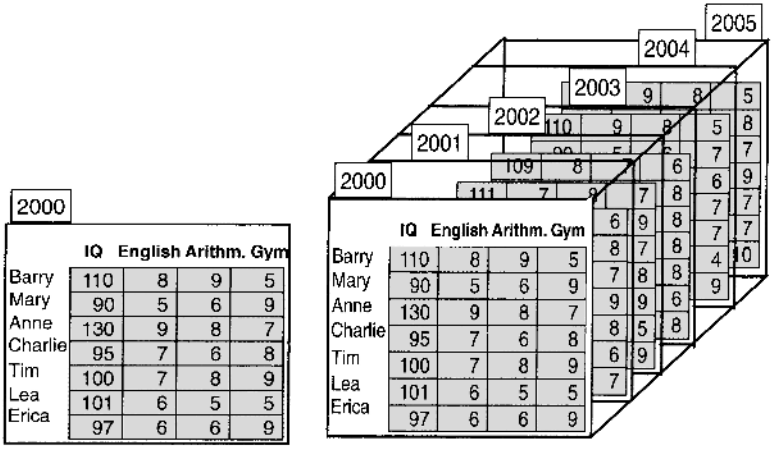
CHAPTER 1

OVERTURE

This book is devoted to the analysis of multiway data, which have a richer and more complex structure than just objects or subjects have scores on a number of variables. Such data are measured several times and/or under several conditions, are ubiquitous, and are collected on a regular basis in many disciplines. They can, for instance, be found in large-scale longitudinal studies and in agricultural experiments and are routinely produced during the study of chemical processes, especially in analytical chemistry. Also, signal processing data and assessments of cracks in the surface of paintings lend themselves to multiway analysis. In fact, it will be shown that scientific research cannot do without multiway analysis, even though not everybody knows this yet. Multiway data can be found in every walk of science, if one only looks. Once alerted to their existence, they can be seen everywhere, in much the same way that first-time prospective parents suddenly notice pregnant women everywhere.

In this introduction, multiway data, in particular, three-way data, are presented and a general introduction is given on how they can be tackled. By means of a detailed example, the value of multiway analysis is brought to the fore as a type of analysis that can be fruitfully used in many disciplines. The introduction will close with a

Applied Multiway Data Analysis. By Pieter M. Kroonenberg



(a) Two-way data (b) Three-way data

Figure 1.1 Storing several two-way data sets in three-way box.

sketch of the origins of multiway analysis and its explosive development since around 1990. The presentation will use primarily three-way examples because they are more common and easier to understand, and because the general practice in multiway data analysis is to count "one, two, multi".

1.1 THREE-WAY AND MULTIWAY DATA[1]

What are three-way data? To put it at its simplest, three-way data are data that no longer fit onto one index card, but need a box to contain them. Looking, for instance, at an example from child studies we note that many data sets there take the form of the scores of a number of children on a number of variables, such as intelligence, and marks for English, arithmetic, and physical education (gym) (Fig. 1.1(a)).

Such data may be written on one index card, in the form of a table with the children in the rows and the variables in the columns. We now have two "ways": Children and variables. If the data form part of a longitudinal investigation in which the children are measured every year on the same variables, we have three-way data, the "ways" being children, variables, and years or occasions. One index card will now no longer suffice, but we need a separate card for each year, and the set of cards is kept in a box: Three-way data fit in boxes, as shown in Fig. 1.1(b).

[1]The text of this and the following section is largely based on Kroonenberg (2005a), which also contains the genealogy of Dutch three-way researchers mentioned in the Foreword. Available from *https://openaccess.leidenuniv.nl/dspace/bitstream/1887/3494/3*. Accessed May 2007.

Once we have realized this, we suddenly notice data in boxes everywhere. A *plant breeder* has planted several varieties of peanuts (first way) in different locations (second way) and measures the characteristics of the harvested plants, such as yield, quality, and the percentage of saturated and of unsaturated oils (third way). A *chemist* has ten containers with different sugar solutions, sends light of varying wavelengths through the containers, and measures the excitation wavelengths emerging at the other side. A *medical pharmacologist* has derived the spatial structures of a number of cocaine variants from crystallographic measurements, and wants to know to what extent their spatial structures are the same, and to what extent they are different. The molecules form the first way, their constituent atoms the second, and the spatial coordinates the third.

It may seem a miracle that a radio mast, which receives signals not only from one's own mobile phone but also from countless other cell phones, knows exactly that the answer to a specific person's outpourings should be sent to her, rather than her neighbor who has nothing to do with the exchange. This process, too, may be analyzed through three-way models.

Extending the concept to multiway data, one might think of the typical plant breeding experiments, which are conducted over a number of years so that four-way data are obtain: peanut varieties × attributes × locations × years. Similarly, any research in which subjects are measured on several variables under various conditions at a number of points in time generates four-way data. Moreover, fMRI scans, which are becoming more and more routine in hospitals and research, also provide multiway data, such as voxels × time points × subjects × trials × task conditions. However, actual situations in which higher-way data are collected are still few and far between. Analyses of such data in their full multiway appearance are even rarer, but they are on the increase.

1.2 MULTIWAY DATA ANALYSIS

Three-way analysis is no more than the analysis of data that fit in boxes, and multiway analysis is the analysis of data that fit in more-dimensional boxes or *hyperboxes*. This statement might not be very enlightening to the uninitiated and it is better to rephrase it as "What type of research questions can be tackled via multiway analysis?" Such a question has the advantage that it is formulated in substantive rather than methodological terms, and we will discuss this in the three-way context.

Let us again look at the example of the children tracked over a number of years. What questions would the researchers have had in mind when they started to collect data? There are of course many possibilities, but let us confine ourselves to those questions that may be handled via three-way analysis. In this case, the *central questions* might be:

- What are the relations between the variables?

- What trends may be discovered over time?
- Are there different types of children?

These are three questions, one for each way. Although these types of questions are interesting, they only apply to one way at a time. Three-way analysis has been devised especially to deal with more complex questions such as:

- Do the relations between the variables change over time? For instance, it is well-known that in very young children intelligence is still very amorphous, but that as they get older some children develop better on some aspects of intelligence than on others. In other words, time brings a change in the structure, that is, the interrelations between the various parts, of an intelligence test.
- An even more complex question is: Does the structure of the variables change over time in a different way for different groups of children, for instance, for boys and for girls, or for children with different levels of mental handicap?

With such complex questions, involving all three aspects, or ways, of the data, three-way analysis really comes into its own. *Plant breeders*, for instance, are interested in the specific adaptation of crops; in other words, they like to know which varieties of a plant will perform well on specific attributes in locations with specific characteristics. In concrete terms, where should one grow what type of peanut plants, in order to obtain peanuts that are specifically suitable for making peanut butter?

One's first acquaintance with techniques for multiway analysis is often a bit of a shock, because of the complications involved in the correct understanding and interpretation of those techniques. However, this is unavoidable, as we are dealing with complex techniques intended to solve complex questions. Testing differences in average lengths between boys and girls is child's play for the average student, but three-way questions are more intricate. However, it is this complexity that can yield deeper insights. The picture opposite the content page of this book, a woodcut by M.C. Escher[2], shows a multiway data analyst intrigued by an apparently unsolvable three-mode problem.

1.3 BEFORE THE ARRIVAL OF THREE-MODE ANALYSIS

How were three-way data analyzed before the arrival of thee-way data analysis techniques? To put it briefly: by flattening the box, or stringing out its contents. In both cases the idea is to make three-way data into two-way data by eliminating one of the ways. Instead of looking at the interactions between three types of units (or ways), one then only needs to analyze two.

[2]M. C. Escher, *Man with Cuboid*. Copyright © 2007 The M.C. Escher Company B.V., Baarn, Holland. *http://www.mcescher.com/*.

Flattening consists of making a two-way matrix out of a three-way array by removing one of the ways. It is typically done by taking the averages over all cards in the box, so that one is left with one card containing means. These may be averages over all years, so that one loses sight of trends over time, but one may also take averages over all subjects, so that individual differences disappear below the horizon.

Stringing out is creating a two-way matrix out of a three-way array by either laying out all the cards in one long row, so that the relation between similar variables at different time points is neglected, or laying out all cards in one tall column, so that the connections between people's scores at different moments in time are lost. The technical term for this procedure is matricization. In all these cases, the data and their analysis are shrunk from three-way to two-way.

Sometimes this may do no harm, because it is possible that a three-way analysis leads to the conclusion that no three-way analysis is necessary: for instance, if nothing changes over time, or if all subjects may be viewed as having been randomly drawn from one single population. However, if this is not the case, the flattening or stringing out of three-way data leads to an unnecessary and sometimes unacceptable simplification.

1.4 THREE-MODE DATA-ANALYTIC TECHNIQUES

Since the beginning of the 1960s, a series of techniques have been devised specifically aimed at doing justice to three-way data, and these techniques were later extended to four-way and higher-way data. They bear such intriguing names as three-mode principal component analysis, multilinear component analysis, three-mode factor analysis, three-way cluster analysis, parallel factor analysis, multiway covariance analysis, multidimensional scaling techniques for individual differences, generalized Procrustes analysis, multivariate longitudinal analysis, and many more of this kind.

Most of these methods have a strongly *exploratory character*, which means that one tries to find the patterns among the elements of the three ways, without *a priori* postulating specific configurations and without applying tests to these patterns. This is partly because it is difficult to specify such patterns beforehand, and partly because hypothesis testing supposes that something is known about the distributions of the scores, which for multiway data is only very rarely the case. It is, however, perfectly possible to determine the stability of the estimated values of the parameters via repeated sampling from the sample in question (*bootstrap* method), but these developments are still in their infancy in three-way analysis.

1.5 EXAMPLE: JUDGING CHOPIN'S PRELUDES

In order to give an idea of multiway analysis, we will look, rather superficially, at a three-mode analysis of data produced by Japanese students, in reaction to listening to (parts of) the 24 preludes by the Polish composer Chopin played by a French

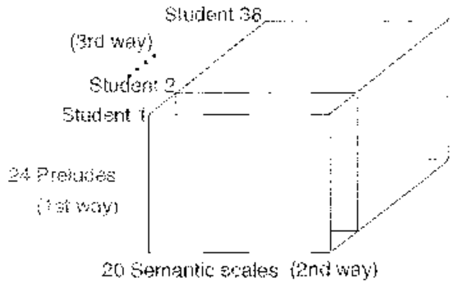


Figure 1.2 The Chopin prelude data set: 24 preludes by 20 semantic scales by 38 students

pianist. The model used for this analysis is the Tucker3 model, which is presented in Section 4.5.3 (p. 54), and a further analysis of these data can be found in Section 8.9 (p. 190). In this example a type of data is used, called "semantic differentials", that actually initiated the development of three-mode analysis, as will be discussed in more detail later (p. 12).

The *research questions* were the following:

1. Classical music may be characterized in *technical* terms such as key, tempo, and mode (i.e., major/minor). To what extent are *nonprofessionals* sensitive to these aspects when we ask them to describe music in everyday adjectives such as loud, gloomy, and tender?

2. Is it possible at the same time to establish, by means of the same adjectives, their preference for specific types of music expressed as key signature, tempo, and mode?

The study was conducted among 38 Japanese students who were familiar with classical music (the first way). The 24 preludes were played to them (the second way), and after every prelude (or the first 80 seconds of it) they were asked to fill out a set of 20 semantic rating scales (the third way); these are given in Table 1.1. Note that these scales consist of two complementary concepts (restless – calm, fast – slow, strong – weak, etc.).

What we are trying to find is the connection between the preludes and the semantic rating scales as related to the individual differences between students. The main result of the investigation was that the students largely agreed on the technical, cognitive aspects of the preludes, but differed as to the affective elements, that is, the sort of music they preferred. This will now be explained in more detail.

1.5.1 Consensus on the musical-technical aspects

Scales. Let us look at the *Semantic space* (Fig. 1.3), in which we see the relations between the 20 bipolar scales as reflected in the students' assessments. Every scale

Table 1.1 Semantic differentials used in the Chopin prelude study

Calm	-	Restless	Gentle	-	Severe
Quiet	-	Noisy	Lyrical	-	Dramatic
Tranquil	-	Vehement	Weak	-	Strong
Slow	-	Fast	Still	-	Loud
Light	-	Heavy	Cheerful	-	Gloomy
Bright	-	Dark	Soft	-	Hard
Happy	-	Sad	Clear	-	Cloudy
Warm	-	Cold	Small	-	Large
Delicate	-	Coarse	Thin	-	Thick
Unattractive	-	Attractive	Uninteresting		Interesting

English translations of the Japanese terms; scores run from 1 to 7.

is represented by an arrow, with one adjective at the tip (*restless*) and its antonym on the other side (*calm*). In the interest of clarity there are only four arrows of which both poles have been drawn. What does it mean that some arrows are close together, and that others are at wide angles to each other? Let us first look at the arrows marked *restless* and *dramatic*. These are close together, because students generally gave preludes similar scores on these two scales. Hence, *restless* preludes are also *dramatic*, and *calm* preludes are also *lyrical*. When the arrows are at right angles to each other, such as, for instance, *fast* and *cold*, this means that according to the students those scales have no relation to each other at all: *fast* preludes may be *hot* as well as *cold*, and the same applies to *slow* preludes.

Preludes. The *Prelude space* shows the configuration of the preludes as a reflection of the scores the students assigned to them (Fig. 1.4). For a correct interpretation of the prelude space, one should imagine the scale space as an overlay on top of the preludes, with the same orientation and the axes aligned.

Using the information from the scales, we can deduce that Prelude no. 16 in b$\flat$, tempo indication *presto*, was judged especially noisy and fast. Prelude no. 20 in c, *largo*, located in the bottom right-hand corner of the Prelude space, was judged by the students as coarse, heavy, sad, and gloomy. However, Prelude no. 15 in D$\flat$, *sostenuto*, in the bottom left-hand corner of the Prelude space, was rated calm and lyrical.

Technical merits. If we now judge the configuration of the preludes in the Prelude space on its technical merits, we note that the students have made a clear division in Tempo (fast and slow) and in Mode (major and minor). In other words, on the basis of the semantic scales the students have arranged the preludes in a pattern that is found to correspond to an arrangement based on musical-technical aspects. This

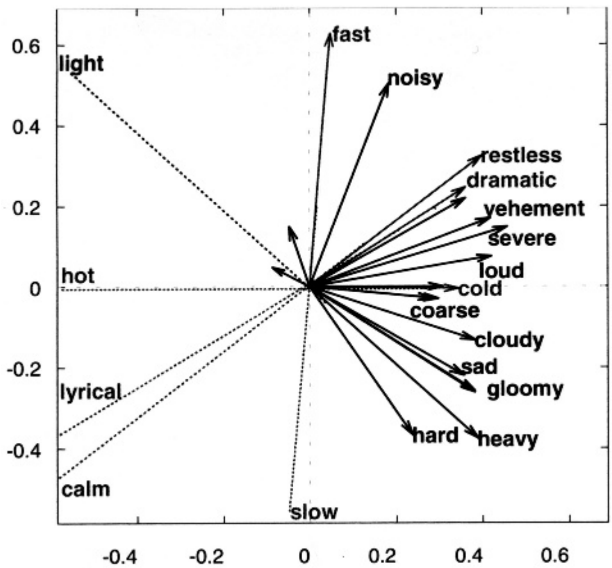


Figure 1.3 Relationships between the semantic differential scales. Narrow angles indicate high correlations between scale; perpendicular scales are unrelated.

is not all, however; closer inspection reveals that the arrangement of key signatures over the area largely corresponds to that of the circle of fifths (see Fig. 1.5). We may now even note two anomalies. These are Preludes nos. 9 and 10: Prelude no. 9 is in a major key and is situated in the "minor" area, and Prelude no. 10 is in a minor key and is found in the "major" part of the Prelude space. The probable cause is that in some longer preludes key changes occur, and the students were offered only the first 80 seconds, which may have given them a distorted impression of these preludes.

1.5.2 Individual differences

Figure 1.4 showed the consensus among the students. What still needs to be discussed is to what extent the students *differed*. There is no need for a figure to illustrate this, because it may easily be said in words. The students especially liked either fast pieces in a major key, or slow pieces in a minor key. Of course, this is as one would expect.

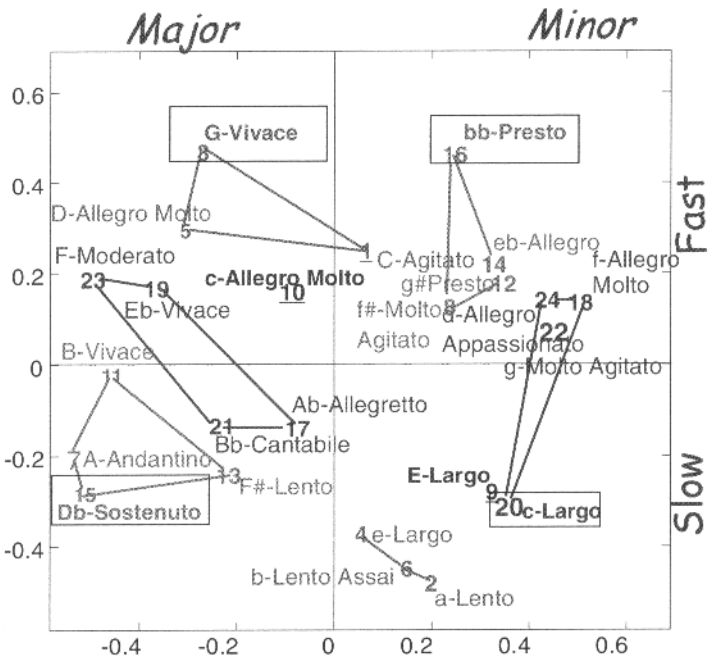


Figure 1.4 Relationships between the preludes. Horizontal axis: Mode -- major, minor; Vertical axis: Tempo -- fast, slow.

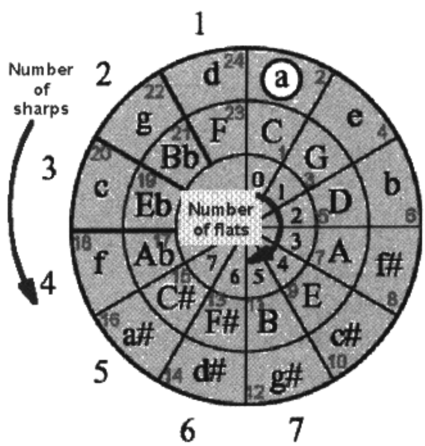


Figure 1.5 Circle of fifths, labeled with the Prelude numbers. Adapted from *http://www.uncletim.com/answers/circle.htm*. Accessed May 2007.; reproduced with kind permission from Tim Gillespie.

1.6 BIRTH OF THE TUCKER MODEL

The history of techniques for three-way data analysis starts with the late Ledyard R Tucker (†2004)[3], who, among other places, worked at the University of Illinois and the Educational Testing Service in Princeton. Another prominent godfather is Douglas Carroll, who spent a large part of his academic career at Bell Laboratories and is now on the staff at Rutgers University. A third founding father of three-way techniques is Richard Harshman, based at the University of Waterloo in London, Canada.

In an interview with Neil Dorans (2004), Tucker had the following to say on this topic:[4]

> Three-way factor analysis grew out of this multidimensional scaling work. While I was at ETS[5], I had observed that Charles Osgood of the University of Illinois had collected data from three modes– concepts, scales, and subjects – in his semantic differential research. I thought that the data should be analyzed differently than it was. He collapsed over people and threw away individual differences data [i.e., he flattened his data box]. *So I developed the 3-mode factor analysis approach, a very general model for evaluating individual differences data. It was able to deal with the variety of people* (p. 8).

Tucker illustrated his objection to "flattening" by the example of the car manufacturer Chrysler, who designed a car for the "average customer" without appreciating the importance of individual differences. Unfortunately, the car did not sell, because there was no such thing as the *average* customer, and Chrysler almost went bust on this enterprise. The necessity of paying attention to individual differences is still the guiding principle for almost all work in the area of three-way analysis.

1.7 CURRENT STATUS OF MULTIWAY ANALYSIS

A full and detailed survey of the development of multiway analysis will not be undertaken here; see, for instance, Section 2.4 (p. 20) and Smilde et al. (2004, pp. 57, 58, for a brief historical introduction). Some idea of its fortunes over the years can be seen in Fig. 1.6, which shows the roughly calculated citation curves for the main protagonists, divided into founding fathers (Tucker, Carroll, and Harshman), psychometricians, and chemometricians. What is most striking in this figure is the steep increase in the number of citations during the 1990s, and the fact that this escalation is almost exclusively due to the chemometricians. Moreover, the rising number of citations for the psychometricians is also due to the stormy developments in chemometrics, because it was in psychometrics that the initial developments had taken place. Especially the models developed by the psychometrician Richard Harshman

[3] Note that R is Tucker's middle name and not an initial, so that it should be written without a full stop.

[4] Reproduced with kind permission from Neil Dorans.

[5] Educational Testing Service, Princeton.

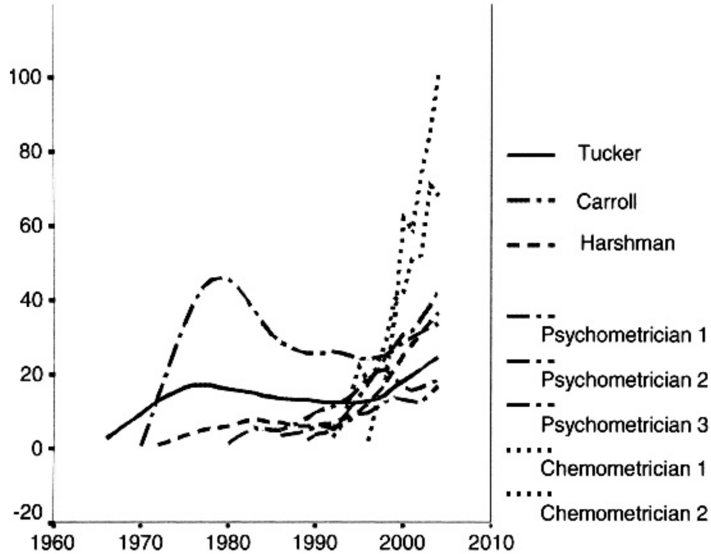


Figure 1.6 Popularity of multiway analysis as expressed in the number of citations per year to work by the founding fathers, three psychometricians, and two chemometricians (1963-2004).

corresponded directly to physical models in use in analytical chemistry. What the chemometricians needed were methods to estimate the parameters of their multiway models and these had been a subject of intensive research in the psychometric field. Moreover, new experiments in chemistry could easily be designed to take advantage of the uniqueness properties of Harshman's Parafac model. Also, important commercial applications immediately offered themselves in the chemistry domain. All these factors led to an increase in publications on multiway analysis.

An interesting aspect of this development is the fact that psychometricians also started to publish in chemometrics journals, because that was where their biggest customers were located. Whereas in the early 1990s applications in the social and behavioral sciences were at a rather low level, applications in applied chemistry were soaring and in other branches of science, such as agriculture, signal processing, medicine, and mathematics, applications can now also be found. This increased interest in applying existing multiway methods also created a considerable drive in further developing technical aspects of the techniques. Foundations were laid and are being laid for mathematical and statistical extensions of what were initially relatively straightforward three-way techniques, to sophisticated multiway methods in various forms.