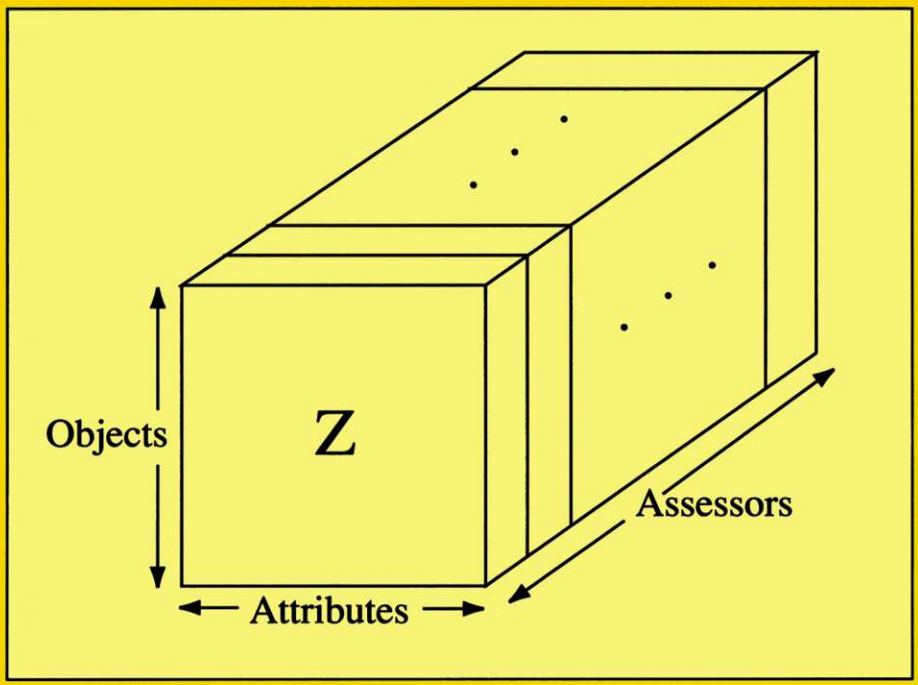
DATA HANDLING IN SCIENCE AND TECHNOLOGY

16

Multivariate analysis of data in sensory science

TORMOD NÆS
EINAR RISVIK
(editors)



ELSEVIER

DATA HANDLING IN SCIENCE AND TECHNOLOGY — VOLUME 16

Multivariate analysis of data in sensory science

DATA HANDLING IN SCIENCE AND TECHNOLOGY

Advisory Editors: B.G.M. Vandeginste and S.C. Rutan

Other volumes in this series:

Volume 1 Microprocessor Programming and Applications for Scientists and Engineers by R.R. Smardzewski
Volume 2 Chemometrics: A Textbook by D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte and L. Kaufman
Volume 3 Experimental Design: A Chemometric Approach by S.N. Deming and S.L. Morgan
Volume 4 Advanced Scientific Computing in BASIC with Applications in Chemistry, Biology and Pharmacology by P. Valkó and S. Vajda
Volume 5 PCs for Chemists, edited by J. Zupan
Volume 6 Scientific Computing and Automation (Europe) 1990, *Proceedings of the Scientific Computing and Automation (Europe) Conference, 12–15 June, 1990, Maastricht, The Netherlands,* edited by E.J. Karjalainen
Volume 7 Receptor Modeling for Air Quality Management, edited by P.K. Hopke
Volume 8 Design and Optimization in Organic Synthesis by R. Carlson
Volume 9 Multivariate Pattern Recognition in Chemometrics, illustrated by case studies, edited by R.G. Brereton
Volume 10 Sampling of Heterogeneous and Dynamic Material Systems: theories of heterogeneity, sampling and homogenizing by P.M. Gy
Volume 11 Experimental Design: A Chemometric Approach (Second, Revised and Expanded Edition) by S.N. Deming and S.L. Morgan
Volume 12 Methods for Experimental Design: principles and applications for physicists and chemists by J.L. Goupy
Volume 13 Intelligent Software for Chemical Analysis, edited by L.M.C. Buydens and P.J. Schoenmakers
Volume 14 The Data Analysis Handbook, by I.E. Frank and R. Todeschini
Volume 15 Adaption of Simulated Annealing to Chemical Optimization Problems, edited by J.H. Kalivas
Volume 16 Multivariate Analysis of Data in Sensory Science, edited by T. Næs and E. Risvik

DATA HANDLING IN SCIENCE AND TECHNOLOGY — VOLUME 16
Advisory Editors: B.G.M. Vandeginste and S.C. Rutan

Multivariate analysis of data in sensory science

edited by
TORMOD NÆS
and
EINAR RISVIK
Matforsk, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway

1996
ELSEVIER
Amsterdam — Lausanne — New York — Oxford — Shannon — Tokyo

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

ISBN 0-444-89956-1

This book is printed on acid-free paper.

Printed in The Netherlands

Contents

This Page Intentionally Left Blank

LIST OF CONTRIBUTORS

P. BAARDSETH	MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway
P.M. BROCKHOFF	Royal Veterinary and Agricultural University, Thorvaldsenvej 40, 1871 Frederiksberg C, Denmark
G. DIJKSTERHUIS	ID-DLO, Institute for Animal Science and Health, Sensory Laboratory, P.O. Box 65, 8200 AB Lelystad, The Netherlands
H. HEYMANN	Food Science & Nutrition Department, University of Missouri, 122 Eckles Hall, Columbia, MO 65201, U.S.A.
D. HIRST	Scottish Agricultural Statistics Service, Rowett Research Institute Bucksburn, Aberdeen AB2 9SB, Scotland, U.K.
E.A. HUNTER	Biomathematics & Statistics Scotland, The University of Edinburgh, Edinburgh EH9 3JZ, Scotland, U.K.
U.G. INDAHL	MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway
T. ISAKSSON	MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway
K. KVAAL	MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway
J. McEWAN	Department of Sensory Science, Campden & Corleywood Food Research Association, Chipping Campden, Gloucestershire GL55, 6LD, U.K.
T. NÆS	MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway
R. POPPER	Ocean Spray Cranberries, Inc., One Ocean Spray Drive, Lakeville/Middleboro, MA 02349, U.S.A.

E. RISVIK — MATFORSK, Norwegian Food Research Institute, Oslovein 1, N-1430 Ås, Norway

P. SCHLICH — INRA, Laboratoire de Recherches sur les Arôme, 17 rue Sully, 21034 Dijon Cedex, France

E. VAN DER BURG — Department of Psychometrics and Research Methodology, Faculty of social and Behavioural Sciences, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands

PREFACE

Data analysis in sensory science has traditionally been performed with univariate statistical tools such as Analysis of Variance. During the last decade, the emerging capacity of computers has made multivariate techniques more appropriate for increased understanding of these complex data. The intriguing data generated by sensory panels and consumers demand models with capabilities to handle non-linear aspects as well as to simplify the large amounts of data, in order to facilitate interpretation and prediction. The development in this sector has been very rapid. From the simple Principal Component models available in mainframe computers in the late seventies/early eighties, 1995 offers a large spectrum of models with very different applications. Common to many of the techniques is the principle of extracting central or common information in large data matrices presented in understandable and simplified formats. Therefore this book starts with a discussion of principles in understanding of results from the Principal Component models, as they can be seen as a conceptual representative for all the families of models represented later in the book.

The present textbook is prepared in collaboration between a group of scientists involved in sensory science at an advanced applied and academic level. Chapters are included to cover basic understanding of the principles behind the methods, experimental design as well as a variety of techniques described in detail. The book has been written to give a reasonable updated selection of *new* methods applied to the sensory field. The editors have intended to generate a book well suited for educational purposes, with sufficient references for complementing the presented text. The authors have all followed the same set of instructions, where theoretical background and practical examples have been given equal importance. The book is made for use by sensory scientists in a practical situation and also in a training situation, to give good and sufficient knowledge about advanced methods applied in sensory science. The examples discussed in the text provide sufficient support for users in a beginner's phase of exploration of the techniques. Statisticians may find the text incomplete, but the references given should provide sufficient additional information for their needs as well. Sensory scientists on the other hand may find the theoretical information on the difficult side - thus, providing room for developing skills/knowledge.

The present text has no ambition to cover all existing techniques available for analysing sensory data. The field is in very rapid development and new and modified methods appear "every day". The chapters illustrate a cross section of what is available on a theoretical and practical level. Not all the presented methods are readily available in practical statistical software, while others exist in many versions implemented in a variety of software. This is a situation in very rapid change, which makes the need for material to help in a process of increased knowledge more urgent than new software. The editors hope this book is a contribution in that direction.

As described above the book starts with a discussion about the conceptual basis for the multivariate methods treated later in the book. One of the main themes in this discussion is the concept of latent variables or latent phenomena. In mathematical terms latent variables can be

referred to as projections or linear combinations of the data, but they can also be given a much broader, deeper and more philosophical interpretation. This is the topic of Chapter 1.

For all scientific investigations, the design of the experiments involved is an extremely important aspect of the whole analytical process. Sensory analysis is no exception and this important topic is covered in Chapter 2. Main emphasis is given to the design of sensory tasting experiments, but treatment designs are discussed as well. The main principles for experimental design are covered and illustrated by examples. The importance of the principles for multivariate analyses is emphasised.

After having presented an account on the philosophical basis for the methods and the main principles for obtaining good data, the next two sections (Part II and III) are devoted to the multivariate methods themselves. Part II focuses on methods mainly developed for analysing aggregated sensory data, i.e. data obtained by averaging data over assessors. Part III on the other hand is devoted to methods which use the individual sensory data for all the assessors in the analyses. The majority of the presented methods in the latter chapter are relatively new and represent an area of research where the goal is to examine both the individual differences and similarities among the assessors and the samples.

Part II has a strong focus on methods for relating sensory data to external information, but techniques for analysing the sensory data themselves are also presented. Chapter 3 gives a treatment of the main principles for so-called preference mapping. This is a family of methods for relating sensory and consumer data based on graphical illustrations of latent variables. Chapter 4 is a study of the potential usefulness of applying neural networks in sensory science. The neural networks are non-linear and flexible methods for relating data matrices to each other. So far they seem to be little used in this area of research. In the same chapter a brief discussion of the well established linear methods PCR and PLS is also given. Chapter 5 treats the important field of extracting information from images which can be correlated to sensory properties of the products. This is a quite new area of research and the present contribution discusses some of the basic principles for it.

Chapter 6 is about MDS. Both methods for aggregated and individual profile data will be presented. The data needed for this type of analysis are distance- or similarity data about samples and can be obtained directly from a specially designed sensory experiment or from computations on data from standard sensory profiling. The important method of GPA is the topic of Chapter 7. This is a method which translates, rotates and shrinks the individual sensory profile matrices in such a way that the match among them is optimal in a LS sense. The STATIS method is another method for individual profile data which looks for similarities and differences. This method is treated in Chapter 9. The method is based on maximising correlations among the profiles and provides plots/maps of a similar type as most other methods treated in the book. GCA is a non-linear generalisation of CCA and is presented in Chapter 8. The method can handle several sets of variables and is based on latent variable plots of the main directions of common information in the data matrices. The different matrices will typically represent different assessors, but can also represent for instance sensory, chemical and physical data from the same set of samples. The last chapter of the book is about 3-way factor analysis methods. These methods are generalisations of the standard PCA and look for common latent structures among the assessors and/or the samples.

Comparisons among the methods are treated in the different chapters of the book. Some of the methods are developed for quite different situations, but in many cases they can certainly

be considered as complimentary to each other. This will also be discussed in the different chapters.

The book ends with an index of all the chapters.

We would like to thank all contributors to this book for their valuable contributions. It has been a pleasure to work with you! We will also thank MATFORSK for allowing us to work with this project. FLAIR-SENS is thanked for financial support to some of the contributions at an early stage of the planning process.

Tormod Næs and Einar Risvik, June 1995.

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

UNDERSTANDING LATENT PHENOMENA

E. Risvik

MATFORSK, Norwegian Food Research Institute, Osloveien 1, N-1430 Ås, Norway

1. INTRODUCTION

The first chapter has the ambition to introduce the multivariate thinking this book is based on, in order to provide a frame of reference for the reader. To be able to understand later chapters the «philosophy» of multivariate analysis is discussed. This is done with two different perspectives. First of all the conceptual understanding is focused, that is the understanding per se. The intention with this has been to exemplify the multi-disciplinary understanding needed in order to utilise the true potential inherent in data from sensory science. To illustrate this, sensory profiling is chosen as an example. This is not the only approach to generate complex data in sensory science. Other approaches can be scaling methods, projective methods, magnitude estimation, questionnaires, observations, consumer responses of different kinds, comparisons, rankings and methods close to profiling like free choice profiling. Conventional profiling is chosen as this is very commonly used, very convenient for this discussion and because the author has the most experience with this method. The second perspective into this material is through a few sensory examples, simple in their interpretation. Non-statisticians' examples are chosen, and these are focused on the understanding of results from conventional profiling, rather than the statistical terminology. This is not a discussion in depth, rather a framework which can be useful when reading later chapters.

Necessary statistical terminology to understand the following chapters will not be introduced here, as the different authors are expected to do this.

For the practical statistical example the frequently used Principal Components Analysis (PCA) is utilised as an example. PCA fits well for sensory profiling data and at the same time illustrates some of the more fundamental concepts involved in analysis. The other methods represented in this book cover a wide variety of approaches, (non-linear artificial neural networks, multidimensional scaling, generalised canonical correlation analysis and three-way factor analysis) and will be introduced by each author.

This chapter does not pretend to give complete answers to any of the very fundamental and complex questions related to understanding of sensory data, but rather to open up for exploratory, creative and critical thinking around the potentials for utilisation of multivariate statistics. Some of the expressed views can be interpreted to be contrasting to each other. This is intentional, as many of the fundamental discussions touched upon provide no answer to these questions. To some extent the «choice of the right model» lies more in a choice of belief rather than fact. The author has therefore made a point of presenting the discussion between philosophies of science, and left the answers to be found by the reader.

The aim of this chapter is to show the multi-disciplinary thinking needed in research where sensory data are analysed by multivariate statistics. This perspective is chosen in order to build the links between cognitive science, psychology, physiology, experimental design and basic philosophy. Several of the very basic problems in sensory science are related to this interface.

2. TASTE, EXPERIENCE AND CHEMISTRY

Drinking a glass of wine can be described in many ways. The wine can have a great potential, be a bit masculine, with a strong body and a good balance between fruity aromas and acids, and it can have a long aftertaste. Said in a different language the wine can be astringent, have strong aromas from black currant, plum, asparagus, and a hint of burnt match, kerosene, vanilla and hazelnut. The same wine can also be characterised with it's absorption spectrum or a gas chromatography profile. Other chemical analyses can also contribute strongly to the chemical characterisation of the wine.

In total there are more than 800 aroma components (Rapp, A. 1988) in wine. Together these compose a complex aroma structure, experienced when wine is consumed. The experience consists of chemical components in interaction with our senses, and the interpretation of the perceived entity by the individual.

The traditional approach by chemists has been to identify and quantify components one by one for at a later stage to ask about their importance. This has been a tedious task, as very often the analytical tool has needed development at the same time. Analytical chemistry today is capable of identifying and quantifying very low concentrations of advanced molecular structure (Rapp, A. 1988). When identification and quantification is no longer a problem, it is relevant to ask questions concerning the importance of the identified components.

After most of the influential variables have been characterised the problem of how the wine tastes is still not solved. The human perception translates and expresses this in a far less explicit vocabulary than chemical analyses, and as such is very difficult to model. It is not at all obvious that the most abundant components have the biggest influence. Very often, in the work with off-flavours, one finds taints associated with components present in very small concentrations, but at the same time giving rise to strong affective reactions.

Aromas are today described with several hundred words, and they show no apparent logic. Several psychologists have tried to classify aromas according to a standard nomenclature. The Zwaardemaker classification of smells into nine classes (1895), refined by Henning (1916) into six classes are examples of this (see also Wada, Y. 1953). The relationship between these proposed structures and an inner and perceived structure is not obvious, and certainly not verified in a satisfactory way. The terminology for wine description in itself shows strong inconsistencies. This is best characterised with examples of two types of descriptors: The descriptive aroma terminology (Noble, A.C. *et al.* 1987) with descriptors like black-currant, nutmeg, peach and black-pepper, and the more complex terminology with attributes described by words like body, potential and harmony in its simple form, and in the most abstract form words like feminine, cosmopolitan and stylish. More than 1700 words (Peynaud, E. 1987; Rabourdin, J.R. 1991) are being used for this purpose from the last groups of complex and integrated attributes.

If the words used to describe a wine were all unique descriptors, each related to one aroma component, our vocabulary would still be sadly insufficient to express the experience of drinking a wine. Like with colours, where the eye can distinguish 6 million different shades of

colours, it is likely that there must be another, underlying and simpler structure which can be used to relate descriptors and aroma components. For colours this is often referred to as the colour space: a three dimensional structure where all colours can be described by a set of three orthogonal variables, the dimensions of the space. This will be discussed in some detail later.

For flavour/aroma, one must conclude, there is no such simple structure known. For texture perception, simplified structures are indicated (Risvik, E. 1994; Harris, J.M. *et al.* 1972), although they cannot be understood to be finalised models and texture must be understood to be of similar complexity as for flavour/aroma.

This chapter will deal with aspects related to the understanding and interpretation of sensory data with two different perspectives:

PART I: Understanding concepts related to multivariate data analysis

This will simplify the process involved in understanding of how statistical tools extract latent structures and how they can be utilised for understanding of products. The second section will thus be:

PART II: How data structures can be extracted and interpreted from a set of data utilising multivariate statistical techniques.

This second section will visualise the mathematical principles involved in analysis, with graphical representations and simple examples to illustrate this.

PART I

3. UNDERSTANDING CONCEPTS RELATED TO MULTIVARIATE ANALYSIS

3.1 The role of perception

The perception of a product can be interpreted with at least 2 different perspectives. The sweetness of a carrot tells the individual that the carrot is sweet, which is obvious, but it also informs, in a general sense, that the senses have the capability to perceive sweetness, which does not have to be the same. One example from sensory science, where this is not true is in the case of bitter blindness (PTC) (Brantzaeg, M.B. 1958; Amerine, M.A. *et al.* 1965), where two humans will perceive different qualities (bitter and no bitterness) in the same sample. The stimulus can be argued to be the same, but the perceived entity is different for the two individuals. As all individuals only experience their own perceived intensities there will be no proof that this is the exact same perceived quality or intensity in all humans, still most of the time we take this for granted. Another example is androstenone, the hormone found in pig meat. This chemical signal component gives strong negative reactions in some humans (mostly females) and is not perceived at all by others (mostly males) (Dorries, K.M. *et al.* 1989; Wysocki, C.J. and Beauchamp, G.K. 1984; Wysocki, C.J. *et al.* 1989).

In sensory experiments these two perspectives, both the chemical signal (the given signal) and the human response (the meaning of the information), have potential interest for the

experimenter, and these are difficult to distinguish depending on the given design of the experiment and also in the interpretation of results. To a large extent this is a similar difficulty as distinguishing between latent and manifest structures. Latent meaning: ...a structure expressed in terms of varieties or variables which are latent in the sense of not being directly observable (Kendall, M.G. and Buckland, W.R. 1957), and manifest meaning: to make evident to the eye or the understanding; to reveal the presence of or to expound, enfold (The Oxford English Dictionary 2. ed. 1989). This needs some explanation. Consider the following example as an illustration.

To purchase an apple can be difficult for someone with no previous concept of what an apple is. Even when having been told in advance, (that is when the concept of an apple has been communicated, but with no previous experience), it will be a very time consuming and difficult task to choose among apples on display in a supermarket shelf. In order to say «it looks juicy» (latent structure), it is necessary with previous experience of juiciness in apples. To be able to explain why an apple «looks juicy» is even more difficult. It demands an understanding of how previous experience relates to visual keys (manifest variables), for there has at this point not been an experience of the juiciness of this actual apple. The only part of experience available for interpretation has come through the eyes.

Most persons do not analyse the situation in this way, they just know which apple to choose. To buy an apple is normally a very fast decision with no conscious evaluation of specific attributes, nor detailed consultation with previous experience. If this was the case, shopping at the green-grocers would be turned into a major undertaking.

Implicit in this example lies the assumption that humans organise experience into simplified structures (latent structures) used for consultation when some decisions are to be made. The more experience collected (manifest variables), the more conceptual structures are being formed. The process is getting very complex when previous experience interacts with perception. When we know the brains ability to reconsider, discuss experience and to change opinions based on exchange of information with other humans, this model becomes not only difficult to understand, but also susceptible to changes over time.

In order to be able to buy «the right» apple it is necessary to know what an apple is. The data available in the experience, the «apple» data, aggregate in «apple related phenomena» or latent «apple» structures in order to simplify the search for the «right» apple. Rather than scanning through all previous times an apple was seen, eaten or talked about, the latent structures or concepts are consulted in an upcoming «apple» situation. This makes the search simpler and faster.

This situation, where «apple» concepts are being formed is very similar to analysis of data from a sensory profiling exercise, where first the experience data base is generated as profiles (apple data are provided by a panel). Then the data base is used as «apple» data in order to describe «apple variation». Finally the data are used (a statistical model) in order to calculate central tendencies in data structures, which can later be used for predictions.

What an apple is will after this, for an individual, default into a conclusion which might be as simple as «dark green apples are always juicy», as no previous experience contradicts this assumption. This conclusion or central tendency does not necessarily represent a conscious or permanent or fixed structure, but stands until the experience data base indicates that this is no longer valid. In a similar manner, central tendencies are extracted from data from sensory analysis. The experience is in this case given by the experimental design and the central tendencies in the data are represented by the latent structures.

Both the perceived and the calculated structures are representations of data collected from the objects. Both structures are simplified and are being used as a reference for conclusions. Both sets of initial data contain sensory information. It is therefore of great interest in sensory science to see how these two paths create similar/different conclusions and to what extent one can be utilised as a tool to understand the other in a better way.

To summarise: Manifest (directly observed information) or «given data» (Idhe, I. 1986), as they are interpreted in the perception of apples, show great similarities to data from a sensory profiling exercise of apples. In both situations latent structures («meant» structures, Idhe, 1986) are extracted. In the one case by human interpretation/translation and in the other by a statistical model.

In a more condensed and maybe complex way this can be expressed as in the following:

In sensory analysis the latent phenomena can be observed through a reflection or a projected structure, performed by a statistical model. Similar analyses are performed by the senses, where the latent structure is a reflection of underlying phenomena in the human being, based on experience in this sample space. The manifest structures are represented by the objective differences between the samples described by basic attributes. Since these two views are difficult to separate it is not obvious how each can be characterised in experiments without influence on the other.

3.2 Data from sensory analysis

In a sensory profile the attributes are rarely made up from simple stimuli, each related to one observable variable. Most attributes utilised in a profile are already complex responses to mixtures of several visual, chemical or structural components (examples are fruity, rancid, juicy and sweet). This turns most sensory attributes themselves into latent structures, as they only can be understood indirectly through observations. In statistical analysis, where attributes contribute to yet another level of latent structures, these are of course also latent.

In our minds eye the product is never perceived as a sum of attributes. Whether we focus on key attributes, aggregate attributes into concepts, perceive holistic forms or make up an iterative process with a mixture of the above is not certain. Most likely our consciousness contains at any time a totality of fewer attributes/concepts than a complete sensory profile, when a product is being perceived. This implies that some form of aggregation of information will take place in our information processing. Whether this takes place in our senses or in the processing in the brain is not implied.

The previous mentioned descriptors utilised as a part of wine terminology exemplify also the degree of complexity involved. Words like nutmeg, vanilla and asparagus can be understood as a descriptor related to chemical components, while «potential» hardly can be characterised as related in a simple way to chemistry. Still it is possible to understand the potential of a wine as either high in the intensities of colour tone and fruity, floral notes, or it can be understood as a wine high in colour intensity and astringency. These two different ways of interpreting the descriptor potential, could both be possible interpretations of the research performed by Sivertsen, H.K. and Risvik, E. (1994). It also shows that more complex descriptors can be related to well defined «simpler» descriptors, which again imply that even more complex descriptors like feminine and cosmopolitan could be defined as latent structures inherent in other latent structures. It is also possible to understand how these complex relationships are susceptible to large individual variations as word rarely are well defined in

their relationship from one level of complexity to another. Another way of expressing this is to say that these are words built on several incomplete layers of latency. In many ways this resembles very much the world described by fractal geometry (Gleick, J. 1987) and as such build a very interesting link between several areas of research including cognitive science, linguistics, chaos research and statistical modelling. This also resembles the philosophy of neural networks or fuzzy logics. The interaction in all these fields are of great interest to sensory analysts. There are two reasons for this. First of all because these tools can be of great importance in simulation of human processing, but also because the interaction between human processing and simulated processing can reveal new knowledge of several of the still open questions in both sensory science and cognitive research.

3.3 Latency, complexity and holism

Trying to understand language development is not the aim of this chapter. Still words are of great importance in sensory profiling. It is impossible and not very practical to avoid attributes chosen from different levels of complexity, simply because levels of complexity do not exist and were never defined. It is therefore of utmost importance to perform analysis of data with capabilities to handle this aspect, and to utilise this information for interpretation in the analysis.

One different and important aspect of sensory profiling language, not yet discussed, lies inherent in the nature of the words. To generate a vocabulary for descriptive analysis, the words are chosen to profile, or to describe a projection of an aspect of the product. The idea is that when a comprehensive vocabulary is developed this will together describe all aspects of the product and thus make up the whole of the product. The assumption that the sum of the parts make up the whole is in this case not necessarily true, as several aspects of product perception also are related to complex words describing so called holistic aspects and these can, because of the nature of these words, rather be understood as semiotic signs, and therefore not always be suitable for profiling. The word Quality is one such word, and one of the few words to describe a holistic experience, others are Form, Essence, Beauty and Preference. In contemporary natural science these play a minor but increasingly important role, maybe because of the influential reductionist tradition most of these sciences have followed since Descartes (1596 - 1650).

The first serious attempt to describe a classification of sciences was written by Francis Bacon (1605, *In: Bacon selections*). This classification incorporates both natural sciences, humaniora and metaphysics. With the reductionist tradition since Descartes these have become contrasts and sometimes in conflict to each other rather than aspects of a holistic scientific view, as intended by Bacon. Bacon's view may in many aspects seem old-fashion in a modern world, but this basic thought, that all sciences join together as parts of a whole can also be interpreted as refreshing (not new, but forgotten). The return back to basic logical deductions, describing Mathematics as a branch of Metaphysics, is to consider Mathematics as apriori representation of attributes while Metaphysics «...handeleth the Formal and Final Causes». This makes the representation in Mathematics a part of the understanding behind the real cause of the experiment. In sensory science, the semiotic representation of the object as a sign has to be interpreted with both these perspectives. First of all, the representation of the attributes, the true description is understood through a mathematical description in sensory descriptive profiles (apriori information). At the same time, this has no interpretation unless the meaning is sought at a metaphysical level (posteriori information).

The relationship to Poetry (on the level with Natural Science by Bacon, that is above Metaphysics) is an even more challenging thought. As Metaphysics seek the formal and final causes for observations, Poetry denote communication (delivery) and interpretation (thinking). For Bacon it is of equal importance to seek language representation of complex ideas, as to be able to communicate the results.

The paradox in sensory science is that already sensory profiles are represented in language. Interpretation of latent structures is already sometimes a metaphysical problem, as it looks for causes behind observed structures and relationships back to formal interpretation. When this is to be communicated it is raised to yet another level of complexity, and a component of individual artistry cannot be avoided.

This brings us back into a circle when sensory profiling perspectives can be defined as a fuzzy latent structure, given the nature of the words and how these are utilised in the language. This is again an interesting observation. Since words of holistic nature can be interpreted as a latent structure built on several hierarchical levels of other latent phenomena, and since they at the same time can be seen as primary attributes with holistic characteristics, this can be interpreted as a network of interrelated attributes with large overlaps and feedbacks, all the characteristics of a typical neural network.

3.4 Latent structures in sensory data

There might be several reasons for the development of sciences in a reductionist direction, and one of these can be related to the complexity of the problems to be explored. The investigation of holistic aspects will always have to deal with a large number of variables with strong influence on the problem. This has been a problem until recently, where computers have made development of statistical tools possible, where large amounts of data can be analysed simultaneously. The statistical methods have been available for quite a long time, as principles (Cattell, R.B. 1952; Horst, P. 1965; Harman, H.H. 1967; Wold, H. 1966), but practical applications have been delayed until computers were manufactured and made the analysis feasible.

Roland Harper (Thomson, D.M.H. 1988) were one of the first to apply a factor analysis on sensory data, as early as in the late forties. In a presentation at the Food Acceptability symposium in Reading in 1987, he told the audience that an analysis of 15 attribute profiles took a month to complete, when this was performed by a group of students, without the aid of electronic devices. In the early eighties a similar analysis would take as much as half an hour on a main frame computer, and in the early nineties, less than one second on a personal computer.

This is also reflected in the amount of literature available. In an overview by Martens, M. and Harries, J. (1983), they report 225 papers with applications of multivariate statistical analysis in food science from 1949 to Sept 1982, one third of which have been published after 1980. Few of these papers are related to sensory analysis. A search in the most commercial bases in 1994 give more than 400 articles related to sensory analysis alone, published after 1980.

Most applications in sensory analysis generate vast amounts of data. To get a good understanding of the information buried in this, a reduction of the information, to a reasonable size, is necessary. In addition, sensory variables, like in a descriptive profile, are always strongly intercorrelated. In a descriptive profile it is not unusual to find 15-20 attributes. This is not because there is necessarily 15-20 unique attributes describing these products. Most of

them will be interacting and overlapping and maybe some will be unique. The majority of the attributes will only show slightly different perspectives into the understanding of the product.

With this perspective on the analysis of profiling data, it is rare to find papers which describe more than 10 unique latent dimensions. Most often 1-3 dimensions contain the essential information in the data (each dimension being a combination of attributes) (Martens 1986). This is of course very dependent of the product and panel in question, but the range between 1 and 10 covers quite well. To some extent this is a reflection of our processing capabilities more than our senses. Trincker, D. claimed in 1966 that our consciousness only perceive one part in a million from what our senses collect of information. Ten years earlier Miller (1956) published an article in Psychological Review where he claims that humans only can handle 5-9 independent phenomena in our consciousness simultaneously. Together they both contribute to the assumption that the complex human sensory perception is reduced to a maximum of 5-9 independent structures in the human consciousness. This opens up for a very interesting discussion on effects of training of assessors for sensory analysis.

In sensory profiling we attempt to train assessors to score intensities of attributes on a scale. In expert evaluations of for instance coffee and wine, the training very often starts with knowledge of coffee or grape varieties. From this the experts are trained to recognise characteristics, typical of grape and bean variety, processing, storage and blending. Later, in a profiling situation it is possible these experts do not only profile the given samples. They may also, by unconscious information processing, recognise the coffee bean and the grape variety, and immediately score the attributes they know by previous learning should be present and not only the way they appear in the samples (Cabernet Sauvignon as having blackberry aroma and dark roasted coffee as not fruity). In this case the consciousness is overruled by preconceived information, established in strong latent structures already available and triggered by the perception of the product. This shows how concept formation can be an important aspect of panel training, as it is in everyday life. And from a sensory point of view it is very important to understand this in detail, in order to reveal conflicting approaches in methodology, such as asking for information related to strongly established concepts such as preference, at the same time as the individual is asked to rate intensities of attributes in a profile. It is obvious that preferences, in this case, can influence strongly the profiling exercise. The nature of this effect is difficult to establish, since this is not necessarily a conscious process, and also with a strong individual component.

As statistical analysis sometimes is supposed to reflect perception of a product, either as a whole or to describe aspects of a product, it should not be far fetched to suggest that the statistical analysis is some sort of analogy or reflection similar to the human perception of food products. Each perceived dimension, latent in the product is then composed of a contribution from the product attributes. It is interesting to note that different products may have very different dimensionality (wine is reported to have up to seven dimensions (Sivertsen, H.K. and Risvik, E. 1994) while whole meat texture have two or three (Risvik, E. 1994). In addition, different persons will be able to perceive different number of dimensions at different times. Some interesting questions for us in analysis of these data will then be:

- Does information from different individuals contain commonalties?
- Is it possible to simulate the human processing in a statistical model?
- Is it possible to describe dimensions in a product which will be understood in very similar ways by humans and in the presentation from a statistical software?
- Are there similarities in the models based on large groups of people?

- Are there common denominators between these groups?
- Or are some dimensions common to all humans?

These questions are not for the author to answer, although it would have been nice to be able to do so. The last will be touched upon as a part of the Colour section, and to some extent under experimental design.

The statistical models calculate latent phenomena from sensory profiles. Most of the time the results are presented without any further comments, as if this in itself contain information. In other cases these are interpreted by individuals seeking resonance in themselves for structures that remind them of a previously familiar concept. When this is recognised it is said to confirm a hypothesis. By tradition experiments have been conducted to confirm/disconfirm pre-set hypotheses incorporated into the experimental design. This implies that the experimenter already before the experiment has made up his/her mind as to what can be deduced from the performed test. In the exploratory nature of multivariate statistics it also opens up for an approach where multivariate statistics can be understood as an interactive tool between experimenter and data. Very often, analysis of sensory profiles generate more hypotheses than they solve. Latent structures appear to be of similar nature to previous experiments, as in the case with wine (Sivertsen, H.K and Risvik, E. 1994; Pagés, J. *et al.* 1987; Heymann, H. and Noble A.C. 1987) and whole meat (Harries, J.M. *et al.* 1972; Risvik, E. 1994). These structures show resonance between papers and also resemble sensible models in a phenomenologist tradition. Still no causal proof in a determinist tradition exist. Meta analysis (Mann, C. 1990; Longnecker, M.P. *et* al. 1988) would have been an attractive tool for further analysis, but this would again add, yet another latent layer in the interaction between the deductive and inductive thinking implied in this approach. In the mid 90`s meta analysis is still only available at this conceptual level, as no practical tool for analysis is made commercial available. Interpretation of results from sensory analysis rely therefore very much on verification through previous experience.

3.5 Processing the information

Different perceived aspects of a product can best be separated for independent discussion, as the information is perceived through independent channels. It can be discussed whether statistical analysis also better could be performed separately. These areas will be the basic aspects for definition of the variables utilised for description of a product. These variables may in their own sense build latent/independent structures reflecting the way the senses have organised their information collection. For some senses like vision this is well described, but for others like odour/flavour perception this is not at all established yet.

3.6 Colour

The first suggested structures for colours are very old. The principle in these systems is illustrated by the colour space from the Natural Color System (NCS) (Figure 1). The best known one is probably suggested by Goethe (1749-1832) in his colour system. The description of the perceived structure is later refined and described in the Munsell Color System (1929), but the initial and general structure is still maintained. This consist of a three dimensional space with directions described by the grey scale from white to black, colour intensity and the hue described by the colour circle. This three dimensional structure corresponds well with the structure of how the sensations are collected from the eye, with one channel for lightness and two for colour separation (red/green and blue/yellow) and as such could be understood as a

manifest structure. The perceptual space can be described as a double cone which is expanded in the areas where the eye is most sensitive (yellow-green area). This space is described in Figure 1. To generate the perceived space from spectral information has demanded a lot of labour as it has shown difficult to find a good transfer function between the standard observer to the perceived space.

An illustrative example on how this structure can reveal itself from not too obvious data is given by Kobayashi (1981). He collected information on colours using coloured samples and had them profiled with a series of attributes which describe strongly emotional aspects of colours, such as: polite, reliable, wild, modern, stylish, safe, forgetful, conservative, happy, vulgar and cultivated. The aim of the study was to look for commonalties in colours with reference to fashion and design.

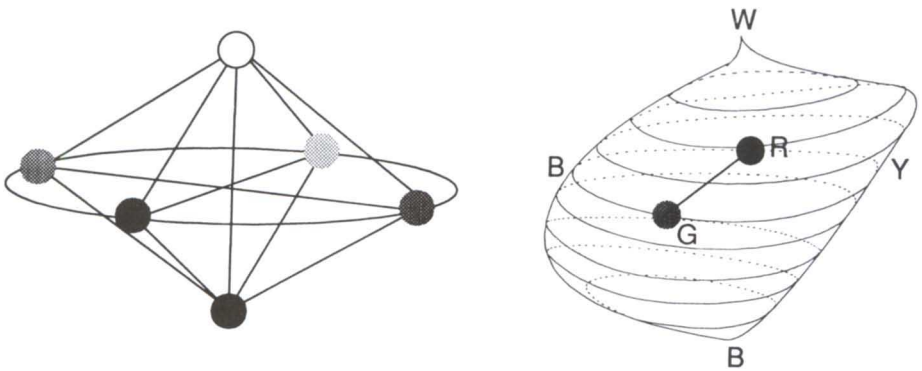


Figure 1 a) Colour space from NCS (Natural Colour System) where all colours at maximum intensity are defined to be 100 % and b) adapted from the Munsell Color System (based on perceived differences). Note the expansion of the space in the green and yellow areas, where the eye is the most sensitive

The first three dimensional solution of a factor analysis revealed an underlying space very similar to the perceptual colour space (Figure 2) and it is remarkable to see this emerge from data based on evaluations of colours with these highly emotional descriptors.

These latent structures extracted from perceptive data resemble very much structures which also can be interpreted as derived from manifest response curves for the eye receptors. This is a very interesting resemblance and further research should be conducted to establish how this has occurred. Multivariate techniques have the potential of revealing this type of information, as the principle of the techniques are based on the calculation of latent phenomena. This fact is surprisingly little used in experimental design and analysis, but the potential for cognitive/sensory research in this field is great with these methods.

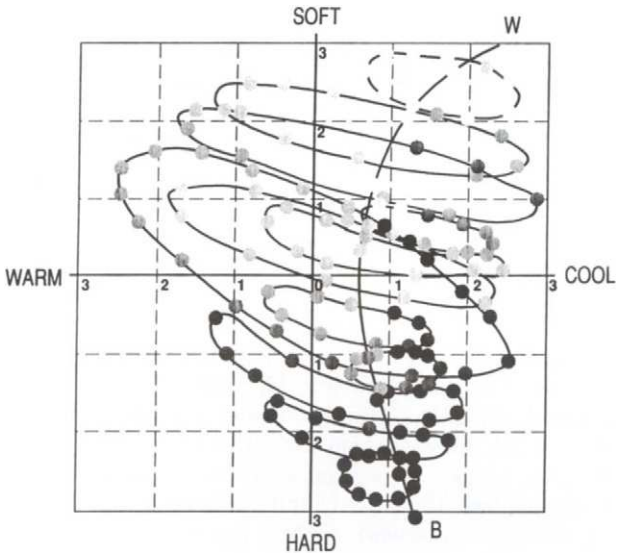


Figure 2 Adapted from Figure 8 plate 3 Kobayashi, S. 1981 The aim and method of the color image scale. Color, 6, 2, 93-107. Adjectives corresponding to color

3.7 Taste

Taste is still maybe the most simple area to describe the structure of. Most people agree to a description of the 4 basic tastes: Sweet, sour, salt and bitter. There is also a discussion whether there are at least two more, and umami and metallic are among the candidates.

Four or six unique tastes make up 4 to 6 dimensions. In practice these dimensions seldom will be orthogonal in a product space, since these attributes always will have certain degree of covariance in a product. For evaluation of a natural product range, such as in the ripening of fruits, it is also possible to observe a high degree of correlation (negative) between attributes such as bitterness and fruitiness.

3.8 Smell/flavour

A generalised structure for odours/aromas have been sought for centuries, as mentioned earlier. So far, to the author's knowledge, no obvious model based on the understanding of the structure for collection of aroma information through the senses, is available. Enormous amounts of information on flavour materials exist, and a model for structure, for food flavours is suggested by Ashirst, P.R. (1991) and Yoshida, M. (1964). These models would become very complex once expanded to all aromas/flavours, but would provide excellent starting points for such research.

If the earlier assumptions hold, a similar approach to the understanding of the colour space can be applied by the sensory scientist, to reveal valuable information into the difficult area of how aromas are organised in simplified mental structures. Generalising the previous assumption that the human consciousness can handle no more than 5-9 independent pieces of

information simultaneously, the final structure should contain somewhere in this range as a maximum of independent latent structures after processing. The amount of information needed might be quite substantial, and the initial structure will be quite complex, but this will have to be rationalised down to few dimensions where the diversity of smell still is maintained.

Recent developments of an electronic nose (Winquist, F. *et al.* 1993) can also contribute strongly towards a more fundamental understanding of how smell/flavour data are organised by human perception.

3.9 Texture

According to Kramer (1973) and Sherman (1969), the term 'texture' has been defined as 'a major division of sensory quality covering all kinaesthetic responses of foods in whatever state they are in'. This is further divided into primary, secondary and tertiary characteristics of the food, and initial perception, mastication and residuals for the human experience. Alternative definitions for texture is given in Texture (Bourne, M.C. 1982).

The number of words utilised to describe sensory characteristics of food texture is enormous, and contains words for everything from particle size to attributes describing the food matrix, like elasticity, gummy, greasy and viscosity, and to mechanical properties like hard, brittle, creamy and powdery. The majority of these descriptors are difficult to attribute to a stringent definition, although lots of work has been put into the effort, especially related to instrumental measurements of these variables. Even when the instrumental definition is exact, the sensory perception of the attribute is not always clearly described. This arrives from the fact that instrumental measures are not always developed in order to reflect perception, but rather to describe a systematic variation in a range of food products. The correlation to perceived entities has sometimes been difficult to establish, like in the case with Instron measurements. For some practical purposes the instrumental measurement have been sufficient, but for a fundamental understanding of perceived texture it has not been adequate. Again underlying limitations in the physiology of the senses do not compare in complexity to the vocabulary utilised to describe sensory perception. Even if this can be explained, it is not obvious how the descriptors fit together in an overall structure. A similar structure to the one of colours is maybe not possible to expect, mainly because the time domain is of utmost importance in texture perception. Still, the relationship between attributes, how they overlap and interact will be easier to understand when this is investigated using multivariate statistics. One good example is given by Yoshikawa *et al.* (1970). A few attempts on meta analysis have been made (Harris, J.M. *et al.* 1972; Risvik, E. 1994) but this must be seen only as initial attempts along this road.

4. PLANNING AN EXPERIMENT

The information collected through the senses reflect two types of information as suggested in the beginning of this chapter. First of all the physiological structure of the senses, that is the channels for information collection. When understood, this normally will be treated as manifest structures like in the case with colours. At the same time the information also reflects the samples in the test, that is the selected space described by the samples, or in other words the latent structures inherent in the experimental design. These two representations cannot be separated in sensory experiments, like in other experiments, and will thus have to be carefully

planned before data are collected. This is one reason why experimental design and interpretation of results are so closely connected in sensory science.

In a sensory profiling experiment the selected attributes limit the amount of information available for analysis. When important attributes are omitted, essential information for the perception of the product will be lost. An important attribute in this context does not necessarily imply an attribute with significant variation. Even attributes with less or small variations between the samples may represent significant information for the perception of the product. The importance of juiciness of apples will not be made available for interpretation if all apples in the experiment have the same juiciness. This is not the same as to say that juiciness is not important.

To design good experiments is not always simple. When manifest variables for observation are all given it is possible to utilise these for experimental design. The most obvious is when physical measures such as length, weight and size can be varied and used in factorial or similar designs. In sensory science, the relationship between the observed variables: the attributes, and the design variables are not always known. In some cases these are also a part of the experimental purpose to be investigated, like in the case of varietal testing of agricultural crops. Here, a part of the test, is to investigate which growth conditions will affect sensory quality; for example of carrots (Martens, M. 1986). The attributes selected are expected to describe variation in carrots, and the experimental design to reflect the variation in crops caused by factors of importance for perceived carrot quality.

This brings the complex decisions to be made, once again, back to the discussions concerning the given and the meant (the relationship between the manifest and the latent).

5. PHILOSOPHY FOR SENSORY SCIENCE

In retrospect, the discussion so far into this chapter can be traced back directly to most of the classical and modern philosophers. This implies a much greater framework for exploration of ideas. To give reference for a few central philosophers a very brief discussion of latent structures in light of philosophy is included.

5.1 Form

Latent phenomena, described in the form of language must be understood as rather fuzzy structures. Words rarely have very specific definitions, and if they do have, they certainly are not used this way in everyday communication. More so, these structures do not have independent definitions without overlap. It is therefore not possible to handle words as if they were orthogonal phenomena. If human communication had to rely on exactness, a simple conversation would hardly be possible.

The Form, Essence or Beauty of an object, as described by Plato (in: The Republic, Symposium and Phaedrus) and Aristotle (in: DeAnima and Metaphysics), can be understood as the physical form of the object, the functionality or better as an abstract form synonymous to the latent structure inherent in the object. It is my allegation that this structure is, through the use of multivariate statistics possible to understand as a form of learned or experienced latent structure, unconsciously triggered as a primary signal or holism when new and more complex experiences are to be characterised. Since our consciousness have limited capacity for simultaneous experiencing, data have to be presented as a compressed structure with reduced dimensionality. This demands efficient processing with strong resemblance to multivariate

analysis of sensory profiles when these are transformed from large complexity down to the simpler latent structures. This is why principal components often trigger resonance structures, when analysed data are being interpreted.

Kvalheim, O.M. (1992) explained the latent variable to be «the «missing link» between Aristotle and Plato in the sense that the latent variable approximates the ideal world by constructs from the real world».

5.2 Phenomenology

Already Kant and Heidegger have strong contributions to a first discussion of a phenomenology of perception. This has been continued by Husserl towards a multidimensional paradigm for perception, as suggested by Idhe (1986). The introduction of the «given» (manifest) and the «meant» (latent) links the object and the observer in phenomenological analysis of perception. This has also been one of the central concerns of Merleau-Ponty (1962), which also brings in ambiguity of perception. In total this can be seen as a development where increasing degrees of complexity have been added to the simplistic models of Plato and Aristotle. Similar or parallel development can be seen from application of Cartesian mathematics towards multivariate statistics, fuzzy algorithms and neural network models, as applied in sensory science.

The philosophers provide paradigms for understanding, the statisticians transform this into practical tools, while sensory science has the unique opportunity to live the interaction of the two realms in experimental settings.

5.3 Poetry

Bacon has suggested a classification of sciences and the intellectual sphere where Poetry was evaluated as a separate branch of science on the same level as Natural Science. Allegorical poetry has always been an advanced form of communication through verbal pictures. The art of creating good poetry, implies the ability to create structures within a strict framework and to communicate, very often a complex idea without being specific. The resemblance between this and the interactive process between statistical presentation and the seeking of resonance from data structures in personal experience is striking. It is not difficult to understand Bacons respect for poetry, which made him suggest this to be on the level with Natural Sciences in his structuring of the intellectual sphere.

PART II

6. HOW DATA STRUCTURES CAN BE EXTRACTED AND INTERPRETED FROM A SET OF DATA UTILISING MULTIVARIATE STATISTICAL TECHNIQUES

6.1 What is a latent structure?

A very simple example first will help to visualise the concept. Imagine a banana (Martens, H. 1985). To understand the spatial information in this banana we need to define a reference system, the space the banana can be found in. If it lies on a table it is possible to use the tabletop as a two dimensional space, with one third dimension going up. Distances along the sides of the table give coordinates for the location of the banana in a two dimensional projection. Depending on how it lies, it will describe different shapes in these two dimensional coordinate system. It can be a straight line (the banana lies with the curvature up), it can be a C-shape (it lies flat on the table), or it can be a sort of circle (the banana hangs in a string from the ceiling. These projections can separately give some information about the banana, but is not complete until all three projections are combined. Since this only is a projection of physical object from 3 to 2 dimensions, it is obvious that information will be lost in the process. This is because the original variables already are orthogonal. When the original space (the reference system) contains variables with a high degree of correlation, a projection from a higher dimensional space down to fewer will not necessarily loose a lot of information. This is what is exemplified in the following discussion.

A constructed example, easy to follow, is chosen to help understanding and to explain theory and concepts in modelling.

Consider car accidents. Most people have a personal view on the most important causes for car accidents. Depending on who you ask, you might get answers like:

- the low standard on vehicles in this country
- because of drinking and driving,
- women drivers,
- young men driving too fast,
- old men driving,
- icy roads,
- because drivers do not respect speed limits and more.

Each of these reflect attitudes and values of the person saying it, and it may also be a potential cause, but not necessarily.

To investigate this further, instead of fighting over who has the right answer, it is possible to perform an analysis on collected data from real accidents. This will generate a data table (matrix) like this:

Table 1
Data from car accidents

	Standard of vehicle	Age of driver	Age of car	Sex	Speed	....
Accident 1	a1	b1	c1	..	...	n1
Accident 2	a2	b2	...			
Accident 3	a3	...				
...	...	...				
...	...					
Accident N	an					nn

When all accidents (samples) are recorded with information on all causes (variables), this makes up a large table or a matrix of information. In a multivariate analysis of this matrix, methods are employed to seek a projection where maximum variation in variables are expressed. This is useful in order to understand how variables are important for a description of causes behind car accidents.

This new and projected space is developed with the aid of an algorithm, where the principles can be exemplified in the following:

First of all a multidimensional space is built from the data matrix. Each variable, that is each registered information (standard of vehicle, age of driver, and so on) is considered to be one dimension. In a three dimensional space this is easy to understand like in Figure 3.

The first accident can be described as a point in this coordinate system with the values a1, b1, c1 on each of the coordinates. When all accidents are introduced into the same space they will make up a cloud of accidents like in the next figure, where all points are described with coordinates in the three dimensions.

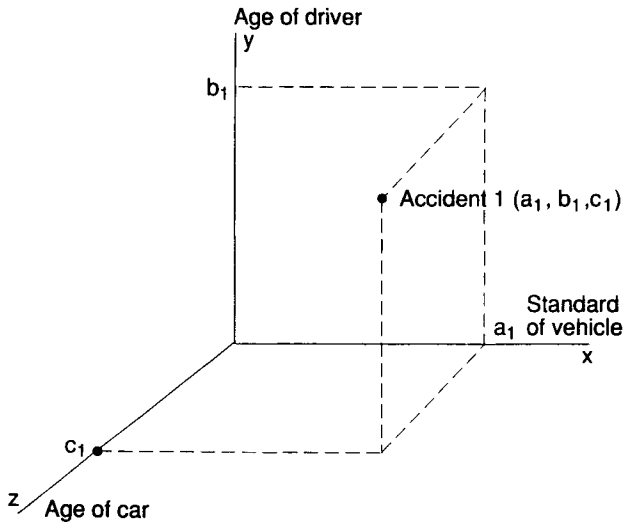


Figure 3. Accident 1 is represented as a point in a 3-dimensional space.

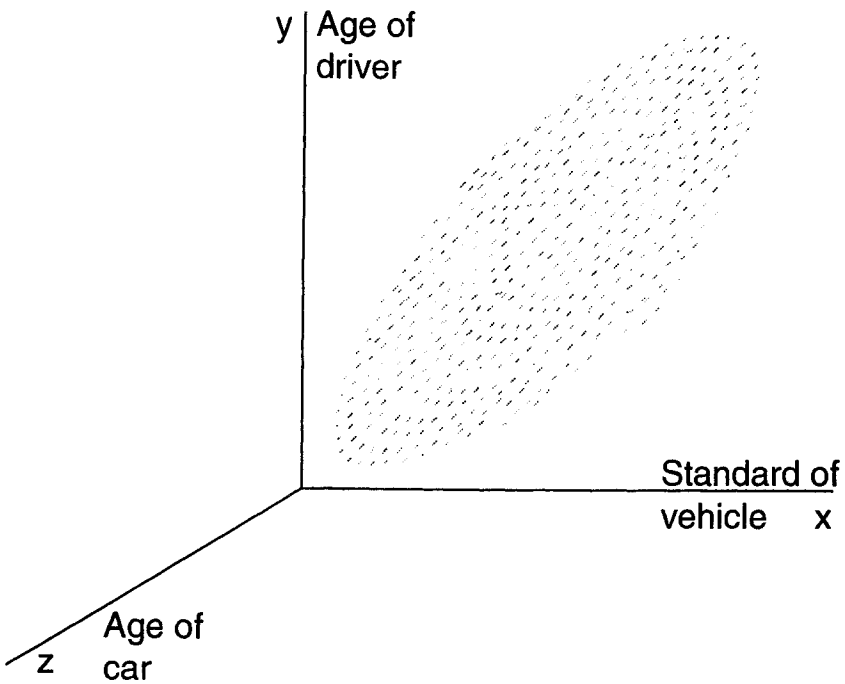


Figure 4. All accidents from Table 1 represented each as a point in a 3-dimensional space (the three first variables only)

This space, for the sake of simplicity, only contains 3 dimensions. In practice there is no reason why this dimensionality cannot be expanded to any number of dimensions. In this case this is the number of variables recorded for each of the car accidents, denominated with "n" in Table 1. For each object (accident) in this n-dimensional space a string of coordinates (a1, b1, c1,.....n1) will be sufficient information to describe the place of this object in the n-dimensional space.

If the variables describing the objects show no systematic variation in this space, the swarm of dots will cover the whole space with an even distribution. In this case the original variables describe unique information in the objects, and there is no relationship between them. Further analysis would, in this case, not be meaningful.

In most cases there will be a relationship, or a covariance between variables. This will be seen as a concentration of points in a specific direction. In a simple and well established relationship between all variables this can be illustrated as in Figure 5, where all points are close to a line going through the space.

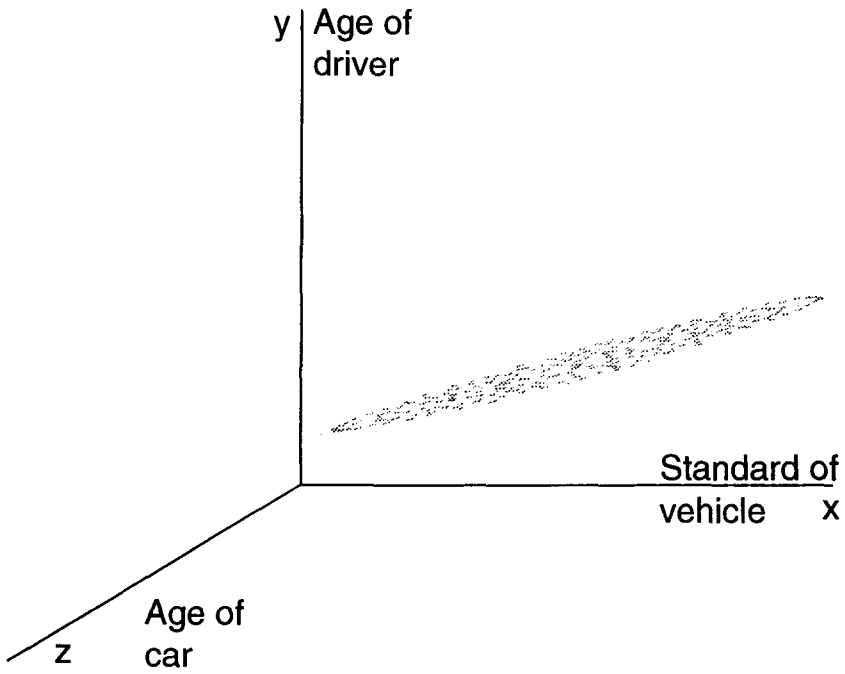


Figure 5. Data for all samples showing a high degree of covariance for all variables

If accidents organise in this plot, they are correlated with an interpretable structure. The way accidents organise along this direction in space will give an indication as to this systematic relationship being an indicator of importance for interpretation, for severity of car accidents.

With few variables in a matrix these plots would very often be sufficient analysis, with the plots, the regression and the correlation coefficients to explain the relationships. With large number of variables two by two plots and correlation coefficients will soon exceed the amount of information possible to hold for interpretation at the same time.

To simplify even further it is possible to utilise the covariance between variables to reduce dimensionality in the original space, or in other words to come up with a projection where important aspects of the original information still is maintained.

STEP 1

A new line is introduced in the original space. The direction of this line is given by the longest distance described in the swarm of data points, that is the best fit to the data. This is illustrated in Figure 6.

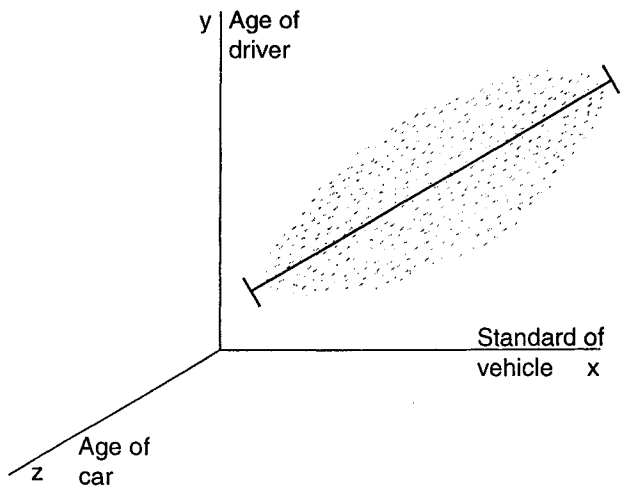


Figure 6. First principal component is introduced, in order to describe the most important direction in space, for interpretation of variance

This new line is characterised by three important aspects illustrated in the next figure.

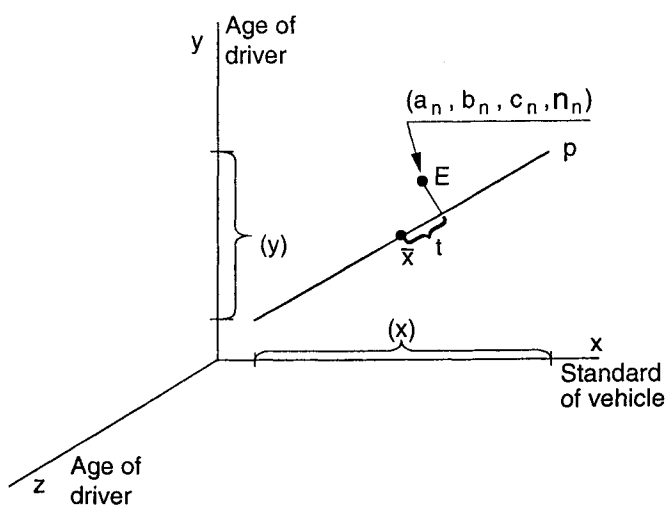


Figure 7. Features of importance for interpretation of a principal component

1) The direction in space relative to the original axes (p) indicates how much is explained of each variable. In this case more of x (standard of vehicle) is explained, while less of y (age of driver). This indicates that in the collected data for «standard of vehicle» show greater importance to explain the variation in this direction than «age of driver». For real data this would maybe not be true unless variables are stretched to a standard length (example: multiplication with xavg/std for all variables (the analysis defaults to analysis of correlations)). In this case variables are comparable, and the relative variation reflect importance of the variable for the found direction in space (there are several concerns connected to this approach, which will be dealt with in later chapters).

   To summarise: this new line in space often called principal component one (pc1), can be characterised by its direction in space relative to the original variables. This tells how much of this original variation is explained by the line. In Principal Component Analysis (PCA) this is called the loadings, and the loadings informs how well variables are explained by the principal component.
2) Each object in the new space is characterised by how far they are from the centre of the space (distance from xavg, the grand mean). These values, or scores as they are called in PCA, explain the relationship between the objects. When objects, in this case car accidents, organise and show systematic variation along a direction in space, this is an indication that the direction is being important for explanation of the way objects organise.

   This statement is also possible to visualise as a move from an external perspective (seeing data from the outside) over to a perspective where the observer is standing in the centre; seeing data from xavg.

   In the case of car accidents, one could assume that the line in Figure 7 describe standard of vehicle as of relative greater importance than age of driver for the described main variation. If scores are organised along this line, so that accidents of less severity can be observed at the lower end and severe accidents at the higher end, it implies that standard of vehicle is of relative greater importance for interpretation of severe accidents than age of driver (since the line is tilted in this direction). If the accidents are organised in a different way along the principal component this will indicate a different relationship.
3) If all objects showed 100% correlation in this 2-dimensional space they would all be exactly on a line between the variables. In the case of a perfect linear relationship, this could be described by the principal component to 100% (non linear relationships will not be discussed in this chapter). For realistic experimental data this will normally not be the case. Each object will also, to be fully characterised, be described by E, the distance away from the principal component (the error in the model). Another way of explaining this is by calling it lack of fit, noise or remaining variance, not accounted for by the model.

   This, left-over information, can be collected for all objects and variables in a new matrix, the error matrix, from the left over information in the calculation of the first principal component.

STEP 2

The error matrix, or unexplained variance from the calculation of the first principal component describes information not accounted for. Since the first pc is calculated to be in the distance of most variation this can be illustrated as in Figure 8.

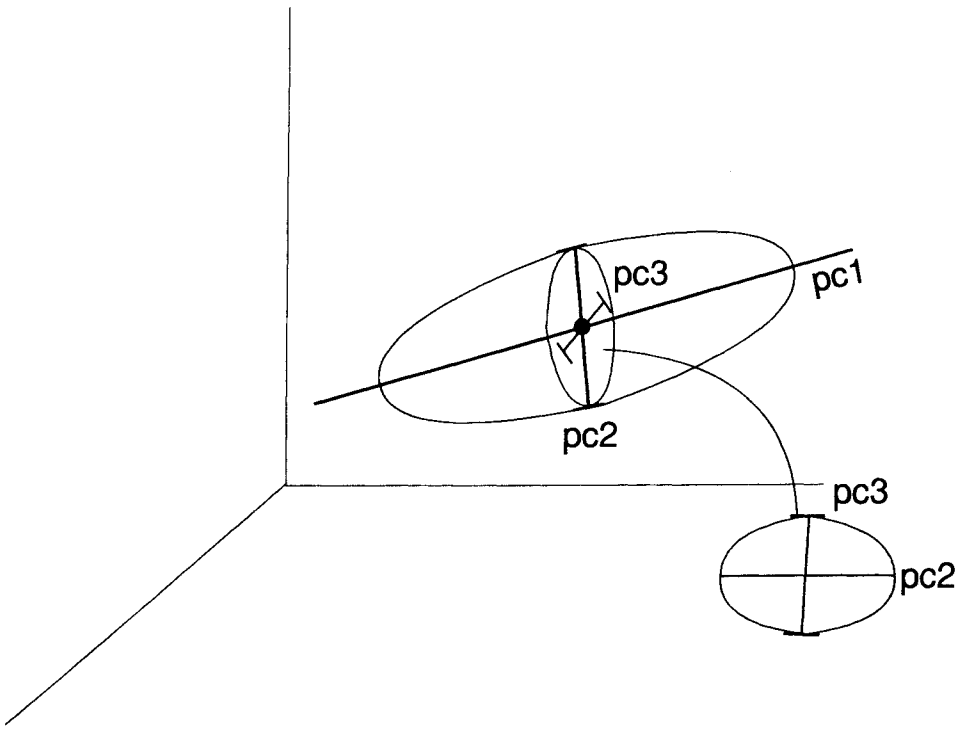


Figure 8. Principal component 1, 2 and 3

The remaining information will often be distributed around pc1 in something like an oval shape. The procedure from the calculation of pc1 can be repeated again and a new dimension will emerge. In most models the criterion for calculating the new dimension is that it will describe as much as possible of the remaining information and at the same time be orthogonal to the previous principal component.

This second pc will be of less importance to the explanation of the variation in the material (importance in this case is different from being important for interpretation of results).

Since it is orthogonal to the first principal component they can be plotted against one another. This information can be presented in two different ways.

1. First of all a loading plot (direction in space with relationship back to the original variables) will give a map of all variables and their relationships.

A score plot will be an illustration of how accidents are distributed in this space, and systematic variation in this space might give an indication of underlying structure or meaning (that is one reason why they are called latent phenomena).

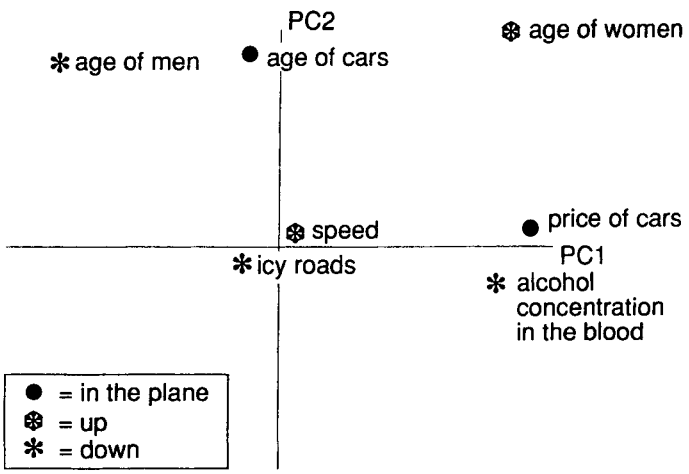


Figure 9. Illustration of possible loading plot for mock car accident data

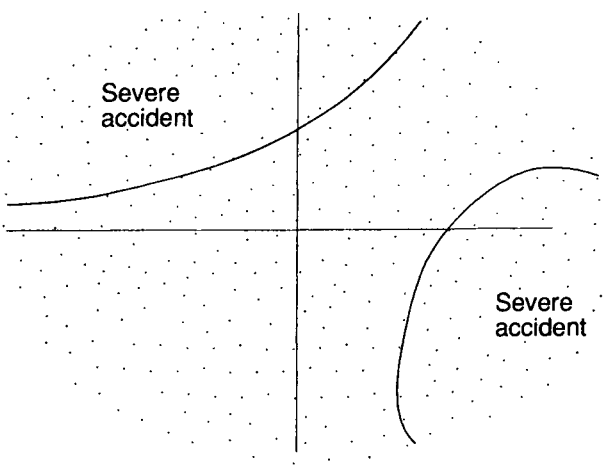


Figure 10.Illustration of possible score plot for mock car accident data

In a biplot these two spaces are superimposed into the same space (not shown here).

The plots are made up to illustrate several difficulties normally met when loading and score plots are to be interpreted. There is no attempt from the author to make the plots resemble what would be the reality in a case study, but rather a more humorist view playing on known prejudices.

7. INTERPRETATION

Interpretation of principal component plots is not always simple. Very often, in literature, only a description is found of how loadings are placed along the principal components. This leaves the interpretation to the reader.

7.1 Principal components

In this case, the first two dimensions are constructed to give an example of a group of two correlated attributes and one other attribute uncorrelated. Correlated means that changes in one variable is parallel to a change in another variable. When these are plotted against each other the objects will fall on a straight line between the two. «Price of cars» and «alcohol concentration in the blood» seem to be correlated in the first two dimensions, while the attribute «age of cars» is not correlated to these. In the score plots each of these factors give rise to a group of severe accidents as indicated in the figure.

Two attributes with no correlation indicate a relationship between the two variables where variation in one, will not show an effect on the other. In this case the price of the car show no relationship to the age of the car, when it comes to accidents. This implies that both old and new cars describe accidents in the same way whether they are cheap or expensive (independent variables).

The two correlated attributes, price and alcohol concentration indicate that expensive cars tend to appear in accidents were alcohol also is important, while cheap cars tend do not. In the plot there is an indication of a separation of these in the third dimension, which also might make the interpretation more difficult.

Two attributes in the plot lie along the diagonals, in-between the principal components. These are «age of men» and «age of women». They both show a high degree of correlation with the attribute «age of cars»; that is: a relationship between the age of the car and the age of the driver in accidents can be seen. At the same time old women (high on age of women) tend to drive more expensive cars, while old men drive inexpensive cars (opposite to high on price of cars). Old women also tend to have a higher alcohol content in the blood, while old men tend to lie on the opposite end of the scale (that implies that young men show the opposite behaviour).

Two attributes are located near to the centre of the plot. They have low loadings and do therefore not play an important role for the interpretation of the relationship between the attributes.

The two first pcs of the analysis always give most of the information available in the data, but there might be several dimensions possible to utilise for interpretation. To establish the optimal number of pcs for interpretation is always difficult. A whole series of techniques called validation techniques are developed for this purpose. Some of these techniques will be introduced in relationship to analysis methods in other chapters of this book.

Let us assume that in this case three pcs were found to be optimal using crossvalidation (Stone, M. 1974). The third dimension, orthogonal on the two first would then create a cube of the loading

plot, where speed and icy roads were to be found on either end of dimension three (indicated in the plot). All the other loadings would also have to move up or down along this axis if they had any correlation with the two attributes. Let us assume that also the old men show some relationships with high alcohol content in the blood, but to a lesser degree than old women. If both (age of men and alcohol content in the blood) were projected down in the structure, it would suggest that there is a grouping of old men, on icy roads with high alcohol content in their blood, important for car accidents. On the other end fast driving older women with expensive cars show another grouping.

So far this has only been a description of the distribution of the loadings along the pcs. To understand whether this gives any meaning it is necessary to compare the loading plot with the score plot.

To recapitulate: The scores describe the relationship between the principal components and the accidents (samples). In this case the distribution of accidents show systematic variation in two areas where severe accidents seem to concentrate, can be identified. These can be related to the loadingplot and explained by attributes pointing in this direction.

Already from the visual inspection of the plots it is possible to see emerging hypotheses in the data. To investigate this further it is necessary to employ other statistical models. One approach could go through the use of Clustes analysis (CA) and or Discriminant Analysis (DA) to look for groupings in objects or variables. For many instances the initial PCA will suggest groups for investigation in complementary and subsequent analyses such as CA or DA. In order to quantify differences and to test for significance, other models like STATIS and Canonical Discriminant Analysis (Schlich, P. 1993) can be of help. Several of these and similar models will be included in later chapters.

8. EXPERIMENTAL DESIGN

Traditional statistical analysis has been developed for descriptive purposes and to support conclusions based on data from experiments. This has very often been factorial designed experiments where few variables have been involved. In standard Analysis of Variance (ANOVA) variables are treated one at a time, and the influence from other variables giving the same effect (inter-correlations) are, in the simplest ANOVA models, ignored.

From Cartesian mathematics to the interpretation of latent structures there is a conceptual jump. Still Cartesian coordinates are the whole basis of multivariate statistics, while interpretations and use of latent structures very often belong to a very different school of thought.

A latent structure is a combination of variables which together make up the main structures in the data in a simpler way. An example can be found in wine profiling (Sivertsen, H.K. and Risvik, E. 1994; Pagés, J. *et al.* 1987; Noble, A.C *et al.* 1987; Heymann, H. and Noble A.C. 1987) where similar structures are found. The attributes of a wine very often aggregate on both sides of the first principal component. In one group the fruity and flowery aromas and on the other side, animal, vegetative and astringent flavours. This, also being the main difference between young and aged wines. It implies that the first pc very often is related, among others, to wine ageing. It is simple to understand how this would be different if all the wines in the experimental design included only young wines, only the one side of the pc being represented in the data.

In the wine example, the structure to be described will focus information from one group of attributes and the distribution of both attributes and samples will appear very different in the loading and the score plot. The relationship between experimental design and the latent structures described

in the analysis is thus obvious. In this case the latent structure observed indirectly in wine can be named ageing. It would be dangerous to suggest a causal relationship, since also many other variables can be confounded or highly correlated with this information. To suggest that principal components describe causal relationships or manifest structures is difficult to say unless relationships is considered together with the information on its experimental design. A few examples will be needed for a sufficient illustration:

** If the intention was to describe underlying structure in car accidents, analysis of all existing car accidents contain all the necessary information to draw causal conclusions. For practical situations this will be a fairly decent sized matrix, even when limiting oneself to one country. Most computers will have problems with this. So to limit the amount of work it is reasonable to draw a sub-set of samples where the main tendencies still are maintained. If this was done with for example all women missing, the results would not hold for this segment of accidents. Similar would be the case for only new cars.*

** In experiments where "the world" is not so easy to describe it is getting increasingly difficult to design sensible experiments. In some cases, like with the difference between organically and traditionally grown vegetables, experimental design can be very complicated (Lieblein, G. 1993) It is not always possible in advance to tell which factor will have the greatest influence in the material, as no response variable in the sensory profile is expected to show a profound "organic" effect. It is very well possible that similar differences in the material also can be caused by other factors such as soil type, weather, latitude, cultivar or pests. In a traditional design this experiment will have to contain variation in all possible factors so that theycan be separated in the analysis. With only 5 factors and 3 levels for each factor this give 243 samples. Only this small experiment, for sensory analysis, is not difficult to perform in an experiment with 12 assessors and 3 sensory replicates.*

** When the purpose of the test describe a limited problem, such as often is the case in product maintenance, the traditional statistical designs may very well be sufficient. To optimise a recipe where one ingredient can be substituted by two or three other ingredients, but the final product is to be as close as possible to the present product on the market, is a typical situation. A systematic variation of the ingredients can be performed according to a factorial design and the results can be analysed in an ANOVA. A discussion of such results is limited to the original design, and it is difficult to interpret causal relationships into the model, as only a very limited number of factors have been chosen for the experiment.*

** In the analysis of a sensory profile it is imperative to notice that the data on each variable have been collected through conceptually very different channels. The colour differences between wines can be simple to score consistently, while bitter taste can be much more difficult (order effects, bitter blindness, fatique, masking effect), and not present in the colour evaluation. In a situation where several attributes are selected to describe highly correlated colour variations in a material, while only one attribute describe bitterness, this will most likely affect the analysis. In a PCA, the first dimension, pc1, will be dominated by the highly correlated colour information, while the bitterness is expressed in later principal components. This is not because bitterness is of less importance in the wine, as already discussed. It is also not because the bitterness is of less importance in describing the variation in the material, but it can be because it is so much easier to evaluate colour, and because the attributes have all been given equal weight, like in a correlation PCA (all variables streched to variance = 1 before the analysis). In this case the structures would better be evaluated in separate analyses, that is colour attributes separate from taste and flavour*

attributes, separate from texture attributes. In its extreme form this opens up for a discussion of whether information collected through separate senses (visual, chemical, auditory kinaesthetic) should be analysed separately more as a rule, rather than the exeption.

** In the analysis of carrot quality, it is easy to understand that the occurrence of a foreign object like tomatoes will cause problems in the analysis. A multivariate statistical model like PCA will in this case concentrate on how the tomato is different from the other samples. Variation within the carrot material will be ignored as most carrots are much more similar to each other than they are similar to a tomato.*

In the example of carrots and tomatoes it is easy to understand why the value of the experiment is reduced unless the odd sample is recognised and removed from the analysis. In reality the odd sample can be difficult to distinguish. In investigations of crop variety among different cultivars, a pest on one sample might cause a similar situation, although not recognised by the experimenter. In this case the analysis will describe the effect of pest and not how cultivars are different.

When all these ifs and buts are taken into consideration: Why is it that even when obviously limited sample sets are chosen, similar structures emerge?

The answer can be one out of many. The structures described by latent phenomena can be very stable, not yet established as manifest structures. It is possible that several confounded effects work together to stabilise certain structures. In a similar way there might be strong indirect correlations (of unknown cause) to causal relationships. And of course the observed similarities might be artifacts. For these reasons the validation of results through specialised techniques or in complementary analyses are of great importance.

Several of the techniques introduced in this book will give complementary views into a set of data, and should be considered as, not in competition, but rather as supplements to each other in data analysis.

One other appropriate question to be asked at this point concerns the interpretability of latent structures. From a deterministic tradition of science we have been trained to seek causal factors or to test pre-set hypotheses. The commonalties or lack of such in similar sensory experiments call for meta-analysis of data, in order to investigate possible manifest relationships between design factors and reoccurring data structures. This is a way of thinking with traditions from humaniora and as such, very often seen as in contrast to the previous. In sensory science the meeting point of humaniora and technology, the opportunity is present for both approaches at the same time. This implies a very exiting research environment with a great potential of new and exiting contributions to contemporary science.

9. GEOMETRICAL REPRESENTATION OF LATENT STRUCTURES

When all considerations concerning experimental design is taken into account, and its effect on the results are handled with care, the actual interpretation still remain. To help in this, real data are used in an example.

9.1 Imposing causal interpretations on latent structures

For the sake of simplicity this example illustrates simple features in multivariate analysis and compares this to analysis of variance. This interpretation goes one step further than before. That is because interpretation on a causal or fundamental level is indicated. In most experiments this level of interpretation is not included, being considered a bit premature based on only one experiment.

Latent structures can be understood as a geometrical representation of data measured indirectly, or a projection of higher dimensional space made up by manifest variables, down into a space with fewer dimensions. A well known example of this is a regular map of an area. The information in the map can contain both the shape (two dimensions), the altitude in the form of contour lines (dimension three) and geological information in the form of colour (dimension four) presented in two dimensional latent structure on a piece of paper. Similar can higher other dimensions be projected down into a representation in fewer dimensions. For sensory data, very often, 2-4 dimensions contain most of the systematic information in a 10-30 dimensional profile.

From sensory science, sensory profiling of raspberries is chosen. The example contains sensory profiles (Risvik, E. 1986; Martens, M. *et al.* 1994) from 12 raspberries harvested at 3 different times for 4 different cultivars. The profiles contain 12 attributes profiled by a panel of 12 in 3 sensory replicates.

Panel data were first averaged over assessors and replicates. In an analysis of variance all attributes came out with significant differences on a 5% level for cultivar, while the harvesting time only showed significant differences for half the attributes. It was also observed interactions between cultivar and harvesting time for two attributes (sweet and viscous). In the PCA, four dimensions were described as possible to interpret after validation with leverage corrections (Figure 11).

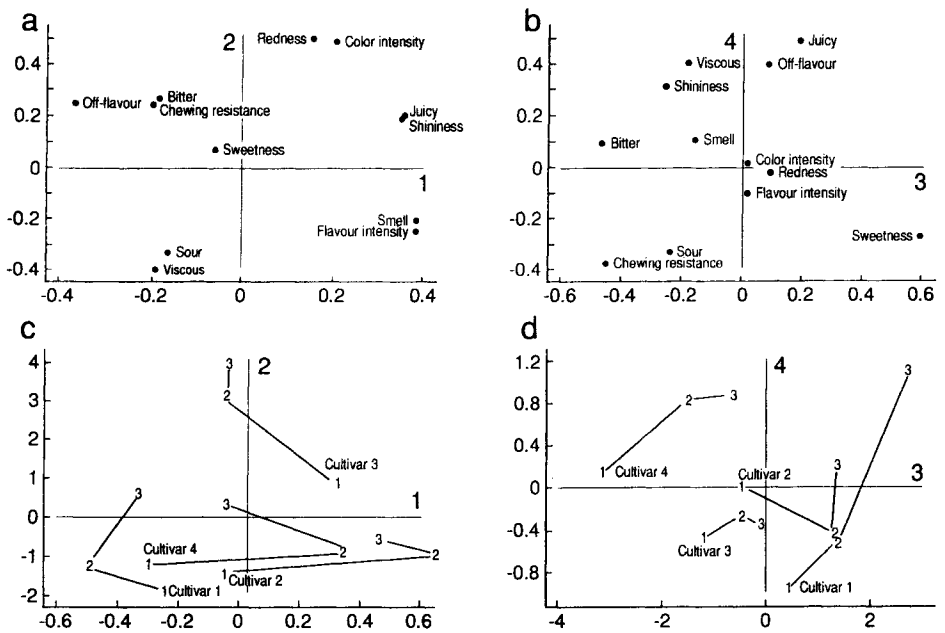


Figure 11. Loadings and scores for principal components 1 - 4 for 4 cultivars of raspberry, harvested on three occasions, and profiled on 12 attributes. Figure 11a&c = pc1&2. Figure 11b&d = pc3&4

In the interpretation of the dimensions, the loading and score plots give simple explanations. In the following the principal components will be discussed one by one and interpretations imposed on to the data structures.

Dimension one, pc1, distinguishes between a group of attributes on the right side: redness, intensity of colour, juicy, shiny, sour, intensity of smell and flavour, on the other side are off-flavour, bitterness and chewy aspects. Cultivars 2 and 4 show large differences between first and second harvest, with a decrease in off-flavour, bitter and chewy and then an increase for the third harvest, while cultivar 1 and 3 show more of a mirrored pattern. Cultivar 1 have more off-flavour, bitterness and chewiness than the other cultivars. It is possible to suspect that the changes along pc1 for cultivar 1 and 4 are examples of how berries within the same field can be at very different development stages when picked at three different times and that the fields also can be different. Maybe even more important for sensory attributes than variety are growth conditions (temperature and water). For berries, ripened with intervals of three weeks, as in this case, the development of juiciness, sourness and also off-flavour can be at very different levels each time, an expression of how the weather has been different in the fields in the mean time.

The second pc describe colour variation and viscosity, that is colour and texture changes with high degree of correlation between them. Sourness correlates well with this direction. When colour intensity increases viscosity decreases. Knowing that more mature raspberries have less viscous texture and more intense and red colour this is explainable, provided the samples organise from less mature to more mature in the same direction as the attribute changes. In the score plot this can be seen to be true. But in addition it is possible to see that the changes are not at all comparable for the 4 cultivars. The order of the cultivars in direction of increasing maturity is the same, but the level and the magnitude (if these can be interpreted) are different. All cultivars increase in redness and colour intensity during ripening, while cultivar 3 already at the beginning at harvest one has more redness and colour intensity and less viscous texture than the other three cultivars. Cultivar 2 and 4 show little change at all. There are at least two interpretations of the results: 1) That the cultivars have different ripening curves, that is a difference in how they ripen. 2) The four cultivars represented here ripen at different times, so that cultivar 3 already at the first harvesting is more ripe than the three others at the third harvesting time. To distinguish between the two interpretations is not possible from the existing design, although an interpretation of pc1 and 2 together would favour the last interpretation. The three harvesting times for each cultivar are linked with a line to show this.

An interpretation of pc3 shows that while 3 cultivars increase in sweetness and redness during maturation, cultivar no 2 show slightly different behaviour. After the second harvest the cultivar decrease in sweetness, which would be expected to cause an interaction term in the ANOVA. This is also seen for the term sweetness.

A very similar explanation can be given for pc4 on sweetness, where sample 2 and 3 have very different orders of sweetness for the 3 harvesting times. For the terms viscous there may not be a difference in the orders of the samples on pc1, but in pc4 where viscous also plays a role the patterns for sample 1 and 2 are different. This may be the cause for the interaction term in the ANOVA.

The interpretation of pc3 and 4 is not so obvious, as for pc1 and 2 with the growth conditions and maturation curves. It seems more linked to differences in patterns for the actual cultivars. Cultivar 1 is more on the sweet and juicy side, while 4 is less sweet, more viscous, but still juicy. This can then be interpreted as the cultivar differences.

As can be seen by the interpretation of pc1 to 4 the information in the loading and score plots can be related to underlying information in very simple ways. It is not the intent with this discussion to indicate causal or fundamental relationships, as this is not yet proven. Still the data in this projection of the data, opens up for such hypotheses to be made. In these data there are indications that it could be possible to recognise effects from growth conditions (pc1), from ripening (pc2) and from cultivars (pc3 and 4). Far fetched interpretation from such a small material, but very interesting if it maintains validity in reproduced experiments.

In order to strengthen the interpretation, the data should be transferred into other models in order to get other perspectives into the interpretation. This is not performed here.

10. CONCLUSIONS

This chapter has discussed sensory profiling with a wide perspective. The intention has been to give a platform for understanding of multivariate statistics in sensory science. In order to do so it is necessary to incorporate a discussion of the more fundamental issues related to the use of multivariate statistics and to the interpretation of results.

Eating a food or drinking a glass of wine is the meeting point between the subject and the object, and as such it touches upon some of the most fundamental discussions of human philosophy, that of life and death, identity, and that of the good versus the bad. Analysis of sensory data will eventually touch upon these or related aspects and the sensory analyst is therefore best prepared when these discussions always are kept alive, as a part of a continuous education in the field.

11. REFERENCES

Amerine, M.A., Pangborn, R.M., Roessler, E.B. 1965. Principles of Sensory Evaluation of Food. Eds G.F. Stewart et al. Academic Press Inc., New York and London, 108-114.

Ashirst, P.R. 1991. Food flavourings, Blackie and Son Ltd, Glasgow and London, 126-127.

Bacon, F. 1605. Advancement of Learning - divine and human. In: Bacon selections. Charles Scribner's Sons, London.

Bourne, M.C. 1982. Food Texture and Viscosity: Concepts and Measurements. Accademic Press, New York and London.

Brantzaeg, M.B. 1958. Taste sensitivity of P.T.C. in 60 Norwegian families with 176 children. Acta Genet. et Statis. Med. 8, 115-128.

Cattell, R.B. 1952. Factor analysis: An introduction and manual for the psychologist and social scientists. Eds Harpers and Brothers, New York.

Dorries, K.M., Schmidt, H.J., Beauchamp, G.K. and Wysocki, C.J. 1989. Developmental Psychobiol. 22, 5, 423-435.

Gleick, J. 1987. Chaos: making a new science. Viking Penguin Inc., New York.

Harman, H.H. 1967. Modern Factor Analysis. The University of Chicago Press, Chicago and London, 2nd ed.

Harris, J.M., Rhodes, D.N. and Chrystall, B.B. 1972. J. Text. Stud., 3, 101.

Henning, H. 1916. Der Geruch, 1st ed, 533 pp, Verlag Barth, Leipzig.

Heymann, H. and Noble A.C. 1987. Descriptive analysis of commercial cabernet sauvignon wines from California. Am. J. Enol. Vitic., 38, 1, 41-44.

Horst, P. 1965. Factor Analysis of Data Matrices, Eds Holt, Rinehart and Winston, New York.

Idhe, I. 1986. Consequences of phenomenology, State Univ of NY Press, 50.

Kendall, M.G. and Buckland, W.R. 1957. A dictionary of statistical terms. Edinburgh: Oliver and Boyd, pp 493.

Kobayashi, S. 1981. The aim and method of the color image scale. Color, 6, 2, 93-107.

Kramer, A. 1973. Food texture - definitions, measurement and relation to other food quality attributes. In: Texture measurements of foods. Eds A. Kramer and A. Szczesniak. D. Reidel publ. Company, Boston, USA.

Kvalheim, O.M. 1992. The latent variable. Chemometrics and Intelligent Laboratory Systems, Elsevier Sci. Publ (publishers), Amsterdam, Editorial, 14, 1-3.

Lieblein, G. 1993. Quality and yield of carrots: Effect of composted manure and mineral fertilizer. Dr.scientarium thesis, Agricultural University of Norway, Ås.

Longnecker, M.P., Berlin, J.A., Orza, M.J. and Charlmers, T.C. 1988. A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA, 260:5, 652-656.

Mann, C. 1990. Meta-analysis in the breech. Science, 249, 476-480.

Martens, H. 1985. Multivariate calibration. Quantitative interpretation of non-selective chemical data. Dr.techn. thesis, Technology University of Norway, Trondheim.

Martens, M., Risvik, E., Martens, H. 1994. Matching sensory and instrumental analyses. In: Understanding Natural Flavours, Eds Piggott, J.R. and Paterson, A., Blackie Academic & Professionals, London, 60-71.

Martens, M. 1986. Determining sensory quality of vegetables, A multivariate study. Dr.agric. thesis. Agricultural University of Norway, Ås.

Martens, M. and Harries, J. 1983. A bibliography of multivariate statistical methods in food science and technology. In: Food Research and Data Analysis, Eds H., Martens and H. Russwurm jr, Appl. Sci. Publ., London, 493-518.

Merleau-Ponty, M. 1962. Phenomenology of perception. Routledge & Kegan Paul Ltd., London.

Miller, G.A. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. In: The Psychological Review, 63, 2, 81-97.

Munsell Book of Color, Standard ed., Munsell Color Company, Baltimore, Md., 1929.

Nickerson, J. 1940. The Munsell Book of Color. J. Opt. Soc. Am., 30:12, 575

Noble, A.C., Arnold, R.A., Buechsenstein, J., Leach, E.J., Schmidt, J.O. and Stern, P.M. 1987. Modification of a Standardized System of Wine Aroma Terminology. Am. J. Enol. Vitic., 38, 2, 143-146.

Pagés, J., Asselin, C., Morlat, R. and Robichet, J. 1987. L'analyse factorielle multiple dans le traitement des données sensorielles. Sciences des Aliments, 7, 549-571.

Peynaud, E. 1987. The taste of wine. In: The Art and Science of Wine Appreciation. Eds. N. Good, Macdonald & Co Ltd., London & Sydney.

Rabourdin, J.R. 1991. Vocabulaire international de la dégustation. Jean R. Rabourdin. - 2me éd. 920225 [Orléans 1991].

Rapp, A. 1988. Wine aroma substances from gas chromatographic analysis. In: Wine Analysis, Modern Methods of Plant Analysis, New Series, Vol 6, Eds Linskens, H.S. and Jackson, J.F., Springer Verlag, Berlin, 29-66.

Risvik, E. 1986. Sensory analysis, Tector AB, Höganäs, Sweden.

Risvik, E. 1994. Sensory Properties and Preferences. Meat Science, 36, 67-77.

Sherman, P. 1969. A texture profile of foodstuffs based on well-defined rhelological properties. J. Food Sci., 34, 458.

Sivertsen, H.K. and Risvik, E. 1994. A study of sample and assessor variation - A multivariate study of wine profiles. Journal of Sensory Studies 7, 293-312.

Stone, M. 1974. Cross-validatory choice and assessment of statistical prediction. J. Roy. Stat. Soc., B, 111-133.

Thomson, D.M.H. 1988. Food Acceptability, Elsevier Applied Science, London and New York.

Trincker, D. 1966. Aufnahme Speicherung und Verarbeitung von Information durch den Menschen, Veröffentlichungen Der Schleswig-Holsteinischen Universitätsgesellschaft, NeueFolge, Nr 44, Verlag Ferdinand Hirt, Kiel.

Wada, Y. 1953. Olfaction. Eds. Takagi-Kido. In: Jikken Shinrigaku Teiyo (Handbook of experimental psychology), Iwanami Shoten, Tokyo, 3, 143-174.

Winquist, F., Hörnsten, E.G., Sundgren, H. and Lundström, I. 1993. Performance of an electronic nose for quality of ground meat. Meas. Sci. Techn. 4, 1493-1500.

Wold, H. 1966. Estimation of principal components and related models by iterative least squears. In: Multivariate Analysis, Eds P.R. Krishnaiah, Academic Press, New York, 391-420.

Wysocki, C.J., Dorries, K.M. and Beauchamp, G.K. 1989. Proc. Natl. Acad. Sci. Genetics, USA, 86, 7976-7978.

Wysocki, C.J. and Beauchamp, G.K. 1984. Proc. Natl. Acad. Sci. USA, 81, 4899-4902.

Yoshida, M. 1964. Studies in psychometric classification of odors (4), Japanese Psychologial Research, 6, 3, 115-124.

Yoshikawa, S. et al. 1970. Collection and classification of words for description of food texture. J. Texture Studies, 1, 437-463.

Zwaardemaker, H. 1895. Die Physiologie des Geruchs 324 pp, Engelmann, Leipzig.

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

EXPERIMENTAL DESIGN

Edward Anthony Hunter

Biomathematics & Statistics Scotland, The University of Edinburgh, Scotland,
United Kingdom.

1. INTRODUCTION

Good experimental design is extremely important in all areas of science; especially where treatment effects are small relative to uncontrolled variation, as in sensory studies. It is generally accepted that a well designed study that is analysed using simple methods will yield more information than a hastily designed study analysed using sophisticated methods. Careful design of sensory experiments is essential in order to derive the maximum amount of useful information from the work of the sensory assessors and the technicians who run the experiments.

Organising a program of sensory research, defining the objectives of each individual experiment and running the experiments efficiently requires an appreciation of many disciplines in addition to the statistical design and analysis of experiments. The book by Stone and Sidal (1993) will be found to offer much useful advice.

1.1 Historical Perspective

Experiments are carried out to test hypotheses and also to estimate parameters in statistical or mathematical models. The sole purpose of experimental work is to provide information. Statistical design of experiments identifies sources of variation, (both random and systematic) and then takes them into account in designing the experiment and in the subsequent analysis. Thus the resources expended in carrying out a well designed experiment result in the maximum amount of information. Statistical design of experiments is widely used in applied biology, medical and clinical science.

The foundation of modern experimental design and analysis is due to the work of R A Fisher at Rothamsted Experimental Station, Harpenden, England in the 1920's. Fisher devised efficient methods of designing and analysing agricultural experiments which have been successfully applied throughout the world. For an account of this work see primarily Cochran and Cox (1957) but also John and Quenouille (1977) and Mead (1988).

Fisher's methods have been adapted for use in the chemical and other continuous process industries and in the engineering industries by G E P Box and colleagues at Madison, Wisconsin, USA. These methods have been used throughout the English speaking world, and many are directly applicable to the food processing industry.

From a very different starting point and with knowledge of Fisher's original work, the Japanese engineer Taguchi has developed a philosophy of quality improvement in

manufacturing industry which incorporates a substantial component of experimental design and analysis, albeit in an industrial context.

Bradford Hill introduced statistical ideas of designing and analysing experiments to clinical studies of new pharmaceutical products and other therapies. The randomised, double-blind trial is the accepted way of testing new treatments.

The use of statistical experiments in perceptual psychology has a long history, Robson (1973). The issues that arise in this area of application also arise in sensory evaluation.

In each area of application, the same statistical principles are applied, but are adapted to meet the special needs of the experimental work and data. It is important that sensory scientists realise that their experiments can be seen as being part of a much wider scenario with a large literature, some of which is relevant to their requirements.

Finally, other authors have considered the problems of experimental design in sensory studies. MacFie (1986) provides check-lists and covers many practical points in his Chapter in a book edited by Piggott. O'Mahony (1985) gives a gentle introduction to the use of experimental design in sensory evaluation and covers some of the points made in this Chapter in more detail. Many practical matters concerned with setting up a sensory laboratory and running sensory experiments are covered in Stone and Sidal (1993).

1.2 Blind Testing

When people are used as experimental subjects, it is important that they are unaware what treatment combination (=sample) they are assessing. In clinical trials neither the experimental subject nor their physician know which treatment is received; this is known as double-blinding.

For example, suppose a supplier of chocolate to a supermarket chain (who then sells it under the supermarket's own label) is required to demonstrate that the product is similar to branded products. This can be achieved, by setting up an experiment to compare the sensory properties of the supplier's chocolate with other brands. In such an experiment all packaging and even the brand names are removed from the surface of the chocolate. Care must be taken to present sub-samples of the same dimensions from each sample. This ensures that assessors are influenced only by the sensory properties of a brand, not by its image or other extraneous factors.

The order of presentation of the sensory tasks within a session is known to systematically affect the results. It is desirable that the sensory technician works from a previously prepared plan giving the order of tasks for each assessor.

It is desirable that sensory assessors are screened from each other during testing so that they are not aware of the reactions of their colleagues to the samples being assessed. Opinions differ about how much feedback of sensory performance should be given to assessors or whether assessors should receive information on the results of experiments. At one extreme, some laboratories routinely provide diagrams for each assessor which show performance relative to other assessors. These are often the laboratories which hold training sessions round a table in which samples are examined and scores for attributes agreed. Other laboratories are more sceptical about the value of trying to align assessors. There is, however, no dispute about the value of external standards and the need to ensure that assessors produce consistent ratings for repeated samples. It is important that assessors are valued for the contribution that their special skills make to the work of the sensory laboratory.

1.3 Randomisation

After systematic sources of variation have been identified and designs devised which take them into account (see later), randomisation is the next step in the process of producing the order of samples for each assessor. This process ensures that the true differences between samples are estimated free of biases and also allows variability to be estimated. If there is laxity about randomisation then it is likely that undesirable systematic effects will bias the sample effects and invalidate the estimates of variation.

One additional hazard of sensory analysis is the ability of assessors to remember previous experiments, to perceive pattern in the samples that they were allocated and then to anticipate this pattern in future experiments. An example will help to clarify the hazards.

Suppose that the sensory laboratory (unwisely) uses the same allocation of samples to assessors each time it compares, say, 4 samples. Suppose also the laboratory performs many experiments on the effect of sweeteners on the sensory properties of products. If the samples are allocated to numbers in increasing order of inclusion of sweetener then, unless there is randomisation of samples to sample symbols or assessors to assessor symbols, assessors will receive the same pattern of sweetness of samples in each experiment. Assessors will quickly learn to anticipate the sweetness of the sample they are assessing and this will prejudice the integrity of their ratings of sweetness and all other sensory attributes.

The only fully satisfactory method of randomisation is to independently randomise the assessors to assessor symbols and the samples to sample symbols for each experiment. This can be conveniently done by a computer program with a different seed for each randomisation. Using this process will ensure that a different order of testing of samples for each assessor is produced every time a design is generated.

1.4 Factors which Influence Success

The factors which influence the success of sensory work are:

1. the clear statement of objectives,
2. the careful design of the treatment structure to satisfy these objectives,
3. the allocation of treatment combinations (=samples) to assessors,
4. the careful execution of the experiment so that no systematic or unnecessary error variation is introduced,
5. careful analysis of data,
6. perceptive interpretation.

By far the most important factor is the clear definition of objectives. The next most important factor is the design of the allocation of treatments to assessors and the design of the treatment structure (dealt with in this Chapter). Good design allows informative univariate analysis of the data, see this Chapter. Advanced multivariate methods, which can do much to summarise voluminous data, also require the experiment to be properly designed despite the misconceptions of some sensory scientists.

1.5 Power of the Experiment

In planning experiments, it is always wise to consider whether the proposed experiment is capable of detecting the differences of interest. A surprising number of experiments fail this test and so are a misuse of resources. In clinical trials, the codes of practice require power calculations to be done. These are illustrated later in section 7.

2. TYPES OF SENSORY EXPERIMENTS

Sensory experiments can be divided into two kinds: difference experiments and profile experiments. In difference experiments, in which the "odd" sample or samples are identified by the assessor, overall differences between samples are assessed. In experiments where the quantitative difference between two samples is assessed, it is possible to ask assessors to rate the difference in aroma, flavour or texture etc. This can be taken one step further and assessors can be asked to rate differences in a particular attribute, such as lemon flavour etc. Where the questions are general, it can be difficult to determine the precise nature of the differences. Reliance must then be placed on assessors notes and possibly a panel discussion. In contrast, in profile experiments the samples are rated for a number of sensory characteristics so that those which define differences between samples are identified.

Difference experiments are extremely useful when a new product is being evaluated for the first time. For example, a laboratory which usually evaluates cheese and other dairy products might well start with difference tests when it begins evaluating a new product, such as fruit cordials. After experience has been accumulated, more detailed information can be gained from profile experiments.

Difference tests can be subdivided into two classes:

1. triangular and other similar tests,
2. quantitative difference tests.

Triangular tests, one of the more common and simpler paired comparison tests, make only light demands on the sensory assessors but have limitations and must be carried out with attention to detail. Quantitative difference tests make greater demands on the assessors but less than profile experiments. When many samples are being compared, they require large quantities of sample and extensive preparation of the sub-samples for the assessors.

In sensory profile experiments, the assessors rate samples for many attributes. This vocabulary can be fixed for all assessors as with almost all profile experiments, or can be personal to each assessor as in free choice profile experiments. In both cases the vocabulary should encompass the differences between the samples. For products which are frequently profiled, a vocabulary will exist and be continuously modified to reflect changes in the products over time and increasing knowledge of the sensory properties.

3. TRIANGULAR AND OTHER SIMILAR TESTS

The duo, duo-trio, triangular, polygonal and polyhedral tests are all variants of tests for comparing two samples.

The best known of these tests is the triangular test in which three sub-samples (two from one sample and one from another) are presented to each assessor who is asked to pick out the odd sub-sample. The nature of the differences between the samples are not defined, only whether or not assessors can perceive a difference.

When viewed from a psychophysical standpoint the triangular test is (surprisingly) subtle and experimenters should refer to the psychological literature for guidance on asking the question in an appropriate way (see for example O'Mahony, 1985).

It is important that the assessors do not receive clues to the odd sub-sample from the sub-sample numbers or mode of presentation. For example, if the sub-samples from one sample are put on plates of a distinctive size or colour, or if the sample is cut into sub-samples of a distinctive size or shape, then the assessors could receive clues from the presentation of the sub-samples.

With a Triangular test there are six possible ways in which the sub-samples can be presented. Suppose the samples are A and B. Then the possible orders of presentation are AAB, ABA, BAA, BBA, BAB and ABB.

Two scenarios will be considered:

1. each assessor performs one test,
2. each assessor repeats the test several times.

The data from each test consist of either a 0 - the assessor identifies the wrong sub-sample as being "odd" or 1 - the assessor correctly identifies the correct sub-sample as being "odd". The data are thus binary data. In the usual form of the test the probability of identifying the "odd" sub-sample entirely due to chance is $p=1/3$. This is called the null hypothesis.

The results from these trials are analysed by using the binomial distribution to calculate the probability of getting such a result or a more extreme result due to chance ie on the assumption that the null hypothesis is true (Type 1 error). In (too) many triangular test experiments a statistically non-significant result is accepted as confirmation that there is no sensory effect of sample. However, small experiments are insensitive to large differences in the value of p. It is instructive to calculate Type 2 errors (Schlich,1993a) by calculating the probability of getting the experimental result or a less extreme result with p set to 0.5 or 0.6.

Simple triangular tests in which each assessor carries out the test only once are only sensitive when large numbers of assessors are available; when (for example) institute staff, students or shoppers in supermarkets are used.

Consider this example. An experimenter wishes to test whether there are sensory differences in the milk produced by two methods of heat treatment of raw milk. Milk is taken from the institute bulk tank (which is continuously stirred) and divided into two parts. At random, one part receives each heat treatment (A and B).

3.1 One Observation per Assessor

Seventy two people working at the institute are asked to carry out a triangular test. All are familiar with triangular testing having previously participated in this form of sensory test. The way in which the sub-samples are presented is randomised independently for each of the assessors. Thirty participants correctly identified the "odd" sample ie p=0.417, compared to 24 that would have been expected to identify it under the null hypothesis of p=1/3. The probability of getting this result or one more extreme ie 30-72 correct results under the null hypothesis can be got from the binomial distribution. This distribution is tabulated in Stone and Sidal (1993) and also in Gacula and Singh (1984). It is also available in many computer programs. The probability is 0.086 which is appreciably more than the usual 0.05 criterion for statistical significance. It is concluded that the sensory differences between treatments A and B are not large enough to be detected by the experiment ie there is no statistical evidence that p is greater than 1/3.

It is instructive to consider the power of the experiment. Suppose the true value of p=0.5 - what is the probability of getting 30 correct or less? Using the binomial calculations this is found to be = 0.097 whereas for 0.6 it is 0.001. From these calculations, it can be seen that the experiment was of sufficient size to detect modest differences in the level of p from 1/3. This test can reasonably be assumed to have tested the consumers ability to differentiate between the treatments. Using the generalized linear model with binomial variation (Collett, 1991), it is possible to explore the effects of the different presentations of sub-samples.

3.2 Several Observations per Assessor

If a trained panel is being used, which seldom numbers more than 15 assessors, the experiment will intrinsically have a poor ability to distinguish small differences between samples. In order to increase the power of the experiment, there is merit in repeating the test several times for each assessor. Six replicates of the test or multiples of six are particularly convenient. This may now cause problems since there are two levels of variation within the system: between assessors and within assessors. If there are no real differences between assessors in their ability to differentiate between treatments, then the assessor component of variance is zero. Satisfactory methods of handling this kind of data are being developed but have yet to be made known to the sensory community.

In fact these results were derived from the sensory panel at the institute. The 12 assessors performed 6 tests each using a design based on Latin Squares. First a Latin Square of order n is defined:

A Latin Square of order n is an arrangement of n symbols in n rows and n columns such that each symbol appears once in each row and once in each column.

Table 1.
A Latin Square of order 6 is given below:

		Column					
		I	II	III	IV	V	VI
Row	A	a	b	c	d	e	f
	B	b	c	f	a	d	e
	C	c	f	b	e	a	d
	D	d	e	a	b	f	c
	E	e	a	d	f	c	b
	F	f	d	e	c	b	a

It can be seen that the symbols "a", "b", "c", "d", "e" and "f" appear once in each row and in each column. The properties of Latin Squares have been extensively studied by mathematicians and special kinds of Latin Squares are frequently used to produce experimental designs with desirable statistical properties.

Two order 6 Latin Squares were used to determine the particular randomisation for each assessor. Rows were regarded as assessors and columns as the order of testing. Symbol "a" corresponded to the set of test sub-samples AAB, "b" to ABA, "c" to BAA, "d" to BBA, "e" to BAB and "f" to ABB. Generalised linear modelling of the binomial response data did not reveal any statistically significant order of presentation or form of test effects. The results for each assessor are:

Table 2.

Assessor	Correct
1	1
2	4
3	3
4	3
5	1
6	2
7	5
8	2
9	3
10	3
11	0
12	3

No assessors detected the "odd" sub-samples correctly 6 times and only one assessor detected them correctly 5 times.

A statistical test is required to evaluate whether there are differences in assessors ability to correctly identify the "odd" sub-sample. Multilevel models with binomial variation are not yet fully developed so a simpler, and arguably less statistically efficient, technique based on randomisation is used.

As a test statistic the variance of the variate of number correct for each assessor is computed. The formula is:

$$\text{Variance} = \frac{\sum_{i=1}^{12} (f_i - f_.)^2}{11}$$

where f_i is the number of correct answers for assessor i.

For the data the value of the variance is 1.909. By randomising the data 100 times and recalculating the variance, a reference distribution is obtained.

Table 3.

Variance	Frequency	Cumulative Frequency
<1.0	16	16
1.0 - 1.2	16	32
1.2 - 1.4	11	43
1.4 - 1.6	17	60
1.6 - 1.8	8	68
1.8 - 2.0	12	80
>2.0	20	100

The test statistic fits into this distribution at the 80 percentile. It is concluded that there is little difference between assessors in their ability to distinguish between the two treatments.

Finally, it should be remembered that the triangular test is not necessarily the most appropriate test, for example the duo-trio test may be more appropriate.

4. QUANTITATIVE DIFFERENCE TESTING

In *Quantitative Difference Tests* two or usually more samples are compared in the same experiment. Difference between each pair of samples is assessed directly. This can be done using an ordered scale with for instance 5, 7 or 9 points or by using an undifferentiated line scale and asking assessors to mark a line at the appropriate point (see also section 5.5 of this Chapter). Schiffman, Reynolds and Young (1981) provide more details. The analysis aims to estimate the magnitude of differences between samples in the underlying sensory dimensions.

The advantage of the quantitative difference experiment over a set of experiments using the triangular test for each pair of differences is that the sizes of the differences between samples

are quantified. The advantage over sensory profiling is that a vocabulary does not need to be developed. The technique is specially useful in the early stages of working with a product when expertise in its sensory properties is still being rapidly accumulated.

A disadvantage of quantitative difference testing, compared to sensory profiling, is that larger quantities of sample are required and that sample preparation is a longer and more exacting task. Each assessor is required to assess each pair of samples, so for 6 samples there are 15 pairs whereas for 8 samples there are 28 pairs and for 10 samples 45 pairs. Thus, for each assessor, 5 sub-samples of each sample are required when there are 6 samples, 7 sub-samples with 8 samples and 9 sub-samples with 10 samples. The underlying concept of the test does not easily permit these levels of sub-sampling to be broken. This technique is therefore constrained to experiments in which modest numbers of samples are being compared. There is merit in replicating the test but most sensory scientists argue that replication requires too many resources. Another disadvantage of the test is that it is not easy to interpret the sensory dimensions.

The number of pairs of samples often exceeds the number that can be readily tested in one session by an assessor. The pairs of samples then have to be broken into subsets that can be tested in a number of sessions. Furthermore, the order of testing within a session requires to be determined. Within a pair the order of presentation also requires to be determined.

In a well organised laboratory full information will be recorded ie assessor, day of testing, session within day, order within session, sub-samples being compared and presentation order within the pair as well as the magnitude of the difference. For a sensory laboratory with a large throughput, computerised data collection is cost effective but pencil and paper methods are perfectly adequate even though they take a great deal of time and effort to manage effectively.

The usual method of analysing this kind of data is by Multidimensional Scaling Methods (MDS) which are dealt with in Chapter 4.1.

4.1 Example

Suppose that the aromas of 9 samples of cheese (A-I) are being compared by a panel of 12 assessors. There are 36 different pairs of samples. This is too many assessments to make in one session, so the experiment is run over 4 sessions in which each assessor evaluates 9 pairs of sub-samples. One of the first questions to be asked is how the pairs of sub-samples for each assessor in each session are to be chosen. The most convenient solution (at least for the sensory technician) is for all assessors in each session to evaluate the same pairs of sub-samples. However, there are many potential hazards to this approach, even when the 36 pairs are allocated to sessions at random. It is possible, perhaps even inevitable, that assessors will experience a learning curve and that pairs of samples assessed in later sessions will be assessed more stringently than those assessed earlier. A more cautious approach is for the 36 samples to be divided into 4 sessions of 9 pairs using a different random process for each assessor. It is known that there are order effects within sessions, the largest difference being between the first evaluation and later evaluations. For each assessor and for each session the order within session should be randomised. Finally the order of testing within a pair should be randomised independently for each assessor by pair combination.

Table 4.
The randomisation process is illustrated below:

n	assessor	sessions	order	pairs	reverse	finally
1	1	1	1	DE	no	DE
2	1	1	2	CG	yes	GC
3	1	1	3	BI	no	BI
4	1	1	4	EF	no	EF
.						
.						
19	1	3	1	BD	yes	DB
20	1	3	2	CD	no	CD
21	1	3	3	AE	no	AE
.						
.						
25	1	3	7	AB	yes	BA
26	1	3	8	EG	yes	GE
27	1	3	9	BC	yes	CB
.						
.						
34	1	4	7	DI	no	DI
35	1	4	8	EI	no	EI
36	1	4	9	EH	no	EH

The above randomisation was produced using a computer program written in the GENSTAT statistical computing language. The assessor, session and order within session structure was set up in systematic order. The 36 sample-pairs were generated for each assessor using the labels AB, AC.....HI. These treatment labels are given in alphabetic order. The sample-pairs were then randomised within assessor. If the sub-samples are given in alphabetic order, this creates a bias which is determined by the initial listing of the treatments. It can be remedied by creating for each assessor a factor "reverse" with 18 "no" and 18 "yes" labels. This is then randomised and determines whether or not the alphabetic order is reversed. The "pairs" and the "reverse" variables then give the final order of the sub-samples. In this particular example the bias from effects of session and order within session have been minimised by randomising over these effects. Given more work and knowledge of the variation in this kind of experiment, it would be possible to produce elegant designs in which each pair of sub-samples is compared three times in each session and which are better balanced for order effects. Nevertheless, the randomisation process illustrated above leads to a valid experiment.

4.2 Analysis

In any experiment, it is important to do a little preliminary work learning about the data before proceeding to the definitive analysis. Here, the data can be regarded as one factor (treatment)

with 36 levels by 12 assessors. One way of looking at the data is to regard the assessors as a block factor and to do a randomised block analysis of variance of the following form:

Table 5.
ANALYSIS OF VARIANCE

Source of variation	df
assessor	11
"pairs"	35
Residual	385
Total	431

This analysis of variance allows a preliminary evaluation of the differences between "pairs". Because of the structure of the 36 "pairs", the means should be displayed in a lower triangular format. Re-ordering of the rows and columns may improve the clarity of the results.

The particular structure of the treatments can be further exploited by taking the lower triangular matrix of mean differences and applying the multi-dimensional scaling (MDS) technique to produce the coordinates in the principal sensory dimensions. This analysis and also the analysis of variance are based on the assumption that assessors perceive the differences in the same way and that differences between the results are solely the result of positional factors or random (uncontrolled) variation. However, it is possible to perform more complicated analysis which allows differences between assessors to be taken account of. The best known of these methods is INDSCAL which is available in the SPSS computer program and elsewhere. This method not only provides information about the samples but also about the assessors. In certain circumstances, it may be reasonable to group the assessors and to perform a separate analysis for each group or alternatively to exclude an aberrant assessor. A fuller account of MDS techniques are given in Chapter 6.

The greatest difficulty in using the MDS technique on directly assessed differences is in attributing meaning to the sensory dimensions. Strictly, all that one can know from a difference experiment is whether or not there are sensory differences in the characteristics on which the assessors are comparing samples.

The value of this technique could be greatly improved for work in food research, if it could be shown that each assessor was required to assess only a part of the possible treatment combinations.

Finally, it may not be necessary to directly estimate differences between samples. Given certain assumptions it is possible to compute them from sensory profile data.

There are two advantages in using this route:

1. it is easier to attribute meaning to the underlying sensory dimensions,
2. many more samples can be tested in one experiment.

However, there is undoubtedly a loss of sensitivity in moving to a less direct form of comparison.

5. SENSORY PROFILE EXPERIMENTS

Sensory profile experiments are the most common form of sensory experiment. In the usual form samples are presented sequentially to the assessors, who rate them for attributes given by a vocabulary. The vocabulary is usually fixed for each experiment (fixed profile) but can be a vocabulary personal to each assessor in the case of free choice profiling. In this special case there is less need for a panel to agree on sensory terms and to use external reference standards to clarify the meaning of terms. Other forms of profiling are based on ranking samples for a number of attributes. This form of profiling is not widely used and will not be discussed further.

5.1 Vocabulary Development

For many sensory experiments, in which familiar products are being profiled, a vocabulary will already be in existence. For new products either an existing vocabulary has to be taken over from another laboratory and adapted or an entirely new vocabulary has to be created. In both cases substantial efforts are required before profiling can commence.

A very common way of developing a vocabulary is for the sensory assessors to have a round table discussion with many samples of the product available for rating. At this meeting assessors suggest appropriate terms and by discussion a vocabulary is agreed.

An alternative procedure is to start with a list of possible sensory terms and to present assessors with a wide spectrum of samples and ask them to identify which terms are relevant to each sample. If the assessors carry out this work under normal sensory conditions of isolation then there is value in analysing the data. An illustration of this approach is given in Hunter and Muir (1993).

5.2 Design of Experiment

Sensory profile experiments can be considered to be special kinds of crossover trials (see Jones and Kenward, 1989), which are widely used in medical and biological science. The special feature of sensory profile experiments is that the experimental subjects (the assessors) are not regarded as replicating the measurements. If replication is required then the whole experiment is repeated. Methods of analysis given in the rest of this Chapter assume that there are only simple differences between assessors in the way in which they rate samples. In the following Chapters more complicated ways of modelling the differences between assessors are described.

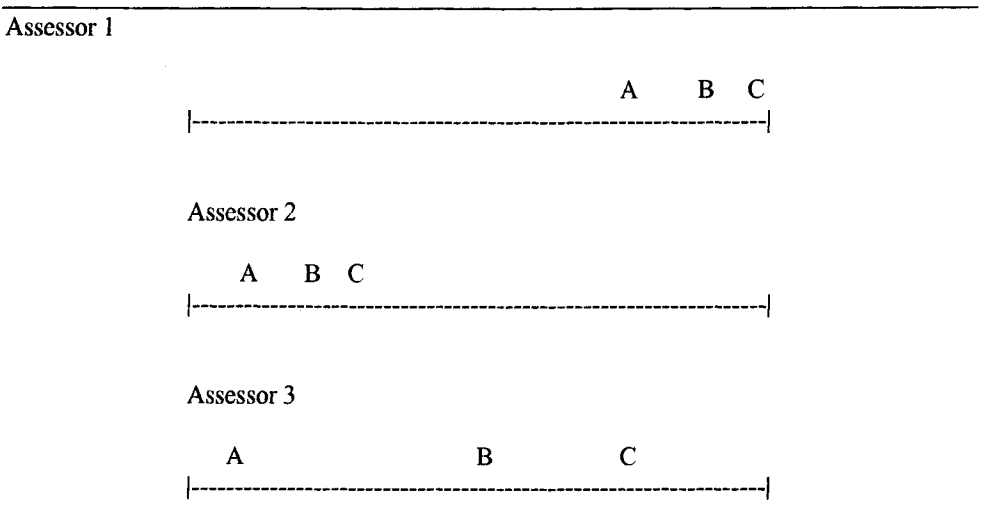
A well designed experiment takes account of known sources of variation by building them into the design. It also randomizes over unknown and uncontrolled sources of variation. In order to maximise the amount of information from the work of the sensory assessors, a sensory scientist must understand how to design an experiment. In addition the experiment must be run in such a way that the design is respected and variation attributable to experimental procedures does not bias the estimates of differences between treatments.

Let us consider an experiment to compare a Cheddar cheese from the institute's experimental dairy A with brands B and C, on sale in the local supermarket. Consider the variable flavour intensity.

The process starts by posing the question, "What is *known* about variability?" From a great deal of previous work it has been established that:

1. assessors use different parts of the scale,
2. assessors use different amounts of the scale.

This can be illustrated by:



It can be seen that Assessor 1 rates the three cheeses higher for Cheddar intensity than does Assessor 2 but nevertheless the differences between the samples are similar. Assessor 3 rates the samples in the same order as Assessors 1 and 2 but uses a larger part of the scale. The question then arises about the design required to minimise these effects. If each assessor rates every sample then the effect of the part of the scale used will have no effect on the differences between samples. By taking means over assessors, the use of different proportions of the scale by different assessors is minimised. It is possible to apply a scaling factor to each assessor's data or to standardise it by calculating normal deviates. Both of these techniques require substantial quantities of data to be effective. Training of assessors and experience in using a vocabulary can help to reduce these kinds of differences between assessors.

Two further features of sensory data are well established:

3. Trends with order of tasting,
4. Carryover effects.

It is well established that there are effects of order of presentation (see for example Muir and Hunter 1991/2). For example in a session in which assessors test 4 products, for positive factors the first tested is often rated higher than justified and the fourth tasted lower. For negative factors the first tested is rated too low and the fourth tested too high. The largest differences are between the sample tested first and those tested later. This effect can be counteracted by asking assessors to make use of a control (unrecorded) sample before the start of a session to familiarise themselves with the product. Alternatively, by ensuring that each sample is tested an equal number of times in each order in a session, it is possible to nullify this effect.

A sample with a strong or otherwise distinctive quality may influence the assessment of subsequent samples (Williams and Arnold, 1991/2). This effect is less well established than the three effects previously discussed. Nevertheless, it is good practice to design sensory experiments with this effect in mind. Special data analysis will then allow this effect to be tested statistically. Schlich (1993b), in a particularly well designed experiment, detected carryover effects in the analysis of the data from an experiment in which four kinds of restructured steaks were compared in a sensory profile experiment. Because the design was balanced for carryover effects, he was able to estimate both the "direct" and "residual" effects of treatment and to calculate the "permanent" effect. Although statistical analysis can adjust for the residual effects of previous treatments, it is preferable that they are minimised by sensory procedures. These procedures include washing out the mouth with water and/or eating a plain (cracker) biscuit between samples, to cleanse the palate. If sensory effects are consistently found then it suggests that the sensory procedures should be modified.

5.2.1 Design possibilities

It is instructive to consider the following possibilities for a design in which 4 samples, with no factorial structure, are compared by 8 or 12 assessors. Below, modules of the designs are given for 4 assessors. The module is repeated twice for 8 assessors and three times for 12 assessors.

Table 6.
Option 1 - Different sample for each assessor

		Order			
		I	II	III	IV
Assessor	A	a	a	a	a
	B	b	b	b	b
	C	c	c	c	c
	D	d	d	d	d

In this design each assessor receives only one treatment. This has two undesirable consequences:

1. Differences between treatments are confounded (confused) with differences between assessors. If assessor A rates sample "a" four times and no other assessor rates this sample and similarly for samples "b", "c" and "d" then any systematic differences between assessors will contaminate the assessment of treatment differences. Since it is known that different assessors use different parts of the scale (even after extensive training) this design will provide very poor estimates of the differences between samples.

2. If a design such as this is commonly used in a sensory laboratory then assessors will soon learn that the same sample is repeated many times. The consequence will be that the second and subsequent ratings will not be independent because assessors will strive to be consistent with the first rating. Consequently they will add very little information to the first rating. Sensory assessors can be very quick to identify pattern in the sequence of samples being presented and can be expected to react to these perceptions.

Table 7.
Option 2 - Same order for each assessor

		Order			
		I	II	III	IV
Assessor	A	a	b	c	d
	B	a	b	c	d
	C	a	b	c	d
	D	a	b	c	d

In this design each assessor receives each sample, so differences between samples are not confounded with differences between assessors. However samples are given to each assessor in the same order. This has two consequences:

1. Differences between treatments are now confounded with order differences. Although this is less serious than confounding sample differences with assessor differences (Option 1), it is not desirable. For some sensory trials of hot foods it may not be possible to use different orders of presentation for each assessor. In these circumstances, it is very important for sensory assessors to receive a priming sample prior to rating the experimental samples.
2. In experiments in which the sample order is the same for every assessor, it is difficult to ensure that assessors are unaware of the samples they are assessing, particularly if the assessments are not done simultaneously. Assessors who have completed the task may pass information to other assessors who will not then make independent ratings of the products.

Table 8.
Option 3 - Latin Square

		Order			
		I	II	III	IV
Assessor	A	a	b	c	d
	B	b	c	d	a
	C	c	d	a	b
	D	d	a	b	c

This design is based on a Latin Square which is produced by cyclic development of an initial row which is in the same order as the first column.

This is a special kind of Latin Square which can always be generated. It allows for assessor and order effects and is thus better than *Options* 1 and 2. Nevertheless, inspection reveals that in this particular form of cyclic Latin Square, sample "a", for example, always follows sample "d". The other defect of this design is that the sequence of treatments is the same for each assessor, though not the order, and so susceptible to anticipation by assessors.

Table 9.
Option 4 - Williams Latin Square

		Order			
		I	II	III	IV
Assessor	A	a	d	b	c
	B	b	a	c	d
	C	c	b	d	a
	D	d	c	a	b

This design, too, is a Latin Square also generated by a cyclic method of construction from an initial row and has the property that each treatment follows every other treatment once. The first row is generated by a method due to Williams (1949). For an even number of rows, columns and treatments balance can be achieved by one square whilst for an odd number two squares are required. This method of design has been promoted in the context of consumer trials by MacFie, Greenhoff, Bratchell and Vallis (1989). In Order II "a" follows "b", in Order III "a" follows "c" and in Order IV "a" follows "d" thus overcoming the defect of *Option 3*.

Only Option 4 is wholly satisfactory for sensory experiments.

5.2.2 Designs based on mutually orthogonal Latin Squares

If a design for 4 treatments and 12 assessors is required then it is possible to generate a design with a higher level of balance for previous treatments than by simply repeating three times the module for 4 assessors given by *Option 4*. Two Latin Squares (of side n) are said to be mutually orthogonal, if, when they are superimposed, for each of the symbols of the first

square the n symbols of the second square are different. At most there can be n-1 mutually orthogonal Latin Squares, however for many integers a full set does not exist. Fisher and Yates tables (1963) give 2 squares for side 3, 3 for 4, 4 for 5, 1 for 6, 6 for 7, 7 for 8, 8 for 9. Mutually orthogonal Latin Squares are available for higher orders in specialised books.

Table 10
Option 5 - Orthogonal Latin Squares

		Order			
		I	II	III	IV
Assessor	A	a	b	c	d
	B	b	a	d	c
	C	c	d	a	b
	D	d	c	b	a
	E	a	c	d	b
	F	b	d	c	a
	G	c	a	b	d
	H	d	b	a	c
	I	a	d	b	c
	J	b	c	a	d
	K	c	b	d	a
	L	d	a	c	b

Inspection of the above design reveals that each assessor rates each sample once and that each sample is tested 3 times in each Order. This design is also balanced for previous treatment in every Order. For example in Order II, treatment "a" follows "b", "c" and "d". However, it should be noted that if the design module for the first four assessors is inspected, treatment "a" follows "b" in Order II, "d" in Order III and "b" again in Order IV. Consequently, although a design based on two or more orthogonal squares may have better properties than a design based on the Williams Latin Square, a design based on one square is not superior.

5.2.3 Replication of assessor by sample allocations

In the sensory literature the meaning of replication is not always as clear as in biological experimentation. If sensory experiments are viewed from this standpoint, then assessors may be regarded as replicate blocks. This leads to experiments in which each assessor rates each sample once and to a randomised block form of analysis of variance for each variable with replicates=blocks=assessors. However, most statisticians working with sensory data would not regard this as being a replicated experiment. If the assessor by sample measurements are replicated by repeating the design, with a different randomisation, then it is possible to quantify the ability of each assessor to reliably measure each attribute, Næs and Solheim (1991). Also, it

is possible to explore the assessor by sample space and so monitor each assessors use of vocabulary and training needs. Three replicates are usually sufficient to allow this to be done.

In a replicated experiment, the sub-samples for all replicates are usually drawn from the same samples. Differences between replicates or interactions between sample and replicate, in addition to sampling and testing variation, may be attributed to the effects of storage of the samples and to the small differences in the environment for each replicate of the experiment.

Sensory experiments which test differences between husbandry or carcass processing treatments on meat yielding animals are particularly difficult to organise. Overall differences between treatments are likely to be small and there is a great deal of variation between animals. This is often increased by the lack of control of important variables at slaughter, during processing of the carcasses and during cooking. Freezing samples of meat and later thawing them may be convenient for the sensory laboratory but it will reduce differences between treatments. A more subtle disadvantage is that the experiment then makes inferences about samples of meat that have been frozen and not about fresh meat. Finally, there are technical problems in carrying out the sensory work where it is known that small differences in temperature at serving can have a major effect on the sensory characteristics. It is recommended that in replicated experiments on meat, samples from different carcasses are used for each replicate, see for example Vipond, Marie and Hunter (1995).

5.3 More Than One Session Per Replicate

Assessors can usually only assess a small number of samples in a session before suffering sensory fatigue and a lowering of the level of performance. Depending on the product, the assessor's experience and the workload of the test, as few as three sub-samples may be rated in a session or as many as eight. In normal circumstances it is not usual to exceed this limit. It is not sensible to restrict experiments to the number of sub-samples that can readily be rated in one session.

Supposing a food manufacturer has commissioned the sensory laboratory to profile all blackcurrant cordials on sale in the local supermarket shelves. Twelve different products are found, coincidently 12 assessors are available. A number of different ways of organising the sensory testing are discussed below.

5.3.1 Three separate experiments

Samples a-d are evaluated in experiment 1 (Orders I-IV), samples e-h in experiment 2 (Orders V-VIII) and i-l in experiment 3 (Orders IX-XII). Three copies of *Option 4* or preferably *Option 5* are used to give the order of sub-samples for each assessor. The advantage of the design is that only four samples are used in each experiment and work can be completed on the first four samples before proceeding to the second four samples etc. The disadvantage of this design is that whilst samples within an experiment are compared with the highest level of precision, differences between samples evaluated in different experiments are confounded by the overall effects of experiment. These effects are unlikely to be negligible relative to the within experiment variation.

5.3.2 Split plot designs

If the design is replicated, say 3 times, then the alternatives are to do all replicates of "experiment" 1, then all replicates of "experiment" 2 and finally all replicates of "experiment" 3 (Design A below). The design thus takes 9 sessions in each of which 4 samples are rated. An alternative arrangement of sessions is to perform all the first replicates, followed by all the second replicates followed by all the third replicates. Possible arrangements of sessions are:

Table 11.

Session	Design A	Design B
1	a-d	a-d
2	a-d	e-h
3	a-d	i-l
4	e-h	e-h
5	e-h	a-d
6	e-h	i-l
7	i-l	i-l
8	i-l	e-h
9	i-l	a-d

Design A completes work on samples a-d before starting on samples e-h and i-l. Assessors will become increasingly familiar with the product with each session and so it is possible that the later samples (i-l) will be more precisely rated than the earlier samples (a-d). Design B allows for trends over time and makes particularly good sense if three sessions are done per day. The sessions form a Latin Square with columns equal to days and rows equal to order in the day. In all but exceptional circumstances Design B should be preferred to Design A.

If ideas about split plots from biology are applied to this sensory experiment, in each session some assessors would test samples a-d, others e-h whilst others tested i-l. This is of little advantage to the sensory technician and there is consequently no reason for using this type of design. Only in exceptional circumstances, related to the nature of the samples, will a traditional split plot design have advantages.

5.3.3 Williams Latin Square designs

Consider a Latin Square of size 12 for 12 assessors, 12 periods and 12 samples. By using the first four columns (Orders I-IV) for the first session, the second four columns (Orders V-VIII) for the second session and the last four columns (Orders IX-XII) for the third session a design is produced which preserves balance for assessor, session and Order within session. Although there is no longer complete balance for previous effects, each treatment appears first in a session three times and follows 9 of the 11 other treatments in the second, third and fourth order within each session.

Table 12.

Assessor	Order											
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	VII
A	a	l	b	k	c	j	d	i	e	h	f	g
B	b	a	c	l	d	k	e	j	f	i	g	h
C	c	b	d	a	e	l	f	k	g	j	h	i
D	d	c	e	b	f	a	g	l	h	k	i	j
E	e	d	f	c	g	b	h	a	i	l	j	k
F	f	e	g	d	h	c	i	b	j	a	k	l
G	g	f	h	e	i	d	j	c	k	b	l	a
H	h	g	i	f	j	e	k	d	l	c	a	b
I	i	h	j	g	k	f	l	e	a	d	b	c
J	j	i	k	h	l	g	a	f	b	e	c	d
K	k	j	l	i	a	h	b	g	c	f	d	e
L	l	k	a	j	b	i	c	h	d	g	e	f
	Session 1				Session 2				Session 3			

This is a very general method of construction for even numbers of treatments. The advantage is that all differences between pairs of treatments are estimated with almost equal precision. The disadvantage is that the number of assessors must equal the number of treatments. Also, another disadvantage of such a design is that the sensory technician has to manage all 12 samples in each session. In a well organised sensory laboratory with trained and experienced sensory technicians, this will usually be possible for products tested when cold but may be difficult or even impossible for products tested when hot. The only sound reason for not testing all samples in each session is that practical considerations do not allow it. An example of this design of experiment is given by Muir and Hunter (1991/2).

5.3.4 Incomplete block designs

From a combinatorial standpoint incomplete block designs have excited the curiosity of statisticians for 60 years. Much work has concentrated on the identification of balanced and partially balanced designs. Apart from their intrinsic interest these designs can be analysed using a calculator. However, partially balanced designs are not necessarily statistically efficient designs nor are they available for all combinations of numbers of samples, assessors and samples per session. Except in exceptional circumstances, data from the sensory laboratory will be analysed using a statistical package, so ease of analysis by hand calculator is not a necessary property. For crop variety trials Patterson, Williams and Hunter (1978) and for consumer trials MacFie, Greenhoff, Bratchell and Vallis (1989) give catalogues of efficient designs. There is a need for the publication of catalogues of designs especially for sensory work. Both balanced incomplete block designs and partially balanced incomplete block designs have been advocated for sensory work by several authors. In general the utility of these designs is compromised by

the lack of balance for order and carryover effects and by constraints on the number of assessors.

In sensory experiments, it is very important that each assessor evaluates every sample even if testing extends over many sessions. However, experiments involving consumers are different; they can only reasonably be asked to rate a small number of samples, probably no more than four, and replication is not possible. The experimental design given above for twelve assessors could be adapted for thirty six consumers (each rating four samples) by defining each consumer as an assessor by session. This design would confound some information between samples with differences between consumers. Nevertheless, it allows twelve samples to be compared simultaneously. Similar designs can be generated for other numbers of samples.

5.4 Tailoring Standard Designs

The designs given in this Chapter are for fixed numbers of assessors. The methods of construction lead to designs in which the number of assessors is determined by the number of samples. This does not correspond to the situation that the sensory scientist experiences in the laboratory. When working with an (trained) external panel, the number of assessors is fixed by those available on the day. It is not feasible to increase (or reduce) this number to utilise a particular design for the number of samples being tested. Also, it is not uncommon for one or more assessors to be absent from a session or to fail to complete it. Both the design and analysis systems must be sufficiently robust to deal with the realities of running a sensory laboratory.

For example, suppose there is a trained panel of fourteen subjects but that the number available for sensory work is thirteen. If the first experiment of the day is to profile 12 cheeses, the design given above for 12 assessors may be adapted by using each row of the design once for an assessor and then choosing an extra row from the design (at random) for the thirteenth assessor. Conversely, suppose that the second experiment requires 16 fruit cordials to be rated. A design can be produced for 16 assessors using the Williams Latin Square method of construction for 16 treatments and 16 assessors. At random 3 rows can be dropped from the design. The design will not be as balanced as the full design but nevertheless will still have good statistical properties. The cost of tailoring the design to fit the circumstances of the sensory laboratory is to make the analysis dependent on using more complicated general statistical methods rather than relatively simple methods.

If it is envisaged that the design, given in the previous section, is to be replicated three times then at present the best advice is to randomise the assessors to rows independently for each replicate. An added precaution is to randomise the samples to the sample labels separately for each replicate. This leads to a three replicate design with good but not optimal properties.

There is a need to improve the designs in use in sensory laboratories. Schlich (1993b) gives a design, based on mutually orthogonal Latin Squares, in which there is an exceptional degree of balance over replicates for each assessor and also within replicates. However, design work must pay attention to the fact that sensory experiments are changeover designs with assessor, assessor by replicate, assessor by replicate by session, order and carryover effects.

5.5 Data

As with the direct assessment of differences (see section 4) assessors often rate samples on a 0-5, 1-7 or 0-9 scale. There is often an attempt to improve the performance of short scales by allowing half points. Very often these half points are infrequently used and the data become potentially more difficult to analyse. These difficulties are seldom recognised and almost always ignored. If a scale is to be used then it is more sensible to use a longer scale ie 0-9 rather than a shorter scale with half points. The most satisfactory method of recording responses is to use an undifferentiated line scale which consists of a line with anchor points at both ends. Assessors record their ratings for a sub-sample by making a mark on the line corresponding to the intensity of the sensory stimulus. The data can be collected using pencil and paper but is more efficiently organised by a computerised data collection system which captures data directly from the assessor by allowing the use of a mouse to move a cursor along a line on a screen. It also simplifies data management and is justified for all but the smallest sensory laboratories.

5.6 Analysis

Data are collected on many variates in each profile experiment. These data deserve to be analysed one variate at a time ie on a univariate basis. Chapter 4 deals with the summary of this data on a multivariate basis.

Both univariate and multivariate analysis have a part to play in understanding the results. Experimenters with little knowledge of available statistical methods for the analysis of data often use only univariate methods. Frequently, even these methods are poorly implemented. In this Chapter, it is assumed that the sensory scientist has access to a computer and that it is loaded with a relatively simple, easy to use statistics package such as STATGRAPHICS, MINITAB, SYSTAT or SPSS. Descriptions of how to calculate simple statistics using a calculator are given in O'Mahony (1985).

All the methods described in this section assume that data are from a continuous distribution and are on an interval scale. Although this is unlikely to be fully true, the methods of analysis advocated are robust to these assumptions.

5.6.1 Methods of analysis

Assuming that data are from a properly designed experiment, six models (Model 1 - Model 6) can be fitted. The simplest method of analysis is to find an (arithmetic) mean and a standard error of mean for each sample (Model 1). However, the error is contaminated by between assessor, order and other effects which inflate the error and reduce the power of tests of significance. Also, if a separate error is determined for each sample, it will be based on few degrees of freedom. Since there is no reason to expect that errors will be different for each sample, pooled errors should be used. Means and pooled errors are conveniently obtained from a one way analysis of variance with sample as the treatment factor. A more precise analysis is obtained by allowing for assessors (Model 2). This can be done using the analysis of variance. More generally, it can be done using a general(ised) least squares method, sometimes called regression with factors. It is also possible to allow for order of presentation (Model 3). Since assessor effects are large, it is reasonable to allow for replicate-by-assessor effects (Model 4). All sensory scientists should endeavour to get to this stage of univariate analysis. They should find the results informative and well worth the extra effort involved. A further stage is to allow

for assessor by replicate by session effects and estimate sample effects solely within sessions for each assessor (Model 5). Information on samples that is partially confounded with sessions is lost. There is thus a trade off between lower mean squares and loss of statistical efficiency.

For those with access to more sophisticated statistical programs (SAS, GENSTAT) there is some merit in fitting a (general) mixed model to the data. It is reasonable to regard the assessor, the replicate within assessor and the session within replicate within assessor as random effects (Model 6). This type of model allows the information on samples confounded with sessions to be properly weighted and thus the standard errors from this analysis are invariably and correctly lower than those from Model 5. The main advantage is that the sources of variation can be carefully modelled and the information can be used to plan more precise experiments in the future.

For replicated experiments, it is useful to obtain estimates of sample by replicate effects for further analysis. Given these effects, it is possible to do a standard two way analysis of variance free of the effects of sensory measurement. When the samples have structure (section 6) the degrees of freedom and sums of squares for sample can be partitioned in an informative way.

Few laboratories routinely estimate carryover effects. However, it is important to check that these effects are relatively small by carrying out a special analysis of the data from time to time. Schlich (1993b) shows how this can be done for a special design using analysis of variance. Using the mixed model approach, an analysis to estimate residual effects can be carried out using either GENSTAT or SAS.

5.6.2 Example

A sensory panel of 13 assessors tested 8 blackcurrant cordials in two replicates each of two sessions. Four cordials were rated on a continuous scale 0-100 in each session. A design based on Williams Latin Square was used to determine the treatment sequences for each assessor. One assessor was not able to attend one of the sessions of the second replicate so the assessor by replicate by treatment table was not complete.

The purpose of analysis is the estimation of sample effects with an appropriate estimate of their variability ie standard error of mean (sem). The results of fitting all six models to the flavour intensity variable are given below.

Table 13.

	Model 1		Model 2	Model 3	Model 4	Model 5	Model 6
Sample	Mean	sem	Mean	Mean	Mean	Mean	Mean
1	53.6	3.94	53.5	53.7	53.3	55.4	54.5
2	58.6	4.00	58.7	58.6	58.1	57.4	58.4
3	60.6	3.67	60.7	60.8	60.2	58.3	60.4
4	59.8	3.72	59.8	59.8	59.4	57.8	60.0
5	59.8	4.33	59.7	59.7	59.3	60.6	60.7
6	65.8	3.55	65.9	65.7	65.1	62.9	64.9
7	65.0	4.06	65.2	65.2	64.6	63.4	64.9
8	68.0	3.89	68.0	68.0	67.6	67.2	68.3
av sem	3.90		3.29	3.31	3.19	3.28	3.17
EMS	372		264	267	248	222	230

The Error Mean Square (EMS) provides overall evidence of how well the model fitted the data. Allowing for the assessor effect (Model 2) reduces the EMS from 372 to 264. Allowing for order of presentation (Model 3) does not reduce the EMS further whereas allowing for a separate assessor effect for each replicate (Model 4) reduces the EMS to 248. By allowing for session (Model 5) the EMS is reduced to 222 but because some information on treatments is lost between sessions the sem rises to 3.28 from 3.19. However, regarding assessor, replicate within assessor and session within replicate within assessor as random effects causes the estimate of the EMS to fall to 230 and the sem to 3.17. Model 6 is arguably the most appropriate analysis.

Because the experiment was carefully designed there are only minimal differences in the estimates of sample effects between the models.

6. TREATMENT DESIGN

In sections 4 and 5 of this Chapter, the assignment of order of testing of samples for each assessor has been considered. In this section, the structure of the sample space is considered. In studies of the sensory properties of products on sale in supermarkets the samples do not have a simple structure but in research or development studies the opportunity exists to impose a factorial treatment structure on samples. In planning the treatment structure, it is very important to define the objectives carefully and not to artificially restrict the problem to one that is assumed to be susceptible to experimentation. It is also very important to review existing knowledge and to separate hard information from conjecture.

It should be remembered that sensory experiments have much in common with other scientific experiments and accordingly, methods in use in other areas of science and technology are relevant. Many sensory analysts instinctively feel that in a multifactor situation, an experiment should be performed with each factor in turn holding the levels of all other factors constant. It has been shown that this is a very bad strategy which uses experimental resources

wastefully and in addition often fails to determine the optimum, Chapter 5 of Cochran and Cox (1957). A better strategy is to evaluate all relevant factors simultaneously in a sequence of experiments. It is recommended that only about 25% of available resources should be allocated to each experiment. In the light of the results obtained later experiments in the series can be planned. This strategy puts a high premium on being able to quickly analyse experimental data, formulate the conclusions and design the next experiment.

6.1 Dose Response Experiments

The simplest treatment structure arises when the experimenter wishes to investigate only one factor. Suppose that the effect of level of sweetener in a fruit cordial is to be investigated and there are resources to run a sensory experiment with 4 samples. It is a relatively simple matter to devise four factor levels. If the experimenter expects a linear response to the factor then the best way to arrange the treatments is at equal intervals. Using existing knowledge a base level is determined and a suitable increment.

Table 14.

Sample	Treatment Level
1	base
2	base + incr*(2-1)
3	base + incr*(3-1)
4	base + incr*(4-1)

If the base is 50g of sugar per litre and the increment is 10g then the levels are 50,60,70 and 80g of sugar per litre.

For many sensory stimuli, the response may be related to the log of the treatment. For example, sensory sweetness may be proportional to the log of the added sweetener. In these cases the treatments levels should have a ratio relationship to each other.

Table 15.

Sample	Treatment Level
1	base + incr
2	base + incr*r
3	base + incr*r^2
4	base + incr*r^3

Suitable values of r can be as large as 10.0 or as small as 1.5. If the base is 0 and the increment 10g of sugar per litre and r=2 then the levels are 10, 20, 40 and 80g of sugar per litre.

The choice of treatment levels depends on hard information and on prior knowledge or conjectures about the shape of the response.

In different replicates of the sensory experiment, it is an advantage to have different realisations of the treatment specification. In this experiment it would mean making up a fresh set of samples for each replicate. This provides a more severe test of the treatments.

6.2 Full Factorial

In a full factorial design the samples consist of all possible combinations of two or more treatment factors each with at least two levels. The number of samples required is given below.

Table 16.

	Levels per factor			
No of factors	2	3	4	5
2	4	9	16	25
3	8	27	64	125
4	16	81	256	625
5	32	243	1012	3125

Few sensory scientists are prepared to contemplate testing more than 30 samples in an experiment. From a sensory viewpoint, it is therefore only feasible to do a full factorial with two, three, four or five factors each at two levels; two or three factors each at three levels and only two factors at four or five levels.

An example will help to illustrate this class of design. Suppose that work is being done on very low fat yogurts. Three factors which are under control of the experimenter are the Type (A or B) of homogeniser, the homogenisation Pressure (low or high) and the Temperature at homogenisation (low or high). All these factors may effect the sensory properties of the final product.

Setting the levels for quantitative factors such as Pressure and Temperature requires some knowledge of the possible operating range. These will usually be defined by existing knowledge but can be defined by a phase of experimentation prior to sensory profiling when physical or chemical measurements are made on the samples.

The process illustrated above can be generalised to factors with more than two levels. It is possible to have several treatment factors with different numbers of levels in one experiment. For design purposes a mixture of factors with 2 and 4 levels are preferable to mixtures of 2 and 3 levels.

Table 17.
The treatment combinations for the samples are:

Sample	Type	Pressure	Temperature
1	A	low	low
2	A	low	high
3	A	high	low
4	A	high	high
5	B	low	low
6	B	low	high
7	B	high	low
8	B	high	high

The interaction between two factors with two levels is defined as the difference in the effect of the second factor between the levels of the first factor or conversely the difference in the effect of the first factor between the levels of the second factor. Three factor interactions are defined similarly. The advantage of a full factorial experiment is that all the degrees of freedom between samples can be uniquely attributed to a main effect of a factor or an interaction.

It is recommended that the analysis of sensory data proceeds as follows. In the first part of the analysis, tables of sample by replicate effects are obtained adjusted for the effect of assessor. These tables are then further analysed by analysis of variance or by regression in the case of response surface data (section 6.5).

If there are three replicates of the 8 samples in the experiment outlined above, the form of the analysis of variance will be:

Table 18.
ANALYSIS OF VARIANCE

Source	df
Replicates	2
Samples	7
Partitioned	
Type	1
Pressure	1
Temperature	1
Type.Pressure	1
Type.Temperature	1
Pressure.Temperature	1
Type.Pressure.Temperature	1
Error or Residual	14
Total	23

An example of the use of this type of design is given by Muir, Banks and Hunter (1992).

6.3 Fractional Factorial

In most research and development projects there are many potential factors which could affect the sensory variables. There is a natural tendency for sensory scientists to simplify the problem or to partition the problem into a number of experiments in order to allow full factorial designs to be used. As shown by the table at the start of the previous section, too many samples are required for a full factorial experiment with five or more factors. However, it is possible to carry out informative experiments requiring a small number of samples by making some assumptions. In all areas of experimentation it is usual to find that main effects are much larger than two factor interactions, which are larger than three factor interactions etc. Only seldom are interactions important when main effects are small. Fractional factorial designs confound information on high order interactions with main effects or low order interactions. Thus if an effect is significant, it is assumed that the main effect or lower order interaction is responsible.

Fractional factorial designs were first discussed by Finney (1945). Box, Hunter and Hunter (1978) give a relatively gentle treatment of this topic. A more comprehensive account is given in Cochran and Cox (1957).

It is instructive to review the design that was considered in the previous section and add the maximum number of factors using the fractional factorial method of construction. From the catalogue of designs given in Cochran and Cox (1957), a fractional factorial design for 5 factors, each at two levels, requiring 8 samples is found. Thus two additional factors, Extra 1 and Extra 2 can be added.

Table 19.

Sample	Type	Pressure	Temperature	Extra 1	Extra 2
1	A	low	low	low	low
2	A	low	high	high	low
3	A	high	low	high	high
4	A	high	high	low	high
5	B	low	low	high	high
6	B	low	high	low	high
7	B	high	low	low	low
8	B	high	high	high	low

The consequences for the analysis of variance are:

Table 20.
ANALYSIS OF VARIANCE

Source	df
Replicates	2
Samples	7
Partitioned	
Type	1
Pressure	1
Temperature	1
Extra 1	1
Extra 2	1
Type.Temperature (=Pressure.Extra 1)	1
Pressure.Temperature (=Type.Extra 1)	1
Error or Residual	14
Total	23

Of the 10 potential two factor interactions only two effects can be estimated orthogonal to the main effects. By simple algebra it is possible to show that each of these effects corresponds to two two factor interactions which are said to be aliased ie inseparable. Of the remaining two factor interactions, four are uniquely aliased to main effects and two are aliased to the same main effect. There is thus only minimal information available on interactions. However, the five main effects are estimable. Provided that the experimenter is willing to assume that interactions are likely to be unimportant then this type of design can be justified.

For 4 and 8 samples it is possible to take the fractional factorial method of construction a stage further and to construct designs in which all the degrees of freedom between samples are uniquely identified with a main effect. Such designs are referred to as saturated designs and are of great utility in the exploratory stage of development studies. A design for 7 factors at two levels for 8 samples is given below:

Table 21.

	Factors						
Sample	1	2	3	4	5	6	7
1	low	low	low	low	low	low	low
2	high	high	high	high	low	low	low
3	high	high	low	low	high	high	low
4	high	low	high	low	low	high	high
5	high	low	low	high	high	low	high
6	low	high	high	low	high	low	high
7	low	high	high	high	low	high	high
8	low	low	low	high	high	high	low

Plackett and Burman (1946) give a more general method of construction for saturated designs which gives designs for factors all with the same prime number of levels .

Taguchi, the Japanese engineering management guru, has popularised the use of designs with many factors and few samples in an engineering context. These designs have come to be known as Taguchi designs or arrays and are given in a number of books (see, for example, Logothetis and Wynn, 1989).

6.4 Response Surface Designs

If the purpose of an experiment is to optimise the settings of a number of quantitative factors, then it is of advantage to use response surface designs. Simple response surface designs are fractional factorial designs with centre points. The so called "star" points can be added in a second replicate. A response surface design allows the data to be analysed by fitting a regression type model to the sample by replicate means. Contour plots can be drawn in the parameter space and an optimum located. An account of these designs is given in Box, Hunter and Hunter (1978) and in Box and Draper (1987). An example of the use of a response surface design in a sensory experiment is given by Muir, Hunter, Guillaume, Rychembusch and West (1993).

6.5 Replication of Fractional Replicate And Response Surface Designs

For these designs, there are advantages in selecting a complimentary set of samples for the second and subsequent replicates. In many cases it will be wise to analyse each replicate as it is completed and to judge whether the treatment levels should be modified for the next replicate.

7. POWER OF EXPERIMENTS

When planning experiments sensory scientists should be aware of the precision and should try to avoid planning experiments which are doomed to failure. Because the differences between two treatments are not significant, it does not mean that differences do not exist but only that they are smaller than the detection threshold. On occasion statistically significant differences will be found which are too small to be of any economic, technological or scientific importance.

Using an estimate of the EMS it is possible to calculate a minimum detectable difference between two samples for an experiment with n_r replicates and n_a assessors from the formula

$$\text{Detectable Difference} = 3 \times \sqrt{\frac{2 \times \text{EMS}}{n_r \times n_a}}$$

The factor 3 is an ad hoc value derived from the "t" value augmented to allow for the fact that the error is estimated from the data. If measurements are on the scale 0-100, then estimates of the EMS vary from 100-500 with 300 representing an acceptable level of variability for a variable with differences between samples. In general the bigger the differences between samples the bigger the EMS. For laboratories which use a different scale, a factor can be derived by dividing the range by 100 and dividing the EMS in the table below by the factor

squared and the table entries by the factor. The table illustrates the consequences of this formula for the detectable difference:

Table 22.

		EMS		
n_r	n_a	100	300	500
1	8	15.0	26.0	33.5
1	12	12.2	21.2	27.4
1	16	10.6	18.4	23.7
2	8	10.6	18.4	23.7
2	12	8.7	15.0	19.4
2	16	7.5	13.0	16.8
3	8	8.7	15.0	19.4
3	12	7.1	12.2	15.8
3	16	6.1	10.6	13.7

Small differences are unlikely to be detected with a modest sized experiment.

8. RELATIONSHIP OF UNIVARIATE METHODS TO MULTIVARIATE METHODS

In addition to assuming that data are from a continuous distribution and are on a linear scale, the univariate methods assume that each assessor measures the samples in the same way. Assessors are known to use different proportions of the scale, and use sensory terms in different ways. Selection, training and reference standard have a part to play in reducing these differences but can seldom eliminate them. Another weakness of presenting many univariate analyses is that there is the implication that there is more information than truly exists. Principal Component Analysis of tables of sample means or replicate by sample means nearly always reveals a very highly related set of variables with few significant components.

The multivariate methods described in the rest of this book model the differences between assessors more realistically than univariate analysis and present the results in the principal sensory dimensions.

The validity of the results from multivariate analysis depends on proper experimental design and particularly on randomisation and blinding. Caution must be exercised in interpreting the multivariate analysis of sets of sub-experiments (section 5.3.1) and split-plot experiments (section 5.3.2) because differences between samples are measured with different degrees of precision.

9. CONCLUSIONS

The key to successful experimentation is clear analysis of the problem followed by careful design of the sensory experiment and skilful analysis of the data which is followed by perceptive interpretation of the results.

Much work remains to be done to develop designs for sensory experiments and in particular sensory profile experiments which are free of artificial constraints on numbers of assessors and numbers of samples per session. Knowledge of good designs should be disseminated by catalogues or preferably by computer programs which produce properly randomised designs with minimal inputs from the sensory scientist.

ACKNOWLEDGEMENTS

My thanks are due to Professor D Donald Muir (of the Hannah Research Institute, Ayr, Scotland) who introduced me to statistical work in Food Science and who has been a most stimulating and demanding collaborator as well as a congenial colleague. The Scottish Office Agriculture and Fisheries Department (SOAFD) funded my work. The EU provided travel and subsistence funds for me to participate in the FLAIR Concerted Action No 2 - SENS and thereby become aware of the work of sensory scientists throughout Europe.

10. REFERENCES

Box G E P and Draper N R, 1987. Empirical Model-Building and Response Surfaces, John Wiley & Sons, New York.

Box G E P, Hunter W G and Hunter J S, 1978. Statistics for Experimenters, John Wiley & Sons, New York.

Cochran W G and Cox G M, 1957. Experimental Designs (2nd edition). John Wiley & Sons, inc. New York.

Collett D, 1991. Modelling Binary Data, Chapman & Hall, London.

Finney D J, 1945. The Fractional Replication of Factorial Arrangements. Annals of Eugenics, 12, 291-301.

Fisher R A and Yates F, 1963. Statistical Tables for Biological Agricultural and Medical Research (6th edition). Oliver and Boyd Ltd, Edinburgh.

Gacula M C and Singh J, 1984. Statistical Methods in Food and Consumer Science, Academic Press, Orlando, Florida, USA.

Hunter E A and Muir D D, 1993. Sensory properties of fermented milks: objective reduction of an extensive sensory vocabulary. Journal of Sensory Studies 8, 213-227.

John J A and Quenouille M H, 1977. Experiments, design and analysis (2nd edition). Griffin, London.

Jones B and Kenward M G, 1989. Design and Analysis of Cross-Over Trials, Chapman and Hall, London.

Logothetis N and Wynn H P, 1989. Quality Through Design. Clarendon Press, Oxford.

MacFie H J H, 1986. Aspects of Experimental Design. Chapter 1 in Statistical Procedures in Food Research, edited by J R Piggott. Elsevier Applied Science, London and New York.

MacFie H J H, Greenhoff K, Bratchell N and Vallis L, 1989. Designs to balance the effect of order of presentation and first-order carry-over effects in hall tests. Journal of Sensory Studies, 4, 129-148.

Mead R, 1988. The Design of Experiments. Cambridge University Press.

Muir D D and Hunter E A, 1991/2. Sensory Evaluation of Cheddar Cheese: Order of Tasting and Carryover Effects. Food Quality & Preference, 3, 141-145.

Muir D D, Banks J M and Hunter E A, 1992. Sensory changes during maturation of fat reduced Cheddar cheese: effect of addition enzymically attenuated starter cultures. Milchwissenschaft 47, 218-222.

Muir D D, Hunter E A, Guillaume C, Rychembusch V and West I G 1993. Ovine Milk. 5. Application of response surface methodology to manipulation of the organoleptic properties of set yogurt. Milchwissenschaft 48, 609-614.

Næs T and Solheim R, 1991. Detection and interpretation of variation within and between assessors in sensory profiling. Journal of Sensory Studies, 6, 159-177.

O'Mahony M, 1985. Sensory Evaluation of Food. Marcel Dekker, inc., New York.

Patterson H D, Williams E R and Hunter E A, 1978. Block designs for variety trials. Journal of Agricultural Science 90, 395-400.

Robson, C, 1973. Experiment, Design and Statistics in Psychology. Penguin Books, Harmondsworth, UK.

Schiffman S S, Reynolds M L and Young F W 1981. Introduction to Multidimensional Scaling, Academic Press Inc., New York.

Schlich P, 1993a. Risk Tables for Discrimination Tests. Food Quality & Preference 4, 141-151.

Schlich P, 1993b. Uses of Change-over Designs and Repeated Measurements in Sensory and Consumer Studies. Food Quality & Preference 4, 223-235.

Stone H and Sidel J L, 1993. Sensory Evaluation Practices, 2nd edition, Academic Press, San Diego.

Williams A A and Arnold G M, 1991/2. The Influence of Presentation Factors on the Sensory Assessment of Beverages. Food Quality & Preference 3, 101-107.

Williams E J, 1949. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Scientific Research, Series A, 2, 149-168.

Vipond, J E, Marie S and Hunter E A, 1995. Effects of clover and milk in the diet of grazed lambs on meat quality. Animal Science 60, 231-238.

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

PREFERENCE MAPPING FOR PRODUCT OPTIMIZATION

Jean A. McEwan

Department of Sensory Science, Campden & Chorleywood Food Research Association, Chipping Campden, Gloucestershire, GL55 6LD, United Kingdom

1. INTRODUCTION

1.1 Background

Product optimization is the aim of every food manufacturer. A company's ability to produce a product which satisfies the consumers' sensory requirements has a distinct lead to success and profitability. Clearly other factors do come into play, such as packaging and brand image, but these are not the subject of this chapter.

In writing about product optimization from the sensory point of view, it is important to realise the complexity of achieving the ideal product, either for an individual consumer, or a group of consumers making up a market segment. It is critical to understand the requirements of the consumers within the market segments of interest to the company, and thus to design and target products to meet these requirements.

Sensory analysis is frequently carried out by companies in the initial steps of product development or as a quality control tool, and provides valuable information in these instances. Consumer information, on the other hand, is routinely used by companies when researching new and existing market products, and this information forms the basis of many important company decisions regarding the launch of new products or the reformulation of existing lines. However, using sensory analysis and consumer information independently does not always enable the company to derive most benefit from available resources. By using these sensory and consumer techniques in conjunction, a more complete picture can be obtained. Preference mapping offers a group of techniques which can be used to relate these two groups of information.

Sensory assessors are sometimes required to give preference or acceptability information, but this is a dangerous practice as sensory assessors are unlikely to be representative of the target population, and by their training are more perceptive, in an analytical sense, than the average consumer. Likewise consumers are frequently asked to give reasons for their judgements or descriptors, but while these can provide some useful information, they need to be interpreted with care. Consumer descriptors are rarely detailed enough or reproducible, and can therefore lead to misleading results due to the difficulties encountered in the interpretation process. Preference mapping techniques offer the opportunity to use information generated from the best source in each instant. It is only the consumer who can realistically provide hedonic data, while a trained sensory panel is able to provide reliable descriptive information.

By relating these two data sets, one compliments the other, thus maximising the information available.

The size, number of samples required, and the apparent complexity of experiments designed for preference mapping are often given by companies as reasons for dismissing it without further consideration. However, these companies will often conduct numerous individual sensory and consumer tests comparing say a new formulation against the original, and perhaps a benchmark such as a competitors product. This approach, when compared with the preference mapping approach appears to be rather hit and miss.

When opting for the preference mapping approach, the experiment is carefully designed, often to consider several parameters at once. Accepting the fact that many product characteristics are interrelated, this approach will enable the product developer, not only to identify the need to adjust one component, but will indicate the effect other characteristics were having on liking. The approach also provides information on consumer segmentation, allowing the product developer to 'target' his product appropriately. While at the outset, the approach may appear to require excessive resource, it may in fact reduce the overall input by scientifically designing the ideal product.

A product resulting from a study using techniques such as preference mapping will be 'designed', and while the structured approach and combination of sensory analysis and consumer research cannot replace the creativity of the product developer, they can assist in identifying and summarizing market place opinions, thus helping the product developer to pin point the 'ideal' product. Therefore this product, given appropriate marketing, should achieve the competitive edge in the market place.

1.2 Use of the Technique

Preference mapping is used to answer a number of questions relating to improving the acceptability of the sensory aspects of a product. In one instance a company may simply wish to identify the attributes of a range of competitive products, which are important to acceptability, with the aim of moving their product into a more desirable position. Preference mapping projects involving market place products, may also enable the company to identify potential market opportunities through product gaps. In another scenario, the company may be working on a range of new product formulations, suggested by market research information, and wish to identify which formulation is most acceptable, and then if and how can the product be improved.

There are many ways in which a company will approach product optimization, and the preference mapping method is only one. The decision to choose this approach will depend on the importance of the project in terms of financial commitment and time, the status of sensory analysis in the company and of course the number and range of samples available for evaluation.

There should be no doubt, the preference mapping approach is expensive, involving both sensory and consumer panels. Equally the cost of product failure on market launch is expensive. Preference mapping as part of a well thought out product development exercise is well worth the expense. However, a decision on the relative feasibility of different options must be taken. While client projects and discussions remain confidential, the author has witnessed numerous examples where the preference mapping approach of utilizing both sensory and consumer information has led to improved formulations and more acceptable products.

2. PREFERENCE MAPPING AS A METHOD

2.1 Internal Preference Mapping

Before considering the method in some detail, it is first necessary to understand the terms metric and non-metric, as these terms are used a lot in relation to the method. A metric method is one where the data are assumed to be linear, or have interval properties. Such data are continuous (e.g., measurement of height). Non-metric methods are used to deal with non-linear data, and for the purposes of this chapter can be considered ordinal. In other words the data are whole numbers, but do have the property of representing an increase or decrease in intensity of a particular attribute.

The method of internal preference mapping (MDPREF) is similar to a principal component analysis (PCA) on a matrix of data, consisting of samples (objects) and consumers (variables). This analysis normally uses the covariance matrix (non-normalized matrix) rather than the correlation matrix. This means that a consumer with small or zero preferences, and consequently a low standard deviation, will not adversely affect the structure of the preference map. However, the correlation matrix is used by some packages, or the user has the option. In order that the geometry of the preference map is correct, it is necessary to normalize the principal components.

The result of internal preference mapping is a sample map, based on the product acceptability information provided by each consumer. A segmentation analysis of consumers is then possible by visually examining the plot of consumer preference directions, or by using a classification algorithm using the PCA parameters.

The more complex the structure of the population preferences, the greater the number of principal components that are required to be interpreted. However, the synthesis power of multidimensional analysis decreases with the number of axes to be interpreted. In fact, the non metric version of PCA allows the user to limit the number of axes to be determined. Non metric PCA involves calculating, for each consumer, the best monotonic transformation (Kruskal, 1964) of the preference data, in order to maximise the variance explained by the first k principal components of the transformed PCA data. It is common practice to choose only two or three preference axes, as after this the solutions become difficult to interpret.

It is an assumption of non metric PCA that the preference data are ordinal, not interval or ratio. When this is the case, it is important to exercise caution, as in effect only product ranks for each consumer are considered. However, the benefit of this is that the variance explained by the first k non metric PCA axes can then be taken as representing only the differences between product preference scores without the distortion to the ranks.

Whether metric or non metric, this form of principal component analysis is commonly referred to as Internal Preference Mapping or MDPREF, as first described by Carroll (1972).

Unlike external preference mapping, this method only uses the consumer data, and thus no information about why the samples are liked or disliked is given. It is possible to link sensory information to the internal preference mapping space, by correlating the mean sample ratings for each attribute with the derived preference dimensions.

2.2 External Preference Mapping

The basic idea behind external preference mapping (Schiffman et al., 1981) is to map acceptability data for each consumer onto an existing perceptual map of the products, usually obtained from profiling. In effect, the profile space is external to the acceptability data.

Preference mapping can simply be thought of as performing regression analysis on the data, where the dimensions of the profile space are the explanatory (or predictor) variables, while acceptability is the response (or dependent) variable (Schlich and McEwan, 1992; Schlich, 1995). It should be noted that the predictor space is, in fact, a decomposed space. This is because it is derived from a multivariate procedure such as principal component analysis or generalized Procrustes analysis, which decomposes the data into a smaller number of dimensions to adequately summarise the data.

In practice, there are two types of preference behaviour; that which fits a linear regression (vector model) and that which fits a quadratic regression (ideal point model). These are described below.

2.2.1 The Vector Model

The vector model pertains to 'the more, the better' type acceptance behaviour. Basically, this means that there is no sample which is perceived as having too much or too little of the characteristics which determine acceptability.

In practice, a multiple linear regression equation is derived for each consumer, and from this a vector depicting the direction of increasing preference can be drawn onto the sample space. The fitting of this model is often referred to as the Phase 4 (or Phase IV) model. This is represented in Figure 1.

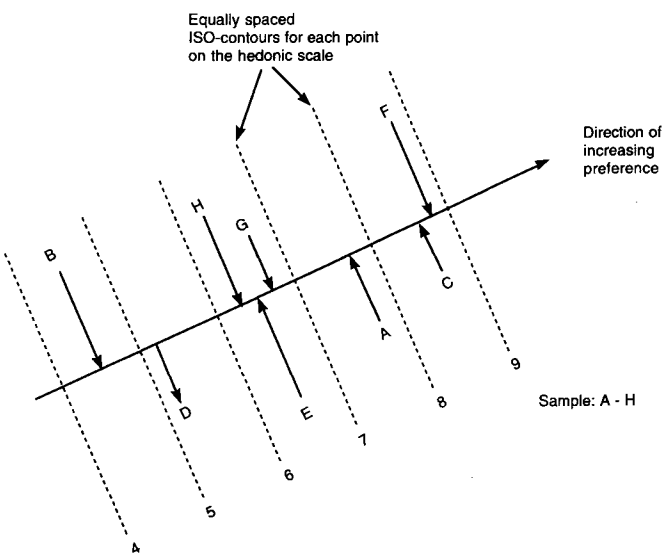


Figure 1. Graphical representation of the Phase 4 vector model.

2.2.2 The Ideal Point Model

The ideal point model pertains to the 'some amount is ideal' type acceptance behaviour. Basically, this means that there are samples in the space which are perceived as having excessive or insufficient amounts of one or more of the sensory attributes. Underlying the ideal point model is the assumption that there is some combination of attributes which make the ideal product. Whether it is realistic to assume that each consumer has only one ideal product is another matter.

There are three types of ideal point model which are often referred to; the circular ideal point model (Phase 3/Phase III), the elliptical ideal point model (Phase 2/Phase II) and the elliptical ideal point model with rotation (Phase 1/Phase I). The lower the phase number, the more complex (and less general) the model. Figures 2 (a) to (c) graphically display the format of these three ideal point models. All ideal points can be either positive or negative. A positive ideal point represents a point of maximum preference, whilst a negative ideal point is a point of anti-preference.

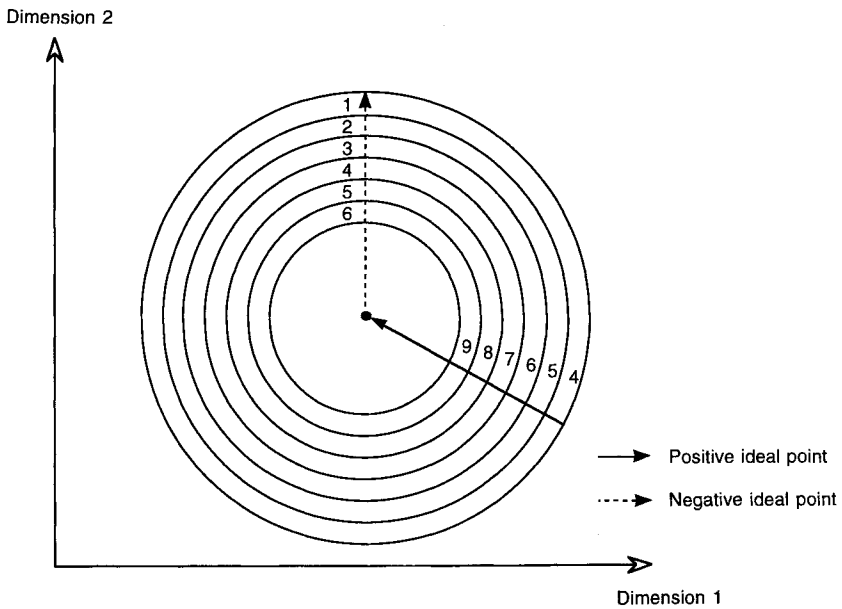


Figure 2a. Graphical representation of the Phase 3 ideal point model.

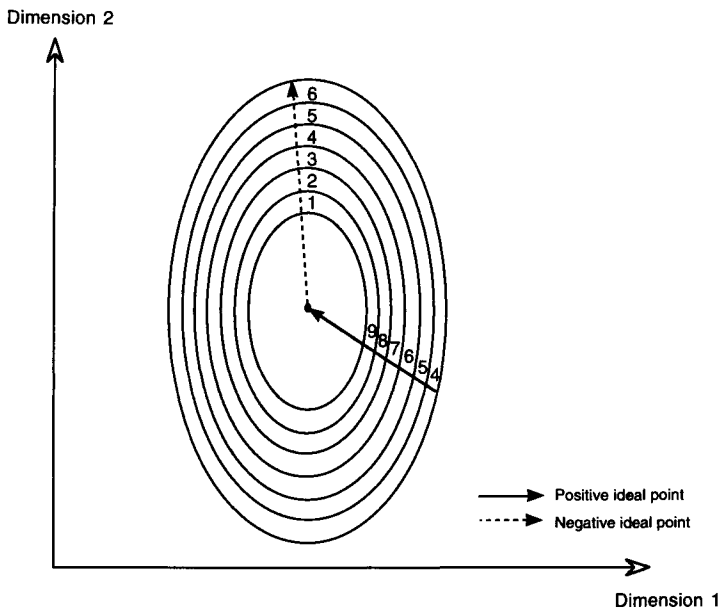


Figure 2b. Graphical representation of the Phase 2 ideal point model.

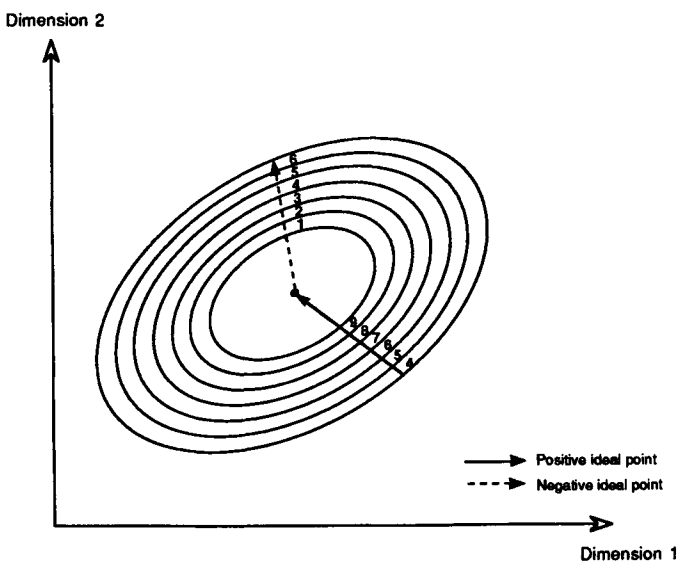


Figure 2c. Graphical representation of the Phase 1 ideal point model.

In practice Phase 1 (quadratic ideal point) tends to be ignored, as it assumes that the sample space is not 'optimal' and so rotates it to try to achieve a better fit with the acceptability data. As this is somewhat complex for interpretation it is seldom used, seldom recommended and therefore will not be discussed in terms of the worked example, presented in Section 6 of this chapter. Phase 2 (elliptical ideal point) is also quite complex as the ideal points can be both positive and negative at the same time, in other words a saddle point. In other words, one dimension optimises preference, while the other dimension optimises anti-preference. For this reason Phase 2 is little used in sensory evaluation, though it is useful if interpreted carefully. Phase 3 (circular ideal point) is the simplest of the ideal point models, and in effect fits a simple quadratic model to the data. In other words, there is an optimum (maximum or minimum) point on the space which is equally influenced by all the sensory dimensions being used to determine it. Thus, the contours round this optimum point are circular.

2.2.3 A Mathematical Explanation

The external preference mapping methods consist of calculating a polynomial regression for each consumer, by utilizing the sensory dimensions X_1, X_2, ..., X_k to explain the response (Y) of preference for each consumer. The X variables are called the independent or explanatory variables, while the Y is the dependent or response variable. Theoretically, the model can be written as shown in Equations 1 to 4. In each equation, α is the intercept term.

$$Y = \alpha + \Sigma_i \beta i Xi \qquad i = 1, ..., k \qquad \text{Eqn (1)}$$

$$Y = \alpha + \Sigma_i \beta i Xi + \delta \Sigma_i Xi^2 \qquad i = 1, ..., k \qquad \text{Eqn (2)}$$

$$Y = \alpha + \Sigma_i \beta i Xi + \Sigma \delta_i Xi^2 \qquad i = 1, ..., k \qquad \text{Eqn (3)}$$

$$Y = \alpha + \Sigma_i \beta i Xi + \Sigma \delta_i Xi^2 + \Sigma_{ij} \gamma ij Xi Xj \qquad i = 1, ..., k \qquad \text{Eqn (4)}$$

Equation 1 shows the vector model, which takes the format of a simple linear regression for each consumer. The βi are the slopes of the regression line (vector) for each consumer.

In Equation 2, a quadratic term, ΣXi^2, is added, where βi and δ are parameters to be estimated. The δ are all the same, as equal weight is attached to all dimensions in the circular model.

Equation 3 is the same as Equation 2, except that δi are different, according to the weighting attached to the dimensions. In other words, in the elliptical model, changes in preference along one dimension may be more important than changes along another dimension.

Equation 4 is Equation 3 with the addition of an interaction term. It is this term that provides the rotation in space of the preference map.

The designation of which phase to use is based on the way the models fit into one another. In fact, it can readily be seen that one model is actually a submodel of the other. Phase 1 (rotated model ideal point) is the most general model, and under this the models fit into each other, by becoming more specific. Therefore, Phase 4 (vector) is the most specific with the most constraints being placed on it. The type of data collected will determine the most appropriate model for a particular application. Fisher's test is normally used to establish which

model is best (Schiffman et al., 1981). The principle behind this is very similar to testing for, say, a sample effect in analysis of variance. In this case the analysis of variance is undertaken on the regression equation to determine how well the fitted model represents the data. By looking up the Fisher (F) tables (Neave, 1978), at the degrees of freedom for the regression equation against the degrees of freedom for the error term, it is possible to identify what significance level can be attached to the fitted model. It is then up to the user to decide the level of significance he is willing to accept.

For the vector model (Phase 4), the least squares estimators for α are used to help define the arrow representing the direction of the consumer's preference on the sensory map. The length of this arrow is proportional to the square of the multiple correlation coefficient R-squared, which indicates the goodness of fit of the model. Small arrows indicate that the consumer preferences cannot be explained by the sensory characteristics of the products, at least through the linear (vector) model. On the other hand, following a long arrow to infinity, it is possible to find the 'ideal' product of the consumer under investigation. The graphical interpretation of the consumers' preference arrows is similar to that of internal preference mapping. The interpretation of the actual sensory correlations is based on the previous multivariate analysis used to obtain the sensory map. In practice the user may choose that the preference vectors of consumers who fit Phase 4 (vector), are scaled to the unit circle round the sensory space. This is usually to make interpretation of consumer segmentation easier.

The other models, circular, elliptical and rotated elliptical, have response areas which are quadratic in shape. A quadratic can have either a maximum or minimum or a saddle point. A maximum represents the consumer's ideal point, whereas a minimum represents a negative ideal point. The interpretation of a saddle point is somewhat more difficult, as the preference signs are reversed, in other words one dimension is positive ideal while the other is negative ideal. In practice, only the maxima and minima are represented on the sensory map with a '+' or '-'.

On the preference plot of ideal points, the situation of the '+' and '-' signs and their density, provides an impression of the distribution of consumers' ideal points. In cases where the ideal point lies outside the sample space, the vector model may well be more appropriate.

The circular model (Phase 3) cannot result in a saddle point, and as its name suggests, the consumers' preference data are represented as circles surrounding the ideal point. With a positive ideal point, the circles direct towards the centre to a common point. However, with a negative ideal point consumer preference increase in any direction away from the centre of the circle.

The elliptical model (Phase 2) will result in a saddle point if the δi estimators have different signs. In this case there is usually no sensible interpretation in terms of identifying the consumer's ideal product. The interpretation of the elliptical model is similar to the circular model, but in this case the preference level lines elliptically surround the ideal point. The sensory attributes on the longest axis of the ellipse are of less importance than the sensory attributes of the shortest axis of the ellipse. Using a large number of consumers in the preference mapping exercise results in great difficulty in building the elliptical model to identify the common ideal point. This is because, consumers may be attaching different levels of importance to each preference dimension, and have different orders in terms of which preference dimension is most important to them. This makes the task of segmenting consumers more difficult, and for the lay user is a very complex and, perhaps, impossible task.

The quadratic (Phase 1), or complete model, takes into account the interactions between the sensory map's components: terms γijXiXj in Equation 4. This results in a rotated PCA, removing the original interpretation of the sensory map. For this reason, this model is not used for sensory analysis external preference mapping.

Although all four models are normally presented together in the literature, in practice only the vector (Phase 4) and circular (Phase 3) models are routinely used by sensory analysts.

2.3 Advantages and Disadvantages of External Preference Mapping

Preference mapping offers a very useful tool in the product optimization process, but perhaps its biggest drawback is that it is often misused due to lack of understanding. It is hoped that this chapter will go some way towards overcoming this problem. The purpose of this particular section is to consider some specific advantages of external preference mapping.

Advantages

- Offers a 'relatively' straight forward procedure for relating sensory and consumer information, for product optimization. Specifically where a preference mapping program has been purchased.
- Helps identify new markets.
- Provides direction for future product development.
- Provides information on market segmentation, with respect to sensory preferences. Can identify the need to make alternative types of product for different market segments.
- Using market samples, the technique can be a first step in looking at products currently available to the consumer, before developing specific formulations for a more detailed study.

Disadvantages

- A fairly large number of samples (e.g., 12-20) are often required to ensure that the preference mapping can be undertaken successfully.
- At present, every consumer must evaluate all the samples put forward to the consumer trial. This can be expensive.
- Can be complex to program the procedure, if the user has not bought a ready written preference mapping program.
- Preference data is not always directly related to the sensory profile map, as the way trained panels perceive products is different from consumers. However, note that lower dimensions of the profile map often relate better to preference than the commonly used first two dimensions.
- Tends to be used for understanding and direction, not prediction.
- Not all consumers well represented by the models.

2.4 Advantages and Disadvantages of Internal Preference Mapping

Internal preference mapping is useful to provide a sample map, based only on preference data. Therefore, it cannot be used to understand reasons for preference on its own. It has been known to use internal preference mapping to produce preference dimensions, and to use external preference mapping to map the sensory attributes onto the preference space. The

advantage of this is that the attributes are directly related to preference dimensions. However, in practice users may feel that this approach detracts from the benefits of linking a profile map directly back to the product processing and formulation parameters. A summary of the advantages and disadvantages on internal preference mapping are listed below.

- Easier to use and understand than external preference mapping, as similar to principal component analysis.
- Allows actual preference dimensions to be determined, as only acceptability data is used.
- Can be used as a screening procedure without sensory profiling, to develop samples worthy of further sensory and consumer work.

Disadvantages

- The program tends to break down after two dimensions in terms of interpretation.
- Percentage variance explained by the dimensions is often very low.

2.5 Software Availability

A number of software packages can be used to undertake preference mapping, though few readily allow the analysis without some work from the user. PC-MDS has a program for both internal and external preference mapping. In addition, the two sensory based statistical packages, SENPAK and SENSTAT also allow preference mapping to be undertaken. In each case running the analyses is fairly straightforward.

The major packages, such as SAS, Genstat, SPSS, S-Plus, SYSTAT and Minitab will allow all or most of the analyses options to be programmed. However, this requires the aid of an experienced statistician. In the event of this option being available, much more flexibility tends to be achieved over the output and graphical displays.

A comprehensive list of packages is provided in Appendix 1, together with the supplier.

3. PRACTICAL CONSIDERATIONS FOR SAMPLES

3.1 Sources of Samples

The sources of samples for a preference mapping study depends very much on the project brief and the objectives within this. If the objective is to characterise the product on the market of interest, e.g., all dairy milk block chocolate bars, then samples will normally be obtained from retail outlets. As with any sensory trial, it is usually advisable that samples within the same batch are purchased for use throughout both the sensory and consumer trials. This is because batch to batch variation may occur within samples. For example, in a crisp trial, the crisps may differ slightly in terms of bake level from batch to batch. Thus, if the consumer trial is undertaken on a different batch of samples from the sensory trial, then a 'true' relationship may not be found. If batch-to-batch variation is a problem, for example seasonal differences in potatoes, it may be important to build this into the design, by undertaking trials at different times of year.

Samples produced by the company themselves are used where the objective is to investigate different formulation and/or processing alternatives, to establish which combination of variables maximise the acceptability of the product. This may be as a follow on to an initial market evaluation as described above, or based on market research information. It can, of course, result from a change of supplier, a legal requirement to reduce an ingredient, or change to a different process. This of course assumes that a difference is wanted. In practice these situations usually require no product change, and therefore difference tests are used.

3.2 Design Considerations

Good experimental design is central to successful experiments, as discussed more fully in Chapter 2 of this book. However, it is important to take a look at what options tend to be available to industry, recognising that these may be improved through technology transfer and by illustrating the benefits that can be obtained.

The first and most important consideration, is establishing if a preference mapping study is to take place. This in turn determines the minimum number of samples that can be used. For example, a descriptive profile can take place on 4 samples, but not a preference mapping study. Recommending the minimum number of samples for preference mapping is fraught with difficulties, as the statistician or sensory scientist must often compromise between what is statistically ideal and what is practically possible within the time and financial constraints of the company undertaking or commissioning the work.

Taking the practical perspective as the starting point, and assuming the samples are well spread on the sample map, the user can sometimes get by with an absolute minimum of six samples for the vector model and seven to eight for the ideal point model (Phase 3). This, from a statistical point of view, allows a few spare degrees of freedom in the regression analysis. However, the point about a good spread of samples is important, as a sample space with one very different sample, and the rest close together, will not allow a good model to be fitted to the data.

Taking the problem, as viewed by the statistician, a larger number of samples than mentioned above is desirable. However, the larger the number the better, is not always true after a point, as no extra information will be gained. For example, if there are thrirty-two possible treatment combinations in a factorially designed experiment, as much information is likely to be obtained on a half replicate of sixteen samples. The number of samples necessary will often depend on whether the samples are produced according to an experimental design and its format, or selected from those in the market. Both cases will now be considered.

Studies with market place samples are often the most difficult, as it is impossible to know in advance whether they will have a good spread on the sample map. Where a wide range of samples is available, it may be necessary to look at twenty to thirty samples on the sensory profile, and then select a representative range of twelve to sixteen samples for the consumer trial. In this case, when undertaking the preference mapping, the sample coordinates to be used will simply be chosen from the results of the multivariate analysis on all the samples used in the profile. It is important not to re-run the multivariate analysis on sensory profile data, using the reduced number of samples before preference mapping, as this will change the sensory map definition. However, the number of samples selected should still be sufficient to allow the preference mapping to be undertaken, as discussed earlier. If the market is small, and only a

few samples exist, then perhaps preliminary acceptability and sensory work can be used as an exploratory tool to setting up a designed experiment, as discussed below.

Preference mapping studies can be carried out on samples which are formulated to a carefully thought out design. For example, while under-used, factorially design experiments are a very efficient way of product development, providing preliminary work has been undertaken to establish the key factors important to preference, and pilot work has been undertaken to establish realistic levels of each of the factors. A simple factorial experiment may produce eight samples, by using three ingredients (A, B, C), each with two levels (1, 2): A1B1C1, A1B1C2, A1B2C1, A1B2C2, A2B1C1, A2B1C2, A2B2C1, A2B2C2. Four or five ingredients may be used, or three levels, thus making more samples. In such cases the concept of fractional factorial experimental designs can be used to reduce the number of samples to a manageable number. With well designed experiments eight samples may well prove to be adequate for preference mapping studies. This approach also has the added advantage that other statistical tools (e.g., factorial analysis of variance and response surface methodology) can be used to extract detailed information as to the best combination of ingredients to use.

A final point to note is that each consumer must taste all the samples put forward to the consumer trial. As mentioned previously, this could be a subset from the sensory trial, or all the samples used in the sensory trial. Work is ongoing to determine whether incomplete designs can be used, as mentioned in Section 2.4. From a sensory point of view, the logistics of tasting a large number of samples in a consumer trial must be considered. Often consumers must be recalled to attend several tasting sessions, or be pre-recruited to attend a half day or whole day tasting, with suitable breaks to prevent sensory fatigue.

3.3 Sensory Methodology

There are a two main sensory techniques which are used to provide a perceptual map of samples: dissimilarity scaling (Schiffman *et al.*, 1981) and descriptive profiling (Stone and Sidel, 1985).

Dissimilarity scaling provides a perceptual map of the samples using the statistical tool of multidimensional scaling. However, it provides no descriptive information about why the samples are different. For this reason, and other practical reasons, descriptive profiling is more widely used both as a tool in its own right as well as for preference mapping studies.

Profiling data is normally analysed by principal component analysis, generalized Procrustes analysis, factor analysis or correspondence analysis. In each case, sample coordinates are produced to position the samples on the map. These coordinates are the input to external preference mapping to define the sample space onto which consumer preference is mapped.

3.4 Consumer Methodology

The key point about collecting the consumer data is that each consumer should evaluate every sample selected for the trial, as discussed in Section 3.1. Acceptability data is normally collected on a 5, 7 or 9 point hedonic category scale (Peryam and Pilgrim, 1957), or on a suitable anchored continuous line scale. In many instance, separate measurement may be made for appearance, flavour, texture (mouthfeel) and overall acceptability, in order that each aspect of the product can be considered in detail. The format of the data for input to external or internal preference mapping is matrix, where the consumers are the rows and the samples the columns.

4. INTERPRETATION AND PRESENTATION OF RESULTS: PREFMAP

The purpose of this section is to illustrate how results from external preference mapping can be interpreted and presented. It utilizes the output format from the PC-MDS package, but can be easily applicable to most other programs. In addition, it was decided to concentrate only on the Phase 3 (circular ideal point) and Phase 4 (vector) models, as Phases 1 (quadratic ideal point) and 2 (elliptical ideal point) are seldom applicable to sensory applications.

4.1 Information from the Analysis

Tables 1 and 2 list the type of information provided by an external preference mapping analysis, and what each refers to, and how useful it is in practice. This is based on specifying Phase 4 (vector) and Phases 3 (ideal), respectively.

Table 1
Comments on phase 4 (vector) output.

Output Description	**Comments**
Original configuration (X-matrix)	These are the scores of the sample space. They should remain the same throughout the analysis.
Vector of scale values (preferences)	These are the normalized preference data, with sum zero and sum of squares equal to 1, as the preference scores are centered by subtracting the preference data from the mean and dividing by the standard deviation.
Dimension cosines of fitted subjects	These are the coordinates representing the point at which to draw the vector from the origin. They are coordinates to enable preference directions to be drawn.

Table 2 Comments on Phase 3 (ideal point) output.

Output Description	**Comment**
Original configuration (X-matrix)	These are the scores of the sample space. They should remain the same throughout the analysis
Vector of scale values (preferences)	These are the normalized preference data, with values (preferences) sum zero and sum of squares equal to 1.
Coordinates of ideal points-with repect to old axes (coordinates of ideal points)	These are the values (coordinates) for plotting the ideal point position on each of the dimensions specified in the analysis.
	These are the weightings for each individual on the axes

Importance of new axes (weights of axes)	(preference dimensions). The weighting on each axis is the same within an individual. The higher the weighting, the greater the profile map structure is in accounting for that individual's preference. A negative set of weights implies a negative ideal point (anti-preference), while a positive set of weights implies a positive ideal point.

At the end of the output the user will find the correlation and F-ratio for the model for each individual and for each Phase. For interpreting the correlation (R) for the regression model, the same principles are used as when using regression methods for other applications. The R can take a value from 0 to 1, with the closer to 1 the correlation is, the better the fit of the model. Correlation tables can be used to establish the significance of the relationship between the preference data and sensory dimensions. In practice many users choose a correlation of, say, 0.6 based on past experience with the method, as representing at least a 5% significant level for sample sizes of greater than 10, or a 10% significance level for sample sizes of greater than 8. This is fine when the method is being used as an exploratory tool to suggest future directions with some confidence, rather than a predictive tool in the mathematical sense. However, it is suggested that individual models with correlations of less than 0.5 are unacceptable to use in the interpretation process. This is because the confidence in such data, even for exploratory interpretation would be low. Consumers not satisfying the required goodness of fit level are removed from the analysis, and not considered further. A lot of information is lost if many consumers fall into this category.

The statistical significance of the regression model can be tested using the F-ratio provided by the analysis, and comparing it to the Fisher tables (e.g., Neave, 1978). This is just another way of measuring how well each consumer's data fits the model used by the preference mapping program. The degrees of freedom to use are given on the output, and therefore can be used to find the critical value for comparison. If the critical value is less than the F-ratio in the output, then the model is well fitted at the level of significance tested. It is usual to use a 5% significance level, but this is often too severe for the purpose of the work being undertaken, many consumers being ignored at this cut-off level.

In addition, the analysis provides a between phase F-ratio, which allows the user to decide if moving to a more complex model offers a better fitting model for a particular individual. In other words the between phase F-ratio helps establish whether a quadratic term should be present in the model. This information is seldom used in practice, as the interpretation process tends to eliminate individuals who could have been represented by the more complex model, as determined by the F-ratio mentioned in the above paragraph. It is important, for segmentation and interpretation purposes, that as many 'good' subjects as possible are represented in each phase, to allow meaningful conclusions to be drawn.

4.2 Presentation of Results

In practice, presentation of the results tends to centre round the plots, both vector and ideal point, as appropriate. In the case of the vector model, the sample plot is produced, with individual directions of preference represented as vectors on the sample map. An example, illustrating this, is provided in Section 6.4.1. If all individuals have preference vectors in the same, or similar, direction, then there is a clear preference for samples with attributes in that

area of the sample map. However, it is more usual to have individuals with different preferences, though often a clear direction for product development will be apparent.

It is also usual, for completeness, to indicate the number of consumers who were actually used for interpretation. This allows the reader (or client) to appreciate better the significance of his results. For example, if only 25% of the consumers fitted the vector model, then clearly a lot of information has been lost, unless these individuals are represented in the ideal point (Phase 3) model.

For the ideal point model, the sample plot is produced with points on the space representing either positive or negative ideal points. In most cases, the negative ideal points are not taken into account, other than to illustrate areas of the space representing samples whose combination of attributes is disliked. As mentioned earlier, a positive ideal point represents a point of maximum preference (or local maxima) on the profile map, whilst a negative ideal point is an anti-preference point (or local minima). Again clear explanation can be found by following the example in Section 6.

4.3 Pitfalls and Misinterpretations

There are many potential pitfalls to the unwary user of external preference mapping. It is tempting to include all consumers in the final presentation of the results, thus providing conclusions which may be misleading. In many instances the Phase 3 or 4 models may just not be appropriate, either because there are too few samples for the analysis, or because the data do not fit the models used, as indicated by the poor goodness of fit measures discussed in Section 4.1. It is important to identify this.

Another pitfall for external preference mapping is the number of dimensions to put into the analysis. All the comments to date have considered the case of a two dimensional sensory map. Clearly three or four dimensions are possible, sometimes more. It is important to remember that increasing the number of dimensions in PREFMAP leads to a requirement to increase the number of samples in the analysis. If the user has only six samples and is running a vector model, only two dimensions can be used.

Experience of using three and four dimensions in PREFMAP has suggested that the interpretation becomes confusing and somewhat unreliable. However, where there is clearly a three or four dimensional sensory map, it would be very wrong to only examine the first two dimensions. This is because the attributes contributing to preference may be best explained in the lower dimensions (e.g., Dimension 4). In the author's experience, it is wise to look at different sets of two dimensions in the PREFMAP analysis. In this way the best solution can be obtained. However, this does involve a lot more work on the part of the user.

There are less pitfalls to using internal preference mapping, but nonetheless certain points should be noted. Firstly, it is often the case that very little of the total variation is explained in the first two dimensions (e.g., 30-40%). However, the MDPREF solution often becomes unstable when more dimensions are used. In addition, not every consumer contributes the same amount of information to the sample map. If the common option of scaling the consumer vectors to unit variance is used, then the apparent segmentation represented on the preference map may be misleading, as could the position of the samples. This is because not all consumers will contribute to the preference map, as will be illustrated in Section 6.

4.4 The Extent of the Conclusions

The extent to which conclusions can be drawn from external preference mapping will depend very much on the number of samples, goodness of fit of the models used, etc. In fact, consideration needs to be given to all the factors previously discussed. Generally speaking, it should be recognised that the preference mapping approach, in a well thought out experiment, will provide excellent direction for future product development. However, it would be unrealistic to expect it to pinpoint the precise level of each product ingredient or process combination to achieve the perfect product from a sensory point of view. Nonetheless from previous work with industry, major product improvements have been achieved. It is also worth noting, that in factorially designed experiments, further value can be added to preference mapping by undertaking response surface analysis.

Preference mapping can, and is used, to identify the major segments of the target market, and can be used to make products for different segments. In saying this, it is important to recognise that a manufacturer cannot go to extremes and make a product for every consumer. Therefore, for a particular segment, the best compromise product can be identified on the basis of preference mapping.

5. INTERPRETATION AND PRESENTATION OF RESULTS: MDPREF

As previously mentioned, internal preference mapping is a form of principal component analysis, but with the option to pre-treat the data in a number of ways, and/or to scale the resultant scores and loadings. In running a MDPREF, there are generally four possible options. There are two possible data pre-treatments, and the choice of whether or not to normalize each consumer's preference vector to fit a unit circle.

Pre-treatment of the data will be considered first. Both methods are forms of centering each consumers preference data. The standard option is to pre-treat the preference data of each individual, by subtracting the mean preference rating from the original sample preference rating. This in effect acts as a translation of the scale used and relocates the data round the centre point, the mean (average) value. The second method uses the relocated data as described, and then divides the sample preference ratings by the standard deviation of the original sample scores. In this way all samples have a standard deviation of 1, which in practice may distort the acceptability data if there are samples with greatly different standard deviations.

Normalization is the second option, and usually the normalization option is chosen to enable the geometry of the preference map to be correct. However, some MDPREF programs allow the user to choose not to normalize. This is not recommended, as the map produced may be misleading.

One option which has not been considered here is whether to scale the preference vectors to unit variance. Some programs do this automatically, while others give the option of allowing the vector lengths to represent the variance contributed by the consumer to the preference map.

In terms of use of internal preference mapping with sensory attributes, a common procedure is to correlate the mean sensory attribute scores with the each preference dimension. The results are simply plotted as vectors on the internal preference map, and this will be illustrated in Section 6.

6. CASE STUDY: ORANGE DRINKS

6.1 Introduction

The example, used to illustrate Phase 3 and Phase 4 of external preference mapping and internal preference mapping, is based on work undertaken on a selection of orange drinks to investigate effect of citric acid and sweetener on product acceptability.

The background to the project was to investigate labelling of diet and regular drinks, and whether this affected the acceptability of the product. The work presented here, to illustrate preference mapping, will concentrate only on blind assessment of the products, which was the first step in the exercise.

6.2 Selection of Samples for Profiling

The samples were selected with the objective of the experiment in mind. Orange drinks of this type are sweetened with sucrose and/or aspartame. Thus, sweetener type was a factor in the experiment with three combinations being chosen. These were sucrose only, aspartame only and a 25% sucrose/75% aspartame mix. The levels of sweetener chosen were based on existing work undertaken at the School of Psychology, University of Birmingham (Booth and Freeman, 1992), using this type of orange drink. However, a preliminary experiment was undertaken to ensure equi-sweetness of the solutions.

A second factor, acidity at three levels, was included in the experiment to investigate the interaction between sweetener and acidulent. In addition, Carboxymethyl-cellulose (CMC) was used with the aspartame samples, to mask differences in viscosity.

The final design of fifteen samples (Tables 3 and 4) was chosen to provide a good range of samples for profiling, as well as to minimize the inter-correlations between the samples from a psychological point of view (Booth and Freeman1992). The rationale is that, for any experiment that seeks to distinguish main effects of two (or more) independent variables, the aim is to minimise the inter-correlations between the levels of the factors in the samples. In this way, for example, problems of multicollinearity in regression are avoided or reduced. In this particular example, if the rank correlation between the sweetener levels and acid levels are calculated from Table 3/4, a value of near zero is obtained This is because the filled cells in Table 3 are 'square', that is high and low levels of one factor are equally represented at high and at low levels of the other.

As this design is somewhat different from the traditional approach of undertaking a factorial design or mixture model design, a few comments on the background will be made. The primary purpose of the orange drink experiment was to compare a psychological approach to product optimization to the traditional sensory approach. In addition, it was important to produce realistic mixtures for assessment, and again this was based on previous unpublished work at Birmingham University. The actual design is based on the recognition that the psychological approach to the analysis is based on an unfolding procedure as described by Conner (1994). Clearly, from a purely sensory point of view, this type of deign runs the risk of not obtaining sufficient good information on the sensory interactions. Ideally, for the traditional

sensory approach a proper factorial or mixture model would have been used. However, this particular example does illustrate well the preference mapping methods.

Table 3
Design used for sample selection. Bold letters - aspartame only, underlined letters - sucrose only.

			Levels Citric Acid		
			Low	Medium	High
	High	7	-	-	I
		6	**E**	D	-
Levels of		5	B	**G**	M
sweetener		4	L	A	**H**
		3	N	O	**K**
		2	**F**	C	-
	Low	1	-	-	**J**

Table 4
Samples used for profiling exercise, where all ingredient quantities were measured per litre of water.

Sample	Orange Powder	Citric Acid (g)	Sucrose (g)	Aspartame (g)	CMC (ml)
A	4.69	6.11	74.92	0.00	115
B	4.69	3.05	104.89	0.00	95
C	4.69	6.11	38.30	0.00	140
D	4.69	6.11	36.76	0.94	140
E	4.69	3.05	0.00	1.52	165
F	4.69	3.05	0.00	0.21	165
G	4.69	6.11	0.00	1.02	165
H	4.69	12.22	0.00	0.55	165
I	4.69	12.22	206.00	0.00	30
J	4.69	12.22	6.83	0.10	160
K	4.69	12.22	0.00	0.33	165
L	4.69	3.05	18.73	0.37	150
M	4.69	12.22	26.22	0.56	145
N	4.69	3.05	53.62	0.00	130
O	4.69	6.11	13.40	0.26	155

Profiling was carried out in the normal way (Lyon et al., 1990), by generating and agreeing on a list of thirteen flavour and mouthfeel terms, all of which were used significantly ($p < 0.001$) to discriminate between the samples. Figures 3 and 4 show the sample and attribute plots derived from generalized Procrustes analysis (GPA). The triangles in Figure 3 represent

the three replicate positions for each sample. GPA was used in order to produce a consensus map, which took into account differences in scale use by the sensory assessors (Arnold and Williams, 1986). The GPA sample map was used as input to the preference mapping in this particular example, but remembering that the attribute plot tells the user why the samples are placed in a particular way on the sample map.

6.3 Selection of Samples

It is often impractical to evaluate a large numbers of samples in a consumer trial, due to time and cost considerations. As mentioned previously, preference mapping works when each consumer has evaluated all samples under investigation.

In this example, it was decided to select eight of the original fifteen samples for the preference mapping study. This was considered the absolute minimum to illustrate the Phase 3 model. Ideally, however, all fifteen samples should have gone through.

Samples were selected to represent the range on the sample map (Figure 3), and these were A, C, D, F, H, I, J, L. Acceptability data were collected using sixty-two staff at Campden, not involved in sensory analysis. Samples were evaluated over two sessions, using a nine point hedonic scale (Table 5) to measure overall acceptability.

Table 5
Nine point hedonic scale used to measure product acceptability (Peryam and Pilgram, 1957).

	Descriptor
9	Like extremely
8	Like very much
7	Like moderately
6	Like slightly
5	Neither like nor dislike
4	Dislike slightly
3	Dislike moderately
2	Dislike very much
1	Dislike extremely

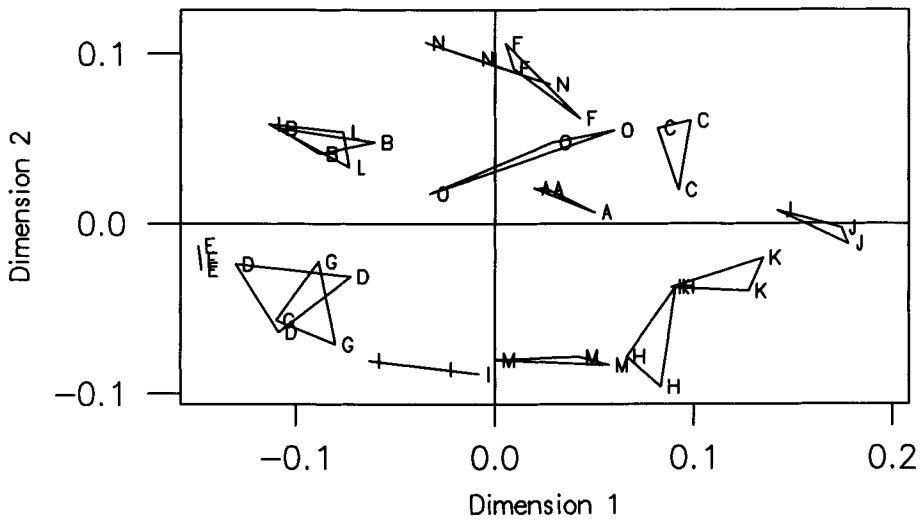


Figure 3. Sample map derived from generalized Procrustes analysis on a conventional profile of fifteen orange drink samples.

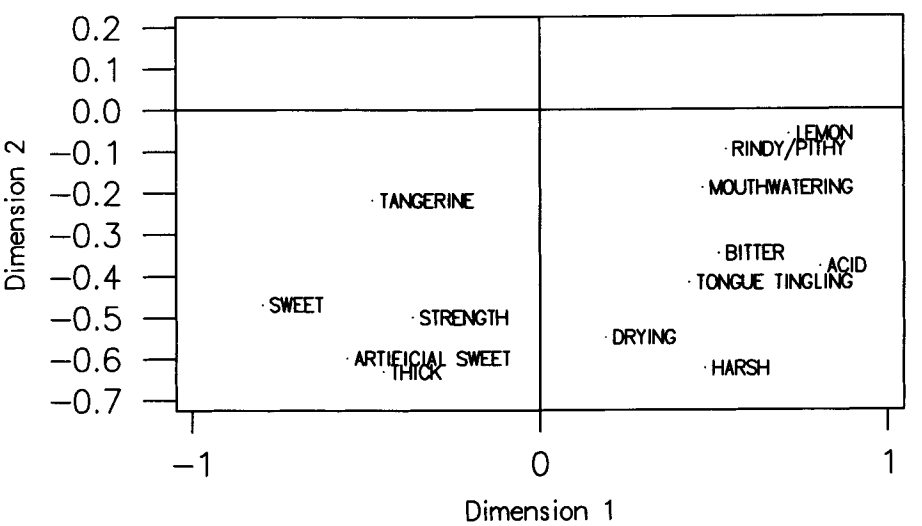


Figure 4. Attribute map derived from generalized Procrustes analysis on a conventional profile of fifteen orange drink samples.

6.4 External Preference Mapping

External preference mapping was run specifying Phases 3 and 4, the ideal point model and the vector model. The sample scores, averaged across replicates, from GPA are shown in Appendix 2, and the acceptability data in Appendix 3.

On running this analysis, the first part of the output examined was the root mean square correlation values for both phases. These take values from 0 to 1, where 1 represents a perfect fit, and 0 represents no fit at all. This measure is an overall measure which takes into account the correlation fits of each assessor's model. For this analysis, root mean square values were 0.758 and 0.693 for Phases 3 and 4, respectively. Both values being over 0.500 are acceptable, and in addition, there was no real improvement in using the ideal point model over the vector model. As is often the case, both models were examined. The choice of 0.5 as the cut-off point is somewhat arbitrary, and tends to be based on number of samples. As the root mean square correlation is analogous to correlation coefficients, the rules for significance apply. At 20% significance a value of 0.507 is required, while for 5% significance a value of 0.707 is needed.

Before considering both models, the correlations for each individual consumer are usually examined, to obtain an idea as to the number of consumers likely to be included in the final plot for interpretation. These values are provided in Table 6.

6.4.1 Vector Model

The first step was to determine which consumers fitted the model. This can be achieved by looking at the correlations at the end of the output (Table 6). In this instance there were eight samples, which required a correlation value of 0.707 for 5% significance or 0.621 for 10% significance. On this occasion, a 10% significance level selection was used. In this way thirty-three of the sixty-two consumer were included for plotting the graphical representation of the results. If a correlation at 20% significance was used (0.507), then an additional twelve consumers would have been included in the final analysis. In many cases this provides further useful information, but at the expense of reducing confidence in a statistical sense.

Table 6
Consumer (subject) correlations for Phases 3 and 4.

Consumer	Phase 3	Phase 4	Consumer	Phase 3	Phase 4
1	0.917	0.914	33	0.706	0.631
2	0.634	0.436	34	0.711	0.696
3	0.984	0.950	35	0.683	0.668
4	0.720	0.683	36	0.616	0.591
5	0.628	0.477	37	0.692	0.462
6	0.708	0.601	38	0.607	0.593
7	0.885	0.870	39	0.533	0.389
8	0.914	0.844	40	0.656	0.599
9	0.890	0.820	41	0.559	0.509
10	0.703	0.700	42	0.824	0.651
11	0.457	0.084	43	0.797	0.757
12	0.754	0.565	44	0.673	0.648
13	0.918	0.904	45	0.780	0.777
14	0.709	0.491	46	0.530	0.281
15	0.904	0.873	47	0.893	0.884
16	0.824	0.823	48	0.817	0.680
17	0.649	0.488	49	0.406	0.402
18	0.841	0.829	50	0.804	0.556
19	0.810	0.809	51	0.849	0.849
20	0.545	0.539	52	0.849	0.836
21	0.530	0.523	53	0.874	0.839
22	0.780	0.780	54	0.979	0.948
23	0.672	0.394	55	0.778	0.763
24	0.625	0.526	56	0.842	0.804
25	0.799	0.421	57	0.503	0.392
26	0.794	0.767	58	0.814	0.801
27	0.726	0.418	59	0.698	0.694
28	0.779	0.772	60	0.473	0.473
29	0.866	0.836	61	0.923	0.784
30	0.730	0.327	62	0.549	0.463
31	0.610	0.595			
32	0.723	0.567	Average	0.906	0.867

The next step is to create a plot of the samples using the coordinates of Appendix 2, and then draw on the vectors using the coordinates shown in Table 7. These are normally identified as 'dimension cosines of fitted subject vectors'.

Table 7
Coordinates for the vector model, for each consumer.

Consumer	Dim 1	Dim 2	Consumer	Dim 1	Dim 2
1	-0.765	0.644	32	-0.795	-0.607
2	-0.955	0.297	33	-0.641	0.768
3	-0.958	-0.288	34	-0.723	0.691
4	-0.663	0.749	35	-0.148	0.989
5	-0.907	-0.421	36	-0.168	0.986
6	-0.468	0.884	37	-0.061	0.998
7	-0.162	0.987	38	-0.288	0.958
8	-0.559	0.829	39	0.508	0.862
9	-0.531	-0.847	40	-0.426	0.905
10	-0.485	0.874	41	-0.751	-0.660
11	0.709	-0.705	42	-0.950	0.311
12	-0.849	-0.529	43	-0.301	0.934
13	-0.439	-0.899	44	-0.999	0.042
14	-0.264	-0.965	45	-0.952	-0.307
15	-0.966	-0.259	46	-0.950	0.312
16	-0.489	-0.872	47	-0.961	0.278
17	-0.676	-0.737	48	-0.346	0.938
18	-0.638	0.770	49	-0.558	0.830
19	-0.965	0.262	50	0.847	-0.532
20	0.058	-0.998	51	-0.719	0.695
21	-0.636	0.772	52	-0.925	0.381
22	-0.998	-0.064	53	-0.617	0.787
23	-0.183	0.983	54	-0.991	-0.133
24	-0.595	0.804	55	-0.456	0.890
25	-0.552	0.834	56	-0.966	0.260
26	0.998	-0.064	57	0.587	-0.810
27	-0.676	0.737	58	-0.826	0.564
28	-0.467	0.885	59	-0.811	0.585
29	-0.787	-0.617	60	-0.998	0.070
30	0.132	0.991	61	-0.266	0.964
31	-0.852	0.524	62	-0.014	0.999

The vector model (Phase 4) plot is shown in Figure 5, where a clear direction of preference towards the left of Dimension 1 can be observed. In fact, Samples D and L are most acceptable overall. By examining Figure 4, the attribute plot, it can be observed that the most acceptable samples are sweet, tangerine, thick and are high in strength of flavour. It is also interesting to observe in this example, that the profile sample map and the internal preference map are very similar. This is not always the case.

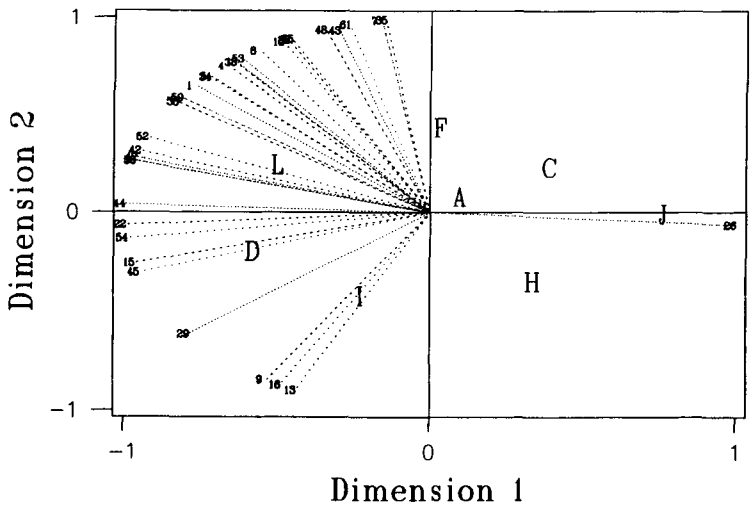


Figure 5. The Phase 4 vector map derived from undertaking external preference mappingusing the sensory profile map in Figure 3.

6.4.2 Ideal Point Model

Again, the first step was to determine which consumers fitted the model satisfactorily by looking at the correlations (Table 6). Taking a 10% significance level, fifty of the sixty-two consumers were included for plotting the graphical representation of the results. If a correlation at 20% significance was used (0.507), then an additional eight consumers would have been included in the final analysis. At this point, another two steps are required, firstly to identify positive and negative ideal points, and then to establish which of these actually lie within the sample space.

To determine whether an ideal point is positive or negative, the user should look at the 'importance of new axes' part of the output (see Table 2 for explanation). This information is provided in Table 8, together with whether the consumer had a positive or negative ideal point. There were forty-two positive and twenty negative ideal points in total. An asterisk by the consumer identity number in Table 8 highlights the fifty consumers who are included for further analysis at the first step.

The coordinates for the ideal points (Table 9) now need to be examined to determine which are to be plotted. A decision is normally taken to plot and interpret only those ideal points within the sample space. This is because ideal points outside the space are really better described by the vector model, as they are ideal points tending towards infinity. In Table 9 a single asterisk represents consumers fitting the model, whilst a double asterisk represents those whose ideal points also fall within the sample space. There were forty consumers to take forward to the plotting stage, of which twenty-eight had positive ideal points and twelve had negative ideal points.

Table 8
Positive and negative ideal points and their coordinates.

Consumer	Importance of New Axes	Sign	Consumer	Importance of New Axes	Sign
1*	(-4.6, -4.6)	-	32*	(26.9, 26.9)	+
2*	(27.6, 27.6)	+	33*	(-19.0,-19.0)	-
3*	(15.4, 15.4)	+	34*	(-8.7, -8.7)	-
4*	(13.8, 13.8)	+	35*	(8.5, 8.5)	+
5*	(24.4, 24.4)	+	36*	(10.5, 10.5)	+
6*	(22.4, 22.4)	+	37*	(30.8, 30.9)	+
7*	(9.7, 9.7)	+	38	(78, 7.8)	+
8*	21.0, 21.0)	+	39	(21.8 21.8)	+
9*	(20.7, 20.7)	+	40*	(16.0, 16.0)	+
10*	(4.0, 4.0)	+	41	(13.9, 13.9)	+
11*	(26.9, 26.9)	+	42*	(30.2, 30.2)	+
12*	(29.9, 29.9)	+	43*	(14.8, 14.8)	+
13*	(9.6, 9.6)	-	44*	(-10.9,-10.9)	-
14*	(30.6, 30.6)	+	45*	(4.1, 4.1)	+
15*	(-14.1,-14.1)	-	46	(26.9, 26.9)	+
16*	(27, 2.7)	+	47*	(-7.5, -7.5)	-
17*	(-25.6,-25.6)	-	48*	(-27.1,-27.1)	-
18*	(8.8, 8.8)	+	49	(-3.5,-3.5)	-
19*	(3.0, 3.0)	+	50*	(34.8, 34.8)	+
20	(4.7, 4.7)	+	51*	(0.8, 0.8)	+
21	(-5.2, -5.2)	-	52*	(-8.7, -8.7)	-
22*	(-0.8, -0.8)	-	53*	(14.7, 14.7)	+
23*	(32.6, 32.6)	+	54*	(-14.5,-14.5)	-
24*	(-20.3,-20.3)	-	55*	(-9.1, -9.1)	-
25*	(40.7, 40.7)	+	56*	(15.0, 15.0)	+
26*	(-12.2,-12.2)	-	57	(18.9, 18.9)	+
27*	(-35.6,-35.6)	-	58*	(8.7, 8.7)	+
28*	(6.4, 6.4)	+	59*	(-4.7, -4.7)	-
29*	(13.3, 13.5)	+	60	(-0.7, -0.7)	-
30*	(39.1, 39.1)	+	61*	(29.2, 29.2)	+
31	(8.1, 8.1)	+	62	(7.8, 7.8)	+

Figure 6 shows the ideal point model plot, from which it is evident that the majority of positive ideal points are to the left of Sample A on Dimension 1. This, therefore, suggests a similar result to the vector model map. This picture suggest that the attributes on this side of the plot are acceptable, but also that the shouldn't be quite as strong as perceived in Samples L and D.

Table 9
Coordinates of the ideal points, for each consumer.

Consumer		Dim 1	Dim 2	Consumer		Dim 1	Dim 2
1*		0.390	-0.321	32**	+	-0.018	-0.038
2**	+	-0.012	0.000	33**	-	0.075	-0.076
3**	+	-0.102	-0.047	34**	-	0.163	-0.147
4**	+	-0.065	0.086	35**	+	-0.016	0.232
5**	+	-0.018	-0.027	36**	+	-0.010	0.164
6**	+	-0.016	0.058	37**	+	0.017	0.037
7**	+	-0.025	0.266	38		-0.047	0.211
8**	+	-0.041	0.081	39		0.043	0.030
9**	+	-0.034	-0.096	40**	+	-0.027	0.088
10*		-0.219	0.420	41		-0.043	-0.065
11		0.025	-0.015	42**	+	-0.024	0.005
12**	+	-0.015	-0.032	43**	+	-0.027	0.138
13**	-	0.131	0.218	44**	-	0.142	-0.015
14**	+	0.008	-0.055	45*		-0.356	-0.131
15**	-	0.143	0.023	46		-0.001	-0.003
16*		-0.369	-0.703	47*		0.260	-0.079
17**	-	0.050	0.024	48**	-	-0.005	-0.076
18**	+	-0.138	0.181	49		-0.144	-0.253
19*		-0.539	0.142	50**	+	0.051	-0.029
20		0.041	-0.369	51*		-1.916	1.860
21		0.186	-0.212	52*		0.213	-0.089
22*		1.958	0.115	53**	+	-0.073	0.109
23**	+	0.013	0.027	54**	-	0.153	0.008
24**	-	0.061	-0.066	55**	-	0.129	-0.224
25**	+	0.004	0.013	56**	+	-0.090	0.020
26**	-	-0.109	-0.002	57		0.053	-0.055
27**	-	0.040	-0.031	58**	+	-0.157	0.111
28**	+	-0.141	0.295	59*		0.300	-0.211
29**	+	-0.090	-0.096	60		1.480	-0.112
30**	+	0.023	0.015	61**	+	-0.002	0.069
31		-0.125	0.079	62		0.019	0.072

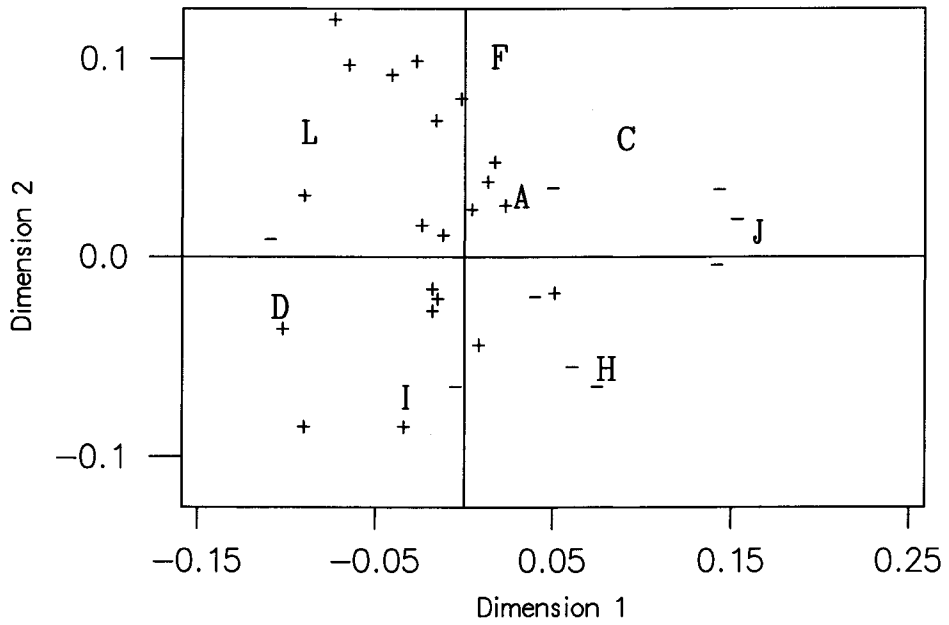


Figure 6. The Phase 3 ideal point map derived from undertaking external preference mapping using the sensory profile map in Figure 3.

6.5 Internal Preference Mapping

6.5.1 MDPREF: First Example

The data were input to the MDPREF program of PC-MDS, in the format shown in Appendix 3, where consumers are rows and samples are columns. In this run of the analysis, it was decided to pre-treat each consumer's acceptability data by subtracting the row mean and dividing by the standard deviation, and then to scale each consumer's preference vector to unit variance, that is so they fit on a unit circle. The correlation matrix was used, as the analysis was run using PC-MDS. It should be remembered that in doing this pre-treatment, some consumers may be given more weight than perhaps they should.

The resulting plot is shown in Figure 7, which shows the samples as letters, and the consumers as directions of increasing preference. All vectors are scaled to fit a unit circle. This example is particularly good, as most consumers prefer samples on the left hand side of Dimension 1, with Sample L being the most acceptable overall.

6.5.2 MDPREF: Second Example

The data were input to a MDPREF program in Genstat, in the format shown in Appendix 3. In this run of the analysis, the data were pre-treated as before by subtracting the row mean and

dividing by the standard deviation. However, in this case the preference vectors were not scaled.

The resulting plot is shown in Figure 8, which shows the samples as letters, and the consumers as directions of increasing preference. The length of the vectors indicate how much information an individual consumer is contributing to the preference map. Consumers with shorter vectors have preference data which are contributing less information to the sample map than those with longer preference vectors. It is clear from Figure 8 that some consumers are contributing less information than others. However, the same general conclusions to those in Figure 7 can be drawn. The fact that the space has rotated 180° is unimportant, as its the structure that is the key aspect for interpretation.

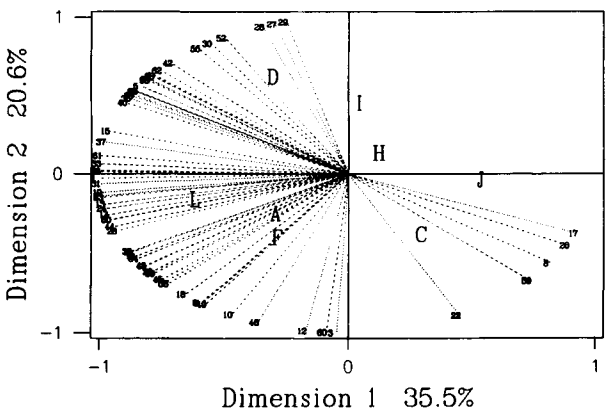


Figure 7. Internal preference mapping plot derived from specifying the option of centering the data and scaling each consumer's preference vector to unit variance.

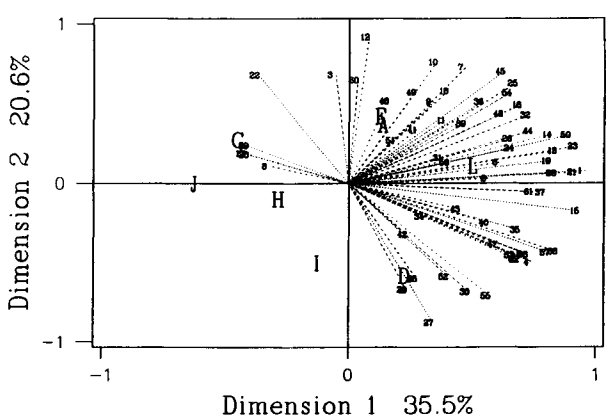


Figure 8. Internal preference mapping plot derived from specifying the option of not scaling each consumer's preference vector to unit variance.

6.5.3 Relationship with Sensory

The correlation coefficients between the sensory attributes and the two preference dimensions were calculated. These are represented in Figure 9, for the preference map of Figure 5. The points are simply obtained by plotting an (x,y) coordinate for each attribute, as two correlation values are obtained for each attribute, one with the first preference dimension (x) and one with the second preference dimension (y). It can be seen that the attributes drying, harsh, tongue tingling, acid, mouthwatering, bitter, rindy/pithy and lemon are all strongly associated with Dimension 1, and the way the samples are separated along this dimension in terms of preference. Examining the direction of the preference vectors on Figure 7 indicates that these characteristics were negative for acceptability.

The attributes sweet and tangerine were associated with preference in both Dimensions 1 and 2, whilst strength of flavour, artificial sweet and thick were more associated with preference Dimension 2. These attributes relate to the preferences of consumers in this upper left quadrant of the plot in Figure 7.

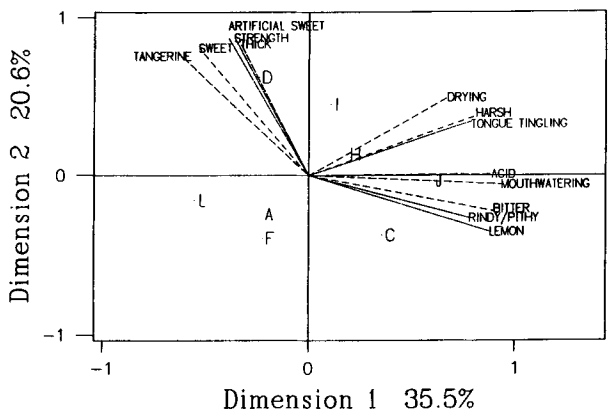


Figure 9. Correlations between preference dimensions in Figure 7 and sensory attributes as rated by a trained sensory panel.

7. CONCLUSIONS

In conclusion, this chapter has highlighted some of the advantages of using both internal and external preference mapping. Emphasis has been on some of the pitfalls awaiting the unwary user. It is therefore worth re-iterating the importance of good experimental design, and of the need to seek advice at the early stages in the learning process, particularly when using external preference mapping. It is hoped that the examples and provision of the necessary data, will allow potential users the chance to experience for themselves the process of running and interpreting data from a preference mapping study.

8. REFERENCES

Arnold, G.M. and Williams, A.A. (1986). The Use of Generalized Procrustes Techniques in Sensory Analysis. In: Piggott, J.R. (Ed.). Statistical Procedures in Food Research. London: Elsevier Applied Science.

Booth, D.A. and Freeman, R. (1992). Personal communication. School of Psychology. University of Birmingham.

Carroll, J.D. (1972). Individual Differences and Multidimensional Scaling. In: Shepard, R.N., Romney, A.K. and Nerlove, S. (Eds.). New York: Academic Press.

Conner, M.T. (1994). An Individualised Psychological Approach to Measuring Influences on Consumer Preferences. In: MacFie, H.J.H and Thomson, D.M.H. (eds.). Measurement of Food Preferences. London: Blackie Academic and Professional.
Economists and the Behavioural and Management Sciences. London: Unwin Hyman.

Kruskal, J.B. (1964). Nonmetric multidimensional scaling. A Numerical Method. Psychometrika, 29, 1-27.

Lyon, D.H., Francombe, M.A., Hasdell, T.A. and Lawson, K. (1990). Guidelines for Sensory Analysis in Product Development and Quality Control. London: Chapman and Hall.

Neave, H.R. (1978). Statistical Tables for Mathematicians, Engineers,

Peryam, D.R. and Pilgrim, F.J. (1957). Hedonic Scale Method for Measuring Food Preferences. Food Technology, 11 (9), 9-14.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981). Introduction to Multidimensional Scaling. New York: Academic Press.

Schlich, P. (1995). Preference Mapping: Relating Consumer Preferences to Sensory or Instrumental Measurements. In: Etievant, P. and Schreier, P. (Eds.). Bioflavour '95. Analysis/Precursor Studies/Biotechnology. France: INRA Editions.

Schlich, P. and McEwan, J.A. (1992). Cartographie des Preferences: Un Outil Statistique Pour l'Industrie Agro-Alimentaire. Sciences des Aliments, 12, 339-355.

Stone, H. and Sidel, J.L. (1985). Sensory Evaluation Practices. London: Academic Press.

APPENDIX 1: SOFTWARE SUPPLIERS

GENSTAT	NAG Ltd., Wilkinson House, Jordon Hill Road, Oxford, OX2 8DR, Great Britain. [Tel: 01865-53233]
MINITAB	CLECOM Ltd., The Research Park, Vincent Drive, Edgbaston, Birmingham, B15 2SQ. [Tel: 0121-471-4199] Minitab Inc., 3081 Enterprise Drive, State College, PA 16801, USA.
PC-MDS	Scott M. Smith, Insitutute of Business Management, Brigham Young University, Provo, Utah, 84602. This has a series of MDS programs including PREFMAP. [Tel: 010-1-801-378-4636/5569]
RS/1	BBN UK Ltd., Software Products Division, One Heathrow Boulevard, 286 Bath Road, West Drayton, Middlesex, UB7 0DQ, Great Britain. [Tel: 0181-745-2800] BBN Software Products, Marketing Communications, 10 Fawcett Street, Cambridge, MA 02138, USA.
SAS	SAS Software Ltd., Wittington House, Henley Road, Medmenham, Marlow, SL7 2EB. [Tel: 01628-486933] SAS Institute Inc., Box 8000, SAS Circle, Cary, NC 27511-8000, USA.
SENPAK	Reading Scientific Services Ltd., Lord Zuckerman Research Centre, Whiteknights, P.O. Box 234, Reading RG6 2LA, Great Britain. [Tel: 01734-868541]
SENSTAT	Sensory Research Laboratories Ltd., 4 High Street, Nailsea, Bristol, BS19 1BW. [Tel: 01275-810183]
STAT-GRAPHICS	Statistical Graphics Coroporation, 5 Indepence Way, Princeton Corp. Ctr., Princeton, NJ 08540, USA. Cocking and Drury Ltd., 180 Tottenham Court Road, London, W1P 9LE, Great Britain. [Tel: 0171-4369481]
S-Plus	Statistical Sciences UK Ltd, 52 Sandfield Road, Oxford, OX3 7RJ. [Tel: 01865-61000]
SPSS	SPSS UK Ltd, 9-11 Queens Road, Walton-on-Thames, Surrey, KT12 5LU. [Tel: 01932-566262] SPSS Inc., 444 North Michigan Avenue, Chicago, IL, 60611, USA.
SYSTAT	SYSTAT UK, 47 Hartfield Crescent, West Wickham, Kent, BR4 9DW. [Tel: 0181-4620093]

APPENDIX 2: SAMPLE COORDINATE SCORES FROM GPA

Sample	Dimension 1	Dimension 2
A	0.3200	0.0162
C	0.0908	0.0455
D	-0.103	-0.0395
F	0.0193	0.0866
H	0.0801	-0.0704
I	-0.0328	-0.0848
J	0.1649	-0.0021
L	-0.0875	0.0485

APPENDIX 3: ACCEPTABILITY DATA

Cons	A	C	D	F	H	I	J	L	Cons	A	C	D	F	H	I	J	L
1	10	0	14	14	3	4	0	22	32	10	1	0	6	1	0	0	22
2	2	9	11	11	8	4	3	15	33	14	5	14	10	11	5	5	15
3	18	16	9	17	13	7	7	6	34	7	3	3	4	8	10	3	8
4	20	5	26	10	8	21	1	17	35	5	0	24	14	4	2	0	10
5	6	5	24	8	10	7	5	22	36	12	13	9	12	10	11	8	15
6	3	6	6	6	6	3	3	17	37	27	3	19	24	13	25	2	21
7	9	11	5	18	8	4	5	16	38	8	0	24	17	7	16	3	22
8	17	24	7	11	25	23	16	22	39	19	23	16	18	8	15	7	19
9	3	14	6	24	11	3	1	16	40	7	2	4	2	0	9	0	6
10	6	10	0	19	0	0	0	10	41	25	1	3	6	17	6	9	15
11	2	7	8	17	5	0	2	8	42	6	23	21	14	12	17	6	18
12	23	25	3	21	17	7	12	23	43	19	8	16	3	6	8	3	8
13	22	6	11	12	1	13	3	26	44	16	10	13	17	3	10	4	13
14	13	0	5	4	0	1	0	13	45	25	11	10	24	6	5	4	17
15	14	8	14	11	9	12	8	15	46	19	22	11	7	11	3	8	18
16	24	2	3	7	0	1	0	20	47	1	0	7	5	0	19	0	21
17	15	13	7	5	11	10	9	6	48	2	3	2	27	4	2	1	28
18	15	18	12	24	22	5	5	21	49	13	7	3	10	9	8	3	7
19	3	0	6	11	3	3	0	9	50	8	1	6	13	2	4	0	18
20	23	24	11	8	27	16	16	16	51	10	4	4	4	7	5	4	5
21	23	4	20	15	1	10	1	22	52	8	0	11	8	10	16	0	5
22	23	18	2	10	4	2	24	10	53	16	6	19	11	21	18	7	22
23	27	5	24	23	3	5	0	28	54	25	7	7	15	10	6	1	16
24	21	5	6	7	15	10	1	24	55	0	0	29	0	0	29	0	29
25	21	9	10	24	10	11	11	22	56	26	9	19	5	13	0	3	10
26	8	8	11	23	19	6	2	27	57	15	10	19	15	14	19	6	18
27	6	3	21	3	7	14	3	6	58	9	5	19	9	10	7	6	14
28	18	4	19	5	21	19	5	10	59	5	14	5	7	3	8	14	11
29	14	3	27	2	24	15	18	22	60	19	14	5	12	12	9	19	22
30	3	5	23	7	19	14	4	20	61	9	11	23	20	10	7	3	18
31	25	10	20	18	6	9	5	4	62	20	6	23	8	16	17	3	17

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

ANALYSING COMPLEX SENSORY DATA BY NON-LINEAR ARTIFICIAL NEURAL NETWORKS

Knut Kvaal[a] and Jean A. McEwan[b]

[a]MATFORSK, Norwegian Food Research Institute, Osloveien 1, 1430 Ås, Norway

[b]Department of Sensory Science, Campden & Chorleywood Food Research Association, Chipping Campden, Gloucestershire, GL55 6LD, United Kingdom

1. INTRODUCTION

Prediction of one data set from another has been the goal of researchers in wide ranging disciplines, including medicine, economic forecasting, market research, physics, chemistry, weather forecasting, quality control, and so on. In the area of sensory science this is also an important goal. The methods used in modelling sensory attributes to physical and chemical measurements have mainly been traditional statistical. These methods are mostly based on a linear approach. Sensory data is often non linear in nature. By introducing non linear methods like neural network, it will be possible to model sensory data in a better way. Neural network modelling is not so much in use in sensory analysis. During the last years there has been a growing interest of using neural networking to describe food quality and preference. By combining statistical methods like PCR and neural nets we will have a new powerful approach of modelling sensory data.

If sensory science is taken in its broadest sense to encompass chemical and instrumental measurements of food, as well as consumer response, then the scope of prediction in this discipline is clear. Predicting the sensory acceptability of new products within a particular product range from sensory information may be cheaper than conducting a full scale consumer survey. It may be cheaper and easier to predict physical measurement from key sensory parameters, or perhaps it may be easier to use instrumental and/or chemical measurements to predict the key sensory parameters which are known to be important to consumer preference.

A variety of regression methods have been used for prediction purposes by those working in sensory science, and related fields. Most commonly these include principal component regression (PCR) and partial least squares regression (PLS Martens et al., 1989). Each of these methods have been shown to work and provide meaningful results for certain types of data. PCR and PLS tend to be used for more serious model building exercises, as programs encompassing these tools allow the flexibility of exploring the best combination of X-variables to provide a good prediction of the Y-data.

The problems associated with prediction in sensory science can be seen as two fold. The first is that data are often non-linear. Simple transformations such as logarithms may help, but

it may be that the non-linearities are more complex. A second problem is that while good models may be obtained where the variation between samples is large, this is seldom the case where variation between samples is very small. In such cases linear models may not provide robust models. It is therefore very important to use data that spans the space we want to investigate and that there is as little collinearity as possible between the variables. To achieve this we will use statistical pre-processing of the data. By using the principal components as inputs to a neural network, it is possible to reduce the collinearity. The principal components are constructed in such way that their bases are orthogonal and that the main variation of the data is to be described in the first components. The noise is being effectively taken away in the higher components.

Working with neural networks is a challenge of trial and error, and it is also very important to have a good knowledge of the history of the data being analysed. There is, however, a growing market for more "intelligent" programs to guide the user in the modelling process. Good programs for building neural networks seems to grow in strength and neural networking will be a good add-on to model building in sensory science.

2. METHODOLOGY

2.1 Neural Networks

"Neural computing is the study of networks of adaptable nodes, which through a process of learning from task examples, store experimental knowledge and make it available for use." (Aleksander and Morton, 1990).

Neural computing is not a topic immediately associated with sensory science, yet its potential, at least from a theoretical point of view, may have far reaching consequences. First it would be useful to have a look at what neural networks are all about. Inspired by biological neuron activity and a mathematical analogy led a group of researchers to explore the possibility of programming a computer to adopt the functionality of the brain (NeuralWare, 1991).

Considering human processing and response (behaviour), it can readily be seen that the brain is constantly learning and updating information on all aspects of that person's experiences, whether these be active or passive. If a person places his hand on a hot plate, then he learns that the result is pain. This response is recorded and his future behaviour with hot plates will be influenced by this learning. There are many such examples, and the reader interested in human processing and cognition should refer to one of the may textbooks on this subject (e.g., Stillings et al, 1987).

It is very important to mention that the neural network philosophy based on biological modelling of the brain is more of an artefact. We will emphasise that the neural network is a method of a mathematical and statistical visualisation based on some fundamental ideas. We will also in this chapter restrict ourselves to a network topology based on function mapping or pattern recogtioning. Discussion will be restricted to the so called *feed forward* layer nets. The information flow between the different neurones in a feed forward layer network always flow towards the output. In feed forward nets, each neurone has its own life getting input and sending the local calculated output to other neurones in the next layer. The training process will force the connection weights to be adjusted to minimise the prediction errors. With all

these neurones processing simultaneously and independently, a computer is needed that has the ability to do parallel task processing. On a sequential computer like the PC, neurone activity needs to be simulated sequentially. Therefore, each neurone activity is calculated in the direction from input to output.

In order to translate the functionality of the brain into a computer environment, it is first necessary to break the processing of information into a number of levels and components. The first level will be the *input* of which there may be several components. For example, an individual is given some chocolate from which he perceives a number of sensory attributes. The chocolate and the individual form the stimulus, and for the sake of argument it will be assumed that the sensory attributes are the input variables, as these can be recorded in the physical world.

At the *output* level, that is the observable response or behaviour, is one component called acceptability, which can also be measured. The *hidden* layers will process the information initiated at the input. The fundamental building block in a neural network is the neurone. The neuron receives input from the neurones in a earlier layer and adds the inputs after having weighted the inputs. The response of the neurone is a result of a non linear treatment in different regions in the inputspace. The neurones in the hidden layer may be identified as *feature detectors*. Several hidden layers may exist, but in practice only one is sufficient. This is represented in Figure 1b. The next problem is how to join the levels of the network. In the human brain there is a complex network of connections between the different levels, and the complexity of their use will depend on the amount and type of information processing required.

So far the concepts of input, output and hidden layers have been explained. The next concept is that of a neurone as the processing element. Each neurone has one or more input paths called dendrites. The input paths to processing elements in the hidden layers are combined in the form of a weighted summation (Figure 1a), sometimes referred to as the internal activation. A transfer function is then used to get to the output level. This transfer function is usually either linear, sigmoid or hyperbolic. The sigmoid transfer function is illustrated in Figure 1c. This transfer function behaves linear or non linear according to the range of input. The function acts as a threshold when low level input values are presented. It acts as a saturating function when high level input values are presented. In between it acts as a linear function. In this way we achieve a very flexible function mapping.

The feed-forward neural network in Figure 1 is defined by an equation of the form

$$y=f[\Sigma_i\, b_i f\,(\Sigma_j\, w_{ij}\, x_j + a_{i1}) + a_2\,] + e \qquad (1)$$

where y is the output variable, the x's are the input variables, e is a random error term, f is the transfer function and b_i, w_{ij}, a_1 and a_2 are constants to be determined. The constants w_{ij} are the weights that each input element must be multiplied by before their contributions are added in node i in the hidden layer. In this node, the sum over j of all elements $w_{ij}x_j$ is used as input to the transfer function f. This is in turn multiplied by a weight constants b_i before the summation over i. The constants b_i are the weights that each output from the hidden layer must be multiplied by before their contributions are added in the output neurone. At last the sum over i is used as input for the transfer function f. More than one hidden layer can be used resulting in a similar, but more complicated, function. The constants a_1 and a_2 acts as bias signals to the network. They play the same role as the intercept constants in linear regression.

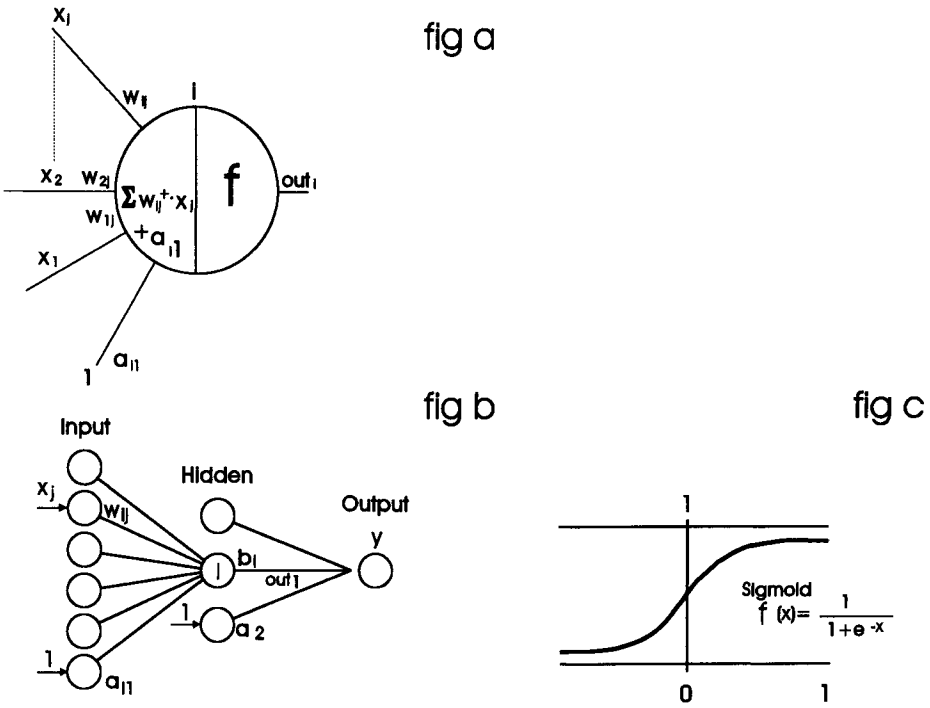


Figure 1. (a) A diagram illustrating the structure of a simple neural network. The fundamental building block in a neural network is the neurone. (b) The connections between different connections in the neural network. The input paths to processing elements in the hidden layers are combined in the form of a weighted summation. (c) A sigmoid transfer function is then used to get to the output level.

2.1.1 Learning and backpropagation

The word *backpropagation* originates from the special learning rule invented by several workers (Hertz et al,1991 page 115). The method is used to optimise a *cost function* (error function) of the squared differences in predicted output and wanted output. In short, information flows from the input towards output, and error propagates back from output to input. The error in the output layer is calculated as the difference of the actual and the desired output. This error is transferred to the neurones in a middle layer. The middle layer errors are calculated as the weighted sum of the error contributions from the nearest layer. The derivative of the transfer function with respect to the input is used to calculate the so called deltas. The deltas are used to update the weights. The derivative of the transfer function will be zero for very small summed inputs and for very large summed inputs. Thus the derivative of the transfer function stabilizates the learning process.

The backpropagation algorithm minimises the error along the steepest decent path. This may introduce problems with local minima. Finding the global minimum of the error in a model is equivalent of estimation of an optimal set of weights. Learning in a neural network is the process of estimating the set of weights that minimise the error. A trained neural network has the ability to predict responses from a new pattern. The training process may be performed by using different learning rules. This chapter will focus on the backpropagation delta rule. Here the errors for each layer will propagate as a backward information in the network. The weights are updated based on these errors.

The weights are calculated in an iteration process. The weights are given initially random values. By presenting a pattern to net network, the weights are updated by computing the layer errors and the weight changes. The learning process will stop when the network has reached a proper minimum error. The learning process is controlled by the learning constants *lrate* and *momentum*. The learning constants are chosen between 0 and 1. Small values slow down the learning. Typical values are 0.5. The lrate controls the update rate according to the new weights change. The momentum acts as a stabilisator being aware of the previous weight changes. In this way the momentum minimises oscillation of the weights. The learning by estimating the weights is described for each layer by

$$W_{(new)} = W_{(old)} + lrate*dW_{(new)} + momentum*dW_{(old)} \quad (2)$$

where W_{new} are the new and updated weights, W_{old} are the weights before updating, dW_{new} are the new deltaweights calculated by the backpropagaton learning rule and dW_{old} are the old deltaweights. The error is calculated as the difference between the actual and calculated outputs. The updates of the weights may be done after each pattern presentation or after all the patterns have been presented to the network (epoch).

There are many modifications of this rule. One approach is to vary the learning constants in a manner to speed the learning process. The method of self adapting constants is considered to be of great value to reduce the computing time during the learning phase. Each weight may also have its own learning rate and momentum term. This approach together with the self adapting learning rates, speeds the learning and therefore it is not so important to choose proper starting values (delta-bar-delta rule). For a more extensive discussion of the mathematics of the backpropagation algorithm, the reader should see the Chapter 6 of Hertz (1991).

2.1.2 Local and Global Minima of the Error

One of the major disadvantages of the backpropagation learning rule is its ability to get stuck in local minima. The error is a function of all the weights in a multidimensional space. This may be visualised as the error surface in a three dimensional space as a landscape with hills and valleys. There is no proof that the global minimum of the error surface has been reached. Starting with different randomised weights leads to different minima if the error surface is rugged. It is important to consider this when analysing the final minimum. The learning is run repeatedly at different starting locations to show that the minimum is reasonable.

2.2 Normalisation

Normalisation is a pre-processing of the data. Many techniques are being used. (Masters, 1995) Here we will present one approach that has shown great effect on the kind of data we are working with.

The data being presented to the network program has to be normalised to a level that does not drive the transfer functions into saturation. Saturation occurs when the change in output of the transfer function is almost zero due to high input values. Another aspect is also to insure that variables with large variations do not overrun variables with small variations. By using the minimum and maximum values of each input and output variable, the network program will normalise the inputs between 0 and 1. The output is normalised between 0.2 and 0.8. If the output transfer function is linear, then there is no need to normalise the output in this way. If several outputs are being used in the modelling, then it is important to normalise according to the variation of the output variables. When the variables are presented to the prediction they are recalculated according to the min/max values. It is important to be aware of the normalisation, but it is mostly handled automatically by the neural network program being used. The limits used here are well designed to the sigmoid transfer function. If a hyperbolic transfer function is used one could normalise between -1 and 1 for the inputs and between -0.8 and 0.8 for the outputs.

2.2.1 Validation of the Performance

Validation of the performance is very important when we want to monitor the generalisation ability of the network. By adding more neurones in the hidden layer, the network becomes a very flexible function mapper. This in turn rises the danger of overfitting. The network may be able to map the learning data perfect, but the predictions on test data may be poor. The validation is by this concept not only to find the iteration count in the learning process, but also a very important process when evaluating the number of nodes in the hidden layers.

Validating the network performance by using a separate test set must be considered. The data is split into two sets, the learning set and the test set. The learning set is used to train the network and find the set of constants that minimises the prediction error. The testing is performed on the test set. It is important to design the learning and test set in such manner that they span most of the variable ranges. The test set is considered to be a set of unknown objects. Ideally a complete independent validation set should be used to test the networks modelling ability. The neural net is a very flexible modelling system. Therefore the test set used in optimisation of network topology may not be satisfactory in validating the generalisation ability of the network. In our presentation we do not use this extra validation set. This is due to the lack of objects and our presentation serves as an illustration. Validation of the network performance is done using the root mean square error of prediction (RMSEP) (Martens et al, 1989).

Another method to be considered is the cross-validation. Cross-validation may be used to validate how single objects are modelled against all the other. By leaving one out to the test set and using all the other objects as learning set, we may get a measure of the average performance of the network. (Leave One Out). (Kvaal et al, 1995). It is also possible to divide the objects into test segments and learn segments in such way that the objects are being tested only once. This will construct network models based on learn and test sets in the way that all the objects in turn will be test objects. One major problem in using cross-validation on neural

nets is the danger of getting into local minima. There will be one new model at each segment validation. This in turn gives different local minima.

This neural network cross validation is similar to PCA/PLS cross-validation but the number of iterations instead of number of factors will be considered. The RMSEP will be a mean for all the models constructed. By using cross-validation the network topology may be optimised. One major problem with the cross-validation applied on neural nets is the huge amount of computing time needed. It is also a problem to interpret the different models being constructed. A main preference of using the cross-validation is the information of the average model performance. It is also a preference to get information of objects that is difficult to model.

2.3 When and Why Neural Networks Should Be Used

With neural network modelling still in its early stages of development and understanding, addressing the questions on when and why neural networks should be used poses some problems. However, through reading and discussion, a number of general guidelines should be considered.

The «when» question will be considered first. Neural computing can be used if the problem is to predict either responses recorded on a continuous scale, or to do classifications. A neural network may be considered as a function mapping device. It may also be considered to be a kind of pattern recognition memory which can generalise on unknown samples. The design of the transfer function is essential in the design of what kind of problems the user wants to solve. Most often the sigmoid transfer function gives useful results. Most software packages have the possibility to change to other transfer functions according to the data taken into account. This process on choosing the right network design is a trial and error process. However some guidelines might be considered.

Different software packages integrate statistical tools and graphic visualisations. Unlike the PCR/PLS where there is a lot of information in the score- and loading plots, it is not easy to interpret the weights of the neural network. There are methods, however, to optimise neurones and find variables that are essential and have an effect on the model. We have already mentioned the cross-validation. It is still an area of research to develop good diagnostic tool used on feed forward networks particularly.

As neural networks may detect non linearities it is a natural choice to use this method. If the data is purely linear methods like PCR and PLS is most likely to be used. The user is recommended to start with PCR/PLS to get a good knowledge of the data set with the tools that this methods have. This will indicate that there might exist non linear relations that a neural network might be able to solve. The flexible nature of neural nets forces the user to be aware of the overfitting problem. A neural net is supposed to model a X/Y relation. The problem is the generalisation on unknown Y's. It is very important to understand this fact, because we often see neural network models that is perfect in respect to the actual X/Y being used in the learning.

The question on «why» use neural networks must be answered on what precision of generalisations the user wants. Neural computing should be used as it provides a powerful alternative to traditional statistical methods for the prediction of responses. If only generalisations are wanted the neural network computes this more easy in a well defined function. If diagnostic tools are essential together with generalisations, a mixture of linear and non linear methods should be taken into account.

2.4 Relationship to Other Methods

The artificial neural network models have strong relations to other statistical methods that have been frequently used in the past. An extensive discussion of the relations will be found in Næs et al, 1993. When a network is trained, the weights are estimated in such way that prediction error is as small as possible. The design of the feed forward network is closely related to multiple linear regression (MLR) when the linear transfer function is used and the hidden layer is removed. In this case the neural network may be solved directly and no training is necessary. If the data has a purely linear relationship, the MLR method may give good results. If, however, the data has non linear relationship, the MLR method will not give satisfactory predictions. Non-linear methods like neural networks should be taken into account to detect the non-linear relations in the data. The transfer function used in the neural net is designed to detect both linear and non-linear relations in the data. Reports in combining MLR and neural nets into one network topology claims success to guarantee optimal solutions on data sets with unknown relationships (Borgaard et al, 1992).

2.4.1 Data Pre-processing and Compression

Near Infrared Spectroscopy (NIR) can be used to directly determine water, most organic molecules and some inorganic molecules by using the absorbance spectra. The variables are described by the wavelength and the absorbance at this wavelength. The variables are strongly correlated and it is generally not possible to select single variables to describe properties. We have to use a multivariate approach to solve this problem. (Hildrum et al, 1992). NIR data may be composed of several hundred variables (wavelengths). This implies large networks and probably very heavy computing when the network is trained. Using methods for compressing the data to fewer number of variables gives networks with lesser number of nodes. Data compression using principal components scores has been reported to be successful in constructing NIR based networks (Borggard et al, 1992, Næs et all, 1993). Typically 250 NIR variables will be compressed to, say, 5 variables. This gives a network with 5 inputs instead of 250. All the major information is described in the principal component scores. Training a network based on scores is very fast. There is often a need to optimise the number of inputs and the number of hidden nodes to give the best predictions. The number of inputs corresponds very often to the optimal factor number in PCR/PLS. This is explained by the fact that the most dominant information is contained in the variables up to the optimal factor number. Higher order variables will contain noise with respect to the attribute we are considering. (Næs et al, 1992). It is therefore a good practice to run PCR/PLS before the network is constructed. Using pc-scores does not always guarantee a better performance. This will be demonstrated by example later. By using a proper number of pc-scores as inputs, the noise is removed more effectively from the data. This ensures that the noise is not a dominant factor in the modelling an overfitting problems may be reduced.

2.5 Advantages, Disadvantages and Limitations

Neural nets are normally easy to implement using a standard program with a user friendly interface. One problem is often that networks based on many input variables like NIR raw data need a long computing time. If time is no problem, then neural networks will be okay.

However, if time is a problem, then data compression is recommended. The main advantage is that a possible solution to a non-linear data problem has good chance of being a success. One solution to a time problem is to run the network on a fast pc and with a floating point coprosessor. One disadvantage using neural nets is the limited set of interpretative tools. It is difficult to interpret the hidden layer weights. Smaller networks, however, are easier to interpret. When training a neural network the initial values of the weights may be randomised differently when the same network is run twice. This gives different results but in the same range. Running PCR/PLS with the same parameters set, will result in a reproduction of the earlier runs. It is important to be aware of this.

2.6 How to Apply the Method

This section provides some guidance on the process in setting up an experiment for analysis by neural computing. A list a key steps is given below and these steps will be further explained later.

1. Prepare the data
2. Optimise the learning rates or use self adapting learning rates.
3. Train the network using raw data or PC scores.
4. Optimise the network model using different number of nodes in the hidden layer.
5. When PC scores are being used, optimise the number of inputs and number of nodes in the hidden layer.
6. Validate the network by consider the RMSEP using a separate test set. (Do a cross-validation of the network if the number of samples are small)

2.6.1 Data Preparation

Data preparation is often a problem when the planning has been bad. Data presented to the neural network needs to be organised in a special way. The data set consists of rows and columns. The rows are the different objects, and the columns are the different variables. The variables are normalised by using a minimum/maximum table as described earlier. This will ensure that variables with great variability do not overrun variables with small variability. The variables describe the pattern that is presented to the network. Input- and output patterns are presented simultaneously to the network. The data set is divided into a training set and a test set. These sets are often separated in two different files, but they might be combined in one file. The training set is the first block of rows and the test set is the last block of rows. This data preparation is often a problem to users of scientific software and statistical packages. Many programs have import facilities to read spreadsheet files to easy handle the problem.

2.6.2 Setting Start values of the Network Parameters

The neural network design needs initial values of learning rates and weights. These initial values are highly dependent on the kind of problem that is being solved. Algorithms exist to set the initial values according to the data set (Demuth H, et al, 1994). If a network has learning rates which are too strong, then the weights will give oscillating and unstable networks. Learn rates which are too low will give extremely long learning time. It is a good practice to start with a relatively strong learning and gradually reduce it as the learning goes on. Typically learn rates and momentum of 0.8 are used. We have so far considered a global learning rate. It is, however, possible to construct learning mechanisms so that each weight is assigned a learning rate. These learning rates are adapted to reasonable values during learning and the network will learn faster.

Initialisation of the weights is often done by randomisation. The weights will typically be initialised to values in the range of ± 0.1. This depends on the number of nodes in use. Some software packages used in neural computing have implemented algorithms to do an initialisation of the weights optimally. This will give a good starting point for further training of the network.

2.6.3 Training the Network

When the network parameters have been initialised the training is done by updating the weights according to a learning rule. In this work the backpropagation learning rule has been used. Commercial network programs often have different modifications of the standard backpropagation. The user need to obtain experience in their own application. By running the network models based on different learning rates, momentum and weights initialisations, number of hidden neurones, etc, the user gets experience of what to by trial and error.

Randomisation gives a new starting point every time it is performed. If a network gets into a local minimum there will be methods for getting out by giving the weights small random variations. It is therefore a common practice to run the network several times to see if different starting points gives different results. It is the randomisation algorithms that decides the different starting points when the network is trained.

The training is stopped when the output error has reached a minimum error on a test set. It is important to have a test set to avoid overfitting of the network. The prediction error will reach a minimum after a set of iterations (calculations). Hopefully the network will generalise well on a separate data set. This validation is important as the network proceeds to learn. This will be discussed more closely later.

2.6.4 Inputs and Nodes

Input variables in the data set are fed to the input nodes. These input nodes are connected to the hidden layer nodes which in turn are connected to the output nodes. How to decide which inputs to use, how many hidden neurones and how many outputs to use simultaneously is not an easy task. This is often done by trial and error. A golden rule is to keep the number of hidden nodes at a low number and vary this number to find an optimum. Using one or several outputs is also done by trial and error. This will be demonstrated by an example later in this chapter.

2.6.5 Validation of the Network

It is important to validate the network on a data set that consist of unknown input/output patterns. The validation is a process that follows the learning of the network. It is also a process that has to be done on the final result of the learning. Many newcomers to neural computing forget about the fact that testing on the learning set will in most cases give very small errors. The network's ability to generalise is, however, not guaranteed because of the possibility of overfitting. If learning is halted at selected points and tested against a separate data set not used in learning one will assure proper learning. A measure often used is the *Root Mean Square Error of Prediction* (RMSEP) (Næs et al,1989). Here the prediction errors of the test set are compared with the wanted outputs of the network. A typical situation during the learning process is shown in Figure 2. Here the error on the learning set gets smaller and smaller as the learning goes on. The error on the test set, however, will reach a minimum at a special number of epochs E_{opt} (number of times all the patterns have been presented to the network). Beyond the E_{opt} we may have underfitting, and above E_{opt} we may have overfitting. The criteria of when to stop is usually set to E_{opt} , but because the network often converges to a stable solution this point might not be so critical. Then we would select an epoch count where the network seems stable (no change in output error from one epoch to the next).

If only a small number of samples are available special techniques for validation of the network should be used. A popular validation method is to select learning and test samples from a population of samples in such way that every sample will be located in a test sample only once ("Jack knifing" or cross-validation). Every time a new learn- and test set is made a network is trained. This can be useful to optimise a network topology.

2.7 Software

Several software packages are available. The software most familiar to the authors, is NeuralWorks Professional II/PLUS package (NeuralWare Inc, USA). This package has the advantage of being a self-contained neural network constructing tool and it has a some diagnostic tools available.

A lot of users, including the authors, do the implementation of special network applications by programming. A lot of textbooks are available telling how to implement a neural network. Some of the standard software tools have, however, the possibilities to be customised. A very popular toolkit is the MathWorks inc Neural Nework Toolbox for Matlab.

Other popular packages on the market are listed in several sources, among them is AI-Expert, February 1993 or updated in later issues. Lists of packages are also available on the Internet. Here there are several neural network packages available for free.

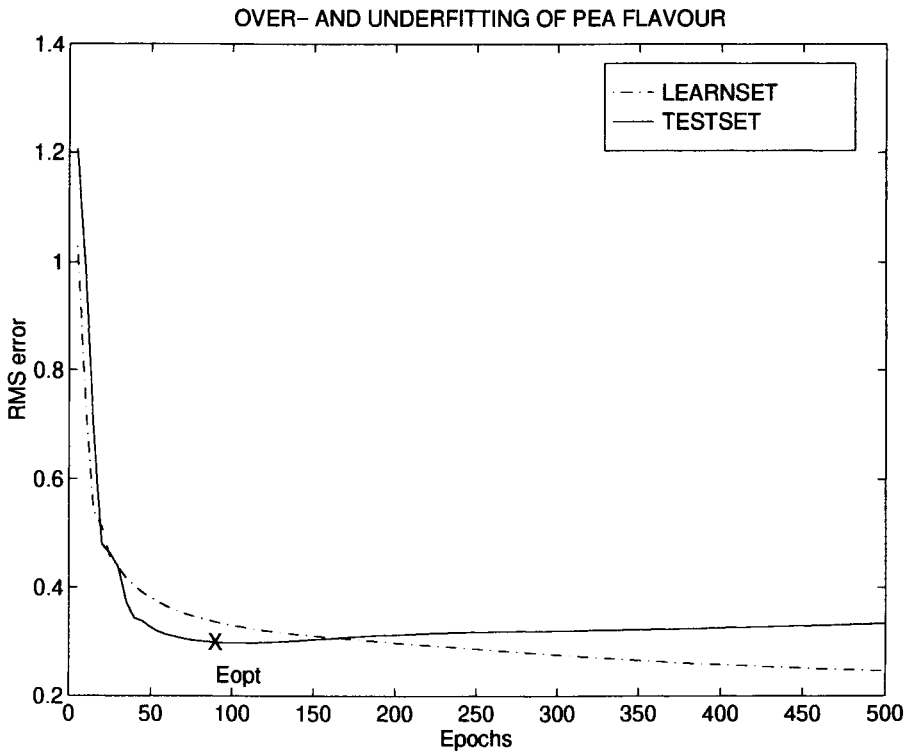


Figure 2. Over- and underfitting of Peaflavour. The learnset is being fitted better and better, but the network has an optimal generalisation ability about 100 epochs.

3. DESIGN CONSIDERATIONS

3.1 Choice and Selection of Samples

Like all prediction modelling problems, it is desirable to have a large number of representative samples. The question of how large is often difficult to answer, as this depends on the selection of samples chosen. A well designed selection of samples can often be less in number and produce a robust and reliable model, than a large number of samples which do not adequately represent the range of available possibilities.

In general, large in the context of neural networks means in the region of 100 samples for initial training, plus a suitable test set of at least 50 samples.

Often the samples are obtained from a factorially designed experiment. In this way more information can be obtained to help understand the reasons underlying the final response. In other applications, a representative range of samples is chosen, such as in the case of predicting the authenticity status of fruit juices.

3.2 Data Collection

The most obvious sensory methodology to use is the method of profiling (quantitative descriptive analysis). This is because from the sensory scientist's point of view sensory attributes are being used to predict some other less easy or more expensive measurement or set of measurements. For example, consumer perception of quality, preference, etc. Within the context of this area, the possibility of wishing to predict sensory characteristics from chemical or other instrumental measures must also be considered. Thus, while profiling is the most common sensory approach, methods for measuring single sensory attributes, sensory quality or sensory difference may be used as the variable to be predicted. It is not intended to cover the collection of the sensory or instrumental data, as this is covered in other textbooks (e.g., Piggott, 1984). The key point about these methods is that they should not only describe the samples, but provide reliable information about the differences between samples.

Instrumental methods include the use of NIRS, GC, Instron, HPLC etc, and collection of these data are covered in texts such as Kress-Rogers(1993).

4. ANALYSIS AND INTERPRETATION OF RESULTS: AN EXAMPLE

4.1 Background

The data used to illustrate the methodology of neural networks in relation to PCR and PLS was provided by MATFORSK (Kjølstad et al., 1988; Næs and Kowalski, 1989). The interested reader should refer to the former reference for a detailed explanation of the samples and data collection procedures.

The reason for including analysis by PCR and PLS is to provide a basis for comparing the performance of neural networks to approaches of known performance ability.

4.1.1 Samples

Samples were 60 independent batches of green peas, collected from 27 different varieties at different degrees of maturity. All samples were freeze dried.

4.1.2 Instrumental Analysis

Near infra red spectroscopy (NIR) analysis was performed on the 60 samples of peas using an Infra Analyser 500 instrument. The data were subjected to a multiplicative scatter correction (MSC) prior to calibration to reduce the effect of different light scatter caused by NIR analysis. This MSC was based on the average spectrum of all 60 samples. A reduction procedure was used to obtain 116 variables from the original 700 wavelengths. The instrumental data forms the X-matrix of 60 samples by 116 variables.

4.1.3 Sensory Analysis

Sensory analysis was carried out using profiling (quantitative descriptive analysis). A trained panel of 12 assessors agreed on and defined 12 sensory attributes to describe a range of pea samples. In previous work (Næs and Kowalski, 1989) six of these attributes were selected as being related to the quality of peas; pea flavour, sweetness, fruity, off-flavour, mealiness and

hardness. The data were averaged over assessors and replicates, thus providing a 60 samples by 6 variable Y-matrix.

4.1.4 The Problem

The problem presented is to predict the sensory attributes of peas from the NIR wavelengths. It is of great interest to see how the neural network may model sensory attributes and NIR data. In this way we may use this method to predict and do classifications on later samples by using NIR and the network and not the assessors. To test the predictability it was decided to select 40 samples to build the calibration model, and use the other 20 as the test set for prediction. Clearly this number is less than that which is ideal, however, this example will serve its illustrative purpose. It is intended to focus on pea flavour as an example for visualisation in this chapter.

4.2 Neural Network Modelling Procedures

4.2.1 Data Pre-treatment

The data were submitted to the neural network program in two forms. The first comprised the raw data, whilst the second used the principal component scores from the NIR data. Using principal component scores has been shown to provide better neural network models than the raw data in some instances (Næs et al., 1993). Forty samples were used as the learning set, and 20 as the test set. The NIR data are normalised between 0 and 1, and the sensory attributes are normalised between 0.2 and 0.8. before being presented to the network. Each variable is normalised between minimum and maximum values. Normalisation of variables may be critical and we recommend the reader to do an extensive investigation of this topic (Masters T, 1993 page 262).

4.2.2 Approach to Data Analysis

The decision of when to use principal component scores as inputs to the network has to be made by trial and error. In this approach, the raw data were used first, then the principal components were used as input variables to the network. It should be possible to have a network model with at least the same degree of RMS. as PCR. The optimal number of scores from PCR were used as the number of score inputs to our network. It is also recommended to optimise the number of score inputs and the number of neurones in the hidden layer to find an optimal network model.

When the network model is constructed, some parameters have to be supplied; the learning constants. This will be the learning rate and the momentum term. The network will try to find the optimal learning rates. In order to achieve optimal conditions prior to fitting the neural network model, a design could be constructed to compare the performance of different levels of learning rate and different momentum rates. As mentioned in Section 2, there is also the possibility to use self adapting learning rates. It is also possible to construct learning based on self adapting learning constants. In this approach it is easier to train a network model and to minimise the prediction error because we do not need to pay a great attention to the starting values of the learning constants and the strength of learning process.

When building the network model, the number of simultaneous outputs have to be taken into account. In our approach, a network to predict all the sensory attributes at the same time

will be sought. Generally it is more difficult to predict several attributes at the same time than predicting only one attribute. This has to do with a more complex error surface, as some attributes may not model well by NIR. When training the network, test at interval iteration counts and stop when the RMSEP is at an optimum to prevent over-fitting.

4.2.3 Optimising Learning Rate and Momentum

The starting point is a network model based on the raw data and all six sensory attributes. The network is run repeatedly by varying the learning rate and momentum systematically from 0.2 to 0.8 in steps of 0.2, and the networks are all trained to a fixed iteration counts. The step length were found by experience. By using a smaller step we did not get more information. The number of hidden neurones is held at a fixed value of 1. Performance is always measured as RMSEP (validation of test set). Previous experimentation has shown that optimisation of learning rates is mostly independent of the number of neurones in hidden layer. (Masters, 1994, page 7)

From Figure 3 it was concluded that learning rate of 0.8 and momentum of 0.8 would converge fast until a point where the weights are unstable. By gradually reducing the learning rates from this point (3000 iterations) it is possible to get stable weights during the rest of the learning process. We see from equation (2) that when the weight change is large we may get an oscillatory change in the weights by using relative large learning constants. We may prevent this effect by gradually reducing the learning constants. We also observe that the momentum term takes care of the old weight change. The momentum term is designed to prevent oscillatory behaviour of the learning. By reducing the momentum it is a danger to reduce original purpose of the momentum. Experience have shown that reducing both the learn rate and the momentum by 50 percent has a good effect. Figure 4 shows the corresponding learning rate and momentum when using four principal components. When the principal component scores are used as input we deal with the main variant part of the data. It can be seen that the convergence is more smooth due to the fact that the noise has been removed from the data. Choosing a learning rate of 0.8 and momentum of 0.8 gives a fast convergence. This corresponds to the values found for raw data.

4.2.4 Optimising the Number of Inputs and the Number of Hidden Neurones

When raw data are being used, the need to optimise the of number of inputs often results in finding the inputs that do not contribute to the network response. It is possible to exclude some inputs in most network programs. In this example the inputs in raw data are not varied. Using PC scores this number has to be optimised. We will also mention the principle of *pruning* the network. In this technique nodes are deactivated or simply destroyed. The resulting network is more optimal and based on lesser neurones. Hopefully it performs as good as the original network topology.

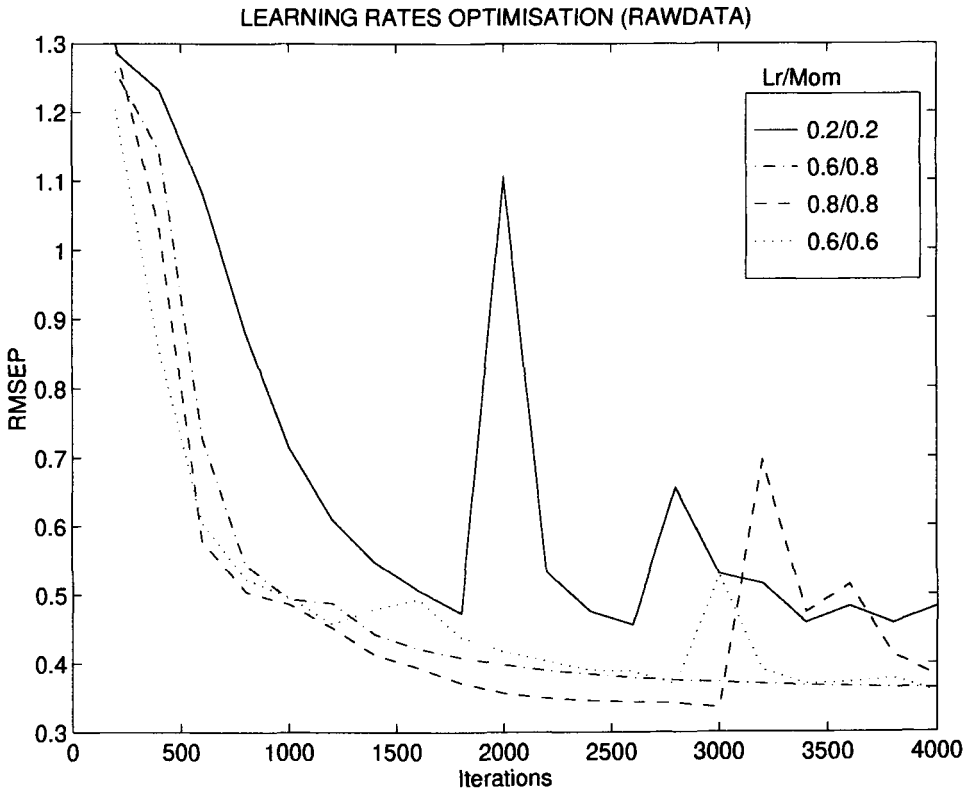


Figure 3. Learning rates optimisation (raw data). The learning rate and momentum are varied systematically from 0.2 to 0.8 in steps of 0.2, and the networks are all trained to a fixed iteration counts. Oscillatory behaviour occurs when the learning is to strong.

Varying the number of hidden nodes between 1 and 30, it can be seen from Figure 5 that there is an optimum at about 10 hidden neurones when using NIR raw data. This indicates that the hidden layer acts as a *feature detector* and that more than one hidden neurone is needed to give optimal results. The least flexible model is obtained by using only one hidden neurone. By adding more hidden neurones this will result in a network that is able to detect more features. It is also experienced that overfitting is not so critical with one hidden as it is with more hidden neurones. This is very important to have in mind when optimising the hidden layer. It therefore important to start the optimisation from 1 and not in the opposite direction. Optimising the number of inputs by varying the number of PC-scores as inputs and using one hidden neurone in the hidden layer will give a similar result. It can be seen from Figure 6 that there is an optimum at four inputs corresponding to the optimal number of factors from PCA.

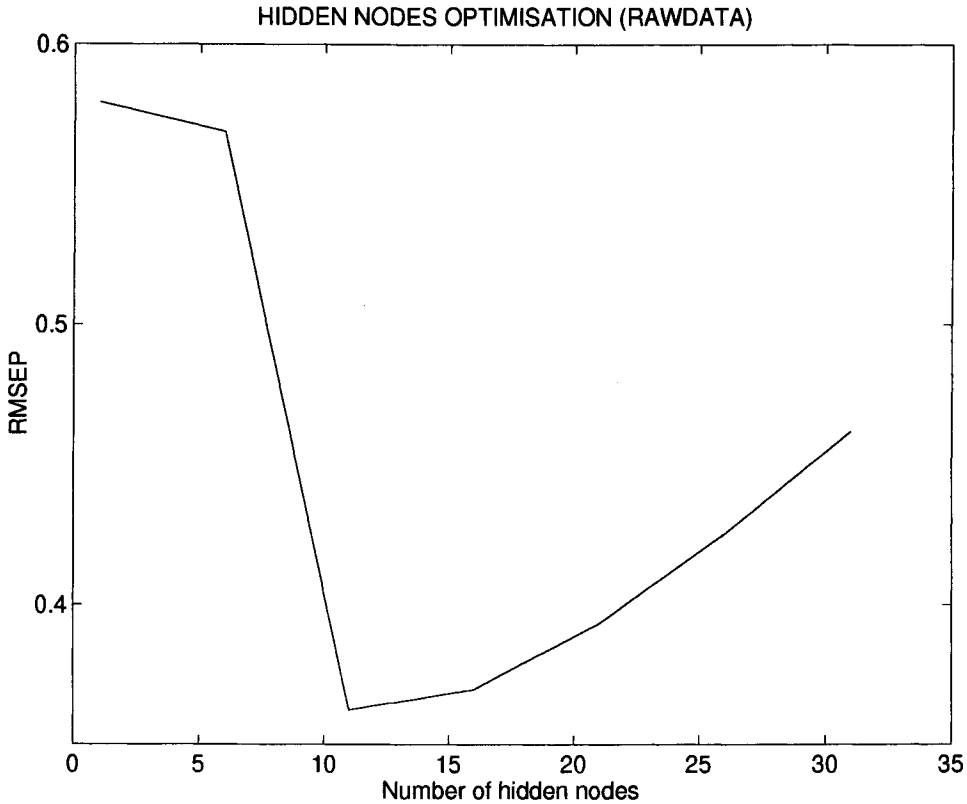


Figure 4. Learning rates optimisation when using four principal components. When the principal component scores are used as input we deal with the main variant part of the data. It can be seen that the convergence is more smooth due to the fact that the noise has been removed from the data. Choosing a learning rate of 0.8 and momentum of 0.8 gives a fast convergence.

Undertaking a variation of the number of nodes in input layer and the hidden layer at the same time gives an indication that four components are optimal. From Figure 7, it can be seen that as many as 8 neurones in the hidden layer gives optimal models. By examining this contourplot we may conclude that there exist an optimal area starting at 4 inputs and 5 hidden neurones. By choosing this topology the 4 inputs corresponds to 4 optimal principal components and the 5 hidden neurones acts as feature detectors for the attributes. We also observe from the contour plot in Figure 7 that there is an area starting at 4 inputs and only one hidden neurone that gives a relative small error. This network topology may be usable because the number of iterations is not so critical when it comes to the danger of overfitting. Later we will see that this is corresponding to the PCR model optimum.

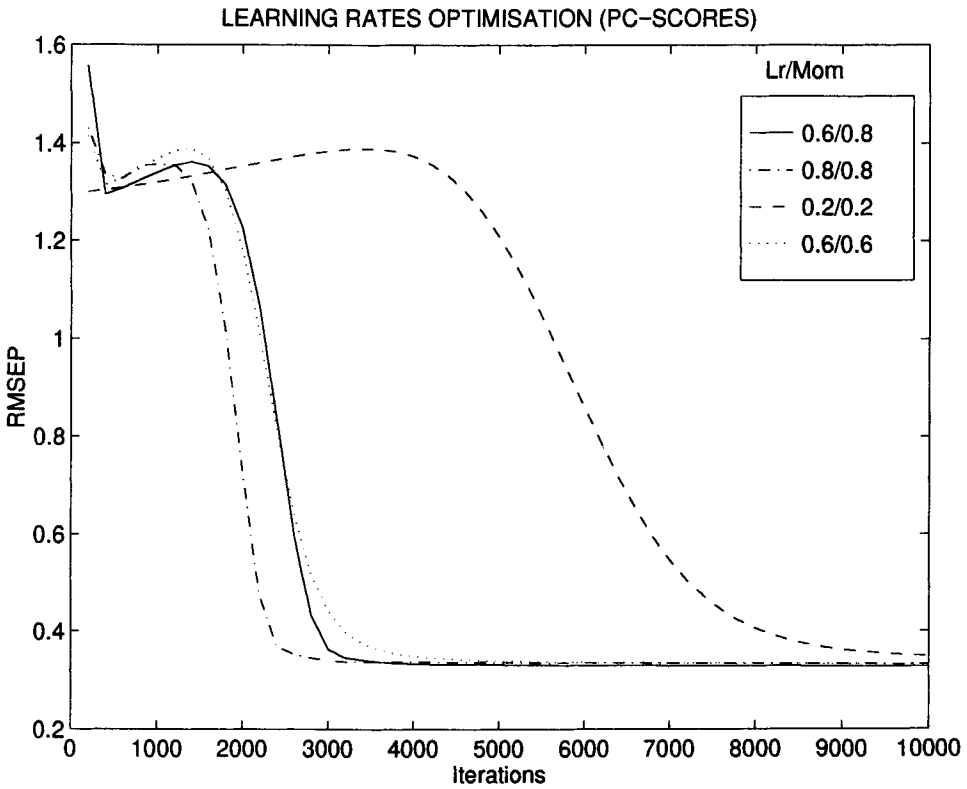


Figure 5. Hidden nodes optimisation (raw data). By varying the number of hidden nodes between 1 and 30, the optimum at about 10 hidden neurones indicates that the hidden layer acts as a "feature detector" and that more than one hidden neurone is needed to give optimal results.

4.2.5 Training an Optimal Topology to Find the Global Error Minimum

By the process we have described we have achieved a network topology that should be optimal for the data being used. It is now possible to run the network with this topology more extensive and elaborate. Varying the starting point by different weights initialisations may be effective. Proper initialisation may speed the learning process and may also enable better performance. New and popular methods is to use methods like genetic algorithms and simulated annealing (Masters, 1993). During learning it is also important to stop the learning when one suspects a local minimum, do some small perturbation of the network weights (jogging) and continue the learning hopefully outside the local minimum.

As stated earlier, it is often a good practice to gradually reduce the learning constants from a point where the weights are unstable. The practice is to reduce the constants at iteration

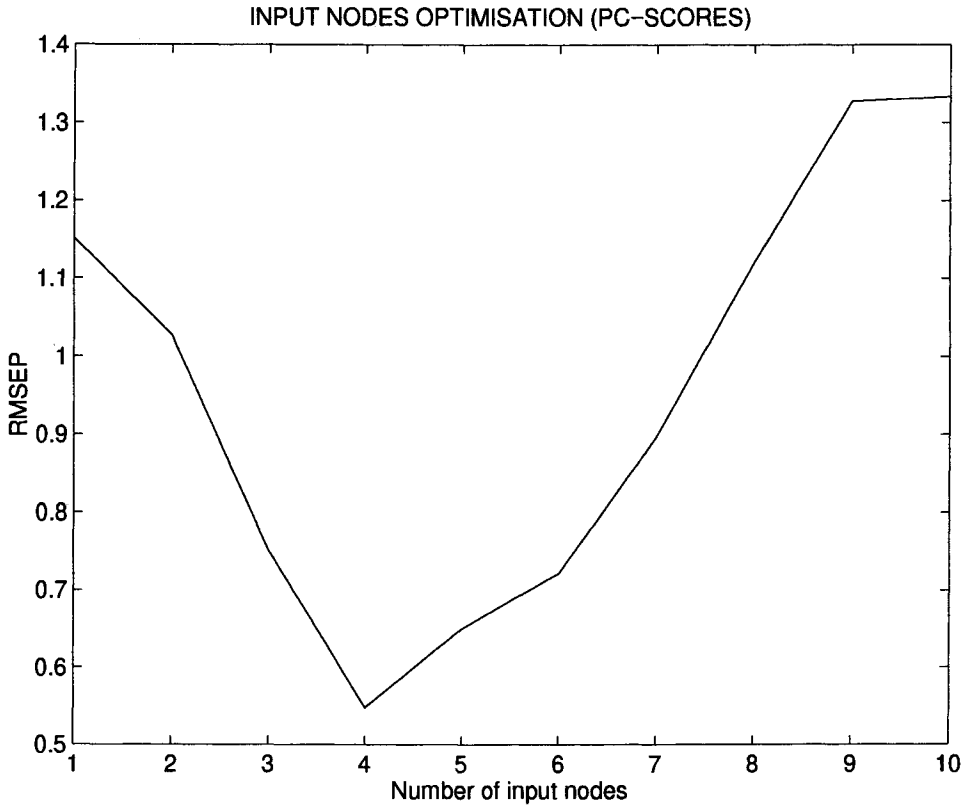


Figure 6. Input nodes optimisation (PC-scores) using one hidden neurone. By varying the number of inputs between 1 and 10, the optimum at about 4 neurones indicates verifies that 4 principal components are optimal.

intervals by a factor of 0.5-0.75. This ensures a more stable weights change.

The results from modelling with test set validation are listed in Table 1. It can be seen that the raw data model gives predictions at the same level or some better than the PC-score model. This shows that it is not always the case that better models are obtained with PC-scores. However, it is a positive point that only four inputs are needed using PC-scores. This will use a lot less computing power to calculate predictions. It is also important to observe that using PC-scores improves interpretation of the results. By doing an analysis by PCR and then building a neural network based on the scores, it is possible to have a better understanding of

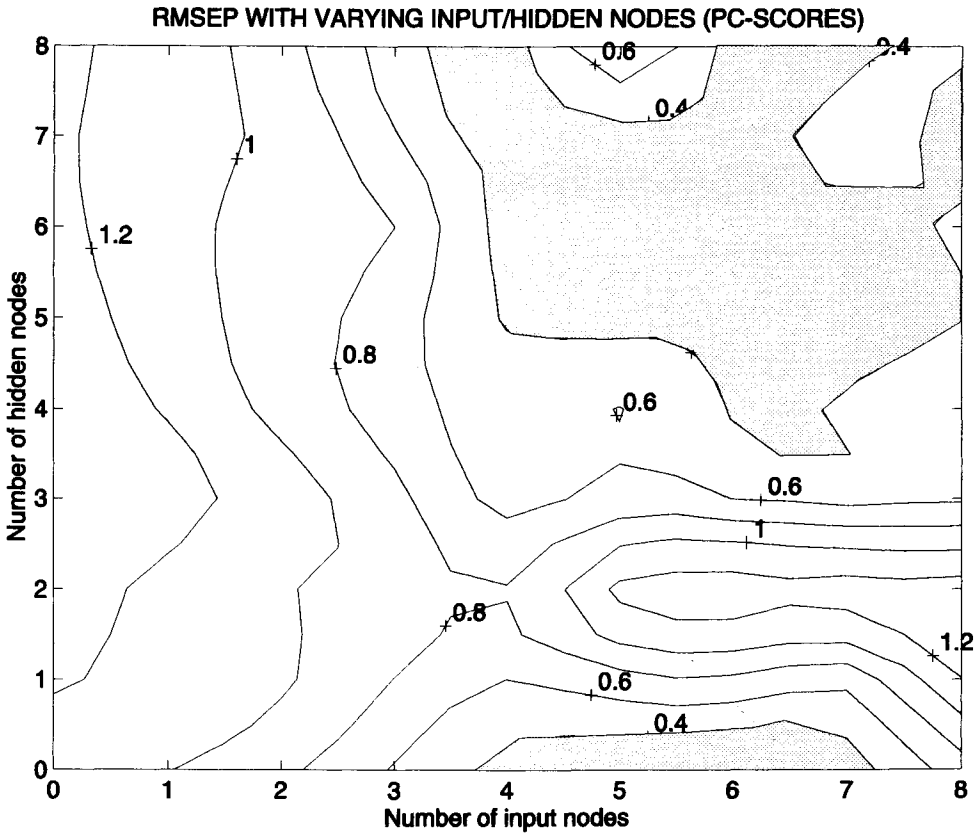


Figure 7. Input/hidden nodes optimisation. (PC-scores). Variation of the number of nodes in input layer and the hidden layer gives the resulting contour plot when the RMSEP is used as response variable. The optimal area is indicated as a shaded region.

the neural network model. The first principal components may contain information about selective sensory attributes. It is therefore possible to relate this information to the role of the input nodes.

Predicting all sensory attributes in the same network model will give an overall optimal minimum of the error function. The overall RMSEP by modelling all attributes at the same time is at the level of PLS shown in a later section. If for some reason some attributes are difficult to model, then models should be made with single predictions of attributes.

The results in Table 1 are obtained by test set validation. It should be noted that a test set validation is more overoptimistic than doing cross-validation. It is also a fact that the neural network model is more flexible than the PCR model. In many cases it seems that the test set validation thus may be more serious to the network when interpreting the results. It should be a good practice to use an independent second test set to do the validation.

Table 1
Root Mean Square Error of Prediction (RMSEP) for each attribute, where each attribute is predicted using the same model based on raw data and PC-scores and for PCR/PLS. The number of factors used in PCR and PLS modelling is given.

ATTRIBUTE	*Raw data Simultan*	*PC-SCORES Simultan*	*Raw data Separately*	*PCR (factors)*	*PLS (factors)*
Pea Flavour	0.24	0.31	0.25	0.42 (4)	0.40 (3)
Sweetness	0.35	0.32	0.33	0.38 (4)	0.37 (3)
Fruityness	0.24	0.26	0.25	0.28 (4)	0.29 (2)
Off-flavour	0.30	0.37	0.27	0.62 (7)	0.43 (7)
Mealiness	0.37	0.44	0.39	0.44 (4)	0.42 (2)
Hardness	0.24	0.28	0.22	0.39 (4)	0.37 (3)
Model RMSEP	0.30	0.34	0.29	0.43	0.38

4.2.6 Cross validation

Standard cross-validation method used in PLS/PCR may be used to validate how well the models describe the data. The main goal of our cross-validation using neural nets is to monitor the modelability of single objects. The data are divided into several test segments. The training set and test set are constructed in such way that the objects will be tested only once. Models are constructed and run until convergence. By using the scores of the combined calibration set and the validation set this would lead to a very overoptimistic model. When PC scores are being used, it is necessary to compute scores for the training set and test set for each segment model. In this way we project the validation set on the principal axes estimated for the calibration set. The network is trained to a fixed number of iterations. We choose the number of iterations where the network seems stable. The Absolute Error (RMSEP of one single object) is then calculated as a mean for all the segments at a fixed number of iterations. Cross-validation in neural network gives the ability to compare modelling performance to the PLS/PCA cross validation method. In addition, this method can be used to monitor single objects and possibly detect outliers. Neural network cross-validation may also be used to optimise the network topology.

In neural network cross-validation single output models are used. This gives simpler interpretations. Some attributes are more difficult to model simultaneously in multi output models.

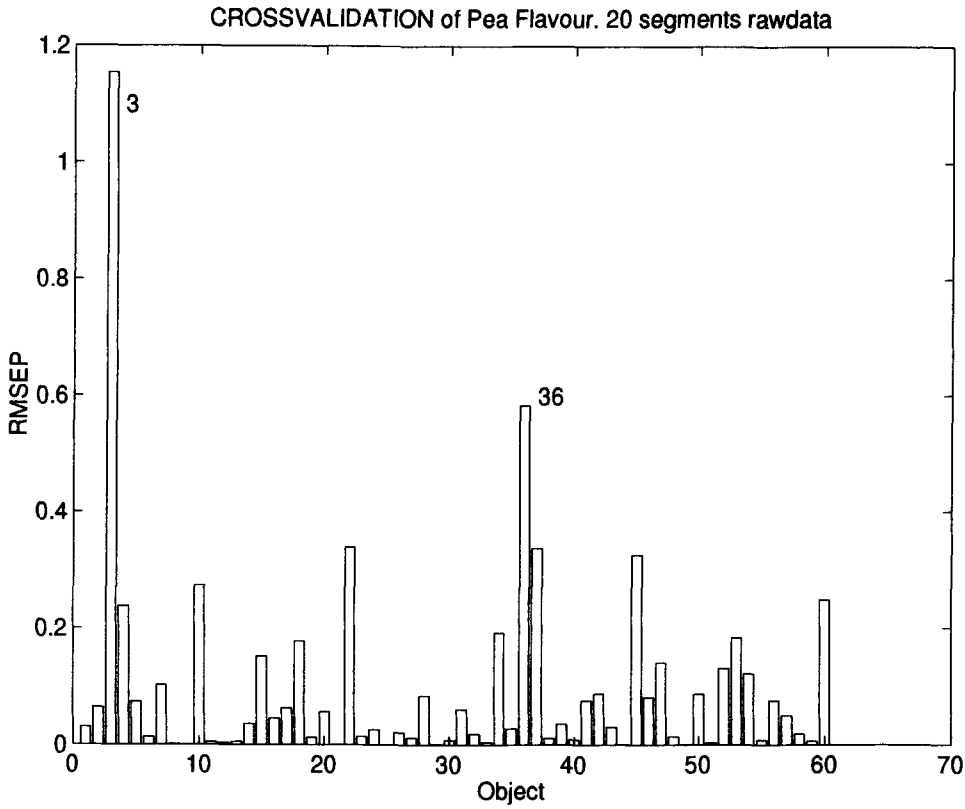


Figure 8. Residuals of each object predicted in cross-validation of 20 segments. The barplot of the Absolute Error as a function of the objects indicates the single attribute's ability to be modelled.

In the first instance, the pea flavour attribute was used as the response variable, and a 20 segment cross-validation run. This gave 3 objects in each test set and 57 objects for learning. The objects were selected randomly. The barplot in Figure 8 of the Absolute Error indicated the single attribute's ability to be modelled. If we choose objects with Absolute Error better than 0.5 we will see objects 3 and 36 to be difficult to model. If these objects fall into the outlier category is somewhat difficult to decide.

4.2.7 Results of Prediction

The results of the predictions from the raw data and the PC scores gives an indication that the raw data may be used directly. The reason for the conclusion can be seen in examining Figure 3 and Figure 4, which shows the RMSEP as a function of the number of iterations in the learning process, for both input types. Table 1 clearly shows this in the model RMSEP for the two input

types. It must be pointed out that the results shown are for single runs and not an average of several runs. The random starting point of the weights gives trained models about in the same order of magnitude.

The network has the ability to predict all the sensory attributes simultaneously. Models that predict only one sensory attribute at a time will, however, give better RMSEP in some instances and worse in other. This is demonstrated in Table 1, which shows the model RMSEP for the prediction of all attributes simultaneously compared to the prediction of each attribute separately. The overall RMSEP is, however, the same for separate and simultan prediction using raw data. Using all attributes simultaneously will give a more complex error surface and a more difficult path to find the global minimum. One will find, however, that correlated attributes might stabilise a model giving better performance for some attributes than using a single attribute model. The predictions of peaflavour using neural net gives a correlation coefficient of 0.97. This indicates a good prediction ability using NIR raw data and a neural network.

5. PCR AND PLS MODELS

5.1 Approach

Principal component regression is a tool which is used to predict one set of variables (Y-matrix) from another set of variables (X-matrix). The procedure is based on undertaking principal component analysis on the X-matrix, and then using the principal components as the predictor variables.

Partial least squares regression (PLS) (Wold, 1982) is an extension of the PCR approach, and was developed by Wold as a method for econometric modelling. It was later applied in chemometrics (Kowalski et al., 1982; Wold et al., 1983; 1984) where it has gained acceptance. Users of PLS argue that this approach is more robust than MLR or PCR, and, hence, calibration and prediction models are more stable. For a more elaborate and tutorial discussion of the PCR and PLS modelling methods we refer to Esbensen et al (1995).

As mentioned above, in a typical «relating data» problem there are two blocks of data, the Y matrix and the X matrix. In PCR a model is formulated to measure the «inner» relation between Y and X, with the aim of explaining/predicting Y. PLS also measures the «inner» relationship, but also uses «outer» relations (Geladi and Kowalski, 1986) of X and Y separately. By using the additional information on the «outer» relations, it is possible to rotate components to lie closer to the regression line, hence providing a better explanation of Y.

In this example, as previously mentioned, there were 116 variables (NIR wavelengths) in the Y-matrix, and 6 sensory variables in the X-matrix. The idea is to determine if the NIR data can be used to predict the 6 sensory variables, one at a time, and together.

In order to do this, and test the model, the same plan as for the neural network problem was used. Out of the 60 samples, 40 were submitted to the calibration matrix, whilst the remaining 20 formed the test set.

5.1.1 Performance of Models

The model was run where the NIR spectra were scaled to unit variance by subtracting the sample mean and dividing by the standard deviation for each sample. The sensory data were not standardised, as it is usual to retain the original variance structure of these data.

On running the PCR analysis where all six sensory attributes were predicted at the same time, four factors were found to provide an optimal solution to calibration model. Table 2 shows the percentage variance explained in the validation X and Y data, after calibration. The PLS model indicated that 3 factors were optimal, and the percentage variance explained in the X and Y data are provided in Table 2. It is clear that the PLS model is performing slightly better than PCR in terms of percentage variance explained.

Table 2. Percentage variance explained in the X (NIR) and Y (sensory) validation data using the PCR and PLS calibration models.

	PCR		*PLS*	
FACTOR NUMBER	***XDATA***	***YDATA***	***XDATA***	***YDATA***
1	42.2	50.7	36.6	63.7
2	64.8	66.1	63.9	86.6
3	77.2	74.2	82.9	89.6
4	93.9	89.6	93.3	89.7
5	96.5	89.3	95.4	89.5
6	97.8	90.0	97.1	90.0

PCR was undertaken, using the NIR data to predict each sensory attribute, one by one. Figure 9 shows the residual variances of some important sensory attributes. Here we see that 4 factors are needed to explain the variables. We have used the Pea Flavour as an example in our neural net part. The Pea Flavour modelled using PCR and PLS is shown in Figure 10. Here it is verified what is shown in Table 2. The PLS performs slightly better than the PCR. The number of factors needed to model the Pea Flavour is 3 using PLS and 4 using PCR. The RMSEP, however, is shown to be nearly equal for both modelling systems as shown in Table 1. The PLS is slightly better. Off-flavour is more difficult to model using PCR, but 7 factors are needed to explain this attribute using PCR and PLS. The idea behind PLS is to take the information in Y into account when the modelling is done. This approach leads to the result of lesser factors in the model than using PCR.

This verifies the results from optimisation of number of inputs to a neural network based on pc-scores. (Figure 6) In the optimal model there was a need of the 4 first pc-scores. Just as many as the optimal number of factors needed in the PCR model.

We have used a test set as validation of the modelling ability. Using full cross-validation to validate the model gives a good indication how the validation method performs. Figure 10 shows the full cross-validation compared to the test set validation in the case of Pea Flavour.

Each object is tested against all the other objects in full cross validation. The resulting RMSEP is calculated as a mean PLS modelling ability of Pea Flavour. We see that the cross-

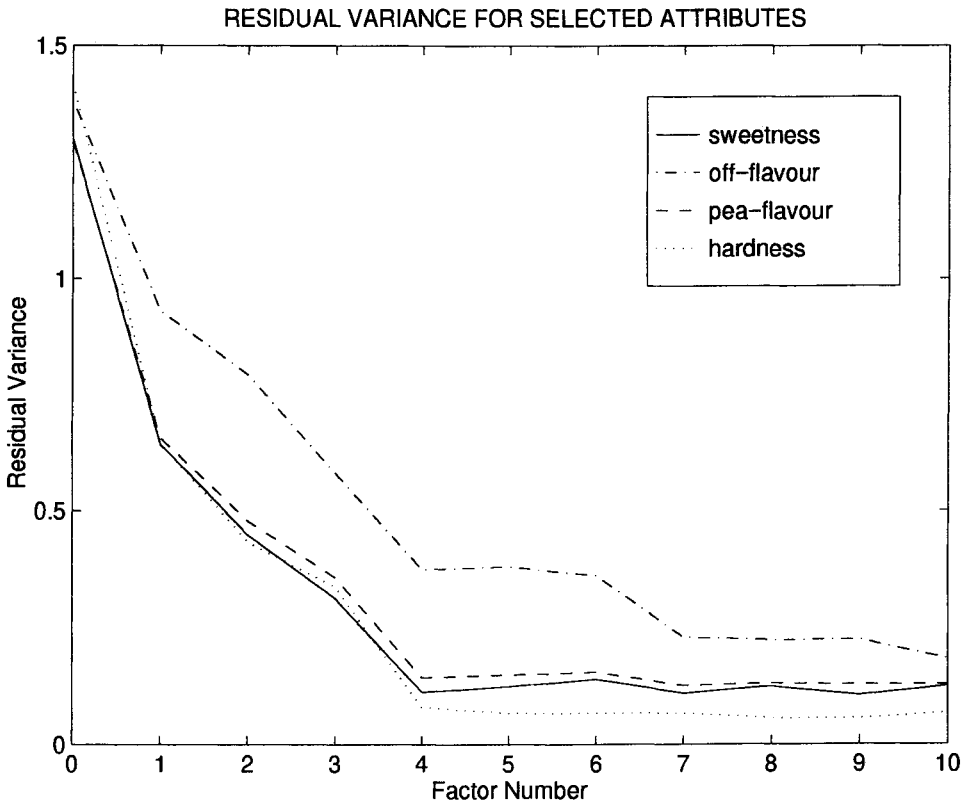


Figure 9. The residual variances of some important sensory attributes using PCR. We observe that 4 factors are needed to explain most variables. Off flavour needs 7 factors.

validation is more conservative when compared to the test set validation. The test set validation in general gives more optimistic results and it is a good practice to compare this to a full cross-validation. If there are very many objects available it is possible to use segmented cross-validation as we have explained earlier in this chapter.

5.1.2 Diagnostic tools. The biplot

The principal components estimated in PCR may be plotted in several ways. A good interpretation tool is the *biplot*. In this plot the relations between the original variables and the different objects may be resembled. The loading plot shows the relations between the variables. Figure 11 shows that the group consisting of hardness, off-flavour and mealiness are strongly correlated. The other group consisting of pea flavour, sweetness and fruitiness are correlated. The two groups are negatively correlated. The score plot shows objects that are positioned relative to these two groups. These objects are described by the relative positioning to the two

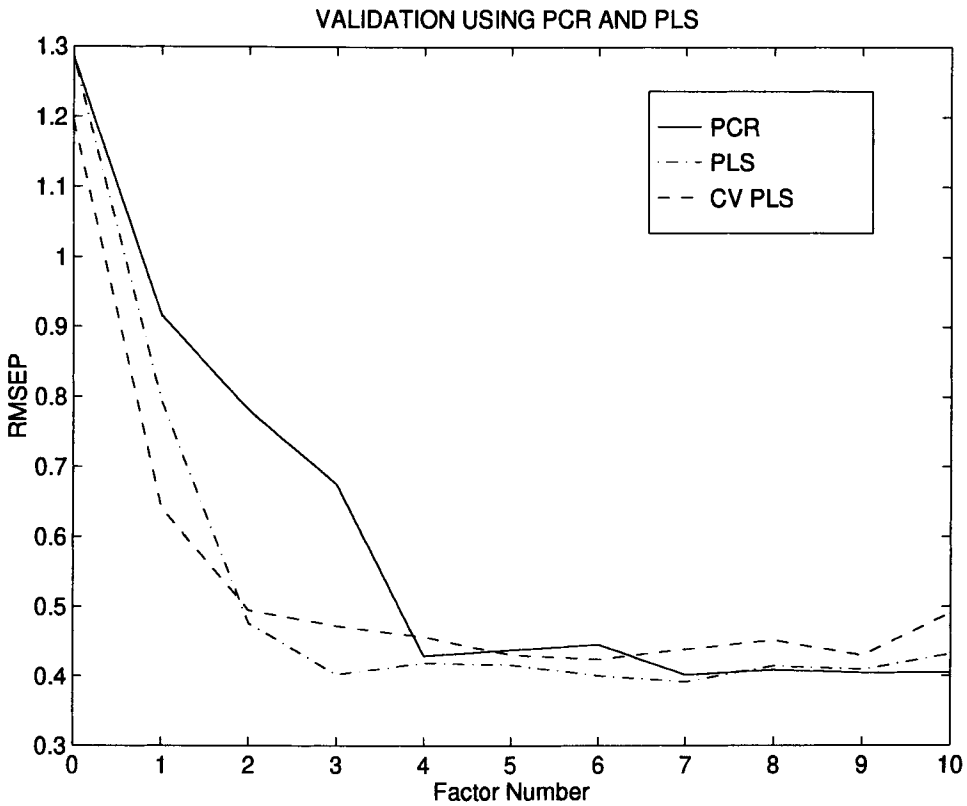


Figure 10. The Pea Flavour modelled using PCR and PLS test set validation and PLS full cross-validaton. The PLS performs slightly better than the PCR. The number of factors needed to model the Pea Flavour is 3 using PLS and 4 using PCR. The cross-validation is more conservative when compared to the test set validation.

groups. This indicates that for instance object 12 has a lot of hardness, off-flavour and mealiness while object 52 has very little of these attributes. Object 52 has very much of pea flavour, sweetness and fruitiness.

There is no correspondence to the biplot when it comes to the feed forward neural net. It is, however, other network topologies that is capable of mapping variables. Just like the principal component variable reduction, the self organising map will reduce multi dimensionality to for example two. The self organising map is designed to act as a feature map like the biplot. (Kohonen, 1988).

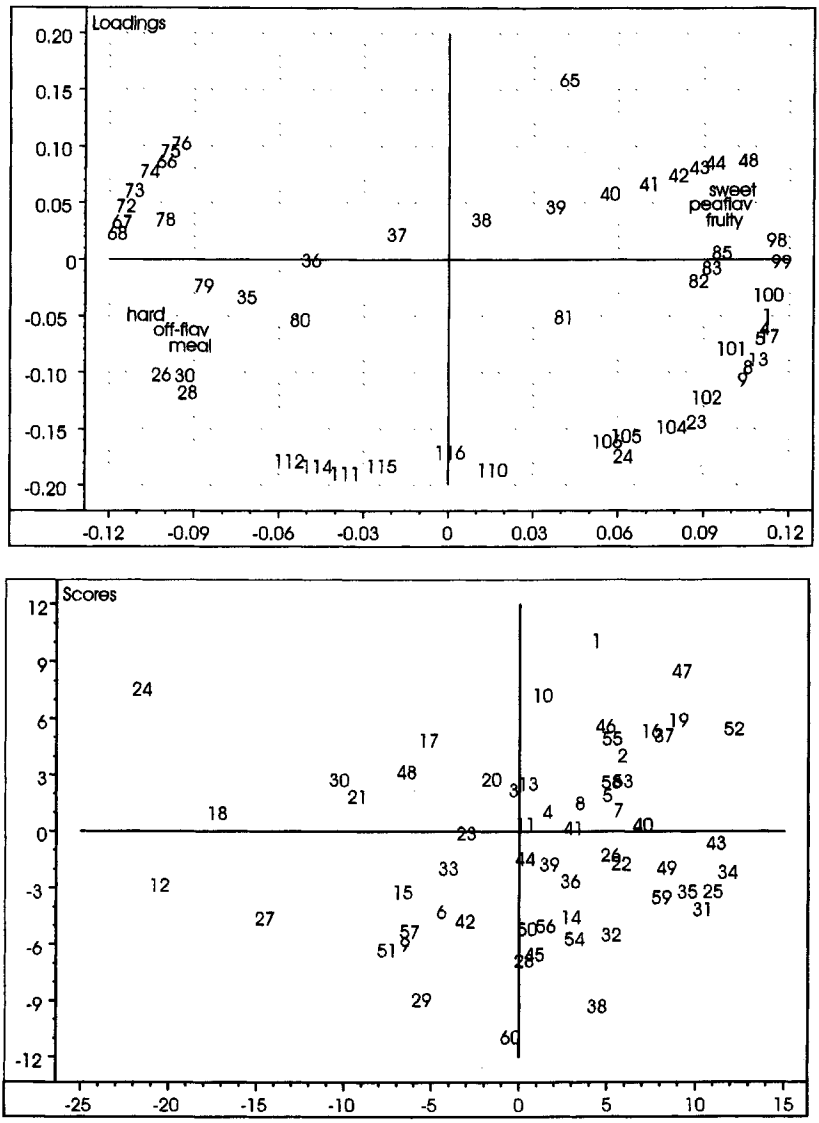


Figure 11. The PCR loading plot shows the relations between the variables (NIR variables are shown as numbers, sensory attributes are shown by name). The group consisting of hardness, off-flavour and mealiness are strongly correlated. The other group consisting of pea flavour, sweetness and fruitiness are correlated. The two groups are negatively correlated. The PCR score plot shows objects that are positioned relative to these two groups.

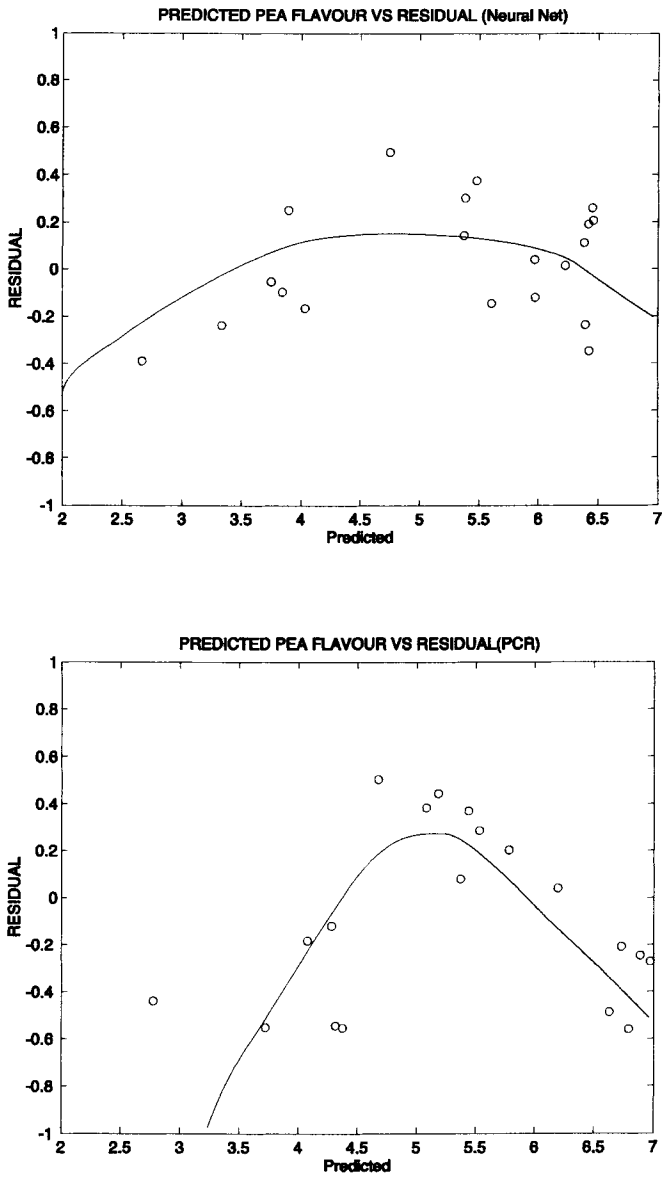


Figure 12. Neural net and PCR predictions plotted against the residual values of pea flavour. This illustrates how well the neural net models a non-linear relationship when compared to the linear PCR model.

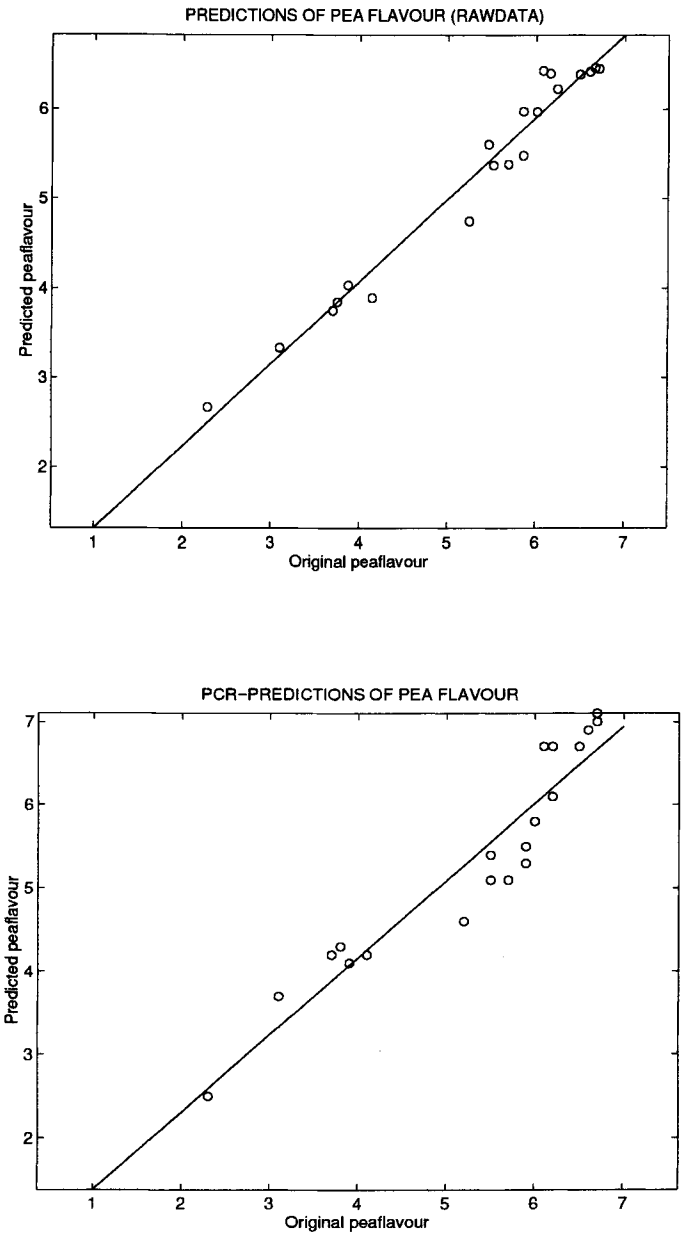


Figure 13. Original pea flavour plotted against the predicted values using PCR and neural net modelling. The correlation coefficient for the PCR model is 0.90 and 0.97 for the neural net model.

5.2 Comparison of Performance

Predictions from PCR compared with the neural network shows a non linear relationship in the data. We select the prediction of pea flavour as an example. Figure 12 shows the residual plotted against the predicted pea flavour for PCR. Here we see a non-linear curvature. Figure 12 also shows the same for neural net raw data model. The network tends to fit the model in such way that the residuals are nearly the same in the whole range of pea flavour. This shows the neural network is good in predicting non linear data relations.

Figure 13 shows the PCR and neural net predictions versus wanted values of pea flavour from raw data. The correlation coefficient estimated using the PCR is 0.90. Compared with the correlation coefficient of 0.97 for neural network predictions this shows that the neural networks is able to detect non-linearities in the data. Comparing the results in Table 1 we see that the neural network also performs better in prediction of all attributes. The non linearity in the data and the fact that neural networks performs much better than PCR/PLS indicates that neural network is a reasonable choice when it comes to predictions of sensory attributes.

6. CONCLUSIONS

Analysing complex sensory data is not a straight forward process. We have shown that using different tools gives corresponding results but with different degrees of accuracy. It is very important for the user of neural nets and PCR/PLS to understand the limitations and pitfalls. In any case it is a very good practice to have a good knowledge of the origin of the data. Neural network should be used when we need a good prediction ability. This is due to the fact that neural network is able to detect non linear relations in the data. When the network is trained it is a simple calculation task to use it as a good predictor. The good diagnostic tools of PCR/PLS are, however, very important when we want to monitor the relations between the different attributes and variables. The feed forward network is specially designed to do a pattern recognition and has its strength in classification. I lacks, however, the diagnostic tools like score plots and loading plots of PCR/PLS. A combination of neural network modelling and PCR/PLS diagnoses gives a more deep understanding of the complexity of sensory data.

7. REFERENCES

Aleksander, I. and Morton, H. (1990). An Introduction to Neural Computing. London: Chapman and Hall.

Borgaard C and Todberg H (1992). Optimal Minimal Neural Interpretation of Spectra. Analy. Chem vol 64, 545

Demuth Howard and Beale Mark (1994). Neural Network Toolbox, The Math Works inc.

Esbensen, K. Schönkopf, S. Midtgaard, T. (1995) Multivariate Analysis in Practice. Camo AS.

Geladi, P. and Kowalski, B. R (1986). Partial Least Squares Regression: A Tutorial. Analytica Chemica Acta, 185, 1-17

Hertz, J. (1991). Introduction to the Therory of Neural Computations. Addison-Wesley.

Hildrum, KI, Isaksson T, Næs T and Tandberg A. (1992) Near Infra-Red Spectroscopy. Bridging the Gap between Data Analysis and NIR Applications. Ellis Horwood LTD.

Kjølstad, L. Isaksson, T. and Rosenfeld, H.J. (1988). Near Infrared Reflectance Prediction of Internal Sensory Quality of Frozen and Freeze-Dried Green Peas. Journal of the Science of Food and Agriculture, 259-266.

Kohonen, T. (1988). Self Organization and Assosiative Memory. Springer Verlag, Berlin.

Kowalski, B. R., Gerlach, R. and Wold, H. (1982). Chemical Systems Under Indirect Observation. In: Jöreskokg, K. G. and Wold, H (Eds). Systems Under Indirect Observation, Part II, pp 191-209. Amsterdam: North-Holland.

Kress-Rogers, Erika (1993). Instrumentation and Sensors for the Food Industry. Butterworth-Heinermann Ltd.

Kvaal, K. Ellekjær, (1995) M. A Cross-validation Algorithm for the Backpropagation Neural Network. Proc. of NNNS 94. The Norwegian Neural Network Seminar 1994.

Martens, H. and Næs, T. (1989). Multivariate Calibration. Chichester: John Wiley and Sons.

Martens, M and Martens, H. (1986). Partial Least Squares Regression. In: Piggot, J. R. (Ed). Statistical Procedures in Food Research, pp 293-359. London: Elsevier Applied Science.

Masters, T (1993). Practical Neural Network Recipes in C++. Academic Press Inc.

Masters, T (1994). Signal and Image Processing with Neural Networks. John Wiley & Sons.

Neural Computing, NeuralWare Inc Technical Publicaton Group, 1991

Næs, T. and Kowalski, B. (1989). Predicting Sensory Profiles From External Instrumental Measurements. Food Quality and Preference, 4/5, 135-147.

Næs, T., Kvaal, K., Isaksson, T. and Miller, C. (1993). Artificial Neural Networks in Multivariate Calibration. Journal of Near Infrared

Pao, Y.H. (1989). Adaptive Pattern Recognition and Neural Networks. Addison-Wesley.

Piggot J R (1984). Sensory Analysis of food. 2nd Edition. Elsvier Applied Science.

Smith, P. M. and Walter, L. G (1991). Neural Networks in Sensory Perception. In: Lawless, H. T. and Klein, B (Eds). Sensory Science Theory and Applications, pp. 207-222, New York: Marcel Dekker.

Spectroscopy, 1, 1-11.

Stillings, Neil A. (1987). Cognitive Science. An Introduction. The MIT Press Cambridge, Massachusetts London, England.

Wold, H. (1982). Soft Modelling: The Basic Design and Some Extensions. In: Jöreskog, K. G. and Wold, H (Eds). Systems under Indirect Observation. Part II, pp 154. Amsterdam: North-Holland.

Wold, S. Albano, C., Dunn III, W. J. Esbensen, K. Hellberg, S. Johansson, E. and Sjöström, M. (1983). Pattern Recognition: Finding and Using Irregularities in Multivariate Data. In: Martens, H. and Russwurm, H. Jr. (Eds). Food Research and Data Analysis, pp 147-188. London: Appliend Science Publishers.

Wold, S. Albano, C., Dunn III, W. J. Esbensen, K. Hellberg, S. Johansson, E. and Sjöström, M. (1984). Multivariate Data Analysis in Chemistry. In: Kowalski, B. R (Ed). Chemometrics: Mathematics and Statistics in Chemistry, pp 17-95. The Netherlands: D. Reidel

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

RELATIONSHIPS BETWEEN SENSORY MEASUREMENTS AND FEATURES EXTRACTED FROM IMAGES

Knut Kvaal, Pernille Baardseth, Ulf G Indahl and Tomas Isaksson

MATFORSK, Osloveien 1, 1430 Ås, Norway

1. INTRODUCTION

Analysing complex sensory data is normally done by using traditional statistical tools like Analysis of Variance (ANOVA) and Principal Component Analysis (PCA). The approach of using imaging techniques and relating images of the products to sensory attributes is, however, not so common. Traditional image analysis of measuring distances, counting objects and looking for hidden phenomena in the images are used on a single variable basis. The multivariate approach by processing several images of the product simultaneously is more elaborate but has some great advantages. This work is a preliminary study of the possibility of relating sensory quality parameters of white bread baguettes to features extracted from image analysis. The success of this work will make it possible to use image analysis to optimise baking processes and select components important to achieve the optimal and best quality products. It is also an important aspect to point at the use of Singular Value Decomposition (SVD) and image analysis in on-line process control.

Sensory analysis and texture analysis of bread are traditionally performed by sensory analysis using trained assessors. The analysis is done by statistical methods like ANOVA and/or multivariate techniques. (Maximo & Singh, 1984, O'Mahony, 1986). The main problem of using the sensory techniques in process optimisation and on-line techniques is the time it takes to get the information. New technology for video cameras connected to computers have given alternative solutions to this problem. This requires fast and precise methods for extracting relevant information from the video images. Traditional image analysis of counting objects, measuring area, performing statistical analysis and combining information in several ways makes it possible to extract information to be handled in further statistical processing (Haralick, 1979). These traditional image processing techniques are, however, rigorous and needs a lot of statistical computations and human interaction (Pratt, 1991). There is an increased interest of using data transformations in pre-processing the images before they are handled by a modelling system. Special focus is put on the Fast Fourier Transform (FFT), the Wavelet transform and the Gabor transform (Masters, 1994). These transforms result in complex numbers, and the modelling needs to take this into account. The SVD is a real number transform and is therefore simpler to handle. Statistical pattern recognition and parameter estimation is handled very extensively by van der Heijden (1994). Success in using SVD in extracting features of different textures have been reported (Ashjari, 1982). Here the singular values of the image are used to identify different textures and the identification is done by

calculating the Bhattacharyya distance of the different SVD texture sets. Looking at image feature extraction in a multivariate way, however, gives new possibilities to extract relevant information in a straightforward way. The singular values estimated from the SVD algorithm identifies a *singular value spectrum* (SV-spectrum) for a particular texture sample. Different images of samples give rise to different SV-spectra. These SV-spectra are used as the X matrix in a multivariate modelling. The SV-spectra may then be modelled together with relevant Y-information like sensory attributes, process variables and image features like object area. Multivariate modelling like PCR and PLS are good diagnostic tools to monitor the hidden relations between X and Y.

We will in the first part of the paper focus on the theoretical aspects of feature extraction of images using the SVD. We will also focus on the multivariate techniques to be used in classification and prediction of sensory attributes, especially the porosity of white bread baguettes and the area of the final bread slices. The paper will also focus on the prediction of a physical measurement of the area of the bread slices. It is of special interest to see how well the area may be modelled at the same time as sensory attributes. If area is of interest as a quality parameter, this reduces the need of doing a physical measurement in addition to sensory measurements.

2. FEATURE EXTRACTION

2.1 Singular Value Decomposition

For a technical description of the SVD algorithm we refer to Press (1992). In our approach the image is considered as a matrix of pixels ordered in rows and columns. We normally consider grey scale images. Colour images can be considered as separate greyscale images in red, green and blue components.

Consider the image A of size (m x n). The SVD theorem states that there exists unitary orthogonal matrices U and V of size (m x r) and (n x r), respectively, and a diagonal matrix S of size (r x r) (where r is the rank of A) such that

$$A = U * S * V' \tag{1}$$

The matrix $S=\{s_{ij}\}$ is considered to be a generalised spectrum of the image (Hansen & Nilsen, 1983). The matrix S can be written as

$$S = \begin{pmatrix} s_{1,1} & 0 & 0 & 0 & 0 \\ 0 & .. & 0 & 0 & 0 \\ 0 & 0 & .. & 0 & 0 \\ 0 & 0 & 0 & .. & 0 \\ 0 & 0 & 0 & 0 & s_{r,r} \end{pmatrix} \tag{2}$$

where the diagonal elements are the singular values of A. The singular values are sorted in descending order. By applying the SVD on the image A this is the equivalent of estimating the principal components of A using the rows as objects and the columns as variables. The

estimation of singular values may also be done on the A' or $A*A'$ matrix as well. We will only consider the first approach (Martens and Næs, 1989).

The SV-spectrum λ of an image sample is composed of the diagonal elements of S estimated from an image:

$$\lambda = \text{diag}(S) \tag{3}$$

A reconstruction, A_p,of an image with p factors is given by

$$A_p = U_p * S_p * V_p' \tag{4}$$

and we have

$$A = A_p + E_p \quad (p<=n) \tag{5}$$

where E_p is a residual image at p factors.

The image A is fully reconstructed by applying the matrix multiplication of (1). By using less principal component factors p<n we have a situation of image compression. This is a lossy compression, but the main structure is described by the first p factors of the image represented by A_p. The number of factors to use is a matter of choice and depends on the property of the images.

Figure 1 illustrates the use of SVD as an image compression technique of the image using different factors. The illustration clearly shows that with p=1 the main information is given by the rectangle convolving the object. With p=5 the shape is being described and with p=50 the poring structure is contained in the restored image. The residuals show remaining structure where p=1 and 5 but only noise when p=50. In this work we will show that the use of less factors than n will enhance the model predictions and classification abilities. Thus the SV-spectrum of an image sample taken into account is described by

$$\lambda_p = \text{diag}(S_p) \quad (p<=n) \tag{6}$$

The SV-spectra estimated using p factors from a set of k images are described by the matrix Λ_p consisting of the λ_p' as the rows of the Λ_p.

$$\Lambda_p = [\, \lambda_{p1}, \lambda_{p2}, \lambda_{p3}, \ldots \lambda_{pk} \,]' \tag{7}$$

The singular values of the image are assumed to contain information of the image texture. The matrix Λ_p is used in classification and prediction using multivariate statistics and neural networks. In supervised classification and predictions the Λ_p is used as the X matrix and sensory and/or process variables correspond to the Y matrix. In unsupervised classifications the Λ_p is used as the data to be classified.

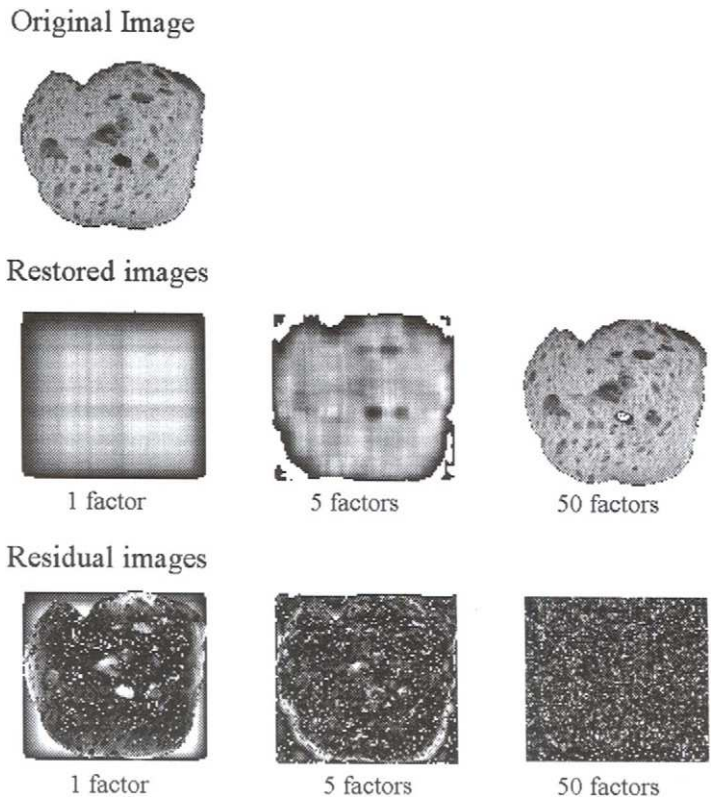


Figure 1. Example of the image compression and restoration using the SVD routine. The corresponding residual images are also shown.

2.2 Modelling techniques

Multivariate methods have been used in spectroscopy and sensory analysis with great success. Principal Component Regression (PCR) and Partial Least Squares (PLS) are evaluated as good tools in classification and predictions of chemical, physical and sensory attributes. (Kjølstad et al, 1990). The variables of the SV-spectra are like the variables of NIR spectra strongly correlated with respect to several variables. It may be possible to use univariate regression to make proper modelling of the data, but the diagnostic strength of PCR and PLS makes these modelling systems very interesting. The use of multivariate techniques enables both linear and non-linear modelling using all, or a range, of variables. The choice of using PCR/PLS is done because of their good diagnostic tools. The modelling is linear. If there is a non linear relationship in the data we could have chosen neural networks. There is a good

correspondence between these methods (Næs et al, 1993). For a good description of PCA/PCR we refer to Martens and Næs (1989). Analysing complex sensory data by artificial neural networks is described by chapter 3 of this book.

Supervised methods like PCR and PLS both have similar estimation procedures. The principal components used in PCA are solely based on maximising the variance of the X matrix. PLSR, however, uses the variation in Y also in the estimation of the components. In this way the modelling ability of PLS highly depends on the Y (Martens and Næs, 1989). Predicting unknown samples by calibration of known samples is done in a similar way by the two methods. A regression equation is obtained by regressing Y on the components. It is of interest to verify if it is an advantage to guide the modelling by taking the information in Y into account in the calibration process.

We will use a standard cross-validation technique to verify the modelling ability of the images in relation to each other. Cross-validation is performed by using all but one object as calibration set and the rest one single object as test set in a calibration. This process is done until all single elements have been used once as a test set. The cross-validation technique used validates single objects against all the other objects and the result is an average prediction ability of the model. We will also use test set validation. By using this technique we will test the modelling ability on a data set that has not been used in the calibration (Esbensen et al, 1995).

3. EXPERIMENTAL

3.1 Design

We will use texture images of baguette slices baked by different process parameters and ingredients as an example. A fractional factorial design was constructed for each flour type. The parameters varied at two levels. This resulted in a 2^{4-1} design for each of the 4 flour types giving 32 samples. The parameters taken into account were Flour type, Carlic Concentration, Mixing Time, Vitamin C Concentration and Baking Process. Special attention was given to the Baking Process and the Flour Type variables. The experiment was also performed to see the effect of garlic on the baking. The design is shown in Table 1.

Table 1. The design of the White Bread baguettes. The -1 and 1 symbolise low and high level.

SampleNo	*FlourType*	*GarlicCons*	*Mixing time*	*Vitamin C*	*Baking Process*
1	1	-1	-1	-1	-1
2	1	-1	-1	1	1
3	1	-1	1	-1	1
4	1	-1	1	1	-1
5	1	1	-1	-1	1
6	1	1	-1	1	-1
7	1	1	1	-1	-1
8	1	1	1	1	1
9	2	-1	-1	-1	-1
10	2	-1	-1	1	1
11	2	-1	1	-1	1
12	2	-1	1	1	-1
13	2	1	-1	-1	1
14	2	1	-1	1	-1
15	2	1	1	-1	-1
16	2	1	1	1	1
17	3	-1	-1	-1	-1
18	3	-1	-1	1	1
19	3	-1	1	-1	1
20	3	-1	1	1	-1
21	3	1	-1	-1	1
22	3	1	-1	1	-1
23	3	1	1	-1	-1
24	3	1	1	1	1
25	4	-1	-1	-1	-1
26	4	-1	-1	1	1
27	4	-1	1	-1	1
28	4	-1	1	1	-1
29	4	1	-1	-1	1
30	4	1	-1	1	-1
31	4	1	1	-1	-1
32	4	1	1	1	1

3.2 Methods

We will consider the images being recorded from wheat bread baguettes based on process data from 4 different types of flour and 2 different baking processes. The different types of flour had different protein content (percent) and water absorbtion abilities. The resulting bread slices were analysed by sensory analysis. The main attribute considered here were *porosity*, *firmness*, *glossiness*, *fresh smell*, *fresh taste*, *saltiness*, *crust breakage*, *juiciness* and *sponginess*. The

area of the baguette slices were calculated from images of size 512 by 512 pixels by counting the number of pixels that the bread sample covered. These areas were used as reference values in the later modelling. The modelling was performed on images resized to 128 by 128 pixels. By resizing the images, we loose some precision of the area estimate.

The images were produced using a modified standard video camera (SilvaCam). This camera has shown to be very well suited in texture recordings. The video signal was captured using a Microway 9000 frame grabber. The SilvaCam is a standard RGB camera (the output is given as red, green and blue signals) with the B channel modified to detect light in the near infrared region. The camera is constructed in such a way that the standard red/green/blue (R/G/B) channels are modified to NIR, red and green (C_{NIR} /R/G). These absorb light in the (760-900 nm), (580-680 nm) and (490-580 nm) regions respectively. The C_{NIR} /R/G components may be used separately or combined in a grey scale image with the channels averaged as (C_{NIR} +R+G)/3. It has been shown in parallel work, however, that Y/C (the ouptut is given as luminance and chrominance signals) video cameras and high quality RGB cameras give equally good results.

Different experimentation with illumination conditions lead us to use 45 degree illumination from both sides. The objects were illuminated using four tungsten lamps, two from each side. This light covers the visible and near infrared spectral region. The production of the baguettes was done on two separate days. The illumination conditions were kept as constant as possible during the recording. Calculations of the images showed, however, a little drift in light conditions giving a small luminance difference. There is also detected a slight gradient in the area distribution of light. This has not shown to be critical, but the results may possibly be better if the lighting conditions were controlled better. In an on-line situation this lighting problem may, however, be existent and robust feature extraction systems are important.

The bread samples were produced by cutting the bread in two parts. One half was analysed by the sensory panel and the other half was recorded by the camera. Thus the surfaces to be analysed were complementary and approximately the same. Twelve trained assessors were used in the sensory analysis. Special attention was given to the sensory attribute of *porosity*. Other attributes like *firmness, glossiness, fresh smell, fresh taste, saltiness, crust breakage, juiciness* and *sponginess* were also measured. Bread samples from three selected assessors were used to produce 3 * 32 = 96 images of bread slices. They were given samples from different productions based on the same design. We used bread samples presented to assessor 2, 11 and 12 in the image recording. These were chosen by random. We did not use the actual values for each acessor, but the mean values of the sensory attributes were used to produce the reference data (Y-data). The reference data for the area of baguette slices was obtained by counting pixels. This was performed using standard image analysis.

The sequence of samples corresponds to the SampleNo in Table 1. The calibration set is composed of the 64 images of baguettes presented to assessor 2 and 11 and the test set is composed of the 32 images of baguettes presented to assessor 12 as shown in Figure 2. In the full cross-validation the calibration set is construcet by selecting 95 images from the total of 96 images of baguettes. The test set is the single left image. The process of cross-validation is performed by calibrating iteratively untill all objects have been tested once.

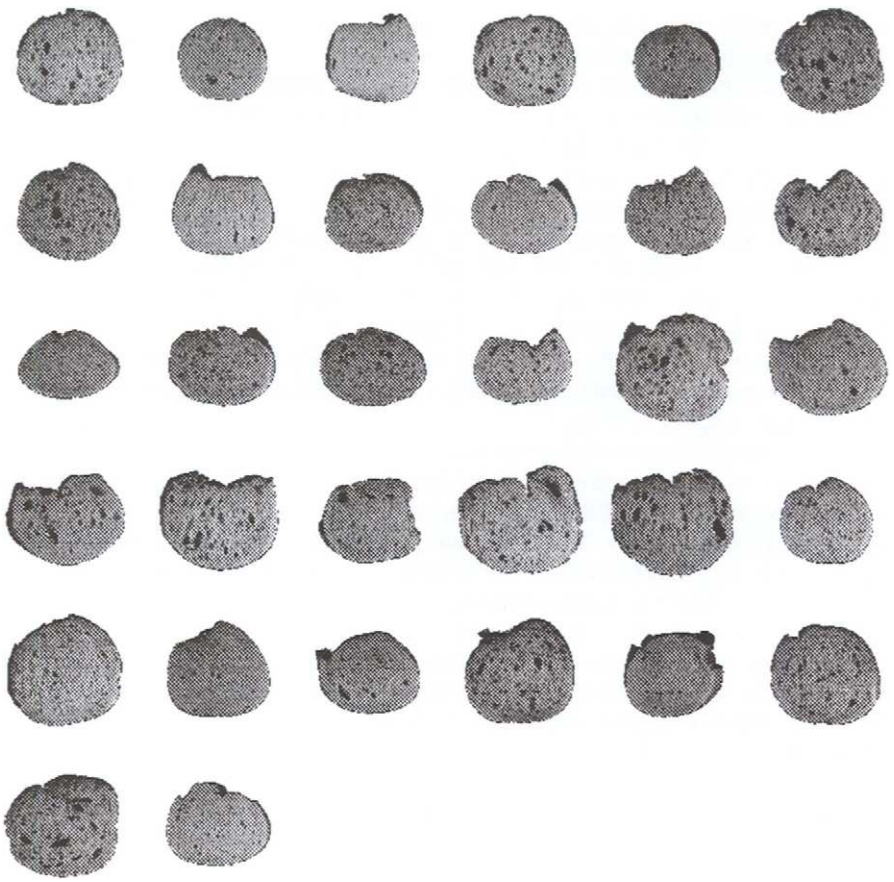


Figure 2. The images presented to assessor 12 and with the design given in Table 1. Variations in size, form and porosity are shown. The rowwise sequence of images corresponds to SampleNo in Table 1. These images are used in predictions.

The images were analysed as two main classes of images. The first class was composed of images of the whole bread object with the background filtered out (OBJ). The second class was composed of cut outs of center parts of breads (CUT). The idea was that the CUT contained texture information about porosity and the OBJ images in addition contained information on the size and shape of the baguettes. The CUT images were 128 by 128 cut outs of 512 by 512 images. The OBJ images were rescaled from 512 by 512 to 128 by 128. In this

way the computing time was reduced significantly. The main poring structure was still remaining in the OBJ images. The reason for dividing the images in these two groups was that we wanted to find out whether the information of process, flour type and porosity was given by the poring structure alone or whether additional size and shape information was needed.

3.3 Computations

Standard image analysis was performed using the ImagePro software package (MediaCybernetics, 1995). The area of the baguettes were estimated by masking the objects from the background and counting the number of pixels contained in the mask. Calculations of the singular value spectra (SV-spectra) were done using the MATLAB software package (The MathWorks, 1995). The PCR and PLS were performed using the Unscrambler software package (Camo AS, 1995).

We mainly considered the prediction of baking process, the different flour types, the area of the baguettes and sensory porosity. The area of the slices of the baguettes are easily computed by counting pixels. This process is time consuming. We wanted to show that this procedure could be included in the modelling and predicted at a reasonable level at the same time as other important sensory attributes. The area of the slice is an important quality parameter. We also considered other sensory attributes like texture, smell and taste to see how these correlated to the main variables mentioned. The modelling was done using PCR and PLS. The validation was performed using full cross-validation and test set validation (see part 2.2, Modelling Techniques). Using test set validation we used the above described 32 samples as prediction set and 64 samples as calibration set. Both the OBJ, CUT and a combination of OBJ and CUT images were considered.

Data pre-processing like sharpening was used to enhance the modelling ability of the SV-spectra. We wanted as small modifications as possible of the original images. A calculation of the greyscale representation was performed before the SV-spectra were calculated. In the CUT representation the NIR component was used. The images were mean centered to equal lighthness due to some drift in the lightning conditions. All channels (C_{NIR}/G/B) were used to produce greyscale images of the OBJ representations. The pixels were not mean centred in this representation. The process of doing the data pre-processing was performed by a rather pragmatic optimisation. By trying out different pre-processing techniques and then doing the modelling we found the method which seemed to be satisfactory. It is therefore possible to enhance the modelling ability by looking at this topic in later work.

The porosity scale ranged from 1 to 8 with 1 as the densest value. Porosity was predicted using PCR and PLS. The area of the baguettes could range from 0 to $26*10^4$ pixels but in practice they waried between $5*10^4$ and $12*10^4$ pixels. Other sensory attributes ranged from 1 to 9 on a nominal scale. The sensory attributes were all calculated using the mean value taken over all 11 assessors. The -/+ 1 values of the baking process variable were coded in a binary way (0 and 1). The flour types were coded by 4 binary variables.

The calculations were first performed in two steps to estimate the number of variables (factors) to use from the SV-specta. The SV-spectra were estimated using equation (6). The number of factors used to restore an image depends on what degree of loss is accepted. How well the attributes are modelled depends on the number of factors to be used in constructing the SV-spectra. Equation (5) gives an estimate of the residual image with p factors. This is visualised in Figure 1 where the residual images based on some selected factors are shown.

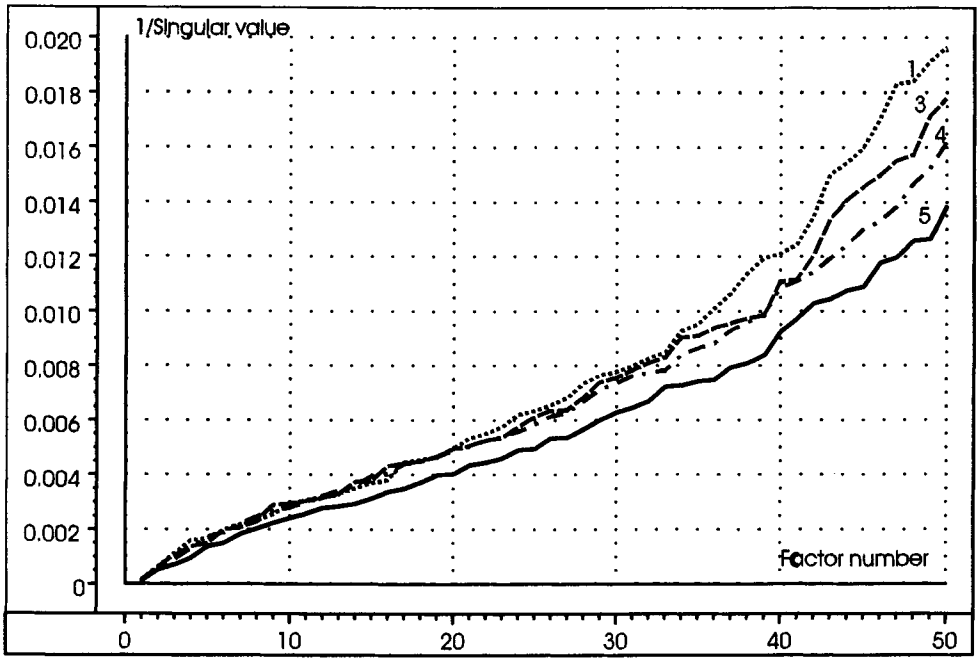


Figure 3. Samples of some estimated SV-specra (1/singular value) from the images presented in Figure 2 are shown in Figure 3. Only the first 50 factors are shown.

Samples of some estimated SV-spectra (1/singular value) from the images presented in Figure 2 are shown in Figure 3. Only the first 50 factors are shown. This indicates that the SVD produces distinct SV-spectra to be modelled.

By looking at the image structures visually in Figure 1, we observe that the porosity structure is gradually becoming visible as more factors are being used. The residuals, however, gradually become more noisy as more factors are being used. Other structures may still be visible in the residual image. We suggest the following method to estimate a reasonable number of factors, p, (variables) to be used from the SV-spectra.

Sensory porosity was used as the Y-variable. Cross-validation was performed by varying the number of SV-spectra variables. We used 10 segment cross-validation and random object selection because this is faster than a full cross-validation and the results will be at the same level. The RMSEP of the cross-validated models were compared to find the optimal model which in turn gives the optimal number of variables to use from the SV-spectra. Figure 4 shows the RMSEP of cross-validation for different number of variables. We observe that there is an optimum at about 90 variables. This optimum is reached after 2 components. Using more

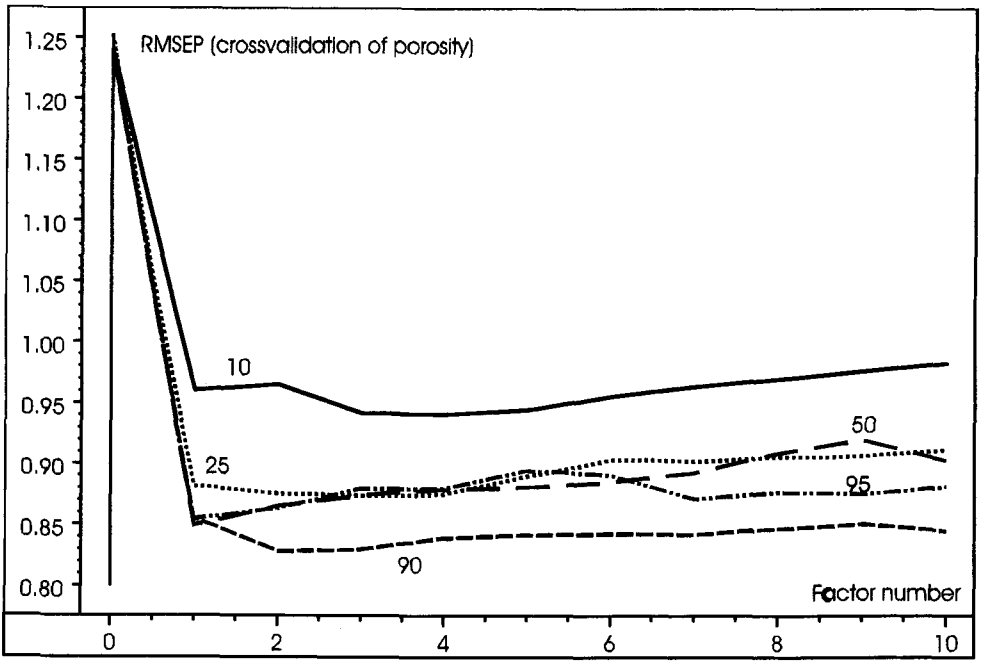


Figure 4. RMSEP for porosity (cross-validated) for different OBJ PCR models. The number of singular values (SV-variables) used in the different models are shown.

than 95 variables introduced only noise. This situation is the same for all other Y variables to be modelled and we conclude that it is reasonable to select the SV-variables 1 through 90.

This illustrates two different noise levels: The Image Level Noise (ILN) is the residual found by the feature extractor (SVD). The ILN should not be input to the PCR/PLS modelling of the sensory attributes and process variables. The Multivariate model Level Noise (MLN) is the residual of the modelling of attributes and process variables using PCR/PLS. By reducing the ILN in front of the modelling, it is possible to enhance the model performance.

4. RESULTS AND DISCUSSION

4.1 Feature extraction and multivariate modelling

The results from PCR and PLS are of the same order and we will focus the discussion on the PCR case only. We will focus on the determination of sensory porosity, the area of bread slices, the baking process and the flour types. Garlic concentration, vitamin C concentration

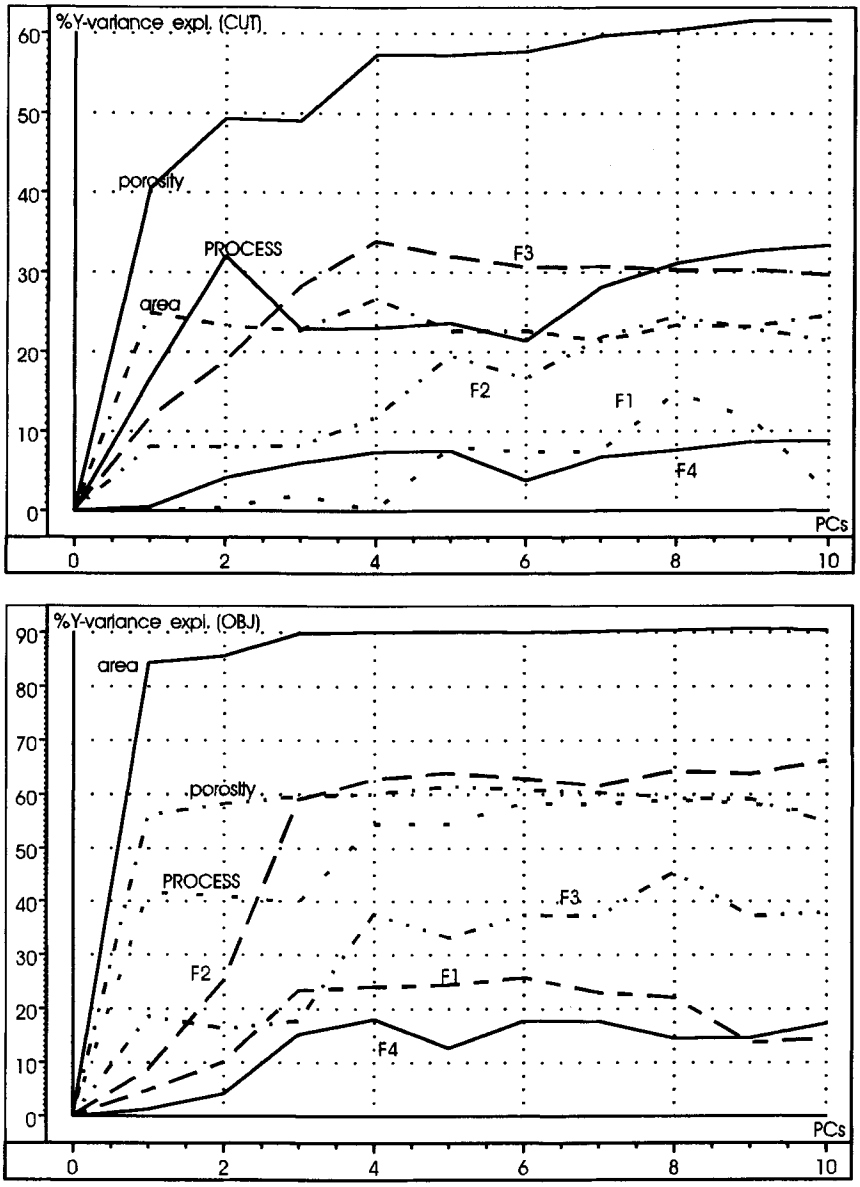


Figure 5. Explained variances of the CUT and OBJ PCR models (test set). The flour types are symboled by F1-F4.

and mixing time had no influence on the texture. These variables are not included in this work. We will also discuss other relevant sensory attributes.

The suggested number of factors to be used is given by looking at the residual Y-variances. The choice of using the test set residual Y-variances as measures for choosing the number of factors, is justified by the fact that test set validation is in correspondence with the full cross-validation. The explained variances of PCR performed on CUT and OBJ data sets are shown in Figure 5. We see that the porosity was modelled equally well using either CUT or OBJ images and the area was modelled best using the OBJ images. The explained variance of the porosity was 60 percent at 3 factors using the OBJ images and 57 percent at 4 factors using the CUT images. The explained variance of area was 25 percent at one factor using CUT images and 90 percent at 3 factors using OBJ images.

The prediction ability of area and porosity are given as root mean square error of predictions (RMSEP) (Martens and Næs, 1989) and the correlation, R, (estimated by The Unscrambler) in Table 2. The RMSEP gives a direct measure in the attributes units of the model error while the correlation factor measures the linear relationship between the measurements and the predicted values. The results show that due to the correlation of area and porosity it is possible to predict the area at a correlation of 0.5 from images that do not contain the shape of the bread (CUT). The corresponding RMSEP is $1.42*10^4$ pixels. The area, however, is predicted with a correlation of 0.95 when the information of size and shape is represented in the images (OBJ). The corresponding RMSEP is $0.52*10^4$ pixels. The prediction of porosity has correlations at the same level (0.76 and 0.78) by using the CUT and OBJ image classes respectively. The corresponding RMSEPs are 0.80 and 0.77. This verifies the results given by the explained variances.

Table 2. RMSEP and correlation coefficient for different models and image classes. CV means full cross-validation.

IMAGE CLASS	*VALIDATION*	*AREA * 10^4 (pixels)*		*POROSITY*	
		rmsep	corr	rmsep	corr
CUT	test set	1.42(1)	0.50	0.80(4)	0.76
	cv	1.30(2)		0.84(4)	
OBJ	test set	0.52(3)	0.95	0.77(3)	0.78
	cv	0.51(3)		0.84(2)	
CUT/OBJ	test set	0.52(5)	0.95	0.67(3)	0.84
	cv	0.60(5)		0.72(4)	

The modelling ability of the process variables (baking process (PR1 and PR2) and flour types (F1, F2, F3 and F4)) are shown in Figure 5 for the PCR models of OBJ and CUT images. The explained variances of the two baking processes are equal as a result of the binary design of PR1 and PR2 and symboled by PROCESS in Figure 5. The explained variances of flour types 2 and 3 are 60 and 37 percent at 4 factors using the OBJ images. The modelling of these variables is more complicated using the CUT images. The explained variances of flour type 3 is

35 percent at 4 factors. For flour type 2 it is 20 percent at 5 factors. The modelling ability of flour type 2 and 3 is reversed for the two image classes.

The explained variances of flour types 1 and 4 are 20 and 18 percent at 4 factors using the OBJ images. The modelling ability of these variables is not so clear using the CUT images. A full cross-validation shows that very little of these variables is explained using the CUT images. This indicates that the OBJ images are better in describing the flour types. It is probably the size and shape that mainly contributes to this information.

The variables describing the baking processes 1 and 2 are best described using the OBJ images. About 30 percent of the variance is explained at 2 factors with the CUT images. As much as 55 percent of the baking process variance is explained at 4 factors with the OBJ images. It seems that the modelling of the baking process needs information of the size and shapes. This is also true for the modelling of flour types.

4.2 Classifications of process variables and sensory quality

The relations between the variables are shown using the PCR loading plot of both X and Y loadings of OBJ class images in Figure 6. The relations between the different bread slice samples are shown using the score plot. The first component PC1 accounts for 85% of the variation. The second and third components account for 4 % of the variation. The corresponding loadings and scores for the CUT class images show a similar structure and we will restrict the discussion to the OBJ class images.

There is a tendency that the porosity correlates strongly to the area of the baguettes. This is expected and this also shows that it is possible to predict the area of the baguettes using the CUT class images that do not visually contain size and shape information. The baking process 2 is negatively correlated to the area and porosity along PC1. We see from the score plot that the baking process 1 is correlated to area and porosity. The flour types are separated in the loading plot. We see that the two flour types 2 and 3 are negatively correlated to each other along PC1. Along the PC1 this shows that flour type 2 gives smaller and dense baguettes and flour type 3 gives bigger and less dense baguettes. The flour types 1 and 4 do not seem to vary along the PC1 but are separated from flour types 2 and 3 along PC2. This is also the case when PC1 is plotted versus PC3.

The score plot in Figure 6 is used to classify the objects baked with different baking processes. Objects baked with process 1 are mainly located in the right half plane separated by the PC2 axis while objects baked with process 2 are mainly located in the left half plane. There is a zone in between consisting of both types of objects. The loading plot in Figure 6 shows that the right half plane describes locations of objects with high degree of porosity and big area slices. This shows that bread baked with process 1 resulted in relative bread slices with relatively large area and with a high degree of porosity. The bread baked with process 2, however, resulted in a low degree of porosity and they were small. This plot gives a good visual map of the resulting quality of the bread samples due to the two different baking processes. Objects labelled with the different flour types are also shown in the same score plot. We see a distinct separation and classification of flour types 2 and 3 in the PC1 direction. Samples with flourtype 3 are mainly located in the area where big samples with high degree of porosity are located. Samples with flour type 2 are mainly located in the area where small samples with low degree of porosity are located. If large area slices and high degree of

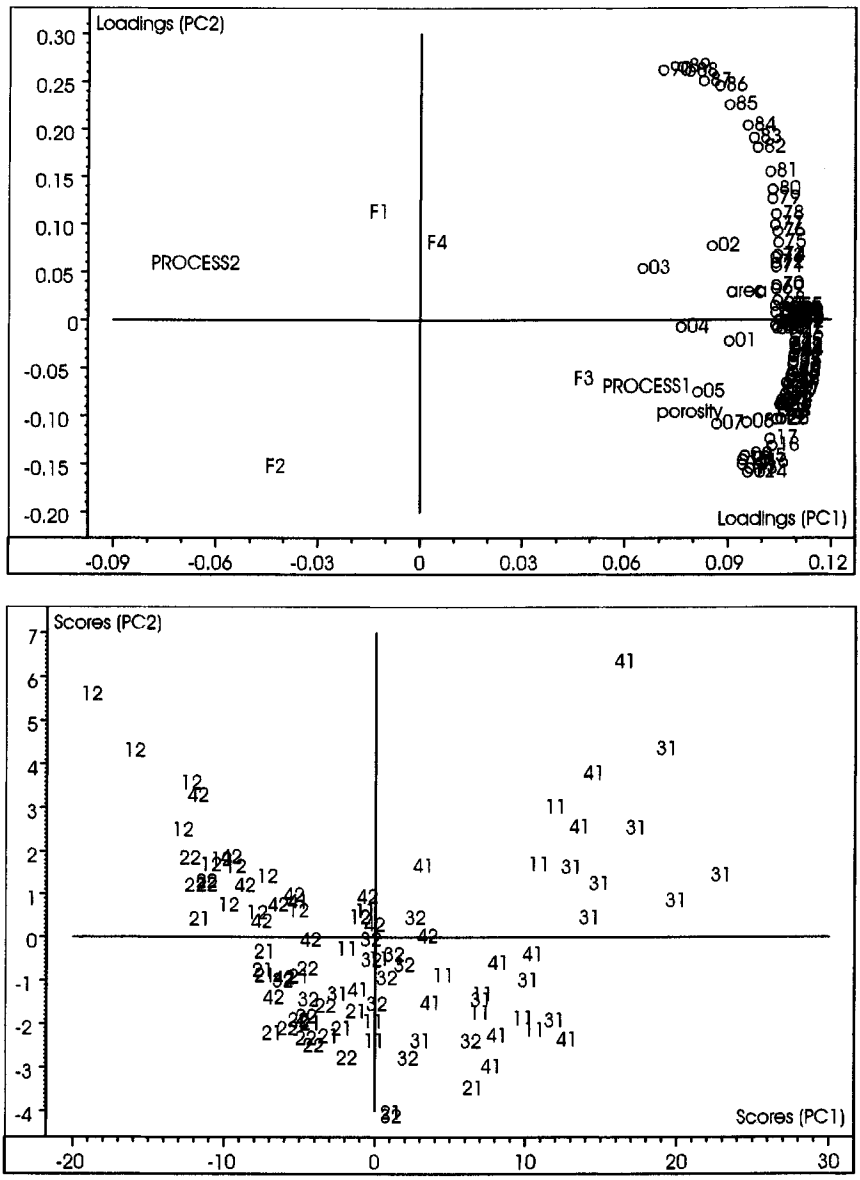


Figure 6. The loading and score plot of the OBJ PCR model. The loading plot shows the X and Y variables. The datapoints are shown by a two character symbol: Flour type (1-4) and baking process (1-2). The variables of the SV-spectra are labelled oxx. The SV-variables are shown as oxx in the loading plot.

porosity are preferred, we conclude that flour type 3 bread gives high quality products despite of baking process while flour type 2 gives low quality products despite of baking process.

The flour types 1 and 4 are separated from flour types 2 and 3 in the PC2 direction. The area and porosity of breads based on flour types 1 and 4 seem to be highly dependent on the baking process. Again, if large area slices of bread with high degree of porosity is preferred then we must use baking process 1 when dealing with flour types 1 and 4.

4.3 Combined image models

The above analysis has shown that the porosity is modelled better by using CUT class images. The area is best modelled by the OBJ class images. This may be due to the fact that the OBJ images have lost some of the information on porosity when they were reduced in size. This was, however, necessary because of the huge amount of computing power needed on large images. Ideally one should use images with good resolution for both porosity and shape. This leads to arranging CUT and OBJ class SV-spectra in one matrix resulting in one model.

The plots of the explained variances shown in Figure 7 indicate the advantage of this approach. The area and porosity are best described using 5 and 3 factors respectively. The porosity is now explained better. The flour types 2 and 3 are described using 5 and 6 factors respectively. The flour types 1 and 4 are best described using 5 factors. We observe that flour

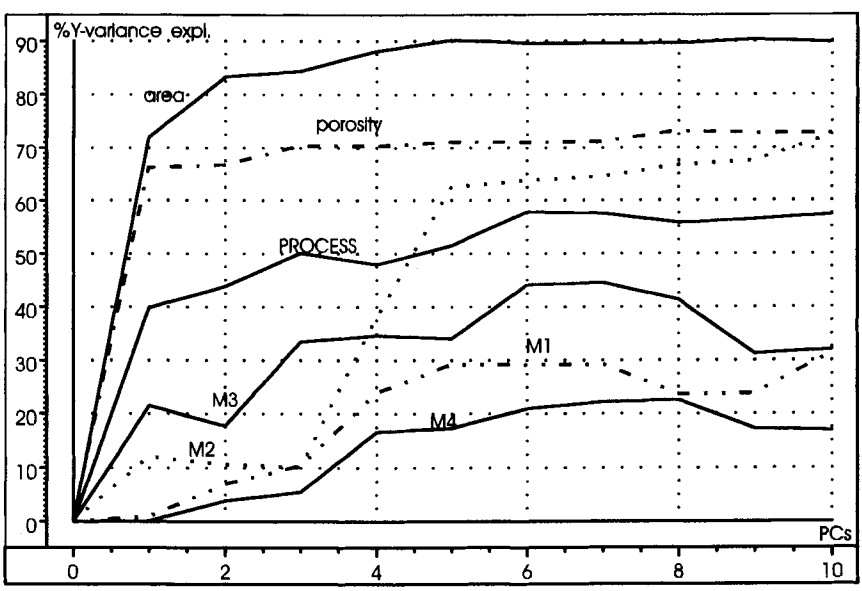


Figure 7. The explained Y variances of the combined CUT/OBJ PCR models. The baking processes PR1 and PR2 are symbolised by PROCESS and flour types (F1-F4).

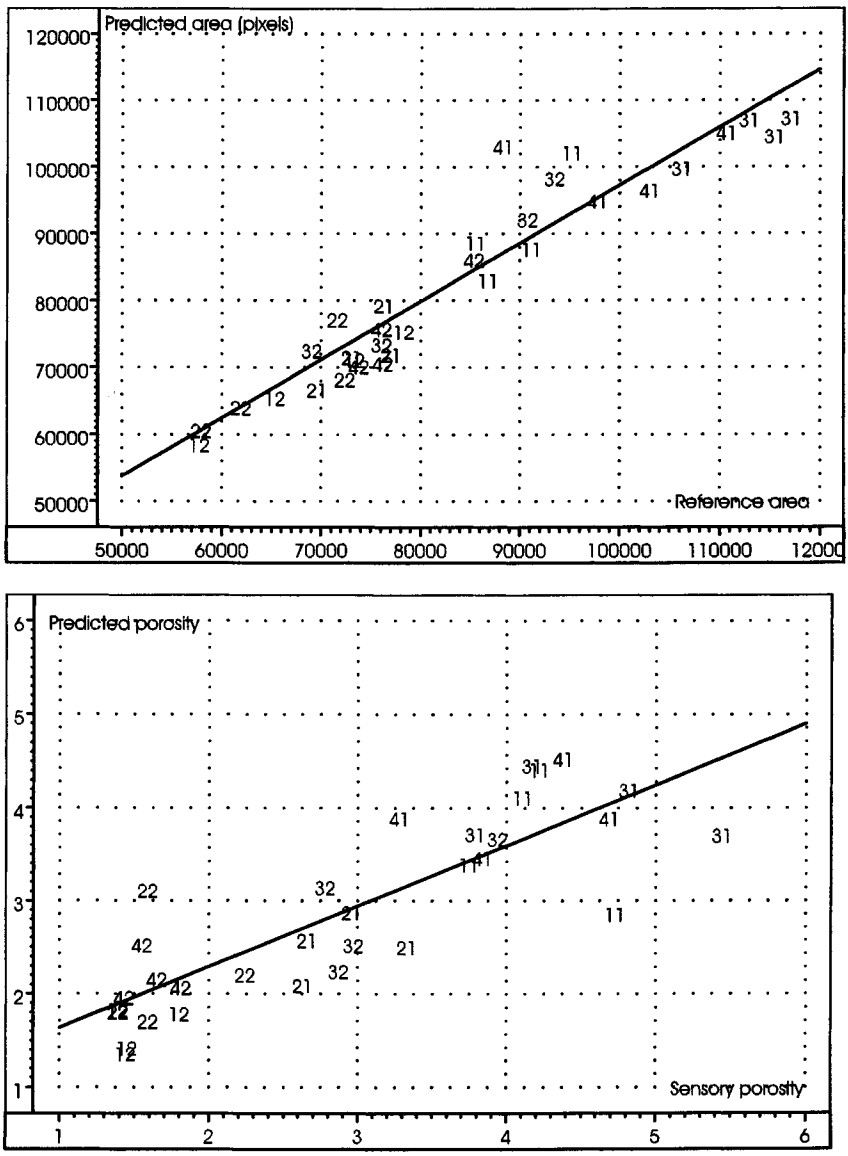


Figure 8. Predictions of area and sensory porosity of the baguettes using the combined CUT/OBJ PCR model. The datapoints are shown by a two character symbol: Flour type (1-4) and process (1-2).

types 1 and 4 now are being explained better than models based on either CUT or OBJ class images. The baking process variables are quite well described using 6 factors.

The porosity is predicted better than by the former CUT class model. The correlation is 0.84 and the RMSEP is 0.67. The area is predicted with the same degree of precision as before. This is shown in Table 2. Figure 8 shows the predictions of area using the test set and a 5 factor model is also shown. The prediction ability of the porosity using test set and a 3 factor model. The labelling of the objects show the flour types (1,2,3,4) as first digit and the baking process (1,2) as the last digit. We observe a classification of bread samples corresponding to the score plot shown in figure 6. This indicates that bread slices produced with baking process 1 are big area bread samples and have a high degree of porosity. Flour type 2 samples are grouped at the lower left and flour type 3 samples at the upper right (Figure 8).

The computing power needed for this suggested combined modelling is considered not to be critical. By splitting the image information in a texture part (CUT) and a size and shape part (OBJ), it is possible to model more features simultaneously. Most of the computing power is for the SVD. Modifications of the SVD routine to compute only the p first factors may be needed. It is then possible to implement larger images that contain more information.

4.4 Sensory analysis based on images

Strong correlations between several sensory attributes make it possible to model several interesting sensory attributes simultaneously.

The loading (X and Y variables) and score plots of the combined PCR model show interesting features of the OBJ/CUT model (Figure 9). The first principal component (PC1) accounts for 62% ov the variation while the second and third principal component (PC2 and PC3) account for 22% and 5 % respectively. The score plot in Figure 9 shows that the OBJ images have information describing the area and porosity. It also shows the correlations between area, porosity, flour type 3 and baking process 1. The CUT images, however, are mainly located in the negative PC1 direction. The first CUT variables (c01-c10) are located along the positive PC1 direction and correlates positively to the baking process 1 and flour type 3. They also vary together with the flour type 2 and baking process 2 in the negative PC1 direction. None of the OBJ variables are located along the negative PC1 direction. This shows that the combination of CUT and OBJ images may enhance the modelling ability.

The process and area variables vary along the PC1 direction. It is not possible to separate the flour types 1 and 4 in the process-area direction. They are slightly separated in the PC2 direction. Flour types 2 and 3 are separated in the PC1 direction and also separated by the PC2 direction. This leads to a suggestion that the first principal component is a baking process/area/porosity dimension. This direction is spanned by the two different image classes.
By adding other sensory variables to the Y matrix we obtain a very informative map of the relations between process variables and the sensory attributes. By the strong correlations of these attributes, it is possible to model taste and smell using SVD and image analysis. In addition to the area and porosity we see that the sensory attributes of juiciness and sponginess are described along the PC1 direction. This is reasonable when compared to the large area samples with high degree of porosity. Firmness is described along the PC2 direction. We observe that firmness is mainly described by the CUT class images. Attributes like fresh taste

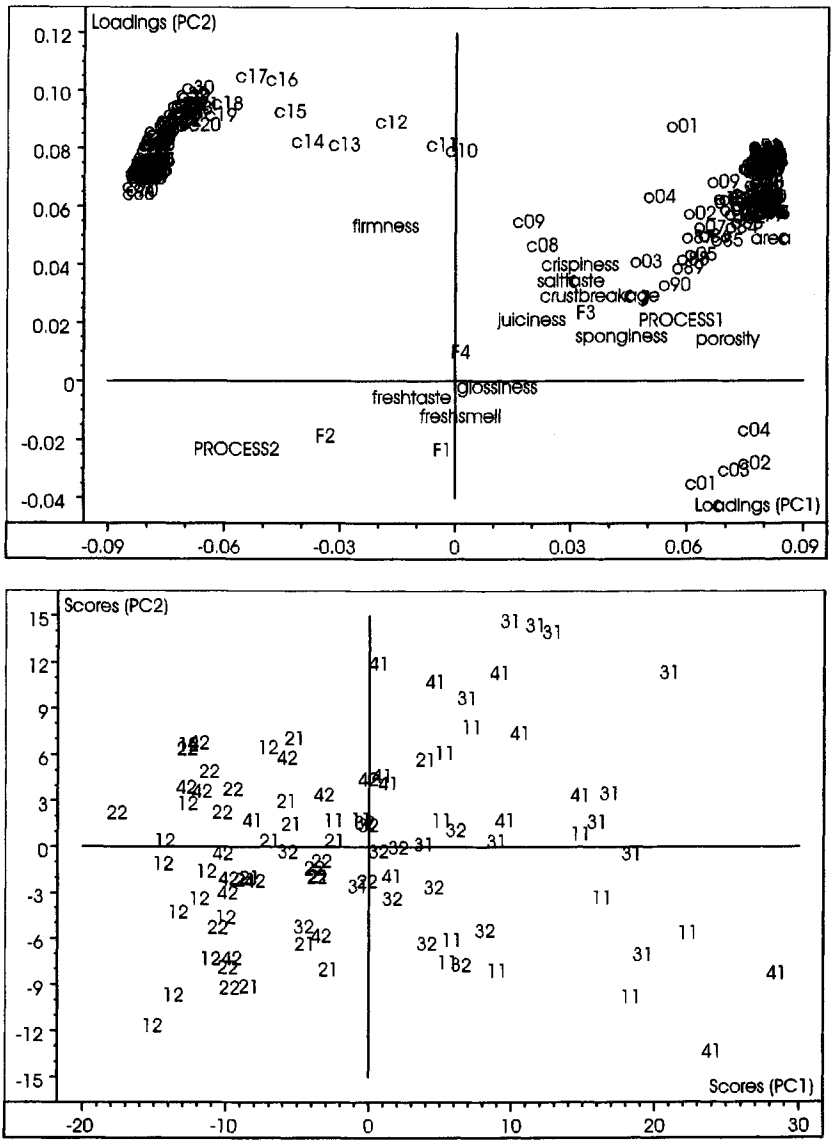


Figure 9. Loading and score plot of the combined CUT/OBJ PCR model. The datapoints in the score plot are shown by a two character symbol: Flour Type (1-4) and process (1-2). In the loading plot the characters oxx and cxx symbolise OBJ and CUT SV-spectra variables respectively.

glossiness and fresh smell is not described in the PC1/PC2 plane. They are described better in the PC1/PC3 plane.

It is interesting to see how the sensory attributes like firmness, glossiness, fresh smell, fresh taste, saltiness, juiciness and sponginess are all being modelled by the SV-spectra (Figure 10). We observe that even though it is quite impossible to detect taste and smell by imaging, the strong correlation of smell and taste variables to the texture makes this possible. It is also possible that other textural properties are related and described by the SV-spectra of the images. The explained variances of some selected attributes are shown in Figure 10. It may be possible that other feature detectors may be better to detect these attributes.

The loading plot also indicates that the flour type 2 gives products much firmness. The flour type 3 may result in products that are crisp and have high degree of crust breakage. (Crust breakage describes the breakage of the outer crust due to the cooling).

This shows an interesting feature in this modelling. By combining sensory analysis and video images, it is possible to build models to be used in on line control of the baking process. In a product optimisation this should be a valuable tool and give a good indication of how to vary the process to obtain an optimum. The proposed method also has a potential of suggesting what ingredients and baking process one should choose to achieve an optimal product. The PCR combined loading and score plot, the *biplot*, is a very useful tool in this process.

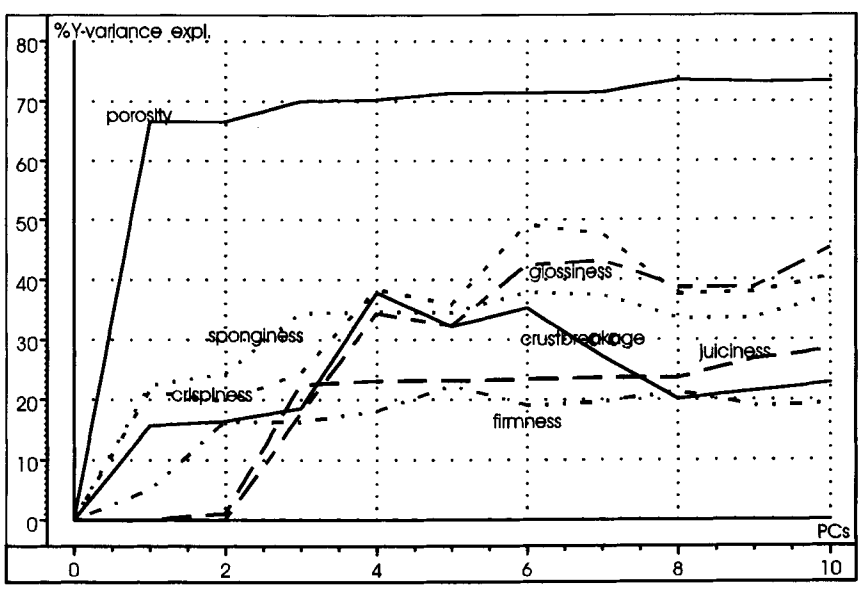


Figure 10. Plot of the explained Y variances of some sensory attributes by applying additional sensory attributes using the combined CUT/OBJ PCR model.

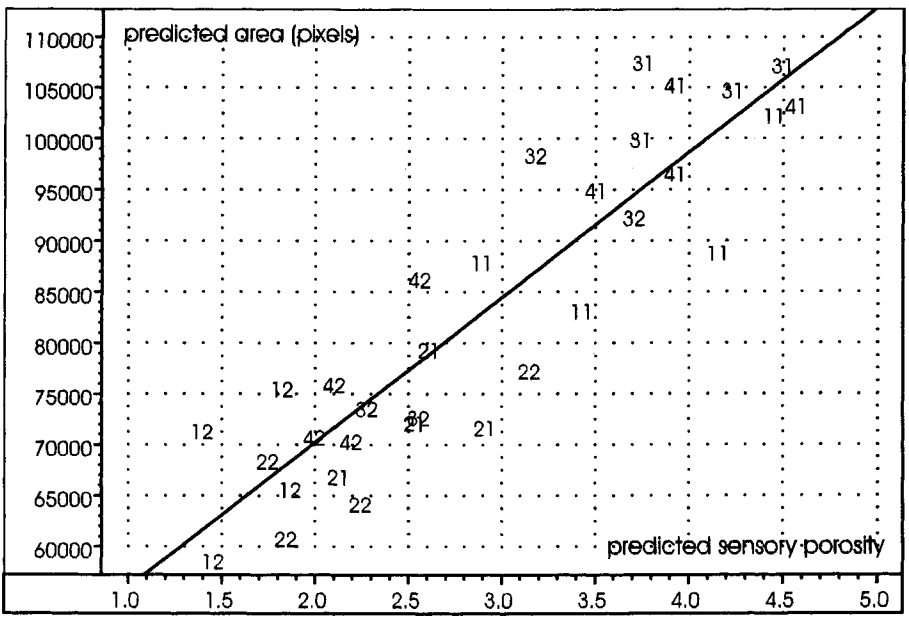


Figure 11. Plot of predicted area versus predicted sensory porosity using the combined CUT/OBJ PCR model. The datapoints are shown by a two character symbol: Flour type (1-4) and process (1-2).

4.5 Alternative predictions of porosity

The prediction ability of porosity and the area of the baguette samples show a strong relation. Figure 11 shows predictions of porosity versus the predicted area using the combined CUT/OBJ model. This shows that it is possible to estimate the porosity from the predictions of the area and vice versa with a correlation of 0.90. Univariate modelling may be used in situations where precision is not so important and we want quick estimates of the relationship. It is also interesting to observe the close relation of the first factor of the SV-spectra to the area and the porosity as well. It is possible to get a quick estimate of the relationship by using univariate regression and the first component of the SV-spectras. The main preference is that multivariate modelling of the images using the SVD also simultaneously models process variables and results in classifications that are very informative.

5. CONCLUSIONS

The SVD algorithm has proved to be a possible feature extractor and a fundamental building block in correlating raw images of bread to sensory attributes, especially the porosity. Strong correlation between the sensory attributes give possibilities to model other sensory attributes like smell and taste. It suggests a way to estimate the area of the final products based on the process variables. The models show a strong correlation between porosity and the size of bread samples. It is also possible to observe the effect the baking process has to the different ingredients of flour types. This leads to the conclusion that this method is possible to use in on-line processes control.

Large images limit the practical use of the SVD. By combining images based on porosity and texture information with images based on both texture and shapes, it may possible to enhance the modelling ability. It is also possible to classify baking process. This information can possibly be used to optimise the products.

The SVD algorithm shows its strength in using raw images with no filtering applied. Standard image analysis often needs a pragmatic way of finding the optimal filter to extract the wanted features of the images. The SVD technique uses SV-spectra and models these with multivariate PCR or PLS. Used in on-line processes this is important because of the speed of computations. It may be possible to optimise the modelling by further investigation of the Image Level Noise and multivariate modelling techniques. We will also point at the possiblity of combining several feature extractors to be used in the modelling.

6. ACKNOWLEDGEMENTS

We want to thank Grethe Enersen and Bjørg N. Nilsen for help during the image recordings. Without the help of the bakers Alf Nielsen and Leif A. Fardal the project would not have been possible. Many thanks to the people of the sensory panel at MATFORSK for doing the sensory measurements. We also want to thank Ellen Mosleth and Marit R. Ellekjær for valuable discussions and Kim Esbensen, SINTEF, for kind permission to use the SilvaCam Camera.

7. REFERENCES

Ashjari B., Singular Value Decomposition Texture Measurement for image Classification. University of Southern California, Department of Electrical Engineering, Ph.D. Thesis (1982).

Camo AS: The Unscrambler ® version 5.5, (1995).

Esbensen, K., Schönkopf, S. and Midtgaard, T. Multivariate Analysis in Practice. Camo AS (1995)

Hansen, P. C. and Nilsen H.B. Singular Value Decomposition of Images. Proc. of The Third Scandinavian Conference on Image Analysis, Copenhagen, pp 301-307 (1983).

Haralick R. M. Statistical and structural approaches to texture. Proc. IEEE. Vol 76, pp 786-804 (1979).

Heijden, F. van der. Image Based Measurement Systems. John Wiley & Sons (1994).

Kjølstad, L.,Isaksson, T. and Rosenfeld, H.J. Prediction of Sensory Quality by Near Infrared Reflectance of Frozen and Freeze Dried Green Peas. J. Sci. Food Agric, (1990) 51, 247-260 .

Martens, H. and Næs, T. Multivariate Calibration. John Wiley & Sons (1989).

Masters, T. Signal and Image Processing with Neural Networks. John Wiley & Sons. (1994)

Mathworks Inc: MATLAB for Windows ®, version 4.12c, (1995).

Maximo, C. G., jr and Singh, J. Statistical methods in food and consumer research. Academic Press (1984).

MediaCybernetics: ImagePro Plus for Windows ®, version 1.2, (1995).

Næs, T., Kvaal, K., Isaksson, T. and Miller, C. Artificial neural networks in multivariate calibration. J. Near Infrared Spectroscopy (1993) 1(1) 1-12.

O'Mahony M. Sensory Evaluation of Food: Statistical Methods and Procedures. Marcel Dekker Inc. (1986).

Pratt, W. K. Digital Image Processing, pp557-598. John Wiley & Sons(1991).

Press, William H. Numerical Recipes in C. Cambridge University Press (1992).

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

ANALYZING DIFFERENCES AMONG PRODUCTS AND PANELISTS BY MULTIDIMENSIONAL SCALING

Richard Popper[a] and Hildegarde Heymann[b]

[a]Ocean Spray Cranberries, Inc., One Ocean Spray Drive, Lakeville/Middleboro, MA 02349

[b]Food Science & Nutrition Department, University of Missouri, 122 Eckles Hall, Columbia, MO 65201

1. INTRODUCTION

Multidimensional scaling (MDS) is a technique employed to display certain kinds of data spatially using a map. The basic concept of MDS is demonstrated in an example of Kruskal and Wish (1991). Consider the intercity flying distances among ten U.S. cities shown in Table 1. This table is easily constructed from a map of the United States by using a ruler and measuring the distances between the cities. Suppose, however, that one is presented with the intercity distances and asked to construct a map based on these distances. This is a more difficult problem, one that MDS is designed to solve. By applying MDS to the intercity distances, one obtains the map shown in Figure 1, which almost perfectly recreates the spatial arrangement of cities from which the distances were derived.[1]

Of course, in real applications of MDS the situation is more complicated. Unlike the intercity distances, real data contain measurement error, so the researcher must make a number of decisions concerning how best to model the data. Although the analysis of the intercity distances is an artificial example, it demonstrates the core idea underlying MDS: based on the distances among a set of objects, MDS constructs a picture in which these objects appear as points on a map.

MDS is applicable to a variety of data, not just actual distances. In fact, MDS can be used to analyze any data that represent how similar (or dissimilarobjects or events are to one another. For this reason, MDS has found application in a broad range of disciplines, including physics, psychology, physiology, linguistics, political science, and market research (Romney, Shepard, and Nerlove, 1972; Green and Wind, 1973; Schiffman, Reynolds and Young, 1981; Golledge and Rayner, 1982; Rosenberg, 1982; Young and Hamer, 1987). In each case, MDS is used to construct a spatial representation of the similarity among objects, with the purpose of discovering relationships or patterns. Usually two or three spatial dimensions are sufficient to reveal the most important relationships among the objects.

---

[1] Section 11 contains a list of the widely used computer programs for MDS. This analysis was conducted using the ALSCAL procedure in SPSS.

Table 1
Intercity flying distances

City 1	2	3	4	5	6	7	8	9	10	
1. Atlanta		587	1212	701	1936	604	748	2139	2182	543
2. Chicago	587		920	940	1745	1188	713	1858	1737	597
3. Denver	1212	920		879	831	1726	1631	949	1021	1494
4. Houston	701	940	879		1374	966	1420	1654	1891	1220
5. Los Angeles	1936	1745	831	1374		2339	2451	347	959	2300
6. Miami	604	1188	1726	966	2339		1092	2594	2734	923
7. New York	748	713	1631	1420	2451	1092		2571	2408	205
8. San Francisco	2139	1858	949	1654	347	2594	2571		678	2442
9. Seattle	2182	1737	1021	1891	959	2734	2408	678		2329
10. Washington, DC	543	597	1494	1220	2300	923	205	2442	2329	

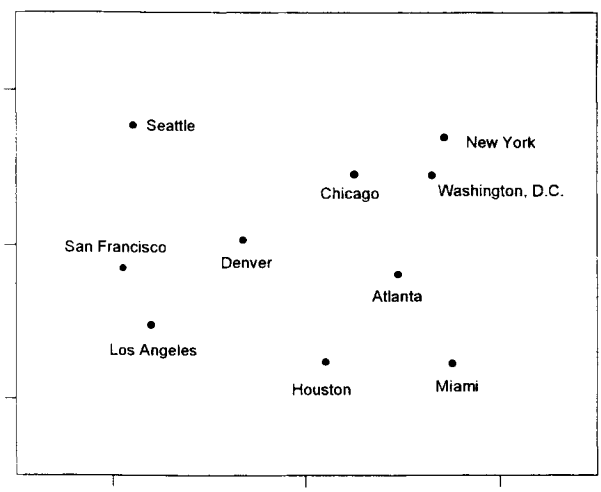


Figure 1. Location of ten U.S. cities as determined by an MDS analysis of intercity flying distances.

The development of MDS was largely motivated by a desire for a psychophysical scaling method that did not presuppose a knowledge of the attributes on which stimuli differ (Torgerson, 1958; Young and Hamer, 1987). MDS is often applied in situations where the researcher may not fully understand what specific attributes distinguish objects from one another. The advantage of MDS is that it requires as input only a measure of overall dissimilarity (or similarity) among objects. Difference measures on specific attributes are not required.

For example, a researcher may be interested in how consumers categorize beverages such as soft drinks, juices, alcoholic beverages, teas, and coffees. By asking consumers to rate the perceived similarity among beverages and by analyzing the ratings using MDS, the researcher can begin to

learn, based on the location of the products in the MDS map, which dimensions are important to consumers in differentiating among beverages. An example of the application of MDS to the study of the beverage market can be found in Hoffman and Young (1983).

2. MDS AND SENSORY ANALYSIS

MacFie and Thomson (1984) list several reasons for applying MDS in sensory analysis. The specific attributes that constitute a complex sensation, such as meat flavor, may not be known. When the attributes are unknown, MDS can be used to differentiate among the products because panelists need only rate dissimilarity. Even when the attributes are known, extensive training might be required for a sensory panel to measure the attributes reliably. Training is not only time consuming, but may be undesirable if a naive response is desired, as from a consumer panel. MDS, which requires the respondent to judge only overall (dis)similarity, provides a potential alternative in these situations.

Another reason for using MDS in sensory analysis is that often only two or three dimensions are needed to depict the important differences among samples. Simply by inspecting the position of the samples in the space and by noting which samples cluster together, the investigator is sometimes able to reach conclusions about the most salient differences and the possible basis for these differences. Other data analysis methods, such as principal component analysis, require the experimenter to collect data on multiple attributes, many of which are redundant or irrelevant to the panelists for distinguishing among the samples.

Finally, certain MDS procedures (Caroll and Chang, 1970) allow for the modeling of individual differences. Individual differences are of great interest, both in descriptive analysis and consumer research. In descriptive analysis, there is often a concern with differences (or inconsistencies) in sensory perception among individual panelists. The existence of such differences may suggest the need for better panel training. In consumer research, the question frequently asked is whether there are segments of consumers that differ in their preference for certain foods. For example, some consumers may prefer a mild tomato sauce, others a spicy one. Later in this chapter, examples will be presented of how MDS can be used to study individual differences among sensory panelists. The multidimensional scaling of preference data is the subject of Chapter 3. of this book.

An example from sensory analysis will help clarify some of the concepts discussed so far. Heymann (1994a) evaluated the aroma differences among four types of vanilla (Pure Bourbon, Bourbon Processed Bali, Indonesian, and Indonesian Nonsmoky), each processed to 3-fold, 10-fold, and 20-fold strength. Vanillin was also included among the samples (at 3-fold strength). Untrained panelists sorted the samples into groups based on their odor similarity, and the results were used to compute similarity scores among the samples (see Section 3 for details on the sorting procedure). A two-dimensional MDS analysis of the similarity scores (using SAS PROC MDS) fit the data well and yielded the map in Figure 2. The results shows that along the horizontal dimension panelists clearly differentiated the Indonesian samples from the Bourbon and Bourbon Processed Bali samples (and vanillin). No differentiation is apparent between the Indonesian and Indonesian Nonsmoky samples, or between the Bourbon and Bourbon Processed Bali samples. Within the Indonesian and Bourbon groups, samples of similar fold tend to group together. The vertical dimension may be related to concentration, although the ordering of fold levels along that dimension is not the same for the Indonesian and Bourbon type samples.

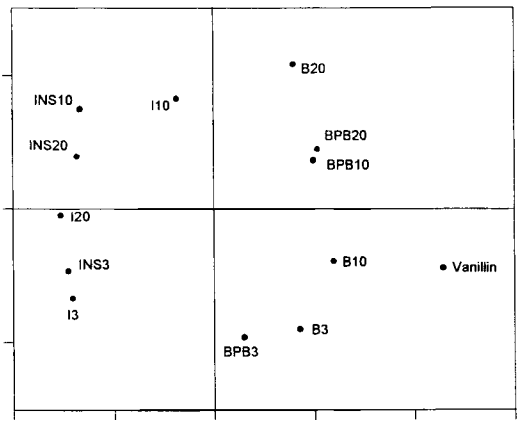


Figure 2. MDS representation of the aroma of vanillin and twelve vanilla samples. B=Bourbon, BPB=Bourbon Processed Bali, I=Indonesian, INS=Indonesian nonsmoky. Numbers represent strength of concentration (3-, 10-, or 20-fold).

The example illustrates several points about MDS. First, using untrained panelists and a similarity judgment task, it was possible to uncover meaningful groupings among the samples. Secondly, these groupings were readily apparent from a two-dimensional map and several conclusions were possible based on the configuration of samples in the plot. However, there is a limitation on one's ability to further interpret these results. Without an *a priori* knowledge of some of the aroma characteristics of these samples, it is not possible to conclude from this study what particular attributes were the basis for the groupings. In the absence of such knowledge, additional information, such as attribute ratings or physical measurements (for example, gas chromatographic readings), is often needed for interpreting MDS results (see Section 8).

MDS has been widely used in the study of chemoreception and in the sensory evaluation of foods and beverages. Schiffman used MDS extensively to map odor and taste quality using simple chemicals and tastants (Schiffman, Reynolds, and Young, 1981; Schiffman, 1984). The perception of alternative sweeteners was studied in model systems by Schiffman, Reilly, and Clark (1979) and Thomson, Tunaley, and van Trijp (1987), and in simple beverages by Schiffman, Crofton, and Beeker (1985). The saltiness of gum solutions was studied using MDS by Rosett, et al. (1995) and Rosett and Klein (1995).

Lawless applied MDS to study odor perception using aroma chemicals and fragrances (Lawless, 1989, Lawless and Glatter, 1990, and Lawless, 1993); to understand mouthfeel attributes (Bertino and Lawless, 1993), and to investigate cheese perception (Lawless, Cheng, and Knoops, 1995). Heymann used MDS in studies of vanilla flavor (Heymann, 1994a), apple essences (Gilbert and Heymann, 1995), and creaminess perception (Skibba and Heymann, 1994a, 1994b; Gwartney and Heymann, 1995).

MDS has also been used to study the different qualities of food sounds (Vickers and Wasserman, 1979, Vickers, 1983), the storage-related changes in orange juice aroma (Velez, et al. 1993), and the sensory characteristics of spreads (Tuorila, et al. 1989, Matuszewska, et al. 1991/2), rye breads (Hellemann, et al., 1987), yogurt (Poste and Patterson, 1988), meat (Francombe and MacFie, 1985), and soft drinks (Chauhan and Harper, 1986).

3. DATA COLLECTION PROCEDURES

A number of methods exist for collecting data for an MDS study. Regardless of how the data are collected, most MDS techniques require that the experimental results be organized in the form of a matrix of dissimilarities (or similarities), as shown in Table 1 for the intercity flying distances.[2] Among the applicable data collection methods, one can distinguish between those that involve an explicit evaluation of sample dissimilarity by the panelist and those that derive a dissimilarity matrix from other measurements (e.g., attribute ratings)[3].

Explicit evaluation methods include pair-wise dissimilarity scaling, conditional rank ordering and sorting. In all cases, the panelist's task is to evaluate dissimilarity among the samples based either on an attribute defined by the experimenter or based on unspecified attributes. For example, an experimenter might ask panelists to judge dissimilarity based on a specific attribute, such as color or odor. On the other hand, the experimenter may choose not to specify an attribute, in which case panelists are free to use their own criteria for judging dissimilarity.

Panel inconsistencies can arise if panelists have difficulty evaluating all samples on the same criteria and change their basis for judging dissimilarity depending on the samples. Such a change in judgment can result in an MDS space which underrepresents the number of dimensions actually relevant to the judgment task. Cohen and Jones (1974) simulated the effects of random error and sub-sampling of dimensions and found that dimensions which panelists consistently observed were well recovered in the final MDS configuration, but those dimensions which were not used consistently were not.

Several preparatory steps need to be completed prior to sensory data collection. These steps, also common to other sensory techniques, are those involved in recruiting and screening of panelists (Stone and Sidel, 1994; Meilgaard, Civille and Carr, 1991; Schiffman and Knecht, 1993). It is important that panelists be able to ascertain differences among samples and that they are capable of making judgments of dissimilarity. Additionally, the researcher has to decide which of the possible techniques, discussed in this section, should be used to collect the data (see also Green and Wind, 1973). During the data collection phase the researcher should use standard sensory methods to ensure the validity of the data (see Stone and Sidel, 1994; Meilgaard, Civille and Carr, 1991, for more information).

Traditionally, panelists evaluate dissimilarity between all possible pairs of products and indicate the perceived dissimilarity of each pair using category or line scales. Schiffman, Robinson and Erickson (1977) used a five inch line scale anchored with 'exactly same' on the left and 'completely different' on the right. In this case larger numbers indicated greater dissimilarity. These pair-wise comparisons provide a set of dissimilarity measurements that are then used in the calculation of the multidimensional spatial distances among samples.

The pair-wise method of sample presentation can quickly lead to an excessive number of evaluations for the panelists, as can be seen by the following calculation. The number of pairs used

---

[2] Not all MDS techniques are based on dissimilarity matrices. In particular, multidimensional unfolding methods (Schiffman, Reynolds and Young, 1981) have been developed that can accept as input sensory or hedonic attribute ratings and do not require the data to be transformed into dissimilarity scores. The unfolding models are not considered here.

[3] In most of what follows, there is no need to distinguish between similarity and dissimilarity scaling. The difference amounts to whether larger numbers reflect increasing similarity or dissimilarity MDS programs often accept data in either form. When only one form is acceptable, similarity judgments are easily transformed to dissimilarities (or vice versa) prior to analysis.

to pair-wise compare all samples in a set of size **n** is n(n-1)/2. With seven samples there are 21 pairs, with ten samples there are 45, and with 25 samples there are 300 pairs. This would lead to excessive sensory fatigue and would be very time consuming for the panelists, especially if the study is to be replicated. However, the technique has been used successfully (see, for example, Schiffman, Robinson and Erickson, 1977; Thomson and MacFie, 1983; Williams and Arnold, 1985; Gilbert and Heymann, 1995).

Due to the problem of sensory fatigue, especially with studies involving odor and flavor, researchers often attempt to find other methods of data collection. Incomplete data designs can be used to reduce the number of comparisons that each panelists must make (Schiffman and Knecht, 1993; Malhotra, Jain and Pinson,1988). These incomplete designs can be as simple as having twice as many panelists with each panelist evaluating half of the sample pairs. The panelists can be assigned sample pairs at random or through the use of a selection scheme (Spence and Domoney, 1974). A simulation by Whelehan, MacFie and Baust (1987) indicated that up to 40% of the dissimilarities in a complete pair-wise design can be omitted, if replications indicated that the error levels are not large. Moskowitz and Gerbers (1974) studied dimensional significance of odors through the use of an incomplete similarity scaling technique. More complex incomplete designs, such as cyclic designs, may also be used (Spence, 1982). For example, Rosett and Klein (1995b) in a study of saltiness perception in sixteen gum solutions, calculated that panelists would have to evaluate 120 pairs of solutions. They used incomplete cyclic designs in which half of the panelists evaluated 80 pairs of a potential 120 samples and the other half evaluated 75 sample pairs. The cyclic designs were chosen to overlap so that 35 of the pairs were the same for both groups.

Conditional rank order can also be used to decrease the number of samples that each panelist evaluates. In this procedure, each sample is used as a standard and the panelist ranks the remaining samples according to their similarity to the standard (Rao and Katz, 1971). For example, consider a study in which panelists are asked to evaluate the similarity of five products: sour cream, cream cheese, ice cream, milk and cream. A panelist first ranks the similarity of the samples using milk as a reference. Next, the panelist ranks the samples using cream as a reference, and subsequently, using sour cream, cream cheese and ice cream as references. Each panelist completes five rank orders, with the order of the standards balanced across panelists. This method works very well with visual stimuli but not with more fatiguing odor or flavor samples. It is possible to make the task less fatiguing by eliminating a sample from the comparison set once it has been used as a standard. In the above example, milk would be eliminated after round one, cream after round two and so on. However, collection of the full data set allows the researcher to check for panelist consistency and reliability (Deutscher, 1982). Special MDS models are required for analyzing conditional rank order data (Schiffman, Reynolds and Young, 1981).

Sorting has also been used to decrease the number of samples that the panelist evaluates. In this case, the panelist receives the entire set of samples at once and then sorts them into mutually exclusive groups based on similarity (Rao and Katz, 1971; Wish, 1976; Rosenberg and Kim, 1975; Rosenberg, 1982). Panelists are often told that they must sort the samples into no fewer than two groups and into no more groups than one less than the total number of samples in the set. This ensures that each panelist creates at least two groups yet cannot place each sample into its own group. The panel leader then counts how often any two samples were placed into the same group, thus deriving a similarity matrix in which larger numbers indicate increased similarity. The assumption underlying this method is that samples occurring in the same group are more similar than samples occurring in different groups. Panelists intuitively seem to understand the task, find it easy and perform it rapidly. This technique has been used extensively by Lawless (Lawless, 1989;

Lawless and Glatter, 1990; Lawless, 1993; Lawless, Cheng and Knoops, 1995) and Heymann (Heymann, 1994a; Gilbert and Heymann, 1995; Skibba and Heymann, 1994a,b) as well as others (MacRae et al., 1990; 1992).

With all of the above methods, the researcher needs to consider how he or she will interpret the dimensions of the spatial configuration that MDS derives (Lawless, 1993). A number of options exist (Hair, Anderson and Tatham, 1984). Researchers can simply use their own judgment, based on prior knowledge of the sample set, to arrive at a dimensional interpretation. In contrast, the researcher can present the final spatial arrangement to the panelists and ask them for suggestions as to its interpretation. The researcher can also ask the panelists, immediately after the data collection, to list the criteria which they used to judge or sort the products. However, panelists are frequently not able to articulate the criteria they used. A separate study, with the same or different subjects, using either consumer test methods or analytic descriptive techniques, can be conducted to generate information helpful to the interpretation of MDS dimensions. Examples of studies using the same panelists for that purpose are contained in Moskowitz and Gerbers (1974), Rosett and Klein (1995b) and Gilbert and Heymann (1995). Heymann (1994a) and Skibba and Heymann (1994a,b) used different panelists.

Instead of measuring dissimilarity directly, it is possible to derive dissimilarity scores from other kinds of data, such as from ratings collected as part of a descriptive study or in consumer research. Data collection procedures appropriate for such studies are described elsewhere (Einstein, 1991; Meilgaard, Civille and Carr, 1991; Heymann, Holt, and Cliff, 1993; Stone and Sidel, 1994). A number of transformations are possible for converting rating data to dissimilarities, including correlations and Euclidean distance computations (see Section 7 for an example).

4. STATISTICAL ASPECTS OF CLASSICAL MDS

An introduction to the statistical aspects of MDS can be found in Kruskal and Wish (1991). The mathematical foundations of MDS are discussed by Davison (1983) and Young and Hamer (1987). Schiffman, Reynolds, and Young (1981), MacFie and Thomson (1984), Schiffman and Beeker (1986), and Schiffman and Knecht (1993) explain in detail the statistical aspects of MDS using sensory applications as examples. In this chapter, only a few of the key statistical concepts will be reviewed.

The simplest type of multidimensional scaling model is called Classical MDS (CMDS) (Young and Hamer, 1987). The majority of applications of MDS involve this model. CMDS analyses a square data matrix, similar to the kind shown in Table 1 for intercity distances. As another example of a data matrix appropriate for analysis by CMDS, consider an experiment in which the investigator has collected pair-wise dissimilarity ratings from several panelists on four samples, using an unstructured line scale of the kind discussed in Section 3.[4] The hypothetical results of this experiment are shown in Table 2, in which the numerical entries represent average dissimilarity ratings for the pairs of samples. According to the results, samples C and B were the most dissimilar, samples D and A the most similar.

[4] The number of samples in an actual MDS study would need to be greater than four, see Section 10.

Table 2 is an example of a square symmetric matrix. The matrix is square, because there are as many rows as columns. The matrix is symmetric, because the dissimilarity of sample A to B equals that of sample B to A. The cells in the diagonal are blank, because panelists were not asked to rate the dissimilarity of a sample to itself.[5]

Table 2
Hypothetical dissimilarity data for four samples

	Sample A	Sample B	Sample C	Sample D
Sample A		2.5	7.6	1.2
Sample B	2.5		13.0	4.5
Sample C	7.6	13.0		10.6
Sample D	1.2	4.5	10.6	

While CMDS analyses only square matrices, other MDS models exist for analyzing rectangular data matrices, such as multiple attribute ratings or preference data, where the columns of the matrix represent samples, and the rows attributes or people. Such models will not be discussed here, but are reviewed by Schiffman, Reynolds and Young (1981).

A CMDS analysis can be either metric or nonmetric. In nonmetric CMDS, the dissimilarity data are treated as ordinal. This means that only the rank order of the dissimilaritiesin the input data matrix is used in determining the spatial configuration. In metric CMDS, on the other hand, the dissimilarities are assumed to have been measured on an interval or ratio level scale. Interval and ratio scales are more quantitative than ordinal scales. Ratio scales, for example, include those commonly used for measuring length and weight. Interval scales are similar to ratio scales, except that they lack a true zero point. Examples of interval scales are the Celsius and Fahrenheit scales of temperature. In CMDS, the researcher can choose whether to treat the data as metric or nonmetric.

It might seem that a nonmetric MDS analysis, which uses only the rank order of dissimilarities, would result in a less precise solution than a metric analysis of the same data. However, Shepard (1962) demonstrated that the rank order of dissimilarities is sufficient to derive a spatial configuration that closely matches that based on a metric analysis. For example, in Section 1, the intercity distances shown in Table 1 were analyzed using nonmetric CMDS, even though the distances represent ratio-level measurements. A metric analysis (not shown) of the same data results in a spatial configuration almost identical to that obtained using nonmetric CMDS.

Shepard's demonstration was important in the history of MDS, because many MDS applications involve scales whose measurement level is probably only ordinal. The level of measurement of sensory scales varies depending on the scale used. Rating scales, including category and unstructured line scales, are often assumed to be interval scales. However, the interval scale properties of these scales have not been demonstrated. MacFie and Thomson (1984) provide an example of why dissimilarity ratings common in sensory analysis may not satisfy the assumptions of a metric MDS analysis. Therefore, in sensory applications, the data are almost always treated as nonmetric. For example, the analysis of the vanilla data discussed in Section 2 was nonmetric.

---

[5] In most sensory studies, symmetry is assumed and the experimenter enters the same numbers in the lower and upper half of the data matrix. However, CMDS programs can accept nonsymmetric data matrices, as well as data matrices with nonzero entries in the diagonal.

A variety of computer algorithms exist for performing CMDS (see Section 11). They have a common objective, namely that of finding the spatial configuration of the samples that best agrees with the dissimilarities in the data matrix. The search for this configuration is an iterative process, one that terminates when further adjustments in the spatial configuration yield only minimal improvements in fitting the data. The degree of fit between the final configuration and the original data is expressed in a number of different ways. Perhaps the most common measure is called "stress", which is a "badness-of-it" measure (lower stress means a better fit). In MDS, stress is defined by the following formula:

$$Stress = \left[\frac{\sum_i \sum_j (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j d_{ij}^2} \right]^{1/2}$$

where d_{ij} represents the distance between objects **i** and **j** in the MDS space and $\hat{d}_{ij}$ the distance that best fits the dissimilarity between **i** and **j**. The formula shown above is often termed "stress formula 1" or "Kruskal's stress formula" (Kruskal, 1964). All MDS algorithms arrive at their final configuration by minimizing Kruskal's stress or a similar quantity.

It is possible to gain an intuitive understanding of stress by considering how MDS evaluates the fit between a spatial configuration and a set of dissimilarities. The distance between any pair of objects in the spatial configuration is compared with the size of the corresponding dissimilarity (a number given by the raw data.) If the spatial configuration fits the data well, a large distance will correspond to a large dissimilarity, a small distance to a small dissimilarity. In metric MDS, the degree of fit is quantified by using least squares regression to fit a straight line to the relationship between distance and dissimilarity. Stress measures the amount of deviation around this straight line. The larger the amount of deviation around that straight line, the poorer the fit and the larger the stress.

In nonmetric MDS, the same stress formula is used, except that instead of using linear regression to fit the distances to the dissimilarities, a least squares monotone regression is used, which fits a curve to the data that preserves the rank order of the dissimilarities, but is otherwise unconstrained.

In addition to stress, another measure of the degree of fit is the squared correlation coefficient between the interpoint distances in the spatial configuration and the dissimilarities (the original data). This correlation, sometimes designated RSQ, can be interpreted as the proportion of variance in the data that is accounted for by the distances in the MDS model. As is the case for any correlation-based measure, RSQ ranges between 0 and 1, where 0 indicates no fit and 1 a perfect fit.

For the intercity flying distances, which were analyzed using nonmetric CMDS, the stress for the two-dimensional solution was 0.008 and RSQ was 1.0 (after rounding). This excellent fit is expected, since the data were error free. For the vanilla data described in Section 2, the two-dimensional space was fit with a stress of 0.12 and an RSQ of 0.93. Kruskal (1964) has stated that a stress below 0.05 indicates a good fit, whereas stress values above 0.20 represent poor fits.

More detailed guidelines exist (Kruskal and Wish, 1991) for determining what level of stress represents a "good" fit. These guidelines take into consideration that several factors, in addition to the amount of error in the data, influence the magnitude of stress. These include the number of samples in the data set and the number of dimensions used to fit the data. In light of the influence of these and other factors, Krzanowski (1988) concluded that Kruskal's (1964) guidelines were overly

these and other factors, Krzanowski (1988) concluded that Kruskal's (1964) guidelines were overly simplistic. Often it is the researcher's past experience with MDS and his or her judgment that ultimately determine whether the fit of a particular MDS solution is acceptable or not.

Another related judgment the experimenter must make is how many dimensions to use to fit the data. Several considerations enter into this decision, including ease of interpretation, the number of samples in the data set (see Section 10 for guidelines) and the level of stress. Stress decreases as the number of dimensions increases. However, there is often a certain number of dimensions beyond which stress does not greatly improve. This point is most easily identified by plotting stress versus the number of dimensions. The point of diminishing improvement in stress appears as an "elbow" in the curve (Kruskal and Wish, 1991). This elbow defines the number of dimensions to be selected for fitting the data.

5. A CASE STUDY: PERCEPTION OF CREAMINESS

The perception of creaminess in foods is very complex. Textural creaminess is not a primary sensory attribute and may encompass thickness/viscosity, smoothness and fatty mouthfeel characteristics (Civille and Lawless, 1987). This case study was part of a larger study whose objective was to gain a more comprehensive understanding of creaminess perception. The case study was exploratory and compared the actual "in mouth" creaminess with expected creaminess based on product concepts as communicated by package labels. Both creamy and non-creamy products were evaluated (Skibba and Heymann, 1994a; 1994b; Gwartney and Heymann, 1995).

Unlike other MDS studies, in this study panelists were asked to evaluate samples based on one very specific, though complex, attribute (creaminess). Twenty food products (Table 3) were chosen to represent a wide range of textural perceptions. For "in mouth" evaluations, the samples were served as 30 ml servings in plastic cups with lids. Panelists received all samples simultaneously and were asked to sort them based on their similarity in textural creaminess. Panelists also received water for use in cleansing the palate. All samples and rinse water were expectorated. The "in mouth" sorting was replicated in two different sessions. For the concept evaluations, the panelists received the actual product labels pasted onto individual cards and were asked to sort them based on the creaminess similarity of the products described on the label. Panelists did not replicate the label sorting. In all cases the panelists were restricted to sorting the products or the labels into no more than 19 and no less than 2 mutually exclusive groups.

The twenty four panelists, all staff and students at the University of Missouri, were familiar with sensory testing but were otherwise untrained. Twelve of the twenty four panelists first sorted the products based on "in mouth" creaminess and subsequently sorted the food labels. The other twelve did the tasks in reverse order. Similarity scores were calculated by counting the number of times a pair of food products or labels was sorted into the same group. The similarity estimates were summarized in similarity matrices and submitted to the SYSTAT (Macintosh Version 3.2) MDS program for non-metric multidimensional scaling using Kruskal's stress formula. The results did not differ depending on the order in which the conditions were run, so only the composite of the data will be discussed.

Table 3
Case study: Food products and labels used in creaminess evaluation

Apple Sauce (Schnucks)
Chocolate Pudding (Del Monte)
Chocolate Syrup (Hershey)
Chocolate/hazelnut spread (Nutella, Ferrero)
Chocolate Milk (Schnucks)
Cream Soda (A & W)
Creamy Peanut Butter (Schnucks)
Evaporated Light Skimmed Milk (PET)
Half and Half (Schnucks)
Marshmallow Creme (Schnucks)
Non-dairy Creamer (CoffeeMate, Carnation)
Non-fat Sour Cream (Land-O-Lakes)
Part-skim Ricotta Cheese (Schnucks)
Ranch Creamy Dressing (Hidden Valley Ranch)
Skim Milk (Schnucks)
Soft Philadelphia Cream Cheese (Kraft)
Sour Cream (Schnucks)
Sweetened Condensed Milk (Meadow Gold)
Water (Culligan)
Whole Milk (Schnucks)

The subjects sorted both the products and the labels into a mean of seven groups (range four to eleven). Figure 3 is a two-dimensional MDS map of the "in mouth" sorting results. The analysis had a stress value of 0.10 and explained 94.5% of the data set variance. Figure 4 is a two-dimensional MDS map of the label sorting results, with a stress value of 0.06 and explaining 98% of the data set variance. According to Kruskal (1964), the stress values indicate a "fair" fit for the product condition and a "good" fit for the label condition.

Based on inspection, the dimensions of the product map (Figure 3) should be rotated by about 45° in a clockwise direction (see vectors). Rotation of MDS configurations, with the exception of individual difference MDS (see Section 6), is permissible if it improves the interpretability of the space (Kruskal and Wish, 1991). After rotation, one dimension can be interpreted as a perceived thickness or viscosity dimension, whereas the other dimension tends to correspond to variations in grittiness or lack of smoothness.

The dimensions of the label map (Figure 4) are less clear-cut, but by inspection it seems that rotating the horizontal dimension by 45° in a counterclockwise direction would lead to it being a contrast of "thin" and "thick". There are also two neighborhoods, one defined by liquids (on the right) and the other by semi-solids (on the left). It is interesting to note that the panelists thought that conceptually (based on the labels) marshmallow creme, chocolate/nut spread and peanut butter would be similar in creaminess to soft cream cheese and chocolate pudding. However, when the panelists evaluated these products "in mouth", the marshmallow creme, peanut butter and chocolate/nut spread were less creamy and less smooth (more gritty) than the cream cheese and chocolate pudding. Based on these exploratory data, panelists appear to respond differently to the perceived creaminess "in mouth" than to the creaminess as communicated by the package labels.

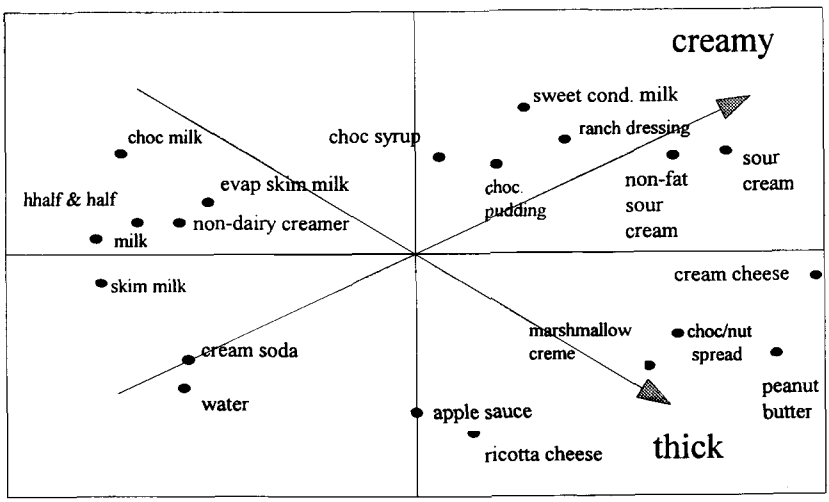


Figure 3. Two dimensional MDS map of the "in mouth" creaminess sorting.

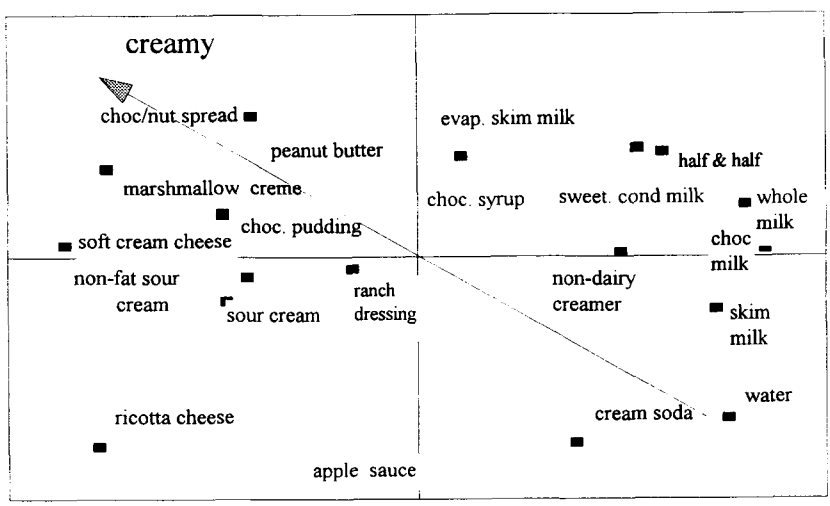


Figure 4. Two dimensional MDS mapof the label creaminess sorting.

6. STATISTICAL ASPECTS OF INDIVIDUAL DIFFERENCES SCALING

So far only applications of CMDS have been discussed. Another important type of MDS model is called weighted MDS (WMDS) (Schiffman, Reynolds and Young, 1981; Young and Hamer, 1987).

The first model for individual differences scaling, called INDSCAL, was proposed by Carroll and Chang (1970), and their work has served as the foundation for the development of most subsequent individual differences scaling approaches (see Young and Hamer, 1987; Krzanowski, 1988). Whereas CMDS analyses only one data matrix of the kind shown in Table 2, WMDSanalyses several such data matrices at the same time. In WMDS, each matrix represents the results of a separate experimental condition, a separate individual, or group of individuals. In MDS terms, the data for CMDS are called "two-way data", because a single data matrix always has two ways: the rows and the columns. When the data consist of a series of such matrices, the data are "three way", the third way corresponding to the factor that distinguishes the matrices from one another. Perhaps the most common application of WMDS is in the scaling of individual differences, where a series of data matrices are submitted to WMDS, one matrix for each individual tested. In that case, the "third way" corresponds to individuals.

In WMDS, differences among individuals are reflected as differences in weights for a set of common underlying dimensions. In addition to a group stimulus space (or consensus spatial configuration), WMDS derives dimension weights for each individual that can range from 0 to 1 and reflect the relative importance of each dimension to the individual.

As an example, consider an experiment in which three individuals are asked to rate the dissimilarity of six colors that vary only in hue. Suppose that the first subject has normal color vision, but the second and third subjects do not. Table 4 presents three dissimilarity matrices, one for each individual, where the numbers are hypothetical dissimilarity ratings. For the results of an actual WMDS analysis of color data, see Helm (1964) and Wish and Carroll (1973).

A two-dimensional analysis of WMDS analysis of these data using the ALSCAL procedure in SPSS results in the group or consensus space shown in the upper left of Figure 5, which reflects the information from all three individuals. In the group space, the colors are arranged in the shape of the familiar color circle, where opposing colors are located roughly opposite one another. WMDS also computes individual subject weights, shown in Table 5 that reflect the salience of each dimension for that individual. Table 5 shows that the three individuals weight the dimensions differently. In particular, subject 1 weights both dimensions about equally, whereas subject 2 and 3 weight one dimension much less than the other. The individual subject spaces, also shown in Figure 5, demonstrate the differences among the three individuals graphically. These individual subject spaces are derived from the group space by multiplying the stimulus coordinates on each dimension in the group space by the square root of the individual subject weight for that dimension (Schiffman, Reynolds and Young, 1981), according to the formula:

$$X_{kia} = W_{ka}^{½} X_{ia} ,$$

where X_{kia} is the coordinate of object **i** on dimension **a** for subject **k**, W_{ka} is the weight for subject **k** on dimension **a**, and X_{ia} is the coordinate of object **i** on dimension **a** of the group space. The differences in weights for the three subjects have resulted in differential stretching (and shrinking) of the two dimensions. Figure 5 shows that for subject 2, distances along the horizontal dimension are reduced, indicative of the low dimensional weight attached to that dimension. For subject 3, distances along the vertical dimension are reduced. Subject 2 suffers from a red-green color deficiency, subject 3 from a blue-yellow deficiency.

Table 4
Three hypothetical matrices of dissimilarity ratings of color*

	red	purple	blue	green	yellow/green	yellow
red	--					
purple	55	--				
blue	95	65	--			
green	105	85	30	--		
yellow/green	90	100	65	46	--	
yellow	60	95	95	90	50	--
red	--					
purple	45	--				
blue	45	40	--			
green	25	55	33	--		
yellow/green	43	90	75	45	--	
yellow	50	95	90	55	15	--
red	--					
purple	34	--				
blue	89	60	--			
green	105	75	15	--		
yellow/green	80	55	22	25	--	
yellow	35	35	65	75	50	--

*Note: The upper half of each matrix is identical to the lower half and has been omitted.

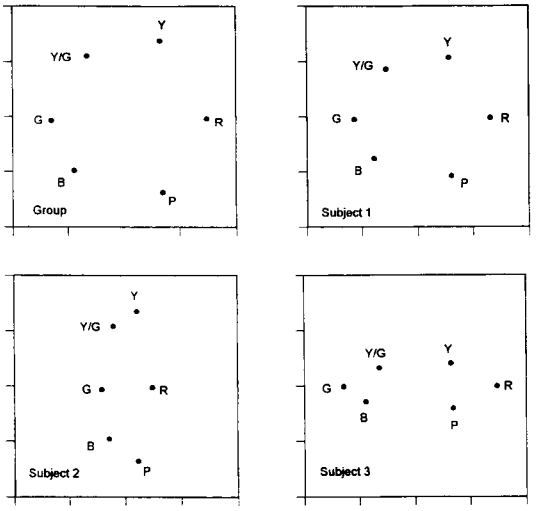


Figure 5. Individual differences scaling of hypothetical color ratings. G=green, Y/G=yellow/green, Y=yellow, R=red, P=purple, B=blue.

In applications of WMDS, individual differences might result from individual differences in sensation, perception or cognition. In each case, WMDS reflects these differences in the weights attached to a set of common underlying dimensions.

Table 5
Subject weights

	Dimension	
Subject Number	1	2
1	0.78	0. 61
2	0.10	0. 96
3	0.99	0 .08

WMDS shares many features in common with CMDS. WMDS can be metric or nonmetric (the analysis of the color ratings was nonmetric). The degree of fit is evaluated in the same fashion as in the case of CMDS, except that there are measures of fit for the group space as well as for the individual spaces.

There is one technical difference between CMDS and WMDS. In CMDS, the dimensions can be rotated as in the creaminess example. In WMDS, the dimensions cannot be rotated. This means that the dimensions in WMDS should be interpreted as is. Schiffman, Reynolds, and Young (1981), however, point out that this non-rotatability is true strictly only when the data contain no error. In the presence of error, some amount of rotation is permissible.

Finally, it should be noted that WMDS is based on a particular view of individual differences, namely that individuals differ in the relative importance they assign to a set of common dimensions. This is the only point of difference among individuals, according to the WMDS model. Mathematical extensions of WMDS models (see Young and Hamer, 1987) include differences among individuals in rotation of the group space and in the number of dimensions of the personal spaces. These extensions of the basic WMDS model, however, have found relatively few applications to date.

7. APPLICATIONS OF INDIVIDUAL DIFFERENCES SCALING

Gilbert and Heymann (1995) reported the results of an experiment in which panelists rated the dissimilarity in aroma among seven apple essences and a reconstituted apple base without essence added. Untrained panelists (N=18) rated the dissimilarity among all twenty-eight possible pairs of samples in three replications, using an unstructured 15 cm line scale. The average dissimilarity ratings were analyzed using CMDS.

In this section, the data are reanalyzed using nonmetric WMDS. Eighteen dissimilarity matrices, representing data from the individual panelists averaged across replications, were analyzed using the ALSCAL procedure in SPSS.

Solutions in two and three dimensions were explored. For the two-dimensional solution, the average stress (across the eighteen matrices) was high (0.30), and RSQ low (0.45). By comparison, a CMDS analysis of the group-averaged data yielded a very good fit (stress = 0.05, RSQ = 0.99).

This indicates that there was significant variation among the panelists that was not apparent in the CMDS analysis. Bertino and Lawless (1993) came to a similar conclusion regarding WMDS vs. CMDS results.

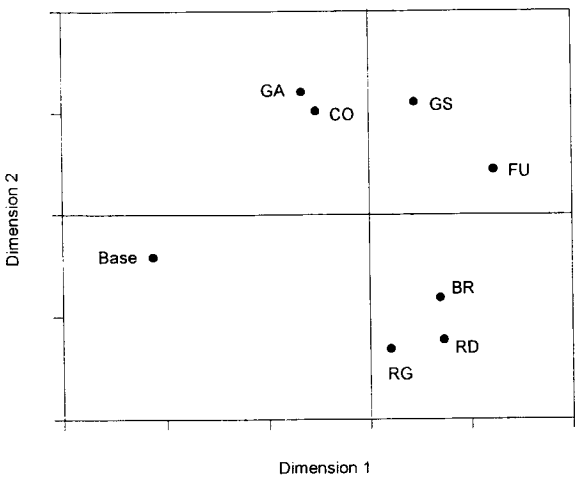


Figure 6. WMDS representation of the aroma of seven apple essences and an apple juice base without essence. See text for explanation of symbols.

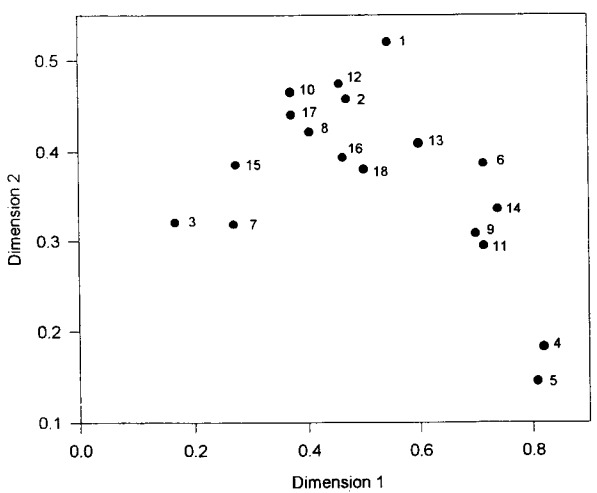


Figure 7. Dimensional weights for 18 panelists evaluating the aroma of apple essences.

Even though the stress was high, the two-dimensional individual difference solution obtained by WMDS was explored further. (The fit for the three dimensional solution was not much better.)

Figure 6 shows the sample space in which dimension 1 separates the apple juice base with no essence (BA) from the samples containing essence. Dimension 2 differentiates among apple essence varieties. Two clusters are apparent along this dimension. One is comprised of the Gala (GA), Cox's Orange Pippin (CO), and Granny Smith (GS) varieties, the other of the Royal Gala (RG), Red Delicious (RD), and Braeburn (BR) varieties. The Fuji (FU) sample may belong to neither cluster.

In Figure 7, individual panelists' weights for each dimension are plotted. Figure 7 shows a fairly wide distribution of weights for dimension 1 (from about 0.2 to 0.8), whereas the weights are more tightly clustered along dimension 2 (note the difference in scales on the vertical and horizontal axes). Panelists 4 and 5 were unusual in that they assigned very low weight to dimension 2. For these two panelists, only dimensions 1 was important, indicating that they attended only to the distinction between sample BA and the other samples. For the remaining panelists, the differences among the essence types were more salient, as indicated by their numerically larger weights on dimension 2. One implication of this analysis is that in the absence of training, the perception of these apple essences is quite variable, and the average solution may not be very representative of any one panelist.

The example above represents an application of WMDS to the scaling of individual dissimilarity matrices. WMDS can also be applied to traditional descriptive data or other attribute ratings. In order to do so, the data must first be transformed to dissimilarity form. The method for deriving dissimilarity data from attribute ratings involves the calculation of the distance between samples, based on the Euclidean distance formula:

$$d_{ij} = \left[\sum_{a=1}^{r} (X_{ia} - X_{ja})^2 \right]^{1/2}$$

where d_{ij} is the dissimilarity between objects **i** and **j**, $X_{ia} - X_{ja}$ is the difference between the two objects on attribute **a**, and the summation extends over all **r** attributes.

This method for deriving dissimilarity scores was applied to data from a study by Heymann (1994b). The study consisted of a descriptive analysis of the thirteen vanilla samples described in Section 2, using a panel different from the one which performed the similarity sorting task. This panel (N=10) rated the vanilla samples on fourteen attributes using standard descriptive methods. The data were reanalyzed by first computing, for each judge, dissimilarity scores among the samples, using the Euclidean distance measure. The ten dissimilarity matrices, one per judge, were then submitted to WMDS.

The average stress for a two-dimensional solution over the thirteen matrices was 0.28, with an RSQ of 0.70, indicating that a substantial amount of the individual variation could be explained by the model. The two-dimensional map, shown in Figure 8 is somewhat different from that shown in Figure 2, as might be expected given the differences in panelists, data collection method, and MDS model. Figure 9 shows the weight space for the ten panelists. There are several notable differences among the panelists. Panelist 2 weights dimension 2 almost exclusively over dimension 1. Panelists 3, 5 and 10, on the other hand, weight dimension 1 over dimension 2. If this situation were encountered early in a project, the panel leader would need to decide whether these panelist differences are ones that should be addressed by further training. An investigation of which descriptive attributes correlated most strongly with dimensions 1 and 2 would help to identify those sample attributes that may need to be clarified. If this kind of a result were encountered at the end of a project, the researcher would need to decide whether to omit certain panelists that are outliers (such as panelist 2) from the analysis.

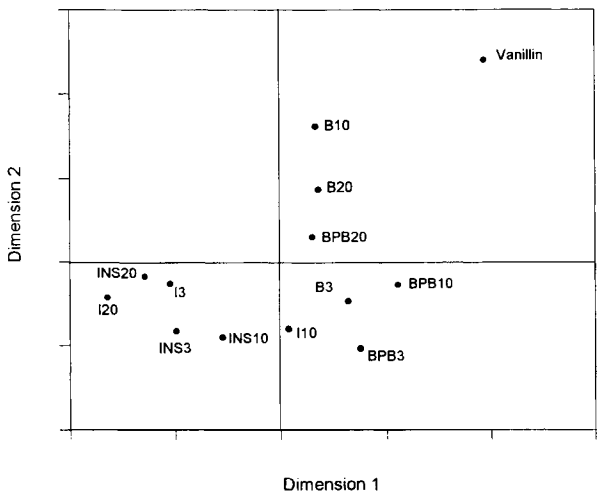


Figure 8. WMDS representation of the aroma of vanillin and twelve vanilla samples. See Figure 2 for legend.

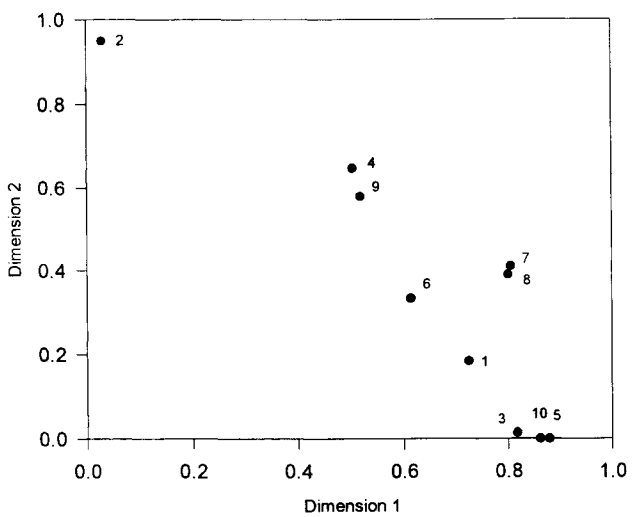


Figure 9. Dimensional weights for ten panelists evaluating the aroma of vanillin and twelve vanilla samples.

The examples above have illustrated the application of WMDS to the analysis of dissimilarity and attribute ratings. Unfortunately, while the similarity sorting procedure described in Section 3 is simple and easy to use, sorting data do not readily lend themselves to an individual differences analysis. This is due to the fact that the sorting data for any one individual consist of a matrix of

zeros and ones, indicating whether the individual grouped two products together or not. Even with replications, the number of times a product was sorted into the same group by one individual may not be a sufficiently graded measure to allow for the derivation of individual subject spaces. Lawless, Sherry and Knoops (1995) discuss a potential variation of the sorting procedure that may yield data that are scalable by WMDS. However, Lawless et al. indicate that to date they have conducted only limited testing using this procedure.

8. ISSUES IN INTERPRETATION OF MDS SPACES

The case studies discussed above illustrate how MDS can be used to derive spatial representations of the similarities and differences among samples. The interpretation of MDS spaces involves a degree of judgment on the part of the researcher. Often the grouping of the samples, together with prior knowledge of sample characteristics, suggests on what basis samples are being differentiated. This can lead to a "naming" of the underlying dimensions of the MDS space, as in the creaminess study (see Figures 3 and 4). In other instances, naming of the dimensions is difficult, but the clustering of samples still can be informative. Without knowing the characteristics on which vanilla samples differ, it is clear from Figure 2 that judges are able to distinguish the Indonesian from the Bourbon types of vanilla samples. This itself can be useful information, for example in deciding whether Indonesian samples can be used as substitutes for the more expensive Bourbon type samples. The results in Figure 2 suggest that there are systematic differences between the two types of vanilla samples, although the salience of that difference might depend on actual product application.

In some cases, the researcher may want additional information regarding why the samples group as they do. This is especially the case when the researcher lacks prior knowledge of the sample characteristics or when visual inspection of the map is insufficient to generate hypotheses about the nature of the underlying differences. In such instances, ancillary data are often used to aid in the interpretation of MDS spaces. These data most commonly consist of ratings of the samples on specific attributes, collected either from the panel that judged the similarity of the samples, or from a separate panel. Instead of attribute ratings, analytical measurements, such as pH, amount of an ingredient present in the sample, etc. also can be used to interpret MDS results. If the same panel is used to collect both similarity and attribute data, the attribute ratings should be collected after the similarity judgments to avoid potentially biasing judges by focusing their attention on a limited set of attributes.

A list of attributes for use in a rating task can be generated in several ways. The panelists who judged the samples for similarity can be asked to identify which attributes they thought most distinguished the samples. Alternatively, the researcher, independent of any panelist feedback, can postulate what attributes are likely to be relevant to judging sample similarity and can collect data from the similarity panel (or another panel) on these attributes. Finally, an independent panel can generate attribute ratings of the individual samples using standard descriptive methods (see Section 3).

There are several methods for relating attribute ratings or instrumental measurements to the MDS space. The simplest (and most frequently used) method is to determine how each attribute is correlated with the dimensions of the MDS space. The mathematical procedure is described by Schiffman, Reynolds and Young (1981) and involves a multiple regression in which each attribute,

taken one at a time, is regressed against the coordinates of the MDS dimensions. Alternatively, multivariate techniques exist for simultaneously fitting several attributes to the dimensions of the MDS space. These techniques include canonical correlation (Schiffman, Reynolds, and Young, 1981; Schiffman and Beeker, 1986) and partial least squaresregression (Schiffman and Beeker, 1986; Popper, et al. 1987).

To illustrate the multiple regression approach, the MDS space of the vanilla samples shown in Figure 2 is interpreted using ratings provided by an independent descriptive panel (Heymann, 1994b). In this analysis, each of the fourteen descriptive attributes served, one at a time, as the dependent variable, while the coordinates of the vanilla samples played the role of the independent variables. There are two independent variables, since the MDS solution was two-dimensional. The analysis can be accomplished using either standard multiple regression software or specialized computer programs, such as PROFIT (which stands for PROperty FITting) or PREFMAP (which stands for PREFerence MAPping). Both these programs were developed by Carroll and Chang at Bell Laboratories and are included, in revised form, in the PC-MDS computer package for multidimensional statistics (see Section 11)[6].

In order to interpret the MDS space, Figure 10 shows the attributes projected as vectors in the space. The attribute vectors point in the direction of increasing magnitude, and their angle indicates the correlation with the vertical and horizontal dimensions. The length of the vectors is drawn in proportion to the magnitude of the correlation between the attribute and the MDS dimensions. Thus, it appears from Figure 10 that the horizontal dimension contrasts the woody, smoky, and nutty aroma of the Indonesian samples (I and INS) on the left with the butterscotch and sweet milk aroma of the Bourbon samples (B and BPB) and vanillin, located on the right. Also, from the lengths of the various vectors it appears that the almond, raisin and rum characteristics are much less relevant to distinguishing among the vanilla samples than the other attributes.

9. RELATIONSHIP OF MDS TO OTHER METHODS

The resemblances of graphical depictions of data spaces derived from multidimensional scaling, descriptive and free-choice profiling data have been studied by Chauhan, Harper and Krzanowski (1983), Williams and Arnold (1985), Lawless (1993), Heymann (1994a), Skibba and Heymann (1994b) and Gilbert and Heymann (1995). These authors all concluded that the data spaces derived by MDS are 'similar' to those derived by the other techniques. In most cases, the similarity was determined based either on visual inspection or correlation.

Chauhan, Harper and Krzanowski (1983) compared the results of similarity scaling of pairs of soft drinks with profiling data derived by the same panelists. They used multidimensional unfolding to compare the results of the two methods and found that the results were essentially identical for five of the seven drinks used in the study.

Gilbert and Heymann (1995) compared the data spaces obtained from multidimensional sorting, multidimensional similarity scaling, free-choice profiling and descriptive analysis of apple essences. They found that the sorting and group-averaged scaling results differed markedly. However, the MDS space derived from the sorting data was very similar to the principal component space derived

[6] PREFMAP, which can be used to fit both attribute and preference data to MDS spaces, includes several different types of models, of which the vector model is applicable here. See Schiffman, Reynolds and Young (1981), MacFie and Thomson (1984) and Chapter 3 of this book for a discussion of preference mapping.

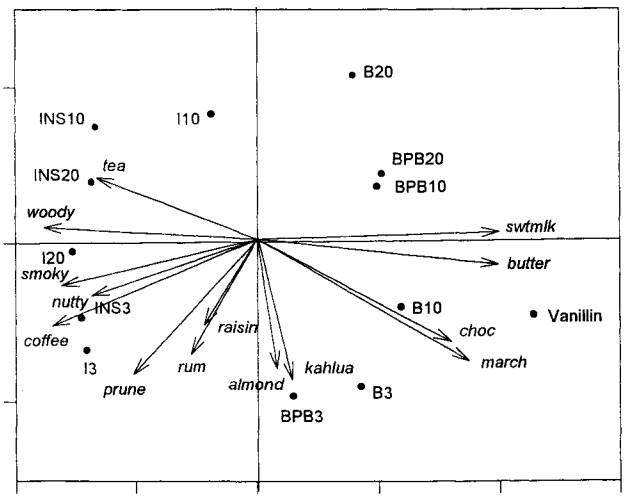


Figure 10. Attribute vectors fit to the CMDS analysis of the vanilla samples shown in Figure 2. Attributes represent different aroma characteristics. [Note: butter=butterscotch, marsh= marshmallow, and swtmlk=sweet milk.]

from the descriptive data of the essences and the Procrustes space derived from the apple essence free-choice profile data. These authors compared the data spaces visually and through the use of correlation and Procrustes analysis.

How do different types of MDS analyses compare to one another and how reliable are the results of any one analysis? As noted above, Gilbert and Heymann found that the MDS maps for sorting and group-averaged similarity scaling differed. Rao and Katz (1971) concluded that multidimensional sorting methods usually performed worse than other multidimensional data collection methods. However, Bertino and Lawless (1993) compared the results of similarity scaling to those of sorting and found that the MDS configurations for the sorting and group-averaged similarity scaling tasks were similar. They also found that an individual differences scaling analysis of the rating data had very large stress, indicating that the group-averaged data did not capture nuances found by individual panelists. Only a few studies have concerned themselves with the reliability of any one MDS method, using either real data or Monte Carlo simulations, and the results have not been conclusive (Golledge and Rayner, 1982; Krzanowski, 1988).

Multidimensional scaling, principal component analysis, cluster analysis, partial least squares (PLS) analysis and Procrustes analysis are multivariate statistical techniques that can all be used to analyze the same dissimilarity scores or attribute ratings. All of these methods can aid in ascertaining latent phenomena in the data. For example, Bieber and Smith (1986) compared multidimensional scaling, factor analysis and cluster analysis and noted both similarities and differences. Krzanowki (1988) explained the connection between principal component analysis and metric multidimensional scaling. He also noted the similarity in results between metric multidimensional scaling and canonical variate analysis using data obtained from a study of British water voles. However, no studies have compared MDS to other methods regarding results on individual differences.

It should be emphasized that the MDS models considered in this chapter are exploratory in nature. These models do not allow for inferential tests of hypotheses concerning the size of sample differences. As stated by Rosett and Klein (1995), MDS and inferential techniques, such as analysis of variance, should be viewed as complimentary techniques in sensory applications.

In more recent MDS models, Ramsey (1982) employed the principle of maximum likelihood in developing tests of statistical significance for the appropriate number of dimensions and the type of MDS model. These models also include confidence regions for samples or subjects (in the case of WMDS). However, the assumptions underlying these models remain controversial (Young and Hamer, 1987), and applications of these models in sensory analysis have not been reported.

10. FURTHER GUIDELINES FOR DESIGNING MDS EXPERIMENTS

Schiffman and Knecht (1993) suggest that the 'it is preferable to use 12 stimuli for two dimensional solutions and 18 stimuli for three-dimensional solutions'. Kruskal and Wish (1991) indicate that the number of samples less one should be at least four times larger than the number of dimensions to the solution. With fewer samples, subtle differences among the samples may not be captured in the solution. Thus, it is better to use more rather than fewer samples.

Schiffman and Knecht also suggest that the number of samples may be decreased if data from more than one subject are used. However, they do not support this statement with data. Unfortunately, the minimum number of panelists needed is not clearly stated by any author.

Green and Wind (1973) point out that depending on the purpose of the study the samples in the stimulus set can be physical objects (like the "in mouth" creamy and non-creamy samples), pictoral or graphical representations of objects (like the labels of the creamy and non-creamy samples), or verbal descriptions of the objects or sensations (Martens, et al., 1988; Bertino and Lawless, 1993).

11. COMPUTER SOFTWARE FOR MDS

The first computer programs for MDS were available only from universities or research centers and were designed to operate on mainframe computers (see Schiffman, Reynolds and Young, 1981 for details on how to obtain and use some of these programs). With the acceptance of MDS as a data analysis tool and with the growth in the power of personal computers, access to MDS software has greatly increased. Several of the original MDS programs have been made available in versions for the PC in a package called PC-MDS (S.M. Smith, Brigham Young University, Provo, UT 84602). A limited version of these programs is available on disks supplied with the book by Green, Carmone and Smith (1989).

Several PC-based statistical packages include MDS procedures. SYSTAT (Systat Inc., 1800 Sherman Avenue, Evanston, IL 60201) performs CMDS, but not WMDS. The PC-versions of SAS (SAS Institute Inc., SAS Campus Drive, Cary, NC 27513) and SPSS (SPSS Inc., 444 N. Michigan Avenue, Chicago, IL 60611) include versions of ALSCAL, a very comprehensive program developed by Forrest Young. ALSCAL offers the full range of options for performing both CMDS and WMDS, as well as other variations of MDS. An excellent introduction to MDS is contained in the documentation accompanying the SPSS-PC program.

12. REFERENCES

Bertino, M. and Lawless, H.T.(1993). Understanding mouthfeel attributes: a multidimensional scaling approach. J. of Sensory Studies, 8, 101- 114.

Bieber, S.L. and Smith, D.V. (1986). Multivariate analysis of sensory data: a comparison of methods. Chemical Senses, 11, 19-47.

Caroll, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckhart-Young' decomposition. Psychometrika, 35, 283-319.

Chauhan, J. and Harper, R. (1986). Descriptive profiling versus direct similarity assessments of soft drinks. J. of Food Technology, 21, 175-187.

Chauhan, J., Harper, R., and Krzanowski, W. (1983). Comparison between direct similarity assessments and descriptive profiles of certain soft drinks. In: Sensory quality of foods and beverages: definition, measurement and control. Williams, A.A and Atkin, R.K. (eds). Ellis Horwood, Chichester, United Kingdom.

Civille, G.V. and Lawless, H.T. (1987). The importance of language in describing perceptions. J. of Sensory Studies, 1, 217-236.

Cohen, H.S. and Jones, L.E. (1974). The effects of random error and subsampling of dimensions on recovery of configuration by non-metric multidimensional scaling. Psychometrika, 39, 69-90.

Davison, M. (1983). Multidimensional Scaling. Wiley, New York.

Deutscher, T. (1982). Issues in data collection and reliability in marketing multidimensional scaling studies -- implications for large stimulus sets. In: Proximity and Preference: Problems in Multidimensional Analysis of Large Data Sets. Golledge, R.G and Rayner, J.N (eds). University of Minnesota Press, Minneapolis.

Einstein, M.A. (1991). Descriptive techniques and their hybridization. In: Sensory Science Theory and Applications in Foods. Lawless, H.T. and Klein, B.P. (eds). Marcel Dekker, Inc., New York.

Francombe, M.A. and MacFie, H.J.H. (1985). Dissimilarity scaling and INDSCAL analysis in the study of flavor differences between normal pH and DFD beef. J. of the Science of Food and Agriculture, 36, 699-708.

Gilbert, J.M. and Heymann, H. (1995). Comparison of four sensory methodologies as alternatives to Descriptive Analysis for the evaluation of apple essence aroma. The Food Technologist (NZIFST) 24, 28-32..

Golledge, R.G. and Rayner, J.N. (1982). Proximity and Preference: Problems in Multidimensional Analysis of Large Data Sets. University of Minnesota Press, Minneapolis.

Green, P.E. and Rao, V.R. (1972). Applied Multidimensional Scaling: A Comparison of Approaches and Algorithms. Allyn & Bacon, Inc, Boston, MA.

Green, P.E., Carmone, F.J. and Smith, S.M., (1989). Multidimensional scaling: Concepts and Applications. Allyn and Bacon, Boston.

Gwartney, E. and Heymann, H. (1995). The perception of creaminess. J. of Sensory Studies (Submitted).

Hair, J.F., Anderson, R.E. and Tatham, R.L. (1984). Multivariate Data Analysis: With Readings. 2nd edition. MacMillan Publ. Co., New York.

Hellemann, U., Tuorila, H., Salovaara, H. and Tarkkonen, L. (1987). Sensory profiling and multidimensional scaling of selected Finnish rye breads. International Journal of Food Science and Technology, 22, 693-700.

Helm, C.E. (1964). Multidimensional ratio scaling of perceived color relations. J. of Optical Society America, 54, 256-262.

Heymann, H. (1994a). A comparison of free choice profiling and multidimensional scaling of vanilla samples. J. of Sensory Studies, 9, 445-453.

Heymann, H. (1994b). A comparison of descriptive analysis of vanilla by two independently trained panels. J. of Sensory Studies, 9, 21-32.

Heymann, H., Holt, D.L. and Cliff, M.A. (1993). Measurement of flavor by sensory descriptive techniques. In: Flavor Measurement. C.T. Ho and C.H. Manley (eds). Marcel Dekker Inc., New York.

Hoffman, D.L. and Young, F.W. (1983). Quantitative analysis of qualitative data. In: Food Research and Data Analysis. Martens, H. and Russwurm, H., Jr. (eds), pp. 69-93. Applied Science Publ., London.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness of fit to a non metric hypothesis. Psychometrika, 29, 1-27.

Kruskal, J.B. and Wish, M. (1991). Multidimensional Scaling. SAGE Publications, Newberry Park, CA.

Krzanowski, W.J. (1988). Principles of Multivariate Analysis: A User's Perspective. Oxford Science Publications, Oxford Press, Oxford, United Kingdom.

Lawless, H.T and Glatter, S. (1990). Consistency of multidimensional scaling models derived from odor sorting. J. of Sensory Studies, 5, 217-230.

Lawless, H.T. (1989). Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis. Chemical Senses, 14, 349- 360.

Lawless, H.T. (1993). Characterization of odor quality through sorting and multidimensional scaling. In: Flavor Measurement. C.T. Ho and C.H. Manley (eds). Marcel Dekker Inc., New York.

Lawless, H.T., Cheng, N. and Knoops, S.S.C.P. (1995). Multidimensional scaling of sorting data applied to cheese perception. Food Quality and Preference, 6, 91-98.

MacFie, H.J.H. and Thomson, D.M.H. (1984). Multidimensional Scaling. In: Sensory Analysis of Foods. Piggott, J.R. (ed), Elsevier, London.

MacRae, A.W., Howgate, P. and Geelhoed, E. (1990). Assessing the similarity of odors by sorting and by triadic comparison. Chemical Senses, 15, 691-699.

MacRae, A.W., Rawcliffe, T., Howgate, P. and Geelhoed, E. (1992). Patterns of odor similarity among carbonyls and their mixtures. Chemical Senses, 17, 119-125.

Malhotra, N.K., Jain, A.K. and Pinson, C. (1988). The robustness of MDS configurations in the case of incomplete data. J. of Marketing Research, 25, 95-102.

Martens, M., Rodbotten, M., Martens, H., Risvik, E. and Russwurm, H., Jr. (1988). Dissimilarities in cognition of flavor terms related to various sensory laboratories in a multivariate study. J. of Sensory Studies, 3, 123-135.

Matuszewska, I., Barylko-Pikielna, N., Tarkkonen, L., Helleman, U. and Tuorila, H. (1991/92). Similarity ratings versus profiling of spreads: do we need both? Food Quality and Preference, 3, 47-50.

Meilgaard, M., Civille, C.V., and Carr, B.T. (1991). Sensory Evaluation Techniques. CRC Press, Inc., Boca Raton, FL.

Moskowitz, H.R. and Gerbers, C.L. (1974). Dimensional salience of odors. Annals of the New York Academy of Science, 237, 1-16.

Popper, R., Risvik, E., Martens, H. and Martens, M. (1988). A comparison of multivariate approaches to sensory analysis and the prediction of acceptability. In: Food Acceptability. Thomson, D.M.H. (ed). Elsevier, London.

Poste, L.M. and Patterson, C.F (1988). Multidimensional scaling - sensory analysis of yoghurt. Canadian Institute of Food Science and Technology Journal, 21, 271-278.

Ramsey, J.O. (1982). Some statistical approaches to multidimensional scaling data (with discussion). J. of the Royal Statistical Society (A), 145, 285-312.

Rao, V.R. and Katz, R. (1971). Alternative multidimensional scaling methods for large stimulus sets. J. of Marketing Research, 8, 488-494.

Romney, A.K., Shepard, R.N. and Nerlove, S.B. (1972). Multidimensional Scaling: Theory and Applications in the Behavioral Sciences. Vol. II: Applications. Seminar PRess, New York.

Rosenberg, S. (1982). The method of sorting in multivariate research with applications selected from cognitive psychology and person perception. In: Multivariate Applications in the Social Sciences. N. Hirschberg and L. Humphreys (eds). Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Rosenberg, S. and Kim, M.P. (1975). The method of sorting as a data gathering procedure in multivariate research. Multivariate Behavioral Research, 10, 489-502.

Rosett, T.R. and Klein, B.P. (1995b). Efficiency of a cyclic design and a multidimensional scaling sensory analysis technique in the study of salt taste. J. of Sensory Studies, 10(1), 25-44.

Rosett, T.R., Shirley, L., Klein, B. and Schmidt, S.L. (1995). Saltiness of gum solutions is affected by the binding of Na+ as measured by 23Na nuclear magnetic resonance spectroscopy. J. of Food Science ,60, 849-853, 867.

Schiffman, S.S. (1984). Mathematical approaches for quantiative design of oderants and tastants. In: Computers in Flavor and Fragrance Research. Warren, C. & Walradt, J. (eds), pp. 33-50. American Chemical Society, Washington, D.C.

Schiffman, S.S. and Beeker, T.G. (1986). Multidimensional scaling and its interpretation. In: Statistical Procedures in Food Research. Piggott, J.R.(ed). Elsevier Applied Science, London.

Schiffman, S.S. and Knecht, T.W. (1993). Basic concepts and programs for multidimensional scaling. In: Flavor Measurement. C.T. Ho and C.H. Manley (eds). Marcel Dekker Inc., New York.

Schiffman, S.S., Crofton, V.A. and Beeker, T.G. (1985). Sensory evaluation of soft drinks with various sweeteners. Physiology and Behavior, 35, 369-377.

Schiffman, S.S., Reilly, D.A. and Clark, T.B. (1979). Qualitative differences among sweeteners. Physiology and Behavior, 23, 1-9.

Schiffman, S.S., Reynolds, M.L. and Young, F.W. (1981). Introduction to Mulitdimensional Scaling: Theory, Methods, Applications. Academic Press, New York.

Schiffman, S.S., Robinson, D.E. and Erickson, R.P. (1977). Multidimensional scaling of odorants: Examination of psychological and physicochemical dimensions. Chemical Senses and Flavor, 2, 375-390.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. Part I and II. Psychometrika, 27, 125-140, 219-246.

Skibba, E.A. and Heymann, H. (1994a). Creaminess perception. Presented at AChemS Annual Meeting, Sarasota, Florida, April 14, 1994.

Skibba, E.A. and Heymann, H. (1994b). The perception of creaminess. Presented at IFT Annual Meeting, Atlanta, Georgia, June 23, 1994.

Spence, I. (1982). Incomplete experimental designs for multidimensional scaling. In: Proximity and Preference: Problems in Multidimensional Analysis of Large Data Sets. Golledge, R.G and Rayner, J.N (eds). University of Minnesota Press, Minneapolis.

Spence, I. and Domoney, D.W. (1974). Single subject incomplete designs for non-metric multidimensional scaling. Psychometrika, 37, 461-486

Stone, H. and Sidel, J.L. (1994). Sensory Evaluation Practices. Second edition. Academic Press, Orlando, Florida.

Thomson, D.M.H. and MacFie, H.J.H. (1983). Is there an alternative to descriptive sensory assessment. In: Sensory Quality of Food and Beverages: Definition, Measurement and Control. Williams, A.A. and Atkin, R.A. (eds). Ellis Horwood, Chichester, England.

Thomson, D.M.H., Tunaley, A. and van Trijp, H.C.M. (1987). A reappraisal of the use of multidimensional scaling to investigate the sensory characteristics of sweeteners. J. of Sensory Studies, 2, 215-230.

Torgerson, W.S.(1958). Theory and Methods of Scaling. Wiley, New York.

Tuorila, H., Matuszewska, I., Hellemann, U. and Lampi, A.M. (1989). Sensory and chemical characterization of fats used as spreads on bread. Food Quality and Preference, 1989, 1, 157-162.

Velez, C., Costell, E., Orlando, L., Nadal, M.I., Sendra, J.M. and Izquierdo, L. (1993). Multidimensional scaling as a method to correlate sensory and instrumental data of orange juice aromas. J. Science Food and Agriculture, 61, 41-46.

Whelehan, O.P., MacFie, H.J.H. and Baust, N.G. (1987). Use of individual differences scaling for sensory studies: Simulated recovery of structure under various missing value rates and error levels. J. of Sensory Studies, 1, 1-8.

Vickers, Z.M. (1983). Pleasantness of food sounds. J. of Food Science, 48, 783-786.

Vickers, Z.M. and Wasserman, S.S. (1970). Sensory qualities of food sounds based on individual perceptions. J. of Texture Studies, 10, 319-332.

Williams, A.A. and Arnold, G.M. (1985). A comparison of the aromas of six coffees charcterized by conventional profiling, free-choice profiling and similarity scaling methods. J. Science Food and Agriculture, 36, 204-214.

Wish, M. (1976). Comparisons among multidimensional structures of interpersonal relations. Multivariate Behavioral Research, 11, 297-327.

Wish, M. and Carroll, J.D. (1973). Concepts and applications of multidimensional scaling. In: Sensory Evaluation of Appearance of Materials, STP545. Hunter, R.S. and Martin, P.N. (eds). American Society for Testing and Materials, Philadelphia, PA.

Young, F.W. and Hamer, R.M. (1987). Multidimensional Scaling: History, Theory, and Applications. Lawrence Erlbaum Assoc., Hilldale, NJ.

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

PROCRUSTES ANALYSIS IN SENSORY RESEARCH

Garmt Dijksterhuis

ID-DLO, Institute for Animal Science and Health, Sensory Laboratory, PO Box 65, NL-8200 AB Lelystad, the Netherlands

1. INTRODUCTION

Since its adoption in the seventies (e.g. Banfield and Harries 1975, Harries and MacFie 1976), Procrustes analysis has become a popular tool for sensory scientists (Williams and Langron 1984, Arnold and Williams 1985), and still the method is used frequently and is studied and extended by several authors (Oreskovich et al. 1991, Dijksterhuis and Gower 1991/2, Wakeling et al. 1992). Procrustes analysis was originally developed as a technique to match the solutions of two Factor Analyses (Hurley & Cattell 1962). The method was generalised to match more than two data sets by Kristof and Wingersky (1971) and Gower (1975). Recently the method has received increasing attention, partly through the availability of software programs for generalised Procrustes analysis (GPA), partly through some critisisms on the method.

In this chapter the kinds of sensory data to which GPA can be applied are introduced, along with the rationale for using the method. Next some background and theory of GPA is provided with special attention for the Procrustes analysis of variance. Finally, two applications of GPA to sensory profiling data, one conventional and one free-choice, are shown.

1.1 Sensory profiling

A very large number of applications of generalised Procrustes analysis is found in the analysis of sensory profiling data. There are two different kinds of profiling data, that can both be analysed by means of generalised Procrustes analysis. Conventional profiling data can also be analysed by averaging and applying factor analysis or PCA to it. Free choice profiling FCP, (Williams & Langron 1984, Williams and Arnold 1985) results in data that can not be averaged over assessors, generalised Procrustes analysis or other, so-called, *K*-sets methods are suited for the analysis of free choice profiling data.

The scores from either profiling technique are derived from the position of marks along a line-scale. The assessor marks his/her perceived intensity of some attribute along a line scale (Figure 1). Often the scores range from 0 to 100, but the range is unimportant, in the following a range from 0 to 100 is assumed.

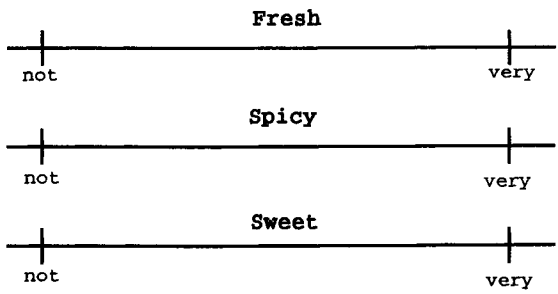


Figure 1. Example of a line-scale often used in sensory profiling experiments.

1.1.1 Conventional profiling

In conventional profiling a fixed vocabulary of descriptive terms is used by the sensory panel to judge the products. A sensory panel is often trained in the use of these terms. In the case of QDA (Quantitative Descriptive Analysis, see Stone & Sidel 1985) the panel starts with the generation of a lot of terms that are thought useful to describe the products under consideration. The whole procedure of attribute generation and training can take considerable time. Because of this training it is assumed that all assessors are able to use the attributes in the same way, so individual differences in use of the attributes are minimized. Because of this the individual judgements are sometimes averaged and factor analysis or PCA is applied to the average scores. However, methods as generalised Procrustes analysis can of course also be applied to conventional profiling data. Such analyses show that the assumption of all assessors using the attributes in the same way is not always justified (see e.g. Dijksterhuis & Punter 1990).

The data from conventional profiling experiments can be seen as a 3-mode data structure built from N products, M attributes and K assessors (see Figure 2).

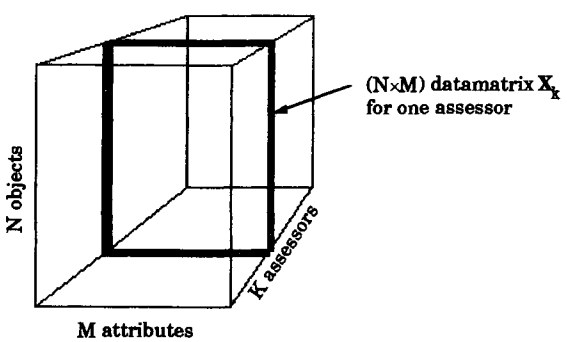


Figure 2. 3-Mode data structure representing conventional profiling data: N products are judged by K judges using M attributes.

The ($N \times M \times K$) data block in Figure 2 consists of K layers, each with the ($N \times M$) datamatrix of one assessor. Other slices of this block may be analysed but generalised Procrustes analysis focusses on the agreement of the K matrices from the individual assessors.

1.1.2 Free choice profiling

In free choice profiling the assessors are free to come up with their own attributes, which they use for judging the products. So between the assessors there is no agreement about attributes. As a result it is impossible to average the individual data, because it makes no sense to combine different attributes. The data from free choice profiling experiments must be analysed by individual difference methods, or rather 'K-sets' methods, of wich generalised Procrustes analysis is one. Unlike conventional profiling data, free choice profiling data cannot be rearranged in some kind of 3-mode data structure. Because each assessor k=1,...,K may have a different number of attributes (M_k), furthermore the jth attributes of the assessors are not the same. Figure 3 shows the structure of a FCP data set.

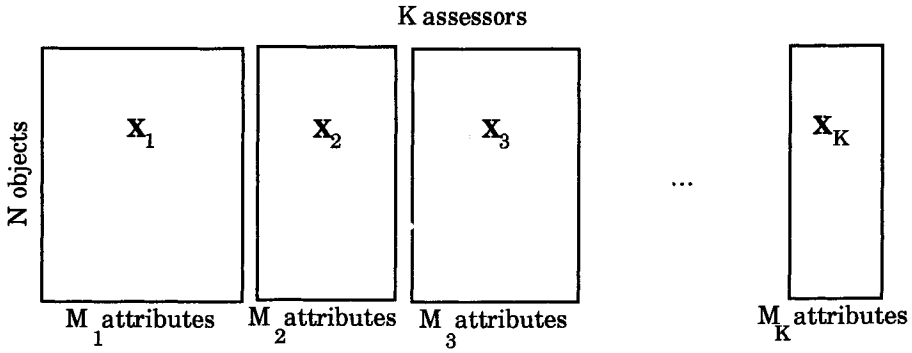


Figure 3. Data structure representing free choice profiling data: N products are judged by K judges using M_k attributes.

Figure 3 shows that the individual datamatrices $\mathbf{X}_k$ cannot be arranged such that the attributes match because each assessor's individual datamatrix has different attributes.

1.2 Sensory-instrumental Relations

One of the fields in which Procrustes analysis can be applied is the study of sensory-instrumental relations. Though Procrustes analysis appears not to be often used in this field it can be a useful method to analyse sensory-instrumental relations (see e.g. Dijksterhuis 1994). The idea behind the study of sensory-instrumental relations is that sensory perceptions have chemical/physical counterparts in the substance under investigation. A simple example is e.g. the amount of caffeine in a certain drink, which of course determines the bitterness perceived by someone drinking it. In real life the sensory-instrumental research is much more complicated, and involves multivariate, not univariate, data, and consequently needs multivariate data-analysis.

The original, not generalised, Procrustes analysis can be applied to sensory-instrumental data, because two-data sets are involved. One data set contains the sensory judgements on a number of, say N, products. The second data set contains a number of instrumental measures on the same N products. These can be results of chemical analyses, physical properties or of other measurements.

1.3 Designed experiments and incomplete data

In some cases it is conceivable that at a profiling experiment, be it conventional or free choice, the data may be gathered according to some experimental design. When the design has been an incomplete one, the datamatrices of the assessors may not all have scores on the same set of N products. In this case it is impossible to analyse these data by means of ordinary generalised Procrustes analysis. Special generalised Procrustes analysis methods that can handle missing data must be used. They are outside the scope of this chapter but can be found in Commandeur (1991) and Ten Berge, Kiers & Commandeur (1993).

2. THEORY AND BACKGROUND OF PROCRUSTES ANALYSIS

In this section generalised Procrustes analysis is introduced in two different ways, first in a geometrical way and next in a somewhat more formal mathematical way.

2.1 A geometrical look

Each assessor's datamatrix, $\mathbf{X}_k$, consists of N rows with scores on M_k attributes. This datamatrix contains elements $\mathbf{X}_{ijk}$, where i is the index over the N products, $j=1,\ldots,M_k$, the number of attributes of the kth assessor and $k=1,\ldots,K$ the number of assessors. In this section no distinction between conventional profiling and Free choice profiling will be made.

The scores in an assessor's datamatrix describe N objects using M attributes. Geometrically the N points can be seen as to lie in an M-dimensional space. With $M=2$ attributes we can draw a plane with the N points in it, but in general M will be (much) larger. Figure 4 shows a configuration of N points from the data of an assessor judging on only 2 attributes. Mathematically high dimensional spaces are no problem, though we may have trouble imagining them, but this we don't need to. When the analysis is done we don't look at the high dimensional space but at a projection onto an imaginable lower dimensional space, often two dimensions, so it can be plotted on paper. This projection is often accomplished by means of performing a principal component analysis and plotting the first two dimensions.

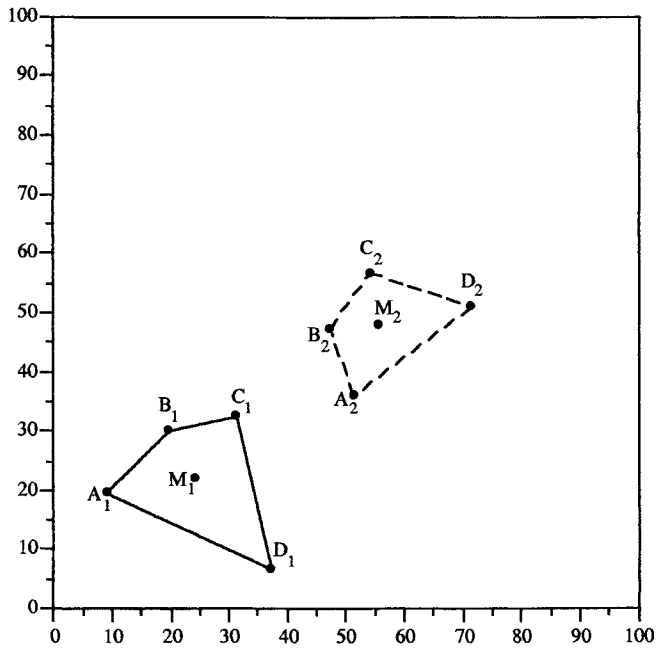


Figure 4. Two configurations with points representing scores on four products from assessor 1 (A_1, B_1, C_1, D_1) and from assessor 2 (A_2, B_2, C_2, D_2) with their centroids M_1 and M_2.

We have M-dimensional configurations of N points for all K assessors. Suppose that we deal with two assessors, to keep this example simple. We can draw the two different configurations of the N=4 points (Figure 4). The objective of generalised Procrustes analysis is to try to get the same objects as close to each other as is possible by shifting entire configurations, rotating them and reflecting them if necessary. The important underlying assumption is that the distances between the N objects for one assessor may not be changed during these transformations. When the configurations are also allowed to stretch or shrink the *relative* distances between the objects remain the same.

The distances between the objects reflect the relations between the objects. Objects close together are similar, objects far apart are different. The reason to keep the *distances* invariant is that in the process of matching, the *relations* between the N objects of one assessor should not change. Similar objects must remain similar, different objects must remain different.

2.2 Transformations

The transformations mentioned above, i.e. shifting, rotating, reflecting and stretching or shrinking, that make up a generalised Procrustes analysis, turn out to correct for a number of assessor effects (see Arnold & Williams 1985).

2.2.1 The Level-Effect: Translation

The so-called level-effect manifests itself by the different average scoring position on a line scale of different assessors. One assessor may give all N products scores that lie between, say, 5 and 25 and another assessor may use scores from 60 to 100 (assuming a 1 to 100 line-scale score). These two extreme assessors could very well perceive the objects identically, and would perhaps agree with one another completely, had not they possessed such different scaling behaviour. This level-effect can easily be corrected for by expressing the scores as deviations from the average score of an assessor on an attribute. Geometrically this results in translating the entire configuration of an assessor such that the centre of the N object-points coincides with the origin of the space (see Figure 5). The centres M_1 and M_2 in Figure 4 are shifted onto each other and this point is labled C in Figure 5. Mathematically this translation operation is known as column-centring, in 'Analysis of Variance' terms the *assessor main-effect* is removed.

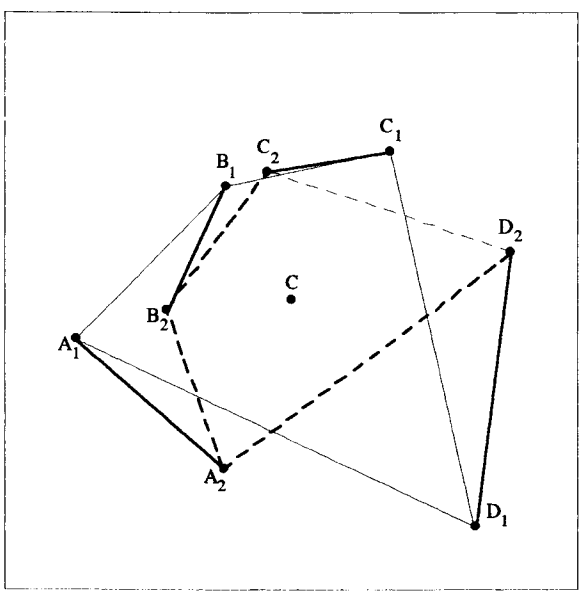


Figure 5. Centred configuration of two assessors.

2.2.2 The interpretation-effect: Rotation/Reflection

The transformations which allow for the fact that the attributes do not have to be the same (the interpretation-effect) for all assessors are rotation and reflection. The entire configuration of an assessor can be rotated to bring the N object-points in agreement with the N points of the other configurations. If necessary the configuration can be reflected in a particular dimension too. As can be seen from Figure 5, the object-points are not very close yet, the lines between the pairs

of points (A_1, A_2), (B_1, B_2) etc. indicate the distance that is to be minimised. Mathematically the rotation and (reflection) are represented in a rotation matrix $\mathbf{H}_k$, for the configuration of assessor k.

Figure 6 shows the two example configurations after rotation. Note that the N points actually are closer (A_1 to A_2, B_1 to B_2, C_1 to C_2, D_1 to D_2) than in Figure 5.

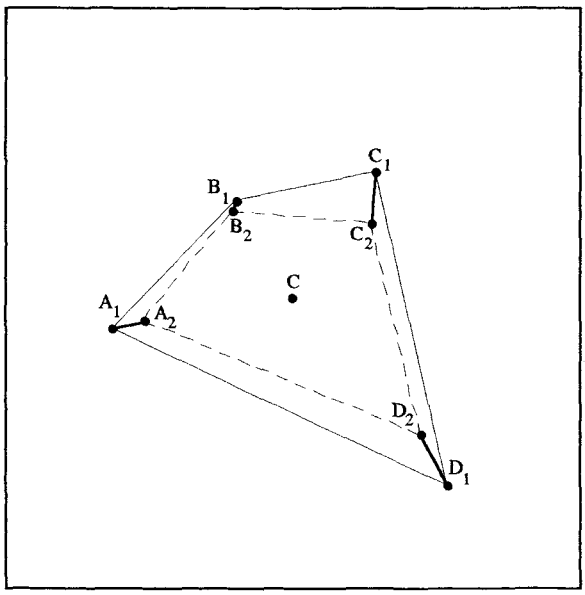


Figure 6. Configurations after centring and rotation.

2.2.3 The range-effect: Isotropic Scaling

Another individual scaling effect is the so-called range-effect. This is shown by the different ranges of scoring that the assessors use. One assessor may give scores ranging between 10 and 95 and another assessor uses scores from 60 to 80. This difference in range is another unwanted effect caused by individual differences in scoring behaviour. The underlying perception is believed not to depend on these differences in scaling range, so the effect is controlled for. The correction that is used is called isotropic scaling, which means that a configuration is shrunk or stretched in its entirity, i.e. alike in all directions of the space.

A different scaling range shows as a different extensiveness of the configurations. Figure 6 showed the two example configurations after centring and after rotation. It can be seen that the second configuration is contained whithin the first. The second assessor must have used a smaller range of the line-scale. The thick lines can now be shortend by stretching the inner configuration a little bit. The result of this operation is shown in Figure 7.

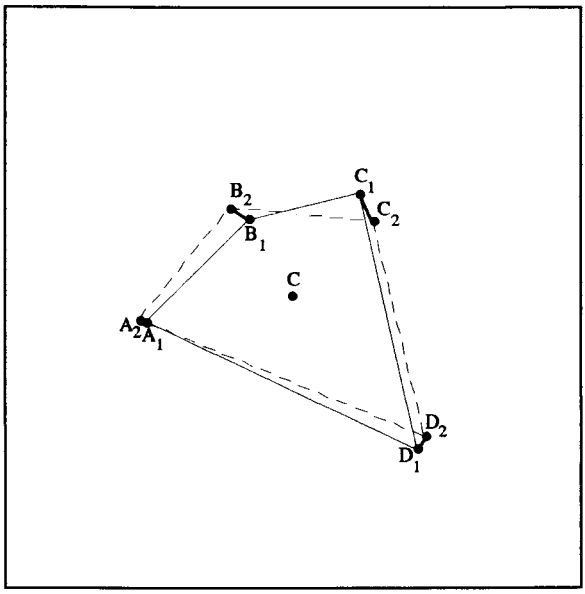


Figure 7. The two example configurations after centring, rotating and isotropic scaling.

The scaling factors are represented by a number ρ_k. A configuration k is shrunk when $0<\rho_k<1$ and stretched when $1<\rho_k$.

2.3 Generalised Procrustes analysis more formally

Mathematically the matching process is expressed by minimizing the distances between the same objects for different assessors, under the conditions that the distances between the objects of one assessor may not change. Gower (1975) gives a mathematical derivation of generalised Procrustes analysis.

The above mentioned distances can be expressed as the differences between the individual matrices:

$$\sum_{k<l}^{K} \left\| \mathcal{T}(\mathbf{X}_k) - \mathcal{T}(\mathbf{X}_l) \right\| \tag{1}$$

$\mathcal{T}(\mathbf{X}_k)$ stands for a certain transformation $\mathcal{T}$ of the matrices $\mathbf{X}_k$, and

$$\|\mathbf{M}\| = \mathrm{tr}(\mathbf{M}\mathbf{M}') = \sum_{i,j} m_{ij}^2$$

for the sum of the squared elements of **M**. The transformation $\mathcal{T}$ has to maintain relative distances between the product-points. Such transformations were introduced in §2.1, now they are presented more formally. Firstly minimising (1) can be shown to be equivalent to minimising:

$$\sum_{k=1}^{K} \left\| \mathcal{T}(\mathbf{X}_k) - \mathbf{Y} \right\| \tag{2}$$

when

$$\mathrm{Y} = K^{-1} \sum_{k=1}^{K} \mathcal{T}(\mathrm{X}_k)$$

the mean of the individudal transformed datamatrices $\mathcal{T}(\mathbf{X}_k)$. The transformations applied in Procrustes analysis are translations, rotations and isotropic scaling and they can be expressed as follows:

$$\mathcal{T}(\mathbf{X}_k) = \rho_k \mathbf{X}_k \mathbf{H}_k + \mathbf{T}_k \tag{3}$$

where ρ_k is the isotropic scaling factor, $\mathbf{H}_k$ the rotationmatrix and $\mathbf{T}_k$ the translation. The translation can be taken care of by column centring the matrices $\mathbf{X}_k$, as was shown by Gower (1975). To keep the formulae in this section from growing long, the translation is not mentioned anymore. It is assumed that the columns in $\mathbf{X}_k$ are expressed in deviations from their means. Removing the means in this way is effectively removing the assessor main effect.

The criterion minimised by generalised Procrustes analysis is the sum of all the squared distances between the individual transformed matrices which by (2) can be written as:

$$\sum_{k<l}^{K} \left\| \rho_k \mathrm{X}_k \mathrm{H}_k - \rho_l \mathrm{X}_l \mathrm{H}_l \right\| = K \sum_{k=1}^{K} \left\| \mathrm{Y} - \rho_k \mathrm{X}_k \mathrm{H}_k \right\| \tag{4}$$

Some constraints are necessary, to assure non-trivial solutions. One constraint is in the $\mathbf{H}_k$ being rotation matrices, which are orthonormal matrices, hence:

$$\mathbf{H}_k' \mathbf{H}_k = \mathbf{H}_k \mathbf{H}_k' = \mathbf{I} \tag{5}$$

A constraint on the isotropic scaling factors ρ_k is needed to prevent them from becoming zero to minimise (4) in a trivial way. The constraint scales the total variance to K, the number of sets:

$$\sum_{k=1}^{K} \left\| \rho_k \mathbf{X}_k \mathbf{H}_k \right\| = K \tag{6}$$

It has been assumed hitherto that all the matrices $\mathbf{X}_k$ are of the same order ($N \times M$), which is the case with conventional profiling data. When Free choice profiling data are analysed this assumption does not hold. In this case the $\mathbf{X}_k$ are made of the same order by padding columns of zero's until all $\mathbf{X}_k$ are of the same order ($N \times \max\{M_k\}$). See Dijksterhuis and Gower (1991/2) for some discussion about this custom. Another possibility is using Projecting Procrustes analysis (Peay 1988) which differs from the classical (Gower 1975) Procrustes analysis. The criterion maximised by Projecting Procrustes analysis is

$$K\left\|\mathbf{Y}^{[p]}\right\| = \left\|\left(\sum_{k=1}^{K} \rho_k \mathbf{X}_k \mathbf{H}_k\right)^{[p]}\right\| \tag{7}$$

where the superscript $[p]$ stands for the first p dimensions of the configuration $\mathbf{Y}$. Note that the variance contained in the resulting, p-dimensional group average space is maximised, while in the classical Procrustes analysis the residual variance beteween the corresponding objects in the entire M-dimensional individual configurations is minimised. The important difference with GPA is that the rotationmatrices are no longer proper *rotation*matrices but they include a projection onto p dimensions as well. This means that it is not necessary to pad all $\mathbf{X}_k$ to the same order. This also means that it is not needed to perform a PCA on the group average space afterwards, because this space already exists in p dimensions.

Another difference proposed by Peay (1988) is the constraint on the isotropic scaling factors ρ_k as follows:

$$\sum_{k=1}^{K} \left\|\rho_k \mathbf{X}_k \mathbf{H}_k\right\| = \left(\max\{M_k\}\right)^{-1} \sum_{k=1}^{K} M_k \tag{8}$$

which is equal to (6) in case all sets are of the same order ($N \times M$). The scaling of the variance, formula (6) or (8), does not influence the GPA solution. Dijksterhuis and Punter (1990) suggest to scale the total variance to 100. This means that all subsequent variances can be read as percentages explained, or residual, variance.

More about GPA and some variants can be found in Ten Berge (1977) and Ten Berge and Knol (1984, see also Dijksterhuis and Gower 1991/2).

2.4 Variables and dimensions

When analysing the raw data from the assessors in a sensory panel the columns of the datamatrices are the variables or attributes the assessors used in judging. When analysing sensory-instrumental data, often only two datamatrices are involved of which one may contain the results of some previous analysis like PCA or even another GPA. Such a datamatrix with the PCA or GPA result does not have variables as columns but dimensions. This is a different situation from the analysis of raw, sensory, datamatrices from different assessors. Different ways of scaling and standardizing are needed when analysing sensory-instrumental data compared to the data from a sensory panel.

The result of prior analyses (e.g. factor analysis, or MDS) will often be normalised configurations, which do not need pre-scaling for the Procrustes analysis. Different instrumental measures (e.g. pH, Instron-measures etc.) will have very different ranges and

levels of scores. In these cases standardisation of each variable may be useful. The sensory scores of a panel are much more homogeneous than different instrumental measures, so they may not need to be standardised individually.

For an application of generalised Procrustes analysis to sensory-instrumental data see Dijksterhuis (1994). Dijksterhuis & Gower (1991/2) also discuss some matters related to the pre-scaling of the datamatrices.

3. RESULTS OF A PROCRUSTES ANALYSIS

This section presents matters related to the results of a GPA. The analysis of variance is an important tool in interpreting the results, as is the PCA which enables inspection of a low-dimensional projection of the group average.

3.1 Analysis of variance

After the analysis, the distances between the corresponding points can be interpreted in the translated, scaled and rotated configurations. These distances are precisely those which are minimized by the generalised Procrustes analysis process. It is not possible to get the objects closer under the assumptions of generalised Procrustes analysis. There are different ways of looking at these distances.

3.1.1 Total fit/loss

Squaring the distances, resulting in 'variances', and adding them, gives a overall measure of loss which can be compared with the squared distances before the generalised Procrustes analysis. It is convenient to express these variances relative to the total variance before the generalised Procrustes analysis (see also Dijksterhuis & Punter, 1990). The thick lines remaining in Figure 7 cannot be made shorter, and these lines represent the *loss*, i.e. that what cannot be modelled by the GPA process. The complement of the percentage loss to 100% gives the *fit* of the obtained solution. Remember that the group average is subjected to a PCA to find a projection onto a low-dimensional space. The aforementioned *fit* can be broken down per dimension, to infer an optimal dimensionality to best represent the results in. Dijksterhuis & Punter (1990) use a scree-graph to infer an optimal dimensionality.

When the variances -squared distances between the objects- are added over the N objects, per assessor, a measure results which shows the agreement of a particular assessor with the group average. When these variances are added over the K assessors, a measure for each product can be obtained, which shows how much agreement there is among the assessors about a particular product. Both outlying assessors and products can be thus identified.

These variance measures for assessors and for objects can be split over dimenions too, this enables identification of assessors or objects which need an extra dimension, or cases in which one assessor or one object accounts for an extra dimension by itself.

3.1.2 Geometry of the variance measures

The different variances in a Procrustes analysis have a clear geometrical meaning. In Figure 8 the different variance measures ('group average', 'Residual' and 'Total') for the different products are illustrated. In this figure the position of product A is shown for three assessors

(A_1, A_2, A_3), their 'group average' point is labelled A. The variances are, as variances usually are, measured relative to the origin, labelled C (Centre).

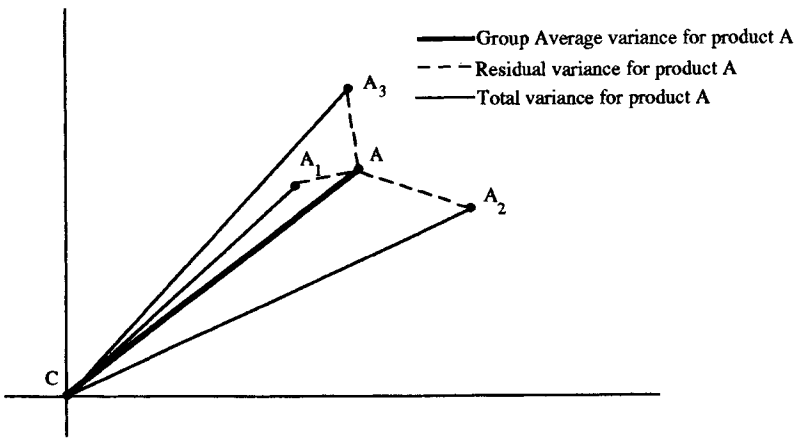


Figure 8. Geometrical interpretation of the group average, residual and total variances for the objects in a Procrustes analysis.

In Figure 8 the lines between the points represent the variances. The squared lenghts of these lines is the variance.

When the residuals or total variances are regarded per subject instead of per product, variance measures for assessors result (see Figure 9). In this figure only three assessors (1, 2 and 3) and two products (A and B) are used to keep the plot from cluttering.

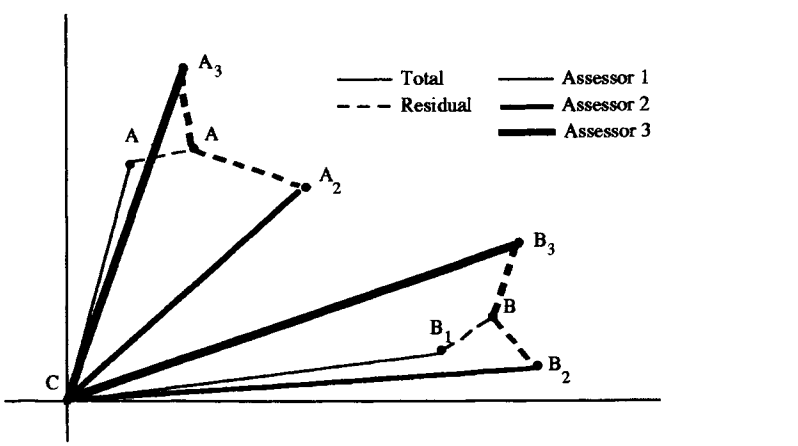


Figure 9. Geometrical interpretation of the group average, residual and total variance for two assessors in a Procrustes analysis.

In Figure 9 the dashed lines represent the residual variances, the plain lines the total variances. The three assessors are represented by different thicknesses of the lines. For assessor 1 the residual is computed by adding all its residual parts from all products, the same holds for the total parts. Note that the residuals and the totals are not part of the same 'average assessor-point', as was the case with the product-points. This is the reason that a group average variance is not available for assessors but only for products.

3.2 Principal component analysis

It is time to expand a little on the matter of the dimensionality of the solution. The classical generalised Procrustes analysis according to Gower (1975) applies all the transformations (translation, rotation/reflection, scaling) in the highest possible dimensionality, i.e. 100% of the data is involved throughout the entire analysis. When the optimal solution is obtained, it is in this high, say M, dimensional space. In order to obtain a convenient representation in a low number of dimenions, say two, a PCA is applied to the M dimensional GPA group average. This final PCA gives a number of dimensions of which the first two can be plotted for inspection. The percentages explained variances of these dimensions can be used to infer a dimensionality of the solution. Perhaps a third or fourth dimension is decided to be needed in order to interpret the results. A scree graph of the explained variance of this PCA can help in deciding on the dimensionality of the representation.

The final PCA on the group average space results in a low-dimensional representation of this space. The PCA gives this space a certain orientation. Because one wants to compare the group average space to all individual spaces, the latter are given the same orientation as the low-dimensional group average space.

3.2.1 Representing the original variables

The original variables, the attributes of a sensory panel or instrumental variables in sensory-instrumental data, can be represented in the GPA group average. Basically there are two ways of doing this. One is to use the coordinates of the rotationmatrices, these are called the loadings of the variables. These matrices, $\mathbf{H}_k$ rotate the individual datamatrices $\mathbf{X}_k$ to $\mathbf{X}_k\mathbf{H}_k$ The matrices $\mathbf{H}_k$ are of order ($M \times M$) and their rows represent the M columns of $\mathbf{X}_k$ on the new -rotated- M dimensions $\mathbf{X}_k\mathbf{H}_k$. These dimensions are represented in the $\mathbf{H}_k$ as their M columns. Plotting the column-points from $\mathbf{H}_k$ thus result in points that represent the original variables in the rotated spaces $\mathbf{X}_k\mathbf{H}_k$.

Another way of representing the original variables is to calculate their correlation with the dimensions of the group average space. Plotting the correlations results in a representation of the original variables often much alike the one using the loadings.

Theoretically the loadings may be interpreted as biplot axes (see e.g. Gower and Dijksterhuis 1994), which can be a reason to prefer the loadings. Others may prefer the correlations because sometimes they may give more explicit results. Which one prefers seems to amount to a matter of taste.

3.3 Statistical matters

There is no formal test of significance available for the results of a generalised Procrustes analysis. In Langron and Collins (1985) such a test is derived, but the assumptions may be unrealistic (cf. Dijksterhuis and Gower 1991/2). GPA is most often used as an exploratory

tool, especially in sensory analysis. Recently some papers are publised which address the matter of significance in a generalised Procrustes analysis context. King and Arents (1991) devise a test based on the analysis of random data-matrices. Their approach is the same as the one used by Langeheine (1982). In this approach random datamatrices, the size of the original data, are analysed, and this is repeated a number, say 100, of times with different random data. The position of the original result in the distribution of the results from the random data analyses, is an indication of the statistical significance of the GPA result. The permutation test approach (see e.g. Wakeling et al. 1992) uses the same distribution as the original data, in fact it uses the very same data, to obtain a measure of the statistical significance. In a permutation test the null hypothesis of no structure in the data, or no relation between the data sets, is simulated by means of permuting the rows of the datamatrices. For the permuted data set a relevant statistic, here e.g. the Procrustes-loss, is calculated. This process is repeated a large number of times, say 100. The empirical result, i.e. the Procrustes loss of the unpermuted, original, data set, is compared to the distribution of the loss-values obtained after permutation of the data sets. Analogously to the random data approach, the position of this empirical loss-value in the distribution of loss-values gives the statistical significance. In Dijksterhuis and Heiser (1995) a brief evaluation of the random data- and the permutation methods is given.

Analytical approaches (Sibson 1978, Langron and Collins 1985) to find a theoretical distribution for the Procrustes loss values, suffer from the fact that the data must follow a multivariate normal distribution, which may not occur in practice. As a result the results of these studies may not work satisfactory in practice.

3.4 Methods for missing data

Generalised Procrustes analysis as an exploratory research tool in sensory analysis presupposes a complete data set for each assessor. Until recently there were no Procrustes models which would hanlde missing values properly. Commandeur (1991) developed a generalised Procrustes analysis in which it is allowed to have arbritrary rows of individual data sets missing. The model is able to fit data sets which are of unequal row-order. This situation could arise in a sensory context when not all assessors tasted or smelled all objects because e.g. some assessors failed to appear at a certain experimental session. Ten Berge et al. (1993) expanded the method of Commandeur to include missing cells. This means that each of the individual data sets can have missing values for some products on some attributes.

3.5 Comparison with other MVA techniques

3.5.1 Procrustes variants

The original generalised Procrustes analysis is developed by Gower (1975). Earlier Procrustes analysis methods were developed to match two data sets. Table 1 presents a concise overview of the most cited contributions to the development of Procrustes analysis.

Table 1
Some papers in the history of Procrustes analysis.

Author	Year
Green	1952
Hurley & Cattell	1962
Cliff	1966
Schönemann	1968
Schönemann & Carroll	1970
Kristof & Wingersky	1971
Gower	1975
Ten Berge	1977
Ten Berge & Knol	1984
Peay	1988
Gower	1995

Another approach to generalised Procrustes analysis is described by Peay (1988). The 'classic' generalised Procrustes analysis of Gower (1975, see also Ten Berge 1977) performs all transformations in the highest possible dimensional space. The results are subjected to PCA afterwards to create a low-dimensional representation. The method according to (Peay 1988) has a different approach to make a low dimensional representation. The rotation/reflection step of the process includes a *projection* onto a low dimensional space. Hence this method will be called *projection Procrustes analysis* in contrast with *orthogonal Procrustes analysis* (see Gower 1995). A PCA is not needed afterwards. A result of the projecting approach of projecting Procrustes analysis is that the dimensions of the result of this method are not *nested.* This means that a P-dimensional solution is not the same as the first P dimensions of a $P+p$ ($p>0$) solution as is the case with classical GPA.

What method is to be preferred is perhaps more a matter of philosophy than of supremacy of one of the methods. Dijksterhuis & Gower (1991/2) compare the 'classical' Gower (1975) method with the Peay (1988) method.

3.5.2 Other MVA techniques

Before talking about the relationship of GPA with other MVA methods there are two distinctions to make:

- between 2-way methods and individual difference methods
- between 3-way and K-sets techniques.

Section II ('Analysing aggregated sensory data') treats a number of different 2-way MVA methods. These methods work on matrices that are aggregated. The aggregation is often done by means of averaging over assessors, so there are no individuals present in the data. It is argued by some (see e.g. Dijksterhuis and Punter 1991, Dijksterhuis 1995a, 1995b) that it is seldom justified to average over assessors in sensory data analysis because the attributes actually are different for each assessor, despite training of the panel.

The methods that respect the individuals in the data are called 'individual difference methods' and they are treated in Section III ('Analysing individual sensory profiles'). Two kinds of individual difference methods must be distinguished: 3-way methods and K-sets methods. There is a fundamental difference between these two methods and between the corresponding two kinds of data: 3-way data and K-sets data. Figure 2 shows the structure of

a 3-way data matrix, in sensory applications this means that all attributes are the same for all assessors. In Figure 3 it is illustrated that the attributes are different for the assessors. 3-Way MVA methods assume that all sets -the assessors- have the same variables, hence it is useless to use these methods for *K*-sets data. *K*-sets methods do not make this assumption, so they are fit for the analysis of *K*-sets data as well as for the analysis of 3-way data. Analysing 3-way data by a *K*-sets method provides a manner to find out if the variables are really commensurate in all sets.

The 3-way factor analytic methods in Chapter 10 ('Analysing individual profiles by three-way factor analysis') are, as their name suggests, 3-way methods. GPA and GCA, Chapter 7 ('Procrustes analysis in sensory research') and Chapter 8 ('Generalised canonical analysis of individual sensory profiles and instrumental data') respectively, are *K*-sets methods. Chapter 6 ('Analysing differences and similarities among products and among assessors by Multidimensional Scaling') treats Multidimensional Scaling methods, which come in a 2-way and an individual-difference variety. The individual-difference MDS methods work differently from GPA and GCA, but they effectively analyse *K*-sets data. This is because individual difference MDS methods study the relationships (distances) between the *objects* of each individual data set, so that the variables dissapear in the process. When the variables dissapear it does not matter anymore whether the data were *K*-sets, or 3-way.

4. CONVENTIONAL PROFILING

In this section a data set is analysed using the program Procrustes-PC v2.2 (OP&P, 1992, Dijksterhuis et al. 1991).

4.1 Data

The cheese data set analysed in this paragraph is made available by Matforsk and is part of a study by Hirst et al. (1994). This data set is also analysed in Dijksterhuis (1995a) in the context of a study of 'panel consonance', i.e. the agreement of the individuals in a sensory panel on each attribute separately.

The data consist of the scores of 10 judges scoring 12 kinds of hard cheese using 19 attributes. The QDA procedure (Stone and Sidel 1985) is used so the data are 'conventional profiling' data. GPA is applied to this data set to study individual differences between judges and to construct a 'group average' configuration of the 12 cheeses. It is the same data set that is studied in the chapter on 3-way factor analysis (chapter 4.4, 'Analysing individual profiles by three-way factor analysis'). The analysis in this chapter is to illustrate the method of generalised Procrustes analysis, it is not meant as a study of the cheese data.

Each of the 12 cheeses is presented twice to each subject. Each replication is analysed as a separate 'product' in the GPA, so 24 'products' are used in the analysis.

4.2 Dimensionality of the GPA group average

Most often the results of GPA are displayed in a two dimensional plot. At this point it is useful to consider the differences between the projection Procrustes analysis according to Peay

(1988) and the original Procrustes analysis according to Gower (1975). The differences between the two methods will be illustrated using the cheese data.

4.2.1 Projection Procrustes analysis

This variant of GPA combines the Procrustes transformations with a projection onto a low dimensional space. This means that when the researcher chooses to calulate a two-dimensional GPA solution, the data are projected onto a 2-dimensional space and that higher dimensions are not used for the calculation of the optimal solution. This does not mean that the solution is sub-optimal, it is the best solution in two-dimensions, but at the cost of loosing sight of any interesting information that could have been captured in the third, forth or higher dimensions. To be sure, in addition a three-, four-, five-, etc. dimensional analysis should be carried out.

4.2.2 Classical Procrustes analysis

The original GPA applies all Procrustes transformations in the full dimensional configuration, and the result of the analysis is a group average in the maximum number of dimensions possible. Any potentially interesting information is available. The resulting high-dimensional configuration is subjected to a principal component analysis in order to be able to give a low dimensional representation of it. The researcher can *a posteriori* decide to use only two or three dimensions of the total result.

4.2.3 Cheese group average

To find-out the optimum dimensionality to represent the group average in, all dimensions are considered. Note that a projection Procrustes analysis with the maximum number of

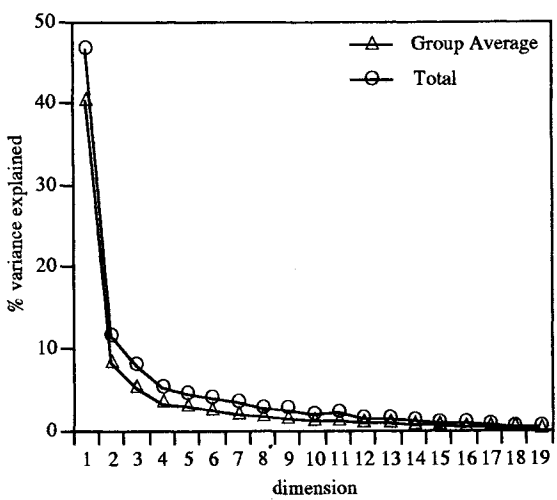


Figure 10. Percentages group average- and total variance explained in the dimensions of the group average space.

dimensions is identical to the classic Procrustes analysis because there are no dimenions left to project onto. In this case the projections are onto the full dimensional space, which is of course the same thing as not projecting at all.Figure 10 presents a scree-graph in which the percentages explained variance of all dimensions are shown.

Figure 10 shows that approximately 40% of the group average variance and 46% of the total variance is explained in the first dimension. Remember that the total variance is the variance explained by the configurations of all the assessors. When these configurations are averaged, becoming the 'group average' configuration, the group average variance remains. The averaging of individual configurations results in the loss of variance. It is exactly this loss, the residual variance, which is minimized by the classical orthogonal procrustes analysis. The projection procrustes analysis maximises the group average variance, in a particular number of dimensions. In this full-dimensional analysis, the two are identical.

Table 2

Cumulative explained variance for the dimensions of the group average and of the individual configurations ('Total') of the GPA result of the (10 × 12 × 19) Cheese data set.

Dimensions	group average	Total
1	40.37	46.62
2	48.59	58.09
3	53.78	66.08
4	57.08	71.19
5	59.92	75.61
6	62.33	79.62
7	64.32	82.98
8	65.97	85.73
9	67.48	88.28
10	68.68	90.25
11	69.81	92.38
12	70.73	93.87
13	71.55	95.33
14	72.20	96.51
15	72.76	97.51
16	73.22	98.40
17	73.59	99.12
18	73.86	99.62
19	74.04	100.00

In Table 2 it can be seen that a two dimensional solution explains 48.59% variance of the group average configuration. The total variance in two dimensions is 58.09%, i.e. the variance explained by the inidividual configurations of all the assessors. Both from Table 2 and Figure 10 two- or three-dimensions seem enough to represent the results in. When we decide that two (or three) dimensions will suffice we can use the first two (three) dimensions of the results of the full-dimensional analysis above. Alternativeley we can perform a new analysis using the projecting Procrustes technique in two (or three) dimensions, which will result in a slightly increased fit in the first two (three) dimensions. The disadvantage is that there are no higher dimensions available, all higher dimensions are explicitely regarded as noise by this decision. Table 3 shows the percentage variance explained by these additional Projecting Procrustes Analyses[1].

Table 3

Cumulative explained group average variance for the separate 2 and 3 dimensional Projection Procrustes Analyses of the Cheese data (corresponding classic GPA percentage from Table 2 between brackets).

Dimension	2D analysis	3D analysis
1	49.764	40.628
2	49.540 (48.59)	49.238
3	-	54.940 (53.78)

Table 3 shows a slight increase in explained variance for the dimensions of the group average compared with the results in Table 2. For the presentation of the cheese data analysis we will choose the result of the two dimensional projection Procrustes analysis.

4.3 Group average configuration

One of the most interesting results from a Procrustes analysis is the 'group average-', or 'Consensus-'configuration. This configuration contains the products, here the 24 cheeses. Figure 11 shows this configuration.

1 Note that the PROCRUSTES-PC v2.2 program, that was used for the analyses in this chapter, allows for both 'classical' orthogonal procrustes analysis and projection procrustes analysis. Most other Procrustes software is based on the 'classical' orthogonal procrustes analysis

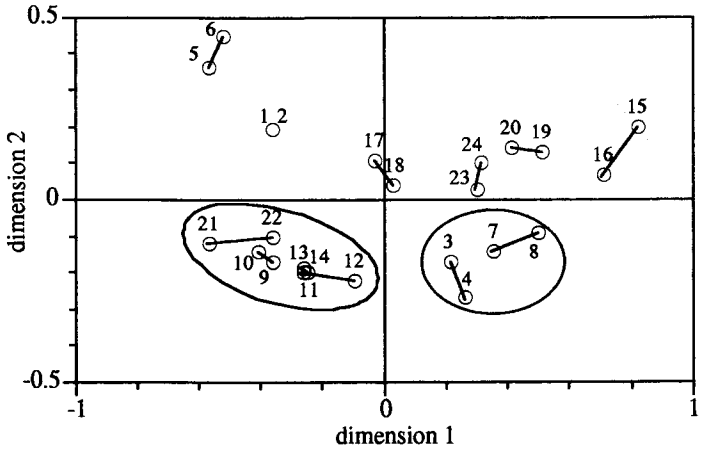


Figure 11. Group average configuration of the Procrustes analysis of the cheese data. Replicate cheeses are connected by a line.

The configuration in Figure 11 shows the 24 cheeses. The replicates are connected by a line. Including replicates in a profiling study is very important, especially when the data are analysed by Procrustes analysis or another multivariate analysis. In Figure 11 the lines connecting the replicates are relatively short, which is an indication that the judges assessed the replicates almost identically, hence an indication of the validity of the obtained result. In this case interpretations of this configuration can be made safely.

Taking a closer look at Figure 11 reveals some groups of cheeses. Two releatively clear groups are indicated in the figure. At the lower left part are the cheeses (9, 10, 11, 12, 13, 14, 21, 22), at the lower right part of the plot are (3, 4, 7, 8). At the upper right part there is a group, though looser than the previous two groups, that seems to consist of the cheeses (15, 16, 19, 20, 23, 24). At the upper left part of the figure clearly the pair (5,6) is different from the other cheeses. Cheese number 1 and 2 lie in that part of the plot too. The numbers 17 and 18 lie almost at the centre of the plot, this usually means that there is no clear agreement between the judges on these cheeses. The numbers 17 and 18 will probably show a relatively high residual variance. The Procrustes 'analysis of variance' can be used to further interpret the results.

4.4 Analysis of variance

In this section the Procrustes analysis of variance tables are shown and interpreted. To illustrate the tables they are plotted as bar-charts (cf. Dijksterhuis and Punter, 1990).

4.4.1 Analysis of variance for objects (cheeses)

Figure 12 shows the group average (explained) variance and the residual (not explained) variance for the 24 cheeses. The 'Total' variance can directly be read from the plot as the total hight of the bars because:

total variance = residual variance + group average variance

The order of the cheeses along the horizontal axis in Figure 12 is in increasing size of their residual variance.

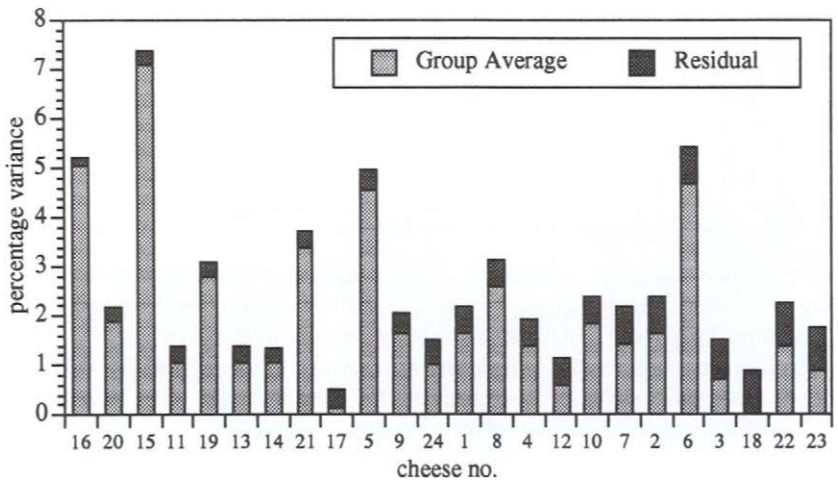


Figure 12. Percentage variance explained (group average) and unexplained (Residual) for the cheeses. The order of the cheeses is in increasing size of Residual variance.

The cheeses at the left hand side of Figure 12 have the smallest residuals. This means that there was not much difference between the scores of the assessors on these cheeses. The panel agreed well on these cheeses.

The cheeses with a larger part of residual variance (right hand side of the picture) did not fit well in the group average, there were differences between the scores of the assessors. In Figure 12 the cheeses 3, 18, 22 and 23 have relatively large residual variances. There must have been less agreement on these cheeses.

4.4.2 Analysis of variance for assessors

In this section the residual variances for assessors are studied. Table 4 shows the residual variance per assessor.

Table 4.

Percentage unexplained ('residual') variance for the assessors.

judge no	Residual
9	0.772
7	0.818
5	0.970
2	1.015
10	1.075
4	1.344
6	1.418
1	1.489
3	1.624
8	1.882

From Table 4 it can be seen that assessors 1, 3 and 8 have the highest residual variances. These assessors' individual configurations of the 24 cheeses differ most from the group average configuration. Assessors 7 and 9 are among the lowest-residual assessors.

When selecting an analytical sensory panel, an homogeneous group of judges is desired. Suppose that a selection of judges is to be made from Table 4, judges with high residual variances will be deleted from the panel, or subjected to extra training.

4.4.3 Individual configurations

To illustrate differences between individual configurations Figure 13 and Figure 14 show the group average position of the 24 cheeses, connected with the position of the same cheeses in the individual result of respectively assessor 7 -a low-residual assessor- and assessor 8 -a high-residual assessor.

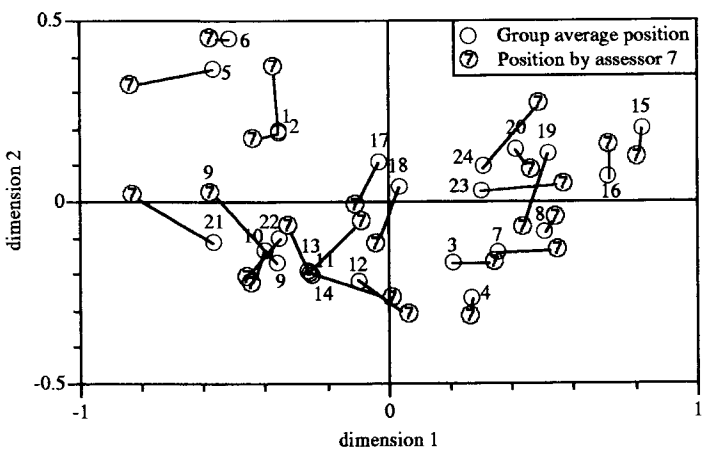


Figure 13. Group average position of the 24 cheeses, connected to the position of the same cheeses according to the configuration of assessor 7.

Though the differences are not large, over-all the lines in Figure 14 are longer than the lines in Figure 13. The sum of the squared lengths of the lines is the residual variance for the assessors 7 and 8, and is shown in Table 4.

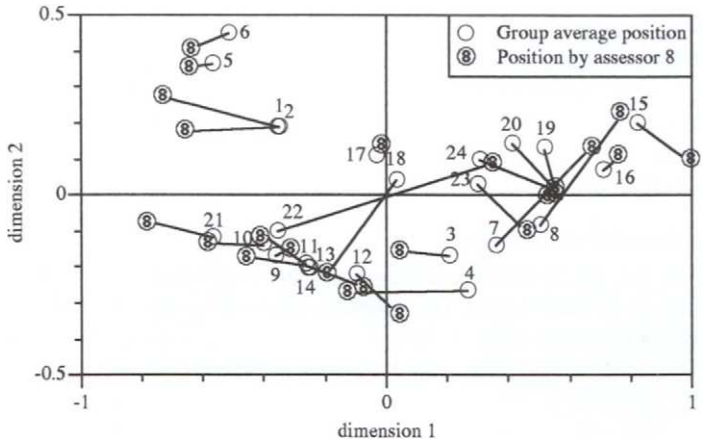


Figure 14. Group average position of the 24 cheeses, connected to the position of the same cheeses according to the configuration of assessor 8.

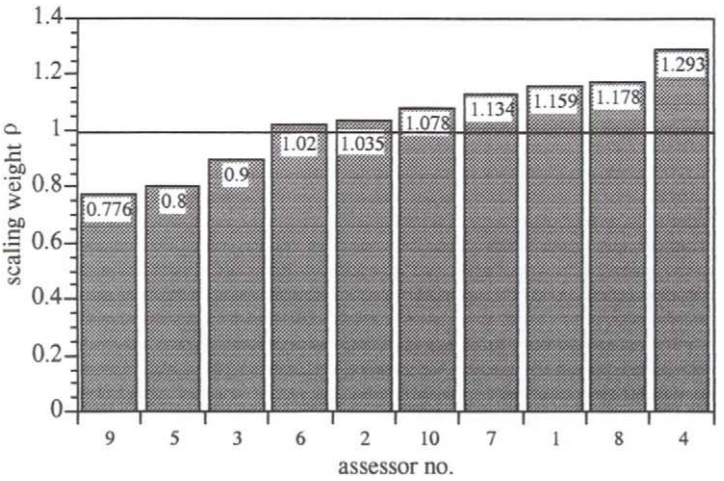


Figure 15. Isotropic scaling factors (sorted) for the 10 individual assessors' configurations.

4.5 Scaling factors

The isotropic scaling factors (see §2.2.3) reflect the amount of stretching or shrinking the individual configurations of the 10 assessors underwent in the Procrustes analysis. Figure 15 presents the 10 scaling weights.

The horizontal line in Figure 15 is at $\rho = 1$. Bars extending above this line show stretched configurations ($\rho > 1$), bars below this line represent shrunk configurations ($0 < \rho < 1$). Assessor 9, 5 and 3 have their configurations shrunk, they used a larger range of scores than the other assessors. It's the other way around for assessors 1, 4 ,7 and 8, their configurations are stretched. They used a limited range of scores. The assessors 6, 2 and 10 had their configurations hardly changed by the scaling.

4.6 Representing the original variables

Until now the objects, i.e. the 24 cheeses, and the assessors are studied. The 19 attributes the assessors used remain to be studied now. The attributes can be subdivided into odour-, flavour and texture attribtutes and are presented in Table 5.

Table 5.

Attributes used in the cheese study (Hirst et al. 1994).

	odour		flavour		texture
1	odour intensity	7	flavour intensity	15	hardness
2	creamy/milky	8	creamy/milky	16	rubbery
3	ammonia/sulphur	9	sour	17	doughy
4	nutty	10	ammonia	18	grainy
5	sour	11	nutty/fruity/sweet	19	sticky
6	other	12	bitter		
		13	salty		
		14	other (cheddar)		

Note that the arrangement in the table does not indicate any relation between attributes in the same row.

The loadings or correlations from the Procrustes analysis output give representations of the original attributes. Both the coordinates of the loadings and of the correlations can be used to draw the original attributes in the group average configuration. In this example the correlations will be used.

Each assessor used these 19 attributes, this means that each individual configuration contains 19 attributes. The total configuration with all judges together will consequently contain $10 \cdot 19 = 190$ attributes. These are far too many attributes to draw in a picture. With conventional profiling data, like this cheese data set, it is possible to average the attributes over the assessors, to make group average attributes. This is analogous to the averaging of the individual product-positions to make group average product points. Figure 16 presents the resulting group average attribute points based on the correlations of the original attributes with the group average dimensions. Of course averaging is only justified with a reasonable fit. When the fit is very low, the group average configuration is to be doubted.

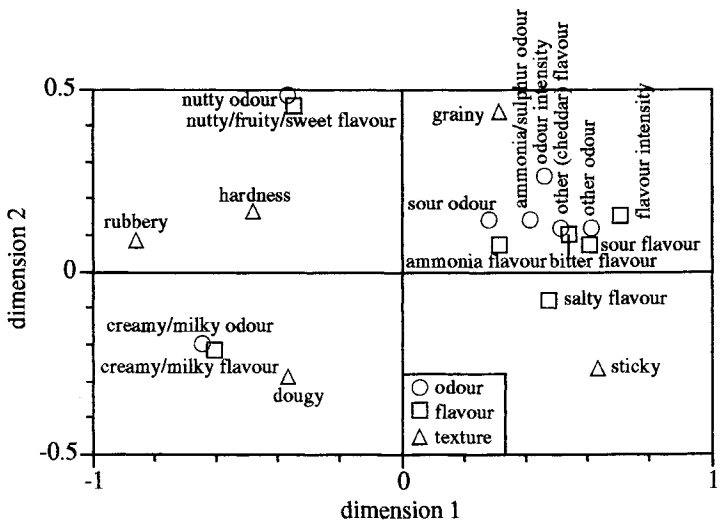


Figure 16. Averaged ('group average') attributes in the group average space of the cheese data.

In Figure 16 the relations between the attributes can be inferred. It shows that some odour and flavour attributes match: nutty odour with nutty flavour, creamy/milky odour with creamy/milky flavour, sour odour with sour flavour, ammonia odour with ammonia flavour, and odour intensity with flavour intensity. The texture attributes seem to divide the cheeses into sticky, doughy, grainy and rubbery/hardness. When dimensions have to be reified the first dimension may be approximately interpreted as a bitter/sour odour/taste and rubbery versus sticky dimension, and the second dimension as a nutty/sweet odour/taste and doughy versus grainy dimension.

Figure 16 can be compared to Figure 11 to infer properties of the cheeses. The lower left group of cheeses (9, 10, 11, 12, 13, 14, 21, 22) appear to be characterised by creamy/milky flavour and taste, their texture is mainly doughy. The cheeses 9, 10, 21, 22 seem to tend to a rubbery texture. The group at the lower right part of the plot (3, 4, 7, 8) has a sticky texture and a somewhat more salty and sour flavour. The cheeses 19, 20, 15 and 16 have a high flavour and taste intensity, a bitter/sour/ammonia flavour/taste. Because these cheeses lay opposite to the texture attributes rubbery and hardness, they do *not* have these properties, they are mainly soft cheeses. The cheeses number 5 and 6 (1 and 2 to a lesser extent) are the nutty/fruity/sweet cheeses. These cheeses are among the harder, more rubbery and grainy cheeses.

The above interpretation of the GPA group average space and the positions of the correlations of the original attributes is a kind of *biplot*-interpretation. For more about biplots in a GPA context see Gower and Dijksterhuis (1992), for biplots in general see (Gabriel 1971, Gower 1992).

5. FREE CHOICE PROFILING

In this section a free choice profiling data set is analysed by means of GPA. For this analysis the Procrustes-PC v2.2 program (OP&P 1992) was used. This same data is analysed by GCA in Chapter 8 too.

5.1 Data

The data consist of the judgements of 20 different mineral waters by eleven assessors[2]. Each assessor used her/his own attributes to judge the waters, so the data are FCP data. FCP data can only be analysed by an individual difference method of the *K*-sets type, or an individual-difference MDS method (e.g. Indscal; Carrol and Chang 1970). In this section GPA is used to analyse this data set. What is presented here is a standard GPA analysis of an FCP data set. The GPA method used is the classic GPA (Gower 1975) so the smaller data sets are padded with zero's to make all sets of the same order.

Some of the 20 mineral waters were presented, blindly, two or three times. These replications are very useful, they will be represented as connected points in the Group Averge plot (compare Figure 11).

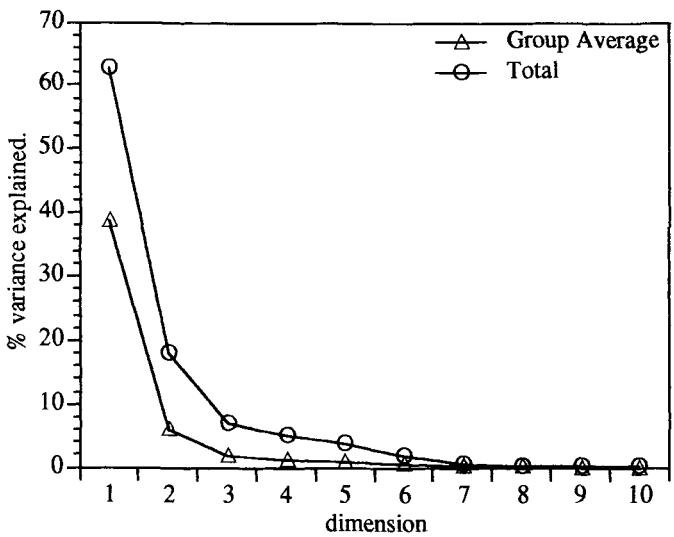


Figure 17. Percentages group average and total variance explained by the dimensions of the GPA group average space.

---

2 The data were made available by Dr. Pascal Schlich, INRA, Dijon, France.

5.2 Analysis of variance

First the dimensionality to represent the results in must be chosen. To this end the explained variance, distributed over the dimensions are needed. Figure 17 presents this scree for the dimensions of the water data.

The scree for the total variance shows that the first dimension explains 63% of the variance, the second dimension adds about 18%, the third adds another 7%. The line for the group average variance has the same shape, but the variances are lower. They should be, the differences are the residual variances per dimension. It seems that a 2-dimensional solution would do as a reasonable approximation of the data. In Table 6 the cumulative percentages explained variance are given, a two-dimensional solution explains 81% in all individual configurations together and 45% in the group average.

Table 6

Cumulative explained variance for the dimensions of the group average and of the individual configurations (Total) of the GPA result of the water data.

dimension	group average	Total
1	38.837	62.908
2	44.729	80.963
3	46.607	88.084
4	47.918	93.231
5	48.722	97.078
6	49.091	99.05
7	49.155	99.46
8	49.202	99.765
9	49.231	99.945
10	49.240	100

Note that in Table 6 the total explained variance in 10 dimensions is 100%, as it should be because in the maximum dimensionality all data are included and of course nothing is lost.

5.3 Configurations

5.3.1 Group average configuration

Figure 18 shows the GPA group average configuration of the 49 mineral waters. Replications are connected by lines. It can be seen from Figure 18 that e.g. water no. 15, 16 and 17, are judged more different than the waters 21 and 22, because the lines connecting the former are much longer than the lines connecting the latter.

Figure 18 enables identification of four approximate groups. At the left are two waters: 15, 16, 17 and 21, 22. Somewhat more to the right are two other waters: 1, 2 and 43, 44. The big cluster of the remaining waters may be subdivided into the waters 25, 26; 47, 48, 49; 10,11, which appear at the rightmost bottom part. This group also includes 45, which is connected by a rather long line to 46, these two waters were not very consistently assessed, they are rather far apart. In the set remaining waters it is hard to distinghuish separate groups. Note that the lines connecting replicates cross through this cluster, so there apear no clearly separated groups of waters.

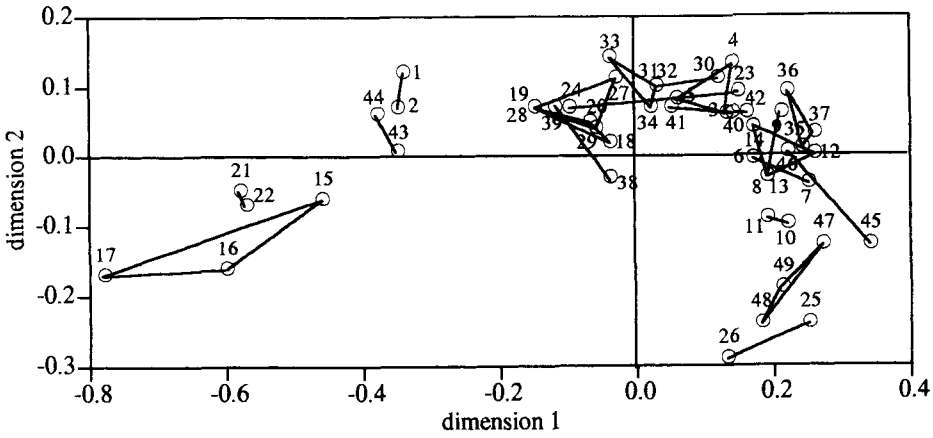


Figure 18. GPA group average configuration of the 49 mineral waters. Replicate waters are connected by lines.

Figure 19 shows the residual and total variances of the ten judges in the mineral water data set. Note that all the residual variances are about equal, and the total variances differ.

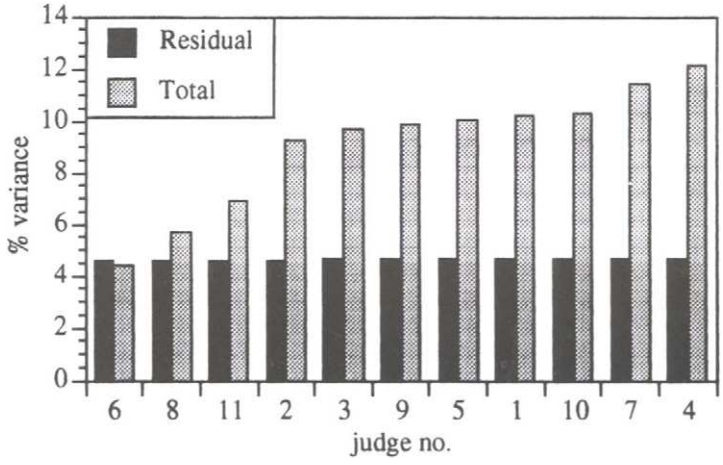


Figure 19. Total and residual variances of the assessors in the water study.

The judges 6 and 4 differ the most in the amount of total variance. It may be interesting here to study the total variances per judge, for dimensions separately. Table 7 presents these variances.

Table 7
Total variance in the first four dimensions, per judge.

judge no.	1	2	3	4
1	8.920	0.509	0.134	0.486
2	5.895	1.831	0.598	0.224
3	5.293	1.819	1.444	0.572
4	9.361	0.725	0.285	0.498
5	4.891	3.540	0.451	0.320
6	1.134	1.933	0.288	0.527
7	7.352	1.401	1.011	0.572
8	2.439	2.210	0.969	0.051
9	7.605	0.514	0.816	0.553
10	6.993	1.811	0.633	0.545
11	3.027	1.761	0.490	0.799

Table 7 shows that some judges (1, 4, 9) have a large proportion of variance in the first dimension, and relative low proportions in the second. In contrast judge 5, 6, 8 have relatively much more variance in the second dimensions, compared to what they have in the first.

5.4 Scaling factors

Table 8 gives the isotropic scaling factors for the individual sets. Two sets (7 and 10) needed to be stretched by a factor 2.6 and 2 respectively. Apparantly the judges 7 and 10 used a rather small range of scores. Judge 11 used a large range of scores, the corresponding configuration is shrunk by a factor 0.6.

Table 8
Scaling weights of the judges in the GPA of the water data.

judge	weight
11	0.627
6	0.769
8	0.785
1	0.855
2	0.920
3	0.959
4	1.007
9	1.065
5	1.389
10	2.001
7	2.571

5.5 Representing the original variables

Table 9 shows the attributes and their use by the assessors. Note that 8 of the 11 judges used the term bitterness, the terms neutral and metal were used by six assessors. There are a lot of

unique attributes, i.e. which were only used by one assessor. Notably assessor 4 and 5 generated most unique attributes.

Table 9

Attributes used in the FCP of the mineral waters and the no. of the judge that used it.

Attribute	judge no.	Attribute	judge no.
bitter	1, 2, 3, 5, 6, 8, 9, 11	balanced	4
neutral	1, 2, 4, 6, 8, 9	persistent	4, 6
taste	1	mineral	5
metal	1, 3, 7, 9, 10, 11	stagnant	5
fluid	1	river	5
salty	2, 4, 7, 8	cool	5
earth	2, 4, 7, 11	sugar	6
hard	2	old	6
acid	3, 4, 11	mushroom	7
paper	3, 10	milky	7
flat	4, 5	energic	9
dry	4	hazelnut	10
pungent	4	soft	11
rubber	4		

Note that the arrangement in the table does not indicate any relation between attributes in the same row.

For the interpretation of the clusters of waters that appeared in Figure 18 it is needed to represent the original attributes in the group average configuration. In Figure 20 the correlations of the original attributes with the dimensions of the group average space are presented.

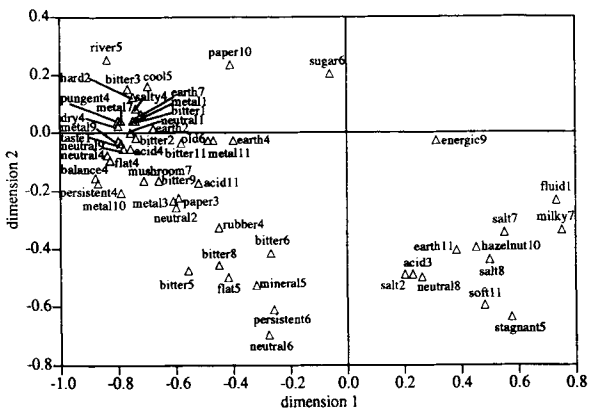


Figure 20. Configurations of the individual attributes based on the correlations of the attributes with the dimensions of the group average. The number following the term indicates the assessor number.

In Figure 20 there appear two main groups of attributes. In the lower right quadrant: fluid, milky, hazelnut, salt (2, 7, 8), soft, stagnant, neutral (8), acid (3), earth (11) (assessor numbers between parentheses) and in the upper left quadrant most other attributes.

Inspecting the configuration in Figure 20 enables one to draw some interesting conclusions with respect to the use of the attributes. The term 'metal' was used by six assessors (1, 3, 7, 9, 10, 11), and there seems to be reasonable agreement between them. This agreement is larger than the agreement found between the use of the term 'neutral', also used by six assessors (1, 2, 4, 6, 8, 9). The agreement on 'bitter' appears less than that on 'neutral', assessors 5, 6 and 8 have scored bitter somewhat differently than assessors 1, 2, 3, 9, 11.

Figure 20 and Figure 18 show the same plane, so they can be superimposed. This results in a kind of biplot, containing both the mineral water object points and the positions of the attributes. This plot is not presented here -it would be too cluttered-, but through comparison of the figures Figure 20 and Figure 18 a biplot-like interpretation can also be given. The set of waters {15, 16, 17, 21, 22} lies in a region in the plane that is characterised by a lot of attributes, including most 'metal' and 'bitter' attributes. The set {25, 26, 47, 48, 49, 10, 11, 45} lies in a region characterised by the attributes salt (for 3 assessors), soft, stagnant, acid, earth, hazelnut, milky, fluid. The remaining waters are mainly characterised as not having a certain property. Most lay opposite the remaining attributes.

It is conceivable that mineral waters have rather low amounts of clear tastes. So, after the attributes are generated, a lot of the waters will turn out not to possess this attribute, or just have it in a very low intensity. In addition, when the tastes are not clear, the differences between the assessors may become rather outspoken. A clear bitter taste may not cause much confusion in a sensory panel, but when the taste is only just above the detection threshold, as it may be in mineral waters, individual differences may arise. This could result in the use of other terms.

6. ALGORITM AND SOFTWARE FOR PROCRUSTES ANALYSIS

6.1 Generalised Procrustes analysis algorithm

The original generalised Procrustes analysis algorithm is presented in Gower (1975). Ten Berge (1977) presents a slightly modified algorithm. These algorithms concern the heart of the Procrustes analysis: the rotation and isotropic scaling. In a somewhat broader view, and in most applied situations a Procrustes analysis consists of three different parts:

- Pre-steps (translation, 'pre-scaling')
- Analysis (rotation/reflection, isotropic scaling)
- Post-steps (PCA, analysis of variance)

6.1.1 Pre

The pre-steps consist of the translation operation which amounts to centering the individual datamatrices $\mathbf{X}_k$. It is also possible to give differential weights to sets or to variables. This is all pre-scaling, it is not part of the actual Procrustes analysis. Depending on the wishes of the analist the data may be pre-scaled to have a certain total variance.

6.1.2 Procrustes analysis

The heart of the analysis consists of the two Procrustes transformations, rotation/reflection and isotropic scaling. The rotation/reflection is computed for all individual matrices $\mathbf{X}_k$ to fit the group average matrix. This computation results in a rotationmatrix $\mathbf{H}_k$. The reflection is a part of this rotationmatrix and will not be mentioned any further for this reason. After each individual rotationmatrix is computed, the new rotated individual matrix is $\mathbf{X}_k\mathbf{H}_k$ and the group average matrix is recomputed (see Ten Berge 1977). This is repeated for all sets k=1,...,K. After one run over the K sets the isotropic scaling is performed.

The isotropic scaling factors ρ_k are computed for each $\mathbf{X}_k$. At this point one iteration of the generalised Procrustes process is completed and a new average matrix, now with inclusion of the scaling factors, is computed. One iteration is seldom enough. The decrease of the sum of squared distances between the individual sets $\rho_k\mathbf{X}_k\mathbf{H}_k$ over two subsequent iterations is taken as the criterion to judge whether a satisfactory result is obtained. This criterion is usually set to a very small value, e.g. 0.001. After a number of iterations the criterion approaches this value and will finally become smaller than 0.001. Then the process is said to have converged and the iterative process is terminated.

6.1.3 Post

As said before, the result is in the highest possible dimensionality and PCA is applied to the resulting average configuration. Suppose we take two dimensions from this PCA to inspect the group average space. In order to be able to compare this two-dimensional representation with the individual sets, the individual matrices $\rho_k\mathbf{X}_k\mathbf{H}_k$ are given the same orientation as the PCA result of the group average. We must assure that we compare the individual sets and the group average in the same plane.

Further post-steps include the computation of several ways of partitioning of the residual, explained and total variance, and the computation of the correlations of the original attributes with the dimensions. Finally the tabling and plotting of the results is the obvious final step of the Procrustes program.

6.2 Software for Procrustes analysis

There are several computer programs available that can perform a GPA. A macro in the GENSTAT language was written by Arnold (1986). Schlich (1989) wrote a GPA macro in the SAS IML language. These programs work fine, but have the disadvantage that they run as macro's whithin a large statistical program. The user needs to be able to 'speak' either SAS or GENSTAT. In 1988 a special Procrustes program for the personal computer was developed, which is called Procrustes-PC (OP&P 1988, Dijksterhuis and van Buuren 1988). At the moment version 2.2 is the latest one (OP&P 1992, Dijksterhuis et al. 1992). Recently a new GPA program -Procrustes for Windows- has been developed (OP&P 1995).

7. ACKNOWLEDGEMENT

The author wishes to thank Paul Arents (Quest, Naarden, the Netherlands) for some comments on an earlier version of this chapter.

8. REFERENCES

Arnold, G.M., 1986. A generalised Procrustes Macro for sensory analysis, Genstat Newsletter, 18, 61-80.

Arnold, G.M., Williams, A.A., 1985. The use of generalised Procrustes Techniques in sensory analysis, In: Statistical Procedures in Food Research, Piggot, J.R. (Ed.).

Banfield, C.F., Harries, J.M., 1975. A technique for comparing judges' performance in sensory tests. J. Fd. Technol., 10, 1-10.

Carroll, J.D., Chang, J.J. (1970). analysis of individual differences in multidimensional scaling via n-way generalization of 'Eckhart-Young' decomposition. Psychometrika, 35, 283-319.

Cliff, N. (1966), Orthogonal rotation to congruence. Psychometrika, 31,33-42.

Commandeur, J.J.F., 1991. Matching Configurations. DSWO Press, Leiden.

Dijksterhuis, G.B., 1994. Procrustes analysis in studying sensory-instrumental relations. Food Quality and Preference, 5, 115-120.

Dijksterhuis, G.B., 1995a. Assessing Panel Consonance. Food Quality and Preference, 6, 7-14.

Dijksterhuis, G.B., 1995b. Multivariate data analysis in sensory and consumer science. Thesis. Dept. of Datatheory, University of Leiden, the Netherlands.

Dijksterhuis, G.B., Buuren, S. van, 1988. Procrustes-PC Version 1.0 Manual. Utrecht: OP&P.

Dijksterhuis, G.B., Gower, J.C., 1991/2. The Interpretation of generalised Procrustes analysis and Allied Methods, Food Quality and Preference, 3, 67-87, Elsevier.

Dijksterhuis, G.B., Heiser, W.J., 1995. The role of permutation tests in exploratory multivariate data analysis. Food Quality and Preference.

Dijksterhuis, G.B., Kraakman, H., Buuren, S. van, 1992. Procrustes-PC Version 2.2. Manual. Utrecht: Oliemans Punter en Partners.

Dijksterhuis, G.B., Punter, P.H., 1990. Interpreting generalised Procrustes analysis 'analysis of variance' tables, Food Quality and Preference. 2, 255-265.

Gabriel, K.R. (1971) The biplot graphical display of matrices with application to principal component analysis. Biometrika, 58, 453-467.

Gower, J.C., 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 3&4, 325-338.

Gower, J.C., 1975. generalised Procrustes analysis, Psychometrika, 40, 1, 33-51.

Gower, J.C., 1992. Biplot geometry. RR-92-02 Leiden: Department of Data Theory.

Gower, J.C., 1993, Orthogonal and Projection Procrustes analysis (draft, july'93)

Gower, J.C., 1995. Procrustes Methods. Manuscript.

Gower, J.C., Dijksterhuis, G.B., 1992. Coffee images: A study in the simultaneous display of multivariate quantitative and variables for several assessors. Is published.

Green, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. Psychometrika, 17, 429-440.

Harries, J.M., MacFie, H.J.H., 1976. The use of a rotational fitting technique in the interpretation of sensory scores for different characteristics. J. Fd. Technol. 11, 449-456.

Hirst, D., Muir, D.D., Næs, T., 1994. Definition of the organoleoptic properites of hard cheese: a collaborative study between Scottish and Norwegian Panels, International Dairy Jounal, 4, 743-761.

Hirst, D., Næs, T., 1994. A graphical technique for assessing differecnes among a set of rankings. J. of Chemometrics, 8, 81-93.

Hurley, J.R., Cattell, R.B., 1962. The Procrustes Program: Producing Direct Rotation to Test a Hypothesized Factor Structure, Behavioral Science, 7, 258-262.

King, B.M., Arents, P., 1991. A statistical test of consensus obtained from generalised Procrustes analysis of sensory data, Journal of Sensory Studies, 6, 37-48.

Kristof, W., Wingersky, B., 1971. Generalization of the orthogonal Procrustes rotation procedure to more than two matrices. Proceedings of the 79th Annual Convention of the American Psychological Association, 6, 89-90.

Langeheine, R., 1982. Statistical evaluations of measures of fit in the Lingoes-Borg Procrustean individual differences scaling, Psychometrika, 47, 4, 427-442.

Langron, S.P., Collins, A.J. (1985). Perturbation theory for generalised Procrustes analysis, J.R. Statist. Soc. B, 47, 277-284.

OP&P, 1988. PROCRUSTES-PC Version 1.0. Utrecht: OP&P.

OP&P, 1992. PROCRUSTES-PC version 2.2. A personal Computer Program for generalised Procrustes analysis. Utrecht: OP&P Software Development.

OP&P, 1995. Procrustes for Windows. Utrecht: Oliemans Punter & Partners.

Oreskovich, D.C., Klein, B.P., Sutherland, J.W., 1991. Procrustes analysis and Its Applications to Free-Choice and Other Sensory Profiling, In: Sensory Science Theory and Applications in Foods, Lawless, H.T., Klein, B.P. (eds.), 353-393, New York: Marcel Dekker.

Peay, E.R., 1988. Multidimensional rotation and scaling of configurations to optimal agreement. Psychometrika, 53, 2, 199-208.

Piggot, J.R., 1985. Statistical Procedures in Food Research. Elsevier Science Publishers.

Schlich, P., 1989. A SAS/IML Program for generalised Procruses analysis, In: 'Seugi '89', Proceedings of the SAS European Users Group International Conference, Cologne, May 9-12, SAS Institute GmbH, 529-537.

Schönemann, P.H. (1966) A generalised solution of the orthogonal Procrustes problem. Psychometrika, 31, 1, 1-10.

Schönemann, P.H., Carroll, R.M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, 35, 245-255.

Sibson, R. (1978) Studies in the robustness of multidimensional scaling: Procrustes statistics. J. Roy. Statist. Soc. B. 40, 234-238.

Stone H., Sidel, J.L., 1985. Sensory Evaluation Practices. Academic Press, Orlando.

ten Berge, J.M.F., 1977. Orthogonal Procrustes rotation for two or more matrices, Psychometrika, 42, 2, 267-276.

ten Berge, J.M.F., Kiers, H.A.L., Commandeur, J.J.F., 1993. Orthogonal Procrustes Rotation for matrices with missing values. Bitish J. of mathematical and statistical Psychology, 46, 119-134.

ten Berge, J.M.F., Knol, D., 1984. Orthogonal rotations to maximal agreement for two or more matrices of different column orders. Psychometrika, 49, 1, 49-55.

Wakeling, I.N., Raats, M.M., MacFie, H.J.H., 1992. A Comparison of Consensus tests for generalised Procrustes analysis. Journal of Sensory Studies, 7, 91-96.

van Buuren, S., Dijksterhuis, G.B., 1988. Procrustes analysis of discrete data, In: Jansen, M.G.H., Schuur, W.H. van (Eds.) The many faces of multivariate analysis. Proceedings of the SMABS-88 conference, held in Groningen, December 18-21, Volume I, RION, Institute for educational research, FPPSW, University of Groningen.

Williams, A.A., Arnold, M.G., 1985. A Comparison of the Aromas of Six Coffees Characterised by conventional profiling, Free-choice Profiling and Similarity Scaling Methods. J. Sci. Food Agric., 36, 204-214.

Williams, A.A., Langron, S.P., 1984. The Use of Free-choice Profiling for the Evaluation of Commercial Ports. J. Sci. Food Agric, 35, 558-568.

This Page Intentionally Left Blank

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

GENERALISED CANONICAL ANALYSIS OF INDIVIDUAL SENSORY PROFILES AND INSTRUMENTAL DATA

Eeke van der Burg[a] and Garmt Dijksterhuis[b]

[a]Department of Psychometrics and Research Methodology, Faculty of Social and Behavioural Sciences, Leiden University, P.O. Box 9555, NL-2300 RB Leiden, The Netherlands

[b]ID-DLO, Institute for Animal Science and Health, Sensory Laboratory, PO Box 65, NL-8200 AB Lelystad, the Netherlands, e-mail: g.b.dijksterhuis@pobox.ruu.nl

1. INTRODUCTION

Generalised Canonical Analysis or GCA is a multivariate data analysis technique that studies the relationship between sets of variables. In sensory research, data often consist of sets of variables, consequently it is worthwhile taking a closer look at GCA. In the GCA model the sets may contain the same variables but also different variables. In the case of a 3-way table, e.g. products × attributes × assessors, the sets contain the same variables (here attributes). Different variables in each set are obtained in free choice profiling, where every assessor chooses individual attributes. Different variables also occur if various sources of variation are studied. For instance external aspects (e.g. package, image and availability) of one type of food products, price and sales figures of the same products, and in addition, taste aspects and quality judgements.

GCA is a technique that gives an answer to the question: 'What is common between the sets'. Put in another way, GCA is a technique that searches for common underlying dimensions in the sets. In Van der Burg and Dijksterhuis (1989) an application of GCA to a three-way data-table is presented. Several brands of smoked sausages were judged by ten assessors on five aspects (e.g. appearance, taste, odour). In that application the data from one judge form a set, and the next question is addressed: 'On which aspects of smoked sausages do the assessors agree?'

The computer program that performs GCA is called OVERALS (Van der Burg, De Leeuw and Verdegaal, 1988; Gifi, 1990, chap. 5; SPSS, 1990, chap. 9). A less technical overview of OVERALS is given by Van der Burg, De Leeuw and Dijksterhuis, 1994. The program OVERALS can handle data measured on different measurement levels (numerical, ordinal and nominal, and a mixture of these levels). Especially ordinal data may occur in sensory research as assessments measured on a category or line scale are used frequently. Nominal data also occur in sensory research, for instance characteristics like the packaging of products (e.g. milk

in glass bottles, plastic bottles or cartons). In general, data are often interpreted as measured on interval level (numerical) although they may be rather measured on an ordinal scale.

The data of the applications in this chapter stem from the assessments of products by different judges, thus individual sensory data. In addition to this type of data, in one case instrumental data are provided. In the following section the different data types and scales occurring in sensory research are discussed. Subsequent sections discuss the data types and data scales and introduce the GCA technique and the corresponding computer program. The relation between GCA and canonical correlation analysis is discussed (section 3.3) and also an overview is given of the relations between the OVERALS program and other multivariate techniques (section 4.4). Following sections describe examples of applying Generalised Canonical Analyses to sensory data.

2. DATA TYPES AND DATA SCALES IN SENSORY RESEARCH

Two types of data occurring in sensory research are sensory profiling data and instrumental data. Both individual sensory profiles and combined sensory-instrumental data can be analysed by means of GCA. In the following paragraphs these two types of data are discussed.

Apart from different types of data, also different scales of data exist. This means that data can be measured on different levels. Usually three measurement levels are distinguished: nominal (e.g. package type), ordinal (e.g. ranking) and numerical or interval (e.g. temperature in degrees). In the literature also the ratio (e.g. weights) and the absolute measurement level (e.g. percentages) are known. However, data measured on the latter two levels are usually treated in a numerical way. Therefore only the first three measurement levels are discussed here.

GCA, as realised in the OVERALS program, is a technique that can handle data measured on nominal, ordinal and numerical measurement levels, both mixed as well as not mixed. Most data analysis techniques presume one type of scale, either nominal (e.g. correspondence analysis) or numerical (e.g. most classic multivariate analysis techniques). Some techniques can handle mixed measurement levels, for instance (M)ANOVA where the dependent set is treated in a numerical way and the independent set in a nominal way.

2.1 *Sensory profiling*

There are two kinds of profiling data, that can both be analysed by means of GCA: conventional profiling data and free choice profiling data. Conventional profiling data are sometimes analysed by averaging and applying for instance Factor Analysis or Principal Component Analysis to the averaged data. Free choice profiling (Williams and Langron, 1984; Arnold and Williams, 1986) is a kind of profiling resulting in data that can not be averaged over assessors (see also chapter 7.2). GCA or other '*k*-sets' methods are suited for the analysis of free choice profiling data. The scores from either profiling technique are derived from the position of marks along a line scale. The marks correspond to the assessor's perceived intensity of some attribute.

2.1.1 Conventional profiling

In conventional profiling a fixed vocabulary of descriptive terms is used by the sensory panel to judge the products. A sensory panel is often trained in the use of these terms. Because of this training it is assumed that all assessors are able to use the attributes in the same way, so that individual differences in use of the attributes are minimised. The individual judgements are sometimes averaged and Factor Analysis or PCA is applied to the average scores. However, individual difference models as GCA can also be applied to conventional profiling data. Analysis results from individual difference techniques show that the assumption of all assessors using the attributes in the same way is not always justified (see e.g. Dijksterhuis and Punter 1990, Van der Burg and Dijksterhuis, 1989, 1993b; Dijksterhuis 1995b, part 1). In the case of untrained assessors, averaging over judges is hardly ever justified and we have to use a technique that can handle individual differences.

The data from conventional profiling experiments can be seen as a 3-mode data structure built from K assessors, N products and M attributes. This $K \times N \times M$ data block consists of K layers, each with the $N \times M$ data matrix of one assessor. Other slices of this block may be analysed, but in sensory research the focus is mostly on the agreement between the matrices of the individual assessors (see Dijksterhuis 1995b, chap. 1).

The 3-mode data matrix can be analysed in its entirety too, e.g. using 3-mode Principal Component Analysis (Kroonenberg, 1983). See also Chapter 10 for this method.

2.1.2. Free choice profiling

In free choice profiling (Arnold and Williams, 1986) the assessors are free to come up with their own attributes, which they use for judging the products. So there is no *a priori* agreement on attributes between the assessors. As a result it is impossible to average the individual data, because it makes no sense to combine different attributes. The data from FCP experiments must be analysed by individual difference methods, of which GCA is one. Unlike Conventional Profiling data, FCP data cannot be rearranged in some kind of 3-mode data structure, because each assessor may use different attributes as well as in number as in meaning.

2.2 Sensory-instrumental relations

One of the fields in which GCA can be applied is the study of Sensory-Instrumental (S-I) relations. Though GCA appears not to be often applied in this field it can be a useful method to analyse sensory-instrumental relations (e.g. Van der Burg and Dijksterhuis, 1993a). The idea behind the study of S-I relations is that sensory perceptions have chemical/physical counterparts in the substance under investigation. A simple example is the amount of caffeine in a certain drink, which of course determines the bitterness perceived by someone drinking it. In real life S-I research is much more complicated, and involves multivariate data, and consequently needs multivariate data analysis.

Usually two data sets are involved in studying sensory instrumental-relations One data set contains the sensory judgements on the products. The second data set contains a number of instrumental measures on the same products. These can be results of chemical analyses, physical properties or results of other measurements. In case of two sets of variables, two-sets-canonical correlation analysis can be applied to study what is common between the sensory assessments and the instrumental measurements.

When the instrumental measurements are divided into several sets, for instance chemical and physical measurements, a three-sets-GCA can analyse the sensory-instrumental relations. It is also possible that even more sets are involved. For instance, suppose that also sales figures and prices of the products are known. Then a four-set-problem has to be analysed.

2.3 Scale types

The most common scale types in research are the nominal, the ordinal and the numerical scale. The nominal scale reflects a classification of the objects. If the objects are food products, we may think for instance of package type (bottle, carton, tin), colour (red, brown, green) or product type (frozen, fluid, dried).

The ordinal scale reflects an ordered classification or a ranking of the objects (products). In our opinion sensory profiling data is probably best considered as rankings. We discuss this later in this section. Other examples of ordinal scales are measurements as size-class (small, medium, large) or storage temperature (very low, low, normal). In the apple study (see section 5) we find these variables.

The numerical measurement level assumes a ranking of the objects and a constant ratio of difference scores. Numerical measurement levels may occur rather seldom in sensory data, although many statistical techniques assume the numerical measurement level. In general only physical or chemical measures are supposed to be on a numerical measurement level. In addition, frequencies and percentages (which are data measured on an absolute level) are usually treated as numerical measurements, for instance the percentage of rotten apples in a sample.

We mentioned already that we will consider sensory profiles as rankings. Although it is common practise to treat sensory profiling data as measured on a numerical level, we do not always support this habit. The task of measuring products on a line scale cannot be done in an exact way by assessors. For instance, if judges assess the sweetness of cups of tea, we cannot always expect the judges to score the exact amount of sweetness. We can expect the judges to say which cup of tea is more sweet than other cups, or to rank the cups of tea according to sweetness. Although the judges will try to guess the amount of sweetness by scoring a very sweet cup very high and a medium sweet cup in the middle, we do not expect that the assessor means a cup of tea is two times sweeter than another cup of tea, if he or she gives a score two times higher. For this reason it may be more appropriate to treat sensory profile data in an ordinal way (i.e. as rankings) than in a numerical way.

Another common practise in sensory research is using averaged data for sensory profiles. Averaging supposes that the various judges use the line scale in a similar way. In addition, averaging supposes the numerical measurement level. If we treat similar attributes for each judge as separate variables, we get rid of the idea that all assessors score in a similar way. If we assume an ordinal measurement level for each line scale, we get rid of the numerical assumptions too. In case of trained assessors, it may be that the judges use the line scales in a similar way, although we will most often not know this for certain. In case of untrained assessors, however, there is no reason to expect the judges to use the line scales similarly. Therefore a separate treatment of the line scales, for each attribute and judge is recommended. In that case the results can show whether judges differ or not.

We can also check if the assumption of a numerical measurement level was correct. For instance, we can compare analysis results obtained under ordinal constraints with results

obtained under numerical constraints. Sometimes this does not give different results, so that we know that the numerical measurement level was not too restrictive. If the solutions differ, it means that relations between the nonlinearly transformed attributes play a role. This pertains to GCA as well as to other multivariate techniques. Van der Burg and Dijksterhuis (1993b) study vegetable soups assessed by 19 judges. They find that the solutions under ordinal and (approximate) numerical assumptions are very similar, showing that a numerical treatment does not restrict the data too much. At the other hand these authors also find that some judges differ from the other judges in using the attributes. Consequently, averaging these data over judges would be a bad idea.

3. THEORY AND BACKGROUND OF GENERALISED CANONICAL ANALYSIS

The original form of generalised canonical analysis is a technique that studies what is common between sets of (numerical) variables. The technique must provide answers on questions like: 'Can we predict the quality of a product from instrumental measures of the same product?' or 'Do several assessors agree in their judgements of a product and on what attributes do they agree?' When considering to use GCA, several (two or more) sets of variables must be involved in the research question.

Sets of variables can be related in many ways. In GCA they are related in a rather straightforward way, namely as weighted sums of attributes per set. Assuming that a weighted sum represents a set, the weights can be made such that the sets (in fact the weighted sums per set) are as similar as possible to each other. Suppose the sets of variables are denoted by matrices $\mathbf{Y}_k$ (of order $N \times M_k$) and the weights by vectors $\mathbf{a}_k$ (of order M_k) with M_k the number of variables in set k (k=1,...,K) and N the number of products. Each set is represented by the weighted sums $\mathbf{Y}_k\mathbf{a}_k$. The aim of GCA is to find the weights $\mathbf{a}_k$ such that we get

$$\mathbf{Y}_k\mathbf{a}_k \text{ as similar as possible to each other for each } k=1,...,K. \qquad (1)$$

By using a weighted sum we let one attribute be more important than other attributes. The magnitude of the weight reflects the importance of the attributes, but this is not straightforward, as we will explain later in this section.

Another way to make K weighted sums as similar as possible to each other is to make them as similar as possible to an unknown vector (see Carroll, 1968 or Van der Burg and Dijksterhuis, 1993b). Let us denote this unknown vector as $\mathbf{x}$ (of order N). The elements of vector $\mathbf{x}$ are scores for each object (product), and we refer to $\mathbf{x}$ as *object scores*. Let us also suppose that the object scores are standardised (mean zero and variance one). In GCA the object scores and weights are such that we get

$$\mathbf{x} \text{ and } \mathbf{Y}_k\mathbf{a}_k \text{ as similar as possible for each } k=1,...,K, \qquad (2)$$

with object scores $\mathbf{x}$ and variables, columns of $\mathbf{Y}_k$, standardised. Note that expression (1) deals with $K(K\text{-}1)/2$ similarities, whereas there are only K similarities in expression (2). Until now we did not specify what 'similar' means, therefore expression (2) can be interpreted in many ways.

For instance, we can take for similarity the squared correlations which leads to Carroll's (1968) formulation of GCA:

$$\sum_{k=1}^{K} \{ \text{correlations } [\mathbf{x}, \mathbf{Y}_k\mathbf{a}_k] \}^2 \text{ maximal.} \tag{3}$$

Another way of defining similarities is by minimising the sum of squared differences or the *loss* between object scores and weighted variables. This gives a formulation of GCA which is equivalent with expression (3) (see for a proof Van der Burg and Dijksterhuis, 1993b). Denoting the sum of squares of a matrix by SSQ (i.e. SSQ(**Z**)=trace(**Z'Z**) if matrix **Z** is in deviation from its column means), we get

$$\text{loss} = \frac{1}{KN} \sum_{k=1}^{K} \text{SSQ } [\mathbf{x} - \mathbf{Y}_k\mathbf{a}_k] \text{ minimal} \tag{4}$$

over $\mathbf{x}$ and $\mathbf{a}_k$, with $\mathbf{x}$ and (the columns of) $\mathbf{Y}_k$ (k=1,...,K) standardised. The GCA formulation in terms of loss is used by Van der Burg, De Leeuw and Verdegaal (1988) and by Gifi (1990, chap. 5).

Of course one might wonder if a one-dimensional solution of object scores suffices to represent what is common between the sets. If one decides that a one-dimensional solution is not enough, a two-dimensional solution can be taken. As it is not interesting to have correlated dimensions, a second dimension of object scores is constructed such that it is uncorrelated to the first one. This means that the axes representing the object scores are perpendicular, and we have a rectangular coordinate system in which we can plot a configuration of products. This configuration is standardised, that is, it has zero mean and unit variance in the directions of the axes.

If one is interested in a solution of more than two dimensions, the same procedure can be followed. The third dimension of the object scores is taken uncorrelated to the first and the second one. The fourth one is taken uncorrelated to all the preceding dimensions, and so forth. If we deal with a P-dimensional solution we find a standardised configuration of products on P perpendicular axes, which is represented by a matrix $\mathbf{X}$ of order (N×P). This matrix consists of uncorrelated and standardised columns that is, $\mathbf{X'X}/N=\mathbf{I}$, with $\mathbf{I}$ the identity matrix. The weights are represented by matrices $\mathbf{A}_k$ (of order M_k×P). The weighted sums for each set k are denoted by $\mathbf{Y}_k\mathbf{A}_k$. Then the GCA problem for P dimensions is written as

$$\text{loss} = \frac{1}{KN} \sum_{k=1}^{K} SSQ\ [\mathbf{X} - \mathbf{Y}_k\mathbf{A}_k] \text{ minimal} \tag{5}$$

over $\mathbf{X}$ and $\mathbf{A}_k$, with the columns of $\mathbf{X}$ uncorrelated and standardised and the columns of $\mathbf{Y}_k$ (k=1,...,K) standardised. The object scores $\mathbf{X}$ can be plotted, just like in PCA. An interpretation of this plot corresponds to an interpretation of the solution. Interpreting the solution can be done via the variables within the analysis, but also by using external variables. To interpret the configuration we have to find out what the various directions in the plot of object scores represent. We do this by checking the *component loadings*, which correspond to the correlations between the object scores (for each dimension) and the variables. The term 'component loadings' is used in analogy with the component loadings from PCA. We can

make a plot of the variables in the object scores space with coordinates equal to the component loadings, which are available for the variables from each set. Usually, in this plot, the variables are represented by vectors from the origin. Interpreting this plot is done in the same way as interpreting the component loadings in PCA. This means that variables far from the origin are more important than those close to the origin. In addition, two variables that are very similar *and* important will have vectors close to each other with comparable lengths. Variables with short vectors are badly represented in the solution. However, this does not mean that these variables have a low explained variance in a PCA sense. It means that such variables represent variance in the data that cannot be found in the other sets.

If a correlation between one attribute and, say, the first (dimension of a) solution of object scores is high, then this attribute has a high contribution to the first dimension. Sometimes it is interesting to reify the object score axes from the contributions of the variables (compare interpreting the principal axes in Factor Analysis). In that case it may be interesting to use a rotation of the solution to facilitate the interpretation of the axes (see Kiers and Van der Burg, 1994). Often, however, only the configuration of products and of variables is interesting in itself, so that axes need not be labelled.

Another way to interpret the configuration of object scores is with the help of an external variable. In some cases there is information provided about the data which is not used in the analysis. For instance, in Van der Burg and Dijksterhuis (1989) sausages are studied. There the GCA solution of the profiles can be (among other things) interpreted by means of the variable 'make' (factory-made versus butcher-made). Another example is found in Van der Burg and Dijksterhuis (1993b), where vegetable soups are analysed. The researchers use the external variables 'package' and 'type' to help interpreting the GCA solution.

Earlier in this section it was mentioned that the weights for the linear combinations indicate the importance of each variable. However, if an assessor behaves similarly on two attributes, the weight is not a good measure for comparing the importance of the various attributes, as the influence of one attribute can be expressed via the other attribute. For instance, it may happen that, although two attributes belonging to the same set measure nearly the same thing (are scored the same), one weight is high and the other one is small. In that case the weight of the first attribute contains the effect of the second attribute. The reason is that the two attributes explain the same variation and that this variation can be explained only once. When there is a lot of multicollinearity between attributes within a set, an attribute can be dropped from the set even without changing the quality of the relation with the other set(s). That is, the weight of this attribute can be zero. As the weights do not always give a good insight in the structure of the sets (unless we keep the correlation matrix in mind), it is much easier to interpret a solution via the correlations between the variables and the object scores: the component loadings. These correlations indicate the importance of every variable to the solution, independent from the contribution of the other variables in the set.

3.1 Optimal scaling

In the preceding discussion on GCA, nothing is said about the measurement level and the corresponding transformation of the variables. If variables are considered to be numerical, this implies that the original scores can be transformed in a linear way without destroying the information in the original data. If a variable is seen as measured on an ordinal level, the scores of this variable can be rescaled in an ordinal way without loss of information. This corresponds

to a rescaling without changing the original order, keeping similar scores equal. There are many possibilities for ordinal transformations, in fact all monotone ascending transformations will do.

Also for nominal variables a transformation is possible without losing information. Such a transformation should keep similar scores equal. A nominal transformation preserves the classification of the products that is induced by a variable. Nominal transformations are less restricted than ordinal transformations. Therefore, there are more possibilities for nominal transformations than for ordinal transformations.

Variables measured on a nominal (or ordinal) level can be treated in two different ways. We can use one transformation for one P-dimensional solution or we can use a different transformation for each dimension. The former is called *single* quantification and the latter *multiple* quantification. The multiple nominal transformation is similar to the scaling in (multiple) correspondence analysis (Nishisato, 1980, chap. 2; Greenacre, 1984, chap. 5). In the applications of GCA shown in this chapter, we only use single nominal transformations. Therefore, we do not discuss multiple nominal transformations here any further. Multiple ordinal transformations are theoretically possible but they are not implemented in any computer program, so they will not be considered either.

We refer to single quantifications (nominal, ordinal or numerical) as *optimal scaling*. In fact optimal scaling implies that the transformations are obtained in a special way, namely in combination with the optimising criterion (Young, 1981).

We need some more notation to include the optimal scaling in our formulation of GCA. Let us denote the sets of transformed (i.e. quantified) variables by $\mathbf{Q}_k$ (of order $N \times M_k$). For these matrices constraints are valid per column (i.e. per variable). If, for instance, the first variable of set k is considered to be measured on an ordinal measurement level, the first column of $\mathbf{Q}_k$ must be a monotone transformation of the first column of $\mathbf{Y}_k$. We call the restrictions to be imposed on the transformations of the variables, including standardisation, *measurement restrictions*. Using this notation, GCA with optimal scaling can be written as

$$\text{loss} = \frac{1}{KN} \sum_{k=1}^{K} \text{SSQ}\,[\mathbf{X} - \mathbf{Q}_k \mathbf{A}_k]\ \text{minimal}, \tag{6}$$

over $\mathbf{X}$, $\mathbf{Q}_k$ and $\mathbf{A}_k$, with the columns of $\mathbf{X}$ uncorrelated and standardised and the columns of $\mathbf{Q}_k$ satisfying measurement restrictions. This formulation of GCA is introduced by Van der Burg, De Leeuw and Verdegaal (1988) and is also used by Gifi (1990, chapter 5). The essence of expression (6) is that optimal scores will be assigned such that the GCA-criterion is maximised and that in addition the measurement restrictions are satisfied. Both Van der Burg, De Leeuw and Verdegaal (1988) and Gifi (1990, chapter 5) describe how to obtain the solutions. Without giving details here, we can say that all the parameters are solved for in an alternating least squares (ALS) manner.

We distinguish various effects in sensory profiling data originating from different assessors. We mention the *level effect*, the *scale effect* and the *interpretation effect* (see also chapter 7.2). Let us consider these effects in GCA as formulated in expression (6). Because, for all types of scaling the measurement restrictions imply standardisation, the level effect is excluded from the assessors' scores by subtracting means. The standardisation also removes the individual scale effect per variable through division by the standard deviation. The interpretation effect is modelled in GCA by using weighted sums of variables or *linear*

combinations. Note that in GPA the interpretation effect was modelled by using rotations (see chapter 7.2).

We can study the interpretation effect by treating the same attributes from different assessors, e.g. all 'sweet' attributes, as different variables. It does not matter which measurement level is used. However, if various assessors have a different definition of one attribute in mind, the impact of ordinal or nominal transformations may be bigger than the impact of numerical -linear- transformations. In case of different definitions of one attribute, the question arises what we are comparing. Of course, we cannot really answer this question. But we can find out that judge A and judge B have different interpretations of an attribute.

Finding an interpretation effect means that averaging scores over judges is not permitted (see also section 2.3). At the other hand, if attributes, assessed by various judges, are interpreted similarly, a justification for averaging over assessors is provided.

Using numerical measurement levels is a very common method in linear multivariate analysis. Note that the term 'linear' can refer either to linear *combinations* of (standardised) variables or to the linear *transformations* of the variables. Here both meanings apply.

Using ordinal and nominal measurement levels (in combination with numerical measurement levels) is rather new. For PCA a nonlinear version for mixed measurement levels exists, both in the form of a model as in the form of a computer program (see the references in section 4.4). Here 'nonlinear' refers to the transformations. Van der Burg and De Leeuw (1983) describe a nonlinear version of two-sets canonical correlation analysis, which was implemented in the CANALS program. A computer program for k-sets was only made available recently although several GCA models were described in the literature many years ago (Horst, 1961; Carroll, 1968; Kettenring, 1971; Van de Geer, 1984). Van der Burg, De Leeuw and Verdegaal (1988) introduced a model for GCA which was implemented in the computer program OVERALS and which, in addition, was provided with possibilities for nonlinear transformations for ordinal and nominal variables, and linear transformations for numerical variables. The OVERALS program is available in SPSS Categories (SPSS, 1990, chap. 9).

If data are measured on an ordinal or nominal scale we prefer to treat the data correspondingly. However, often it is interesting to compare a linear analysis (only numerical measurement levels) with a nonlinear analysis and to see the similarity or the difference. If the resemblance is very high, we know that the linear analysis makes sense. If there is a difference, we have to accept that different judges define the same line scale in a different way. We may, in addition, interpret the optimal quantifications to find out what causes the differences.

3.2 Loss and fit measures

The results of a nonlinear GCA analysis can be evaluated by the loss and fit measures. The loss shows the lack of fit of a solution. In case of a P-dimensional solution, the minimum loss is 0 and the maximum P (see Van der Burg, De Leeuw and Verdegaal, 1988, p. 184). The loss can be divided over dimensions, and one minus the loss per dimension corresponds to the eigenvalue (maximally one and minimally zero). The eigenvalue corresponds to a goodness-of-fit measure and the sum of eigenvalues is called the total fit. The loss or *total loss* is equal to P minus the total fit. In formula we write

$$\text{(total) loss} = \sum_{p=1}^{P} \text{loss}(p) = \frac{1}{KN} \sum_{p=1}^{P} \sum_{k=1}^{K} \text{SSQ}\,(\mathbf{x}_p - \mathbf{Q}_k \mathbf{a}_{kp}) \qquad (7)$$

with $\mathbf{x}_p$ and $\mathbf{a}_{kp}$ the p-th column of $\mathbf{X}$ or $\mathbf{A}_k$. Note that the loss in (7) is equal to the loss in (6). For the fit we get

$$\text{total fit} = \sum_{p=1}^{P} \text{eigenvalue}(p) = \sum_{p=1}^{P} \{ 1 - \text{loss}(p) \} = P - \text{loss}. \tag{8}$$

The eigenvalue represents the mean variance of the sets (i.e. the weighted sums of variables), accounted for by the object scores. For more details on the properties of these eigenvalues we refer to Van der Burg, De Leeuw and Verdegaal (1988) or Van der Burg, De Leeuw and Dijksterhuis (1994). The eigenvalues do not necessarily correspond to the average of explained variances of the optimally scaled variables per set. If two optimally scaled variables have a perfect correlation, but are located in different sets, these two sets can be predicted perfectly from each other, irrespective of how much variance of the other variables in the corresponding sets is explained. It means that, in GCA, we always have to check the meaning of a high fit. If a high fit corresponds to only a little explained variance of the optimally scaled variables, we may decide to investigate the higher dimensions of the solution or to drop one of the variables that causes the high fit. It is a result of the fact that the GCA method focuses on correlations between variables in different sets, irrespective of associated variance per set.

Especially because of the optimal scaling we have to beware of unique patterns. Unique patterns are correspondences between sets shared by very few objects. For instance, if there is only one product packed in glass and this product is the only one that is judged as breakable, we have a unique pattern. The OVERALS program may fit this pattern by scaling all categories of 'package' into zero, except the glass score, which may get a high quantification. If the program does this for the scores on 'fragility' too, the two optimally scaled versions of 'package' and 'fragility' (in different sets), are highly predictable from each other, resulting in a high fit. Thus in case of a high fit, the optimal scores have to be checked for such degeneracies.

As was described in section 2.3 on 'scale types', the transformations of the variables, or the 'scaling of the categories', always satisfy the measurement restrictions. For ordinal variables it means that the order of the original scores is maintained, for nominal variables it means that similar original scores get the same transformed scores. So different *original* scores may get similar transformed values. In the above example with the two nominal variables 'package' and 'fragility', the nominal restrictions are satisfied, although it provides a unique pattern.

3.3 Canonical Correlation Analysis

The term Canonical Correlation Analysis (CCA) usually refers to a two-sets multivariate technique that maximises the correlations between linear combinations of two sets of variables (Hotelling, 1936). CCA and two-sets GCA are similar although CCA exists much longer. In fact, CCA is a special case of GCA, namely the case of K=2 and only numerical measurement levels. The usual representation of CCA (e.g. Tatsuoka, 1988, chap. 7; Gittins 1985, chap. 2) is formulated in terms of the correlations between the linear combinations per set (canonical axes or canonical variates) and the variables. For each set, the scores on the canonical axes are uncorrelated, just like the object scores. The projections of the standardised variables onto the canonical axes are equal to the correlations between variables and canonical axes or 'structure correlations' (c.f. Ter Braak, 1990), or 'intraset' and 'interset' correlations (Gittins, 1985, p. 38), and a plot of correlations in CCA is comparable to a plot of component loadings in GCA.

However, in CCA there are two sets of correlations, of all variables with the canonical axes of each set, and in GCA there is only one set of component loadings, of all variables with the object scores. This is because the object scores in 2-set GCA correspond to the average scores on the canonical axes in CCA (Van der Burg and De Leeuw, 1983). In K-set GCA, the object scores correspond to the mean over K canonical axes.

3.4 Representing the original variables

The variables are represented by the weights and the component loadings. As was mentioned already, the weights can, in case of multicollinearity within a set, include the effect of the other variables of the set. Therefore, the component loadings give a better view on the solution. They provide a measure for the relation between a transformed variable and the object scores for each dimension. The squared loadings represent the explained variance of the variables by the object scores. If we denote a column of $\mathbf{Q}_k$ by $\mathbf{q}_{l_k}$ (with l_k=1,..,M_k), the component loadings, collected in the matrices $\mathbf{C}_k$ (of order $M_k \times P$) , are defined by

$$c(l_k,p) = \text{component loading } (l_k,p) = \text{correlation } [\mathbf{x}_p , \mathbf{q}_{l_k}]. \tag{9}$$

In the output of the OVERALS computer program, variables are individually represented by the *centroids*, which are the mean object scores averaged over the products in the same category of a variable. These centroids correspond to the so-called multiple nominal transformations. As we do not use these type of transformations in the applications, we will not expand on them further. In addition, for each variable there are the so-called *projected centroids*, which are optimally scaled scores $\mathbf{q}_{l_k}$ (called *category quantifications*, with as many different values as there are different categories of variable l_k) multiplied by the corresponding component loadings ($\mathbf{c}'_{l_k}$= rowl_k of matrix $\mathbf{C}_k$ with l_k=1,..,M_k):

$$\text{projected centroids } (l_k) = \mathbf{q}_{l_k} \mathbf{c}'_{l_k} \tag{10}$$

These are found on a line in the object scores space. These scores are called projected centroids because they can be seen as centroids projected on a line (with measurement restrictions). The direction cosines of this line are the component loadings. Thus centroids and projected centroids refer to the space of object scores, so that, for interpretation of quantifications one needs these scores. (see e.g. the plot of projected centroids in Figure 4). However, often an interpretation of quantifications is not necessary. In that case we use the category quantifications only to check for degeneracies. Figure 3 and Figure 8 show a plot of category quantifications, furthermore Van der Burg, De Leeuw and Verdegaal (1988) discuss an example with interpretation of the centroids.

3.5 Statistical matters

The GCA technique as implemented in the OVERALS program, is not equipped with statistical tests. As there are no assumptions about a distribution, we have to use permutation or randomisation methods to test the stability of a solution. For such tests we only have to assume a multinomial distribution of the profiles (possible score patterns for an object/product), which is true in case of random sampling. One can use the Bootstrap or the

Jackknife (Efron, 1982; Miller, 1974) to study the stability of a solution by computing confidence intervals. In addition, permutation tests can be used (Edgington, 1987; Good, 1994; Van der Burg and De Leeuw, 1988) to find the significance of results.

The bootstrap is a method that resamples from the data (a sample from a larger population) such that inferences can be made on the population. The sampling takes place with replacement, keeping the number of observations equal to the number in the original sample. Thus, the randomly produced sample (called the bootstrap sample) may contain the same product several times and other products not. Then the GCA technique is applied to the bootstrap sample. The procedure of taking a bootstrap sample and performing a GCA is repeated many times. Then, for each statistic under study (for instance the fit), we have as many instances of this statistic as there are bootstrap samples. From these values the variance of the statistic can be estimated. In addition, an estimation of the population mean can be made, so that a confidence interval can be computed (Van der Burg and De Leeuw, 1988).

The Jackknife is similar to the Bootstrap except that the Jackknife samples are drawn in a different way. With the Jackknife the new samples are the same as the old ones save one product or save s products (s a small number). Every product is dropped once. The Jackknife sample contains (N-1) or (N-s) products. If one product is dropped, there are exactly N Jackknife samples. As with the Bootstrap, the Jackknife sample values of the statistic under study provide an estimation of the variance and the population mean, so that a confidence interval can be computed.

Permutation tests are made by permuting the data, that is, randomly reordering the products separately within each set (see De Leeuw and Van der Burg, 1986). In the case of two sets with one nominal variable per set, if we organise the two variables in a cross table, permuting the data comes down to changing all the cells of the cross table while keeping the marginals at a constant level. Each table represents a permutation sample to which the technique can be applied. Thus every table provides a value of the statistic under study and together they form a distribution from which the original sample is one. Using order statistics then provides a significance level for the statistic. For two nominal variables we have Fisher's exact test. For more complicated cases a permutation test can also be made. However, if the number of variables grows, the number of possible randomisations grows too, so that it becomes rapidly impossible to compute the exact permutation distribution.

4. COMPUTER PROGRAM FOR GCA

In this section characteristics of the computer program OVERALS are introduced. It is also shown how the implemented GCA method relates to other multivariate methods and programs.

4.1 Categorising data

The input for the computer program OVERALS is restricted to data containing a small number of positive integers. This implies that continuous data have to be recoded into data with a relatively small number of categories. The restriction that data for OVERALS have to be discrete, has to do with the way the computer program is made. The computer program works with scores for categories. If every variable has as many categories as there are objects, the program may become rather slow. Therefore, the number of categories per variable, for each

set (assessor) should be reasonably smaller than the number of objects, which, in our experience, is not a severe restriction.

When one analyses the rank-order scores of the objects, there are as many categories as there are objects in the data. An ordinal analysis of these rankings can be a good way to reduce the number of categories.

The assumption that the input of the program contains discrete data, implies that, if the data do not satisfy this assumption, the researcher has to recode the data until the assumption is satisfied. For instance, if line scales scoring from zero to 100 are used, the information has to be compressed into a small number of categories. If we do not want to loose too much information we can take for example 10 to 15 equidistant categories. If we are not too concerned about details, often 3 to 6 categories are sufficient to retain the information that determines the relations between the sets.

Two analyses, one with 10 to 15 categories (and numerical measurement restrictions for all variables) appeared hardly different from the results from the analysis of 3 to 5 categories (and ordinal measurement restrictions for all variables) (see Van der Burg and Dijksterhuis, 1993b).

4.2 Missing data

The OVERALS computer program can handle missing data. The theory of the GCA-model including missing data is described by Van der Burg (1988, p.107). If a product is not scored for one variable, the computer program treats the scores for all the variables in the same set as missing. This means that the product does not contribute to the fit for the set at hand. Many missing scores will make it easier for unique patterns to arise, thus with a lot of missing data the outcome has to be checked for degeneracies. In normal cases missing data do not give rise to problems. As missing scores do not contribute to the fit, there is no optimal scaling either for missing scores. Thus no estimation of the missing score is provided by the OVERALS program.

4.3 Dimensions

The number of dimensions for the OVERALS analysis has to be specified by the researcher. He or she has to decide for him- or herself how many dimensions are needed. In practice this often means that solutions of different dimensionality are computed and that the best one is chosen to report. One of the most important considerations in choosing for P dimensions, is that all P dimensions are interpretable. An argument for taking one or two dimensions is that it is easy to plot. An argument to prefer P dimensions above P+1, is that there is hardly any difference between the P- or the (P+1)-dimensional solution. If P is larger than two it can be hard to interpret the solution. In that case a rotation may help to decide how many dimensions must be taken (Kiers and Van der Burg, 1994).

4.4 Relations of OVERALS to other MVA techniques

In this section we illustrate the relations between GCA as realised in OVERALS (Van der Burg, De Leeuw and Verdegaal, 1988) and other multivariate techniques. Sometimes we use the name of an author to indicate a model, sometimes we use the name of the technique, but in addition we also use the name of a computer program to indicate a model. We hope this will not lead to confusion.

OVERALS is a very general technique that comprises many techniques as a special case. If the measurement level of all variables is numerical and there is only one variable per set, then we are dealing with ordinary PCA. In this case the fit of a solution (the eigenvalue) corresponds to the mean explained variance of the variables because in this case variables and sets are identical. This is the usual definition of the eigenvalue in PCA. If we keep the number of variables per set equal to one, but allow for different measurement levels, we obtain a nonlinear version of PCA. In the literature we encounter this model under the name PRINCALS (Gifi, 1990, chap. 4; SPSS, 1990, chap. 8). Other nonlinear PCA models can be found in PRINCIPALS (Young, Takane and De Leeuw, 1978) and PRINQUAL (Kuhfeld, Sarle and Young, 1985; SAS/STAT, 1990, p. 1265).

If all variables are considered as multiple nominal and we still have one variable per set, we arrive at a technique called dual scaling (Nishisato, 1980, chap. 2) or multiple correspondence analysis (Greenacre, 1984, chap. 5; Gifi, 1990, chap 3). This technique can be viewed as PCA for nominal data. For instance the computer program HOMALS (SPSS, 1990, chap. 7) performs this type of analysis, also the program CORRESP (SAS/STAT, 1990, p.615). In case there are only two multiple nominal variables, we get correspondence analysis, for instance implemented in the ANACOR program (SPSS, 1990, chap. 6).

If we restrict OVERALS to two sets of variables and only single measurement levels we get canonical correlation analysis with optimal scaling. This technique is very similar to CANALS (Van der Burg and De Leeuw, 1983). If, in addition, one of the sets contains only one variable, we get nonlinear multiple regression or MORALS (Young, De Leeuw and Takane, 1976). Two-sets OVERALS with only numerical measurement levels gives ordinary canonical correlation analysis.

If OVERALS is restricted to numerical measurement levels only, we get the technique described by Carroll (1968) (see also section 3).

If we relate OVERALS to three way techniques that generalise PCA we can compare it, for instance, to the class of models discussed by Kroonenberg (1983). In particular the TUCKER2 model (see Chapter 10) is obtained from the OVERALS model by constraining the weights. In addition, the TUCKER2 model restricts the variables in all sets to be similar, as it supposes a three way table, which is not the case in OVERALS. Thus we see that TUCKER2 is a special case of OVERALS.

As discussed already in section 3.1, other techniques have been proposed for K sets analyses. (Kettenring, 1971; Horst, 1961). In particular Van der Geer (1984) compares several techniques.

4.5 Post Hoc Rotations

Rotations are not provided in the OVERALS computer program. Of course, if the user is interested in naming the axes or interpreting a high dimensional solution, a rotation may help. In case of 'single' variables only a simple varimax rotation can be applied to the object scores space. Oblique rotations are not allowed as they will severe the orthogonality constraint on the object scores. In case of mixed multiple and single measurement levels a simple varimax does not satisfy because the 'multiple' variables do not have component loadings. Kiers and Van der Burg (1994) propose a rotation algorithm in which the so-called discrimination measures for 'multiple' variables are used and the component loadings for 'single' variables. They also give an illustration of their algorithm.

5. ANALYSING SENSORY-INSTRUMENTAL RELATIONS USING GCA

The example to illustrate sensory-instrumental relations concerns research on apples[1]. Van der Burg and Dijksterhuis (1993b) use these data for the prediction of assessments from instrumental measures, illustrating nonlinear Redundancy Analysis. We will illustrate nonlinear GCA with the help of the apple data.

Table 1

Variables Measured on Cox Apples

Background variables		*categories*		
Ca	origin	low, high Calcium		
Per	picking-date	early, middle, late		
Si	size	small, large		
Temp	storage-temperature	3, 13, 23 degrees Celsius		
Instrumental variables		*min*	*max*	*# cat*
Pread	penetrometer: red side	2.50	5.80	5
Pgreen	penetrometer: green side	3.50	5.50	5
Pmean	penetrometer: mean	3.70	5.90	5
Moist	expelled moisture	5.51	43.06	6
Drymat	dry matter	12.25	17.08	5
Acid	total titratable acid	3.55	8.07	5
Ithick	Instron: thickness at failure	1.33	3.05	6
Ifrac	Instron: force at failure	26.21	71.67	4
Isurf	Instron: area	11.51	56.64	5
Islope	Instron: slope	7.48	97.95	6
Imod	Instron: modulus	1.27	3.66	6
Catac	Catalase activity	6.90	19.40	6
Sensory variables		*min*	*max*	*# cat*
Mealy1	mealiness judge1	0 (not mealy)	100 (very mealy)	4
Mealy2	mealiness judge2	0 (not mealy)	100 (very mealy)	4
Mealy3	mealiness judge2	0 (not mealy)	100 (very mealy)	4
Firm1	firmness judge 1	0 (firm)	100 (soft)	4
Firm2	firmness judge 1	0 (firm)	100 (soft)	4
Firm3	firmness judge 1	0 (firm)	100 (soft)	4

min= lowest score; *max*=highest score; # *cat*=number of categories after recoding.

5.1 Data on apples

The data consist of measurements on Cox apples (Koppenaal, 1991). The measurements can be divided into three sets of variables: background variables, instrumental measures and sensory variables (Table 1). The first set has been created to acquire a variety of Cox apples. The background variables are: origin, picking date, size and storage temperature. The instrumental variables consist of several types of physical or chemical measures like the amount of expelled moisture, the catalase activity and several Instron measures. Table 1 shows all the instrumental variables. The sensory variables consist of the assessments of three trained judges on the characteristics 'mealiness' and 'firmness'. These two characteristics represent aspects of the quality of the apple. Originally the researchers were mainly interested in a prediction of

[1] The ATO-DLO Institute for Agrotechnology (Wageningen, The Netherlands) is thanked for making the data set on apples available.

'mealiness' and 'firmness' from the instrumental variables (Koppenaal, 1991). They performed two linear multiple regressions on the accumulated scores for mealiness and firmness. The question addressed next is: 'How do the instrumental variables and the sensory variables intermingle with the background variables, taking into account the various measurement levels?' In fact we want an answer to the question: 'Under which condition is an apple judged as 'nice' and which instrumental measures are good in indicating the quality of the apple?' Note that the background variables were manipulated as factors in an experimental design, so that these variables are independent.

For our secondary analysis the sensory and instrumental measures have been recoded to reduce the number of categories. Our experience is that this hardly influences analysis results (see Van der Burg and Dijksterhuis, 1993b). In Table 1 the minimum and maximum original score is given plus the number of categories used for the recoding. The recoding always refers to equidistant divisions of the original scores, except for the lowest and the highest category. The latter was done to avoid unique patterns.

OVERALS was applied to the recoded data. The measurement levels of background variables were considered single nominal and the measurement levels of the instrumental variables and the sensory assessments (single) ordinal.

5.2 Fit and Loss measures

A four dimensional OVERALS solution, with all variables treated as single, gives eigenvalues per dimension of 0.865, 0.771, 0.682 and 0.666 respectively. The first dimension appeared to be completely dominated by storage temperature (TEMP), the second dimension by origin (CA), the third by SIZE and the fourth by picking date (PER). This shows the independence of the background variables. The sensory assessments and various instrumental measures load mainly on the first two dimensions. So apparently TEMP and CA are related to the sensory and instrumental variables.

We repeated the analysis in two dimensions. This gives eigenvalues per dimension of 0.864 and 0.779 respectively (see Table 2).

Table 2

The loss per set, eigenvalues and fit of the apple data for a two-dimensional OVERALS solution

LOSS PER SET	*dimension*		
	1	2	SUM
Background variables	0.183	0.255	0.438
Instrumental variables	0.073	0.093	0.166
Sensory variables	0.150	0.314	0.465
MEAN	0.136	0.221	0.356
FIT			1.644
EIGENVALUE	0.864	0.779	

We also performed the two-dimensional analysis with the background variables considered as multiple nominal. This gives eigenvalues of 0.863 and 0.797 which is hardly better than in

the single nominal case. We decided to report the two-dimensional solution with all variables treated as single. The loss, fit and eigenvalues are given in Table 2. This table shows that all sets have a very low loss in the first dimension which means that the assessments are well predicted by background variables and instrumental measures at the same time. The instrumental variables also have a low loss in the second dimension telling us that they do well in (are much related to) this dimension. We will see which variables are related, and to what extent, in the plot of component loadings.

5.3 Component loadings and object scores

The component loadings are plotted in Figure 1. Such a plot is comparable to the loading plot or correlation plot of PCA, although this plot is obtained in a different way. We see from the plot of component loadings that the first dimension is dominated by TEMP from the set of background variables, by MOIST, CATAC and ACID from the instrumental set and by all the assessments of the sensory set. The second dimension is dominated by CA and ITHICK. The sensory variables hardly play a role in this dimension. Thus storage temperature is dominating the taste, in the sense that a high temperature corresponds to mealy and soft apples, and a low temperature to a good taste, not mealy and firm apples. The vectors point to the direction of high scores, in case of the sensory variables to a bad taste, i.e. mealy and soft (see Table 1). A good taste -firm and not mealy- also corresponds to a lot of expelled moisture, a high catalase activity and a high amount of titratable acidity. Specially MOIST is a very clear indicator for a good taste.

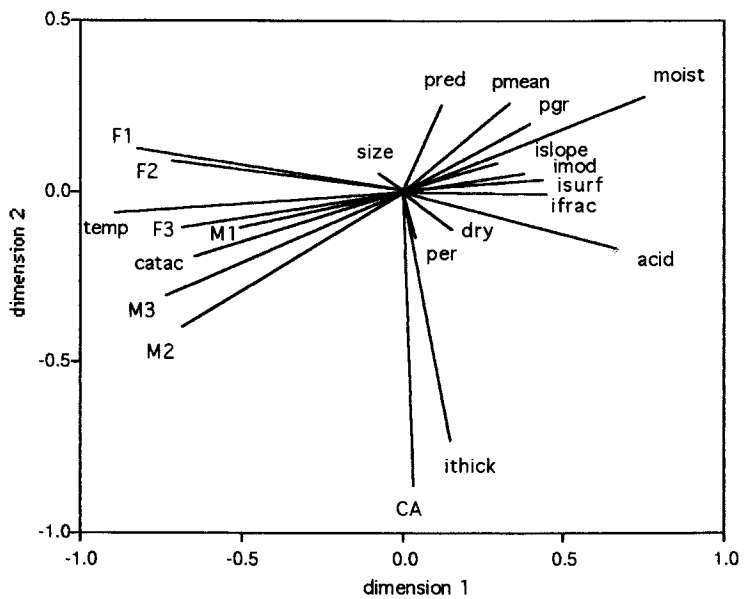


Figure 1. Component loadings of OVERALS applied to the data set on Cox apples.
(F = Firmness, M = Mealiness).

If we take a look at the object scores (Figure 2) labelled by temperature, we find that all apples stored at low temperature (3 °C) are located on the right side. These are the apples of good quality. The apples stored at 13 and at 23 degrees C are found at the left side. These apples are judged to be of lower quality. In general the apples at the left have low moisture, low titratable acidity and low catalase activity.

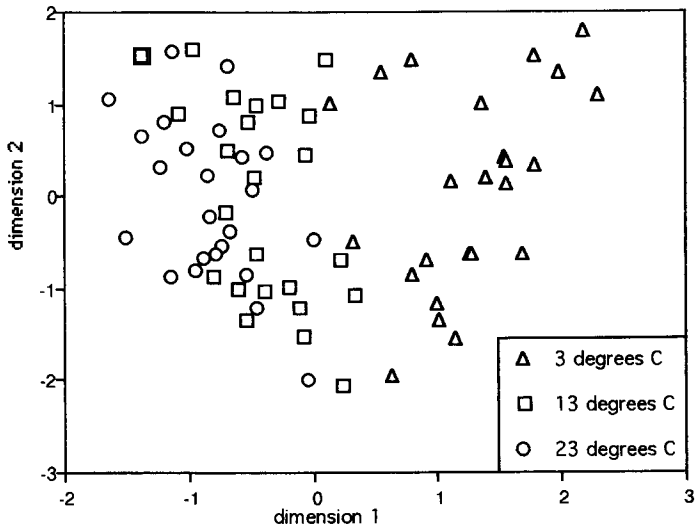


Figure 2. Object scores of OVERALS applied to the data set on Cox apples. The objects, i.e. individual apples, are marked by storage temperature.

The apples found in the lower half of Figure 2 are characterised by a high Ca condition and also by a large thickness at failure (ITHICK). In the higher part we find the apples with a low Ca condition and a low ITHICK. The position in the lower or higher part of Figure 2 is hardly related to quality judgements.

From figures 1 and 2 we see that mainly the first dimension is of interest for the quality of the apples. We also saw this in Table 2. The contribution of the sensory variables to the loss is small for the first dimension and much higher for the second one.

5.4 Variables and categories

Plots of the category quantifications of three important variables (TEMP, Catalase Activity, MOIST) are given in Figure 3. The term 'original' score, refers to the input scores of OVERALS, i.e. the scores after the recoding of the numerical data. For TEMP this means: 1=3 °C, 2=13 °C and 3=23 °C. For CA the recoding implies 1=low calcium and 2=high calcium, and for MOIST, 1=5.51...10.0, 2=10.1...15.0, 3=15.1...20.0, 4=20.1...25.0, 5=25.1...30.0 and 6=30.1...43.06.

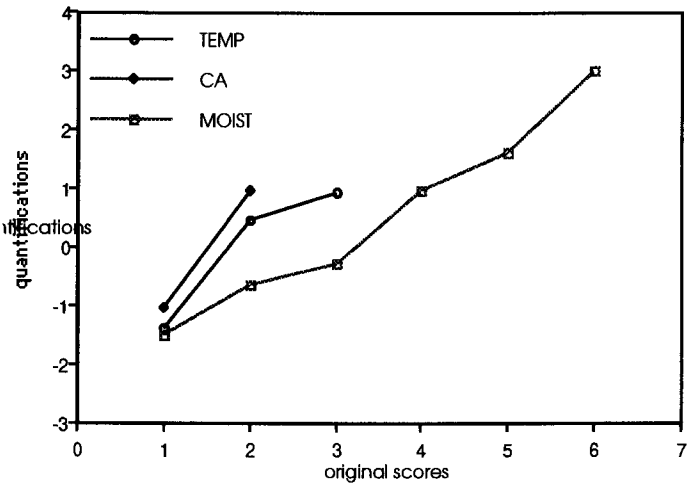


Figure 3. Category quantifications of some important variables in the OVERALS result of the analysis of the Cox data set. The variables are TEMP, CA and MOIST.

The quantifications of temperature are respectively -1.38, 0.47 and 0.94. Thus the main difference is between the low temperature and the other two higher temperatures, meaning that mainly the low temperature of 3 degrees Celsius can be taken responsible for a good taste. This corresponds to what we saw in the plot of object scores (Figure 2). The transformation of CATAC contains three ties, original scores 3 (=10.1...12), 4 (=12.1...14) and 5 (=14.1...16) (not shown in figure), thus no difference is made in the solution between these scores. Originally there also was a -recoded- category 1 for FIRM3. However, as this category was scored with a very low frequency, we recoded it into a 2, to avoid a unique pattern.

It can sometimes be difficult to interpret the transformation plots. It may be easier to imagine what happens if the quantified categories are plotted on the lines through the component loadings. Such a plot is made for the most important variables (Figure 4). We clearly see the spread of the (recoded) categories over the object scores space.

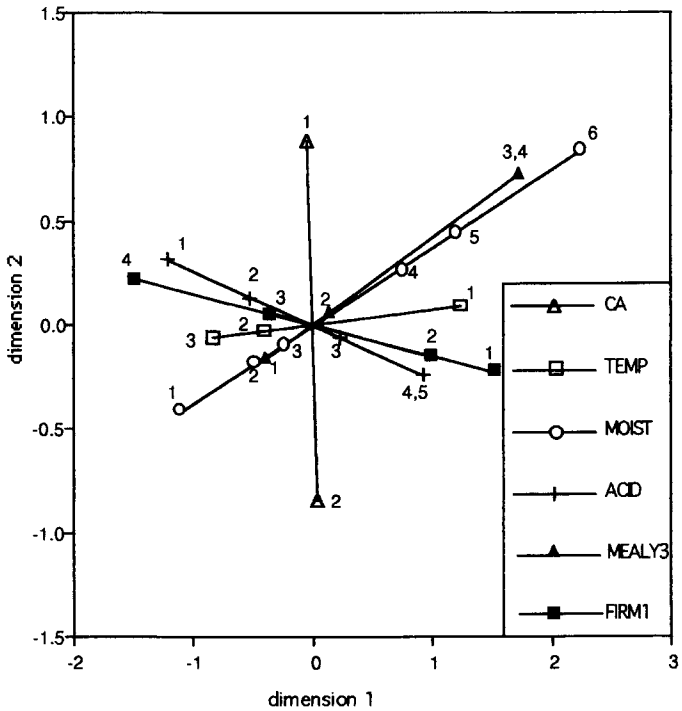


Figure 4. Projected centroids of the most important variables in the analysis of the Cox data set.

5.5 Conclusion

The nonlinear GCA analysis of the Cox apples shows that in relating background variables, sensory assessments and instrumental measures, we obtain a mainly one-dimensional solution as the sensory variables play a role nearly only in the first dimension. In fact Van der Burg and Dijksterhuis (1993a) had a similar conclusion though they applied a different technique and did not use the background variables. The second dimension combines CA with ITHICK, but the sensory variables are not important in the second dimension.

The first dimension differentiates the apples with respect to taste. This goes together mainly with the amount of moisture, but also with the amount of titratable acidity and catalase activity, which makes these variables good indicators for the quality of an apple. To reach a high quality -low mealiness and high firmness-, apples should be stored at a low temperature.

6. PERCEPTIONS OF LUNCHEON MEAT

The data used in the following analysis were originally collected in a study of the perceptions of 27 kinds of luncheon meat.[2] For the analysis presented here, a selection of the original 33 questions was made to illustrate the use of GCA. Seven questions related to health and image matters were selected for the analysis. We want to find out how these matters are related and which meat-products have a healthy or an unhealthy image. The seven questions selected are shown in Table 3.

Table 3

Seven questions about image-related matters, from the luncheon meat study.

	This luncheon meat...
1	is a healthy product.
2	is a meagre/light product.
3	is a natural product.
4	is a luxury product.
5	is a craftsman's product.
6	is bad for your figure.
7	contains a lot of nutrients.

The questions were answered using Likert scales. The categories from these scales were converted into category-numbers from 1 to 5 (meaning respectively: 1:disagree completely, 2:disagree, 3:neither disagree nor agree, 4:agree, 5:agree completely). In this study 13 assessors participated, but some of them failed to answer the questions for the products they did not know. Because of this, 7 products were deleted from the analysis because there were too many missing scores for these products. The 20 remaining kinds of luncheon meat in the study are shown in Table 4.

2 The data were kindly made available by Oliemans Punter & Partners, Utrecht, The Netherlands.

Table 4
Meat types in the study of luncheon meat.

grilled ham	rare cooked lean beef
cured port belly	cooked liver
corned beef	spreadable liver sausage
course raw dried fermented sausage	minced lean beef
raw dried fermented sausage	liver sausage
coarse liver sausage (Farmer's quality)	cooked ham
raw cured ham	fried minced meat ('meat loaf')
cooked shoulder	cooked chicken filet
smoked raw cured beef	liver sausage (Butcher's quality)
cooked (cured) ham	finely comminuted liver paste

6.1 Results of the analysis: fit and object scores

The GCA analysis, performed with the OVERALS program, was carried out treating the 5 categories of all variables as ordinal. This seemed the most natural choice for the categories of the scale used.

Firstly the dimensionality of the solution had to be chosen. It is advisable to first try an analysis with a high number of dimensions, and then identify the dimensions with a substantial fit. We did two analyses, one in two and one in three dimensions. Table 5 shows the corresponding fit-values. In GCA the fit-values are equal to the eigenvalues per dimension (see section 3.2), so both terms may occur in a GCA context.

Table 5
Eigenvalues per dimension and total fit for a two- and a three-dimensional OVERALS solution for the luncheon meat data.

number of	Eigenvalues for dimension			Total fit
dimensions	1	2	3	
2	0.923	0.839	-	1.763
3	0.917	0.860	0.826	2.604

Keeping in mind that a P-dimensional solution has a maximum fit of P, the column 'Total fit' may help in deciding the dimensionality. The choice of dimensionality is entirely for the data-analyst. There is no clear method for determining the dimensionality, one has to balance between 'parsimony and interpretability' (see e.g. Hofmann and Franke, 1986).

After inspecting the fit one wants to take a look at the space of object scores. This space contains the configuration of the objects from the data, in this case the 20 luncheon meat types. As an illustration the three-dimensional object space is presented first, but in the remainder the two-dimensional result will be used. Two dimensions are easier to display than three, though in practice the third dimension, or even higher dimensions, may have a good interpretation. In such cases these dimensions should of course be taken into consideration too.

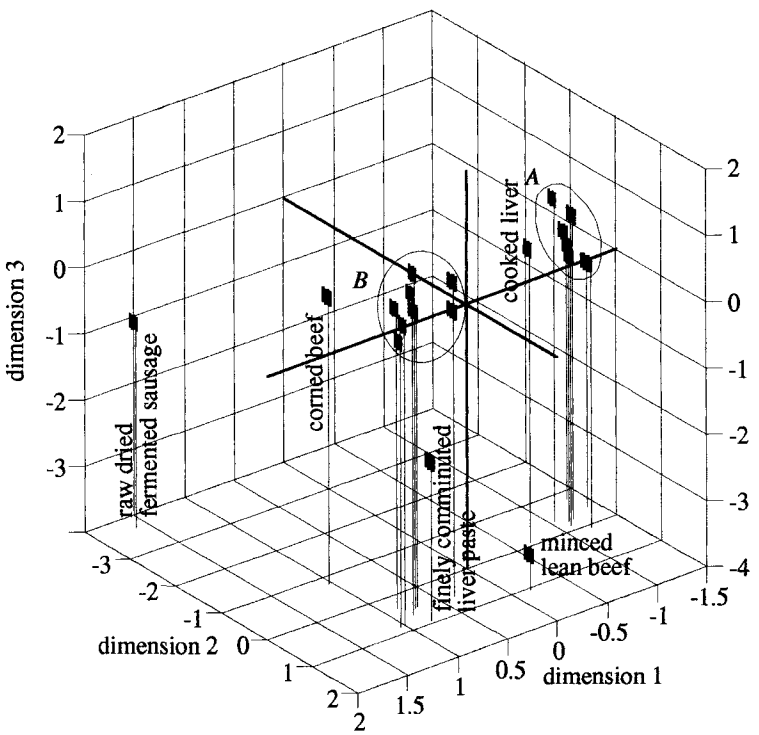


Figure 5. Three-dimensional OVERALS representation of the space of object scores of luncheon meat data.

Figure 5 shows the three-dimensional OVERALS result. This figure shows that, apart from the 'loners' raw dried fermented sausage, corned beef, finely comminuted liver paste and minced lean beef, there are two main clusters of objects, they are indicated *A* and *B* in the figure. *A* contains cured pork belly, coarse liver sausage, spreadable liver sausage, liver sausage, fried minced meat, liver sausage (Butcher's quality), and cooked shoulder; *B* contains raw cured ham, smoked raw cured beef, cooked chicken fillet, cooked ham, grilled ham and rare cooked lean beef.

To be able to see why these clusters appear we need to study the positions of the questions, i.e., to study the component loadings. Because the main object of this analysis is to illustrate

the GCA technique we continue with the two-dimensional result because two-dimensional plots are easier to present and to look at than three-dimensional plots. One could of course plot dimensions 1 against 2 and 1 against 3, but this would mean twice as much plots, which we believed not helpful for understanding the ideas behind GCA analysis and the interpretation of its results, for the illustratory purpose of this analysis. Furthermore the two-dimensional solution captures the most salient aspects of the analysis.

In Figure 6 the two-dimensional space of object scores is presented, this space is the basis for the remainder of this example. Apart from the finely comminuted liver paste and the minced lean beef, two main clusters of objects appear, they are much like the clusters in Figure 5.

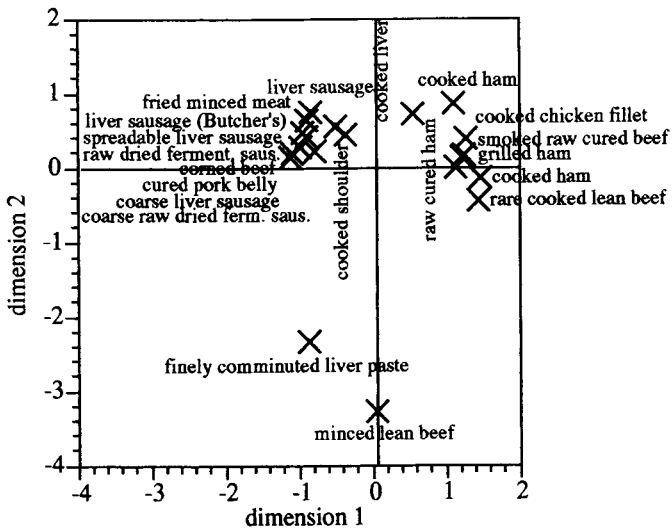


Figure 6. Two-dimensional object scores space from OVERALS applied to the luncheon meat data. The exact position of the objects cannot be seen in this figure, however, the plot is clear enough to illustrate the features of this object scores plot.

6.2 Looking at the questions

Because in GCA each assessor is treated separately from the others, one has the opportunity to study differences in the interpretation of the questions between the assessors. In this study 12 assessors answered 7 questions, one question for an assessor was deleted from the analysis because this assessor gave the same answer for each product on this particular question. This results in one variable which does not vary over products. It has zero variance, so it cold not be used in the analysis. For each question a configuration showing the position of this question for each assessor can be made (plot of component loadings), so that the assessors' use of the questions can be compared (see Figure 7). Note that each plot belongs to the same space of object scores. Also note that we did not connect the points with the origin, to keep the plots clear.

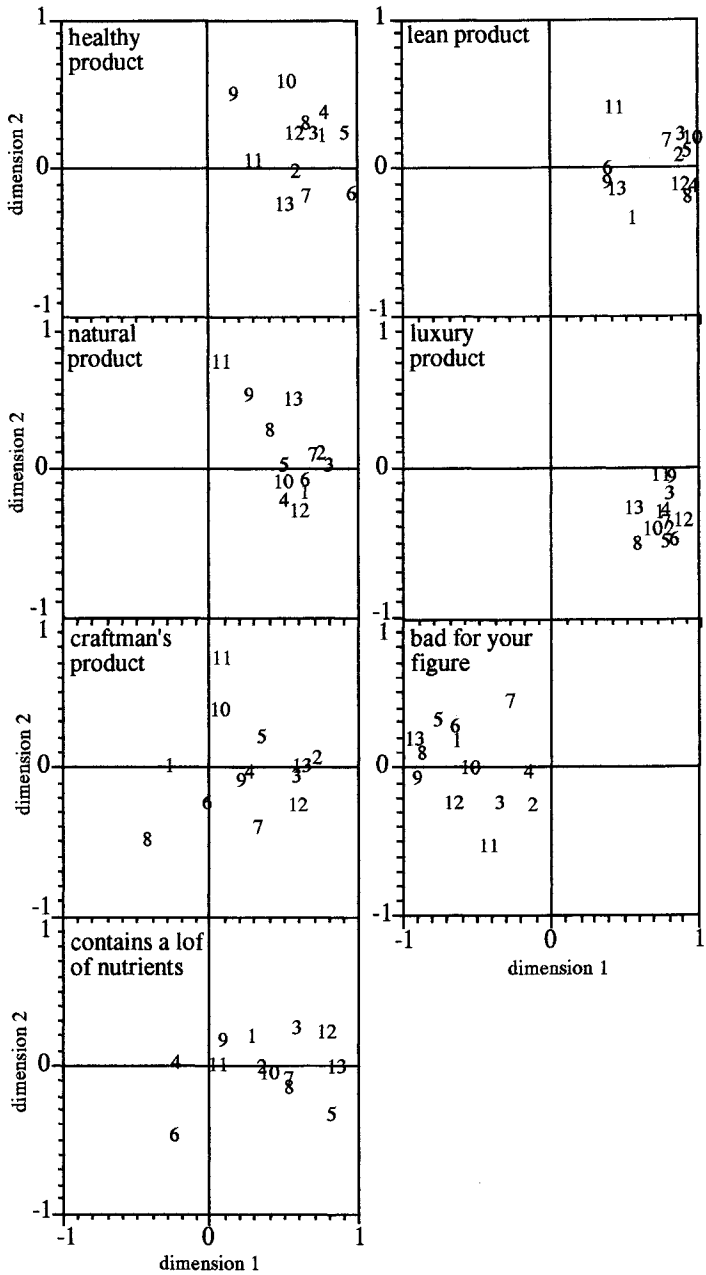


Figure 7. Component loadings of the 7 questions for the 13 assessors. The numbers refer to the assessors.

Figure 7 can be used to assess the homogeneity of the assessors with regard to their interpretation of the questions. Each of the seven panels in Figure 7 contains points representing the 13 assessors. The thighter the cluster of these 13 points, the more homogeneous the assessors were in answering the question. It can immediately be seen from Figure 7 that the assessors agree very well regarding the question about the meats being a luxurious product. They seem to agree the least on the question about the craftsmanship of the make of meat. The question about meagre/light seems to result in one outlier (11) and two groups, agreeing, but to different extents. The same assessor (no. 11) can be identified as a kind of outlier for the question 'bad for your figure'. The question on the healthiness shows a loose kind of cluster, which is hard to separate into sub-clusters. However it is clear that the agreement on the health issue of the meat-products is not clear-cut. Regarding the nutritiousness there are two assessors (no. 4 and no. 6) with a view opposite to most assessors' view. Assessor no. 11 is in the centre of the plot, she/he does not use this question to distinguish between the meats. The question on the naturalness of the meat seems to result in two groups of assessors, one group containing 1, 2, 3, 4, 5, 6, 7, 10 and 12, the other group 8, 9, 11, 13. It could be interesting to study this question more closely, which is done in the next section.

Another way of presenting the questions is to group the questions per assessor instead of grouping the assessors within one question, as is done in Figure 7. In that case for each assessor a plot is made which shows the positions of the question for that particular assessor. This way of presenting the questions can be seen in Van der Burg and Dijksterhuis (1989).

Of course Figure 7 can be used to interpret the directions in the space of object scores. With the help of the component loadings in Figure 7, the properties that distinguish the two clusters of meat types in Figure 6 can be found. In Figure 6, two main clusters of meat types were found, one in the left part, and one in the right part of the plot. From Figure 7 can be inferred that the left part of the space is mainly characterised by 'Bad for your figure', and not by the other questions. The right part of the space in Figure 6 is characterised by the questions about the health, lean-ness, naturalness, luxuriousness and nutritiousness of the meats. With the help of the component loadings in Figure 7 it can be concluded that the two main clusters of luncheon meats obtained have a different image. The meat types cured pork belly, coarse liver sausage, spreadable liver sausage, liver sausage, fried minced meat, liver sausage (Butcher's quality), corned beef, coarse raw dried fermented sausage and cooked shoulder (left in the object scores space, see Figure 6) have a rather negative image. The meat types raw cured ham, smoked raw cured beef, cooked chicken fillet, cooked ham, cooked (cured) ham, grilled ham, cooked liver and rare cooked lean beef, appear to have a positive image.

6.3 Quantifications of categories

To study the questions and the categories in more detail, the quantification of the categories can be studied. Remember that the original questions were presented in five categories, and that these categories received quantifications by the optimal scaling algorithm used in the OVERALS program. These quantifications provide a means to study the questions more closely which is illustrated here using the question about the naturalness of the meat. Figure 8 shows the quantifications of the five categories in the analysis of the luncheon meat data set, for all 13 assessors.

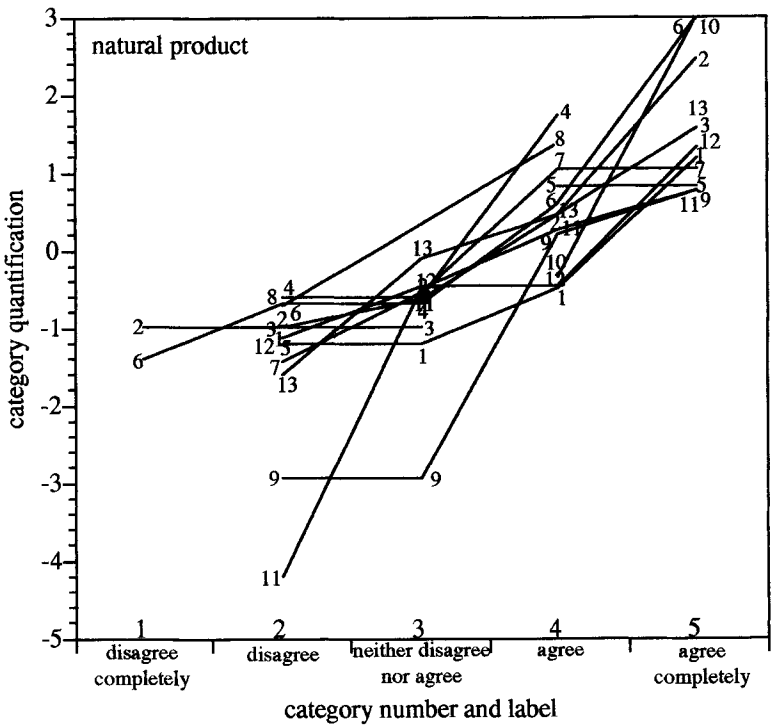


Figure 8. Category quantifications of the question on natural product, for each assessor. The numbers in the figure refer to the assessor numbers (sets).

Figure 8 shows clearly that only two assessors (2 and 6) used category 1 ('disagree completely'). They received about the same quantification. Assessors 9 and 11 have quantifications for category 2 ('disagree') different from the quantification of this category for the other assessors. Assessor 9 has another different quantification. For this assessor the categories 2 ('disagree') and 3 ('neither disagree nor agree') have much lower quantifications compared to these categories for the other assessors. The fourth category ('agree') is the most homogeneously quantified for this question. The fifth category ('agree completely') has three assessors whose categories have somewhat larger quantifications than the other assessors, no. 2, 6 and 10. The numbers 2 and 6 were the only assessors using category 1, number 10 used only the categories 4 and 5. Clearly, assessors 2 and 6 can be identified as the 'extrovert' users of the categories, they use all categories, and the quantifications of category 5 even amplify the 'extremeness' of their use of the categories. Concluding, the assessors 9 and 11 are most different in 'using some' categories of 'natural product'. The categories of most other assessors have received more or less equal quantifications.

The quantifications of the categories for one, or more, questions enables one to study the analysis result in considerable detail. However, a quantification plot like in Figure 8, can be made for each of the 7 questions in this example. Studying them all is a tedious task, which is

useful when one is particularly interested in comparing categories of different questions, or, like in Figure 8, in the use of categories by different assessors.

Another way of looking at the differences between the categories is inspecting the so-called projected centroids (see section 3.4). Figure 9 shows these projected centroids. The distances of the category points along the lines represent the quantifications of the categories. Two categories with the same quantification have the same position. The category points are projected onto the lines connecting the component loadings of each variable with the origin (see also section 3.4).

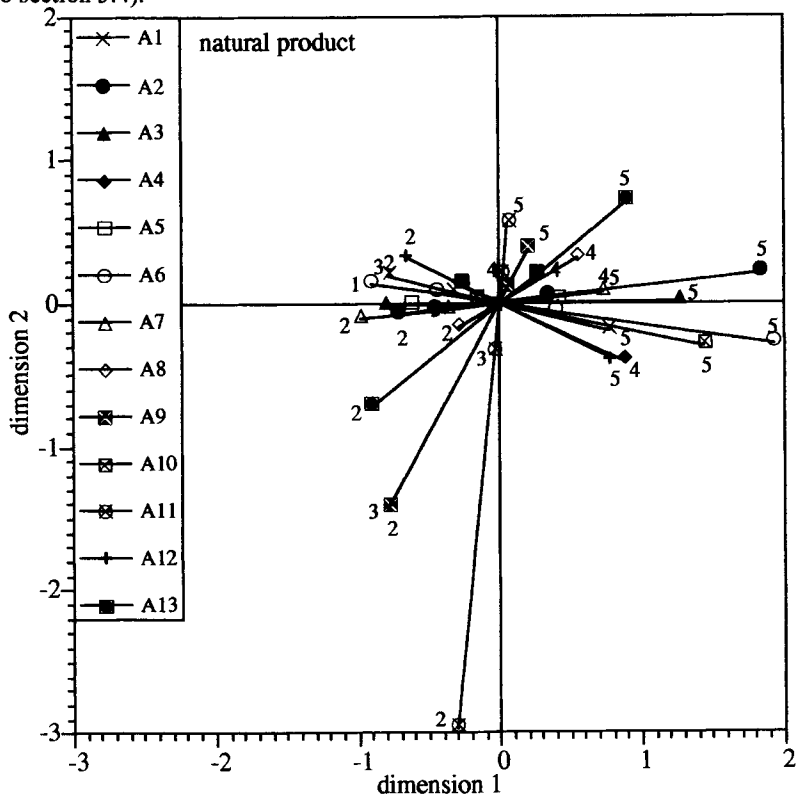


Figure 9. Projected centroids of categories (1 to 5) of the question on natural product. The assessors are coded with the symbols A1 to A13 in the legend.

We saw in Figure 7 for question 3 ('Natural product') that there were differences in answering. The assessors 9 and 11 are different from the other assessors. In Figure 9 the categories are shown as positions on a line with the same direction as the line representing the assessor in the corresponding panel of Figure 7. The individual positions of the categories of question 3 can be seen in Figure 9. There is a line with the 5 categories for each assessor in this plot. The category-numbers 5 ('agree completely') lie mostly in the right part of the plot together with some categories 4 ('agree'). The lowest category (1, 'disagree completely') is not seen much in the plot. This category is not often used by the assessors. Assessor 6 did use

it but the use is comparable with the use of category 2 ('disagree') used by most other assessors. Assessor 1 and 9 used category 3 ('neither disagree, nor agree') not different from category 2. In the plot the positions of these categories coincide.

The positions of the categories for the deviant assessors 9 and 11 show that they differ mainly in the use of the lower categories, compared to the other assessors. The low categories of these assessors lie in the lower-left part of the plot. These assessors apparently have a more negative interpretation of the products.

6.4 Loss and fit of assessors

Each assessor is represented by a set of variables in the analysis. Some assessors' scores are typical for the group of assessors, and other may be different. Assessors with deviating scores do not *fit* very well with the other assessors. The fit values tell how well assessors fit in the solution. The *loss* value measures the lack of *fit*.

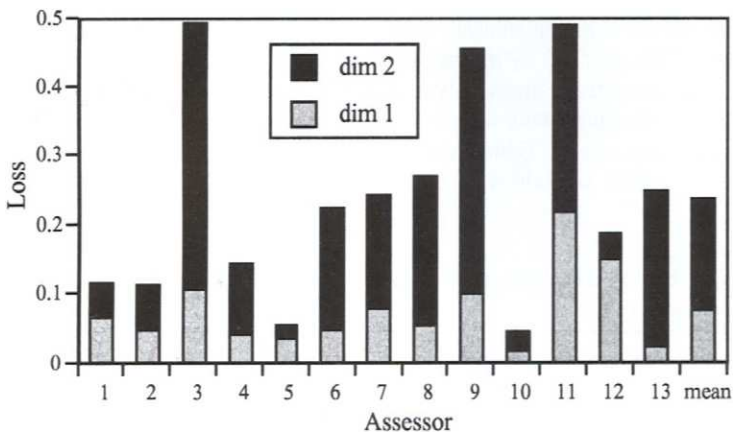


Figure 10. Loss per assessor for the two-dimensional OVERALS solution of the luncheon meat data.

Fit and loss values can be computed for each dimension of the solution. In this case there is a loss value for the first and for the second dimension. The sum of these two loss values indicate the lack of fit in the two-dimensional solution obtained in this example. For each assessor this loss value is presented, for both dimensions, in Figure 10. The figure shows that assessor 11 has the largest loss value in the first dimension. This assessor was already identified as an outlier in Figure 7 and Figure 8. Figure 10 shows that assessor 11 did not fit very well with most other assessors. Another observation from Figure 10 is that assessors 3 and 9 have relative large loss values in the second dimension. Looking at the sum of the loss values it shows that assessor 3, 9, and 11 fit relatively poor in the analysis. The best fitting assessors are 5 and 10, they have the lowest loss value. Assessor 13 has a low loss value in the first dimension, this assessor 'agrees with' the most important features of the solution. The loss

value for this assessor in the second dimension is among the largest loss values, so she/he does not 'agree' with the second dimension of the solution.

6.5 Conclusion

The analysis of the data on the questions on the image of the different types of luncheon meat illustrates the use of an ordinal analysis. GCA helped to study some items in considerable detail, while at the same time, providing an interpretable configuration of all the meat types based on two underlying dimensions. The obtained configuration was shown to contain mainly two groups, one with a positive and one with a negative image. The negative image was 'bad for your figure', the positive image contained the items health, meagre/light, natural, luxury and nutritious.

7. FREE CHOICE PROFILING OF MINERAL WATERS

A number of twenty different mineral waters is judged by K=11 assessors.[3] The same data were analysed in Chapter 7.2 by means of GPA. In the experiment some mineral waters were presented twice, some three times, to the assessors, totalling to 49 presentations. Each judge used his/her own attributes, thus we are dealing with FCP data. The number of attributes per assessor ranges from 3 to 10. Table 6 shows all attributes. Note that the same attribute used by different assessors does not indicate similar ideas of the assessors about this attribute.

Table 6

Attributes used in the FCP of the mineral waters and the no. of the judge that used it (see also Table 9 of Chapter 7.2).

Attribute	judge no.	Attribute	judge no.
bitter	1, 2, 3, 5, 6, 8, 9, 11	balanced	4
neutral	1, 2, 4, 6, 8, 9	persistent	4, 6
taste	1	mineral	5
metal	1, 3, 7, 9, 10, 11	stagnant	5
fluid	1	river	5
salty	2, 4, 7, 8	cool	5
earth	2, 4, 7, 11	sugar	6
hard	2	old	6
acid	3, 4, 11	mushroom	7
paper	3, 10	milky	7
flat	4, 5	energetic	9
dry	4	hazelnut	10
pungent	4	soft	11
rubber	4		

Note that the arrangement in the table does not indicate any relation between attributes in the same row.

---

3 The data were made available by Dr. Pascal Schlich, INRA, Dijon, France.

The perceived intensities of the attributes were scored on a line scale, labelled 'weak' and 'strong' at respectively the left and right end. The original scores ranged from 0 to 100. To use the OVERALS program these scores had to be recoded into a small number of categories. This recoding was such that the chances of the occurrence of a unique marginal frequency was low. When as a result of this recoding a particular category occurs only a few times, a different recoding should be chosen. The OVERALS algorithm is sensitive for categories with low marginal frequencies (see section 3.2). Such categories will receive an extremely high, or low, quantification in the optimal scaling step of the algorithm. The recoding shown in Table 7, produced marginal frequencies in acceptable balance, i.e. each category appeared sufficiently often.

Table 7
Recoding of the original scores of the mineral water FCP data set.

original score	recoded category	approximate meaning
0	1	not perceived/not applicable
1-25	2	weak*
26-75	3	intermediate
76-100	4	strong*

* The line-scales were anchored at the left and right ends by 'weak' and 'strong' respectively.

The recoded data were analysed with a two- and a three-dimensional OVERALS analysis. There are 11 sets, one for each assessor. The data consist of 49 products, since the replicates were taken into the analysis as separate objects. This enables a check of the similarity of the replications in the final configuration of the 49 mineral waters.

7.1 Objects and attributes

First the three-dimensional OVERALS solution was computed, the eigenvalues were 0.682, 0.492 and 0.411, respectively. Because the eigenvalues were rather low, we proceeded with a two-dimensional analysis. The results from this analysis are presented next. The fit of the solution is 1.192, the maximal fit of a two-dimensional solution is 2. The two eigenvalues are 0.691 and 0.502, which sums to the fit. Though the fit is not particularly high it may be worthwhile to inspect the results to find out the reason for this relatively low fit. The first step is to inspect the loss of the individual sets. Figure 11 presents the loss per assessor for the two dimensions of the OVERALS analysis. It is clear from Figure 11 that assessor 4 fits best in the solution. There appear no assessors with an extremely high loss, so no assessors need be deleted from subsequent analyses.

Figure 12 shows the space of object scores containing the 49 mineral waters and Figure 13 gives the component loadings of the attributes in the same space. Like in the results of the analysis of the luncheon meat data, we did not connect the points with the origin. Clearly there are some outlying objects in Figure 12. The mineral waters 15, 16, 17 and the two pairs 1, 2

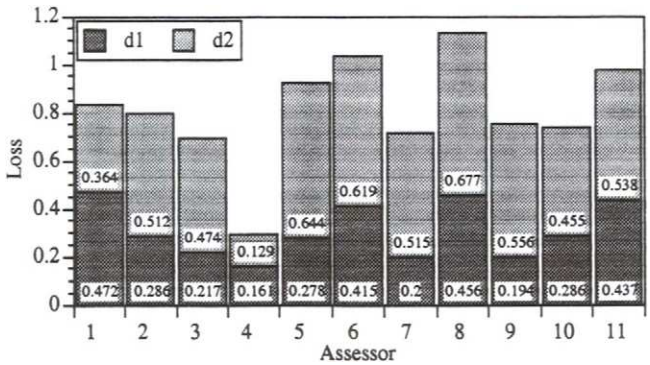


Figure 11. Loss per assessor for the two dimensions of the OVERALS analysis of the mineral water data.

and 12, 22 have rather extreme object scores. They are connected by lines because they are replicates. The component loadings in Figure 13 enable to see the attributes that apply to these outlying objects. It appears that the mineral waters 15, 16 and 17 are characterised by the attributes old6, rubber4, and the cluster of attributes paper3, flat4, dry4, metal3, mushroom7, bitter9, bitter8, and neutral4. The mineral waters 1 and 2 are mineral waters with extreme scores on the attributes salty4, acid11 and paper10, mineral water 22 is characterised by balance4, bitterness1, metal10, bitter5, river5, metal11, earth4 and mineral water 21 by the cluster of attributes metal9, bitter11 earth2, neutral2, metal7, earth7, bitter3, pungent4, metal7, taste1, neutral9, hard2, and neutral1. We do not give an interpretation of this.

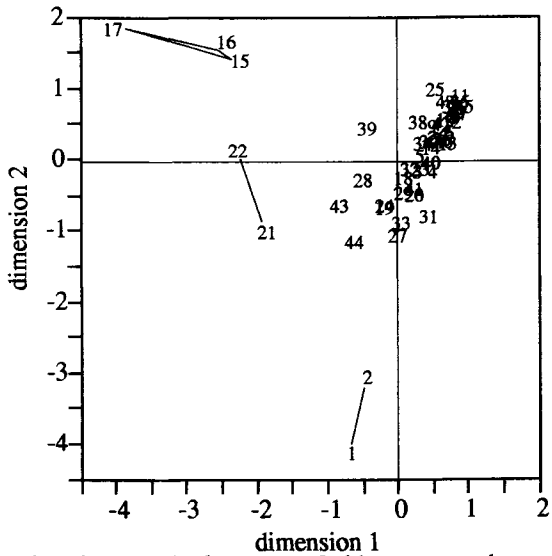


Figure 12. The 49 mineral waters in the space of object scores, the replicated mineral waters that are outliers, are connected by a line.

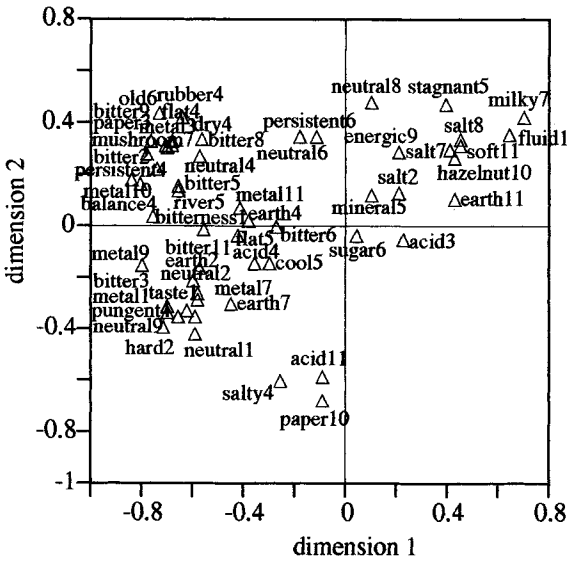


Figure 13. The component loadings of the attributes from all assessors. The number attached to the name of the attribute refers to the number of the assessor who used the attribute.

To reveal structure in the remaining cluster of objects, the analysis is repeated without the outlying mineral waters, leaving 42 objects in the analysis. The fit of this analysis is 1.059, the first two eigenvalues are 0.566 and 0.493. Note that these values are somewhat lower than previous results. Figure 14 shows the losses per assessor of this analysis. The losses are somewhat more evenly spread over the assessors than in the previous analysis.

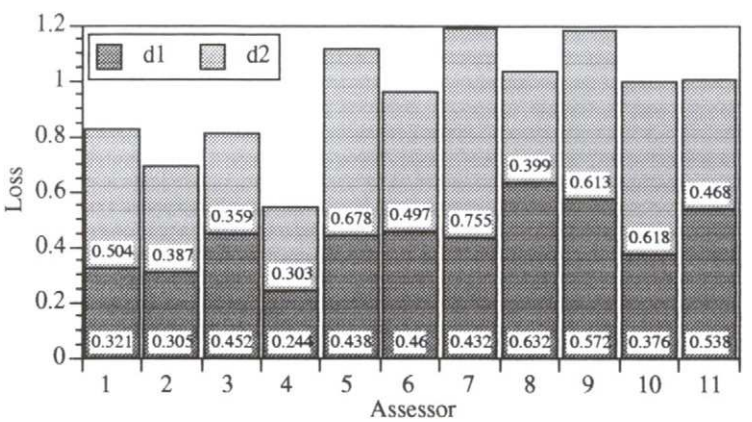


Figure 14. Loss per assessor for a two-dimensional OVERALS analysis after removal of some outliers.

Figure 15 shows the position of the remaining 42 mineral waters in the space of object scores. Replicate mineral waters are connected by a line in Figure 15. The connected points are close together in the direction of the first dimension, they are further apart in the direction of the second dimension. This may reflect the fact that the second dimension contains a certain amount of noise. Also the low second eigenvalue suggests this in Figure 15. Clearly the mineral waters 43 and 44 and the pair 25 and 26 are distinct from the other objects. The mineral waters no. 45, 46 and the trio 47, 48, 49 are close together and distinct from the other mineral waters. Higher in the figure is the pair of mineral waters 38 and 39. In addition to these and disregarding the distances between corresponding mineral waters in the second dimension - mentally replacing each pair or trio of mineral waters by its centroid-, the following two main clusters of mineral waters can be distinguished. Left in Figure 15 we find the numbers 10, 11; 35, 36, 37; 12, 13, 14; 40, 41, 42; 8, 9 and the pair 6, 7. In the right part of Figure 15 there are number 18, 19, 20; 3, 4, 5; 32, 33, 34 and the trio 27, 28 29.

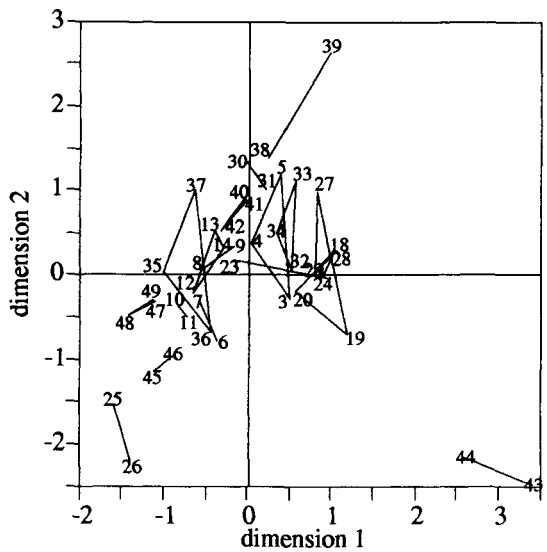


Figure 15. Object scores of a two-dimensional OVERALS after removal of some outliers.

Figure 16 presents the positions of the individual attributes in the two-dimensional OVERALS solution. This plot can be used in conjunction with Figure 15 to find out the properties of the different clusters of mineral waters. When we try to identify trends of attributes it appears that most attributes 'metal' and 'bitter' lie in the right part of Figure 16. The attributes cool, river, pungent and hard are found here too. The left part of Figure 16 contains salt, hazelnut, soft, milky and stagnant. This may indicate a distinction (in Figure 15) between two kinds of mineral waters: the first could be coined strong and fierce, the latter soft and easy.

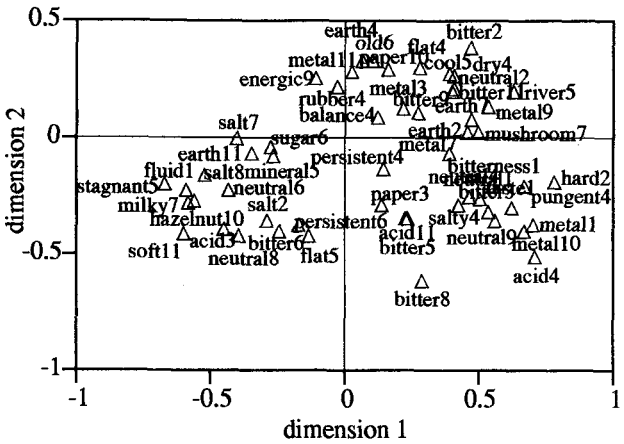


Figure 16. Component loadings of a two-dimensional OVERALS analysis after removal of some outliers.

8. CONCLUSION

In this chapter the GCA method is illustrated in three applications from sensory science. In analysing sensory-instrumental relations the ability of GCA to analyse more than two data sets, and different measurement levels together, proved useful. In the analysis of conventional profiling data, in the example on luncheon meat, of an ordinal level, GCA helped to study some items in considerable detail, while at the same time providing an interpretable configuration of all the meat types based on two underlying dimensions. The FCP data set was, after recoding, analysed using ordinal transformations. Some outliers were removed and the final, two-dimensional configuration of mineral waters, and of attributes, revealed some interesting ideas about the judgement of mineral waters.

Generalised Canonical Analysis, implemented in the OVERALS program, is a useful method to analyse sensory profiling data. Especially the ability of the method to analyse nominal, ordinal and numerical, and mixed, measurement levels proves useful. The combination of nonlinear transformations of the variables and the *K*-sets character of the method, makes it a powerful tool for the analysis of individual sensory profiling data, which is often of an ordinal, rather that a numerical, measurement level. For the same reasons GCA has also proved useful in the analysis of the relations between sensory and instrumental data.

9. ACKNOWLEDGEMENTS

The authors thank Patrick Groenen (Department of Datatheory, Leiden University, The Netherlands) for reading the chapter and providing valuable comments. P.C. Moerman (TNO Nutrition and Food research, Zeist, The Netherlands) is thanked for kindly providing the English translations of the originally Dutch names in the luncheon meat study.

10. REFERENCES

Arabie, P. Carroll, J.D., DeSarbo, W. (1987). Three-way scaling and clustering. Sage University Paper series on Quantitative Applications in the Social Science, no. 65. Beverly Hills: Sage Publications.

Arnold, G.M., Williams, A.A., (1986). The use of Generalised Procrustes Techniques in sensory analysis, In: Statistical Procedures in Food Research, Piggot, J.R. (Ed.). London: Elsevier.

Carroll, J.D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. Proceedings of the 76th Annual Convention of the American Psychological Association, 5, 227-228.

De Leeuw, J., Van der Burg, E. (1986). The permutational limit distribution of generalized canonical correlations. in: E. Diday, Y. Escoufier, L. Levart, J.P. Pagès, Y Schektman, R. Tomassone (Eds.). Data Analysis and Informatics IV, 509-521. Amsterdam: North Holland

Dijksterhuis, G.B., (1995a). Assessing Panel Consonance. Food Quality and Preference, 6, 7-14.

Dijksterhuis, G.B., (1995b). Multivariate Data Analysis in Sensory and Consumer Science. Ph.D. Thesis. Dept. of Datatheory, Leiden University, The Netherlands.

Dijksterhuis, G.B., Punter, P.H., (1990). Interpreting Generalized Procrustes Analysis 'Analysis of Variance' tables, Food Quality and Preference, 2, 255-265.

Edgington, E.S. (1987). Randomization Tests. New York: Marcel Dekker, Inc.

Efron, B. (1982). The Jackknife, the Bootstrap, and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics. no. 38. Philadelphia: SIAM.

Gifi, A. (1990) Nonlinear Multivariate Analysis. Chichester: J. Wiley and Sons.

Gittins, R. (1985). Canonical Analysis. A Review with Applications in Ecology. Berlin: Springer Verlag.

Good, P. (1994). Permutation Tests. Berlin: Springer Verlag.

Greenacre, M. (1984). Theory and Applications of Correspondence Analysis. New York: Academic Press.

Hoffman, D.L. Franke, G.R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. Journal of Marketing Research, 23, 213-227.

Horst, P. (1961). Relations among m sets of measures. Psychometrika, 26, 129-149.

Hotelling, H. (1936). Relations between two sets of variates. Biometrika, 28, 321-327.

Kettenring, J.R. (1971). Canonical analysis of several sets of variables. Biometrika, 56, 433-451.

Kiers, H.A.L., Van der Burg, E. (1994). Simple Structure Rotation for Nonlinear Canonical Correlation Analysis and Related Techniques. Dept. of Psychology. University of Groningen, The Netherlands.

Koppenaal, E. (1991). Project: Objectieve meting voor detectie en voorspelling van de sensorische kwaliteiten stevigheid en meligheid in appellen. (Project: Objective measurements for detecting and predicting the sensory qualities firmness and mealiness of apples. In Dutch) Internal Report ATO-DLO Institute for Agrotechnology, Wageningen, The Netherlands.

Kroonenberg, P.M. (1983). Three-mode Principal Component Analysis. Leiden: DSWO Press.

Kuhfield, W.F., Sarle, W.S., Young, F.W. (1985). Methods in generating model estimates in the PRINQUAL macro. SUGI-Proceedings, 962-971. Cary, NC, USA: SAS-Institute Inc.
Miller, R.G. (1974). The jackknife: A review. Biometrika, 61, 1-15.
Nishisato, S. (1980). Analysis of Categorical Data: Dual Scaling and its Applications. Toronto: University of Toronto Press.
SAS/STAT (1990). User's guide. Cary, NC, USA: SAS-Institute Inc.
SPSS (1990). SPSS Categories™. Chicago, USA: SPSS Inc.
Stone, H., Sidel, J.L. (1993). Sensory Evaluation Practices, second edition, San Diego: Academic Press, Inc.
Tatsuoka, M.M. (1988). Multivariate analysis. Techniques for educational and psychological sciences (second edition). New York: Macmillan Publishing Company.
Ter Braak, C.J.F. (1990). Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika, 55, 519-531.
Van de Geer, J.P. (1984). Linear relations between k sets of variables. Psychometrika, 49, 79-94.
Van der Burg, E., De Leeuw, J. (1983). Nonlinear canonical correlation, British Journal of Mathematical and Statistical Psychology, 36, 54-80.
Van der Burg, E. (1988). Nonlinear Canonical Correlation and Some Related Techniques. Leiden: DSWO Press.
Van der Burg, E., De Leeuw, J. (1988). Use of the multinomial Jackknife and Bootstrap in generalized nonlinear canonical correlation analysis. Applied Stochastic Models and Data Analysis, 4, 159-172.
Van der Burg , E., De Leeuw, J., Verdegaal, R. (1988). Homogeneity analysis with k sets of variables: an alternating least squares method with optimal scaling features. Psychometrika, 53, 177-197.
Van der Burg, E., Dijksterhuis, G.B. (1989). Nonlinear Canonical Correlation Analysis of Multiway Data, In: Coppi, R., Bolasco, S. (Eds.) Multiway Data Analysis, 245-255, Amsterdam, North-Holland, Elsevier Science Publishers B.V.
Van der Burg, E., Dijksterhuis, G.B. (1993a) An application of nonlinear redundancy analysis and canonical correlation analysis. In: Psychometric Methodology. Proceedings of the European Meeting of the Psychometric Society, Trier, Germany. In: R. Steyer, K.F. Wender, K.F. Widaman (Eds.), 74-79. Stuttgart: Gustav Fischer Verlag.
Van der Burg, E., Dijksterhuis, G.B. (1993b). Nonlinear generalized canonical analysis: Theory and an application from sensory research. In: Advances in Longitudinal and Multivariate Analysis in the Behavioral Sciences. Oud, H., Blokland-Vogelesang, R. (Eds.), 193-203. Nijmegen: ITS.
Van der Burg, E., De Leeuw, J., Dijksterhuis, G.B. (1994). OVERALS Nonlinear canonical correlation with k sets of variables. Computational Statistics & Data Analysis, 18, 141- 163.
Williams, A.A., Langron, S.P. (1984). The use of free-choice profiling for the evaluation of commercial ports. J. Sci. Food Agric., 35, 558-568.
Young, F., De Leeuw, J., Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. Psychometrika, 41, 505-529.
Young, F.W., Takane, Y., De Leeuw, J. (1978). The principal components of mixed measurement multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 43, 279-281.

Young, F.W., Takane, Y., De Leeuw, J. (1978). The principal components of miced measurement multivariate data: An alternating least squares method with optimal scaling features. Psychometrika. 43, 279-281.

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

DEFINING AND VALIDATING ASSESSOR COMPROMISES ABOUT PRODUCT DISTANCES AND ATTRIBUTE CORRELATIONS

Pascal Schlich

INRA, Laboratoire de Recherches sur les Arômes,
17 rue sully, BV 1540, 21034 Dijon cedex, France

1. INTRODUCTION

Sensory profiling of food is the process by which assessors give scores to a number of products for several attributes. The statistical analysis of profiling data requires multivariate techniques in which the attributes are the different variables. Principal Component Analysis (PCA; Jolliffe, 1986), the most basic multivariate technique, is widely used by sensory scientists to describe the data set composed of the product mean scores as the observations and the attributes as the variables. This is a reduced view of the data as it does not take into account the variance of the product mean scores due to individual differences. In the univariate framework, it is fairly accepted that no mean should be computed without its standard deviation. We argue that the same should be required on the multivariate side. Moreover, many people are used to performing this PCA on the basis of the correlation matrix, that is to perform the so-called normalized PCA. With this practice, an attribute having product means not significantly different, which is stated by means of an analysis of variance (ANOVA; Scheffe, 1959), has the same weight as a discriminant attribute among the products. To overcome this problem, one can include in PCA only the attributes being significant for the product effect. Although this practice protects to some extent against the previous problem, it suggests that the required method would be the Canonical Discriminant Analysis (CDA; Mardia *et al.*, 1979) of the product effect. CDA is indeed the natural multivariate extension of the one-way ANOVA. The input of this CDA is the full data set, where the assessors stand as replicates. But due to psychological and/or physiological reasons, the assessors may use the scale in different ways, and should therefore be considered as a block effect. Consequently, we would recommend centering each attribute to a zero mean by assessor as a pre-treatment of CDA. Although less important than centering, one could also consider standardization of each attribute to a unit variance by assessor. Unfortunately, CDA and these pre-treatments seem to have been overlooked by sensory scientists.

But the above mentioned flaws are even less important than the basic problem of sensory profiling, which can be summarized in a single statement the attributes are not statistical variables. A statistical variable is a vector of n measurements of a random variate obtained on n experimental units characterized by their values for different controlled factors. Our point is to say that the meaning of the random variate is the same over the experimental units. In agronomy for instance, the yield or the plant size has the same meaning for every experimental

unit, whatever the levels of treatment and block factors are. In sensory profiling on the contrary, it is likely that the measurements of a given attribute from two different assessors measure two different sensory concepts. In such a case, computing for a given attribute a mean over the assessors can make no sense. Although good, but expensive, training of the assessors can substantially reduce this kind of problem, it is almost impossible to completely prevent it. Free-choice profiling (FCP; Williams & Langron, 1984), in which each assessor scores his own list of attributes, was a definite contribution to solving this basic problem. The variables being different among the assessors, neither PCA nor CDA can be applied to FCP, unless a separate analysis is performed by assessor. Although some individual analyses can be useful to elucidate particular points during an analysis of data, they are not practical enough to manage to be recommended as a basic analysis. Therefore, Williams and Langron (1984) proposed Generalised Procrustes Analysis (GPA; Gower, 1975; Arnold & Williams, 1986) to analyse FCP. The FCP-GPA coupling became popular in the sensory field and people realized that GPA could also be applied to conventional profiling data. Less well known is the fact that GPA makes possible a posteriori individual attribute selection leading to what can be called a "simulated free-choice profiling" (Schlich, 1993). Recent improvements in GPA were proposed thanks to applications in sensory analysis (Dijksterhuis & Punter, 1990; Dijksterhuis & Gower, 1991/2). Although a chapter in this book (4.2) fully describes GPA, one can say that GPA defines a consensus among assessors on the product differences by means of an iterative algorithm. Wakeling *et al.* (1992) define a test of the significance of the consensus based on permutations of the product labels by assessors. Although GPA is a significant improvement over PCA of the mean scores, a number of questions are still open concerning the analysis of sensory profiles:

- How can the dimensionality of individual sample space be estimated?
- How can the similarity between two individual sample spaces be measured?
- How can the agreement between two individual sample spaces be statistically tested?
- How can a consensus sample space be derived by an analytical solution?
- How can the significance of this compromise be tested without computing numerous permutations?
- How can several panels profiling the same products be compared?
- How can exchangeability of assessors among panels be tested?
- How can assessors be compared on the basis of attribute correlations instead of sample distances?

The present chapter aims to introduce in the sensory field a French statistical framework allowing firstly to deal with FCP and secondly to answer these questions. The basis of this framework is the RV coefficient (Escoufier, 1973; Robert & Escoufier, 1976), which is a generalised correlation coefficient between two sets of variables recorded from the same samples. It gives a way to quantify the agreement between two assessors about the sample differences. The RV coefficient is also useful in sensory science for relating sensory to instrumental measurements (Schlich *et al.*, 1987) and for analysing gas-chromatography data (Schlich & Guichard, 1989). The French acronym STATIS stands for "Structuration des Tableaux A Trois Indices de la Statistique", which could be translated into "structure of 3-way data sets in statistics". The technique was originated by L'Hermier des Plantes (1976) and was

fully described in Lavit (1988) and Lavit *et al.* (1994). It is a non-iterative 3-way multivariate analysis, based on the RV coefficient, which defines a compromise among assessors about the sample differences. Kazi-Aoual (1993) and Kazi-Aoual *et al.* (1995) defines an exact permutation test for RV coefficient which does not require numerous permutations to be actually performed. Another exact permutation test makes it possible to state whether the assessors from several panels profiling the same products can be considered as exchangeable among panels (Kazi-Aoual, 1993; 1992).

In the present chapter, these techniques are applied to an interlaboratory sensory analysis of 16 coffees evaluated by 11 different trained panels located in 8 different European countries. This experiment was organized and conducted by the European Sensory Network (ESN). ESN is currently completing a book presenting the results on that experiment, which also included consumer trials in 8 countries with half of the same coffees. Only a part of the profiling data is used in this chapter. The ESN book will make the whole set of coffee data available to the reader. More information about ESN can be obtained from the author.

2. METHODS

2.1 Raw data, sample weights and metrics

Let k be the total number of assessors and assume that assessor i scored p_i attributes $(i = 1, ..., k)$ for n samples. Let X_i be the matrix containing the scores of assessor i. The n rows of X_i are the samples, whereas the p_i columns of X_i are the attributes. Throughout this chapter, it is assumed that the attributes are centered within each X_i, that is per assessor. With the exception of Dual STATIS, every technique described in this chapter allows the attributes to be different among assessors, making it possible to handle FCP or any kind of individual selection of attributes.

Although in most applications every sample has the same weight $(1/n)$, the techniques described in this chapter work with unequal weights. In this case, these weights are arranged into a diagonal matrix D of size n.

The question of choosing the attribute weight system, also called metrics, is more important. Metrics is a way of computing distances between samples. Generally speaking, it is a positive symmetric matrix Q of size p (the number of attributes) :

$$Q = (q_{lm})_{l,m=1,p} \tag{1}$$

As Q can be different among the assessors, sometimes it will be called Q_i. The squared distance between two samples $x = (x_l)$ and $y = (y_l)$, $(l = 1, ..., p)$, is given by :

$$d^2(x,y) = \sum_{l=1}^{p}\sum_{m=1}^{p}(x_l - y_l)q_{lm}(x_m - y_m) \tag{2}$$

For instance, in the framework of PCA two different metrics are commonly used . For the first one, Q is the identity matrix containing 1 as diagonal elements and 0 elsewhere, which

corresponds to the PCA of the covariance matrix and to the usual Euclidean distance :

$$d^2(x,y) = \sum_{l=1}^{p} (x_l - y_l)^2 \tag{3}$$

For the second one, Q is a diagonal matrix composed of the inverse of the attribute variances (σ_l^2, $l = 1, ..., p$), which corresponds to the PCA of the correlation matrix, also called standardized PCA, and to the weighted distance :

$$d^2(x,y) = \sum_{l=1}^{p} (x_l - y_l)^2 / \sigma_l^2 \tag{4}$$

As in the introduction section, where non standardized PCA was recommended for analysing the mean score products, we again recommend the non standardized PCA for analysing individual data sets. Many reasons can justify this choice, a few others can be found against it. It is out of the scope of this chapter to open such a discussion, because the techniques proposed in this chapter can accommodate any individual metrics. Furthermore, other metrics can be used with profiling data. For instance, we advocated (Schlich, 1993) the use of the individual Mahalanobis distance :

$$Q_i = (X_i' D X_i)^{-1} \tag{5}$$

when dealing with free-choice profiling, or even with conventional profiling as soon as different attribute correlation structures are expected among the subjects.

Finally, as Q is positive, it is possible to find a square matrix T of size (n,n) such as :

$$Q = T\,T' \tag{6}$$

Analysing X with metrics Q is thus equivalent to analyse :

$$Y = T\,X \tag{7}$$

with the classical identity metrics. Therefore, without loss of generality, it is assumed right now that the metrics Q is the identity and that the sample weights are all equal to $1/n$.

To summarize this section, the reader should remember that :

- Individual attributes are centered
- RV coefficient and STATIS work with any set of product weights (the same for each assessor) and with any individual metrics (which can be different by assessor)
- The selection of attributes and the choice of the metrics are two essential steps in the analysis of sensory profiles which are not addressed in this chapter.

2.2 Association matrix of individual sample space

For each assessor, the samples can be seen as a centered cloud of n points in a multidimensional space spanned by the attributes. The dimensionality of this individual sample space is equal at most to P_i :

$$P_i = \min(n-1, p_i) \tag{8}$$

The distance between two samples in this space measures the magnitude of the sensory differences between them. Because of correlations between attributes, it is likely that the main sample differences can be summarized by fewer dimensions than P_i. These new variables, called the principal components, are given by the successive eigenvectors of the association matrix of size (n,n) :

$$W_i = X_i X_i' \tag{9}$$

which contains the usual scalar products among the samples. The distance d_{uv} between samples u and v is linked to their scalar product W_{uv} through the formula:

$$d_{uv}^2 = w_{uu} + w_{vv} - 2w_{uv} \tag{10}$$

Inversely, because the columns of X_i are centered the formula (10) can be reversed into :

$$w_{uv} = -\frac{(d_{uv}^2 - d_{u.}^2 - d_{.v}^2 + d_{..}^2)}{2} \tag{11}$$

where :

$$d_{u.}^2 = (1/n)\sum_v d_{uv}^2 \qquad \text{and} \qquad d_{.v}^2 = (1/n)\sum_u d_{uv}^2 \tag{12}$$

$$d_{..}^2 = (1/n^2)\sum_{u,v} d_{uv}^2 \tag{13}$$

Finally, it should be remembered that the association matrix W_i contains the full information about the multidimensional differences among samples found by subject i. The association matrices are the basis of the subject comparisons in STATIS, which can therefore be done even if the assessors did not score the same attributes.

2.3 Estimating the dimensionality of an individual sample space

What PCA does is to decompose the total multidimensional variance of the sample space into successive non-correlated components which account for the maximum of this information. Precisely, the amount of variance given by the l-th principal component is the eigenvalue λ_l, of $(1/n)W_i$ $(l = 1, ..., P_i)$. A classical, but still difficult, question is to decide how many components should be analysed. We think that it is best to use resampling techniques such as

cross-validation or bootstrap (Efron & Tibshirani, 1993). However these techniques are not widely available within the statistical softwares yet and can be time-consuming. The β_i coefficient can actually be understood as an estimation of the dimensionality of the individual sample space from subject *i*:

$$\beta_i = \frac{(\text{trace}(W_i))^2}{\text{trace}(W_i^2)} \tag{14}$$

where the trace of a matrix is the sum of its diagonal elements. It is possible to write β_i as a function of the eigenvalues λ_l:

$$\beta_i = 1 + \frac{(2\sum_{l<m}\lambda_l\lambda_m)}{\sum_l \lambda_l^2} \tag{15}$$

The following properties can be derived from equation (15):

$$1 \leq b_i \leq P_i \tag{16}$$

$\beta_i = 1$ if and only if a single eigenvalue is not null (17)

$\beta_i = P_i$ if and only if the P_i eigenvalues are all equal (18)

Property (17) says that the lowest dimensionality (a single axis) is obtained when all the attributes are fully correlated, whereas property (18) says that the highest dimensionality is obtained when no correlation at all exists among the attributes. These properties make it clear as to why β_i can be seen as a dimensionality coefficient. Although we do not trust this coefficient as being the exact truth about the number of dimensions involved in sensory evaluation of a set of products, we strongly rely on it for comparing dimensionality of several individual sample spaces. This concept of dimensionality is really important for the panel leader for choosing the number of attributes to be included in the profile. The β coefficient suggests to the panel leader a minimal number of ideal attributes which should be sufficient to span the sample differences. However it is almost impossible to define such ideal attributes, therefore it is recommended to include a number of attributes being at least about the double of β.

To summarize this section, it should be remembered that the β coefficient makes it possible to compare dimensionalities of individual sample spaces, which can be understood as individual complexities of assessment.

2.4. Comparing two individual sample spaces by means of the RV coefficient

The quantity :

$$\langle W_i, W_j \rangle = \text{trace}(W_i W_j) = \sum_{l,m} w_{lm}^i w_{lm}^j \tag{19}$$

is the natural scalar product between two matrices, where w^i_{lm} is the (l,m)-th element of matrix W_i. It is a generalised covariance coefficient between W_i and W_j matrices. The greater $<W_i,W_j>$ is, the more similar assessors i and j are in terms of their raw product distances. The quantity :

$$<W_i,W_i> = \text{trace}(W_i^2) = \sum_{l,m} {w^i_{lm}}^2 = \sum_l \lambda_l^2 \tag{20}$$

is consequently the norm of W_i or a generalised variance for subject i. The greater this quantity, the more different the products are for this subject. In this context, the RV coefficient is defined as :

$$\text{RV}(W_i,W_j) = \frac{< W_i,W_j >}{\sqrt{< W_i,W_i >.< W_j,W_j >}} \tag{21}$$

and appears as a generalised correlation coefficient between W_i and W_j matrices; it is worth pointing out that the RV coefficient is the classical Pearson correlation coefficient between the association matrices arranged into vectors of size n^2. One can prove that $\text{RV}(W_i,W_j)$ is between 0 and 1. The closer the RV is to 1, the more similar assessors i and j are in terms of their standardized product distances.

Therefore, the comparison between two assessors can be based either on generalised covariance given by formula (19) or on generalised correlation given by formula (21). The author recommends the latter, because the former depends on the use of the scale. For instance, an assessor who tends to use a small portion of the scale for every attribute will get smaller covariance with the other assessors but not automatically a smaller RV. But the reader must understand that this matrix standardization is different from the classical attribute standardization, which is done by a PCA of correlation matrix. Contrarily to this normalized PCA, RV takes into account the differences between attribute variances for a given assessor.

To summarize this section, one should remember that the RV coefficient is a measure of the similarity between two individual sample spaces. RV is the classical correlation coefficient between the two square matrices of sample scalar products of size (n,n) having been previously arranged within two vectors of size n^2.

2.5. Testing significance of a RV coefficient by an exact permutation test of the products

Testing the significance of a given RV value would require complicated parametric assumptions. Therefore, a non parametric alternative (Schlich, 1993) consists in permuting the product labels within X_i without permuting correspondingly the product labels within X_j and to recompute the RV coefficient. Providing that the two assessors agree to some extent about the sample differences, one can expect this "permutated RV" to be lower than the actual RV. This process is repeated a large number of times (say 100) in order to derive the 95 % quantile from the distribution of RV under permutation. If the actual RV is greater than this empirical quantile, it can be concluded that the agreement between the two assessors is better than what can be obtained by chance. The computations required by this permutation test can be too time-consuming to be easily implemented on a micro-computer. Kazi-Aoual (1993) proposed an efficient way to avoid computing a number of permutations. This author established

formulas (22) and (23) giving respectively the mean and the variance of all the *n!* possible permutated RV coefficients:

$$E_p[RV(W_i,W_j)] = \frac{\sqrt{\beta_i \beta_j}}{n-1} \tag{22}$$

$$V_p[RV(W_i,W_j)] = \frac{(n-1-\beta_i)(n-1-\beta_j)(2n(n-1)+(n-3)c_i c_j)}{(n+1)n(n-1)^3(n-2)} \tag{23}$$

where :

$$c_i = \frac{(n-1)(n(n+1)\delta_i - (n-1)(\beta_i+2))}{(n-3)(n-1-\beta_i)} \tag{24}$$

with :

$$\delta_i = \frac{\sum_l w_{ll}^{i\,2}}{\text{trace}(W_i^2)} \tag{25}$$

Therefore, a normalized deviation between the actual RV and its distribution under permutation can be computed as follows:

$$NRV[W_i,W_j] = \frac{\mathbf{RV(W_i,W_j)} - \mathbf{E_p[RV(W_i,W_j)]}}{\sqrt{V_p[RV(W_i,W_j)]}} \tag{26}$$

this normalized RV coefficient is a measure of the agreement between assessors *i* and *j*. Assuming a normal distribution of the permutated RV coefficients, formula (26) defines a test statistic for the null hypothesis of no better agreement between assessor *i* and *j* than what can be obtained after permutation of the label products. One can expect this value to be roughly greater than 2 if the agreement between the assessors *i* and *j* is better than what can be obtained by chance. Although the normal assumption has not been proved till now, it has been observed in practice when performing 100 permutations (Schlich, 1993). Anyway, the exact probability level of this test is not necessary, because the experimenter is most interested in detecting when two assessors do not agree more than what can be obtained by chance.

To summarize this section, one should keep in mind that because the magnitude of a RV coefficient depends on both the number of observations and the number of variables in the two data sets, there is a need for a statistical test of RV significance. An exact non parametric permutation test is available, it makes it possible to declare whether two assessors agree more than what can be just obtained by chance.

2.6 Defining and interpreting a compromise sample space among the assessors by means of the STATIS method

A natural way to define a compromise among subjects on the sample differences would be to compute:

$$W = (1/k) \sum_i W_i \tag{27}$$

and to analyse this matrix by means of a principal co-ordinate analysis (PCO; Gower, 1966) in order to map the samples in accordance with the distances induced by $W_.$. In most cases, the author would recommend this analysis instead of PCA of the mean scores, because it does not require "attribute alignment" among subjects and consequently is able to cope with FCP. Unfortunately, this simple method for analysing sensory profiles seems to be unknown to sensory scientists.

The STATIS compromise W differs from this natural compromise $W_.$ because the latter is a classical mean, whereas the STATIS compromise is a weighted mean of the W_i :

$$W = \sum_i a_i W_i \tag{28}$$

where $(a_i)_{i=1...k}$ is the first eigenvector of the matrix of size (k,k) containing the RV coefficients between assessors. The components of this vector are positive and normalized to get a sum equal to one. This vector represents the "principal agreement among assessors". Thus, the greater the a_i, the more assessor i agrees with the panel on the sample differences. The strategy of STATIS is therefore to put weights on subjects proportionally to their agreement with the panel. Therefore, the weight of an outlier should be close to 0.

Replacing W_i by W in formula (14) makes it possible to estimate a dimensionality of the compromise, which can be an indication about the number of dimensions to be interpreted.

The product coordinates on the axes of the compromise are obtained by a PCO of W. The q compromise components are the q first eigenvectors of W being standardized to have a variance equal to the corresponding eigenvalues. These components are arranged as the columns of a matrix C of size (n,q). The interpretation in terms of sensory attributes can be conducted thanks to the covariances or correlations between the individual attributes and these compromise components. It seems to us that the use of covariances is more logical when no attribute standardization was initially applied to the data as in a PCA of the covariance matrix. Being computed on an individual basis these covariance or correlation coefficients are numerous and it is sometimes necessary to summarize them by averaging scores over assessors having the same attributes before computing these coefficients. As no biplot property holds in this context, we are used to producing a covariance or a correlation plot not superimposed on the compromise plot. Conversely, it is possible to superimpose individual sample spaces on the compromise plot by a classical technique of projection of supplementary elements in multivariate data analysis. The sample coordinates from assessor i on the q compromise components are given by the columns of the following matrix C_i of size (n,q) :

$$C_i = W_i C E \tag{29}$$

where E is a diagonal matrix of size (q,q) containing on the diagonal the inverses of the square roots of the eigenvalues of the compromise.

The compromise location of a given product is the barycenter, for the STATIS weight system, of the k individual locations of this product. The smaller the dispersion of the individual locations of a product around its compromise, the better the agreement among assessors on this product is. In order to better visualize this dispersion, one can draw the convex hulls gathering individual locations of the same products. When two convex hulls do not overlap too much, a logical rule of thumb is to consider the two associated products as different. Alternatively, one can draw for a given product a 95 % confidence convex hull, which is the smallest convex set gathering at least 95 % of the individual locations of this product. Drawing and looking at a convex hull can also be simplified by drawing and looking at a confidence ellipsoïde, which requires a bi-normal assumption to be fulfilled. Although the number of assessors usually included in a trained sensory panel does not make it possible to check this assumption, we do think that drawing confidence ellipsoïdes or convex hulls on a compromise and individual plot is a powerful descriptive tool. Here, as most often in data analysis, it is not a strict and true p-value which is required but some evidence that a pattern makes or does not make sense.

For comparing assessors two ways exist to derive an assessor map. The first one is obtained with the two first eigenvectors of the RV matrix of size (k,k). Each of these eigenvectors is normalized so that its sum of component squares is equal to the corresponding eigenvalue. The first axis of this map represents the direction of an assessor being equal to the compromise and the corresponding eigenvalue can be understood as the proportion of inter-individual variance explained by this compromise. The assessor coordinates on this first dimension are proportional to the STATIS weights a_i and are therefore positive. An assessor is represented on this plot as an arrow joining the assessor point to the origin. The angle between this arrow and the first axis is proportional to the disagreement of this assessor with the compromise. The length of this arrow is proportional to the quality of the representation of the assessor on this plot. In some applications, it can be necessary to produce the subsequent plots (1,3), (1,4) ... and, in such a case, the visual interpretation becomes difficult.

The second way to derive an assessor map is based on the first two eigenvectors of the RV matrix being previously and simultaneously centered in lines and in columns (that is subtract from each RV coefficient the means of the line and of the column from those it belongs to and add the grand mean of the RV matrix). Therefore, this assessor map is centered and is useful for showing whether different groups of assessors could exist on the basis of the sample differences. In some applications, more than two axes can be necessary for correctly describing the assessor structure. Obviously the RV matrix or the double centered RV matrix can also be taken as the input of any clustering algorithm.

To summarize this important section, it is worth recalling that :

- STATIS derives a compromise association matrix W which is a weighted mean of the individual association matrices W_i
- These weights are given by the first eigenvector of the RV matrix among assessors, which means that the weight of an assessor is proportional to its agreement with the panel
- The compromise sample plot is obtained thanks to a PCO of W

- Sensory interpretation is conducted through covariances or correlations between compromise components and individual attributes
- Assessor maps are derived from PCO of the RV matrix and from PCO of the doubly-centred RV matrix.

2.7 Testing panel homogeneity and significance of compromise

For testing panel homogeneity, Schlich (1993) compared the observed mean of the $k(k\text{-}1)/2$ RV coefficients to the 95 % quantile of the distribution of the mean RV coefficient when a permutation of the product labels is randomly and independently chosen for each assessor. In order to estimate this quantile, 100 sets of permutations were sampled independently. A faster solution for testing significance of STATIS compromise consists of computing the mean of the $k(k\text{-}1)/2$ normalized RV coefficients (obtained by formulas (22) to (26)) and checking whether it is roughly greater than 2. This faster solution relies on normality of the RV distribution under permutation. As soon as the number of products is larger or equal to 6, the number of possible permutations (n !) becomes very large and therefore the normality assumption, observed in practice, should hold.

This test of panel homogeneity is an average of homogeneity computed over pairs of assessors. This approach can be too demanding, because at the end the data is summarized by a compromise. Therefore, it seems sensible to test a weaker hypothesis, that is the individual agreement with the panel compromise. A normalized RV coefficient is computed between each assessor (W_i) and the panel compromise (W) and this paper proposes to average these k values to get a test for compromise significance.

As a summary, one should remember that the exact permutation test defined in section 2.5 makes it possible to test whether the assessors agree among themselves and whether they agree with the STATIS compromise defined in section 2.6. The strength of these tests is that they are exact and that they do not require any permutation to be actually performed.

2.8 Comparing two panel compromises about the same products

The similarity between two panel compromises can be evaluated by their normalized RV coefficient. If this coefficient is about 1 or less, then the panels disagree dramatically; if it is between 1 and 2, then the panels agree rather poorly; if it is greater than 2, then the panels agree and the interpretation should lead to the same conclusions about the sample differences. But the reader must be aware that the interpretation of these differences, in terms of the attributes, may not be equivalent from panel to panel, firstly because the attributes may not be the same among panels and among assessors within panels (FCP or individual selection of attributes), and, secondly because the scalar products in W are sums over attribute contributions making it possible to obtain equal sums composed of different attribute contributions.

This test can be based either on the whole compromise spaces or on the compromise subspaces spanned by the interpreted dimensions. The former is done directly by using the two compromise matrices in formulas (22) to (26), whereas the latter requires first to recompute new W matrices on the basis of the selected components. The advantage of the latter is to

provide insurance that no noise can destroy a significant agreement in the interpreted space, which is at the end the only information retained.

The idea to be retained from this section is that the exact permutation test makes it possible to test agreement between compromises drawn from two different panels having profiled the same samples. Finally, the technique could also be applied to check whether two competitive data analysis techniques lead to the same sample space or to the same sample plot.

2.9 Testing exchangeablity of assessors among panels

When several panels have profiled the same samples and when the test described in the previous section is significant, one could wish to go further by testing whether the permutation of assessors among panels would provide us with the same amount of panel differences, as a null hypothesis. If this test is significant, it means that discrimination holds and therefore, it informs the panel leaders that they cannot exchange their assessors.

Assume that g panels were available and that the l-th panel includes k_l assessors ($\sum_{l=1}^{g} K_l = k$). Having gathered the k assessors into a single data set, the RV matrix of size (k,k) among these assessors is computed. A PCO of the doubly centred RV matrix is performed in order to keep part or all of the assessor components. From this system of assessor coordinates it is now possible to apply a CDA of the panel factor. The panel discrimination can be measured thanks to a classical statistic in CDA :

$$H = \frac{\text{trace}(B)}{\text{trace}(T)} \tag{30}$$

in which B is the between-panel covariance matrix and T is the total covariance matrix. The closer to 1 H, the more different the panels are. The test proposed in Kazi-Aoual (1992) consists of permuting the assessor labels in the assessor coordinate table obtained from the PCO, without permuting the corresponding assessor coordinates and then computing a CDA of the new permutated partition of the k assessors into g panel of k_l assessors. If the null hypothesis of assessor exchangeability holds, one can expect the real H statistic to be not greater than the same statistic under permutation. Instead of performing numerous permutations in order to estimate the distribution of H under permutation, the following formulas from Kazi-Aoual (1992; 1993) give respectively the mathematical expectation and the variance of H under permutation :

$$\mathrm{E_P}[H] = \frac{g-1}{k-1} \tag{31}$$

$$\mathbf{V_P}[H] = \frac{2((k-1)/\beta - 1)(g-1)(k-g)(1+(k-3)\cdot c\cdot f/2k(k-1))}{(k+1)(k-1)^2(k-2)} \tag{32}$$

where β and c are defined as β_i and c_i in equations (14) and (24) replacing W_i by the association matrix among the assessors computed from the assessor scores obtained by CDA.

The constant f is given by :

$$f = \frac{(k-1)(k(k+1)\sum_l (1/k_l) - (k-1)(g-1)(g+1))}{(k-3)(g-1)(k-g)} \tag{33}$$

making it possible to derive a normalized H statistic :

$$H_N = \frac{H - \mathrm{E_P}[H]}{\sqrt{\mathrm{V_P}[H]}} \tag{34}$$

when this H_N is greater than 2, one can decide that the assessors are not exchangeable across panels.

To conclude this section, it is worth pointing out that, contrary to the previous permutation tests, this one is based on permutations of the assessors instead of the products. It appears as a competitor of the multivariate analysis of variance tests such as the Hotteling-Lawley trace or the Wilks ratio (Mardia *et al.*, 1979), but contrary to these parametric tests it does not require the assumptions of multinormality and homogeneity of within-panel covariance matrices.

2.10 Defining a compromise about attribute correlations by means of the Dual STATIS method

Dual STATIS, proposed in Lavit (1988), is the STATIS method applied to the covariance matrix $X'X/n$ (or to the correlation matrix) instead of the association matrix XX'. The aim of Dual STATIS is to compare the assessors on the basis of their individual structure of attribute covariances or correlations, which can be understood as their own way to understand attributes and to link them together. The STATIS compromise becomes a weighted mean of the individual covariance matrices. Such an analysis can be very interesting for the panel leader when training his panel to conventional profiling. Furthermore, this analysis can be run even when the assessors scored different samples but for the same attributes, making it possible to investigate simultaneously correlation structures on the basis of different kinds of products.

Unfortunately, the analytical permutation tests described in Kazi-Aoual (1993), do not seem to work with Dual STATIS, because of the non centering of the columns of X'. Therefore, as in GPA, one should run numerous permutations to derive a test. Note that in this context, the permutations are applied on the attributes instead of the products.

To conclude this section, it can be underlined that with a conventional profiling data set both STATIS and Dual STATIS can be performed, making it possible to compare the assessors on the basis of both the sample differences (client need) and the attribute relationships (panel leader need). If on both aspects the agreement among the assessors is good enough, the panel leader can trust that the sample differences perceived by the assessors (significant STATIS) were also described in the same way by these assessors in terms of the sensory attributes (significant Dual STATIS).

3. COMPARISON WITH OTHER METHODS

3.1 Principal Component Analysis (PCA), Simple and Multiple Correspondence Analysis (CA and MCA) and Canonical Discriminant Analysis (CDA)

The limitations of the classical PCA of the mean score products have already been mentioned in the introduction section of this chapter. There are two other ways of performing PCA on a profiling data set, called TUCKER1 in chapter 10 of this book. The first TUCKER1 method consists of a PCA of a data set composed of n times k observations and p attributes; this data set gathers the individual data sets vertically. It cannot be applied to free-choice profiling and the author thinks that it can be quite non robust to outlier observations. The second TUCKER1 method consists of a PCA of a data set composed of n observations and $\sum_{l=1}^{k} p_l$ attributes; this data set gathers the individual data sets horizontally. This technique is more interesting than the previous one, because it is a solution for analysing free-choice profiling when only a PCA program is available. Nevertheless, the weight of an assessor in this analysis can be strongly inflated or deflated according to his number of attributes. One can avoid this problem by dividing the column centered scores of a given assessor by the square root of the sum of squares of these scores and by doing a PCA of the covariance matrix.

The TUCKER2 and TUCKER3 methods, also described in this book (chapter 10), are more interesting but cannot be applied to free-choice profiling either.

The use of CA and MCA with profiling data, proposed by McEwan and Schlich (1991/2), considers sensory measurements at ordinal level instead of interval level, which is definitely not possible with STATIS or with any technique based on covariance or correlation computations. Whether this point improves significantly the interpretation of the sensory data is still not obvious. Another advantage of CA and MCA, linked to the previous one, is their ability to discover non linear relationship between attributes. Most of the linear techniques could accommodate non linear transformations of the data such as spline fonctions. But till now there has been a lack of published examples of these techniques in the sensory field. Finally and once again, the analysis of free-choice profiling with CA or MCA is not straightforward.

The introduction section of this chapter has suggested why CDA together with some pre-treatment of the data could be useful for analysing sensory profiles. Nevertheless and like the other methods evoked in this section, with the exception of the second TUCKER1 method, CDA cannot deal with free-choice profiling or with the problem of non attribute alignement in conventional profiling.

3.2 Generalised Procrustes Analysis (GPA)

The introduction section of this chapter has recalled that GPA is the historical leader technique for the analysis of sensory profiles because it was presented as the dedicated technique for free-choice profiling. The present book includes a chapter (7), which fully describes this technique. In the past, the author of the present chapter advocated the use of GPA through his SAS/IML® software for GPA (Schlich, 1989). But now, in the light of the validation concern of sensory profiles, the author prefers the STATIS framework for several reasons :

- It is a non iterative technique

- Its weighting system of assessors deals smoothly with outliers
- The β coefficient gives simple dimensionality estimation
- The analytical permutation tests of RV coefficient provide a straightforward panel homogeneity estimation, a compromise validation and a panel comparison
- Dual STATIS is a unique tool for investigating individual correlation structures

3.3 INDSCAL

INDSCAL (Carroll & Chang, 1970), described in this book in chapter 6, is the MDS (Schiffmann *et al.*, 1981) technique dedicated to three-way data analysis. It analyses a set of individual product dissimilarity matrices, in that respect it is a little more general than STATIS, which analyses a set of scalar product matrices. INDSCAL iteratively defines an a priori fixed number of optimal dimensions mapping the products. The strength of INDSCAL, not available in STATIS, is that each assessor can weight differently each of these dimensions. These individual vectors of weights are defined in order to minimize the STRESS which is a least square criterion between the observed individual dissimilarities and the compromise product distances in the fitted space.

Being a more general model than STATIS, INDSCAL can cope with free-choice profiling as soon as a dissimilarity function is chosen to be derived from the attributes scores. Unfortunately, this technique does not seem to be widely used in our field. Furthermore, to the knowledge of the author no analytical validation technique exists in the INDSCAL framework, certainly because of its iterative algorithm.

4. APPLICATIONS

4.1 The ESN coffee experiment

The European Sensory Network (ESN) was launched in 1989 as a basis for close collaboration between major food research centers in Europe. Nowadays, ESN gathers about 20 academic sensory scientists coming from about 10 different countries in Europe. The aim of ESN is to exchange ideas and to transfer results from basic research to the food industry. In order to check consistency of sensory profiling, ESN organized an interlaboratory study of 16 coffees evaluated by 11 different panels managed by ESN members in their own institutes.

The samples, the beverage preparation and service and the experimental design were absolutely identical across panels. For practical reasons, the same samples had to be assessed during a session by every assessor and no more than 4 products could be presented during a given session. Therefore, 4 sessions were conducted to complete a replicate. Three replicates were done. The allocation of the samples to the sessions was identical across panels and determined thanks to previous knowledge about the expected coffee differences. The strategy was to span the coffee space as much as possible within each of the 4 sessions of the first replicate. Second and third replicates were then defined according to the same rule and in order to respect a pair balance condition, which was to ask that any pair of samples must not be present in more than one session. Concerning the selection and the training of the assessors, as for the vocabulary development, no instruction was given to the panel leaders. Depending on

the panels, the number of assessors was between 8 and 12, whereas the number of attributes was between 14 and 56. As said in the introduction, a book will be published soon by ESN presenting the complete analysis of the coffee data and making it available to the reader. For the present chapter, only a small part of the profiling data is used to illustrate the application of the above proposed techniques.

The 16 coffee samples were provided by the International Coffee Organization (ICO) in London, who was an ESN member. The sensory panel from ICO was composed of 12 subjects highly trained to profile coffee for many years. ICO panel can be considered as an expert panel. On the contrary, one of the French panel, called F2, was poorly trained to profile coffee. Because of this opposition, we decided to present a STATIS analysis of the ICO data (section 4.2) and another one of the F2 data (section 4.3). For the sake of simplicity, we also decided to analyse only a subset of attributes (12 for ICO and 9 for F2), chosen from attributes being scored by most of the panels. Presenting the analysis of the "best" and of the "worst" panels, we aim to convince the reader that our statistical framework is actually able to detect such a diagnostic.

In order to illustrate the techniques of panel comparisons together with the Dual STATIS method, we defined a panel called S49 including 49 assessors from 5 differents panels, called 1F for France (different panel than F2), IC for ICO (UK), No for Norway, Po for Poland and Sw for Sweden. These panels share the property of including 4 basic attributes for describing coffee : bitterness, acidity, astringency and body/mouthfeel. Therefore, we applied STATIS (section 4.4) and Dual STATIS (section 4.5) on the S49 panel data restricted to the above 4 attributes.

Each of the three data sets analysed were first averaged over the 3 replicates by product times assessor, making the number of observations being equal to the number of products (16) times the number of assessors (12, 9 or 49).

4.2 STATIS of the ICO panel

Table 1
ICO panel. RV coefficients between subjects

	A	B	C	D	E	F	G	H	I	J	K	L
A	1.00											
B	0.81	1.00										
C	0.76	0.83	1.00									
D	0.82	0.87	0.77	1.00								
E	0.91	0.83	0.84	0.81	1.00							
F	0.86	0.91	0.84	0.87	0.86	1.00						
G	0.86	0.76	0.72	0.77	0.79	0.81	1.00					
H	0.91	0.77	0.77	0.84	0.87	0.87	0.87	1.00				
I	0.88	0.82	0.88	0.81	0.91	0.87	0.80	0.88	1.00			
J	0.78	0.69	0.70	0.76	0.73	0.78	0.82	0.88	0.78	1.00		
K	0.76	0.81	0.86	0.75	0.84	0.83	0.70	0.77	0.84	0.70	1.00	
L	0.86	0.76	0.70	0.75	0.74	0.80	0.87	0.87	0.80	0.78	0.73	1.00

Table 2
ICO panel. Normalized RV coefficients between subjects

	A	B	C	D	E	F	G	H	I	J	K	L
B	8.56											
C	7.84	8.76										
D	8.60	9.38	7.93									
E	9.69	8.73	8.85	8.60								
F	9.10	9.75	8.73	9.24	9.10							
G	9.23	8.04	7.34	7.93	8.41	8.48						
H	9.77	8.09	8.02	8.87	9.35	9.27	9.29					
I	9.40	8.54	9.29	8.55	9.71	9.18	8.55	9.49				
J	8.18	7.15	7.15	7.88	7.63	8.11	8.66	9.41	8.24			
K	7.90	8.55	9.05	7.60	8.93	8.65	7.06	7.99	8.94	7.09		
L	9.26	8.08	7.30	7.86	7.87	8.50	9.18	9.28	8.63	8.20	7.63	
Mean	8.87	8.51	8.21	8.40	8.81	8.92	8.38	8.98	8.96	7.97	8.13	8.34

As we recommend most often, no attribute standardization was performed, but the assessors were compared on the basis of their RV coefficients (Table 1). These RV coefficients range from 0.69 to 0.91 denoting a very good agreement between panelists on the sample differences. The permutation test statistics of these coefficients, which are the normalized RV coefficients given in Table 2, range from 7.15 to 9.77. According to the normal distribution, we can therefore conclude that each pair of assessors strongly agrees about the sample structure. Such a high homogeneity in a trained sensory panel is rather unusual in practice.

Table 3
ICO panel. PCO of the non centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	9.9181	82.65	82.65
DIM 2	0.6041	5.03	87.69
DIM 3	0.3237	2.70	90.38
DIM 4	0.2682	2.23	92.62
DIM 5	0.2418	2.02	94.63
DIM 6	0.1560	1.30	95.93
DIM 7	0.1294	1.08	97.01
DIM 8	0.1163	0.97	97.98
DIM 9	0.0879	0.73	98.71
DIM10	0.0685	0.57	99.28
DIM11	0.0517	0.43	99.72
DIM12	0.0337	0.28	100.00

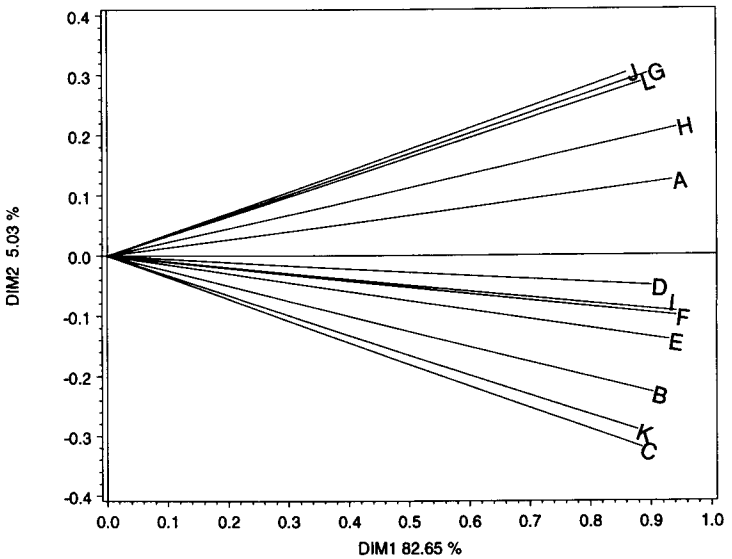


Figure 1 : ICO panel. PCO of the non centered RV matrix

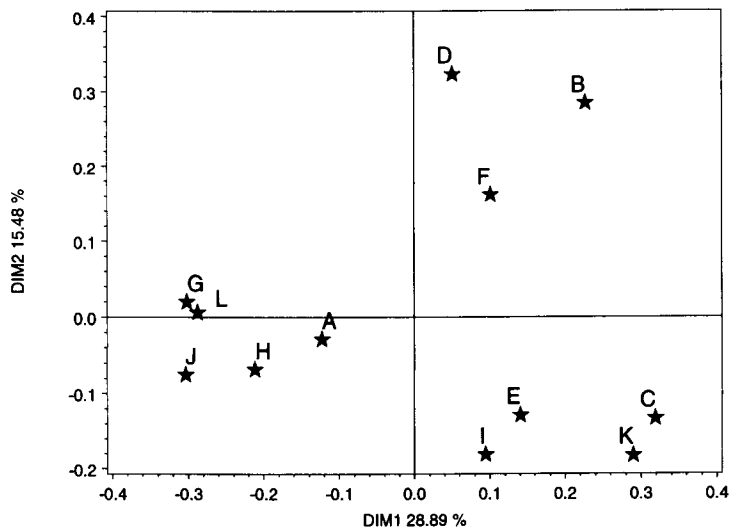


Figure 2 : ICO panel. PCO of the centered RV matrix

The first eigenvector of the RV matrix accounts for 82.65 % of the panel variation (Table 3). This can be understood as follows : the compromise association matrix W defined by STATIS will account for 82.65 % of the variation between the 12 individual association matrices W_i. Figure 1 is the first assessor plot drawn from this PCO. One can see that the assessors who might agree the least with the compromise (first axis) are assessors J, G and L at the top of the plot and assessors K and C at the bottom of the plot. One can check this point by reading the last line of Table 2, in which these assessors actually get the smallest, but still excellent, average normalized RV coefficients.

Table 4
ICO panel. PCO of the centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	0.6042	28.89	28.89
DIM 2	0.3237	15.48	44.38
DIM 3	0.2715	12.98	57.36
DIM 4	0.2445	11.69	69.05
DIM 5	0.1560	7.46	76.51
DIM 6	0.1296	6.20	82.71
DIM 7	0.1172	5.61	88.31
DIM 8	0.0894	4.28	92.59
DIM 9	0.0688	3.29	95.88
DIM10	0.0517	2.47	98.35
DIM11	0.0345	1.65	100.00

Thanks to the PCO of the centered RV matrix, one can draw a centered assessor map (Figure 2), in which the compromise would be located at the origin. On this map, one can observe three groups of assessors : D, B and F at the top of the plot, C, K, E and I at the bottom of the plot and the remaining assessors on the left of the plot. This grouping of assessors might correspond to very slight differences in sample perception, as permutation tests demonstrated strong homogeneity in each pair of assessors. Moreover, as this map accounts for only 44.38 % of the total variation (Table 4), this visual grouping might also be artificial. Comparing Tables 3 and 4, one can note that the total number of dimensions in the first analysis was equal to the number of assessors, that is 12, whereas in the second analysis centering the RV matrix removed one dimension.

Because of the high homogeneity of the ICO panel, the assessor maps given by Figures 1 and 2 were not really necessary. The point is that without looking at the magnitude of the RV coefficients and without testing them thanks to the normalized RV coefficients, it would have been difficult to assess on the basis of any assessor map whether homogeneity held or not.

Table 5 is the most interesting table printed by our STATIS program. The lines of this table are the assessors. The column called 'p' gives the number of attributes. In this case, standard deviation of one or two attributes were nil for subjects A, D, G and L. The 'Scaling' column gives the coefficient by which each individual data set should be multiplied to achieve a common and equal global dispersion over both samples and attributes. This column demonstrates that subjects B and I had a clear tendency to concentrate their scores on a small

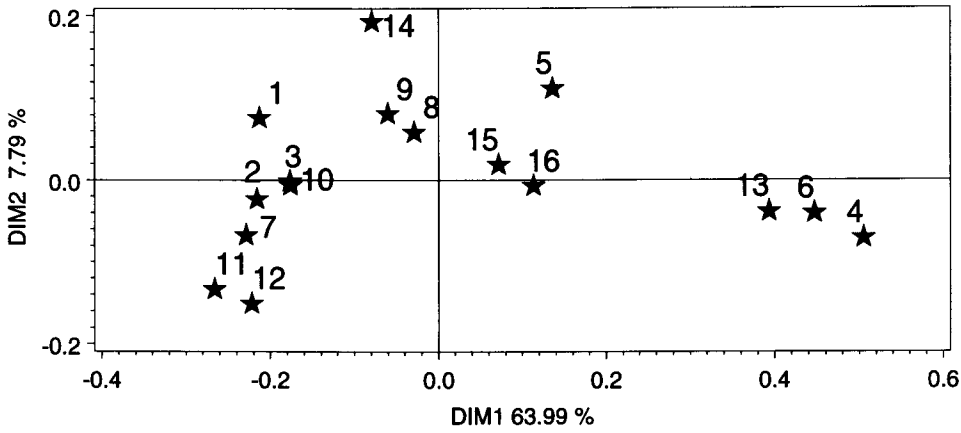


Figure 3 : ICO panel. PCO of the sample compromise among subjects

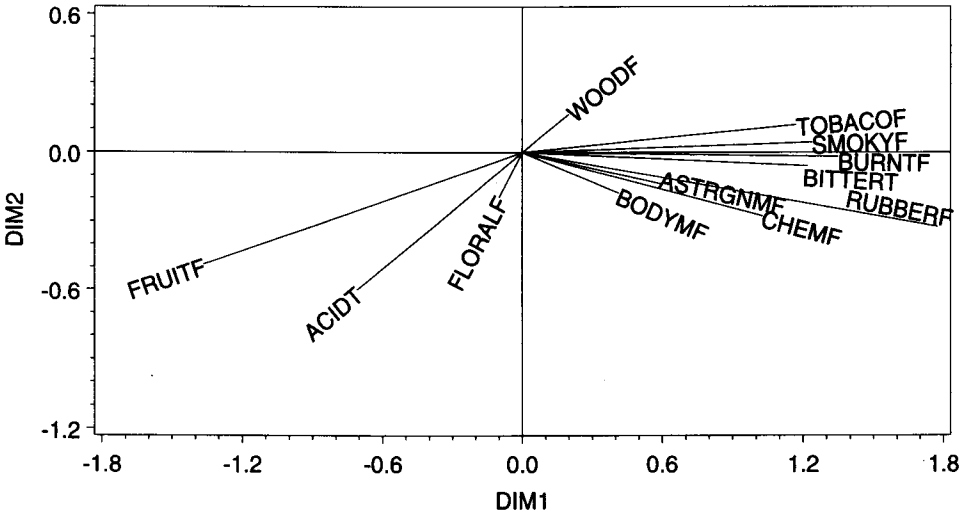


Figure 4 : ICO panel. Covariance plot of attributes

part of the scale (scaling coefficients equal to 2.131 and 2.190), whereas subjects D and G clearly behaved in the opposite way (scaling coefficients equal to 0.325 and 0.293). The next 'NRV' column is the 'Mean' line of Table 2, that is the mean of the normalized RV coefficients between the current subject and the others. Looking at this column, one can immediately detect panelists who would disagree with the rest of the panel. The 'Weight' column in Table 5 contains the STATIS weights of the assessors being normalized in such a way that their sum is equal to the number of panelists. Therefore, an individual weight greater than one corresponds to a subject who is in better agreement with the panel than the others. Because of the high level of homogeneity, the weight range is very narrow around one, denoting that there is no reason to trust some panelists more than others. The next 'BETA' column gives the dimensionality coefficient defined in section 2.3 of this chapter. This coefficient ranges from 1.467 to 2.786 with a mean value of 1.979. On this basis, one can postulate that two dimensions would be enough to span the coffee space as it is perceived by this panel. Therefore, only the two first dimensions of the compromise sample space will be interpreted below. The last column of Table 5, called 'NRVC', gives the normalized RV coefficient between each assessor and the STATIS compromise. We suggested in section 2.7 that the mean of this column can be accepted as a normal test of the null hypothesis of chance against the alternative hypothesis of compromise significance. In that respect, the ICO STATIS compromise is found highly significant : 9.752. This last column provides a check of individual agreement with compromise. In the ICO panel, NRVC is always much greater than 2, denoting that every assessor agrees with the compromise.

Table 5
ICO panel. Summary of individual STATIS statistics

Subject	p	Scaling	NRV	Weight	BETA	NRVC
A	10	0.674	8.865	1.030	1.734	10.100
B	12	2.131	8.512	0.995	1.539	9.698
C	12	1.229	8.206	0.975	2.321	9.393
D	11	0.325	8.404	0.992	2.786	9.618
E	12	1.434	8.808	1.024	1.627	10.020
F	12	0.799	8.918	1.037	2.407	10.140
G	11	0.293	8.378	0.985	2.424	9.607
H	12	0.528	8.984	1.038	1.732	10.230
I	12	2.190	8.956	1.037	1.600	10.150
J	12	0.996	7.973	0.947	1.708	9.171
K	12	0.907	8.127	0.967	2.399	9.325
L	10	0.493	8.344	0.973	1.467	9.575
Mean	11.5	1.000	8.540	1.000	1.979	9.752

The product map in Figure 3 is obtained from a PCO of the compromise association matrix, whose eigenvalue decomposition is given in Table 6. This plot is dominated by the first dimension which accounts for 63.99 % of the information. This first axis splits samples 4, 6 and 13 from the others and the second axis seems to distinguish sample 14 at the top of the plot from samples 11 and 12 at the bottom of the plot. The interpretation of this sample structure is conducted by means of an average covariance plot (Figure 4) as explained in

section 2.6. It is clear from this plot that the first axis is a gradient of coffee strength positively correlated to the bitter taste and to the tobacco, smoky, burnt and rubber flavours and negatively correlated with the acid taste and the fruity flavour. Therefore, samples 4, 6 and 13 were judged as the "strongest" coffees, whereas sample 11 and 12 were perceived as more acid and fruity and finally coffee 14 would be the weakest sample for both bitterness and acidity. The ESN book will explain why these findings make sense considering the origin of the samples.

Table 6
ICO panel. PCO of the sample compromise among subjects

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	0.9833	63.99	63.99
DIM 2	0.1197	7.79	71.78
DIM 3	0.0690	4.49	76.27
DIM 4	0.0620	4.04	80.31
DIM 5	0.0509	3.32	83.62
DIM 6	0.0473	3.08	86.70
DIM 7	0.0390	2.54	89.24
DIM 8	0.0338	2.20	91.43
DIM 9	0.0287	1.87	93.30
DIM10	0.0233	1.51	94.82
DIM11	0.0216	1.41	96.23
DIM12	0.0189	1.23	97.45
DIM13	0.0170	1.10	98.56
DIM14	0.0129	0.84	99.40
DIM15	0.0093	0.60	100.00

As suggested in section 2.6, Figure 5 locates on the previous compromise sample map (C followed by the sample numbers) the individual assessments of each coffee (sample numbers) and for one half of them (from left to right samples 12, 1, 14, 8, 16, 5, 13 and 4) draws the corresponding convex hulls. Drawing all the 16 convex hulls would have made the plot unreadable. This picture suggests that the following seven groups of coffees : (11, 12), (7, 2, 10, 3, 1), (14), (8, 9), (15, 16), (5), (13, 6, 4) may be different between groups and similar within groups.

4.3 STATIS of the F2 panel

The RV coefficients (Table 7) range from 0.13 to 0.63, showing that the best RV coefficient in F2 panel is lower than the worst in ICO panel. The normalized RV coefficients in Table 8 are often not significant (lower than 2) and quite often strongly not significant (lower than 1). From the 'Mean' line of this table, it can be concluded that only assessors B, D, E and H seem to agree with the panel.

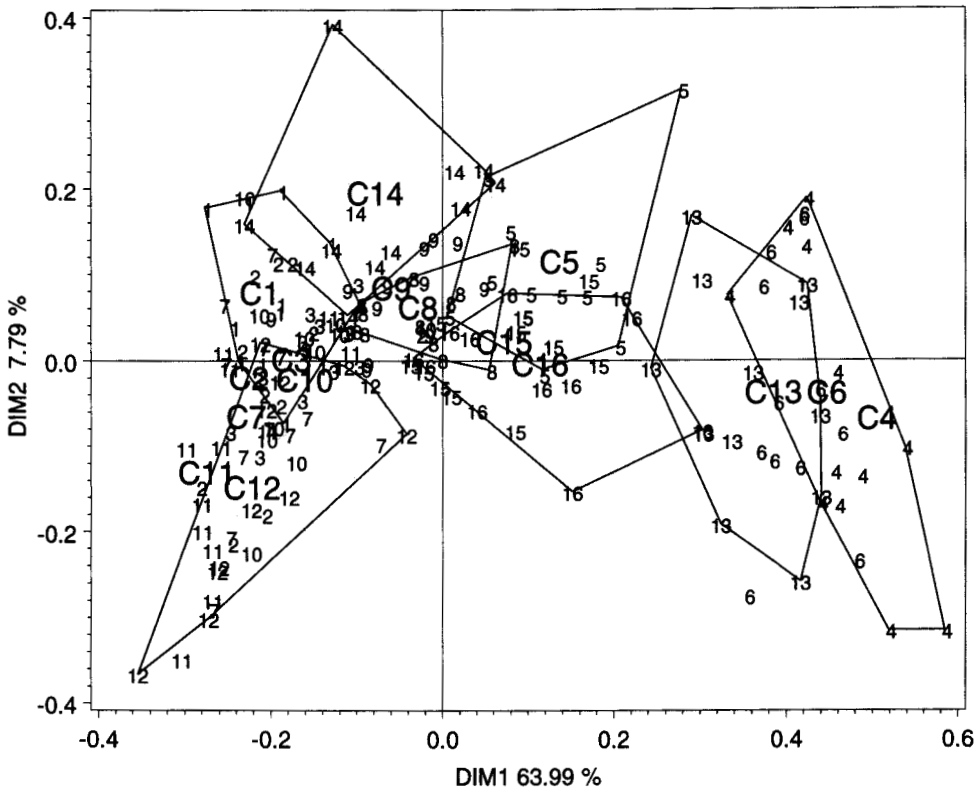


Figure 5 : ICO panel. Convex hulls of samples on the compromise plot (1, 2)

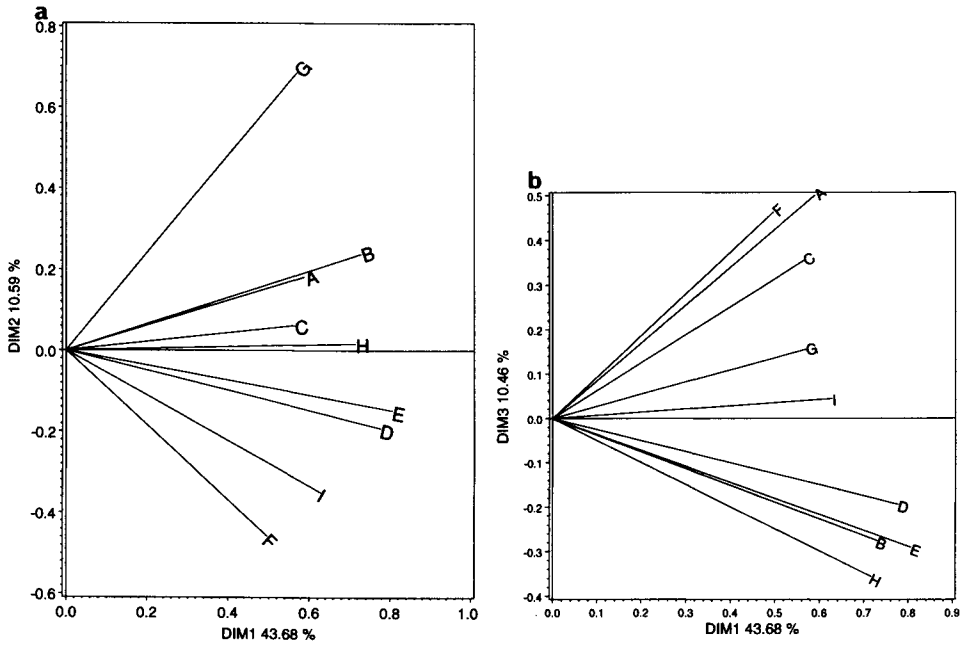


Figure 6 : F2 panel. PCO of the non centered RV matrix
a. Plot (1, 2) b. Plot (1, 3)

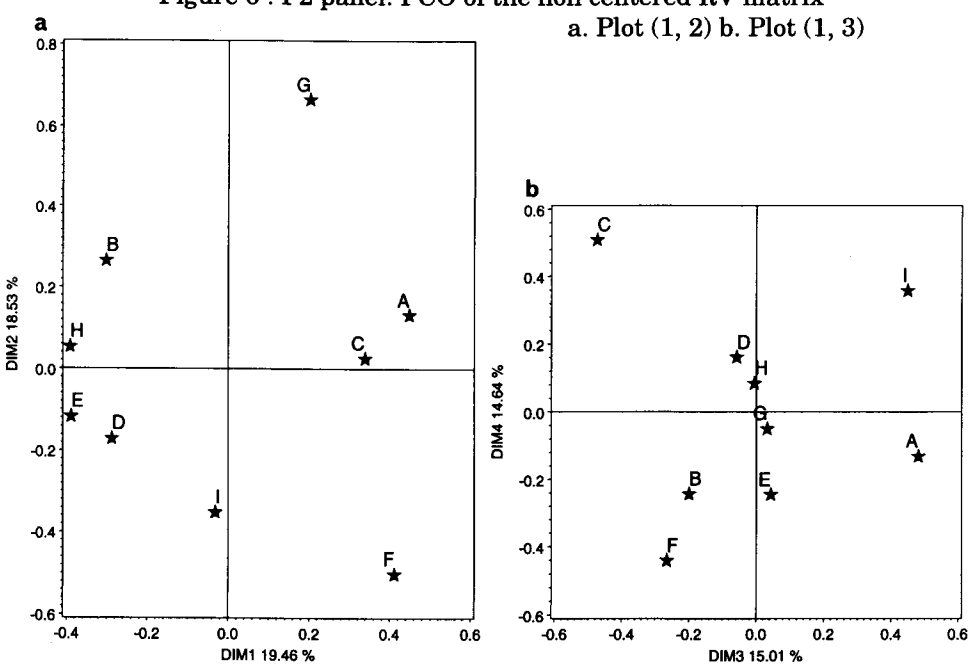


Figure 7 : F2 panel. PCO of the centered RV matrix
a. Plot (1, 2) b. Plot (3, 4)

Table 7
F2 panel. RV coefficients between subjects

	A	B	C	D	E	F	G	H	I
A	1.00								
B	0.29	1.00							
C	0.28	0.33	1.00						
D	0.29	0.47	0.40	1.00					
E	0.38	0.63	0.26	0.63	1.00				
F	0.30	0.25	0.29	0.32	0.37	1.00			
G	0.41	0.45	0.31	0.33	0.31	0.13	1.00		
H	0.29	0.48	0.32	0.52	0.58	0.20	0.33	1.00	
I	0.37	0.29	0.29	0.49	0.44	0.25	0.20	0.38	1.00

Table 8
F2 panel. Normalized RV coefficients between subjects

	A	B	C	D	E	F	G	H	I
B	1.48								
C	0.88	2.16							
D	1.17	4.29	2.94						
E	2.71	6.11	1.38	6.43					
F	1.18	1.24	1.28	1.86	2.85				
G	2.58	3.84	1.46	1.81	1.99	-0.91			
H	0.98	4.19	1.68	4.46	5.51	0.16	1.72		
I	1.66	1.38	0.88	3.91	3.62	0.43	-0.54	2.17	
Mean	1.58	3.09	1.58	3.36	3.82	1.01	1.49	2.61	1.69

Table 9
F2 panel. PCO of the non centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	3.9313	43.68	43.68
DIM 2	0.9528	10.59	54.27
DIM 3	0.9413	10.46	64.73
DIM 4	0.7711	8.57	73.30
DIM 5	0.7525	8.36	81.66
DIM 6	0.5152	5.72	87.38
DIM 7	0.4617	5.13	92.51
DIM 8	0.4167	4.63	97.14
DIM 9	0.2574	2.86	100.00

As a consequence of this poor homogeneity, the first axis of the PCO of the RV matrix accounts for only 43.68 % (Table 9) of the variation among the assessors, whereas we obtained 82.65 % for the ICO panel. Because the third axis accounts for about the same amount of variation as the second one, Figure 6 presents both plots (1,2) and (1,3) of this PCO. But as these three axes account for only 64.73 % of the information (smaller arrow lengths in Figure 6 compared to Figure 1) one must interpret Figure 6 cautiously. Nevertheless,

this figure underlines assessor F as being the worst because of the large angles its arrow makes with the first axis in the two plots. This is confirmed by the lowest mean normalized RV coefficient of 1.01 (Table 8) obtained by subject F. The best assessors B, D, E and H are almost the closest to the first axis in the plot (1,2) and are gathered in the plot (1,3), suggesting that they may have directed the compromise.

Table 10
F2 panel. PCO of the centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	1.0003	19.46	19.46
DIM 2	0.9527	18.53	37.99
DIM 3	0.7718	15.01	53.00
DIM 4	0.7525	14.64	67.63
DIM 5	0.5152	10.02	77.66
DIM 6	0.4658	9.06	86.71
DIM 7	0.4256	8.28	94.99
DIM 8	0.2574	5.01	100.00

This last point is clarified by the PCO of the centered RV matrix, whose distribution of eigenvalues (Table 10) suggested that four axes should be interpreted to take into account 67.63 % of the total information. These panelists B, D, E and H are clustered on the left part of plot (1,2) in Figure 7, whereas the others do not gather, and they are close to the origin on plot (3,4), whereas the others are farther to the origin. In fact, this panel is composed of four assessors (B, D, E and H) who fairly agree among themselves and of five other assessors who disagree with the group of four and among themselves. Therefore, the compromise cannot be anything else than a rough mean of assessors B, D, E and H. Interestingly this point could have been detected in Table 8, where the 6 normalized RV coefficients among these four subjects are the only ones to be above 4.

The 'Scaling' column in table 11 points out assessor G as having spanned his scores much more than the others. The 'NRV' column of Table 11 has already been interpreted above, but note that the mean of all the normalized RV coefficients is equal to 2.248, which can be considered as just significant, although it is only one fourth of the corresponding statistic in the ICO panel (8.540). The STATIS weights range from 0.760 to 1.235, but the mean weight of assessors B, D, E and H is 1.160, whereas the mean weight of the others is 0.872, meaning that in average STATIS gave to an assessor from the group (B,D,E,H) a weight being 33 % greater than the weight given to the other assessors. Interestingly, the 'BETA' column of Table 11 has a mean value of 2.876, whereas it was only equal to 1.979 for the ICO panel. Therefore, the French assessors seem to be more complex than the British assessors.... The French would perceive 3 dimensions whereas the British would perceive only 2. The reader, who may be British, should first noticed that the author is French... and secondly should not take these two last findings as the absolute truth, but rather as a useful indication of a pattern present in this data. Again more interesting is the fact that the mean b coefficient is equal to 2.306 within group (B,D,E,H), whereas it is equal to 3.331 within the other subjects, meaning that the more homogeneous assessors are at the same time the least complex subjects. The author found this kind of relation in many datasets; it is quite logical : the more dimensional,

the more chance to disagree. In our application, it is not difficult to agree about the coffee strength (first dimension common to everyone), it is much more difficult to agree about the coffee flavour (second and third dimensions not identically perceived by everyone). Finally, the 'NRVC' column in Table 11 shows that assessors from group (B,D,E,H) strongly agree with the STATIS compromise (mean normalized RV with compromise equals to 7.815) and the remaining assessors fairly agree with this compromise (mean normalized RV with compromise equals to 4.329). This is the "STATIS miracle" : although in this data set some neat disagreements held between some assessors, at the end everybody significantly agreed with the compromise. The F2 panel was definitely less homogeneous than the ICO panel, but STATIS was still able to define a valid compromise. At this point of the discussion, the reader may think that the null hypothesis in these permutation tests is too weak, making these tests artificially powerful. The author agrees that a null hypothesis which is actually true would be a disaster for the panel leader. In such a case, the only thing to do is to bin the data. In some situations, which have to remain anonymous, the author actually observed mean normalized RV coefficients between assessors around 0 and between assessors and compromise around 1.

Table 11
F2 panel. Summary of individual STATIS statistics

SUBJECT	p	Scaling	NRV	Weight	BETA	NRVC
A	7	0.700	1.580	0.902	3.758	4.383
B	8	0.910	3.086	1.117	1.963	7.527
C	8	1.330	1.581	0.865	2.943	4.446
D	8	1.361	3.359	1.191	2.580	8.061
E	7	1.545	3.824	1.235	1.752	8.825
F	8	0.709	1.010	0.760	2.860	3.421
G	8	0.305	1.495	0.872	3.137	4.441
H	8	1.231	2.608	1.096	2.930	6.846
I	8	0.909	1.687	0.962	3.958	4.955
Mean	7.77	1.000	2.248	1.000	2.876	5.878

From the distribution of the eigenvalues of the compromise (Table 12), it was decided to analyse the first four dimensions of the compromise sample space (Figures 8 and 9). The overall structure of plot (1,2) (Figure 8) is roughly the same as ICO. Nevertheless, some differences are noticeable. This structure is more complex as exemplified by the fact that the first dimension accounts for about half of the total information than for ICO. Sample 4 is now slightly split from samples 6 and 13, which are now gathered with sample 16. Looking at the convex hulls (Figure 10), it appeared that two assessors had sent this sample 4 very far from the others. The larger surfaces of these hulls, compared to ICO, lead to a less confidence about the following visual grouping of samples : (2,11), (1,3,7,9,10,12), (5,8,15), (14), (6,13,16), (4). The covariance plot (Figure 12) confirms the interpretation made with the ICO data, namely the opposition of bitterness and acidity along the first axis and the weakness of both aspects at the top of the second axis opposed to the bottom correlated to the astringency of the coffee.

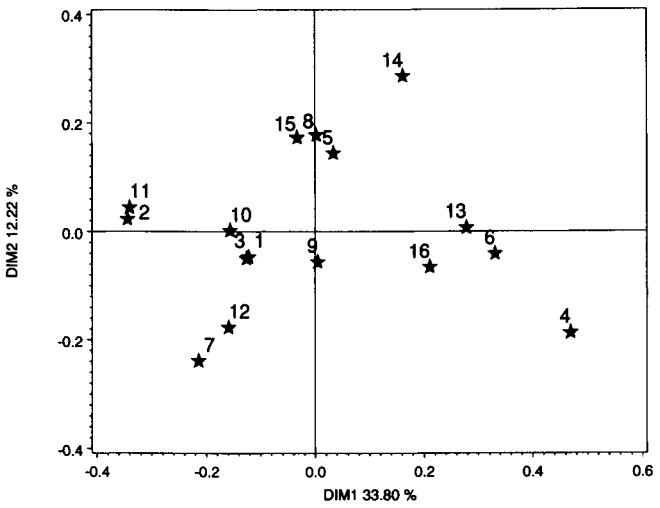


Figure 8 : F2 panel. PCO of the sample compromise among subjects. Plot (1, 2)

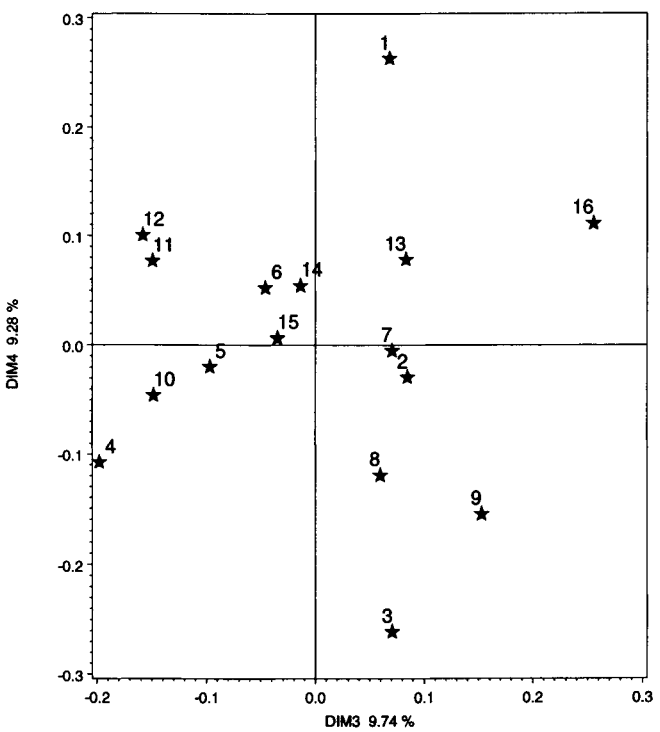


Figure 9 : F2 panel. PCO of the sample compromise among subjects. Plot (3, 4)

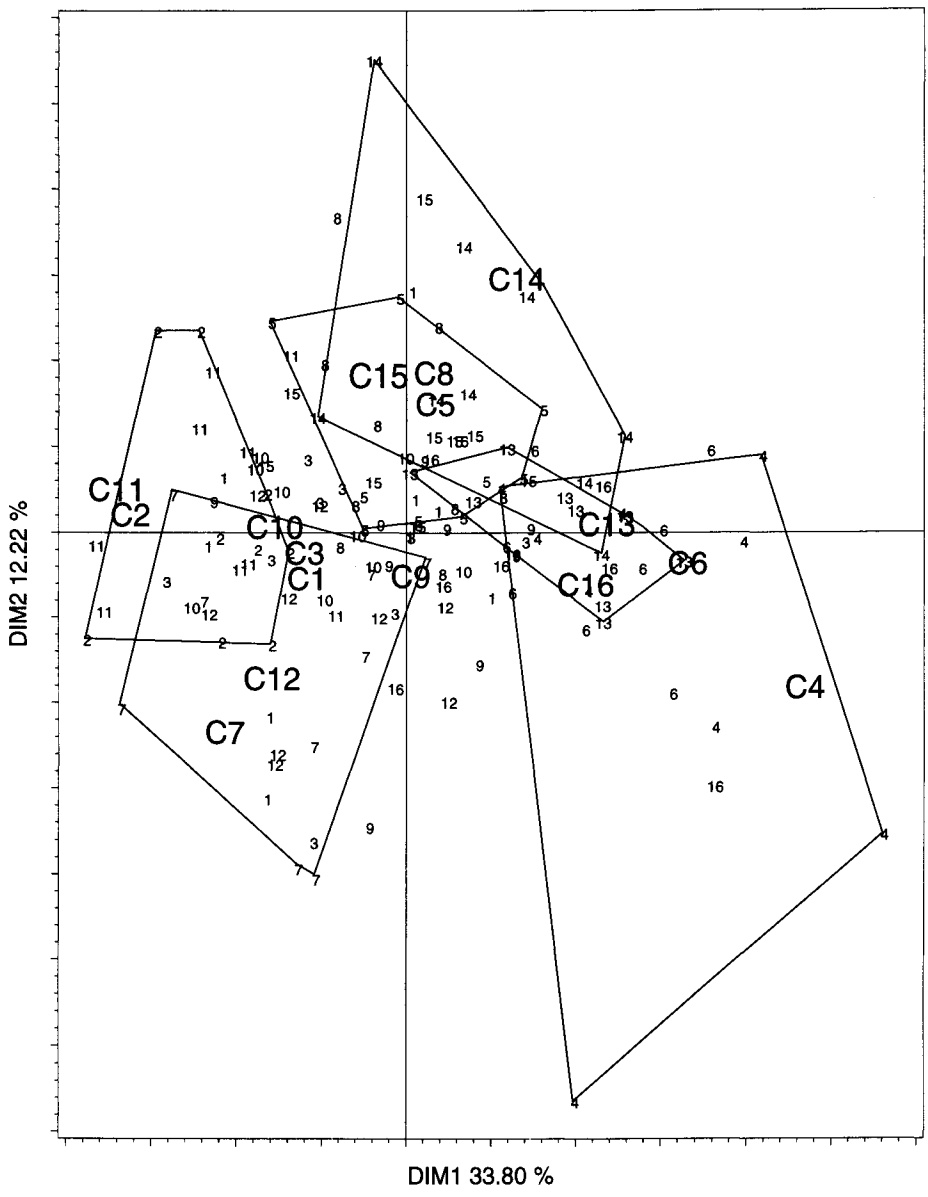


Figure 10 : F2 panel. Convex hulls of samples on the compromise plot (1, 2)

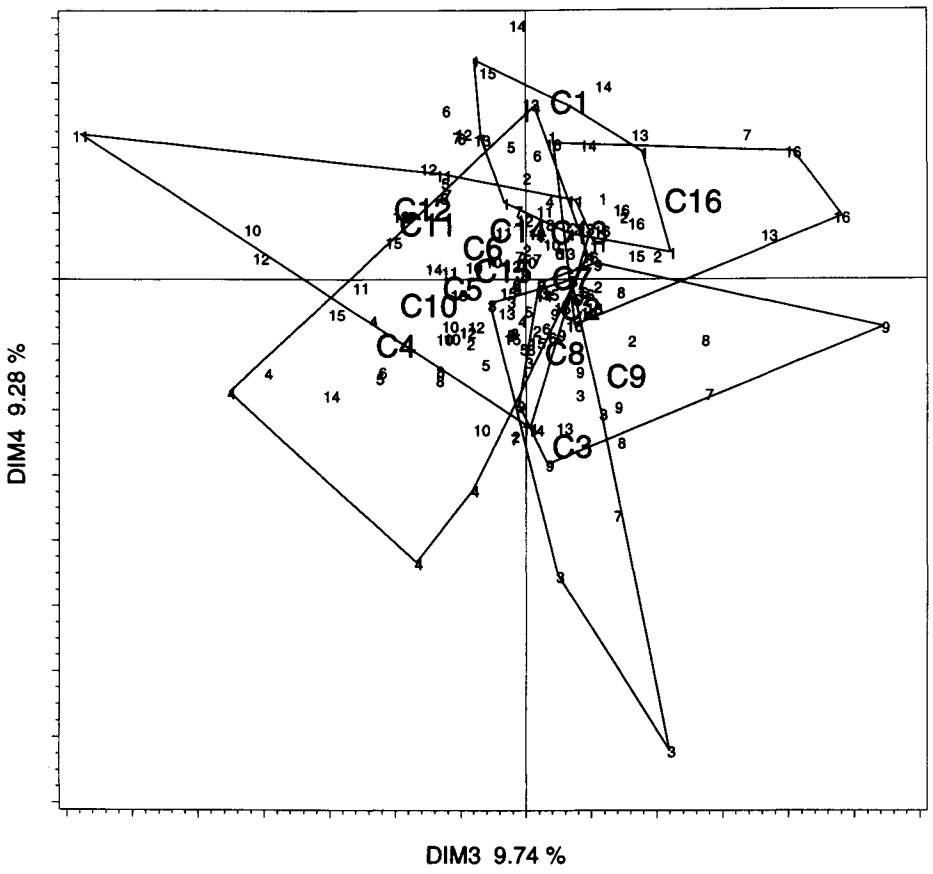


Figure 11 : F2 panel. Convex hulls of samples on the compromise plot (3, 4)

Sample and covariance plot (3,4) (Figures 9 and 13) suggest that sample 1 would have a more burnt flavour, as sample 16 which would also be more astringent and that samples 11 and 12 would have a more chemical flavour. But looking at the corresponding convex hulls in Figure 11, it is clear that the last finding about samples 11 and 12 may not be shared by most of the panel and that the previous conclusion about samples 1 and 16 is not universal.

Table 12
F2 panel. PCO of the sample compromise among subjects

Axis	Eigenvalue	Variance (%)	Cum. Var. (%)
DIM 1	0.8402	33.80	33.80
DIM 2	0.3038	12.22	46.02
DIM 3	0.2421	9.74	55.76
DIM 4	0.2306	9.28	65.04
DIM 5	0.1817	7.31	72.35
DIM 6	0.1239	4.98	77.33
DIM 7	0.0972	3.91	81.24
DIM 8	0.0934	3.76	85.00
DIM 9	0.0858	3.45	88.45
DIM10	0.0767	3.09	91.54
DIM11	0.0672	2.70	94.24
DIM12	0.0495	1.99	96.23
DIM13	0.0407	1.64	97.87
DIM14	0.0340	1.37	99.24
DIM15	0.0190	0.76	100.00

Figure 14 is a biplot from a covariance PCA of the mean score products. It is clear that this classical analysis leads to almost the same structure and to almost the same sample plot interpretation as in STATIS. Therefore, why should STATIS be used ? Firstly because of the permutation tests provided by STATIS, secondly because of the information about assessor similarity provided by STATIS and thirdly because with less homogeneous panel than F2 the output of STATIS can be different from that obtained by means of a PCA of the mean score products. The biplot of assessor E data (Figure 15) is again very similar to the STATIS compromise illustrating why this subject had the largest normalized RV coefficient with the compromise. On the contrary, biplots from assessors G and F (Figures 16 and 17), who were the subjects who disagreed the most with the compromise, clearly show numerous discrepancies with the compromise plot.

4.4 STATIS of the S49 panel

This analysis gathers 49 assessors coming from 5 different panels having scored the same four attributes : bitterness, acidity, astringency and body/mouthfeel. The discussion will focus on panel comparisons (see sections 2.8 and 2.9 of this chapter). The mean RV coefficients within

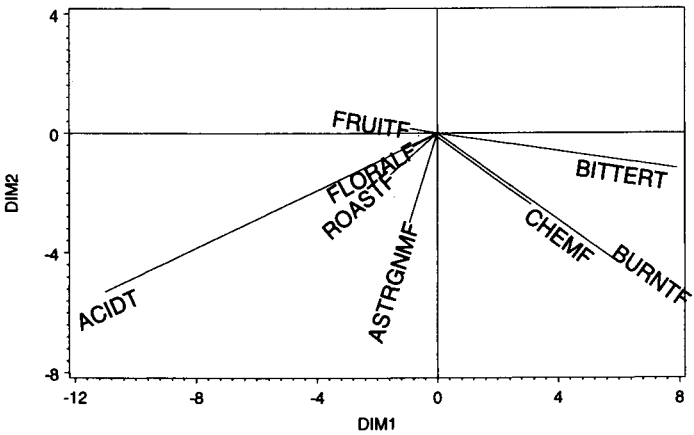


Figure 12 : F2 panel. Covariance plot (1, 2) of attributes

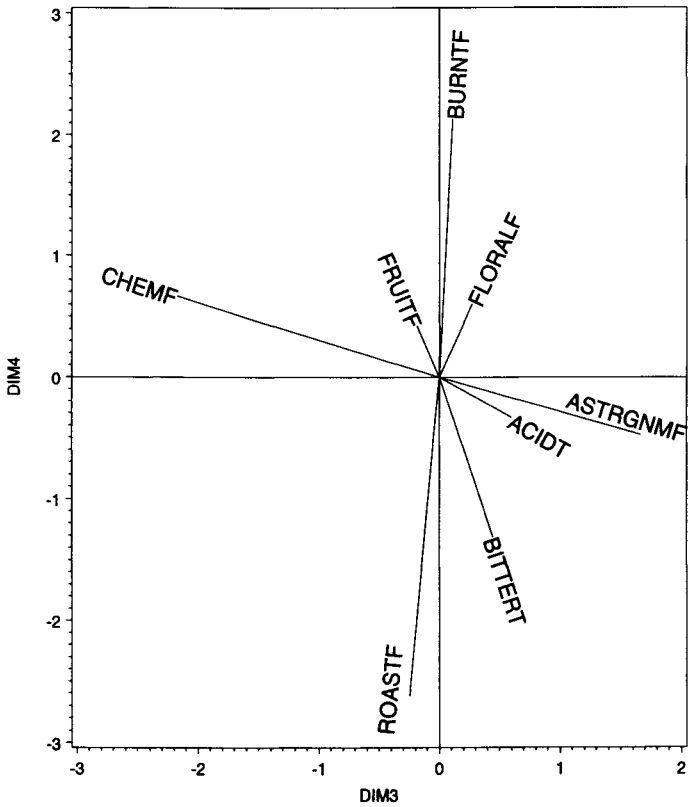


Figure 13 : F2 panel. Covariance plot (3, 4) of attributes

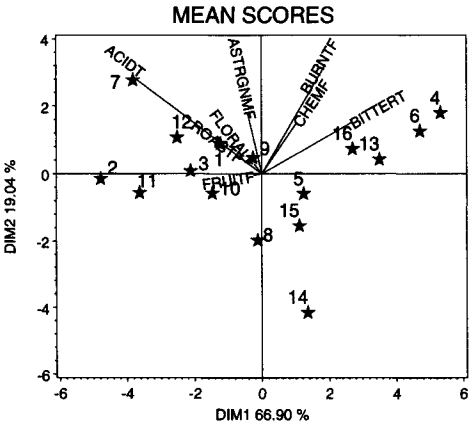


Figure 14 : F2 panel. Covariance PCA of mean score products

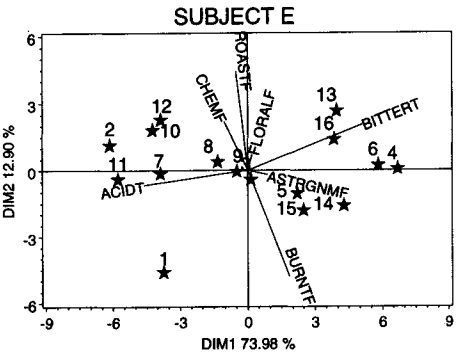


Figure 15 : F2 panel. Covariance PCA of scores from subject E

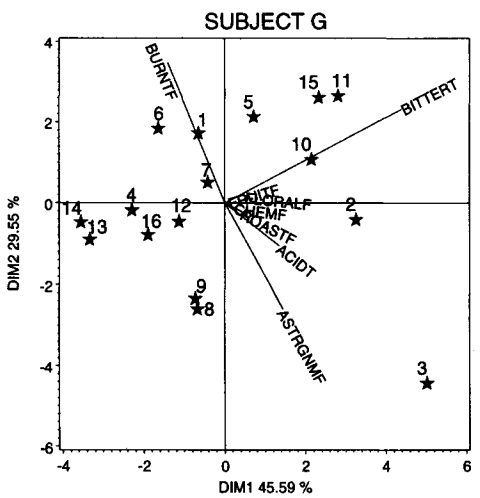


Figure 16 : F2 panel. Covariance PCA of scores from subject G

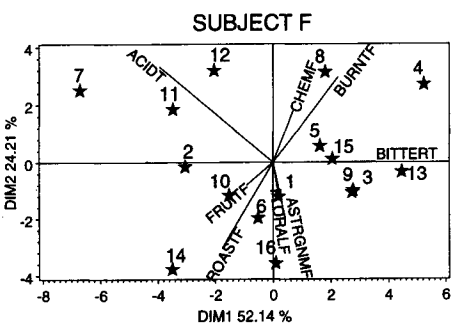


Figure 17 : F2 panel. Covariance PCA of scores from subject F

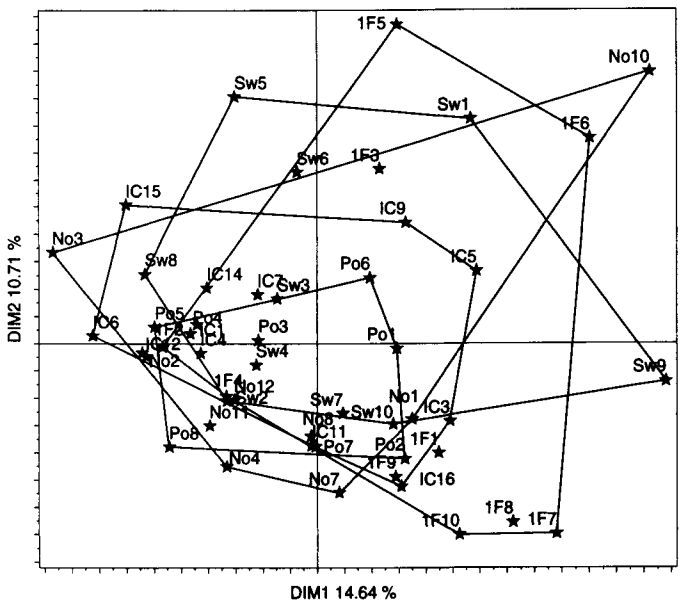


Figure 18 : S49 panel. PCO of the centered RV matrix

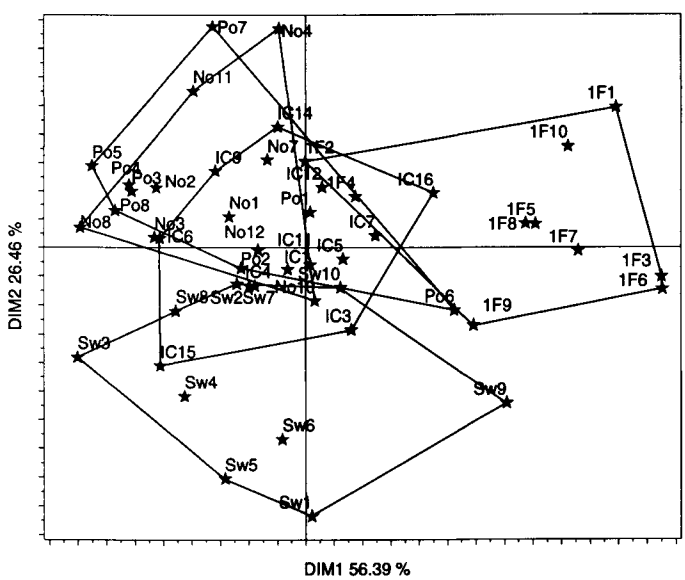


Figure 19 : S49 panel. CDA of panels 1F, IC, No, Po and Sw

Table 13
S49 panel. Mean RV coefficients within and between panels

	1F	IC	No	Po	Sw
1F	0.48				
IC	0.50	0.57			
No	0.49	0.57	0.56		
Po	0.49	0.56	0.57	0.56	
Sw	0.44	0.51	0.51	0.50	0.47

Table 14
S49 panel. Mean normalized RV coefficients within and between panels

	1F	IC	No	Po	Sw
1F	4.43				
IC	4.64	5.50			
No	4.65	5.58	5.54		
Po	4.63	5.46	5.69	5.55	
Sw	3.96	4.87	4.86	4.74	4.36

and between panels (Table 13) are rather similar and a little bit smaller when involving Swedish assessors (Sw panel). The corresponding normalized RV coefficients (Table 14) are all significant, meaning that the 5 panels are homogeneous and also agree among themselves about the sample structure, which does not automatically mean that the assessors are exchangeable among panels.

The first eigenvector of the RV matrix, which is the compromise, accounts for 54.74 % of the assessor variation (Table 15), which is a fairly good result considering that 48 axes exist in this PCO. The first plot of the PCO of the centered RV matrix does not show panel discrimination as exemplified by the panel convex hulls on Figure 18. But this plot accounts for only 25.35 % of the total information (Table 16). A good technique for mapping panel discrimination is the CDA suggested in section 2.9. For the S49 panel, the first two axes of this CDA accounts for 82.84 % of the panel discrimination (Table 17) and splits panel 1F and Sw between themselves and from the three other panels (Figure 19). The Norwegian panel (No) and the Polish panel (Po) do not seem different, meaning that the assessors might be exchangeable. Finally, the ICO panel appeared as the central panel on this map. In order to test significance of panel discrimination visually observed on this map, the permutation technique of section 2.9 was performed on the basis of the 16 first axes from the centered PCO and its output was 2.521, which is fairly significant. Therefore, one cannot exchange the assessors among the five panels, even if, on the average, they fairly agree about the sample structure. This conclusion was met by the author with several other data sets; it raises a controversial point associated to the permutation techniques described in this chapter and already mentioned above : are these tests too powerful because of a too unrealistic null hypothesis ?

Table 15
S49 panel. PCO of the non centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	26.8248	54.74	54.74
DIM 2	3.1572	6.44	61.19
DIM 3	1.8522	3.78	64.97
DIM 4	1.7188	3.51	68.48
DIM 5	1.4765	3.01	71.49
DIM 6	1.3574	2.77	74.26
DIM 7	1.2311	2.51	76.77
DIM 8	1.0794	2.20	78.97
DIM 9	1.0355	2.11	81.09
DIM10	0.9631	1.97	83.05

Table 16
S49 panel. PCO of the centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	3.3992	14.64	14.64
DIM 2	2.4869	10.71	25.35
DIM 3	1.7418	7.50	32.85
DIM 4	1.4779	6.36	39.21
DIM 5	1.3591	5.85	45.07
DIM 6	1.2317	5.30	50.37
DIM 7	1.0883	4.69	55.06
DIM 8	1.0385	4.47	59.53
DIM 9	0.9846	4.24	63.77
DIM10	0.8503	3.66	67.43

Table 17
S49 panel. Panel canonical discriminant analysis

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	2.4898	56.39	56.39
DIM 2	1.1684	26.46	82.84
DIM 3	0.5238	11.86	94.71
DIM 4	0.2337	5.29	100.00

Table 18
S49 panel. PCO of the sample compromise among subjects

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	0.7322	56.64	56.64
DIM 2	0.1416	10.95	67.59
DIM 3	0.0518	4.01	71.60
DIM 4	0.0470	3.64	75.24
DIM 5	0.0425	3.29	78.53
DIM 6	0.0412	3.19	81.72
DIM 7	0.0392	3.03	84.75
DIM 8	0.0331	2.56	87.31
DIM 9	0.0326	2.52	89.83
DIM10	0.0281	2.18	92.01
DIM11	0.0250	1.94	93.94
DIM12	0.0242	1.87	95.82
DIM13	0.0212	1.64	97.46
DIM14	0.0172	1.33	98.78
DIM15	0.0157	1.22	100.00

From the distribution of the eigenvalues of the compromise (Table 18) and from the mean b coefficient over the 49 assessors which was equal to 1.75 ('Mean' line and 'BETA' column in Table 19), it was decided to interpret the two first axes of the sample space given in Figure 20. This structure is once again similar to those already met with ICO and F2 panels, namely an opposition along the first axis of acidity and bitterness (Figure 21). Note that the astringency and the body/mouthfeel attributes are found correlated to bitterness.

The magnitude of the 'Scaling' column in Table 19, or in Table 20 which is Table 19 averaged by panel, is highly dependent on the panels simply because of the different scales used by these panels. In such a case, STATIS must be done with the global matrix standardization recommended in section 2.4. The 'NRV' and 'NRVC' columns in Table 19 point out the assessors No10 (assessor 10 from Norway) and assessor 1F6 (assessor 6 from France) as presenting a strong disagreement with the panel and with the compromise. The individual biplots of these two assessors (Figures 22 and 23) are indeed extremely different from the sample structures we have met up to now. For instance, assessor No10 found sample 8 very astringent and assessor 1F6 seems to correlate positively bitterness and acidity, therefore he did not make many differences between the acid samples 11 and 12 and the bitter samples 4, 13 and 16. One assessor for each of the three other panels (Sw1, IC9 and Po1) is also analysed by means of a biplot (Figures 24, 25 and 26). These assessors were chosen because they had the lowest agreement with the whole panel, but being still significant they do not present a very big difference with the compromise sample structure. Interestingly, all of them seem to present a null correlation between bitterness and acidity instead of a negative one. But comparison of attribute correlation structures is a matter of Dual STATIS which will be performed in the following section.

Table 19
S49 panel. Summary of individual statistics

Subject	Scaling	NRV	Weight	BETA	NRVC
1F 1	0.009	4.086	0.869	1.696	4.216
1F10	0.012	5.241	1.035	1.423	5.521
1F 2	0.015	6.129	1.197	1.729	6.475
1F 3	0.005	3.217	0.749	2.556	3.074
1F 4	0.017	6.361	1.232	1.932	6.695
1F 5	0.009	1.992	0.549	2.295	1.682
1F 6	0.011	1.227	0.442	2.609	0.769
1F 7	0.016	5.109	1.001	1.264	5.334
1F 8	0.012	5.306	1.048	1.644	5.497
1F 9	0.012	5.917	1.140	1.383	6.239
IC 1	0.017	6.264	1.211	1.951	6.568
IC11	0.011	5.993	1.178	2.163	6.249
IC12	0.057	5.835	1.153	1.579	6.209
IC14	0.016	4.304	0.907	1.464	4.543
IC15	0.037	4.808	0.997	1.825	5.024
IC16	0.023	5.672	1.111	1.843	5.939
IC 3	0.026	4.971	0.995	1.606	5.140
IC 4	0.061	5.658	1.119	1.531	5.994
IC 5	0.009	3.796	0.821	2.224	3.708
IC 6	0.037	6.031	1.179	1.309	6.477
IC 7	0.038	5.480	1.084	2.294	5.663
IC 9	0.005	3.543	0.817	3.121	3.327
No 1	0.010	5.006	1.005	1.616	5.214
No10	0.009	0.863	0.323	1.544	0.466
No11	0.037	6.312	1.219	1.462	6.717
No12	0.018	6.259	1.196	1.212	6.674
No 2	0.016	6.378	1.226	1.356	6.824
No 3	0.049	5.649	1.110	1.097	6.084
No 4	0.032	5.825	1.130	1.085	6.301
No 7	0.015	5.676	1.117	1.500	5.998
No 8	0.015	5.257	1.066	1.930	5.494
Po 1	0.651	3.787	0.845	2.458	3.785
Po 2	3.173	5.460	1.066	1.246	5.739
Po 3	0.977	5.436	1.086	1.985	5.653
Po 4	2.510	4.921	1.012	1.744	5.188
Po 5	3.605	5.605	1.110	1.575	5.955
Po 6	0.730	4.350	0.909	1.963	4.393
Po 7	2.914	5.114	1.014	1.237	5.462
Po 8	3.631	6.867	1.303	1.219	7.376
Sw 1	2.583	2.252	0.587	2.177	2.005
Sw10	2.051	5.438	1.048	1.089	5.769
Sw 2	7.994	5.672	1.116	1.237	6.045
Sw 3	1.462	4.862	0.996	2.119	5.029
Sw 4	3.983	6.143	1.190	1.651	6.462
Sw 5	1.306	3.484	0.803	2.582	3.392
Sw 6	1.392	3.239	0.771	2.749	3.123
Sw 7	4.015	5.524	1.091	1.617	5.767
Sw 8	2.270	5.400	1.069	1.341	5.764
Sw 9	3.096	3.583	0.759	1.503	3.493
Mean	1.000	4.925	1.000	1.750	5.113

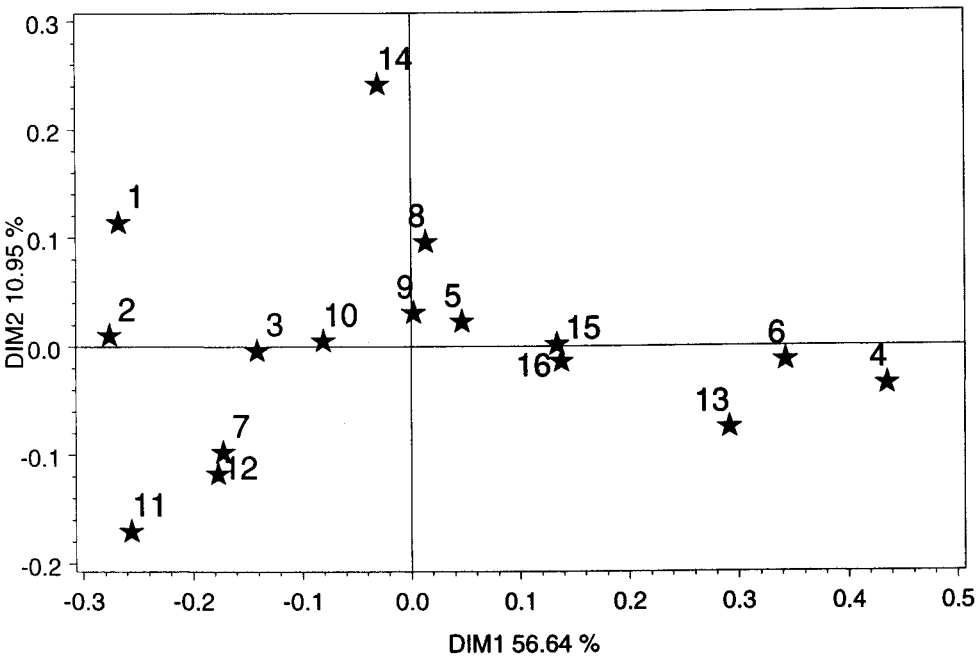


Figure 20 : S49 panel. PCO of the sample compromise among subjects

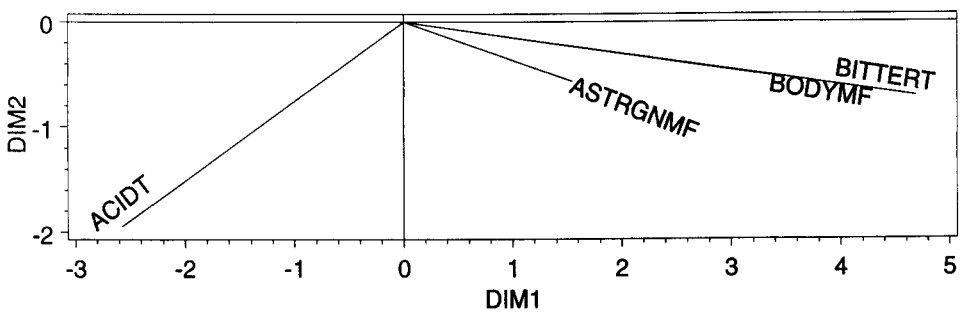


Figure 21 : S49 panel. Covariance plot of attributes

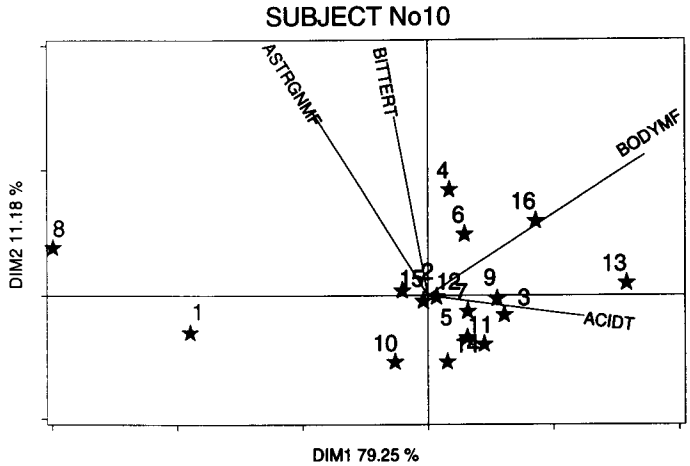


Figure 22 : S49 panel. Covariance PCA of scores from subject No10

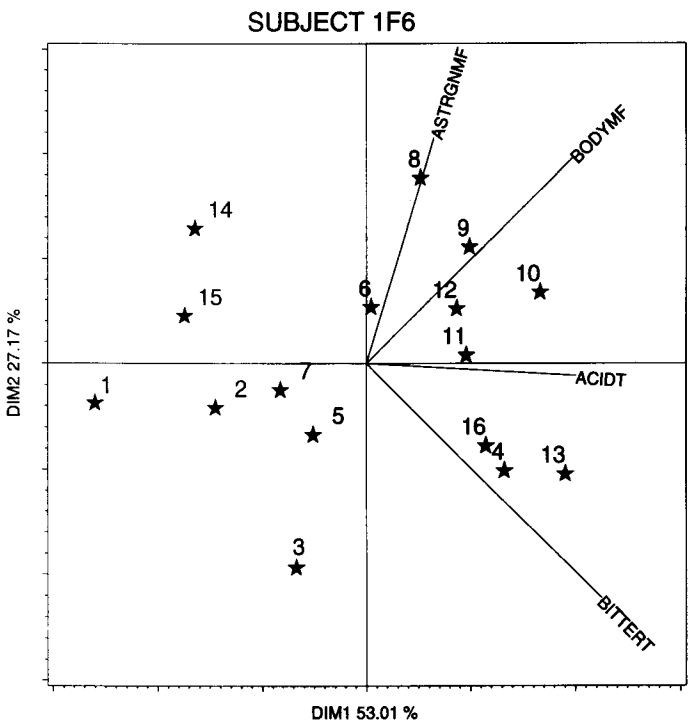


Figure 23 : S49 panel. Covariance PCA of scores from subject 1F6

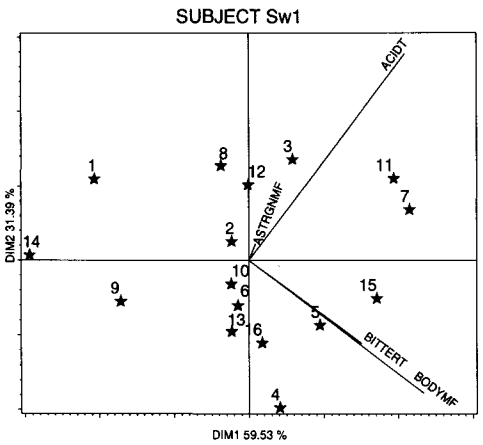


Figure 24 : S49 panel. Covariance PCA of scores from subject Sw1

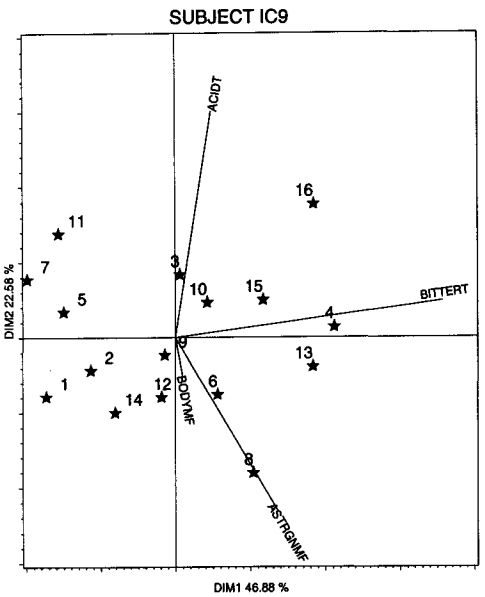


Figure 25 : S49 panel. Covariance PCA of scores from subject IC9

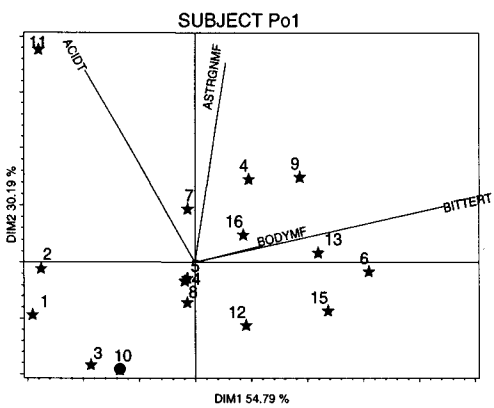


Figure 26 : S49 panel. Covariance PCA of scores from subject Po1

Table 20
S49 panel. Summary of individual statistics averaged by panel

Subject	Scaling	NRV	Weight	BETA	NRVC
1F	0.012	4.459	0.926	1.853	4.55
IC	0.028	5.196	1.048	1.909	5.403
No	0.022	5.247	1.044	1.422	5.53
Po	2.274	5.192	1.043	1.678	5.444
Sw	3.015	4.56	0.943	1.807	4.685

4.5 Dual STATIS of the S49 panel

Table 21
S49 dual panel. Mean RV coefficients within and between panels

	1F	IC	No	Po	Sw
1F	0.77	0.72	0.56	0.73	0.70
IC	0.72	0.78	0.74	0.79	0.78
No	0.56	0.74	0.75	0.73	0.75
Po	0.73	0.79	0.73	0.82	0.80
Sw	0.70	0.78	0.75	0.80	0.77

Table 22
S49 dual panel. PCO of the non centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	36.9344	75.38	75.38
DIM 2	5.7214	11.68	87.05
DIM 3	2.5591	5.22	92.28
DIM 4	1.8882	3.85	96.13

The mean RV coefficients within and between panels on the basis of the individual correlation matrices (Table 21) are greater than those obtained on the basis of the sample spaces (Table 13), suggesting that the assessors agree more on attribute correlations than on sample differences. As explained in section 2.10, no analytical permutation test exists in the framework of Dual STATIS. The first eigenvector of the RV matrix accounts for 75.38 % of the PCO (Table 22), which again denotes a better compromise than that obtained by STATIS (54.74 %, Table 15). The first three axes of the PCO of the centered RV matrix (Table 23) explain almost all the variation, which once again was definitely not the case in classical STATIS (Table 16). But the reader should remember that this Dual STATIS compares matrices of size (4,4), whereas the classical STATIS compared matrices of size (16,16);

Table 23
S49 dual panel. PCO of the centered RV matrix

Axis	Eigenvalue	Variance (%)	Cum.Var. (%)
DIM 1	5.8397	46.54	46.54
DIM 2	2.5852	20.60	67.15
DIM 3	1.9749	15.74	82.89
DIM 4	0.8342	6.65	89.54

Table 24
S49 dual panel. Summary of individual statistics

Subject	Weight	Subject	Weight	Subject	Weight	Subject	Weight	Subject	Weight
1F 1	1.119	IC 1	1.112	No 1	0.798	Po 1	1.050	Sw 1	0.968
1F10	0.795	IC11	1.083	No10	0.611	Po 2	1.024	Sw10	1.075
1F 2	1.109	IC12	1.060	No11	1.078	Po 3	1.064	Sw 2	1.041
1F 3	0.950	IC14	1.096	No12	1.024	Po 4	0.853	Sw 3	1.041
1F 4	0.874	IC15	1.090	No 2	0.978	Po 5	1.067	Sw 4	1.090
1F 5	0.991	IC16	1.072	No 3	0.994	Po 6	1.130	Sw 5	1.107
1F 6	0.819	IC 3	0.984	No 4	0.990	Po 7	1.022	Sw 6	1.017
1F 7	0.926	IC 4	1.068	No 7	1.063	Po 8	1.138	Sw 7	1.122
1F 8	0.987	IC 5	0.708	No 8	1.119			Sw 8	1.021
1F 9	0.797	IC 6	1.000					Sw 9	0.795
		IC 7	1.084						
		IC 9	0.994						
Mean	0.937	Mean	1.029	Mean	0.962	Mean	1.044	Mean	1.028

Table 24 gives the Dual STATIS weights of the 49 assessors. One can observe that they are quite similar among assessors and also among panels. These weights allow the compromise correlation matrix (Table 25) to be derived as a weighted mean of the 49 individual correlation matrices.

Table 25
S49 dual panel. Compromise correlation matrix

	ACIDT	ASTRGNMF	BITTERT	BODYMF
ACIDT	1.00			
ASTRGNMF	-0.05	1.00		
BITTERT	-0.30	0.55	1.00	
BODYMF	-0.24	0.42	0.63	1.00

therefore the better agreement raised in Dual STATIS may be due, or partly due, to this difference. Figure 27 scatters assessors according to Dual STATIS weights and classical STATIS weights. Therefore, assessors located at the bottom of this plot disagree the most

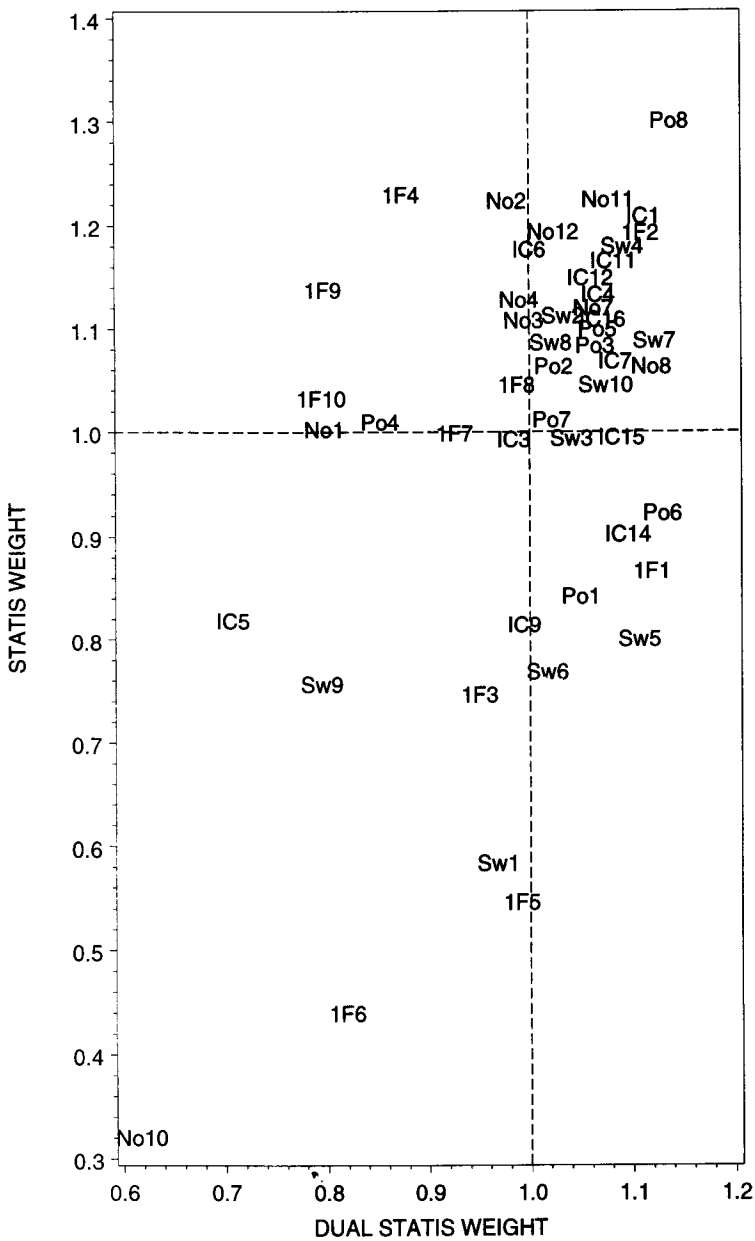


Figure 27 : S49 panel. STATIS and DUAL STATIS weights of subjects

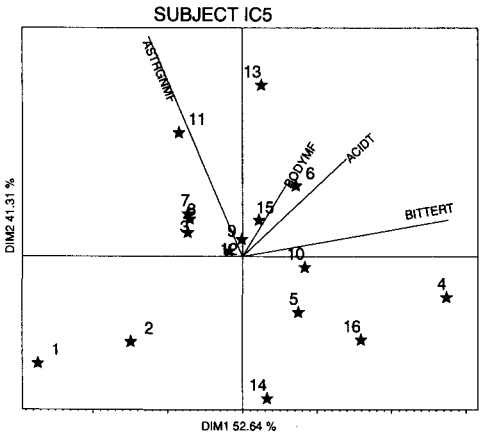


Figure 28 : S49 panel. Covariance PCA of scores from subject IC5

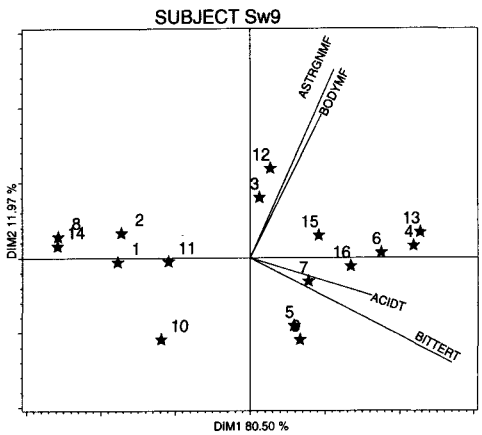


Figure 29 : S49 panel. Covariance PCA of scores from subject Sw9

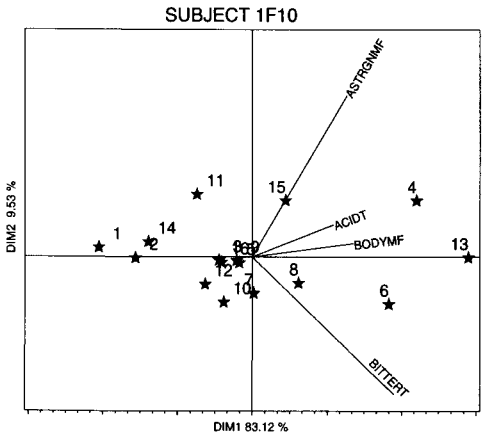


Figure 30 : S49 panel. Covariance PCA of scores from subject 1F10

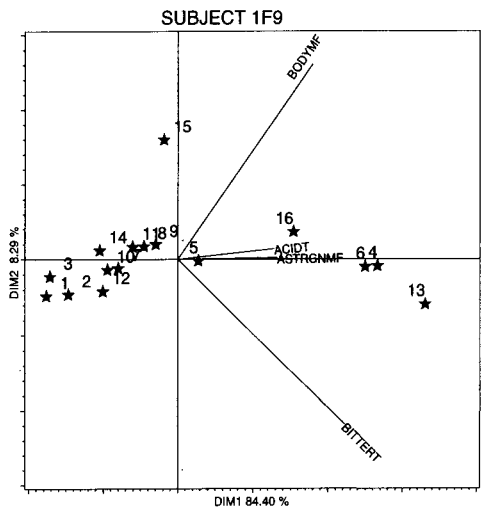


Figure 31 : S49 panel. Covariance PCA of scores from subject 1F9

with the panel on sample differences, whereas assessors located on the left of this plot disagree the most with the panel on attribute correlations. The worst assessors are obviously those located in the bottom left quadrant of this plot. The worst assessor is definitely No10, whose biplot has already been given in Figure 22. On the contrary, the best assessor seems to be Po8. The biplot of four assessors, who poorly agree with the Dual STATIS compromise, are given in Figures 28, 29, 30 and 31. IC5 and Sw9 have in common to correlate positively bitterness and acidity. The two French subjects 1F10 and 1F9 have very similar biplots, unless for acidity which is highly correlated with astringency for 1F9, whereas it is with body/mouthfeel for 1F10.

5. SOFTWARE

Every computation and graph presented in this chapter was done on a SUN workstation under the UNIX system and within the SAS® software, thanks to several macros developped by the author. These macros should be announced elsewhere.

6. CONCLUSION

Advantages and limitations of the techniques proposed have been discussed throughout the chapter.

The most important advantages were :

- Ability to take free-choice profiling into account
- RV coefficient for measuring similarity between two sample spaces
- Permutation tests for validation of the panel homogeneity with almost no computation
- β coefficient of individual dimensionality (complexity)
- STATIS compromise on sample distances or attribute correlations
- Compromise obtained as a mean of assessors weighted by their individual agreement with the panel
- Cross-panel comparisons
- Analytical, instead of iterative, techniques

The principal limitations were :

- Possible over-powerful tests
- No analytical permutation tests with Dual STATIS
- Individual weights only depend on agreement with the panel.

The comparison of these two lists makes it clear why the author does think that the RV related techniques have a great potential in the sensory field. Nevertheless, this cannot be recognized by the sensory community before they have been sucessfully applied by many sensory scientists in a wide range of situations. This could not happen without papers such as this chapter and without availability of dedicated softwares.

7. ACKNOWLEDGMENT

The author would like to thank Yves Escoufier for introducing the RV coefficient to him long time ago and Frédérique Kazi-Aoual for sharing as soon as possible her useful theoretical results with praticians. The author is also very grateful to his colleagues from ESN who allowed this paper to be illustrated by the coffee data. Finally the help of Patricia Collin with the English language and the help of Stéphanie Degoud with tables and figures were greatly appreciated.

8.REFERENCES

Arnold, G. M., & Williams, A. A. (1986). The use of generalized procrustes techniques in sensory analysis. In J. R. Piggott (Ed.), Statistical procedures in food research (pp. 233-253): Elsevier Applied Science.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35(3), 283-319.

Dijksterhuis, G. B., & Gower, J. C. (1991/2). The interpretation of generalized procrustes analysis and allied methods. Food Quality and Preference, 3, 67-87.

Dijksterhuis, G. B., & Punter, P. (1990). Interpreting generalized procrustes analysis 'Analysis of Variance' tables. Food Quality and Preference, 2, 255-265.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. New-York: Chapman & Hall.

Escoufier, Y. (1973). Le traitement des variables vectorielles. Biometrics, 29, 751-760.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53(3 and 4), 325-338.

Gower, J. C. (1975). Generalized procrustes analysis. Psychometrika, 40(1), 33-51.

Jolliffe, I. T. (1986). Principal component analysis. New-York: Springer-Verlag.

Kazi-Aoual, F. (1992). Apport de la méthode STATIS pour étudier l'homogénéité d'un panel. Paper presented at the 3èmes journées européennes "Agro-Industrie et Méthodes Statistiques", Montpellier.

Kazi-Aoual, F. (1993). Approximations to permutation tests for data analysis (Report n° 93-06). E.N.S.A. Montpellier, unité de Biométrie.

Kazi-Aoual, F., Hitier, S., Sabatier, R., & Lebreton, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. Computational Statistics and Data Analysis, in press.

L'Hermier des Plantes, H. (1976). Structuration des tableaux à trois indices de la statistique. Thesis, Montpellier II.

Lavit, C. (1988). Analyse conjointe de tableaux quantitatifs. Paris: Masson.

Lavit, C., Escoufier, Y., Sabatier, R., & Traissac, P. (1994). The ACT (STATIS method). Computational Statistics and Data Analysis, 18, 97-119.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate analysis. London: Academic Press.

McEwan, J. A., & Schlich, P. (1991/2). Correspondence analysis in sensory evaluation. Food Quality and Preference, 3, 23-36.

Robert, P., & Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods : the RV-coefficient. Appl. Statist., 25(3), 257-265.

Scheffe, H. (1959). The Analysis of Variance. New-York: John Wiley & Sons.

Schiffmann, S. S., Reynolds, M. L., & Young, F. W. (1981). Introduction to multidimensional scaling. Theory, methods and applications. London: Academic Press.

Schlich, P. (1989). A SAS/IML® program for generalised procrustes analysis. Paper presented at the "SEUGI' 89" SAS European Users Group International Conference, Cologne.

Schlich, P. (1993). Contributions à la sensométrie. Thesis, Paris XI Orsay.

Schlich, P., & Guichard, E. (1989). Selection and classification of volatile compounds of apricot using the RV coefficient. Journal of Agricultural and Food Chemistry, 37(1), 142-150.

Schlich, P., Issanchou, S., Guichard, E., Etievant, P., & Adda, J. (1987). RV coefficient: a new approach to select variables in PCA and to get correlations between sensory and instrumental data. Paper presented at the 5th Weurman Flavour Research Symposium, Oslo.

Wakeling, I. N., Raats, M. M., & Macfie, H. J. H. (1992). A new significance test for consensus in generalized procrutes analysis. J. of Sensory Studies, 7, 91-96.

Williams, A. A., & Langron, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. J. Sci. Food Agric., 35, 558-56

Multivariate Analysis of Data in Sensory Science
Ed. by T. Næs and E. Risvik

ANALYSING INDIVIDUAL PROFILES BY THREE-WAY FACTOR ANALYSIS

Per M. Brockhoff[a], David Hirst[b] and Tormod Næs[c]

[a]Royal Veterinary and Agricultural University, Thorvaldsensvej 40, DK-1871 Frederiksberg C Denmark,

[b]Scottish Agricultural Statistics Service, Rowett Research Institute Bucksburn, Aberdeen AB2 9SB Scotland

[c]MATFORSK, Oslovegen 1, 1430 Ås Norway.

1. INTRODUCTION

1.1 Advantages of three-way methods in sensory analysis

Three-way factor analysis (TWFA) techniques first appeared in the psychometric literature, see for instance Tucker (1966), Kroonenberg and De Leeuw (1980) and Kloot and Kroonenberg (1985), and have been used in several applications (Henrion et al. (1992), Leurgans and Ross (1992)). So far, however, there are few applications within the field of sensory analysis. The aim of this chapter is to discuss these methods within a sensory context and show that they can be useful for analysis of individual sensory profile data.

TWFA techniques are generalizations of principal components analysis (PCA) but while PCA works on two-dimensional matrices, TWFA techniques can be used to analyse three-dimensional matrices with three 'directions'or 'ways' of information. Therefore, they can be used to investigate similarities and differences between objects, assessors and attributes at the same time. The kind of questions that can be answered by these techniques are for instance:

- Do the assessors use the attributes or the measurement scales differently?
- Are some of the assessors more sensitive than others to some of the attributes?
- Are some of the assessors better at tasting differences among certain groups of objects?
- Do all assessors distinguish equally well between the objects?
- Do the assessors use the same attributes to distinguish between the objects and to span the underlying variable space?

All these questions are of interest to the panel leader who is responsible for the quality of the panel and may wish to retrain or remove some of the assessors, to the data analyst who has to make decisions about which analysis technique is most appropriate, and to the manufacturer since they can highlight variability among consumers' perceptions of the objects. The results of a TWFA can be presented in simple two- or three-dimensional scatter-plots, which may be relatively easy to interpret. In the following sections several techniques will be discussed, emphasizing applications and the relationship between TWFA methods and other techniques in this book.

1.2 The structure of profile data

Assume there are m assessors in the sensory panel measuring p attributes for n objects. The data can then be collected in a three-way table y_{ijk}, $i = 1,...,m$, $j = 1,...,n$, and $k = 1,...,p$. Replicates will here be denoted by $l = 1,...,q$. The handling of replicates is discussed in Section 7. They can either be averaged over or treated separately, in which case each of the m x n x p cells of the three-way matrix of data consists of q elements. This type of data can always be described by an analysis of variancemodel, see Searle (1971),

$$x_{ijkl} = \mu_k + \alpha_{ik} + \beta_{jk} + \delta_{ijk} + \varepsilon_{ijkl} \tag{1}$$

The main effects α_{ik} for assessor i (and attribute k) represent the differences between this assessor's average score for that particular attribute and the overall average for the same attribute. The main effect β_k describes how the average score for object j and attribute k deviates from the overall average for the same attribute. The interactions δ_{ijk} represent the differences between assessors in measuring differences between objects. Note that individual differences among assessors are present both in the main effects a_{ik} and in the interactions δ_{ijk}. The error terms ε_{ijkl} represent variation due to replicates under the same experimental conditions.

The TWFA methods in this paper will model both these types of individual differences if no pretreatment of the data is used. There exist preprocessing techniques, however (see below), which eliminate the main effects α_{ik} from the analysis and only concentrate on the interactions.

2. DIFFERENT TWFA MODELS

2.1 TWFA as a generalisation of PCA

Standard PCA of an n x p matrix X is based on the following 'model'

$$X = TP' + E \tag{2}$$

where T ($n \times a$) is the matrix of object scores (defined to have orthogonal columns), P' ($a \times p$) the variable loadings (orthogonal rows) and E ($n \times p$) the matrix of residuals, corresponding to those direction in principal component space that have little variability and which are frequently interpreted as noise. The loadings P are defined so as to describe as much of the variation in X as possible given the dimension a, normally with $P'P=I$, and T is found as the projection of X on P.

Alternatively, this can be stated as the problem of finding the T and P matrices that minimize the residuals E, i.e. the T and P that minimize the least squares criterion

$$\|X - TP'\|^2 \tag{3}$$

The T and P matrices are usually plotted in low-dimensional scatter-plots to reveal structures among the objects and among the attributes.

Three-way factor analysis techniques are generalizations of PCA developed for matrices with an extra way (or order), see Figure 1. Each slice in the stack of matrices corresponds to one particular assessor and contains objects-by-attributes information for that particular assessor. Of course it is equally possible to slice the matrix in two other ways, with the slices then corresponding to either individual objects or attributes. It would be possible to do a separate PCA on each slice of the matrix, which would be to ignore any similarities between the assessors (or objects or attributes depending on how the matrix was sliced) or to take a mean over the slices and do a PCA on the resulting matrix, which would ignore any differences between them. TWFA is a form of PCA for the slices of the matrix which takes account of these similarities and differences.

2.2 Tucker-1 modelling

If we call the $n \times p$ slice of the three way matrix corresponding to assessor i's individual objects-by-attributes matrix X_i where $i = 1,..., m$, then one possible way to analyse the data is to model X_i as

$$X_i = T_i P' + E_i, \tag{4}$$

where P has dimension p x a ($a < p$). The number a is chosen to give a low dimensional approximation to the data as in PCA, and T_i and P are found for any a by minimization of the least squares criterion

$$\sum_{i=1}^{m} \|X_i - T_i P\|^2. \tag{5}$$

There are no constraints on the T_i here, but P is usually constrained to have orthogonal rows, i.e. $P'P = I_a$. This can be seen as a PCA of each X_i where each PCA is forced to have the same variable loadings matrix P, though the scores T_i are allowed to vary. An interpretation of this model is that the assessors perceive the same underlying variables but rate the objects differently to obtain individual scores matrices. It is generally known as the common loadings Tucker-1 model.

It is useful to note here that if we let the m x p slice corresponding to the assessors-by-variables matrix for object j be Y_j, then we can write

$$Y_j = U_j P' + E_j. \tag{6}$$

Then minimizing $\Sigma_{j=1}^{n} \|X_j - T_j P'\|^2$ will give exactly the same common loadings matrix P (and the same fit).

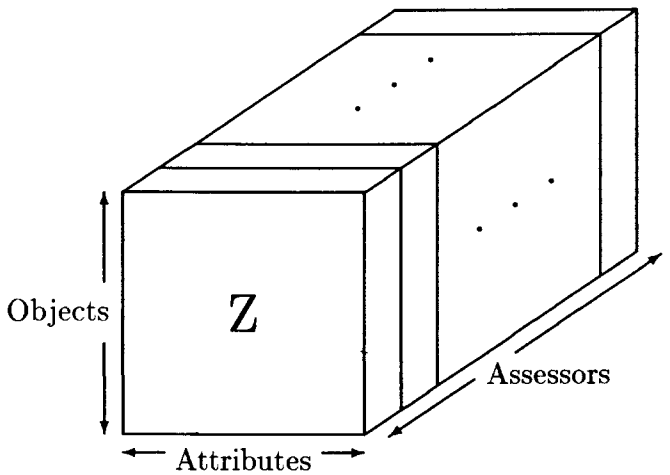


Figure 1: Three-way data matrix

Alternatively, TWFA models can be based on the model

$$X_i = TP_i' + E_i, i = 1,\ldots,m \tag{7}$$

where now the loadings P_i differ from assessor to assessor. T has dimension n x b where b is the reduced dimensionality of the model. T is generally constrained to have orthogonal columns, i.e. $T'T = I_b$, but the P_i are unconstrained. The assessors have a common scores matrix T, which describes relationships among the samples, but differ in the way they perceive the variables. This is known as the common scores Tucker-1 model. It is equivalent to writing $Z_k = TV_k + E_k$, $k = 1,\ldots,p$, where Z_k is the n x m slice of objects-by-assessors for variable k.

There is also a third Tucker-1 model formed by writing or

$$Z_k' = QW_k + E_k \tag{8}$$

$Y_j = QR_j + E_j$. Here Q is the 'assessor scores' matrix with dimension m x c, c (< m) being the reduced dimension. In general the three different models will give different fits to the data.

Which of the three models one uses depends on the aim of the analysis. For instance if one is interested primarily in the relationships among the objects, i.e. which of the objects are similar and whether or not they can be represented in a low dimensional 'object space', then the common scores model is appropriate. A possible interpretation of this model is that the b new 'object dimensions' represent 'ideal object types', and that each real object is made up of a linear combination of these types. For example in the example discussed later it might be possible to represent the objects in only one dimension going from 'ideal cheddar' to 'ideal Norwegian'.

Mature cheddar would have a high score in this dimension, Norwegian a low score and Norwegian Cheddar would lie somewhere in between.

Note that this model says nothing about the relationships among the attributes or among the assessors. In fact the assessors could all use their own individual sets of attributes without the analysis being changed.

If interest is primarily in the relationships among the variables, e.g. whether there are some 'underlying factors' perceived by all of the assessors, then the common loadings model is appropriate. The interpretation is exactly analogous to that for the common scores model, i.e. that the attributes can be represented in a lower dimensional space, with the new dimensions being interpreted as 'ideal' or 'underlying' attributes, perceived by all of the assessors. Taking the cheese example again, perhaps one of the underlying variables could relate to texture, going from firm and rubbery to crumbly and grainy. Again nothing is said about relationships among the assessors or objects.

If interest is in the relationships among the assessors, then the third Tucker-1 model is the best. The 'common assessor scores' Q can be plotted to look for relationships among the assessors. The implication is that the assessors can be represented in a lower dimensional space, i.e. there are a few underlying 'assessor types', with each assessor being a linear combination of some or all of them.

If there is interest in more than one mode, e.g. in both assessors and attributes (as is often the case), then there are two possible approaches. The first is to take the individual scores matrices from a common loadings Tucker-1 model, and to look for similarities among them. This can be done by 'stringing out'the rows of each matrix into long rows of length na, joining these rows into one new matrix of dimension m x na and doing a PCA on this matrix. The scores on the first few PCs of this matrix can be plotted to look for relationships among the assessors, and the eigenvalues examined to decide on the dimensionality of the assessor space. This is equivalent to a Tucker-1 analysis on the individual scores matrices.

This is a two stage process, first the attribute dimension is reduced to approximate the raw data, and the resulting 'underlying attributes' are examined. Then the assessor dimension is reduced to find an approximation to this approximation, and the resulting assessor dimensions examined. This means that the relationships between the attributes are modelled as well as possible (in the chosen reduced dimensionality), and the assessors are modelled less well. This is a sensible approach if the variables are considered of primary interest. If the two modes are of equal interest, then a Tucker-2 model is more appropriate.

2.3 Tucker-2 modelling

Tucker-2 modelling is a generalization of Tucker-1 modelling to reduce the dimensionality of two modes simultaneously. There are three versions, one for each pair of modes. The most usual is probably the one having common scores T, common loadings P and individual assessor matrices W_i, $i = 1,..., m$. These W_i relate T and P through a different linear transformation for each assessor. This model is written as

$$X_i = TW_iP' \tag{9}$$

where the W_i have dimension b x a. T (n x b) and P (p x a) are found to minimize the least squares criterion

$$\sum_{i=1}^{m} \left\| X_i - TW_iP' \right\|^2 \tag{10}$$

Note that this model can be written both as an individual loadings model and an individual scores model. In the former case, the individual loadings are $P_i = PW_i'$ and in the latter case, the individual scores are $T_i = TW_i$. For the individual scores model, the individual scores $T_i = TW_i$ can be interpreted as products of a common score matrix multiplied by the individual transformation matrices W_i, but this method will not in general give the same fit as the Tucker-1 model.

The interpretation of this model is that the objects can be represented in a b ($< n$) dimensional space, and the variables can be represented in an a ($< p$) dimensional space. In other words there are a 'underlying attributes' which describe b 'ideal object types'. Each assessor uses the underlying attributes in a different way to describe the ideal objects. The individual difference matrices W_i describe how each assessor does this. The matrix T gives the scores of the objects in the object space, and the first two dimensions (for example) can be plotted to examine their structure. P gives the loadings of the underlying attributes on the attributes, and is interpreted in the usual way. Of course it is not possible to link the object scores to the attribute loadings in any meaningful way, as the link is different for each assessor. As with the Tucker-1 models, this Tucker-2 model is not well suited to provide information about the assessors. It is possible to do a Tucker-1 analysis of the W_i matrices in order to look for associations among the assessors, in the same way as it is possible to analyse the individual scores matrices from a Tucker-1 model. However, it is more sensible to choose a Tucker-2 model to investigate the modes of interest directly. Hence if the attributes and assessors are of interest, it is possible to write a Tucker-2 model as

$$Y_j = QO_jP' \tag{11}$$

Q is now an $m \times c$ matrix of 'assessor scores' and P an $p \times a$ matrix of attribute loadings. The Q matrix then gives information on the relationships between the assessors (common for each object), and the O_js are the object difference matrices that link together the 'object-common' loadings and scores. Alternatively, if there is interest in all three modes, the Tucker-3 model is appropriate, as is the PARAFAC model described later.

2.4 Tucker-3 modelling

The Tucker-3 model is the natural generalization of Tucker-2. There is only one Tucker-3 model, and it can be represented as the Tucker-2 model in equation (9), where the W_i are expressed as linear combinations of a limited number, c, of fixed matrices C_j (a different linear combination for each assessor). This model can equally well be written as equation (11) where the O_j are linear combinations of fixed matrices. This model links together all three modes in an interpretable way. It can also be written as

$$Z = TC(Q' \otimes P'), \tag{12}$$

where $\otimes$ is the kronecker or direct product. Z is the data unfolded to form an $n \times mp$ matrix of objects-by-(assessors x attributes), with each assessor's attributes kept together in a block. T is the $n \times b$ matrix of object scores, P the $p \times a$ matrix of variable loadings, Q is the $m \times c$ matrix of assessor scores, and C is the $b \times ac$ matrix made up of the core matrices placed side by side. The interpretation is as follows: The objects lie in a b-dimensional space the axes of which represent ideal object dimensions'. Each object can be described as a linear combination of these ideal objects. The attributes lie in an a-dimensional space, the axes of which represent 'underlying attributes'.

Each attribute can be described as a linear combination of the underlying attributes though it is more usual to consider the underlying attributes as linear combinations of the original attributes. The assessors lie in a c-dimensional space, the axes of which represent 'ideal assessor types' or underlying ways of perceiving the samples. Each assessor is a linear combination of these types.

2.5 Interpreting the core matrices in a Tucker-3 model

The three modes are linked through the core matrix, and it is sometimes possible to interpret this matrix in a helpful way. Suppose we have reduced each mode to two dimensions, and so there are two 'assessor types', two 'object types' and two 'underlying attributes'. The core matrix is a three way matrix so consider the slice corresponding to assessor type 1. This is a 2×2 matrix which relates the object types to the underlying attributes. Suppose the underlying attributes have been interpreted as sweet/salt and rubbery/creamy, and the first object type is Norwegian/cheddar. The first slice of the core matrix may be

$$\begin{pmatrix} 1 & 0 & 2 \\ 4 & & 8 \end{pmatrix}$$

The first row corresponds to the weight assessor type 1 gives to the two underlying attributes in describing object type 1, in other words he/she describes ideal Norwegian cheese mainly as sweet, but also with an element of rubberyness. Ideal cheddar would then be described as very salty with a hint of creamyness.

Interpreting the core matrix can be very difficult, especially if the dimensions in the three modes cannot be interpreted. One technique that can be helpful is drawing a separate biplot for each assessor type, i.e. in each plot the scores would be given by T and the variables by C_jP'. This gives a picture of how the assessor types relate the actual objects to the measured attributes. Similarly biplots could be drawn for each object type or each underlying attribute. The former would give a picture of which attributes different assessors considered important in describing the object types, the latter a picture of which objects each assessor considered to have the ideal attributes.

2.6 The PARAFAC model

The other three mode method is the PARAFAC model which is defined by equation (9) where the W_i are forced to be diagonal with only positive elements on the diagonal. This is also known as the CANDECOMP model, see for instance Carrol and Chang (1970), and Harshman and Lundy (1984). This model is no longer symmetrical in the three modes, and has a slightly different interpretation. This is that the assessors perceive the same underlying attributes, but weight them differently when scoring the objects. This model can be useful if the assessors disagree on which attributes are most important for describing differences among objects. There are of course three different versions of the PARAFAC model, corresponding to the three different Tucker-2 models. Note that in order for the W_i to be diagonal, two of the modes are forced to have the same dimension. For example for the Tucker-2 model (9) the object and attribute dimensions would have to be equal. This is not the case with the Tucker-3 model.

2.7 Three mode analysis using single mode methods

As mentioned above, it is possible to move from a single mode to a two mode model by successive application of a Tucker-1 model. It is then clearly possible to obtain a three mode model by another

application of the Tucker-1 model. This has computational advantages since a standard principal components analysis program can be used (see below), rather than specialized software. The procedure is: First a Tucker-1 model is applied to the raw data, for example the common loadings model in equation (4). This results in an $a \times p$ common loadings matrix P, and m individual $n \times a$ scores matrices T_i, $i = 1,...,m$. These T_i can now be analysed using Tucker-1, using either the common object scores or common assessor scores model. As an example the former of these will result in an $n \times b$ common scores matrix T, and m individual $b \times a$ matrices Q_i. These Q_i can now be analysed by the common assessor scores Tucker-1 model to give an $m \times c$ matrix Q, and a $c \times b$ matrices W_i, the core matrices.

This procedure can be followed in 6 different ways depending on the order in which T, P and Q are found, and in general they will all give different results. Also, each will give a poorer fit to the original data than the Tucker-3 model, since this directly minimizes the sum of squared residuals, equation (10). For these reasons this method is not to be recommended if Tucker-3 programs are available.

3. FITTING THE TWFA MODELS

3.1 Tucker-1

The Tucker-1 model is based on unrestricted minimization over T_i and P of the quantity

$$\sum_{i=1}^{m} \left\| X_i - T_i P' \right\|^2, P'P = I \tag{13}$$

for the common loadings model, and minimization of an analogous expression for the other two models. If the three way matrix is unfolded to give an $mn \times p$ matrix, as in Figure 2, then it is easy to see that the minimization is achieved by a standard PCA or SVD of the unfolded matrix. Notice that the eigenvectors of the unfolded matrix are identical to the eigenvectors of the sum of the S_i, $i = 1,...,m$, where S_i is the covariance matrix for assessor i.

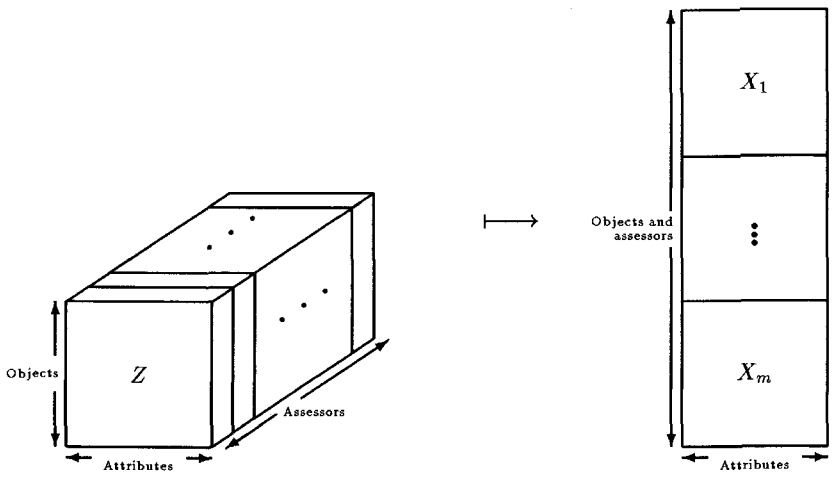


Figure 2: Unfolding of three-way data matrix

3.2 Tucker-2

For this case the minimization is over W_i, T and P of the equation

$$\sum_{i=1}^{m} \|X_i - TW_iP'\|^2 \tag{14}$$

(or one of the other two forms) where T is a set of common scores, P the common loadings and W_i are the individual difference matrices to be estimated.

The solution to this is more complicated than for Tucker-1 and must be done by numerical optimization. A solution based on alternating least squares (ALS) was proposed in Kroonenberg and De Leeuw (1980). The optimization works by finding the best solution for T given P, then the best P is found given the value of T. This procedure continues until convergence. Then finally the W_i that minimize (14) are found. Using ALS ensures that an improved fit is obtained for each cycle and so convergence is guaranteed. There is, however, no guarantee that the global minimum value of the criterion is obtained.

In more detail, the solution for P, T and W_i can be found from the following algorithm.

1. Construct starting values of P (e.g. from a Tucker-1 solution).
2. Compute $D = \sum_{i=1}^{m} X_iPP'X_i'$
3. Put the eigenvectors associated with the b largest eigenvalues of D into the columns of a matrix T.
4. Compute $Q = \sum_{i=1}^{m} X_i'TT'X_i$.

5. Put the eigenvectors associated with the a largest eigenvalues of Q into the rows of P.

6. Repeat 2-5 until convergence.

7. Put $W_i = T'X_iP$.

This algorithm gives a solution in which P and T have orthogonal columns or rows, since they are formed from eigenvectors. This is not, however, a constrained minimization, the solution is the minimum over all T and P (though it may be a local rather than global minimum). Any solution to the minimization of (14) is in fact unidentified, since any of the matrices P, W_i and T can be multiplied by linear transformation matrices without consequence for the fit, if the other two matrices are corrected accordingly. For instance P can be multiplied by F, and W_i by F^{-1} without changing the fit. It should be mentioned that even when constraining the columns of P and T to be orthogonal the solution is unidentified.

3.3 PARAFAC-CANDECOMP

In this case the optimization criterion is the same as above, namely

$$\sum_{i=1}^{m} \left\| X_i - TW_iP' \right\|^2 \tag{15}$$

where P and T are unrestricted, but now the W_i's are diagonal matrices. The solution must be found by numerical methods such as the ALS method mentioned above. An exact eigenvector-based estimation procedure for the parameters has been proposed for certain chemical applications of the model, Sanchez and Kowalski (1990), but this exact solution does not optimize the LS criterion.

The ALS solution, see for example Carrol and Przuzanski (1984), is found in a similar way to that for the Tucker-2 model above. One starts with initial values of T and P and estimates W_i, then T is reestimated before P is reestimated. One continues until convergence. The exact eigenvector solution mentioned above can be used to find starting values. In more detail the algorithm is as follows:

1. Construct starting values of P and T.

2. Find W_i as diagonal matrices with the same diagonal as $T'X_iP$.

3. Compute $D = \sum_{i=1}^{m} X_i (PW_i) (PW_i)' X_i'$.

4. Put the eigenvectors associated with the b largest eigenvalues of D into the columns of a matrix T.

5. Compute $Q = \sum_{i=1}^{m} X_i' (TW_i) (TW_i)' X_i$.

6. Put the eigenvectors associated with the a largest eigenvalues of Q into the columns of P.

7. Repeat 2-6 until convergence.

It should be mentioned that in this case the solution is only unidentified with respect to scalar multiplication of the matrices. This means for instance that no rotation of the matrices is allowed. This was proved by Kruskal (1977) and is an interesting feature of the model.

3.4 Tucker-3

For the Tucker-3 model, each W_i is assumed to be a linear combination (dependent on i) of matrices which are independent of k. In other words,

$$W_i = \sum_{j=1}^{n} c_{ij} C_j \tag{16}$$

where C_j are matrices independent of i, and c_{ij} are constants. Alternatively this can be written as $Z = TC(Q' \otimes P')$, as described in Section 2.4. T, Q, P and C are found by an ALS procedure similar to that for the previous models. The algorithm is as follows:

1. Unfold the three way data X in three ways to form three matrices:

   · Z_1 is the $n \times mp$ matrix formed from the m objects-by-attributes slices.
   · Z_2 is the $m \times np$ matrix formed from the n assessors-by-attributes slices.
   · Z_3 is the $p \times nm$ matrix formed from the m attributes-by-objects slices.

2. Obtain starting values for T and P:

   · T is formed from the first b eigenvectors of Z_1Z_1'.
   · P is formed from the first a eigenvectors of Z_3Z_3'.

3. Q is formed from the first c eigenvectors of $Z_2\,(TT' \otimes PP')\,Z_2'$.

4. P is formed from the first a eigenvectors of $Z_3\,(QQ' \otimes TT')\,Z_3'$.

5. T is formed from the first b eigenvectors of $Z_1\,(QQ' \otimes PP')\,Z_1'$.

6. Repeat steps 3 to 5 until convergence

7. Put $C = T'Z\,(Q \otimes P)$

As before the solutions are unidentified, and the orthogonality of T, Q and P is just for convenience. Note that there is not complete freedom in choosing the dimensions a, b and c: The scores matrix for any mode cannot be estimated if its dimensionality is greater than the product of the dimensionalities in the other two modes. This can be seen in step 3 for example where $T \otimes P$ has dimension $np \times ab$, and so c cannot be greater than ab.

4. RELATIONSHIPS TO OTHER WORK

4.1 Generalised Procrustes Analysis

The Procrustes rotation method discussed in Chapter 7 of this book also models individual differences among assessors and is designed to obtain information about assessors, attributes and samples simultaneously. In fact it can be regarded as a special case of the Tucker-1 common scores model.

Recall that in Section 2.2 we wrote the common scores model as $X_i = TP_i' + E_i$, where T is the matrix of common scores and P_i the individual loadings, found to minimize

$$\sum_{m=1}^{m} \|X_i - TP_i\|^2 .$$

In this case P_i is a general matrix, but if it is forced to be orthogonal, then we can write

$$X_iP_i = T + E_iP_i, \tag{17}$$

i.e. the common scores are found by rotating the original 'configurations' X_i to minimize

$$\sum_{i=1}^{m} \|X_i - T_iP'\|^2 \tag{18}$$

This is the GPA criterion apart from two points: in GPA the dimension of P_i is not usually restricted, and the configurations are translated as well as being rotated. This second point can however be regarded as a standardization, and included in the TWFA model, see later. It is worth recalling at this point that in fitting this TWFA model the fact that the assessors all measure the same variables is not used, as in GPA which is often used for free choice profiling. It can therefore be seen that GPA is simply the common scores Tucker-1 model with the individual loadings constrained to be orthogonal. It should also be mentioned that the isotropic scaling of each assessor used in GPA is already a part of the TWFA model, since W_i always can be multiplied by a constant without changing the model.

The TWFA model is clearly more general than GPA, and so will in general give a better fit. In fact, if the dimensionality is not reduced at all, it will give a perfect fit which is not the case with GPA. We leave a full discussion of GPA to the GPA-chapter, but it is worth considering the following point: In choosing whether to use GPA or TWFA it is obviously necessary to decide whether or not the orthogonal transformation in GPA is sensible. Although it may look unnatural in many cases, certain types of confusion problems can be modelled very well by this transformation, as described in Arnold and Williams (1987). For instance, switching of two attributes by one of the assessors can be accounted for by an orthogonal transformation. This aspect may indicate that GPA is best suited for detecting confusion and scaling problems related to names, definitions etc. (Arnold and Williams (1987)) and TWFA for modelling more general individual differences. Very briefly we can state the following: Procrustes rotation is best suited for detecting errors in the data while TWFA is best suited for modelling individual differences. This may indicate that Procrustes rotation is better suited for situations with untrained assessors and TWFA is best suited for error-free reliable data.

4.2 Individual differences MDS versus TWFA

Consider the common scores and common loadings Tucker-2 model (9). The 'profile' of object j for assessor i, x_{ij}, is the jth row of matrix X_i, the objects-by-attributes matrix for assessor i. This is approximated by f_{ij} where

$$f_{ij} = t_j W_i P' \tag{19}$$

where t_j is the jth row of T. The squared Euclidean distance $D_{ij_1j_2}$ between the approximate profiles of samples j_1 and j_2 for assessor i is

$$\begin{aligned} D_{ij_1j_2} &= (t_{j1} - t_{j2}) W_i P' P W_i' (t_{j1} - t_{j2})' \\ &= (t_{j1} - t_{j2}) W_i W_i' (t_{j1} - t_{j2})' \\ &= (t_{j1} - t_{j2}) V_i (t_{j1} - t_{j2})' \end{aligned} \tag{20}$$

where V_i is a general symmetric matrix. Hence we can write the Tucker-2 model as

$$(x_{ij_1} - x_{ij_2})(x_{ij_1} - x_{ij_2})' = (t_{j_1} - t_{j_2}) V_i (t_{j_1} - t_{j_2})' \tag{21}$$

This is identical to the generalized subjective metrics model for individual differences MDS.

If we consider the PARAFAC model the same way and in addition assume that P and T are orthogonal matrices we obtain

$$D_{ij_1j_2} = (t_{j_1} - t_{j_2}) W_i P' P W_i' (t_{j_1} - t_{j_2}) = (t_{j_1} - t_{j_2}) V_i (t_{j1} - t_{j_2})' \tag{22}$$

where now V_i is diagonal with nonnegative diagonal elements. Therefore we have

$$D_{ij_1j_2} = \sum_{k=1}^{b} (t_{j_1k} - t_{j_2k})^2 v_{ki} \tag{23}$$

which is exactly the INDSCAL model used for individual differences MDS.

The individual differences MDS models are treated in Chapter 6 of this book and will not be considered further here.

Whether there exists a similar analogy between Tucker-3 and an MDS model is not known to us.

4.3 Relations to models for spectroscopy

Above it was mentioned briefly that the PARAFAC model is also used in some chemical spectroscopy examples. The reason for this is that the PARAFAC model is exactly Beer's law for mixtures extended to two dimensions. This kind of model is relevant to, for instance some applications of multivariate chromatography and two dimensional NMR. In such cases, P and T are interpreted as pure spectra for the two dimensions and the W-values are interpreted as the chemical concentrations. In for instance chromatography, T can be interpretated as the time profiles for the

different constituents and the P can be considered as the chemical spectrum matrix of the wavelengths observed.

This type of model has usually been approached by a so-called rank annihilation technique, see Ho et al. (1978). There exist iterative versions of it and direct eigenvector based methods, the so-called GRAM methods (Sanchez and Kowalski (1990)). These methods represent solutions to the general PARAFAC model structure, but they are not least squares solutions as is the classical PARAFAC solution.

The GRAM methods are often applied to calibration problems of two-dimensional instruments. They are particularly useful in cases where the unknown prediction samples contain unknown interferences that were not present in the set of calibration samples. Because of the uniqueness of the different directions, information about the concentrations of the interesting constituents in one particular sample is enough to estimate the concentration for the same constituents in any unknown sample, even if this sample has unknown interferences. The drawback with the technique however is that, at least in its present form, it puts quite strong assumptions on the data, which sometimes can be inadequate.

4.4 Common principal components models

The common loadings Tucker-1 model is closely related to the common principal components model, see Flury (1988) and Krzanowski (1988). This model was developed for the situation where the same variables are measured on different groups of objects, and it is believed that although the group covariance matrices are not equal, they do share common principal axes. This is essentially the same model as the common loadings Tucker-1 model, where although the objects are actually the same for each assessor, this information is not used in the estimation procedure. Flury (1988) gives a maximum likelihood method for estimating the common loadings, and Krzanowski (1988) shows that sensible alternative estimates can be obtained from the eigenvectors of a weighted sum of the individual covariance matrices. If the attributes are standardized within assessors, by subtracting assessor means, this is exactly equivalent to the Tucker-1 solution.

5. DATA PRETREATMENT IN TWFA MODELS

As for most multivariate analyses, centering and scaling of the raw data will affect the results of a TWFA. Therefore it is important that the problems are properly understood by the user of the techniques. Indeed in TWFA, pretreatment can be done in many different ways and so the problem is much more difficult than for standard PCA. In the following we consider the most common pretreatments and discuss the relationships between them.

5.1 Centering

If there is no centering of the raw data then a large proportion of the variation will be due to differences in assessor means and attribute means. These are often considered to be of little interest, and so are removed from the analysis. Two types of centering are usually considered; centering of attributes over all objects and assessors, and centering of attributes for each assessor separately. The first option only standardizes the attributes with respect to mean, and so the analysis will include variation due to differences in assessor mean scores. This is sensible if this kind of difference between assessors is of interest, but more often it is regarded as noise, and so removed from the

analysis by means of the second centering. This has the same effect as the centering in Procrustes rotation, i.e. the elimination of translation effects. It is also equivalent to estimating and removing main effects in the ANOVA model (1).

5.2 Weighting

In addition to standardizing the data by removing variation due to differences in attribute and assessor means, it is often sensible to standardize variation. This can be done by dividing each attribute by its standard deviation, and as with centering there are two options: the standard deviation can be computed over the whole sample or for each assessor separately. As above, the two options have quite different effects on the results. The first option considers each assessor to be using the same scale, so that if he/she uses a smaller part of the scale than the others, he will still after weighting have less influence on the TWFA solution than the rest. In other words, this type of weighting will only have an effect on the relative importance of the different attributes, with no reference to the difference in scale among the different assessors. The second option on the other hand also has an effect on the relative importance of the different assessors by weighting them all equally. In this way, we can say that each assessor is transformed to the same scale. The choice between the two weightings depends on what is believed about the assessors' performance: if it is thought that an assessor will use a large part of the scale if he/she is confident about there being a large difference between the samples, and that a small difference means he/she perceived very little difference, then the weighting should be across all assessors. If on the other hand it is believed that each assessor perceives differences in the same way, and simply chooses to use the scale differently, then the standardization should be done within assessors.

This gives rise to another possible scaling, in which each assessor is given weight proportional to his ability to detect differences among the objects. One way to do this, if there is replication within assessors, is to give each assessor a weight proportional to his average F-value for the different attributes. This could for instance be combined with centering the different attributes within each assessor. Another possibility is to give each assessor and attribute combination a weight proportional to its particular F-value.

6. RELATING THREE-WAY MODELS TO OTHER DATA

Sometimes it is of interest to predict sensory profile data from external measurements. This may be to improve understanding of the sensory data and the individual differences, or to replace the sensory measurements by some fast and reliable instrumental measurement. In the first situation one would typically use chemical or physical measurements, while in the second instrumental measurements such as near infra-red spectral data are often more suitable. In both cases there is a situation as indicated in Figure 3. There is a matrix Y of external information to be related to the individual profiles Z. If the aim is improved understanding of Z it may be of interest to see the relationship between the external measurements and each individual assessor. If the aim is replacement of sensory data by instrumental measurement, prediction of the average score is often more relevant. This can be done by standard multivariate regression techniques such as principal component regression and partial least squares regression, although there are some indications that even in this case improved prediction may be obtained by treating the assessors as individuals, Næs and Kowalski (1989).

The simplest way to use TWFA models to link sensory data with external data is to compute the score matrix T and relate it to the external data Y by some regression technique, i.e.

$$T = BY + E. \tag{24}$$

The matrix T is estimated first, then related to Y to get a relationship between Z and Y. This approach can be used for both prediction and understanding. An alternative which is more goal-oriented and also sometimes easier to compute is to apply the restriction $T = BY$ directly in the factor model. In other words, the restricted matrix $T = BY$ is substituted into the general model $X_i = TW_iP'$ and the parameters W, B and P are optimized by for instance the least squares criterion

$$\sum_{i=1}^{m} \|X_i - TW_iP'\|^2 \tag{25}$$

Writing TW_iP as $(BY + E)W_iP$ with E being the error term in the regression equation $T = BY + E$, we see that the error in the restricted model is the sum of the error in the unrestricted model and EW_iP. The restricted approach certainly represents a more direct and goal-oriented solution to the problem, but because of the more complicated model error structure, it is likely that the unrestricted model better satisfies the usual least squares(LS) requirements of equal variance etc. In practice the Y variables may often be highly collinear. In order to obtain stable solutions they can be replaced by the principal components corresponding to the most interesting information.

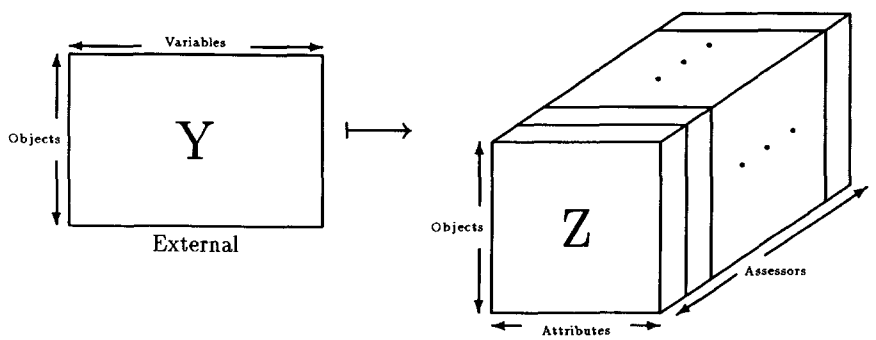


Figure 3: Data setup with external information.

CANDELINC, Carrol et al. (1980) is a method that is designed for optimization of equations like (25). As shown in Carrol et al. (1980), if the W_i's satisfy a PARAFAC or Tucker-2 model, LS optimization can easily be reduced to a minimization of the same type as the unrestricted optimization. In the Tucker-2 model the solution can be found as a simple eigenvector solution (Kloot and Kroonenberg (1985)). Therefore, the restricted approach is solved much more easily than the unrestricted approach.

It should be mentioned that instead of doing any simultaneous modelling of the scores before relating to Y, one could simply relate Y to each of the X_is separately. Using the simultaneous model

is however, a way of obtaining better prediction ability and better interpretation possibilities. As always, if the model is correct, the results are better. If not, they are poorer.

7. HANDLING OF REPLICATES

If there are q replicates for each assessor in the experimental design there are several options. The simplest are averaging over replicates before analysis, using the replicates as extra assessors and using the replicates as extra attributes. The first of these is easiest, but represents a loss of information. It is for example impossible to tell whether an assessor fits badly because he is generating a lot of noise, or because he has a different opinion to the other assessors.

The second approach can be used to distinguish between differences in opinion and noise. After fitting of all the mq 'assessors' one can compare the q replicates for each assessor on an assessor plot. Those of the assessors creating little noise, on the set of variables as a whole, should be close together. If an assessor has a different opinion to the others but is consistent in his view, he should have q replicates close to each other but some distance away from the other assessors. It would be possible to examine a separate assessor plot for several subsets of the variables.

The third option is used to examine which of the attributes are recorded with little noise and which are very noisy, 'averaged' over all assessors. The variable plot should be examined in the same way as the assessor plot above.

The same information on an individual attribute basis can be obtained by ANOVA techniques. For instance one can compute residual errors and F-values for the different attributes and assessors and plot them as advocated in e.g. Næs and Solheim (1991). In this way, assessors' performance for the different attributes can be compared and used to get information about the reliability of each particular assessor.

From the point of fitting the model, taking means over replicates would usually be the most sensible choice. The only point in doing otherwise (apart from the diagnostic reasons given above) would be if there was some useful information in the replicates, e.g. if they represented different orders of tasting and so there was a systematic reason why the replicates should be different. If the only reason for differences between the replicates is noise, then it makes little sense to model this noise and replicates should be averaged over.

8. DETECTION OF OUTLIERS

It is important to realize that the aim of the TWFA models is to look for and describe similarities in structure among the representatives of each mode. For example in the common scores Tucker-1 model it is assumed that each assessor perceives the relationships among the objects in the same way, i.e. that they all regard the same objects as similar and the same ones as different, though they may use different variables to describe these relationships. It is quite possible that for one or more assessors this is not a valid assumption, and the best way to investigate this is to examine the residuals. Any structure in the residuals implies that the model is not adequate, and that the dimensionality is too low in one or more of the modes, or possibly that the data pretreatment was inappropriate. Isolated large residuals however can reveal interesting unusual cases. It is also possible to sum residuals over assessors or attributes or objects to see which fit the model badly.

Note that an assessor, say, who is an outlier on the assessor plot need not have a large residual. This kind of outlier fits in with the model, i.e. perceives the underlying variables and the relationships between the objects, but relates the two in an unusual way. An assessor with a large total residual either does not fit in with the model, or possible generates an unusual amount of noise.

9. MISSING VALUES

In practice when working with large data-sets, there is always a chance that some data will be missing. They could be individual data-points or whole vectors, for instance one whole sample for one particular assessor. There is little advice about what to do about this in the literature, but a few simple solutions are obvious. It should, however, be remembered when using one of these techniques that the solution is always 'wrong', i.e. different from that obtained from a full data matrix. If there are replicates available, and for instance only one of the replicates is missing, a solution to the problem is simply to replace the empty cell by the average of the other replicates. If there are no replicates available, a possible solution is to replace each empty cell by the LS-mean of a main effect ANOVA model. In terms of the model (1) in the introduction, this means that interactions are left out, α_{ik}'s and β_{jk}'s estimated and the missing value is replaced by the corresponding estimate of $\mu_k + \alpha_{ik} + \beta_{jk}$. In a balanced model this is equal to

$$\bar{x}_{\cdot\cdot k} + (\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot k}) + (\bar{x}_{i\cdot k} - \bar{x}_{i\cdot\cdot k}) \tag{26}$$

This is identical to taking the sum of the mean over the assessors and the mean over the samples and subtracting the grand mean.

10. VALIDATION OF THE MODEL

TWFA methods can be seen as purely descriptive ways of examining the data at hand, but sometimes it is useful to know something about whether they have any relevance to other data sets, for example whether the same groupings of samples (or variables or assessors) will appear if other variables (or samples or assessors) are used. Also it is useful to know how much the final model depends on one or two odd observations. One method used for this kind of investigation is cross-validation (Stone, (1974)). Each observation in turn is omitted from the data set, and the model fitted to the remaining data. The residual for the omitted data point is then found. This gives an estimate of how representative of the data set each omitted observation is.

If there is no replication, there are three different ways of doing the cross-validation, corresponding to the three possible definitions of an 'observation', i.e. object, attribute or assessor. These three methods give information on the 'unusualness' of samples, attributes and assessors respectively. Also, if any of these groups can be regarded as a random sample from some population, then the appropriate method can be used to estimate the proportion of the variance of that population that the model would explain. Depending on the model fitted, it is possible to treat one or two (but not all three) of these groups as the observations to be omitted.

The principle is as follows: suppose a Tucker-1 common loadings model has been fitted, i.e. the individual samples-by-attributes matrices X_i have been modelled as $X_i = T_i P' + E_i$, where P is the

common loadings matrix. Since P has orthogonal columns, i.e. $P'P = I$, for any assessor matrix X_i, we can calculate the individual scores matrix T_i as $T_i = X_iP$. Hence the approximation of X_i is $\hat{X}_i = X_iPP'$ and the residuals E_i from this model are $X_i - X_iPP'$. If we now omit assessor z from the data, we can still fit the model, but we will get a different common loadings matrix P_z. We then calculate the residuals E_z for this assessor as $X_z - X_zP_zP_z'$. Usually the squared elements of this matrix are summed, to give the total squared cross validated residual for assessor z. This procedure is repeated for all of the assessors.

If it is desired to omit objects rather than assessors in the cross validation, the procedure is to fit the model as $Y_j = U_jP' + E_j$ where Y_j is the assessor-by-attribute matrix for object j (recall that this gives the same P as previously). The residuals for an omitted object w are then found in the obvious way, as $Y_w - Y_wP_wP_w'$. It is not possible to omit attributes in this model, they can only be cross-validated if one of the other two models is fitted, i.e. common object scores or common assessor scores.

In general it is only possible to cross-validate a group that has not been reduced in dimensionality in the model. Therefore in the common scores-common loadings Tucker-2 model, it is only possible to cross-validate the assessors. The procedure is as follows: model assessor i's objects-by-attributes matrix X_i as $X_i = TW_iP' + E_i$, where T ($n \times a$) are the common scores, P ($p \times b$) are the common loadings and W_i ($a \times b$) is the individual difference matrix for assessor i. Since $T'T = I_a$ and $P'P = I_b$, the residuals for assessor i are $E_i = X_i - TT'X_iPP'$. Hence any assessor can be omitted from the model, the new T and P calculated and the cross-validated residuals found as before. Clearly for the other two possible Tucker-2 models there is only one possible way of cross-validation. Without replication it is not possible to cross-validate a Tucker-3 model.

If there is replication there is a wider choice of validation methods. All of the above methods are available, as is the option of omitting the replicates one at a time. This can be done even for the Tucker-3 model. An alternative is to regard one set of replicates as a test set, fit the model on the other set and find the residuals for the test set.

11. DISCRIMINATION AMONG MODELS

Choosing and validating a model are closely connected, as a poor validation result could lead to the choice of another model. Choice of model refers here to choice of underlying dimensionality. This is a problem that even in standard PCA has no clearcut solution. It can be argued that a PCA or TWFA merely is a low dimensional projection of the data picturing as much variation as possible. Since we can only easily look at two- or three-dimensional plots, we simply choose two or three dimensional models and note how much variation is explained by them. This is how standard PCA is often used. It would however be convenient to have some criteria for the choice of dimensionality. A method commonly used in PCA is a plot of residual variation against number of components, the so-called scree diagram. The 'elbow' or point on this plot where this variation stops decreasing rapidly is chosen as a reasonable dimensionality. Generalizing this to Tucker-1 is straightforward. For Tucker-2 however, there is a different model/dimension for each combination of a and b leading to a 3-dimensional scree diagram, and for Tucker-3 the general scree diagram would become 4-dimensional. It is unfortunately not possible to use separate scree diagrams for each mode as the choice of dimension for one mode effects all of the other modes. In other words two dimensions for the assessor mode may be appropriate if other two modes are also two dimensional, but if the object

mode is then increased to three dimensions it may be necessary to increase the assessor dimension also.

One approach is to restrict the dimensionality according to some other criterion. One possibility is to set $a = b$. This has the consequence that the assessors 'configurations' or fitted values are all linear combinations of each other. This makes TWFA more similar to Generalized Procrustes analysis and may in some cases be helpful. It reduces the scree diagram by one dimension and makes it a practical proposition, although the concept of an 'elbow' in three dimensions is a little difficult.

Any scree diagram can be based on cross-validated residual variance, and there is a tendency for these plots to level out more quickly, and so lower dimensionalities tend to be chosen. This is usually a good thing as there is no benefit in modelling dimensions that are merely noise.

Table 1: The 12 cheeses with the name used in plots.

No	Description	Name
1	Jarlsberg FHS	Jarl_FHS
2	Marks & Spencer Mature	Marks
3	Jarlsberg Lite H30	Jarl_H30
4	Tesco canadian extra-mature	Tesc_mat
5	Norvegia H30	Norv_H30
6	Safeway home produced mild	Safeway
7	Vel-Lagret Norvegia	Norv_Vel
8	Anchor mature	Anchor
9	Norsk Cheddar skorpefri	Cheddar
10	Tesco reduced fat	Tesc_fat
11	Skorpefri F.45 (Norvegia F45)	Norv_F45
12	Tesco mild reduced fat	Tesc_mil

12. ILLUSTRATION BY AN EXAMPLE OF A CHEESE TASTING EXPERIMENT

Twelve cheeses were selected for this study, six Norwegian and six Cheddars. A list of the brand names is given in Table 1. They were assessed by a Norwegian and a Scottish panel, but for this example only the data from the Norwegian panel are considered. Full details of the experiment are given in Hirst et al (1994). The panel consisted of 10 trained assessors. The attributes are given in Table 2. They were scored on a continuous line scale anchored at 1 and 9. The experiment was balanced for order of tasting and session effects. There were two replicates, which have been averaged throughout the example.

12.1 PCA of the cheese data

In order to compare TWFA with more conventional methods a principal components analysis was performed on the objects-by-attributes matrix averaged over all assessors and replicates. The averaged attributes were centered and scaled to zero mean and a standard deviation of unity. In Figures 4(a), (b) and (c) some results from the analysis are presented. The first two principal components explain respectively 71% and 10% of the variation. From the score plot for the first two components it is clear that the panel roughly discriminates the Norwegian from the Cheddar cheeses along the first component, with the exceptions that the Safeway mild cheese seems more 'Norwegian' than Cheddar, and the Norwegian Cheddar is closer to the other Cheddars. The Norwegian Jarlsberg H30 is separated out by the second component, which appears from the loading plot to be a texture component spanning from firmness/grainyness to pastyness/moath coating texture. The Jarlsberg H30 is apparently more firm than the other cheeses. The first principal component includes together with texture properties creamy odour/flavour in one direction, characterizing the Norwegian cheeses, and the remaining flavour/odour properties in the other direction characterizing the Cheddars. The real distinction between the cheeses appears therefore to be that compared to the Cheddars the Norwegian cheeses have a pronounced creamy flavour/odour together with a more rubbery texture.

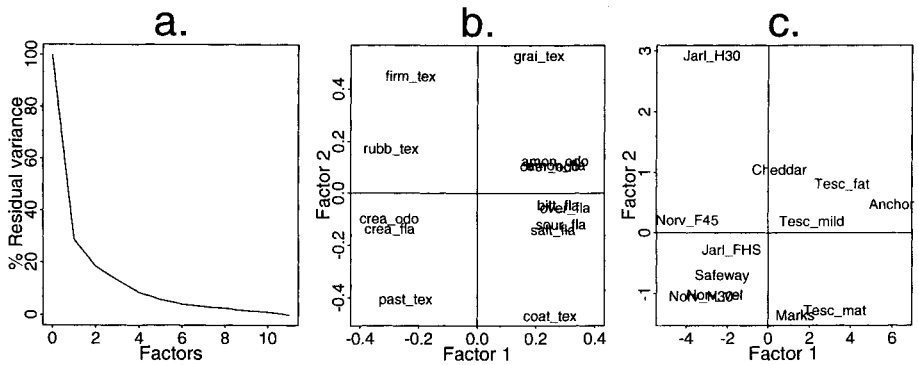


Figure 4: Results from PCA on assessor mean scores. (a) Residual variance, (b) loadings and (c) scores for the first two factors.

12.2 Tucker-1 analysis of the cheese data

As described above three possible approaches can be taken corresponding to the three possible ways of 'unfolding' the three-way data matrix. We will here show some results from the common scores model and the common loadings model: The common scores model assumes that the assessors all perceive the relationships between the cheeses in the same way. This is probably sensible here, though if the Scottish assessors had been included in the analysis this may not have been valid as it is possible that they would perceive different 'underlying cheese types'. The common loadings model assumes the existence of common underlying sensory attributes for cheese which again may be valid for the Norwegian panel but possibly not if the Scottish panel were

included. In both cases the data were pretreated by centering and standardizing all variables within assessors.

Table 2: The 14 common attributes with the names used in plots.

	Description	Name
1	Overall odour	over_odo
2	Creamy/milk odour	crea_odo
3	Ammonia odour	ammo_odo
4	Overall flavour	over_fla
5	Creamy/milk flavour	crea_fla
6	Sour flavour	sour_fla
7	Ammonia flavour	ammo_fla
8	Bitter flavour	bitt_fla
9	Salt flavour	salt_fla
10	Firmness texture	firm_tex
11	Rubbery texture	rubb_tex
12	Pasty texture	past_tex
13	Grainy texture	grai_tex
14	Mouth coating text.	coat_tex

Scores and loadings for the first two factors are plotted in Figures 5(a) and (b) (common scores model) and Figures 6(a) and (b) (common loadings model). First note that the proportion of variation explained by two factors are 51% in the common scores model and 53% in the common loadings model. This demonstrates firstly that the two fits are not the same and more importantly that more variability remains unexplained compared to the mean score PCA of the previous section. This is to be expected as a lot of the variability in the PCA analysis was lost when the assessors were averaged over.

Neither the common scores nor the common loadings plot show great differences from the PCA plots. This indicates that averaging over assessors does not conceal major relationships for this particular data set. However some changes do appear: the Safeway cheddar has moved outside the group of Norwegian cheeses on the second component, and the ammonia flavour/odour has moved upwards along the second component.

The interpretation of a changed position of a sample is that the assessors do not entirely agree on the use of certain attributes. In the mean score PCA the assessors are 'forced' to agree on the attributes as an average value is used, but the common scores model allows the assessors to use the

attributes individually. A similar consideration holds for the change of position of an attribute in the common loadings model. We will return to this in further detail in the section on Tucker-2 modelling.

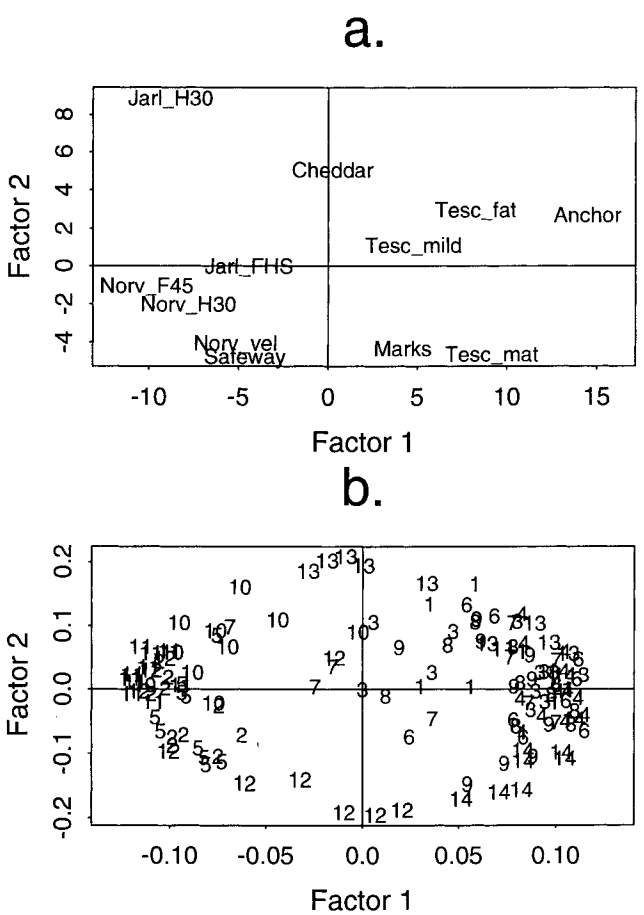


Figure 5: Scores (a) and loadings (b) for the first two factors in the 'common scores' version of the Tucker-1 model. The numbers in the loadings plot (b) refer to the attributes, cf. Table 2.

It is now useful to relate the common scores to the attributes (or common loadings to the objects). One way to do this is to plot all 140 assessor loadings on the common scores plot (Figure 5(b)) (or all 120 assessor scores on the common loadings plot, Figure 6(b)). These plots contain so many points they are almost impossible to interpret, though there are clearly similarities between the assessors. An alternative is to produce a separate plot for each assessor, for whichever model is chosen. Again there is too much detail to be interpreted, though it is highly likely that all assessor plots would be similar. Therefore a Tucker-2 model to investigate both objects and attributes is

sensible. Note that the superposition of the common scores and common loadings plots is not possible as they are the results of different models.

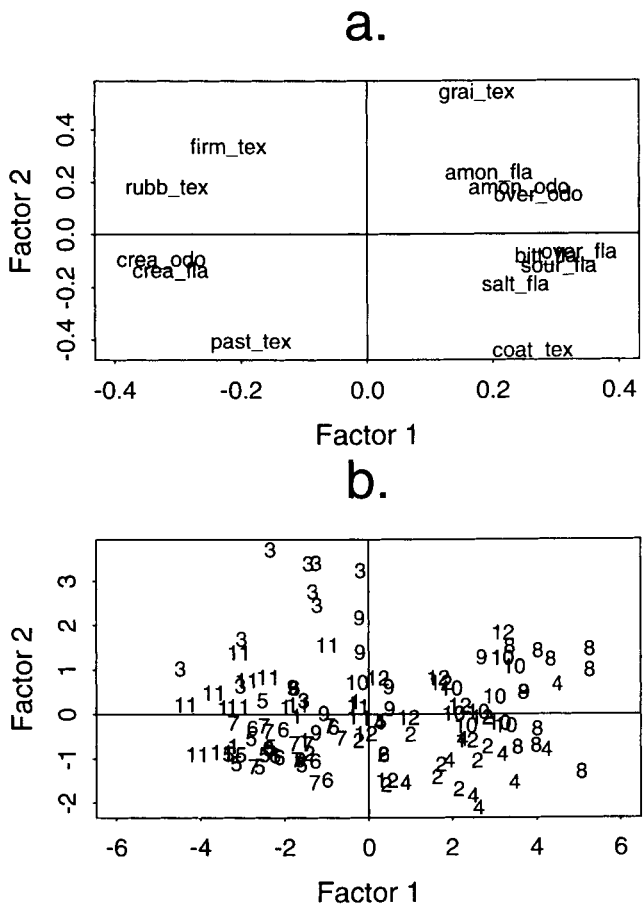


Figure 6: Loadings (a) and scores (b) for the first two factors in the 'common loadings' version of the Tucker-1 model. The numbers in the loading plot (b) refer to the cheeses, cf. Table 1.

12.3 Tucker-2 modelling of the cheese data

As above the data is centred and standardized for each assessor and attribute. A Tucker-2 model with $a = b = 2$, cf. Section 11, was fitted by performing the algorithm of Section 3.2. The amount of variation explained by fitting a model with two factors in both assessor and object modes is 51.1%, approximately the same as for the Tucker-1 models. Note that what we have done is to reduce the object dimension to two as compared with the common loadings Tucker-1 model, which involved no reduction of dimension for the objects (or equivalently the variable dimension has been reduced

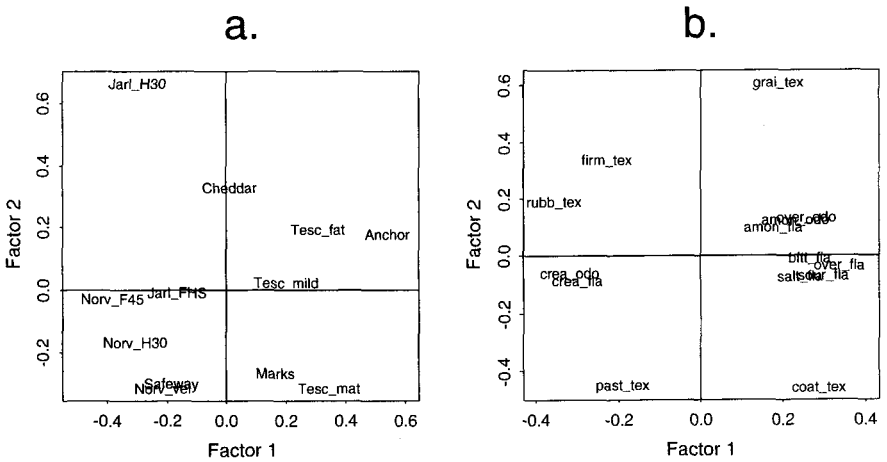


Figure 7: Common scores (a) and common loadings (b) for the two factors in the Tucker-2 model with $a = b = 2$.

compared with the common scores model). The fact that the variance explained hardly changes means that the assumption of two underlying object types is probably valid.

In Figure 7(a) and (b) the common scores and loadings, P and T, are plotted. Again they look very much like those for the Tucker-1 model. The two plots cannot be superimposed, unlike in PCA, as the connection between cheeses and attributes can only be made through the individual 2×2 matrices W_i. These matrices describe how the assessors use the underlying attributes to describe the object types. In order to investigate this further we consider the individual 12×2 scores matrices TW_i, though it would be equally relevant to consider W_iP'. The Figures 8 (a)-(j) show these individual scores plots, which can be directly interpreted together with the common loadings plot, Figure 7(b). The individual scores can be interpreted as the way each assessor places the 12 cheeses in the common attribute space defined by the common loadings. Along the first axis, the component separating Norwegian from Cheddar cheese, the assessors agree to a large extent, and the interpretation of the scores plot from the mean score PCA, Figure 4(b) seems to be valid. There are however differences between the assessors on the second axis. Assessors 3, 4, 7, 8, and 10 rank the cheeses differently to the other 5 assessors on this axis. Recall that the second axis was mainly a texture variable with firmness/graininess at the positive end, and pastiness/moath coating texture at the other. There seem to be two different ways of using these four attributes, maybe due to confusion. In the section on the effect of different pretreatments of the data below we interpret this further. We now proceed to investigate these differences between assessors by Tucker-3 modelling.

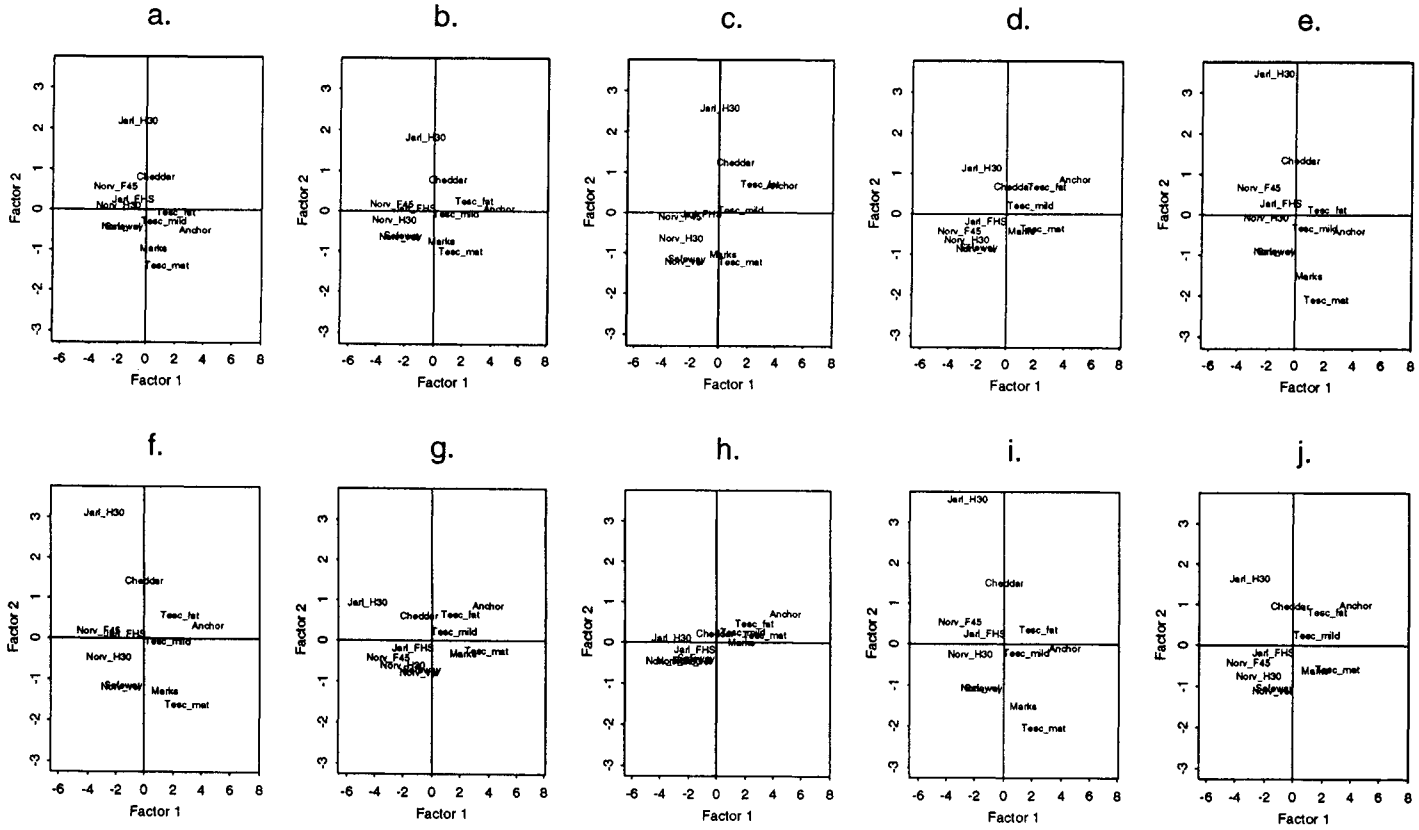


Figure 8: Individual scores for the 10 assessors for the two factors in the Tucker-2 model with $a = b = 2$.

12.4 Tucker-3 modelling of the cheese data

Based on the same centering and standardization as above we used a Tucker-3 model with $a = b = c = 2$, i.e. two components in each of the three modes. This gives a 2×14 common loadings matrix P, a 12×2 common scores matrix T and a 10×2 assessor scores matrix Q, together with the two 2×2 core matrices C_1 and C_2. These scores are plotted in Figures 9(a), (b) and (c).

In the assessor plot Figure 9(c) we can see that the assessors all have similar scores on the first dimension, indicating agreement about the main source of variation in the cheeses, but there is a range of values on the second dimension, indicating considerable disagreement about the less important sources of variation. This difference can be interpreted by examining the core matrices. They are:

C_1	att 1	att 2
sample 1	24.8	-1.1
sample 2	1.1	9.5

C_2	att 1	att 2
sample 1	-1.5	-2.9
sample 2	4.5	3.0

These two matrices represent two assessor types, with each assessor being partly one and partly the other. The first type, C_1, is fairly simple. Sample type *1* is described by attribute type 1, and sample type 2 by attribute type 2. Referring to the sample and attribute plots (Figures 9(a) and (b)) it is clear that sample type 1 represents a Cheddar-Norwegian difference, and sample type 2 seems to separate out the high fat Jarlsberg. Attribute type 1 is a contrast between strong flavours such as bitter, salt and overall flavour, and creamy flavour, and attribute type 2 seems to be a texture variable contrasting sticky and doughy with hard, rubbery and grainy. Assessor type 1 therefore would describe cheddar cheese as being strongly flavoured, compared to Norwegian cheese which is creamy. He/she would distinguish Jarlsberg by its hard and rubbery texture.

Assessor type 2 is more complex. He/she would say that although the strength of flavour is important in distinguishing Cheddar and Norwegian cheese, the texture seems more important and the other way around for the separation of Jarlsberg.

The range of individual core matrices can be investigated by noting that all assessors have a weight of about 0.3 on W_1, but weights from 0.4 (assessors 1 and 3) to -0.6 (assessor 7) on W_2. This pattern, seen in Figure 9(a), do not express any large explanational power of assessor type 2 as compared to type 1, but merely expresses that the assessors have different amounts of the less important assessor type 2 in them. These weights correspond to core matrices ranging from

$$\begin{pmatrix} 6.8 & -1.5 \\ 2.1 & 4.1 \end{pmatrix} \textit{ to } \begin{pmatrix} 8.3 & 1.4 \\ -2.4 & 1.0 \end{pmatrix}$$

First note that as the core matrices C_1 and C_2 listed in the beginning of this section were representing 'ideal' assessor types the two matrices here represent actual assessors. The two matrices both have large values on the upper diagonal, indicating agreement that variation in sample type 1 is largely due to attribute type 1.

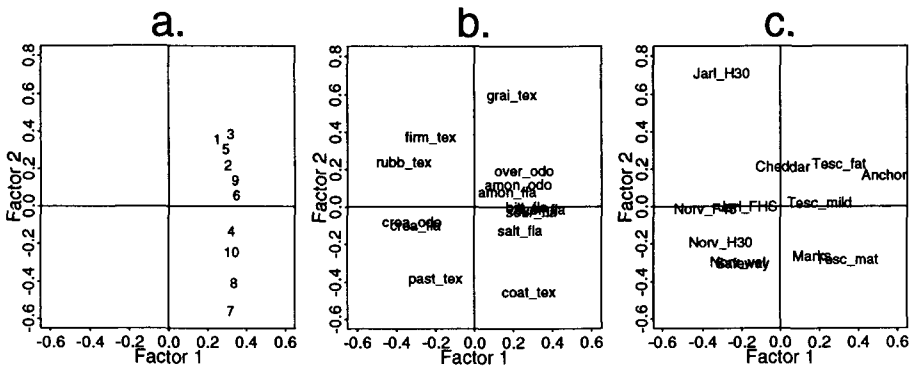


Figure 9: Results from the two-factor Tucker-3 model. (a) Assessor scores, (b) loadings and (c) cheese scores.

The assessors 1 and 3 also have a large value in the lower diagonal indicating that variation in sample type 2 is largely due to attribute type 2, but this is not the case for assessor 7.

Also there is disagreement in how important the other attribute should be in each case. The change of sign indicates a significant difference between the assessors - assessors 1 and 3 think cheddars should have negative scores on the texture variable, i.e. that they are sticky and doughy, represented by the attributes past_tex and coat_tex in Figure 9, whereas assessor 7 would describe them as grainy and hard, corresponding to the positions of rubb_tex, firm_tex and grai_tex in Figure 9. There is a similar difference in the sign of the strength of flavour attribute in describing the Jarlsberg.

The two matrices C_1 and C_2 have the additional useful property, that they give information about the amount of variation explained by each type of assessor mode. This means that the sum of the squares of the four elements of C_1,

$$24.8^2 + (-1.1)^2 + 1.1^2 + 9.5^2 = 708$$

is the amount of variation explained by the Tucker-3 model with $a = b = 2$ and $c = 1$ (Kloot and Kroonenberg(1985)). In an analogous way the sum of the squares of the eight elements of C_1 and C_2,

$$708 + (-1.5)^2 + (-2.9)^2 + 4.5^2 + 3.0^2 = 748$$

is the amount of variation explained by the Tucker-3 model with $a = b = c = 2$. These amounts must be seen relative to the total amount of variation in the data, which due to the pretreatment of the data is a fixed number, determined by the number of assessors, objects and attributes alone. With the standardization in use, the data consists of 140 'variables' (assessors-by-attributes) of 12 observations divided by the standard deviation of these 12 observations. Letting x_{ijk} denote the original data before pretreatment and SD_{ik} the standard deviation of the 12 samples for assessor i and attribute k, the total variance can be found as

$$\sum_{k=1}^{14}\sum_{i=1}^{10}\sum_{j=1}^{12}\frac{(x_{ijk}-\bar{x}...)^2}{SD_{ik}^2}=\sum_{k=1}^{14}\sum_{i=1}^{10}\frac{11SD_{ik}^2}{SD_{ik}}=140\bullet 11=1540.$$

The total percentage of explained variation for the Tucker-3 model with $a = b = c = 2$ is thus 748/1540 = 49%. This is almost the same as for the Tucker-2 model, indicating that two assessor dimensions is probably reasonable.

12.5 Validation and choice of underlying dimensionality

In this example the dimensions of the attributes and samples have been kept the same. There is no particular reason why this should be done, but it does mean that only one dimension needs to be chosen for the Tucker-2 model, and two for the Tucker-3 model. Hence scree diagrams can be constructed.

Consider the Tucker-2 model first. In Figure 10 the accumulated percentage residual variance is plotted together with the same for two different cross-validation principles: assessor-wise, replicate-wise. The replicate-wise cross-validation variance starts to increase from dimension 2. The residual variance and assessor-wise cross-validated residual variance also seems to have leveled off at factor 2, maybe even at factor 1.

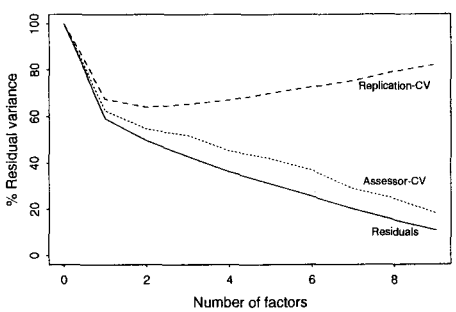


Figure 10: Regular and cross-validated scree plots for Tucker-2 models with $a = b$.

Fixing $a = b = 2$ we now turn to choice of dimension in the Tucker-3 model. In Figure 11 the accumulated percentage residual variance is plotted together with the same for the replicate-wise cross-validation, this being the only cross-validation possible in the Tucker-3 model. This again suggests that a choice of 2 for each dimensionality seems sensible, maybe even only 1 factor is needed, but two is definitely reasonable.

After choosing the dimensionality there are still some validatory tools of interest, as mentioned in section 8. The Figures 12(a), (b) show how well the Tucker-2 model with $a = b = 2$ explains the variation in each attribute and for each assessor. We see that among attributes creamy odour, overall flavour and rubbery texture are best and amonia flavour and salt flavour most poorly explained by the model. The attributes with the highest amount of explained variation are the ones with the most structure related to the cheeses. The actual structure could, however, differ from assessor to assessor. Among the assessors number 1 seems to be poorly described by the model

compared to the others. Looking at assessor number 1's individual score plot, Figure 8(a) we see that number 1 is the assessor with the least spread of the cheeses in the two-dimensional attribute space given by the common loadings. Number 1 is thus the assessor that along the estimated common attribute components is worst at distinguishing between the cheeses.

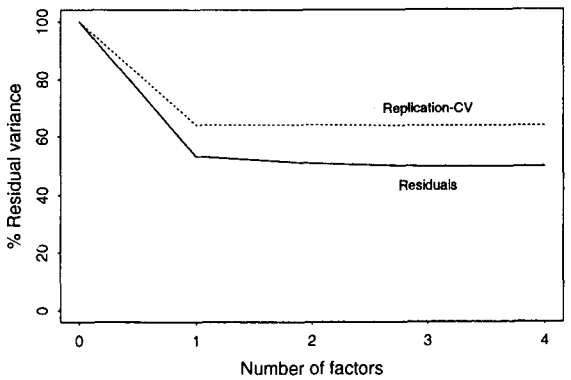


Figure 11: Regular and cross-validated scree plots for Tucker-3 models with $a = b = 2$.

Similar plots could be made cheese-wise, and in the Tucker-1 and 2 cases the assessor-wise and cheese-wise plots might be substituted with plots of corresponding cross-validated variance.

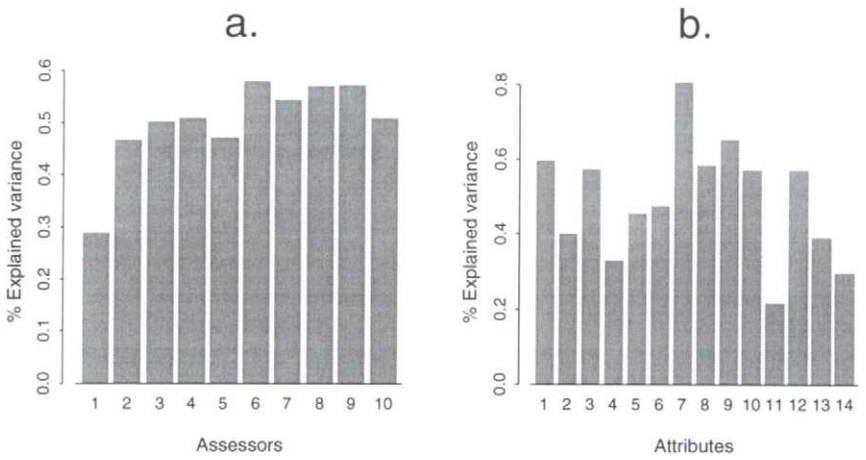


Figure 12: Assessor-wise and attribute-wise relative explained variance by the Tucker-2 model with $a = b = 2$.

12.6 The effect of pretreatment of the data

In the interpretations made above we must bear in mind that each attribute for each assessor was standardized to have unit variance. As discussed in section 5.2 this helps to remove differences in the assessors use of the scale and assumes implicitly that such differences do not express real differences between the cheeses. If however we want to put some emphasis on differences in use of scale two possibilities arise: Firstly the data can be pretreated as above, and then the scale differences investigated by other means. This could be done by estimating a scale parameter for each attribute, eg. the 'stretching and shrinking' values in Næs & Solheim (1991) or the 'maximum likelihood' values in Brockhoff & Skovgaard (1994) together with some kind of plots summarizing the information for all attributes as done in the former of the two mentioned papers. This is, however, a univariate approach to the investigation of scale differences. A multivariate approach could be to choose the second weighting option mentioned in section 5.2, namely to weight each attribute with the inverse standard deviation computed over all assessors, i.e. based on 120 observations. This way the individual scaling differences will be included in the TWFA modelling. We still centre the data for each assessor before weighting, as we do not want to include the differences in assessor levels. With this pretreatment we fitted a Tucker-2 model with $a = b = 2$. Figures 13(a) and (b) show the common scores and loadings.

Comparing with the Tucker-2 model for the former pretreatment, Figures 7(a) and (b), and mean score PCA, Figures 4(b) and (c), we see that the difference between the current Tucker-2 common loadings/scores and the standard PCA loadings/scores are more distinct. This goes together with the fact, that by introducing more individual variability, by allowing the assessors to use different portions of the scale, the standard PCA becomes less representative for a 'typical' assessor.

The individual score plots, Figures 15(a)-(j), show the same patterns as do Figures 8(a)-(j), but the differences between the assessors are more clear. Especially the spread of the cheeses are varying quite a bit now. The spread is directly related to the actual variation in the data for a particular assessor. Note that in the former pretreatment of the data, the variation in the data for the assessors were equal. Figure 14(a) shows how much each of the 10 assessors contributes to the total variation in the data, and we observe that the heights of the bars in Figure 14(a) are directly related to the spread of the cheeses in the individual score plots, Figures 15(a)-(j). The 'directions' in the individual score plots are for the individual assessor determined by the attributes for which he/she has a particular sensitivity. We have documented this by examining the F-statistics from ANOVA's for each assessor and attribute. For example for assessor number 6 the attributes with the four largest F-values are crea_odo, over_odo, crea_fla and rubb_tex. Taking the positions of these four attributes in the common attribute plot Figure 13(a), they span the direction of the individual scores of assessor 6. This tendency is observed for all the individuals.

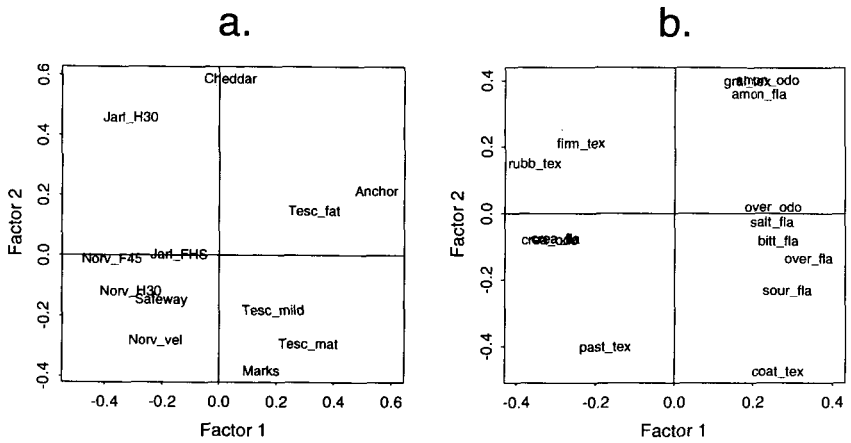


Figure 13: Common scores (a) and common loadings (b) for the two factors in the Tucker-2 model with $a = b = 2$ with the alternative pretreatment of the data.

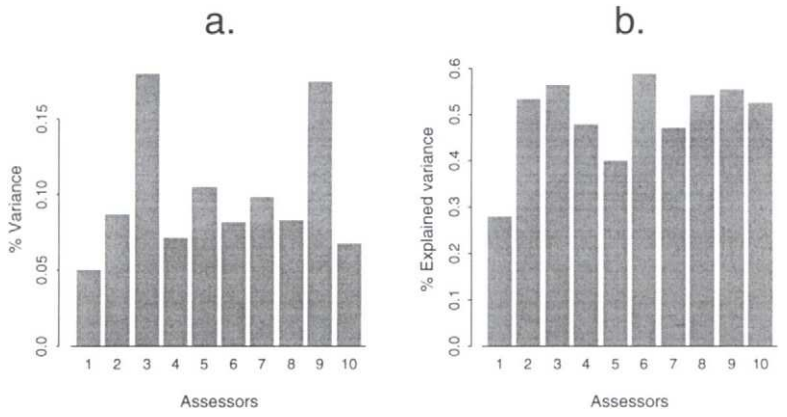


Figure 14: Variability in the alternative pretreated data. (a) Total variability for each assessor, (b) relative explained variance for each assessor by the Tucker-2 model with $a = b = 2$.

13. CONCLUSION

We have presented the concept of TWFA modelling in the setup of sensory profile data. The fitting procedures and interpretations are thoroughly treated in a way that should make it possible for the reader to adopt and apply the methods without further literature search.

From Tucker-1 to Tucker-3 models we have outlined how these models embrace most known multivariate methods of investigating sensory profile data: PCA, GPA, INDSCAL, PARAFAC and 'common principal components'. This generality could be stressed to be both the strength and the weakness of the 'Tucker-approach' we have taken in this chapter. The strength lies in the general principle of not making any model selection errors, when the modelling is started at a sufficiently general level and subsequently letting the data decide which simplifications can be assumed. The weakness comes up due to the substantial number of possible models to fit and investigate, which together with the various data pretreatment approaches requires a considerable task of the analyst. Also formal statistical testing of model simplifications are not performed. Re-sampling methods, such as permutation tests and bootstrapping, definitely has a role to play in that context. We leave this area open here.

In spite of these weaknesses we believe, and have illustrated by the cheese data example, that the TWFA methods as applied here offers additional information and insight in a typical sensory profile data set.

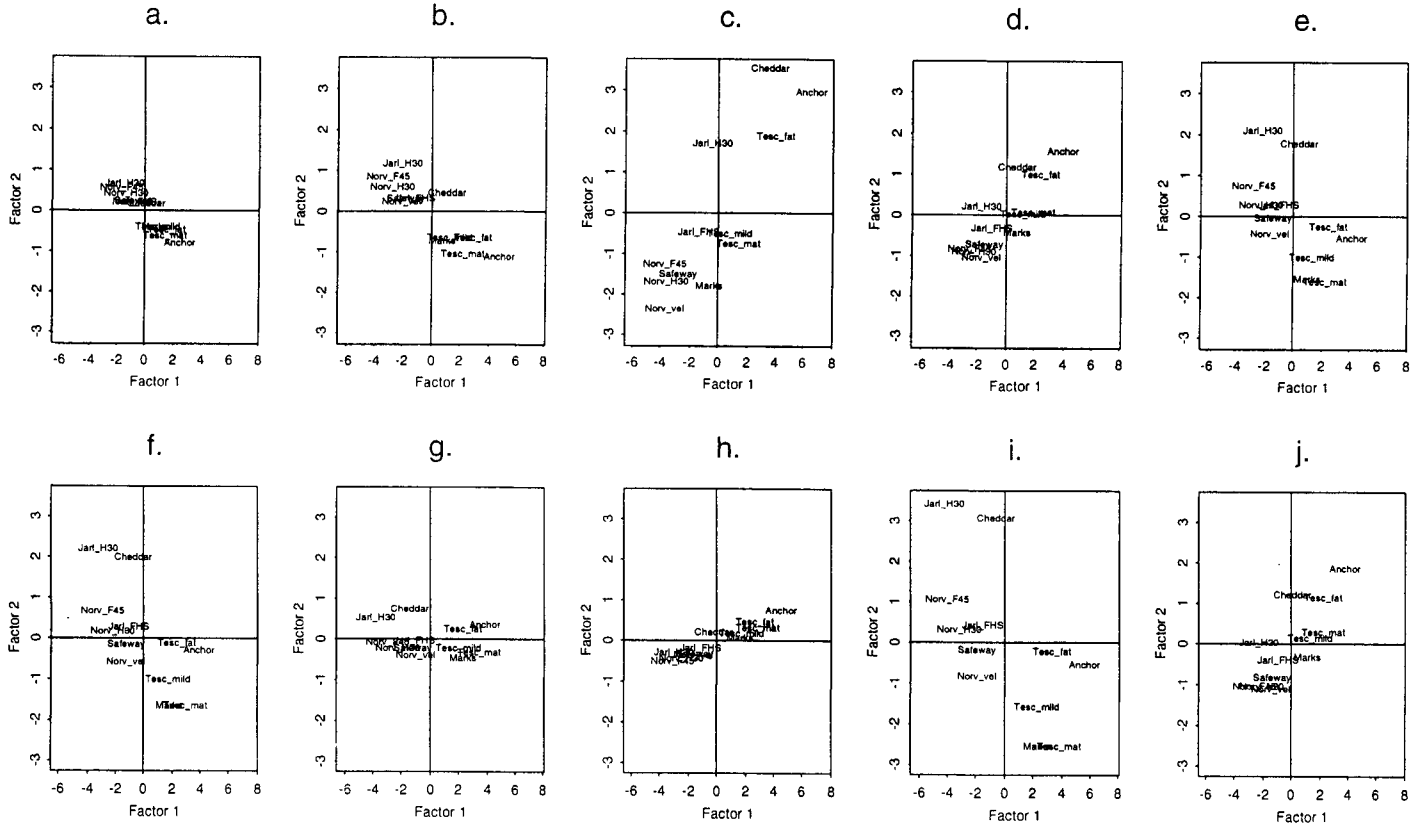


Figure 15: Individual scores for the 10 assessors in the Tucker-2 model with $a = b = 2$ for the alternatively pretreated data

14. REFERENCES

Arnold, G.M. and Williams, A.A. (1986). The use of Generalized Procrustes techniques in sensory analysis. In Statistical Procedures in Food Research. (J.R. Piggott, ed.), Elsevier Applied Science, London.

Brockhoff, P.M. and Skovgaard, I.M. (1994). Modelling individual differences between assessors in sensory evaluations. Food Quality and Preference 5, 215-224.

Carrol, J.D. and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. Psychometrika 53, 283-320.

Carrol, J.D. and Pruzansky, S. (1984). The CANDCOMP-CANDELINC family of models for multidimensional data analysis. In H.G. Law, C.W. Snyder Jr., J.A. Hattie & R.P. McDonald (Eds.), Research methods for multi-mode data analysis (pp. 372-402). New York: Praeger.

Carrol, J.D., Pruzansky, S. and Kruskal, J.B. (1980). CANDELINC: A general approach to multidimensional analysis of many-way arrays with linear constraints on parameters. Psychometrika 43, 3-24.

Flury, B.F. (1988). Common Principal Components & related Multivariate Models. Wiley, New York.

Harshman, R.A. and Lundy, M.E. (1984). The PARAFAC model for three-way factor analysis and multidimensional scaling. In H.G. Law, C.W. Snyder Jr., J.A. Hattie & R.P. McDonald (Eds.), Research methods for multi-mode data analysis (pp. 122-215). New York: Praeger.

Henrion, R., Henrion G. and Onuoha, G.C. (1992). Multi-way principal components analysis of a complex data array resulting from physiochemical characterization of natural waters. Chemometrics and Intelligent Laboratory Systems 16, 87-94.

Hirst, D., Muir, D.D. and Næs, T. (1994). Definition of the organoleptic properties of hard cheese: a collaborative study between Scottish and Norwegian panels. International Dairy Journal 4, 743-761.

Ho, C.N., Christian G.D. and Davidson, E.R. (1978). Application of the method of rank annihilation for quantitative analyses of multi-component fluorescence data from the video fluorometer. Anal. Chem. 50, 1108-1113.

Kloot, W.A. van der and Kroonenberg, P.M. (1985). External analysis with three-mode principal component models. Psychometrika 50(4), 479-494.

Kroonenberg, P.M. and De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 45(1), 69-97.

Kruskal, J.B. (1977). Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. Linear Algebra and its Applications 18, 95-138.

Krzanowski, W.J. (1988). Principles of Multivariate Analysis: A User's Perspective. Clarendon Press, Oxford.

Leurgans, S. and Ross, T. (1992). Multilinear models: Applications in spectroscopy. Statistical Science 7(3), 289-319.

Næs, T. and Kowalski, B. (1989). Predicting sensory profiles from external instrumental measurements, F. Qual. Pref. 4, 135-147.

Næs, T. and Solheim, R. (1991). Detection and interpretation of variation within and between

assessors in sensory profiling. Journal of Sensory Studies 6, 159-177.
Sanchez, E. and Kowalski, B.R. (1990). Tensorial resolution: a direct trilinear decomposition. Journal of Chemometrics 4, 24-45.
Searle, S.R. (1971). Linear Models. John Wiley, New York.
Stone, M. (1974). Cross-validatory choice and assessment of statistical prediction. J. Roy. Stat. Soc. ser. B, 111--133.
Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis Psychometrika 31, 279-311.

Index