Chapter 1

Here's Looking at Multivariables

Jan de Leeuw

1 Introduction

I don't really understand what "visualization of categorical data" is about. This is a problem, especially when one is supposed to write the opening chapter for a book on this topic. One way to solve this problem is to look at the other chapters in the book. This empirical, data analysis–oriented, approach is based on the idea that the union of all published chapters defines the topic of the book.

For this book, the approach of looking at the titles of the chapters produced somewhat disappointing results. Whatever "visualization of categorical data" is, it consists of about 50% correspondence analysis, about 10% multidimensional scaling, about 10% cluster analysis, about 20% contingency table techniques, and the remaining 10% other ingredients. It is unclear from the titles of the chapters what they have in common. When writing this introduction I assumed, and I have since then verified, that every author in this book shows at least one graph or one plot. But this is a very weak common component. Not enough to base an opening chapter on.

Thus the empirical approach fails. Alternatively, I can try to define my way out of the problem. This is intellectually a more satisfying approach.

We start with *data*, more precisely *categorical data*, and these data are analyzed by using a (data-analytical or statistical) *technique*.

Such a technique produces an *image* or a *smooth* or a *representation* of the data. The title of the book indicates that we are particularly interested in *visual* smooths or representations of the data.

Thus our course is clear; we have to define *data*, *technique*, and *representation* and then single out categorical data and visual representations. After that exercise, we can go back and see if and how the contents of the conference fit in.

This paper can be seen as the next member of the sequence of de Leeuw (1984, 1988, 1990, 1994). The general approach, in an early form, can also be found in the first chapter of Gifi (1990).

2 Data

The data D are an element of the *data space* $\mathcal{D}$. The design of the experiment, where both "design" and "experiment" are used in a very general sense, defines the data space.

If you distribute a particular questionnaire to 1000 persons and your questionnaire has 50 items with 7 alternatives each, then the data space has $1000^{7^{50}}$ possible elements, and if you allow for missing data and nonresponse, it has even more. In general, these data spaces, which are the sets of all possible outcomes of our experiment, tend to be very large.

The same thing is true if you make measurements in a particular physical experiment, or if you plant a number of seeds in a number of pots, or if you watch a number of infants grow up. Even with a limited number of variables, the possible number of outcomes is very, very large.

In all these cases the data space is defined before the observations are actually made, and the possible outcomes of the experiment are known beforehand as well. Is it possible to be surprised? I guess it is, but that is a flaw in the design.

2.1 Coding

We do not find data on the street. Data are not sense impressions, which are simply recorded. Data are *coded*, by which we mean that they are entered into a preformatted database. This is not necessarily a computerized database; it could simply be the codebook given to interviewers or to data-entry persons, or it could be an experimental protocol.

The important notion is that data are already categorized and cleaned and that the protocol tells us how to reduce data from a data space of quadri-zillions of elements to one of trillions of elements. We know, for instance, that we can ignore the look on the face of the person filling in the questionnaire, and the doodles on the student's examination forms are not counted toward the grade.

Another key point is that usually the greatest amount of data reduction goes on in this coding stage. The really important scientific decisions, and the places where the prior knowledge has the greatest impact, are not necessarily the choice between the normal distribution and Student's t or between frequentist and Bayesian procedures.

2.2 Example

This could perhaps be illustrated by an actual example. One of the clients of University of California Los Angeles (UCLA) Statistical Consulting is the California Department of Corrections. There is a gigantic project set up to study whether the classification of prisoners into four security categories actually reduces within-prison violence. The data for answering this study are the prison careers of all individuals who were in one of the state prisons in California in the last 10 years. It will not surprise you to hear that these are hundreds of thousands of individuals. Many of them have been in and out of prison for 20 or more years. They have been shifted between security levels many times, often on the basis of forms that are filled in and that have objective cutoffs, but often also on the basis of "administrative overrides" of these objective results.

It is generally known that propensity to violence in prison is related to age, to previous prison career, and to gang membership. Clearly, there are potentially thousands of variables that could be coded because they might be relevant. Wardens and other prison personnel observe prisoners and make statements and judgments about their behavior and their likelihood to commit violent acts while in prison. Presumably, many of these judgments and observations change over time for a given prisoner and maybe even for a given prison guard.

The observations and classifications of the prison personnel, however, are not data and not variables. They become data as soon as they are organized and standardized, as soon as variables are selected and it is decided that comparable information should be collected on each prisoner, over time, over changing security levels, and perhaps over changing institutions. It is decided that a study will be done, a database will be constructed, and the integrity and completeness of the database become so important that observations in different institutions and time periods by different observers on the same individual are actually coded uniformly and combined. Without reducing the chaos of impressions and judgments to a uniform standard and format, there really are no data.

2.3 Categorical Data

Classically, data are called categorical when the data space is discrete. I think it is useful to repeat here that all data are categorical. As soon as we have set the precision of our measurements, the grid on which we measure, and the mesh of our classifications, then we have defined a discrete and finite data space.

Statistics has been dominated by mathematics for such a long time that some people have begun to act as if "continuous" data is the rule. Continuous data is a contradiction. Continuity is always part of the mathematics, that is, of the *model* for the data. The question whether continuity "really" occurs in nature is clearly a metaphysical one, which need not concern us here. We merely emphasize that continuity is used mostly to simplify computations, in the same way as the normal distribution was first used to simplify binomial calculations.

The codebook, or the rules for entry into the database, also contains rules for coding numerical information. It has to be categorized (or rounded), because our data entry persons and our computers cannot deal with infinite data spaces.

Thus:

All Data Are Categorical

although perhaps some data are more categorical than others. This suggests that, in a strict sense, it is impossible to distinguish "categorical data" from "other data." In actual practice, however, we continue to use the distinction and speak about categorical data when our variables are non-numerical and/or have only a small number of discrete values.

2.4 Multivariables

The most important type of data that science has been able to isolate is the variable or, if you like, *multivariable*. This is closely related to the "fluents" in Newtonian physics, the random variables of statistics, and the variables in mathematical expressions. For some useful philosophical discussion of these concepts we refer to Menger (1954, 1955, 1961) and quite a few other publications by the same illustrious author.

In the case of a multivariable, the data space is the product of a number of functions defined on a common domain, with different images. Table 1 shows a simple example of a bivariable, describing the nine faculty members in the new UCLA statistics department. Two variables are used: department of origin and country of origin.

If we look at the types of data spaces most frequently discussed at this conference, we find the multivariable in various disguises. In formal concept analysis multivariables are called *many-valued contexts (mehrwertige Kontexte)*, the variables are *attributes (Merkmale)*, and the domain of the variables is the *objects (Gegenstände)*—see Wolff and Gabler (Chapter 7) and Frick *et al.* (Chapter 6).

In cluster analysis, multidimensional scaling, contingency table analysis, and multivariate analysis, the multivariable is often preprocessed to form a *distance*

Table 1: A Multivariable

	Department	**Born in the United States**
Ferguson	Mathematics	Yes
Li	Mathematics	No
Ylvisaker	Mathematics	Yes
Berk	Sociology	Yes
DeLeeuw	Statistics	No
Mason	Sociology	Yes
Bentler	Psychology	No
Muthén	Education	No
Jennrich	Mathematics	Yes

matrix, a *covariance matrix*, or a *cross-table*. This often involves data reduction, although sometimes the map is one-to-one. This particular step in the data reduction process can be thought of as either the last step of coding or the first step of the statistical analysis.

Also, in some cases, we observe dissimilarities or measure distances directly. This can be coded as a single real variable on $I \otimes I$ or as three variables, the first two being labels.

3 Representation

The process of coding maps the possible outcomes of an experiment into the data space, which is defined by the design. Although in some experiments coding may be relatively straightforward, in others it involves many decisions.

The mappings used in coding are not often studied in statistics, although perhaps they should be analyzed more. Design in the narrow sense is generally seen to be a part of statistics, but codebooks and experimental protocols are usually assumed to be part of the actual science.

What definitely is a part of statistics is the next phase, the mapping of data into representations. We take the data, an element of the data space, and we compute the corresponding element of the representation space. This mapping is called a *statistical technique* (see Figure 1).

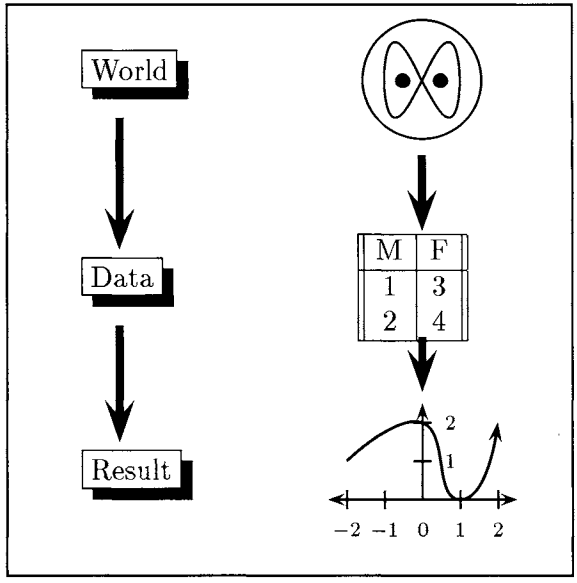

Figure 1

Not surprisingly, many types of representations are used in statistics. In formal concept analysis, data are represented as lattices or graphs; in cluster analysis as trees, hierarchies, or partitionings; in correspondence analysis (CA), multidimensional scaling (MDS), and biplots as maps in low-dimensional spaces. In regression analysis and generalized linear modeling we find many types of scatterplots, either to picture the structural relationships (added variable plots, smooths) or to portray the residuals and other diagnostics. There are very many variations of these mapped plots, and new ones are continually being invented.

In contingency table analysis we now also have graphical models, familiar from path analysis and structural equation modeling. Residuals from contingency table analysis are modeled with Euclidean techniques. We should remember, however, that 500 pages of computer output also defines a representation space and that people look at the tables in CROSSTABS output from SPSS as primitive visualizations as well.

4 Techniques

We have seen that techniques map data space into representation space. What are the desirable properties of the techniques? We mention the most important ones.

- A technique has to be as *into* as possible; that is, it should be maximally data reducing.
- A technique should incorporate as much prior knowledge from the science as possible (this could, however, be prejudice or fashion).
- A technique should separate the stable and interesting effects from the background or noise.
- A technique should show the most important aspects of the data.
- A technique should be stable, that is, continuous and/or smooth.

Some qualifying remarks are in order here. Data reduction cannot be the only criterion, because otherwise we could replace any data set with the number zero, and this would be a perfect technique. In the same way, stability cannot be the only criterion either (same example).

We also need some notion of fit, and this is embedded in what we think is interesting (i.e., in our prior knowledge). In homogeneity analysis (or multiple correspondence analysis) we apply a singular value decomposition to a binary matrix of indicators (also called dummy variables). In analyses using the ordinary singular value decomposition, fit is defined as least-squares approximation to the observed matrix by a matrix of low rank. But in homogeneity analysis we do not want to approximate the zeros and ones, we want to make a picture of the qualitative relations in the data. Thus we look at partitionings, coded as star plots (Hoffman and de Leeuw, 1992).

Also observe that continuity of a technique requires a topology on $\mathcal{D}$ and $\mathcal{R}$, and smoothness in the sense of differentiability even requires a linear structure. This

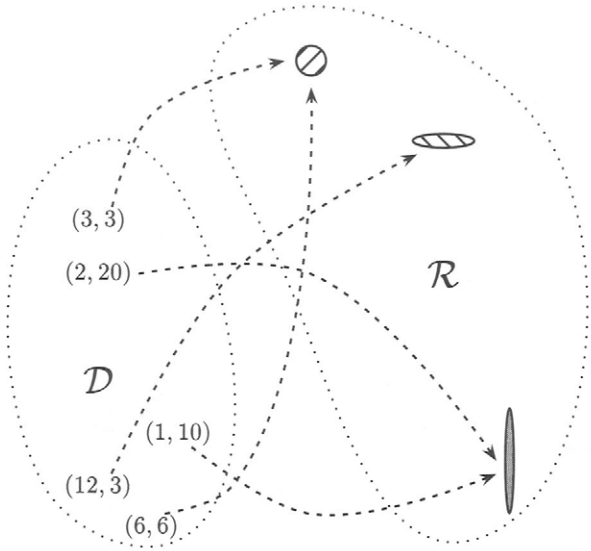

Figure 2

already provides so much mathematical, in fact geometrical, structure that we can almost say we are visualizing the data.

Tentatively, we could maintain the working hypothesis:

All Statistical Techniques Visualize Data

There seems to be some idea that visualization takes place by using representations that are geometrical and that maybe even have the geometry of Euclidean space. This is certainly suggested by the contents of this book, given the predominance of CA and MDS.

But this point of view is certainly much too narrow, because the notions of geometry pervade much of analysis, algebra, discrete mathematics, topology, and so on. Even the real numbers (the real line) are geometrical, and computing a one-dimensional statistic means mapping the data space into the real line (think of confidence intervals, for instance). Again, as with the notion of categorical data, all analysis is visualization, but some analyses are more visual than others.

As we mentioned earlier, it is difficult to draw the line between coding and analysis. Both involve data reduction, and both involve following certain rules. But usually there is a decision to ignore a part of the process and to consider the outcome of this ignored part of the data, which will be fed into the technique.

Very often the technique has multiple stages. We start by reducing the data to a contingency table, or a covariance matrix, or a set of moments, or an empirical

distribution function. This stage is often formalized by the optimistic concept of sufficient statistics, which gives conditions under which we do not lose information.

Only after this stage of preliminary reduction, the serious data analysis starts. Such a serious analysis is often based on a model.

5 Additional Tools

We have discussed data and the forms they take, emphasizing multivariables. We have also discussed the most common types of representations, including simple statistics, tables, graphs, plots, lattices and other ordered structures, partitions, and Euclidean representations. And finally, we have discussed the maps of the data space into the presentation space, which associate the outcome of the statistical analysis with the data in the study.

There are some ideas that can guide us in the construction of visualizations. If the data themselves are spatial, we do not need such guides (and this is recognized more and more often in the use of geographical information systems, or GISs). But otherwise we can use models, and we can try to represent properties of the data as well as possible in our visualizations (using some notion of fit).

5.1 Role of Models

Figure 3 illustrates the use of a model. In this particular case, the model is that gender and size of the ego are independent. The data P are in the upper left-hand corner;

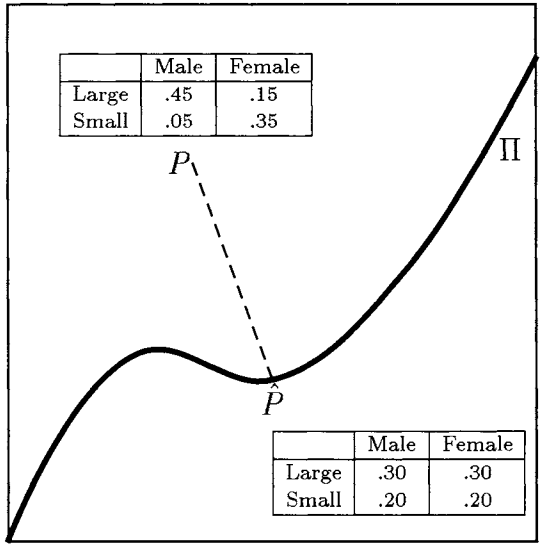

Figure 3

they are proportions in a 2×2 table. The model is the set of all 2×2 tables with independence, a curved surface in three-dimensional space, represented in the figure by the curved line Π. To find out whether the model fits the data, we look at the distance between the model and the data. The statistical technique actually projects the data P on the model Π and comes up with the fitted (or reduced) data $\hat{P}$.

Models are convenient tools with which to capture prior information and to construct statistical techniques. The idea is that a model is some subset of the representation space $\mathcal{R}$ and that prior information tells us that the data, suitably reduced perhaps to a table or covariance matrix, should be close to the model.

This discussion of the role of models covers maximum likelihood methods, the linear model, the t-test, and much of nonparametric statistics as well. It works, provided we are willing to specify a model in the representation space, that is, a subset of that space that we are particularly comfortable with (for scientific reasons, but often only for aesthetic reasons).

5.2 Fit

Figure 3 illustrates one notion, the distance between suitably reduced data and the model. More generally, we may want to know how good or faithful a visualization of the data is. Sometimes representations are very faithful, in fact one-to-one.

Some of this is also illustrated in the pictures that follow, where we first make a graph of the data (Figure 4) and then modify the graph by using multiple correspon-

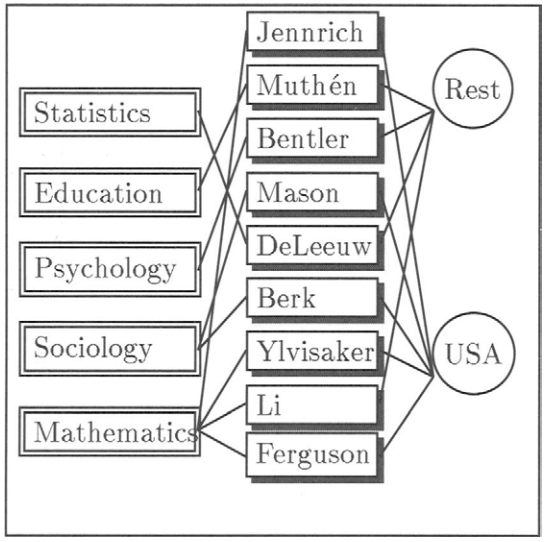

Figure 4

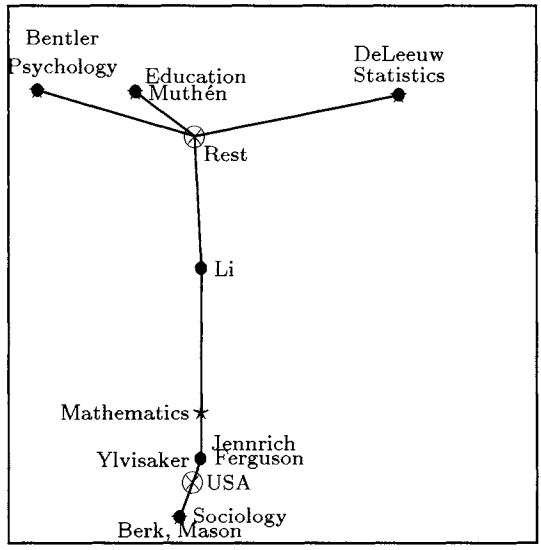

Figure 5

dence analysis (to "make the lines shorter"), shown in Figure 5. As long as the lines are still there, we do not lose information. If we leave them out and interpret on the basis of proximity, we have to guess, and we'll sometimes guess wrong.

In Figure 5 the distances between the statisticians approximate the "chi-squared distances," while the categories of the variables are plotted using the "centroid principle."

6 Visualization of Statistics

Due to the fast personal computer and the bitmapped screen, our day-to-day use of statistics is changing. We can replace assumptions by computations and long lists of tables by graphs and plots.

But, even more profoundly, our interpretation of statistics has been changing too. Moving away from specific calculation-oriented formulas has led to a much more geometrical approach to the discipline (most clearly illustrated in the differential geometric approach, but also in the more applied graphical models approach and of course in the use of GISs).

In a sense, this is nothing new, because modern analysis has been thoroughly geometrized as well. And even in the area of the greatest rigor, that of proofs, a picture is sometimes worth a thousand numbers.

To close with an example of this, an illustration is shown in Figure 6. This is a familiar picture, and it can be used to illustrate many of the basic regression principles.

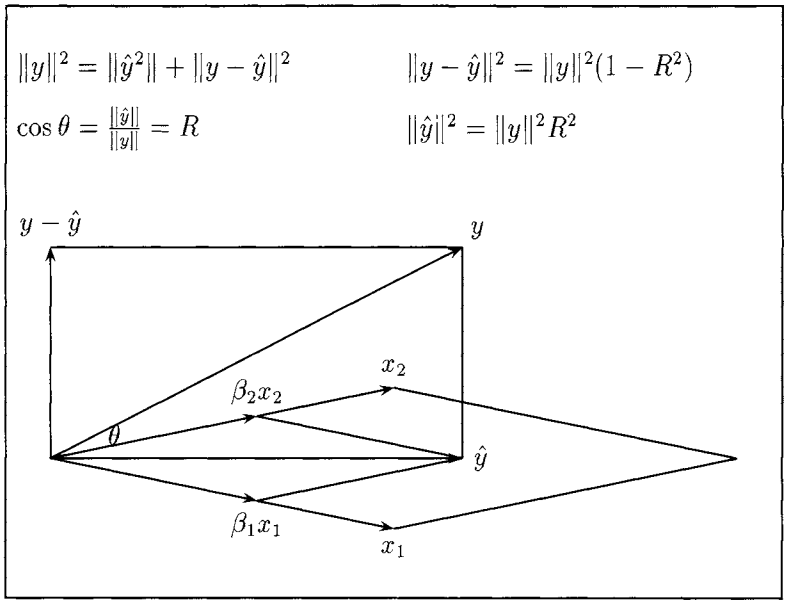

Figure 6

In my regression course, I use it perhaps in 10 of the 30 lectures. It portrays projection, orthogonality *à la* Pythagoras, the regression coefficients, the residuals, the predicted values, the multiple correlation coefficient, and the residual sum of squares. Thus it provides a picture of the basic regression statistics that are printed out by all packages, in a form in which we can use quite a few lectures, of 1000 words each, to explain to the students what is actually going on.