity analysis of incomplete data.
Leiden, DSWO-press, 1982.

Nishisato, S. Analysis of categorial Data: Dual Scaling and its applications.
Toronto, University of Toronto Press, 1980.

Schriever, B.F. Scaling of order-dependent categorical variables with correspondence analysis.
Int. Stat. Review, 51, 1983, 225-238.

Tukey, J.W. The future of data analysis.
Ann. Math. Statist., 33, 1962, 1-79.

Ontvangen: 3-1-1984