DATA MODELING AND THEORY CONSTRUCTION

JAN DE LEEUW

This paper was originally presented at the symposium *Operationalization and Research Strategy*, jointly organized by the Netherlands Association for the Advancement of Pure Research and the International Sociological Association, in Amsterdam, September 8-9, 1988. It was published previously as Chapter 13 in J.J. Hox and J. de Jong-Gierveld (eds), *Operationalization and Research Strategy*, Swets & Zeitlinger, 1990

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. data analysis, description vs inference, exploratory and confirmatory, correspondence analysis. multidimensional scaling.

> Among physicists at large, there is comparatively little inquiry into why or how they do what they are doing, and this is not to be deprecated, because human activities are inhibited by introspection.
>
> J. SYNGE, 1960, p.3

One of the purposes of this chapter is to compare, and to a certain extent contrast, two different approaches to the use of models in science. The two approaches are illustrated with Figure 1, a picture of a scientist arriving at the Scientific Forum. In the first case the scientist starts by selecting a model from the model box, he then grabs data from the data box, and holds them against the model. If they don't fit, he rejects the model and throws it away. The Forum applauds. On the other hand if the data do fit he holds on to the model. There is some applause, but not as much. If he has time he grabs new data, and he goes on until either they do not fit the model, or until his time is up. In the second interpretation of the picture the scientist begins by selecting the data. He then grabs a model from the model box, and holds it against the data. If it does not fit, he throws it in the trash. Not much applause. He then grabs another model, and so on, until he has found one that fits, or until his time is up. He holds on to the one that fits, if he finds one. There is also some applause in that case.

The picture and its interpretations illustrate a somewhat more serious quotation from a recent textbook on system identification

The term system identification may be somewhat unfamiliar, but it is the engineering equivalent for "fitting a model" or "choosing a model from a class of models". In engineering these are usually models for multivariate time

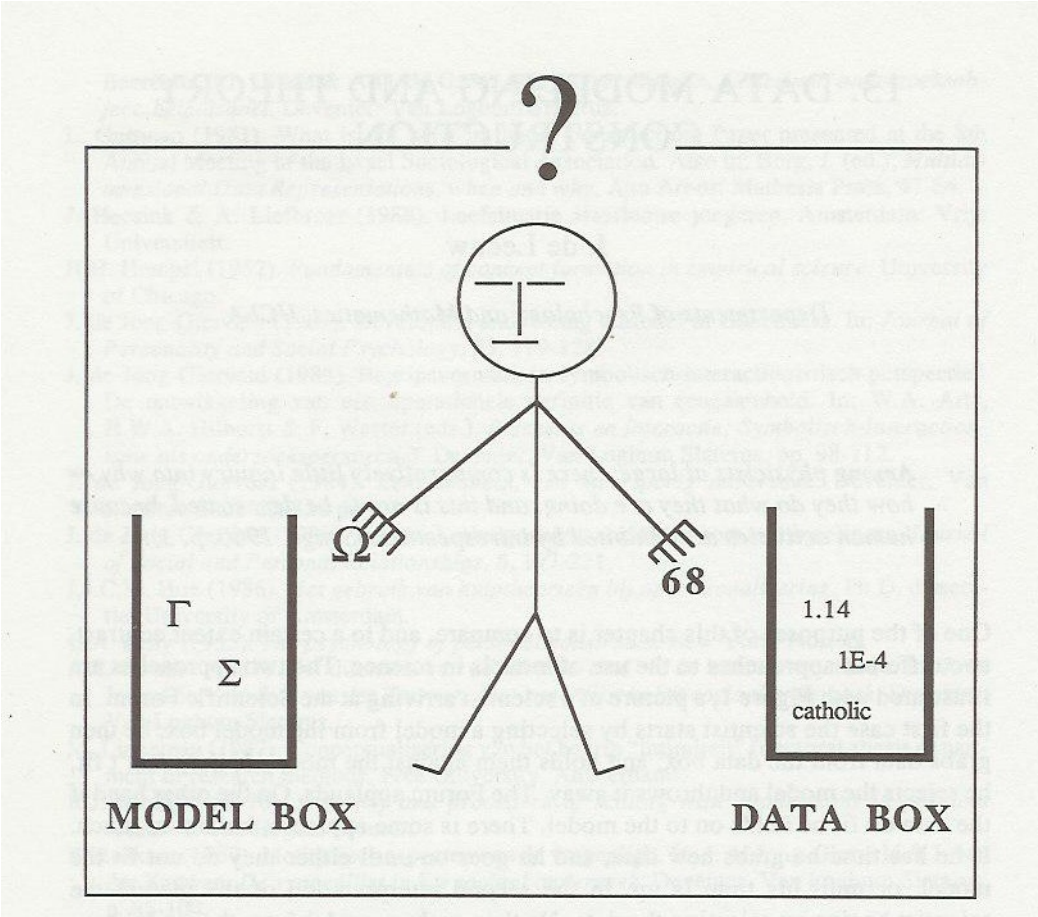

FIGURE 1. The Scientist between Data and Model

series, that describe the motion of one or more objects through space-time. But in the general discussion of system theory we can replace these dynamic systems by general mathematical-statistical models. Because modeling in the exact sciences takes place under somewhat different conditions than in the social sciences, a comparison of the two is interesting.

"Mathematical models may be developed along two routes (or a combination of them). One route is to split up the system, figuratively speaking, into subsystems, whose properties are well understood from previous experience. This basically means that we rely on "laws of nature" and other

well-established relationships that have their roots in earlier empirical work. These subsystems are then joined mathematically and a model of the whole system is obtained. This route is known as modeling and does not necessarily involve any experimentation on the actual systems... The other route to mathematical as well as graphical models is directly based on experimentation. Input and output signals from the system... are recorded and subjected to data analysis in order to infer a model. This route is system identification." [Ljung, 1987, p. 6].

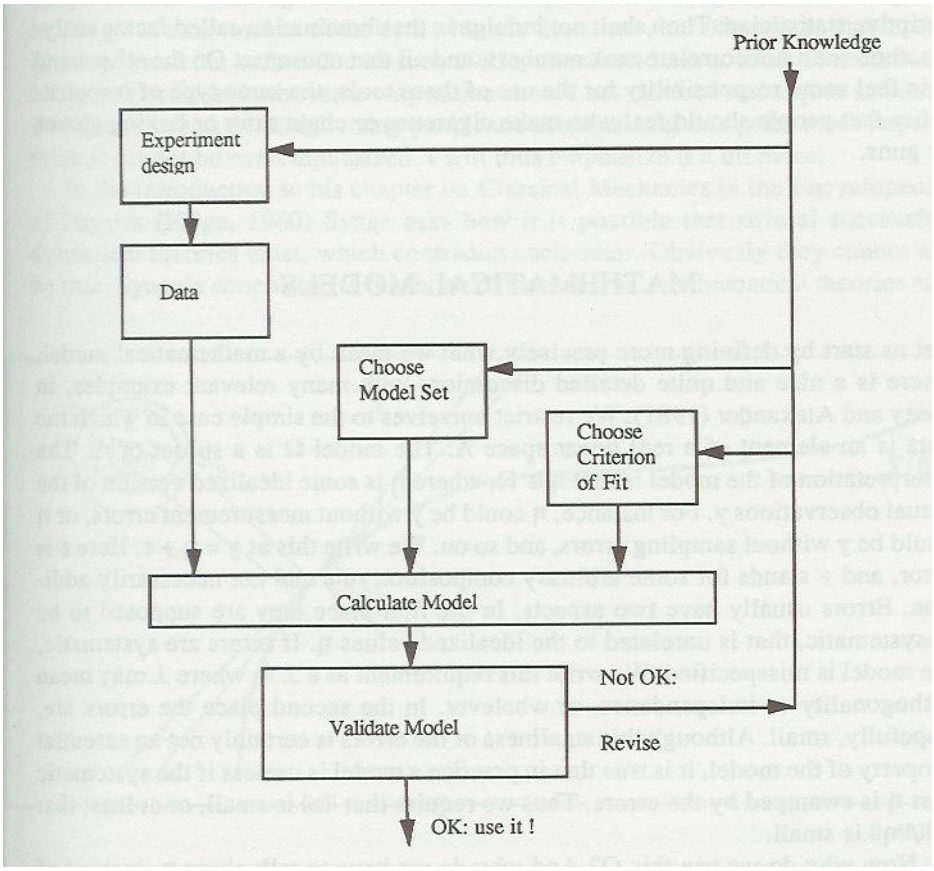

FIGURE 2. The System Identification Loop. [Ljung, 1987, p. 9]

Figure 2 illustrates system identification according to Ljung. It is clear that to some extent the two interpretations of Figure 1 are integrated here, because there is feedback in the loop, but the emphasis is on the second interpretation. Models are actually calculated. In system identification models are often thought as black boxes, as long as they describe the input-output behavior of the system it is not really interesting or relevant if their functional form is correct or true. In recent years passionate pleas for using system identification instead of modeling, in social sciences areas such as economics and psychology, have been published by Kalman [1982a,b, 1983]. The argument is that in these 'soft' sciences there is not enough prior knowledge to use modeling. Modeling in social sciences will inevitably be based on prejudices, not on valid prior knowledge derived from experience. We shall return to this discussion further on in the chapter.

The two interpretations of Figure 1 discuss possible behaviors of the *scientist*. On the other hand we also have to acknowledge that there is another profession, that of *statistician*, which is at least to some extent independent of any one particular science. One can be a statistician without being a psychologist, or a physicist, or a biologist. Statisticians are also concerned with the relation between data and models, but on a different level. In the discussion of the figure we have argued that the scientist 'holds the data against the model' and 'draws data from the data box'. This is vague, and can be made concrete in various ways. The statistician develops tools to carry out these activities. Tools vary in quality, in degree of sophistication, and also in price. It is very important to remember this analogy with tools: if you go to the statistician, you are going to buy a tool. Some people will try to sell you a tool which is far too expensive and elaborate for your purpose, and no

matter where you go there will always be commercials. Bayesian commercials, frequentist commercials. Many statisticians will try to convince you that what you have been doing in the past is incorrect, that using tools other than the ones offered by them is irresponsible, or even incoherent. But these are all commercials, and they should be evaluated as such.

Statisticians are instrument makers, tool builders. This is a difficult and honorable profession, but it is not science. Of course scientists can be part-time statisticians. But it is important to distinguish the two types of activities, and to recognize that statisticians do not make statements about the truth or generalizabilty of results. This is not their responsibility. Not our responsibility, I should say, because I am a statistician. This chapter looks at the way some of our tools are being used these days. I will do my utmost to avoid being an old-fashioned prescriptive statistician. Thou shalt not indulge in the abomination called factor analysis, thou shalt not correlate rank numbers, and all that nonsense. On the other hand I do feel some responsibility for the use of these tools, the same type of responsibility that people should feel who make cigarettes or chain saws or boxing gloves or guns.

MATHEMATICAL MODELS

Let us start by defining more precisely what we mean by a mathematical model. There is a nice and quite detailed discussion, with many relevant examples, in Saaty and Alexander [1981]. We restrict ourselves to the simple case in which the data is an element of a real linear space Λ. The model Ω is a subset of Λ. The interpretation of the model is that $\eta \in \Omega$, where η is some idealized version of the actual observations y. For instance, η could be y without measurement errors, or η could be y without sampling errors, and so on. We write this as $y = \eta + \epsilon$. Here ϵ is error, and $+$ stands for some

arbitrary composition rule and not necessarily addition. Errors usually have two aspects. In the first place they are supposed to be unsystematic, that is unrelated to the idealized values η. If errors are systematic, the model is misspecified. We write this requirement as $\epsilon \perp \eta$, where $\perp$ may mean orthogonality or independence, or whatever. In the second place the errors are, hopefully, small. Although this smallness of the errors is certainly not an essential property of the model, it is true that in practice a model is useless if the systematic part η is swamped by the errors. Thus we require that $||\epsilon||$ is small, or at least that $\frac{||\epsilon||}{||\eta||}$ is small.

Now why do we use this Ω? And why do we have to talk about η, instead of just being satisfied with y? There are many reasons to use models, in particular mathematical models, and there must be at least 1000 books each year which mention these reasons. A few keywords are consequently enough for our purposes. Models provide links with previous knowledge in the subject field, thus making it possible for science to be cumulative and for scientists to communicate efficiently. Merely submitting the data y is not considered satisfactory by the editors of most journals. Models provide interpolations and extrapolations, thus making it possible to predict data which have not been observed (yet). We can apply deductive reasoning methods to models, and derive consequences which may not have been obvious. And models are filters of our data. They make it possible to weed out the errors, and thus to provide us with more stable or more reliable information. This stabilizing property of models, which is very important, is illustrated in Figure 3, and discussed in De Leeuw [1988a].

This figure refers to a model for the covariance matrix, such as factor analysis or a LISREL model. The model Ω is the surface $\Sigma(\theta)$, we use Σ_0 for the hypothetical 'truth', and T_n and S_n are two observed covariance matrices

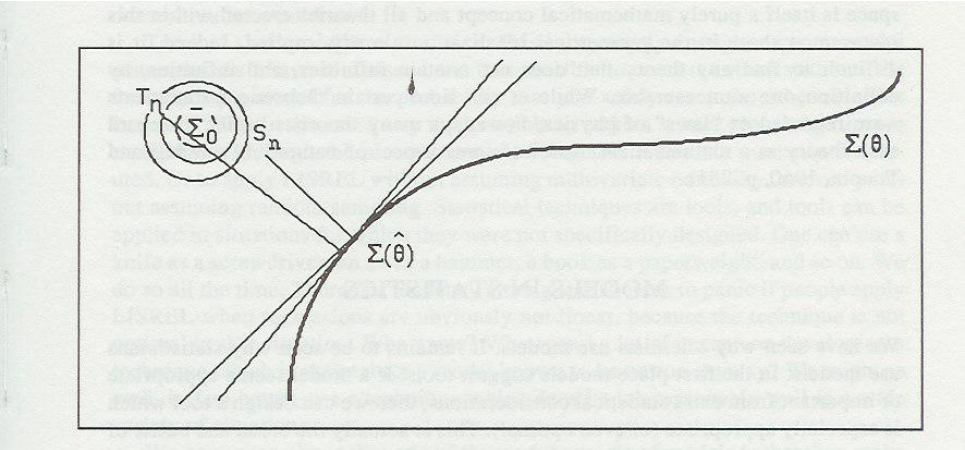

FIGURE 3. A Covariance Model in which Projection Increases Stability

based on n observations. Fitting the model means projecting the observed matrices on the model. Projecting the data on the model gives us a version of the data which is 'less true', but often more stable. As I have pointed out before [De Leeuw, 1983], one should never think of models as being true or false, in the same way as one should not think of techniques as being wrong or correct. Although many people are emphasizing this point these days, I think it cannot be overemphasized. I will thus emphasize it a bit more.

In the introduction to his Encyclopedia of Physics chapter Synge [1960] asks how it is possible that several successful dynamical theories exist, which contradict each other. Obviously they cannot all be true. Synge's conclusion is that none of them is true. Mathematical theories are no more than maps of nature, indeed the connection between the equations and the physical reality is even more remote than that between the map and the country (which at least both exist in the physical world). In the words of Ljung [1987, p. 6] there is an impenetrable but transparent screen between the world of mathematical descriptions and the real world. Synge [1960, p. 4] uses another model: there is a three-column dictionary, with a name in

the first column, a mathematical concept in the second column, and a physical concept in the third. "The hypothetical dictionary is used as follows. A physical problem is first formulated in terms of physical concepts. It is then translated into mathematical concepts by using the same words now with their mathematical meanings. Mathematical laws (usually differential equations) are found by a similar translation of physical laws, first stated in terms of physical concepts. The application of these laws to the problem in question then presents a problem in pure mathematics, and, when this problem has been solved, the conclusion is translated into terms of reality by restoring to the words their physical meanings." Synge goes on to remark than this description would have seemed ridiculously elaborate to physicists and mathematicians of the nineteenth century, who did not have such a clear distinction between physical and mathematical concepts. But it does explain why different maps can be useful. As Ljung points out, it also indicates that we should not think about models in terms of 'truth' but in terms of 'usefulness'. The concept of a 'true system' is a fiction, which is sometimes useful. In fact assuming that there is something like the truth, which we are trying to discover, is in itself a model of our scientific activities. "Any mathematical theory of physics must idealize nature. That much of nature is left unrepresented in anyone theory is obvious; less so, that theory may err in adding extra features not dictated by experience. For example, the infinity of space is itself a purely mathematical concept and all theories erected within this space must share in the geometrical idealization already implied. Indeed, it is difficult to find any theory that does not contain infinities, and infinities, by definition, are unmeasurable. While at one time certain theoretical statements were regarded as "laws" of physics, nowadays many theorists prefer to regard each theory as a mathematical model of some aspect of nature." [Truesdell and Toupin, 1960, p. 231].

MODELS IN STATISTICS

We have seen why scientists use models. It remains to be seen why statisticians use models. In the first place models suggest tools. If a model seems appropriate or important from extra statistical considerations, then we can design a tool which is especially appropriate (or even optimal). This is actually the bread and butter of a whole generation of mathematical statisticians. There are some general statistical principles, such as least squares or maximum likelihood, which are applied to particular models to produce techniques. It has been emphasized (for instance by Gifi [1984]) that models suggest tools in this way, but conversely tools can also suggest models. As long as models and techniques correspond one-to-one by the optimality relation, it is as easy to go from the model to a technique which is optimal for it, as it is to go from the technique to a model for which it is optimal. Gauss, for instance, started with the technique least squares and proved that it was optimal for normally distributed errors.

Figure 3 also suggests clearly why models are useful to improve techniques. If there is prior information about the scientific situation, then it is useful to take this into account in defining the technique, because using this prior information will improve the precision and stability of the technique. We have to remember here that if the prior information is merely prejudice in the sense of Kalman, then using it may introduce bias. And this bias can easily offset the gain in stability.

It is important to realize that building models is not really the task of the statistician, it is the task of the scientist. The scientist is supposed to know if certain assumptions are realistic, and if they apply to his/her situation. It is strange that a statistician has to assume for the scientist that the regressions

between the variables are linear, and that the disturbance terms are normally distributed. It has certainly been true in the past that statisticians have built stochastic models for scientists. This is not really a problem, as long as we continue to distinguish the two types of activities that are going on here. We have already seen that the scientist can be a part-time statistician, and certainly the reverse is true as well. But the point of view of statisticians has usually been to apply the general principles, such as maximum likelihood, and to suggest functional forms which are not too complicated, in the sense that they lead to feasible techniques. These are not necessarily the same considerations as the ones a scientist interested in the content matter would use.

There is also model-free statistics. This is, strangely enough, a controversial statement. As I have said elsewhere, it is possible to cross the street without first formulating a model for the probability that you safely reach the other side [De Leeuw, 1988b]. In the same way it is possible to make a scatter plot and to draw a straight line through the plot without assuming that errors are normally distributed. Or to apply LISREL without assuming multivariate normality, or even without assuming random sampling. Statistical techniques are tools, and tools can be applied in situations for which they were not specifically designed. One can use a knife as a screwdriver, an ax as a hammer, a book as a paperweight, and so on. We do so all the time. There is a tendency among statisticians to panic if people apply LISREL when regressions are obviously not linear, because the technique is not optimal in that situation. Who cares? Why spend a lot of energy on developing a technique which is optimal for a model known to be untrue anyway? We might as well, in fact better, use a technique which does its job reasonably well in a wide variety of situations.

We have seen above that one of the useful aspects of models is that they make communication between scientists easier. Publishing the data is usually not possible. On the other hand we can, and often do, publish parts of the data in the form of cross tables, regression equations, box plots, ANOVA tables, and so on. Statisticians would like to make us believe that if we publish regression weights we really have assumed 'implicitly' that our errors are normally distributed. I think this is nonsense. It is possible to vote without being a member of a political party. If we report a t-statistic we have merely taken the viewpoint that if we compare two means it makes sense to divide by the pooled standard deviation in order to get a scale free measure of the difference. Thus publishing summaries of the data, the results of applying statistical tools, does not presuppose the use of models.

CASE STUDY: WE START WITH THE MODEL

At this point it may be interesting to look at some case studies of modeling, on the one hand, and system identification, on the other side. The modeling interpretation of Figure 1 is certainly the most familiar one. It is what Suppe [1977] calls the "received view". It is associated with Popper's 'conjectures-and-refutations' philosophy of science, and with traditional statistical modeling. It assumes that we start with a model (theory, hypotheses). This model leads to certain predictions about reality. We then investigate whether these predictions are true or false. This particular philosophy does not pay any attention to the fact of where the model comes from. We could have adopted it by careful deductive reasoning in some axiomatic system, but also by reading tea leaves or by using hallucinatory drugs. There are no laws in the context of discovery, anything goes. Some procedures are perhaps more successful than others, but why this is the case

remains a mystery. All that matters is the context of justification, and this is where some very strict rules must be observed.

Let us suppose that there is a physical scientist Robert Hooke, who has an idea about spring balances. After long deliberation, using a great deal of prior experience, and perhaps some metaphysical ideas about the relation between measuring experiments and the time of the day, he has come up with the idea that the extension of the spring balance will be proportional to the weight applied to it, with the proportionality factor depending on the particular balance he is using. I am not saying, by the way, that the historical person Robert Hooke did indeed find the law of the spring by such reasoning. In fact the actual history of the discovery suggests that the system identification description of how this model was found is much closer to the truth. The model is that $x_i = \beta a_i$, with a_i the weight of object i (in kg), with β the constant describing the spring balance, and with x_i the extension caused by applying object i to the balance (in cm). How does Robert Hooke find out if his law fits the data? He simply plots the x_i against the a_i and he looks if the resulting points are on a line through the origin. The slope of the line gives the value of β. If a different spring is used, then we simply find a different line through the origin.

For our second example, the psychologist Charles Spearman had the idea that scores on intelligence tests will be proportional to the intelligence of the individual taking the test, with the proportionality factor depending on the particular test. Thus he also assumes (presumably taking Robert Hooke and physical scientists like him as his role model) a linear law of the form $x_i = \beta a_i$ with a_i the intelligence of individual i, with β the constant describing the test, and with x_i the intelligence test score (in number of items correct,

for instance). Although the models of Robert Hooke and Charles Spearman look very similar, there are some far reaching differences.

The first obvious difference is that Robert Hooke can use any number of additional independently validated instruments to find out what a_i is. They do not have to be spring balances (although they could be), they can be other types of scales as well. Basically we assume that a_i is known, and can be used in the verification of the law of the spring. We can actually plot x_i against a_i. For Charles Spearman the situation is more complicated. His a_i, the intelligence of person i, cannot be measured by more basic measurement procedures and we cannot make the actual plot of test scores against intelligence. Thus it seems that the law proposed by Spearman can never be falsified or verified.

But early in this century psychologists realized that if we have two tests, then the two laws $x_{i1} = \beta_1 a_i$ and $x_{i2} = \beta_2 a_i$ can easily be combined to $x_{i1} = (\beta_1/\beta_2)x_{i2}$. This new form of the law does not involve the unmeasurable a_i any more. This basic idea is really all there is to factor analysis, most subsequent developments are really only technical refinements. If we have m tests, then in a similar way $x_{ij} = a_i\beta_j$ merely says that the $n \times m$ matrix $X = \{x_{ij}\}$ has rank one. This rank condition on the matrix of observed test scores does not involve the concept of intelligence, and does not suppose that independent measures of intelligence are available. Instead, it defines intelligence. We call a variable, defined on our population of individuals, the intelligence of these individuals if the test scores of these individuals are linear functions of this variable. Observe that Robert Hooke, by using m spring balances, could have defined the weight of an object in exactly the same way. If weight is interpreted as a latent variable, then it is still possible

to measure it given at least two spring balances. It seems that, in the second analysis, Charles Spearman is no worse off than Robert Hooke.

But this is deceptive. Although the measurement theoretical properties of weight measurement and intelligence measurement seem to be identical, the practical implementation of the two programs rapidly showed the enormous differences. In the first place tests do not have a natural unit. Number of items correct is easy enough to use, but it seems to use the idea that each item is equally important (it is like using the number of equal weights needed to get equilibrium in a pan balance as a measure of weight). In the second place the psychologists have to agree that the tests they use are all tests of intelligence. This presupposes a lot of agreement on the nature of intelligence, and there is no such agreement. Not then and not now. Thirdly there is the unfortunate fact that all measurements have a certain error associated with them. For Robert Hooke this is no real problem. His errors are small, and can be made much smaller by various technological refinements. Spearman perhaps tries to minimize the influence of measurement errors by increasing the number of tests. But he can only do this by adding tests that are valid, i.e. tests that measure intelligence and nothing else. Fourthly it turns out that Robert Hooke's model indeed describes the behavior of spring balances, at least under relatively normal circumstances. Charles Spearman's model does not describe the behavior of intelligence tests at all well. Psychologists have tried to repair this unfortunate situation by blaming measurement errors. This was not very convincing. The next step was to blame the selection of tests. The conclusion of this phase was aptly summarized by Wolfle [1940]. If we take a battery of tests and remove all tests that do not satisfy the Spearman model, then the remaining tests do indeed satisfy the Spearman model. It is clear that such manipulations are

suspect, and also that they are possible only because there is no agreement on the nature of intelligence. If a spring balance does not satisfy the Hooke model, then we reject the model. We are not going to argue that this object is not really a spring balance.

What do we learn from this example? The most ambitious attempt so far in psychology to build a mathematical model from various bits and pieces has failed rather miserably. In the end we can see that although Spearman certainly did take some empirical evidence (mostly of a qualitative nature) into account while building his model, the main components were indeed prejudices. There was the (very important) eugenic prejudice, which made it desirable to rank persons on a single scale [De Leeuw, 1986]. There was the desire to be respectable scientifically, and to work with measurable one-dimensional scales. And after twenty five years of sloppy mathematics and analysis of often very small data sets, the model consisted almost completely of prejudices, and the arguments around it became fundamentalist quarrels. It may be true, by the way, that factor analysis is a poor example. But the other examples that I am familiar with are equally poor. The quantitative genetic models have failed to teach us anything about the transmission of human traits, even of traits as simple as body weight. It seems too they have not been very successful in controlled breeding experiments with plants and animals. Economic models, based on rational considerations, have been quite unsuccessful in predicting economic developments.

CASE STUDY: WE START WITH THE DATA

Now let us look at the second interpretation of Figure 1, that of system identification. The chemist Beer is interested in spectroscopy, i.e. the absorption of photons by a chemical specimen as a function of the energy

of the photon. The chemical specimen consists of various quantities of a number of distinguishable light-absorbing entities, or chromophores. Let us now measure x_i, the negative logarithm of the fraction of the intensity of a beam of light of wavelength i which passes through a solution with concentration b of a certain chromophore. Chemists use Beer's law, which says that $x_i = a_i b$. If we use different solutions j, with different concentrations of the chromophores, then $x_{ij} = a_i b_j$. If the solutions contain p different chromophores, in different concentrations, then Beer's law becomes $x_{ij} = \sum_{s=}^{p} a_{is} b_{js}$. In many instances in spectroscopy the number of chromophores p is unknown, and so are their concentrations b_{js} and their extinction coefficients a_{is}. In chemometrics factor analysis is often used to estimate all these quantities [Malinowski and Howery, 1980].

There is a comparable situation in psychology. Suppose we have measurements of n individuals on m tests where all tests seem to have something to do with aptitude. Collect them in an $n \times m$ matrix $X = \{x_{ij}\}$. The psychologist Leon Thurstone made Spearman's theory empirical in the sense that he did not suppose there was only one factor determining the relationship of the tests; the number of factors had to be determined empirically from the data. In the model box there are various factor models, with various numbers of factors, and we look until we find one that fits our data. In the same way the engineers using system theory start with a box filled with linear dynamical systems (or ARMAX models). These models vary in the dimensionality of the state space, and we look until we have found a system which fits the data.

This identification approach, which clearly starts with the data, has been quite unpopular, at least among post-positivist philosophers and sophisticated social scientists. Thurstone's multiple factor analysis may have taken

us away from Spearman's apriorism, but it actually led to anarchy. The latest estimate is that there are more than one hundred factors of intelligence, and they can all be measured independently by constructing suitable tests. This is clearly the end of a research program. Nevertheless it is useful to point out that there are certain areas in science in which this particular approach is indeed taken seriously. We start with a quotation from Isaac Newton (1687). "But hitherto I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypothesis; for whatever is not deduced from the phenomena is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy. In this philosophy particular propositions are inferred from the phenomena, and afterwards rendered general by induction." Similar ideas were quite influential early in this century because of the philosophical work of Pearson and Mach, who thought that scientific laws were efficient summaries of long lists of sense data. In general they seem to appeal to physical scientists and engineers. Applications to the social and behavioral sciences are quite rare, perhaps because of the debacle of factor analysis. Only more recently we find attempts to revive the empiristic approach there as well. "Le model doit suivre les données, non l'inverse." [Benzécri, 1980, p. 6]. We shall review a recent example.

CAUSAL MODELS EX MACHINA

With the increasing sophistication of data analysis techniques, the fact that much better and larger data sets are available, and the increases in computer power, there have been new hopes for the empiristic program from the second interpretation of Figure 1. Dynamic systems models and Kalman

filtering techniques of prediction and control have been very successful in helping to put men on the moon, keeping satellites in orbit, regulating hydroelectric plants, and so on. As we have seen, these models often take a simple black box approach to modeling. We measure input and output of the system, and we select from the model box until something fits. Then this surviving model is used for prediction and control.

The failure of factor analysis to produce a coherent model for intelligence testing can, of course, be attributed to the limitations of the factor analysis model, which does not take any external variables into account, and to the particular subject area, which may be very complicated. Models which do not have the limitations of the factor analysis model, and which combine features of factor analysis with the simultaneous equations models of econometrics and the path analysis models of genetics, have been very popular recently. They are called causal models, or simultaneous equations models with latent variables or, quite inappropriately, LISREL models. The basic idea is probably familiar. A path diagram is drawn with arrows connecting the variables. Arrows are then translated into linear relationships between the variables, and a technique such as LISREL or EQS is used to fit the model to the data.

Perhaps the best introduction to simple path analysis and the problems connected with it is Freedman [1987]. In LISREL and EQS there is the additional complication that some variables, which seem relevant from theoretical considerations, cannot be measured directly. Instead we measure one or more indicators of this variable. Intelligence is the prime example, but social economic status and permanent income are similar constructs. In econometrics the related notion of a variable measured with error has been studied in detail. In system theory Willems [1987] has recently pointed

out that latent variable terminology can also be used to describe state space realizations of linear dynamic systems. The number of path models for a given number of variables is very large, especially if we admit latent variables and correlated measurement errors. Now we can imagine applying the empiristic strategy here as well. We have data, usually in the form of a covariance matrix Sn' and we look in the box of LISREL models until we have a model with a satisfactory fit.

How do we search in the space of causal models, an essentially discrete structure with zillions of elements? There has been a lot of work which seems relevant. to this question. In statistics some tools have been introduced which make it possible to compare models in terms of fit. Not straightforward fit, of course, because then a model with more parameters would always be preferable to a simpler model with less parameters. Various combinations of fit and simplicity have been proposed, for instance by Akaike, Schwartz, and Hannan (compare De Leeuw, 1988, for a discussion). Willems [1986a,b,c] discusses abstract versions of the same type of procedures under the name approximate modeling in the context of linear systems. Unfortunately there seems to be a proliferation of model selection tools, so that it seems likely we will need a tool to select model selection tools in the not too distant future. The major computer programs for fitting causal models with latent variables all have tools built in to search through the space of models. This is a feature they are required to have for commercial reasons. We shall look in more detail at an even more recent attempt to generate theory by computer. Glymour et al. [1987] have programmed a search procedure in their computer program TETRAD. Their commercial is based on another intellectual catch phrase of these troubled times: Artificial Intelligence. I do not want to sound prejudiced, and certainly not

old-fashioned, but I suggest that anything which prominently features the label Artificial Intelligence should be approached with a great deal of mistrust. If we look at the implementation details of TETRAD we find a simple objective function which is minimized by straightforward search techniques over the space of graphs. The choice of the loss function is not explained very well, and the function itself looks quite unattractive and clumsy.

However, in Spirtes et al. [1988] the performance of TETRAD is compared with the model search procedures in the LISREL and EQS programs, which are based on modification and fit indices. TETRAD recovers true models better than these other procedures. I am not impressed. In the first place this is a Monte Carlo study, and we have no idea how these results will hold up in general. In the second place the fact that one dubious technique outperforms two other dubious techniques does not make it any less dubious. Thirdly it is clear that in situations like this the starting model, the place where the search begins, will be very important. This means that the choice of the starting model still requires the same sort of knowledge that choice of model requires in the case where one does not apply search. Fourthly it is unknown, and anybody's guess, what these search procedures do to the stability of the results.

Glymour et al. are no fools. They know their philosophy of science, and they are able to demolish most of the arguments against causal modeling of Freedman and others [De Leeuw, 1985; Freedman, 1987]. This is because these arguments are in absolute terms, and take the standard framework of statistics for granted. That linear causal models are literally false, for instance, is not really an objection. We have already seen that. That the assumptions are not tested is merely a reflection on actual practice. They

could be tested, and sometimes they are. Glymour et al. argue, quite convincingly, that the situation in the social and behavioral sciences is not really different from that in physics. As we have seen in comparing Hooke and Spearman, there is something to be said for that.. The difference is more subtle than the critics of causal modeling suppose. The argument should not be that the assumptions are not tested, the argument should be that the assumptions are often untestable because replication is impossible. Moreover questions on whether social science data are normally distributed or not are usually quite irrelevant. In many cases probabilistic models do not make sense at all. The level of error in social science modeling, which is not essential from a methodological point of view, destroys both the credibility and the generalizability of many results. Unlike bridges and other simple physical structures, social and behavioral science models crumble and collapse. Not only if you try to use them, but even if you look at them carefully. It is usually unclear how one could use these models in the first place. Although it may be attractive to specify one's prior information in the form of a graph, it usually requires far too much prior knowledge to do this in any detail. There may be some global properties of the graph which seem quite certain (relations based on order in time, for instance), but details will be largely arbitrary. We can use programs such as LISREL, EQS, or TETRAD to fill in the details, but the stability of the resulting solutions will be doubtful and the usefulness of the detailed aspects of the fitted model even more so.

DATA ANALYSIS

If we look at Figure 1 and its two interpretations in more detail, the differences between the two approaches turn out to be mainly questions of

emphasis. The interpretation in which the scientist walks in with a model, and tests it, does not say where he got this model from. This means that he may very well have found it by using system identification techniques, i.e. by using the second approach. The difference then is that the first interpretation does not talk about the phase in which people construct their models because it feels that this does not properly belong to science, or rigorous science. Nevertheless there is little doubt that off stage modelers do it too. On the other hand system identification people do not really talk very much about the contents of the model box. They seem to think that every conceivable model is in that box, but in actual practice the box contains only finite dimensional linear systems, or factor analysis models, or LISREL models. This means a great deal of modeling has already gone into the filling of the model box before the identification starts. In fact fitting a statistical model, also in the first approach, often amounts to choosing a number of free parameters. Thus we do not really hold the data against the model, but we look if there is an instance of the model close enough to the data. In Figure 3 we project on the model because the model is not a single covariance matrix, but a surface in the space of covariance matrices. Or, put differently, in comparing data and models we also perform selection from the model box. The model is not a black box, about the contents of which we cannot say anything, it is a gray box, whose contents are known up to values of a number of free parameters. Thus we see that the modeling approach and the system identification approach can easily be interpreted as two different phases in the same cyclic process. There is no confirmation without exploration, no induction without deduction, no inference without description. Instead of talking endlessly about Aristotelian dichotomies like this, we had better get to work. Some of you may have some modeling to do, and I can perhaps construct and sell another useful tool.

A final word about the nature of the tools. We have seen that some tools are data presentation tools. Graphics, plots, tables and so on are simple examples, but even complicated LISREL models derive most of their popularity from the graphical component, the path diagram. The statistical superstructure is used to sell LISREL to unsuspecting audiences. but it is largely irrelevant and its appropriateness is highly debatable. If LISREL or a similar program is used to describe the dependence/independence structure between variables. then the relevant question is whether it is the best tool for that purpose. We know that there are a number of special cases of the model, basically the recursive or block-recursive path models with independent errors, in which there is a nice probabilistic interpretation of the graph in terms of partial independence. In such situations the graph seems to give useful information in a compact and pleasant form. Because detailed values of path coefficients are usually not very reliable, it seems wise to limit oneself to block-type models in which either no between-block paths or all between-block paths are drawn. This makes the figures particularly simple to interpret, and actually by following this strategy we stay close to classical regression, canonical analysis, and factor analysis. A detailed search, as performed by TETRAD. focuses on uninteresting and unstable aspects of the description.

SUMMARY

Complicated fitting procedures with many parameters are dangerous techniques, especially if they masquerade as inferential techniques. We use very little information from the data, and we do not impose restrictions of a strong type on the representation. This type of program appeals greatly to many social scientists who are very unsure about the value of their prior

knowledge. They prefer to delegate decisions to the computer, and they expect techniques to generate knowledge. All too often, this strategy leads to chance capitalization, triviality and degeneracy. Hypotheses are never rejected, and investigators are constantly making errors of the second kind. We impose so little prior knowledge that the data, including all outliers, stragglers, idiosyncrasies, coding errors, missing data, completely determine the solution. As a consequence results can, of course, never be replicated. This is the empiristic and technological approach, popular in applied psychology. On the other hand it is well known that if we pay too much attention to errors of the second kind social scientists can say absolutely nothing. This is also considered to be an undesirable state of affairs. It can be circumvented by introducing vast quantities of prior knowledge, as in clinical psychology, personality psychology, and some sociology. Of course in many cases the prior knowledge is nothing but prejudice, and it so dominates the investigation that the results become almost independent of the data. These two extremes define the dilemma of much applied empirical social science. According to the canons of scientific respectability we can say almost nothing, and the things we can say are likely to be trivial.

There is one quite legitimate way out of this problem, one which in fact many applied researchers already use. If we have done a large scale investigation on the relationship between intelligence tests, we do not assume the factor analysis model. We merely describe the correlation between the tests as good as possible using all the tools provided for this purpose by statisticians. Perhaps factor analysis, interpreted as a data presentation tool, is one of them. But there is no reason, so far, to take it seriously as a model that describes what really is going on. The same thing is true if you study formation of attitudes on the basis of information and background. Do not

assume that the Fishbein model is true, but simply present the data in such a way that it becomes clear that some groups of partial correlations are systematically small. Model testing is far too pretentious in this case. We are still finding out what the facts are, it is too early to define the building blocks for the model. The same applies to the analysis of school careers. Throughout the last fifteen years I have followed the development of 'causal' school career models for the Netherlands. This research program has been progressive, I think, because all unnecessary and unstable branches have been trimmed from the models. What is left, however, is quite trivial. The scores on the intelligence test and the advice of the teacher in the sixth grade determine the choice of secondary education. The choice of secondary education determines the career in secondary education. As a consequence I find the tables and graphs published by the Central Bureau of Statistics far more informative than the rather pitiful LISREL models fitted by the social scientists. The IQ debate is another example. It has somewhat died down now, but the last summaries published by the various opponents show that they do not even agree (after 100 years of research) what the basic facts are.

Thus descriptive statistics, in various forms and of various degrees of sophistication, are not only quite sufficient in many cases, they are at the moment all that we can responsibly do. It can be argued that the true purpose of any data analysis is providing generalizations and predictions, and that a descriptive analysis does not show how to generalize. This is not true. Any respectable tool gives information about its precision, Le. about its stability under various circumstances. The confidence intervals and significance tests in the output of programs as LISREL or TETRAD are a very primitive and unconvincing sort of stability information, because they are only relevant under hopelessly unrealistic circumstances. I have argued earlier [De

Leeuw, 1988c] that if you want information about replication stability, then you must replicate your experiment. Or you must wait until somebody else replicates it. If it turns out that replications behave as statistics expects them to behave, then there is no need to replicate further, because statistical models can take over the burden of additional replications. But you cannot assume the usual statistical models for replication stability if they are a priori highly unlikely, and if you are never going to test their reasonableness in the first place. If your model falls apart if somebody replicates the investigation, or if you decide that it is impossible to replicate the investigation anyway, then replication stability is irrelevant and there is no need to worry about it. Building models, identifying systems, making generalizations and inferences is not the task of the statistician but of the scientist. Hammers don't build houses, and books are useless until somebody reads them.

References

J.P. Benzécri. Les Principes de l'Analyse des Données. In J.P. Benzćri et al., editor, *Analyse des Données*, volume I. Dunod, Paris, 1980.

J. De Leeuw. Review of Four Books on Causal Analysis. *Psychometrika*, 50:371–373, 1985.

J. De Leeuw. Models and Methods for the Analysis of Correlation Coefficients. *Journal of Econometrics*, 22:113–138, 1983.

J. De Leeuw. Individuele Verschillen en Ongeljkheid [Individual Differences and Inequality]. In J. Berting, editor, *Sociale Ongelijkheid [Social Inequality]*. Coutinho, Muidernerg, Netherlands, 1986.

J. De Leeuw. Model Selection in Multinomial Experiments. In T.K. Dijkstra, editor, *On Model Uncertainty and its Statistical Implications*. Springer, Berlin, 1988a.

J. De Leeuw. Models and Techniques. *Statistica Neerlandica*, 42:91–98, 1988b.

J. De Leeuw. Multivariate Analysis with Linearizable Regressions. *Psychometrika*, 53:437–454, 1988c.

D.A. Freedman. As Others See Us: a Case Study in Path Analysis (with Discussion). *Journal of Educational Statistics*, 12:101–223, 1987.

A. Gifi. *Nonlinear Multivariate Analysis*. DSWO Press, 1984.

C. Glymour, R. Scheines, P. Spirtes, and K. Kelly. *Discovering Causal Structure*. Academic Press, New York, 1987.

R.E. Kalman. Identification from Real Data. In M. Hazewinkel and A.H.G Rinnooy Kan, editors, *Current Developments in the Interface: Economics, Econometrics, Mathematics*. Reidel, Dordrecht, 1982a.

R.E. Kalman. System Identification from Noisy Data. In A.R. Bednarik and L. Cesari, editors, *Dynamical Systems*, volume II. Academic Press, New York, 1982b.

R.E. Kalman. Identifiability and Modeling in Econometrics. In P.R. Krishnaiah, editor, *Developments in Statistics*, volume IV. Academic Press, New York, 1983.

L. Ljung. *System Identification: Theory for the User*. Prentice Hall, Englewood Cliffs, 1987.

E.R. Malinowski and D.G. Howery. *Factor Analysisin Chemistry*. Wiley, New York, 1980.

T.L. Saaty and J.M. Alexander. *Thinking with Models*. Pergamon Press, Oxford, 1981.

P. Spirtes, R. Scheines, and C. Glymour. Simulation Studies of the Reliability of Computer Aides Model Specification Using the TETRAD, EQS, and LISREL Programs. Technical Report CMU-LCL-88-3, Laboratory of Computational Linguistics, Carnegie Mellon University, 1988.

F. Suppe. *The Structure of Scientific Theories*. University of Ilinois Press, Urbana, 1977.

J.L. Synge. Classical Dynamics. In S. Flügge, editor, *Encyclopedia of Physics*, volume III(1). Springer, Berlin, 1960.

C. Truesdell and R. Toupin. The Classical Field Theories. In S. Flügge, editor, *Encyclopedia of Physics*, volume III(1). Springer, Berlin, 1960.

J.C. Willems. Models for Dynamics. Technical report, Department of Mathematics and Computer Science, University of Groningen, 1987.

J.C. Willems. From Time Series to Linear System. Part I: Finite Dimensional Linear Time Invariant Systems. *Automatica*, 22:561–580, 1986a.

J.C. Willems. From Time Series to Linear System. Part II: Exact Modeling. *Automatica*, 22:675–694, 1986b.

J.C. Willems. From Time Series to Linear System. Part III: Approximate Modeling. *Automatica*, 23:87–115, 1986c.

D. Wolfle. *Factor Analysis to 1940*. Number 3 in Psychometric Monograph. University of Chicago Press, 1940.

Departments of Psychology and Mathematics UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024