Statistical Analysis of Sensory Data

Ph.D. Thesis

Per M. Brockhoff

The Royal Veterinary and Agricultural University
Copenhagen

1994

Contents

Preface

The thesis is a partial fulfillment of the conditions for the Ph.D. degree at the Royal Veterinary and Agricultural University (KVL), Denmark under supervision of Docent Ib Skovgaard. The work was carried out at the Department of Mathematics and Physics, KVL, Division of Statistics, at the University of California, Davis, USA(UCD) and at MATFORSK, the Norwegian Food Research Institute, Ås, Norway. It was supported by the Danish Ministry of Education as part of the FØTEK Programme, the Danish Research Academy and NorFA, the Nordic Research Education Academy.

I am greatly indebted to numerous people for inspiration, help and cooperation: the staff at the Quality Department, MATFORSK who made the cold north seem warm and comfortable, in particular I would like to thank Tormod Næs for offering me some of his time, the staff at the Division of Statistics, UCD, who made the hot south-west seem mild and swallow, in particular I would like to thank Hans-Georg Müller for giving so much interest to my work, the sensory staff at UCD, in particular Barbara Guggenbühl for patient e-mail correspondence, Leif Poll for teaching me basic sensory analysis, Jens Ledet Jensen, David Hirst, Hanne Refsgaard and Stine Møller for fruitful discussions, my office mates over the period, Inge Sandholt, Martin Hansen and Morten Larsen, my colleagues at the Department of Mathematics and Statistics, KVL, and last but far from least Ib Skovgaard for offering guidance, advice and discussion time beyond any possible expectations and my better halves Lene Bruun Hansen for invaluable support in spite of weekly long distance commuting.

Copenhagen, December 1994.

Per M. Brockhoff.

Resumé

Afhandlingen omhandler statistiske og data analytiske problemstillinger i forbindelse med sensoriske eksperimenter. Der gives en kort introduktion til begrebet sensorisk analyse i lyset af statistisk tankegang. I levnedsmiddelvidenskaben anvendes i dag en lang række af klassiske statistiske modeller og metoder. Med udgangspunkt i data fra sensoriske forsøg præsenteres i afhandlingen alternative og supplerende metoder til løsning af flere centrale problemer, og disse sammenholdes med metoder af mere konventionel art.

Metoder til at prædiktere sensoriske vurderinger på basis af kemisk/fysiske målinger er sammenlignet, og det nyligt udviklede 'continuum regression' princip vises at fungere omtrent som den mere velkendte 'partial least squares' teknik. Der foreslås at fortolke prædiktionsfejlen som størrelsen af det sensoriske panel, der skal til for at 'prædiktere' den gennemsnitlige sensoriske vurdering lige så godt.

En model for sensoriske profil data, der explicit modellerer individuelle forskelle både i brug af skala og i reproducibilitet, præsenteres som et alternativ til almindelig variansanalyse uden tilfældige effekter. En generalisering af den foreslåede model til at inkludere tilfældige dommereffekter foreslås som alternativ til den tilsvarende sædvanlige variansanalyse med tilfældige effekter.

Den multivariate generalisering af samme variansanalyse med tilfældige dommereffekter anbefales som et udgangspunkt for at studere korrelationskomponenter i et multivariat sensorisk eller sensorisk-kemisk/fysisk data materiale. Likelihood ratio testet for simultan uafhængighed hen over flere strata er udledt, og en saddelpunktsapproximation til fordelingen af $-2\log Q$ teststørrelsen baseret på Gamma-fordelingen foreslås og vises at give gode resultater.

Inden for rammerne af generaliserede lineære modeller udvikles der en quasi-likelihood inferens procedure for en tilfældig tærskelværdimodel som anvendes i et dosis-respons forsøg med gentagne målinger.

Trevejs principal komponent analyse præsenteres i sensorisk analyse sammenhæng og sammenholdes med andre multivariate teknikker. De vises at bibringe information om de underliggende strukturer i relation til individuelle forskelle i et multivariat sensorisk data materiale.

Endelig gives en kort oversigt over teorien for saddelpunktsapproximation og en Gamma baseret approximation til fordelingen for linearkombinationer af χ^2-fordelte stokastiske variabler vises at have relativ fejl, der konvergerer mod nul i begge haler for ethvert endeligt antal observationer. Der vises endvidere hvordan dette kan udnyttes til at approximere fordelingen af F-teststørrelser i non-standard situationer, der ofte forekommer i sensoriske eksperimenter.

Summary

The thesis addresses statistical and data analytical issues in connection with sensory analysis experiments. A brief introduction to the field of sensory analysis is given with a view towards points of statistical interest. Commonly used sensory experiments produce data of diverse kinds ranging from binary data through ordinal to continuous data, often of multivariate nature. Several classical statistical models and techniques are frequently used in the food research literature. With real sensory data as starting points the thesis presents alternative and supplementary approaches to a number of central problems and relates those to methods of more conventional nature.

Methods to predict sensory perceptions from chemical/physical measurements are compared and the newly developed principle of continuum regression is shown to provide comparable results to the widely used partial least squares regression, and a new way to interpret prediction ability in terms of an equivalent panel size is presented.

A model for sensory profile data explicitly modelling individuals' different use of scale and different reproducibility is presented as an alternative to standard fixed effects analysis of variance, and a random judge effects version is suggested as an alternative to the mixed model analysis of variance.

The multivariate generalization of the mixed model analysis of variance is recommended as a basis for studying sources of correlation in multivariate sensory and/or sensory-chemical/physical data. The likelihood ratio test for overall independence over several strata between sets of variates is formulated, and a saddlepoint approximation to the distribution of the $-2\log Q$ test statistic based on the Gamma distribution is suggested and indicated to give reliable results.

In the framework of generalized linear models a quasi-likelihood inference procedure is proposed for handling a random threshold model for dose-response designs with repeated measurements.

The methods of three-way principal components analysis are presented in the context of sensory profile data, related to other multivariate methods and shown to provide complex information about the underlying structures in relation to the assessors in a multivariate sensory data set.

Finally saddlepoint approximation methods are given a brief review, a Gamma based saddlepoint approximation of the tail probability for linear combinations of chi-square distributed random variables is shown to provide limiting exactness in both tails for any finite number of observations. It is revealed how this can be used for F-testing in non-standard situations frequently occurring in sensory data.

Chapter 1

Introduction

The research field of statistical analysis of sensory data, or shortly sensometrics, will by nature encompass development of statistical methodology and theory. But just as important is the communication and promotion of obtained knowledge and developed methods. In acknowledgment of this, major parts of the present work was written for a target group including as well the sensory analysts and sensometricians as mathematical statisticians.

An overview of the thesis construction is given in this introductory chapter. For the reason of broadening the target group it is natural to begin with a brief clarification of what, why, where and how questions with respect to sensory data.

1.1 Sensory analysis

Broadly the notion of sensory analysis or science stands for any kind of investigation of sensory properties of any imaginable object, where 'sensory' may refer to any of the human senses. In the context of the present work objects are food related only and sensory properties are odour and taste related. As such sensory science is a contribution to flavour research in general. Pangborn (1987) pointed out that advances in the sensory science field as opposed to research in analytical chemistry of volatile components had been modest for the preceding three decades. One reason for this was difficulties with proper handling of the inherent variability of behavioral responses. Since 1987 progress has been done and this thesis can be seen as another step in the direction of overcoming these inherent problems of sensory analysis.

Historically sensory analysis as an independent research discipline has been shaped by the confluences of many fields of scientific research, Moskowitz (1992). One of its origins is psychophysics, the study of the relation between stimulus variables and sensory variables, that again dates back to the nineteenth century with Fechner (1860) being a classical reference, see Pangborn (1981). Fechner's famous law of sensation says that the sensory perception (P) is linearly related to the logarithm of stimulus intensity ($\log I$). Stevens' Power Law, see for instance Norwich (1992), states that independent of sensory property investigated the power relation $P = \text{constant} I^n$ holds. Historically the psychophysical influence has played an important role, for instance by bringing the experimental approach of paired comparisons as a tool for difference testing into the sensory science. Today the psychophysical laws of perception, at least for the sensory science in the food research field, are of less importance. But the psychometrical research continues to offer relevant methodology to sensory and sensometrical science. Also the interest in sensory-instrumental correlations is in line with psychophysical thinking.

The modern flagship of sensory analysis, the Descriptive Analysis, can be traced back in history along another stream. It basically began with an 'expert approach' for product evaluation in food corporations. In the 1950's this evolved to small panels agreeing on a qualitative description of products through taste and discussion sessions, the Flavour Profile, see for instance Stone and Sidel (1985) and references therein. The more quantitative extensions of the Flavour Profile, QDA (Quantitative Descriptive Analysis) and Spectrum were developed in the 1970's, and are today well-established and frequently applied methods. Both can be considered as an 'instrument' to measure, quantify and profile sensory properties in products in a precise and reproducible way. The products may represent various factors under investigation, for instance brands, ingredients, storage time etc. The modern descriptive analysis of a given set of products is performed by having a trained panel of assessors/judges, typical 8-12 persons, evaluate the products one at a time with respect to a number of sensory properties under controlled and designed conditions. The QDA and Spectrum methods differ in the principles set up for panel training and sensory property vocabulary determination. Free Choice Profiling is another common descriptive analysis technique where assessors use vocabulary of their own choice.

Sensory analysis will in the present work stand for an analysis performed with the 'sensory instrument', the trained panel. Apart from general food research benefits such a sensory instrument has obvious opportunities for the food industry in connection with product development and marketing. However, let us once and for all point out the important distinction between sensory analysis and consumer preference studies: a sensory analysis gives basically no information about consumer preferences. Combining sensory analyses with consumer preference studies though, is likely to be a powerful tool in the hand of the product developer and the marketing manager. In this thesis, however, sensory data coming from trained panels are considered only.

1.2 Sensory methodology with a view towards statistics

It is useful to consider sensory data in the light of the major pure measurement scales, as pointed out in McCullagh and Nelder (1989), p. 150: the nominal, ordinal and interval scales. Roughly viewed sensory experiments can be arranged into the following four groups:

1. Nominal categorizing (difference testing)
2. Ranking
3. Ordinal categorizing
4. Line scale scoring.

The numbering represents an increasing level of information but also an increasing difficulty in the task required by the assessor.

Category 1 experiments provide data on nominal scale, and category 2 on ordinal scale. Category 3 data is truly ordinal but in some instances the categories may be sufficiently numerical interpretable to regard the data as interval scale data. The ordinal scale is commonly used as a structured n-point intensity scale, n being relatively small, 7–10, see for instance Amerine *et al.* (1965). Category 4 data ought to be interval scale data, but it may be argued that the underlying psychophysical scale is truly ordinal and hence even line scale data should be considered ordinal. Nevertheless it is today generally accepted to consider also line scale data as interval scale data. A fourth scale, the ratio scale, is also met in sensory category 3 and 4 data, O'Mahony (1986), but in the present work we shall concentrate on binary nominal scale data and interval scale data.

The difference testing, based on classical binomial testing, usually formed as a K-alternative forced choice experiment, see Frijters (1988), has, as mentioned, a long tradition in this field. Repeated difference testings on varying concentrations of a stimulus is also a widely used approach, that is used to determine stimulus threshold concentrations. If more differentiated information is sought often a category 4 experiment is chosen right away and a descriptive analysis is performed, where the actual assessment typically is done by putting a mark on a line segment. Consequently this approach provides multivariate continuous (interval scale) data and the tool box of almost all thinkable statistical methods is potentially relevant. Conventional univariate analysis of variance (ANOVA) for each property, see e.g. Lea *et al.* (1991), is widely used for making inference about the product differences. Multivariate data analysis of almost any kind has been used to explore and analyse sensory profile data, see for instance Piggott (1986).

The consideration of category 1-3 data is important also for the reason that data can always be viewed to be of lower category, decreasing the risk of overrating the level of information. This may be of particular importance in the light of the sensory scale discussion above. Examples of using a lower category approach to data originally considered as interval scale data are seen in McEwan and Schlich (1991) and Hirst and Næs (1994). In the former correspondence analysis is suggested as alternative to principal component analysis for multivariate sensory evaluations, viewing the data in a sense as nominal and in the latter plots based on cumulative ranks are developed, viewing the data as derived ranks.

As for any experimental research the design of experiments is central in sensory analysis. Recently cross-over designs were promoted in the sensory literature, see Schlich (1993), as order and carry-over effects may be likely disturbers in sensory experiments. Response surface techniques are also relevant in this context, see Vuataz (1986). In the present work we do not consider any design problems at all.

The inherent problem of assessor variability has been acknowledged in the food research literature only to a limited extent. For inference about the products in question the assessor variability has to be 'accounted for' properly, and for panel evaluation it is relevant to make inference about the assessors.

Basically the 'assessor problem' in a univariate setting may be viewed as one of repeated measurements, using this expression in a broad sense. Consider the following typical setup that we will return to repeatedly throughout the thesis: A panel consisting of A assessors have assessed P products R times. Let X_{apr} denote a univariate assessment of product p, the r'th time by assessor a. Accepting that the assessors represent some population the natural model for the response may be the two-way analysis of variance model with random assessor effects,

$$X_{apr} = \mu + \nu_p + A_a + B_{ap} + \varepsilon_{apr}, \tag{1.1}$$

where ε_{apr}, A_a and B_{ap} are independent normal distributed random variables,

$$\varepsilon_{apr} \sim N(0, \sigma_E^2), \quad A_a \sim N(0, \sigma_A^2), \quad B_{ap} \sim N(0, \sigma_{AP}^2).$$

Model (1.1) is a simple form of a repeated measures model for this setup, and has as a consequence that observations for the same individual are modelled to be correlated. This classical approach is well-known in the sensory science, although maybe not fully understood and it may go under other names. Below contributions of this thesis will be discussed in the light of the univariate model (1.1).

1.3 Thesis structure overview

The thesis is built up around the five papers:

Brockhoff, P.M., Skovgaard, I.M., Poll, L. and Hansen K. (1993). A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements. *Food Quality and Preference* **4**, 215-222. (Chapter 2)

Brockhoff, P.M. and Skovgaard, I.M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference* **5**, 215-224. (Chapter 3)

Brockhoff, P.M. and Müller, H.G. (1994). Random effect threshold models for dose-response relations with repeated measurements. Submitted to: *J. Royal Stat. Soc. Ser. B.* (Chapter 4)

Brockhoff, P.M. and Guggenbühl, B. (1995). Two-dimensional covariance component models applied to sensory data. To be submitted to: *Journal of Sensory Studies.* (Chapter 5)

Brockhoff, P.M., Hirst D. and Næs, T. (1995). Three-way factor methods in sensory analysis. In: *Multivariate Analysis of Data in Sensory Science*, Ed. T. Næs and E. Riisvik, Elsevier Science Publishers, Amsterdam, The Netherlands. (Chapter 6)

Each paper names a chapter (Chapters 2-6) and the papers are presented in the exact form they appear at their present status. This spans from published papers to late stage working papers, and means that notation is only consistent within chapters, although overlap occurs. However, section numbering were standardized throughout and placement of figures and tables were reorganized to fit the format of the thesis in general. At the beginning of each chapter there is a brief presentation of the status of the paper (December 1994) and of the way it was or is intended to be published. After the presentation each paper begins with a title-page with title and author information together with an abstract, if present, as (to be) published.

At the end of the papers in Chapter 2–5 there are some additional sections with supplementary work and/or discussions, that are of direct relevance for the papers in question. These sections are indicated with **Add:** in the section titles. Chapter 7 contains material written specifically for this thesis. As mentioned most of the material is written also for readers who are not mathematical statisticians, but the following parts are exceptions: Section 3.10, the paper Brockhoff and Müller (1994) of Chapter 4, Section 5.13 and Chapter 7. There is an extensive presentation in the beginning of Chapter 4, though, that aims at less statistically minded readers. It should be noted, however, that some basic knowledge in statistics is needed throughout to capture the messages of this work.

1.4 Thesis contents overview and related work

Often statistical inference is based on approximate distributional assumptions relying on asymptotic results. In sensory analysis the number of replications is typically far from infinity, say $R = 1, 2, 3, 4$, and there is a need for good small sample approximations. Chapter 7 presents the saddlepoint approximation technique as a general tool for approximating densities and distribution functions, see for example Reid (1988). It has in recent years developed in statistics among other things as an alternative to Bartlett correction techniques, see Bartlett (1937). The 'magic formula' of Barndorff-Nielsen (1983) is a saddlepoint approximation formula for the density of the maximum likelihood estimate in the one-parameter case. In Chapter 7 the emphasis is put on the developments of approximations based on the Gamma distribution rather than the Normal distribution, a research direction much influenced by J.L. Jensen, Aarhus University, see e.g. Jensen (1988, 1991). For some cases a Gamma based saddlepoint approximation has particularly nice small sample properties. One situation is the test for (single-stratum) independence between sets of variates in multivariate (mixed) analyses of variance models, see for instance Anderson (1958). A multiple-strata independence test together with a Gamma based saddlepoint approximation of the $-2\log Q$ test statistic is developed in Section 5.13 and applied in Brockhoff and Guggenbühl (1995). Another case is the distribution of linear combinations of χ^2-distributed random variables, where the saddlepoint approximation is an alternative to the classical method due to Satterthwaite (1946). This becomes relevant in sensory analysis whenever the effects of replication is modelled in addition to (1.1) as random effects. A theoretical result of relative limiting exactness in the tails for any sample size is outlined in the latter case, a result that applies to the more general subclass of Gamma convolutions considered in Jensen (1992).

In Brockhoff and Guggenbühl (1995), Chapter 5, the bivariate extension of (1.1) is considered together with the just mentioned random replication effects situation. The paper is to a large extent of expository nature promoting this in statistics well-known but sparsely used model to sensory and sensometrical scientists. This amounts much to a promotion of the 'Danish' ANOVA thinking with the factor structure diagrams of Tjur(1984, 1991) as a lubricating aid. Although it is a study of sensory-instrumental relations, the focus is more on the detection and interpretation of the various sources of correlation rather than on a sensory-predictive approach, as often otherwise the case, see for instance Martens and Martens (1986). The estimation of covariance components is based on the old approach of equating matrices of observed mean squares with the expected dittos, although better methods exist also for multivariate models, see Dempster *et al.* (1981).

The sensory-predictive approach is taken in Brockhoff *et al.* (1993), Chapter 2, where the situation is that of a large number of physical measurements taken for each product in the sense of gas chromatographic spectra. These are used for predicting assessor mean scores on three sensory properties. Facing the classical problem of multicollinearity the problem is embedded into the field of multivariate calibration that is highly connected to chemometrical research, see Næs and Martens (1989). In the present work a comparative study of prediction methods was performed with as much emphasis on the comparison methodology as the comparison itself. The principle of continuum regression, see Stone and Brooks (1990), is introduced as a new method in this context. Being based on a generalized factor selection criterion embracing ordinary linear regression, principal component regression and partial least squares regression as points in a continuum of possible choices, the method is almost bound to work just as well as any of these methods used separately, apart maybe from an over-fitting effect due to the additional estimated continuum parameter. This expected performance of continuum regression is confirmed in the present study. In Kowalski (1990) a stepwise multiple linear regression approach was conjectured to be superior to any of the known biased regression techniques. The present work confirms other work, see Cruciani *et al.* (1992), in that the conventional one-at-a-time cross validation method used in Kowalski (1990) is unsuited to compare these regression techniques; questioning highly the validity of the conclusions in Kowalski (1990). In the light of the sensory scale arbitrariness and the effect of design on overall correlation structure it is a hard task to quantify the predictive ability of the gas chromatographic measurements. It is here suggested to interpret the size of the prediction error in relation to the variability of the sensory panel as a prediction method cannot ever be expected to predict a quantity with higher precision than the 'measurement error' of the quantity.

The papers of Chapter 3, 4 and 6 are all specifically devoted to the 'assessor problem'. It is not attempted to give an inclusive review of data analytical methods in food research literature in connection with sensory panels. Some basic features are, though, the wish to assess the performance of the panel(lists) by measures of reproducibility and sensitivity, and possibly weigh assessors differently for the inference about the products involved. The F-statistics and variance estimates from the individual one-way analyses of variance are common and quite sensible choices of sensitivity and reproducibility measures, see Næs and Solheim (1991). However, using for instance F-statistics for weighted inference about the products a severe selection bias must be contemplated.

The idea underlying e.g. the F-statistic approach is that assessors can differ in three ways, that is, with respect to level, range of scale and variability. The contribution of Brockhoff and Skovgaard (1994), Chapter 3, is the explicit formulation of the statistical model encountering these major ways that assessors can differ. In relation to the repeated measurements discussion above this 'assessor model' can be seen as model (1.1) conditional on the assessors with the generalization that variances may differ, and the restriction that the assessor-product interaction is modelled multiplicatively. The model is a variance heterogeneous version of Mandel's 'bundle of straight lines' model, Mandel (1961). A formal likelihood approach leading to an iterative partial maximization algorithm is performed, and is shown that this leads automatically to a sensitivity weighted inference about the products.

In the additional sections of Chapter 3 the asymptotic results used for the assessor model are thoroughly verified, based on Lehmann (1986). The algorithm is shown to converge to the unique maximum likelihood estimate with a probability tending to 1 as $R \rightarrow \infty$. The convergence properties for finite R is discussed in the light of a straightforward extension of the global algorithm convergence result by Jensen *et al.* (1991). Further a small simulation study was performed to investigate the small sample validity of the used χ^2-approximations. Finally, Chapter 3 is ended by a formulation of a random extension of the 'assessor model', which may then be seen as an alternative to (1.1). This elaborate modelling of individual dissimilarities based on formal statistical thinking is not very common in the sensory science literature.

In Chapter 4 the focus is on sensory thresholds and generalized linear models in the sense of McCullagh and Nelder (1989). Odour thresholds of aroma compounds, say, in strawberries vary extremely from compound to compound, see Larsen and Poll (1992) and to assess the relative importance of the compounds in a food product the concentrations should be seen in relation to the odour/taste thresholds, see Teranishi (1991). A comprehensive presentation is given prior to the paper Brockhoff and Müller (1994) including two things: A brief introduction to generalized linear models and a sensory methodological review of threshold determination. The latter will indicate that the statistical approach can be seen as unifying with respect to threshold definition and determination. In the additional section of Chapter 4 the use of generalized linear models in sensometrics in general is discussed, indicating that the features of such an approach is quite relevant for many sensory situations.

Thresholds are typically determined from binary dose-response experiments. The generalized linear models pop up then since the model (1.1) for the unobservable thresholds leads to a probit regression model for the binary response, see Finney (1971), and in fact a model with random effects. Generalized linear models with random effects are far from as well understood and widely used as their linear equivalents, and the statistical research on the topic is still active, see Breslow and Clayton (1993). The over-dispersion approach also suggested in McCullagh and Nelder (1989) (and in the first edition of 1983) may in some cases be a way to account for random effects.

Brockhoff and Müller (1994) is a contribution to this research area. A marginal based quasi-likelihood procedure is proposed specifically designed to allow for (varying) baseline probabilities, corresponding to correct-by-guessing chances in forced choice experiments, and allowing for general random assessor intercept and error distributions. The important consequence of introducing these models in sensory threshold determination is the acquisition of reliable confidence limits for the estimated thresholds. In the paper the proposed method is shown through simulations to be superior to the maximum likelihood approach of Anderson and Aitkin (1985) with respect to correctness of length of confidence bands.

The final paper, Chapter 6, is of quite different nature than the rest of the thesis. It is a fully multivariate data analytical approach to the typical sensory profile data set. It reflects the fact that when you leave the nice univariate world you tend to loose the detailed modelling and testing principles of classical statistics somewhere between dimension 2 and p. (Chapter 5 indicates that by dimension 2 we are still able to hold on to the principles.) Multivariate analysis of variance with canonical variates, see Krzanowski (1988), is an approach where the principles in general are not lost, and it is indeed suggested in the sensory literature as a tool for discrimination, see Powers and Ware (1986). But still there is a tendency to use the bi-plots of principal components analysis on assessor mean scores for exploratory analysis, see e.g. Piggott and Sharman (1986). Maybe partly due to the fact that canonical variates analysis still does not in its basic setup allow for the individual scale and variability differences modelled in Chapter 3, and the exploratory opportunities of the approach are not so well established.

Apart from the univariate dissimilarities assessors might now also disagree multivariately, for example due to property confusion effects. The method of Generalized Procrustes Analysis, see Gower (1975), is particularly well suited to account for and investigate confusion effects. The three-way factor methods utilized in Chapter 6 represents a general approach to the investigation of multivariate dissimilarities. The methods are developed in psychometrical research, see for instance Tucker (1966), and can be seen to include many other techniques for similar tasks: Generalized Procrustes Analysis, PARAFAC-CANDECOMP models, multidimensional scaling. The paper is expository with discussions of these mentioned relations to other approaches and detailed discussion of interpretations based on a real data example.

Chapter 2

A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements

The paper was published in *Food Quality and Preference*, 1993. This journal has as aim to “bridge the gap between research and application by bringing together authors and readers in marketing, consumer research, sensory science and nutrition, as well as in food research, development and quality assurance.” The journal's coverage includes: “sensory and motivational studies, sensory/instrumental correlation, mathematical modelling in relation to food acceptability.”

The journal is central in what could be called the European sensometrics society, for instance as publisher of contributed papers at the European Sensometrics meetings. The ‘Second Sensometrics Conference’ was held in Edinburgh, September 1994.

Table 2.6 of this paper is here in a corrected version compared with the published one.

A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements

Per Brockhoff, Centre of Food Research,
Ib Skovgaard, Dep. of Mathematics and Physics,
Leif Poll and Keld Hansen, Dep. of Dairy and Food Science,

Royal Veterinary and Agricultural University,
Thorvaldsensvej 40,
DK-1871 Frederiksberg C, Denmark.

(Received 13 April 1993; accepted 13 September 1993)

Abstract

A comparative study of linear methods for prediction of sensory profiles from gas chromatography (GC) measurements was performed. The data used came from an experiment on the effect of storing apples at various oxygen concentrations. Partial least-squares regression and continuum regression showed the best performance, measured by a two-step cross validation principle. The traditional prediction error sum of squares (PRESS) overestimated the predictive ability of a multiple linear regression approach. The quality of the predictions of sensory properties from GC analyses was measured in terms of a 'panel size equivalent'. Thus, the predictions obtained in the present study were as accurate as predictions from an assessor panel consisting of 2 to 6 assessors, depending on the sensory property in question.

Keywords: Continuum regression, cross validation, gas chromatography, linear prediction methods, multiple linear regression, partial least squares, prediction ability, principal component regression, ridge regression, sensory analysis.

2.1 Introduction

The idea of calibrating instrumental measurements with sensory information is of major interest for food research and the food industry. This paper will compare and discuss linear calibration methods based on data from sensory and gas chromatography (GC) analyses of 'Jonagold' apples stored at various O_2-concentrations.

In building up sensory predictive models based on GC measurements of aroma components we face the classical problem of multicollinearity, as these components are closely related. A conventional multiple linear regression approach is not advisable owing to extreme uncertainty about the parameter estimates. Researchers have turned to the biased regression techniques, of which the most common are principal component regression, ridge regression and partial least-squares regression (see Hoerl & Kennard, 1970; Næs & Martens, 1989).

More recently, the approach of continuum regression was proposed by Stone & Brooks (1990). The continuum regression embraces principal component regression, partial least-squares regression and ordinary least-squares multiple linear regression. This is achieved by the introduction of a parameter, varying in a continuum, in a generalized factor selection criterion. The possible values of this parameter represents a continuum of possible regression models with principal component regression and multiple linear regression at the two extremes and partial least-squares regression as an intermediate point.

Sundberg (1993) revealed a close connection between first stage continuum regression and ridge regression, which we have used in this paper to implement a version of continuum regression as described in the Appendix. The continuum regression and its relationships to the well-known methods are briefly outlined in the section on statistical methods. We also include two variants of a multiple linear regression approach where model selection is based on the PRESS statistic from cross validation. This approach was taken by Kowalski (1990), and showed multiple linear regression to be superior to any of the biased techniques for five data sets.

The main objectives of this paper are to compare the predictive abilities of these methods and to find a suitable prediction model for the sensory properties of Jonagold apples. We used two-step cross-validation to evaluate the predictive ability of the various models. The idea of this is to partition the data into, say, three groups, and then in turn treat two of the groups as 'training sets' and one as the 'test set'. This is referred to as the second cross-validation step. The first step refers to the use of cross-validation for model/factor selection in each training set. In this way, we avoid the criterion for the predictive performance being deflated by a selection based on the same criterion.

Finally, we present a new way of reporting this predictive ability of a model based on a given data set from a designed experiment, by reporting the number of assessors needed in a sensory panel to predict the sensory attribute in question equally well.

As the purpose of the present paper is to investigate the statistical aspects of the prediction problem, the description of experimental materials and methods will be brief and partly covered by references to other publications. For the same reason the statistical method is presented in a section by itself.

2.2 Experimental materials & methods

2.2.1 Storage

Apples of the variety 'Jonagold' were harvested on 11 November 1987. Immediately after harvest, the apples were transferred to four containers where the they were stored at 2°C in (1%O_2— 99%N_2), (2%O_2—98%N_2), (4%O_2—96%N_2) and (21%O_2—79%N_2), respectively. Below, only oxygen concentrations are used as a shorthand notation. Two storage periods, of 109 days and 190 days, were considered. Further details on the apples and storage have been given by Hansen *et al.* (1992). After storage, the apples were removed from the containers and allowed to ripen in normal atmosphere at 20°C for up to 40 days (post-storage period).

2.2.2 Analysis of the volatiles

Gas chromatography (GC) measurements of the volatiles were performed by dynamic headspace sampling as described in detail by Poll and Hansen (1990) and Hansen *et al.* (1992). The volatiles produced by the apples were collected 10 or 11 times during the post-storage period by a Poropac trap, eluted with ether and injected to the gas chromatograph. However, as sensory evaluations and GC analyses were not made on the same days, linear interpolations were performed on the GC data to obtain GC values corresponding to the days of sensory evaluation. This seemed reasonable as the time profiles of the GC results were fairly smooth.

As sensory evaluations were performed six times during the post-storage period, the GC data set consists of 48 samples (six times, four storage conditions and two storage periods). For computational reasons, only GC values for the 15 esters producing more than about 1% of the total volatile production were chosen for further statistical analysis.

2.2.3 Sensory analysis

The apples were evaluated by a trained sensory panel of 8–10 assessors six times during the post storage period. Approximately six apples from each treatment was placed in

3-litre glass jars with lids. The apples were covered to prevent assessors from gaining any visual impression of the fruits, and were evaluated by profile analysis, where the assessors were asked to evaluate the smell produced by the apples. The assessors were instructed to give points on a 0–5 point scale for the properties: intensity, green, banana, pineapple, anise, musty and preference. These properties were chosen by the panel on the basis of discussions during the training sessions. For convenience we refer to 'preference' as a sensory attribute as well, although it is not a descriptive property.

2.3 Statistical methods

2.3.1 Analysis of variance

For each of the two storage periods (109 days and 190 days) the sensory data were initially analysed by analysis of variance of the original assessor scores, according to the model

$$\mathbf{E}(\text{score}) = \text{Assessor} + \text{Day} + \text{Treatment} + \text{Day} \times \text{Treatment} \tag{2.1}$$

where we have used a notation indicating the main effects and interaction included. In this model 'Treatment' refers to the storage conditions and 'Day' refers to the post-storage time.

To assess the overall treatment effect on the sensory attribute in question we then tested the reduction to the model

$$\mathbf{E}(\text{score}) = \text{Assessor} + \text{Day} \tag{2.2}$$

by an F-test. Only attributes exhibiting clearly significant treatment effects were considered in the further analyses.

2.3.2 Two-step cross-validation

The 48 samples in X-matrix were divided into three subsets of 16 observations. The three subsets were chosen systematically to be similar with respect to short- or long-term storage, average post-storage and average oxygen treatments. Each subset was then in turn regarded as a test set, and the remaining two subsets as a training set. We denote the three 32×15 dimensional training set matrices by X_1, X_2 and X_3, respectively.

For each of the three training sets, full cross-validation was performed to 'optimize' each of the prediction methods. This means that each of the 32 samples was in turn left out of the estimation and then predicted. The 'optimization' is to be understood in a broad sense including model selection, variable selection and estimation, the details depending on the method being investigated. For example, this kind of cross-validation was used to include or exclude variables in multiple linear regression, and to choose the number of factors in principal component regression and partial least-squares regression.

We denote the mean score for the sensory attribute in question by y_i, where i is the observation number for the data set. Each method leads to a prediction function and consequently to a 'predicted value' $\hat{y}_i$.

For each training set, indexed by $k = 1, 2, 3$, the PRESS statistic is

$$\mathrm{PRESS}_k = \sum_{i=1}^{32} (y_i - \hat{y}_i)^2$$

and we define the mean root prediction error sum of squares for the training sets as

$$\mathrm{MRPRESS1} = \frac{1}{3} \sum_{k=1}^{3} \left(\frac{1}{32} \mathrm{PRESS}_k \right)^{\frac{1}{2}}$$

In a comparison of the methods we want to use error in predicting the test sets, as a measure of predictive ability. To obtain this prediction error the model in question was fitted to all 32 observations in the training set and used for prediction in the test set. Thus, for each test set we obtained a prediction error sum of squares

$$\mathrm{PRESS2}_k = \sum_{i=1}^{16} (y_i - \hat{y}_i)^2$$

leading to the mean root prediction error sum of squares

$$\mathrm{MRPRESS2} = \frac{1}{3} \sum_{k=1}^{3} \left(\frac{1}{16} \mathrm{PRESS2}_k \right)^{\frac{1}{2}}$$

MRPRESS1 gives the cross-validatory index traditionally reported in analyses, averaged over the three partitions. Instead of this, we used MRPRESS2. Apart from giving a more realistic estimate of prediction error, this gave a comparison of the various models for equal conditions. In particular, in a comparison between multiple linear regression (MLR) methods and principal component analysis (PCR) and partial least-square regression (PLS) methods, MRPRESS1 can be expected to favour MLR methods, as both variable selection and number of variables are determined by cross-validation. For PCR and PLS only the number of factors are found by cross-validation.

2.3.3 Linear prediction methods

Assessor mean scores of the sensory attributes were related to GC measurements of flavour volatiles by several methods; these included MLR, PCR, PLS, ridge regression (RR) and continuum regression (CR). Each of these methods obtains a linear function of the GC measurements, predicting sensory attribute.

The 32×15 GC data matrices, the X_k-matrices, were in all applications centred over samples; i.e. the mean of the 32 measurements was subtracted for each of the esters.

Models with standardized as well as original data were used, except in MLR for which scaling makes no difference. For the standardized versions two slightly different standardizations were used. In the standardized PCR and PLS the X_k-matrices, $k = 1, 2, 3$, were column-wise standardized by column standard deviation; i.e. for each X_k-matrix the original (centred) x_{ij}-element was replaced by the standardized element

$$\tilde{x}_{ij} = \frac{x_{ij}}{s_j}$$

where

$$s_j^2 = \frac{1}{31} \sum_{i=1}^{32} x_{ij}^2$$

For RR and CR the column root sums of squares were used to convert X_k to correlation form; i.e the standardized element was

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\mathrm{SS}_j}}$$

where

$$\mathrm{SS}_j = \sum_{i=1}^{32} x_{ij}^2$$

We used the latter standardization for RR and CR in accordance with standard ridge regression (see Hoerl & Kennard, 1970).

For MLR two approaches were applied. By MLR1 we denote a forward selection procedure which at each step includes that variable among the remaining variables giving the smallest cross-validation index (PRESS). This was continued until the PRESS statistic increased. By MLR2 we denote a corresponding backward selection procedure starting with all variables included; this procedure was continued until the last variable was excluded, and then the model with smallest PRESS statistic was chosen.

The prediction of the sensory score by its average, before centring, for the 32 cases in the training set is included for comparison, and is referred to below as the 'constant' prediction method. As this method does not use the GC measurements at all, it provides a reference from which we may see how much is gained by use of the GC data.

2.3.4 Continuum regression

We let X denote a column mean centered $n \times p$-matrix of GC data and $\mathbf{y}$ an n-vector of assessor mean scores for one sensory attribute. The ordinary least-squares predictor (MLR) and any component of PCR and PLS is of the form $\mathbf{c}^t x$, where $\mathbf{c}$ is a p vector of parameter estimates and x is a p vector of GC measurements. For one stage of the CR, the vector $\mathbf{c}$ is chosen to maximize the criterion

$$T_\alpha = (\mathbf{c}^t \mathbf{s})^2 (\mathbf{c}^t S \mathbf{c})^{(2\alpha-1)/(1-\alpha)} \tag{2.3}$$

with $\mathbf{s}= X^t\mathbf{y}$, $S = X^tX$ and $0 \leq \alpha \leq 1$. Here α is the continuum parameter which can be chosen by cross-validation. In Stone & Brooks (1990) it is shown that $\alpha = 0$ corresponds to MLR, $\alpha = \frac{1}{2}$ corresponds to PLS and $\alpha = 1$ corresponds to PCR. Moreover, Stone & Brooks (1990) pointed out that the predictors for these specializations may be referred to as 'canonical correlation', 'canonical covariance' and 'canonical variance' respectively.

Thus CR chooses the 'best' predictor among a huge set of possibilities, including the standard methods, and a pre-decision of which to use is not necessary. If more than one stage of CR is applied, every $\mathbf{c}$ is chosen to maximize the criterion (2.3) under the constraints of being orthogonal to all of the previous chosen $\mathbf{c}$s. However, we consider first stage CR only.

Sundberg (1993) argued that first-stage CR is no worse than the general approach of RR. In the Appendix, the theoretical result by Sundberg relating first-stage CR to RR is given, as this is applied directly in this paper to find the optimal α-value and corresponding predictor $\mathbf{c}$. It should be noted that this implementation of CR only involves a search in the continuum ranging from MLR ($\alpha = 0$) to PLS ($\alpha = \frac{1}{2}$), and the part ranging from PLS to PCR ($\alpha = 1$) is omitted.

2.3.5 Panel size interpretation of prediction ability

When attempting to predict the average score for a certain sensory attribute, say banana flavour, it makes sense to ask how precisely this value is measured. Even a perfect prediction function cannot hit the target without error, owing to the assessor score variability. The variance of the mean score caused by this variability can be estimated from the data by use of model (2.1). Let us consider the given experiment. For each storage time and each combination of ripening times and oxygen concentrations, n assessors judged the banana flavour, n being 8 or 10 depending on storage period. The mean score $\bar{Y}$ for a given treatment and ripening time then has the variance

$$\mathbf{Var}(\bar{Y}) = \frac{1}{n}\sigma^2$$

where σ^2 is the individual assessor score variance estimated as the residual variance in the analysis of variance based on model (2.1). This would be the resulting prediction error for a 'perfect' method exactly hitting the mean μ. More realistically, an unbiased predictor $\hat{\mu}_{\text{pred}}$ with a non-zero variance would result in the mean squared prediction error

$$\mathbf{Var}(\hat{\mu}_{\text{pred}} - \bar{Y}) = \frac{1}{n}\sigma^2 + \mathbf{Var}(\hat{\mu}_{\text{pred}})$$

Let us suppose now that the mean score from the sensory evaluation were to be predicted by another panel with m members assessing apples from the same combination of treatment and storage time. The resulting mean squared prediction error would be

$$\mathbf{Var}(\hat{\mu}_{\text{pred}} - \bar{Y}) = \left(\frac{1}{n} + \frac{1}{m}\right)\sigma^2 \tag{2.4}$$

which may be used as a measuring stick for the prediction errors based on the GC measurements. As n is the number of assessors in the investigation, and σ^2 is estimated from the analysis of variance, m may be determined so that the expression above matches the mean squared prediction error achieved in each case. The result is the number, m, of panel members required to obtain a prediction of the same quality as that based on the GC measurements.

2.4 Results

2.4.1 Analyses of variance

The F and P values for tests of model (2.2) against model (2.1) are listed in Table 2.1. Flavour intensity, banana flavour and preference are seen to be the only attributes showing clear significances for both 109 and 190 days of storage, and consequently only prediction of these from the GC analyses was attempted.

2.4.2 Prediction results for intensity, banana and preference

The 15 esters chosen for predictions are listed in Table 2.2, together with their concentrations and relative headspace distributions at two particular combinations of storage time and treatment. This shows the relative importance in magnitude of the chosen volatiles, and also that a considerable percentage of the concentration of flavour volatiles is not taken into account in the predictions.

Table 2.1: F statistics and P values for the significance of storage treatment effects based on analyses of variance of raw assessor scores for 109 days and 190 days, respectively.

Variable	109 days		190 days	
	F	P	F	P
Intensity	11.34	<0.0001	10.52	<0.0001
Green	0.68	0.83	0.83	0.66
Banana	10.36	<0.0001	11.58	<0.0001
Pineapple	1.11	0.35	2.99	<0.0001
Anise	0.91	0.57	1.34	0.17
Musty	1.36	0.16	1.77	0.033
Preference	5.90	<0.0001	7.48	<0.0001

Table 2.2: Effect of O_2 concentration on 15 flavour volatiles. Production and relative distribution in headspace are values after 190 days in ($1\%O_2$—$99\%N_2$) or ($21\%O_2$—$79\%N_2$) (ambient atmosphere) and post-storage ripening for 7 days in ambient atmosphere at 20°C.

No	Compound	Production of volatiles $(\mu g/(kg \cdot L))$		Relative distribution in headspace (% of total)	
		1 % O_2	21 % O_2	1 % O_2	21 % O_2
1	Propyl acetate	0.02	0.74	0.75	4.62
2	2-methyl-propyl acetate	0.05	0.19	1.38	1.04
3	Propyl propanoate	0.01	0.20	0.18	1.31
4	Butyl acetate	0.21	5.37	5.74	24.09
5	2/3-methylbutyl acetate	1.44	2.57	38.30	14.92
6	Butyl propanoate	0.03	0.85	0.88	4.19
7	Pentyl acetate	0.02	0.21	0.56	1.13
8	Butyl butanoate	0.03	0.65	0.93	2.76
9	Butyl 2/3-methylbutanoate	0.05	0.30	1.42	1.56
10	Hexyl acetate	0.16	3.00	4.27	14.40
11	Propyl hexanoate	0.01	0.17	0.47	1.05
12	Hexyl propanoate	0.01	0.42	0.32	2.18
13	Butyl hexanoate	0.06	0.99	2.03	4.56
14	Hexyl 2/3-methylbutanoate	0.07	0.78	2.35	3.65
15	Hexyl hexanoate	0.03	0.59	1.27	2.53
	Total	3.50	19.85	60.85	83.99

Table 2.3: Prediction results for flavour intensity. Optimization parameter, which is averaged over the three training sets, refers to the number of factors for PCR and PLS, the number of variables for MLR1 and MLR2, the ridge parameter for RR and the continuum parameter for CR.

Model	Standard- ization	MRPRESS1	Optimization parameter	MRPRESS2
PCR	1	0.476	1	0.464
PCR	1/s	0.490	2	0.483
PLS	1	0.474	1	0.458
PLS	1/s	0.493	1.3	0.481
MLR1	1	0.463	2.7	0.466
MLR2	1	0.435	6.3	0.640
RR	1	0.480	17.65	0.460
RR	$1/\sqrt{SS}$	0.489	1.03	0.469
CR	1	0.474	0.62	0.457
CR	$1/\sqrt{SS}$	0.489	0.25	0.472
Constant	1	0.716	0	0.702

The degree of multicollinearity for the 32 × 15 training data matrices, measured by the matrix condition numbers for the corresponding 15 × 15 matrices of cross products, were approximately 996 000, 663 000 and 1 048 000, respectively, for the three training sets. These numbers indicate a high degree of multicollinearity.

The prediction results are presented in Tables 2.3–2.5. In terms of the MRPRESS2 results, the unstandardized PLS and CR performed best, with a slight superiority of the PLS. Also, the number of factors was smallest for the PLS models. This confirms previous PLS experience. In fact, only one factor is needed for the optimal unstandardized PLS. Thus, this single-factor PLS is a special case of the first-stage CR, which is reflected by the smaller MRPRESS1-values for CR as compared to PLS. Nevertheless, we note that for two out of three test sets the prediction errors (MRPRESS2) for PLS were slightly smaller than for CR, reflecting an increase in prediction error as a result of the estimation of an extra parameter. An interesting fact is that the unstandardized models in general perform better than the standardized ones.

The MRPRESS2 values can be compared with the value corresponding to the 'constant' prediction which does not use the GC measurements at all. This shows that even by optimal choice of prediction model, the reduction in standard prediction error is only between 35 and 40 % compared with the most naive choice of prediction — an illustration of the high degree of uncertainty in sensory data of this kind.

Also striking is the fact that if prediction ability were to be judged from the

Table 2.4: Prediction results for banana flavour. Optimization parameter, which is averaged over the three training sets, refers to the number of factors for PCR and PLS, the number of variables for MLR1 and MLR2, the ridge parameter for RR and the continuum parameter for CR.

Model	Standard- ization	MRPRESS1	Optimization parameter	MRPRESS2
PCR	1	0.514	1.3	0.538
PCR	1/s	0.561	1.3	0.577
PLS	1	0.515	1	0.519
PLS	1/s	0.561	1.3	0.578
MLR1	1	0.524	3	0.578
MLR2	1	0.493	5.7	0.743
RR	1	0.526	18.56	0.535
RR	$1/\sqrt{SS}$	0.558	0.89	0.564
CR	1	0.513	0.75	0.525
CR	$1/\sqrt{SS}$	0.555	0.40	0.565
Constant	1	0.858	0	0.855

Table 2.5: Prediction results for preference. Optimization parameter, which is averaged over the three training sets, refers to the number of factors for PCR and PLS, the number of variables for MLR1 and MLR2, the ridge parameter for RR and the continuum parameter for CR.

Model	Standard- ization	MRPRESS1	Optimization parameter	MRPRESS2
PCR	1	0.462	1.3	0.466
PCR	1/s	0.494	1.3	0.479
PLS	1	0.466	1	0.451
PLS	1/s	0.495	1	0.475
MLR1	1	0.469	2.3	0.505
MLR2	1	0.417	9.3	0.938
RR	1	0.474	19.67	0.469
RR	$1/\sqrt{SS}$	0.495	1.17	0.489
CR	1	0.465	0.77	0.456
CR	$1/\sqrt{SS}$	0.490	0.45	0.485
Constant	1	0.751	0	0.735

MRPRESS1 values, the MLR2 approach would clearly be the winner, and MLR1 would be comparable with PLS and CR. However, in MRPRESS2 the MLR1 is clearly worse than PLS and CR, and MLR2 is disastrous. For the preference attribute it even performs much worse than using the training set preference average as test set prediction.

2.4.3 PLS Analysis of banana flavour

In this section, a brief report of our sensory-instrumental study is presented. Figures 2.1–2.4 contain the results of an unstandardized PLS-analysis of banana flavour. We base this presentation on 'full' cross validation on all 48 observations. Variables number 4, 5 and 10 have large loadings (see Table 2.2), consistent with the fact that these are the major components measured in concentration. The scores scores from the first PLS factor seem in particular to separate high and low oxygen treatments. PLS analyses of flavour intensity and preference give almost identical pictures. Figure 2.4 shows the prediction function together with the 48 observations.

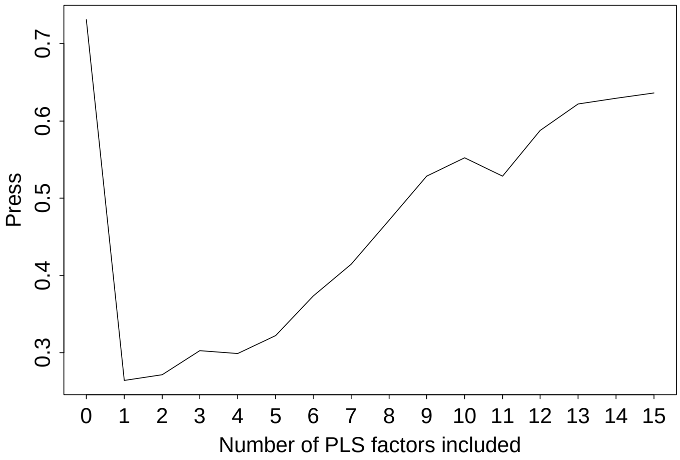

Figure 2.1: Cross-validated residual variance from the set of 48 observations from an unstandardized PLS analysis of banana flavour and the 15 aroma components.

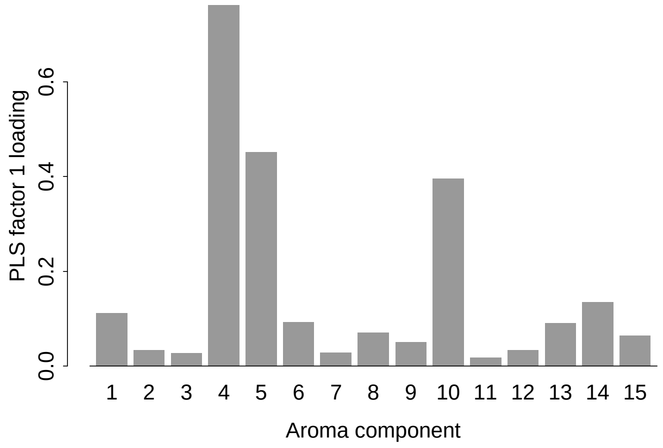

Figure 2.2: Loadings of the 15 aroma components on the first PLS factor from an unstandardized analysis predicting banana flavour.

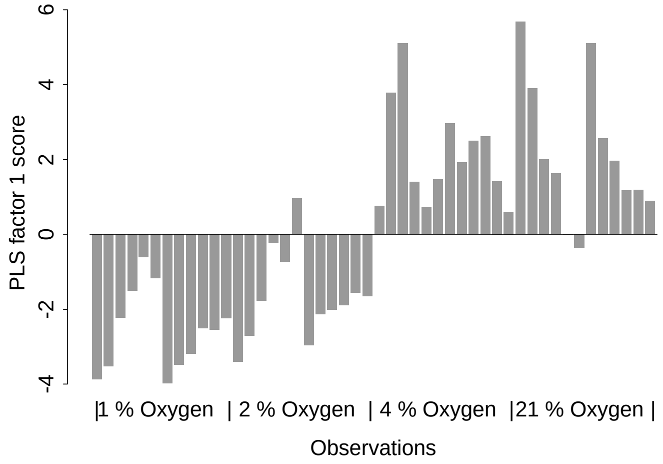

Figure 2.3: Scores relative to the first PLS factor from an unstandardized PLS analysis of banana flavour and the 15 aroma components. The 48 observations are ordered according to the four storage conditions.

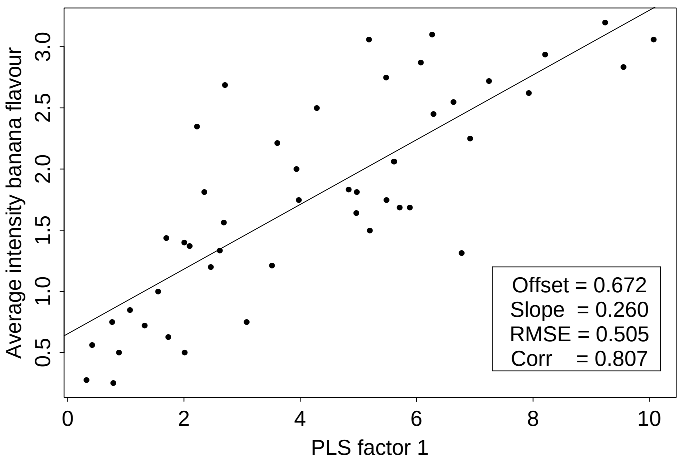

Figure 2.4: Fitted linear relationship between factor 1 and banana flavour from an unstandardized PLS analysis of banana flavour and the 15 aroma components; 72 % of the variation in aroma component space is explained by factor 1.

2.4.4 Panel size interpretation of prediction ability

By substitution of the MRPRESS2 for the unweighted PLS for, say, banana flavour (see Table 2.4), on the left hand side of (2.4) we get

$$\frac{1}{m} = \frac{0.519^2}{\sigma^2} - \frac{1}{n},$$

and the corresponding panel size m can be calculated. The results are summarized in Table 2.6 for the sensory attributes in question and the PLS predictions. We see that the GC measurements lead to predictions comparable with the results from a sensory panel of between two and six assessors.

2.5 Discussion

The important sensory attributes for ‘Jonagold’ apples in the present study were flavour intensity, banana flavour and preference. The PLS and CR methods were shown to provide the best predictions of the three attributes in question, based on GC measurements of 15 aroma volatiles. That PLS performs slightly better than (or

Table 2.6: Panel size interpretation of PLS predictions.

Variable	109 days ($n = 8$)		190 days ($n = 10$)	
	$\hat{\sigma}^2$	Panel-size	$\hat{\sigma}^2$	Panel-size
Intensity	0.46	2.8	0.39	2.4
Banana	0.62	3.3	0.60	3.1
Preference	0.70	6.1	0.66	5.4

similar to) PCR with a small number of factors is in good agreement with previous experience (see Næs & Martens, 1985; Næs *et al.*, 1986; Kowalski, 1990). The CR is new in this context. It is important to note that it was only used here in a very restricted version. Only one factor was used and only the part of the continuum ranging from MLR ($\alpha = 0$) to PLS ($\alpha = \frac{1}{2}$) was examined. This leaves CR in its complete form as a strong candidate among the linear prediction approaches.

There are, of course, several other possible methods. For the given data, logarithmic transformations and MLR models with multiplicative and quadratic terms were tried at an earlier stage of our analysis, without any indications of improved results. These approaches were therefore not pursued further. Non-linear and non-parametric methods were not applied. Such methods have been developed, not only in the statistical literature, but also in chemometric literature (see Cruciani *et al.*, 1992, and references therein). Recently Sutter *et al.* (1992) advocated the selection of factors in PCR based on predictive ability rather than from the top down; an idea already proposed by Næs & Martens (1985).

A different approach to the predictions of sensory profiles is to build up models based on raw data rather than pre-averaging over judges. Næs & Kowalski (1989) present several ways of performing such analyses, including unfoldings and factor models. This approach seems to be very attractive as the interaction between assessors and treatment effects may be included in the modelling.

A two-step cross-validation procedure was applied to evaluate predictive ability. One step was used to 'optimize' a particular method and the other to compare the methods. This approach is different from that of Næs *et al.* (1986), who used a single fixed test set for prediction evaluation. Moreover, Næs *et al.* (1986) did not use cross-validation in their MLR modelling.

It has been shown that the classical PRESS-statistic based on 'full' cross validation (MRPRESS1 values) favours the MLR models. Cruciani *et al.* (1992) reached the similar conclusion that this approach cannot be used to compare regression methods. Nevertheless, Kowalski(1990) used this method to conclude that MLR models give a better prediction than biased methods. This conclusion therefore seems highly questionable.

The generality of the conclusions obtained from a single experimental study, such as the one presented here, may be questioned, of course, and our findings should be seen in conjunction with those presented by other workers. Concerning the experimental conditions, the use of only four containers from which apples were taken at varying times might cause some difficulties in the separation of treatment effects from 'container variation', although the comparison of prediction methods should not be affected systematically by such confounding factors.

Finally, we have presented an interpretation of the predictive ability in terms of the number of assessors required to obtain a prediction of equal accuracy. This approach seems new, and offers a way of reporting and interpreting the results that reflects the underlying idea of replacing sensory evaluations by instrumental measurements.

2.6 Acknowledgment

This work was supported by the Danish Ministry of Education as part of the FØTEK Programme. A number of helpful comments from two referees are gratefully acknowledged.

2.7 References

Cruciani, G. , Baroni, M. , Clementi, S. , Costantino, G. , Riganelli, D. and Skagerberg, B. (1992). Predictive ability of regression models. Part I: Standard deviations of prediction errors (SDEP), *J. Chemometrics*, **6**, 335-346.

Hansen, K. , Poll, L. , Olsen, C.E. , and Lewis, M.J. (1992). The influence of oxygen concentration in storage atmospheres on the post storage of 'Jonagold' apples, *Lebensm.-Wiss. u.-Technol.*, **25**, 457–461.

Hoerl, A.E. and Kennard, W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12** (1), 55-67.

Kowalski, K.G. (1990). On the predictive performance of biased regression methods and multiple linear regression, *Chemometrics and Intelligent Laboratory Systems*, **9**, 177-184.

Næs, T. , Irgens, C. and Martens, H. (1986). Comparison of linear statistical methods for calibration of NIR instruments, *J. R. Statist. Soc. B*, **35** (2), 195-206.

Næs, T. and Kowalski, B. (1989). Predicting sensory profiles from external instrumental measurements, *F. Qual. Pref.*, **4**, 135-147.

Næs, T. and Martens, H. (1989). *Multivariate calibration*, pp. 116-165. Wiley, Chichester.

Næs, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data, *Commun. Statist.-Simula. Computa.*, **14** (3), 545-576.

Poll, L. and Hansen, K. (1990). Reproducibility of headspace analysis of apples and apple juice, *Lebensm.-Wiss. u.-Technol.*, **23**, 481–483.

Stone, M. and Brooks, R.J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression, *J. R. Statist. Soc. B*, **52** (2), 237-269.

Sundberg, R. (1993). Continuum regression and ridge regression, *J. R. Statist. Soc. B*, **55** (3), 653-659.

Sutter, J.M. , Kalivas, J.H. , Lang, P.M. (1992). Which principal components to utilize for principal components regression, *J. Chemometrics*, **6**, 217-225.

2.8 Appendix

The relationship between CR and RR

In the notation from the section on CR, we let $\gamma = \frac{\alpha}{1-\alpha}$. Moreover, we let $b^{CR}(\gamma)$ denote the vector **c** that maximizes (2.3), i.e. the CR estimates corresponding to continuum parameter γ, and let $b^{RR}(\delta)$ denote the RR estimates with ridge parameter δ, i.e.

$$b^{RR}(\delta) = (S + \delta I)^{-1}s, \quad \delta \geq 0$$

where I is the n-dimensional identity matrix. Sundberg (1993) showed that for $0 \leq \gamma < 1$

$$b^{CR}(\gamma) = \left(1 + \frac{\gamma}{1-\gamma}\right) b^{RR}(\delta) \tag{2.5}$$

with

$$\begin{aligned} \delta &= \frac{\gamma}{1-\gamma}\frac{b^{CR}(\gamma)^t S b^{CR}(\gamma)}{b^{CR}(\gamma)^t b^{CR}(\gamma)} \\ &= \frac{\gamma}{1-\gamma}\frac{b^{RR}(\gamma)^t S b^{RR}(\gamma)}{b^{RR}(\gamma)^t b^{RR}(\gamma)} \end{aligned} \tag{2.6}$$

This means that letting the ridge parameter δ vary from 0 to infinity (while calculating $b^{RR}(\delta)$) we obtain from (2.6) the continuum parameter γ varying between zero and one, and from (2.5) the CR estimates $b^{CR}(\gamma)$.

2.9 Add: Panel size interpretation of prediction ability

In this section we consider the uncertainty in the estimation of the panel size equivalent m of a prediction error v. Rewriting (2.4) we can express m as a function of v and σ^2,

$$m(v,\sigma^2) = \left(\frac{v}{\sigma^2} - \frac{1}{n}\right)^{-1} \tag{2.7}$$

Both v and σ^2 are estimated with error in applications. Based on the standard error of estimation of σ^2 from the ANOVA, we will consider the functions

$$m(v,\hat{\sigma}^2), \quad m(v,\hat{\sigma}^2_{Lo}), \quad m(v,\hat{\sigma}^2_{Up})$$

where $\hat{\sigma}^2_{Lo}$ and $\hat{\sigma}^2_{Up}$ are the lower and upper endpoints of the 95% confidence band for σ^2. In Figure 2.5 we see the confidence bands for m given the observed prediction errors as the intersection of the solid vertical line with the dotted curves. The Figure also gives an impression of the additional uncertainty owing to 'moving' the prediction error around the observed value. We see that large values of m will be poorly estimated. This has the consequence in these cases that the prediction error must be quite small for us to claim that the prediction method is better than the panel prediction.

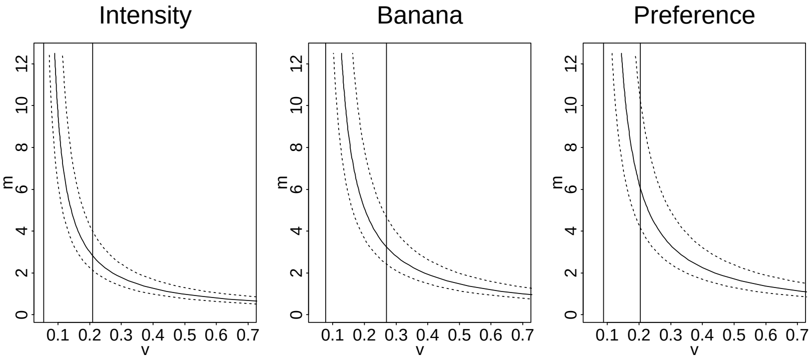

Figure 2.5: Panel size versus prediction error $m(v,\hat{\sigma}^2)$ (109 days). The dotted curves are the lower and upper functions, $m(v,\hat{\sigma}^2_{Lo})$ and $m(v,\hat{\sigma}^2_{Up})$, and the vertical lines indicate the best possible ($\hat{\sigma}^2/n$) and observed PLS prediction error in each case.

Refsgaard *et al.* (1994) calculate panel size equivalents for PLS predictions of geranium leaves taste and odour in a storage experiment of aroma in aquavit. The design structure was basically the same as in the Brockhoff *et al.* (1993) and the calculation of the panel sizes were performed equivalently and yielded 2.5 and 2.0 respectively. However, in the aquavit experiment a 'true' replicate of all treatments were present, and a more reasonable choice for the variance estimate (and uncertainty distribution) may be the Treatment×Storage-time×Assessor interaction mean square (and degrees of freedom). From Figure 2.6 we see that the latter, by disregarding the 'within' variance, gives a little larger equivalent panel sizes. Figure 2.6 moreover shows that small changes in prediction error gives less changes in m than for the apple experiment. Relative to the panel used in each experiment the predictions therefore seem to be better in the apple experiment. We should note, that to compare values along the x-axis of these diagrams from experiment to experiment the sensory scales must be the same. An interesting thing to note also, is that the diagrams can be made from an ANOVA only, and may therefore serve as a comparative tool with respect to what may be expected from a future chemical/physical prediction attempt. The scope of such an approach would need further study, though.

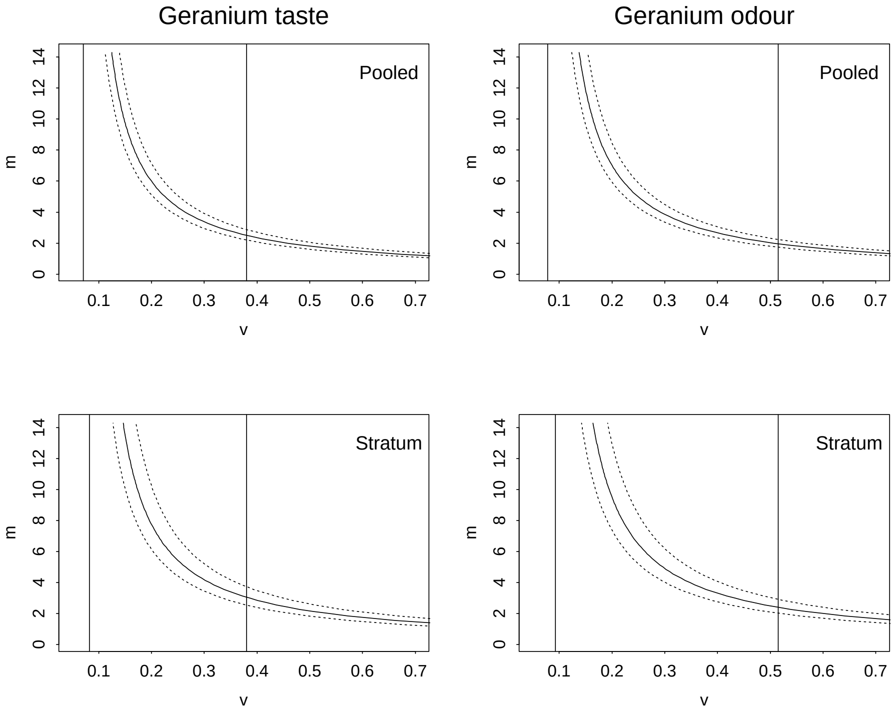

Figure 2.6: Panel size versus prediction error $m(v, \hat{\sigma}^2)$. The dotted curves are the lower and upper functions, $m(v, \hat{\sigma}^2_{Lo})$ and $m(v, \hat{\sigma}^2_{Up})$, and the vertical lines indicate the best possible $(\hat{\sigma}^2/n)$ and observed prediction error in each case. 'Pooled' means that the interaction effect is left unmodelled, and 'Stratum' that the interaction mean square is used as variance estimate.

Chapter 3

Modelling individual differences between assessors in sensory evaluations

The paper was published in *Food Quality and Preference*, 1994.

Modelling individual differences between assessors in sensory evaluations

Per M. Brockhoff
Ib M. Skovgaard

Department of Mathematics and Physics
and Centre of Food Research,
Royal Veterinary and Agricultural University,
Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark

(Received 13 September 1993; accepted 4 February 1994)

Abstract

A parametric model for sensory panel data is presented. The model takes scale differences between assessors into account as well as reproducibility differences. Parameter estimates are derived leading to an iterative partial maximization algorithm, and the scale parameters are shown to be closely related to the 'stretching and shrinking' constants of a 1-dimensional Procrustes analysis. A measure of assessor precision is defined within the model, and the use and interpretation of the model are illustrated by a real data example.

Keywords: Analysis of variance; assessor precision; individual differences; maximum likelihood analysis; pairwise comparisons; reproducibility differences; scale differences; sensitivity weights; sensory analysis.

3.1 Introduction

Sensory panel data, where individuals evaluate different products on a continuous scale, are often blurred by extensive individual variations. The approach to the handling of such differences may affect the results of a statistical analysis considerably. The purpose of the present paper is to present one possible approach to this by means of a particular statistical model, together with theoretical as well as practical issues associated with the model. The emphasis will be on the statistical methodology and applications leaving more theoretical results concerning asymptotics and algorithm convergence to a future publication, as these issues are parts of ongoing work by the authors.

We will consider just one sensory property at a time. Let Y_{apr} denote rth replicate of a score given by assessor a for product p, where $r = 1, \ldots, R_{ap}$, $a = 1, \ldots, A$ and $p = 1, \ldots, P$. Assume for simplicity that $R_{ap} = R$ for all a and p, and that replications are included in the randomized sequence of assessments so that block effects are not necessary in the model.

A straightforward approach for such a situation would be to use the model

$$Y_{apr} = \alpha_a + \nu_p + \varepsilon_{apr}, \qquad \mathrm{Var}\, \varepsilon_{apr} = \sigma^2, \tag{3.1}$$

where $\{\varepsilon_{apr}\}$ are independent random variables. This is the usual model for two-way analysis of variance with additive effects. Different values of the α_as correspond to different basic 'levels of assessment' for the assessors, while the ν_ps represent the product 'values' with respect to the property in question.

In reality, however, results from an assessor panel are often more complex. Even trained assessors differ in their ability to distinguish certain tastes and flavours. This may be reflected in the results in two ways. Firstly, some assessors separate the products by more units on the subjective scale—an interactive effect referred to as 'different use of scale' in what follows. Secondly, the individual variances, measured in terms of replications of the same experimental unit, may vary between assessors. In practice, a higher total variation between the scores from a particular assessor does not imply that he or she has a higher degree of uncertainty and therefore should be given a lower weight, nor does it imply that this assessor spreads the products more out on the scale, because the variation might be unsystematic. An important aspect of the model presented here is that it includes all of the effects mentioned above, i.e. different levels, different uses of scale, different individual variances, and, of course, different product values as in (3.1).

In the following section the model is presented and initially discussed; parameter estimates are then derived prior to a section on successive hypotheses testing. These techniques are then illustrated by an application of the model to real sensory data. Finally, the last two sections consist of suggestions of possible extensions of the model and a discussion of open questions and directions for future research. The more technical parts of the presentation are placed in the Appendix at the end of the paper.

3.2 The model

Let $\{Y_{apr}\}$ denote the assessor scores as in the Introduction. The products are assumed to have some (unknown) values ν_p, $p = 1, \ldots, P$, with respect to the property assessed. The assessors are assumed to score in agreement with these values, but possibly with different baselines, uses of scale, and standard deviations. Thus we assume that for each assessor, a, there are three values: α_a, β_a and $\sigma_a > 0$, such that

$$Y_{apr} = \alpha_a + \beta_a \nu_p + \varepsilon_{apr} \qquad \operatorname{Var} \varepsilon_{apr} = \sigma_a^2 \tag{3.2}$$

where the $\{\varepsilon_{apr}\}$s are independent normally distributed random variables with zero expectations.

For convenience in connection with the estimation of the parameters entering the model we impose the restrictions

$$\bar{\nu} = 0 \qquad MS_\nu = 1 \tag{3.3}$$

where $\bar{\nu}$ and MS_ν are the averages over products of the ν_ps and the ν_p^2s, respectively. These restrictions ensure that the model (3.2) is uniquely parametrized by the triplets $(\alpha_a, \beta_a, \sigma_a)$ and the ν_ps whenever at least two products differ. This parametrization will be used throughout the paper. The number of parameters in the model (3.2) is thus $3A + P - 2$ and we must assume that

$$3A + P - 2 \leq ARP \tag{3.4}$$

which is satisfied if $A \geq 2$, $P \geq 2$ and $R \geq 2$.

A technical problem arises when all the products are alike, i.e. when all the ν_ps are identical. In this case the 'regression parameters', α_a and β_a, are not identifiable, essentially because we are regressing on just one value of the covariate.

3.2.1 Interpretation and related models

Model (3.2) is a submodel of the model

$$Y_{apr} = \gamma_{ap} + \varepsilon_{apr} \qquad \operatorname{Var} \varepsilon_{apr} = \sigma_a^2 \tag{3.5}$$

where $\{\varepsilon_{apr}\}$ are independent, i.e. the model involving general interaction effects between assessors and products, and with different assessor variances. Model (3.5) is thus equivalent to A independent one-way analyses of variance. Model (3.2) can be seen as a linear regression of the scores for each assessor on the unknown product values.

If $\beta_1 = \cdots = \beta_A$ and $\sigma_1^2 = \cdots = \sigma_A^2$ model (3.2) reduces to model (3.1) and it is therefore possible to test these models successively starting from the general interaction model (3.5). We return to this issue in a later section. Note further that the model (3.5) with variance homogeneity, $\sigma_1^2 = \cdots = \sigma_A^2$, is the conventional analysis of variance model, which Næs (1990) has stated to be generally applicable to sensory profile data.

An approach related to the one described here was taken by Yates & Cochran (1938). They interpreted a significant interaction by regressing the interaction effects on factor level means. Three points differ from the approach here: we 'regress' the total product effect rather than just the interaction, our regression is done on 'unknown' product levels incorporated into the model, and finally we allow assessor variances to differ. The second difference arises because Yates and Cochran did not specify a model corresponding to their regressional interpretation.

Mandel (1961) formulated the constant variance case of model (3.2) explicitly. He referred to the model as representing a bundle of straight lines corresponding to the A individuals, and differing from each other in both parameters, α_a and β_a. Further, the model include Tukey's 'one degree of freedom' interaction model as a special case (Tukey, 1949). Later in Mandel (1969; 1971), a more general interaction model, including the counterpart of model (3.2) with variance homogeneity as a special case, was developed, where a number of multiplicative terms were used to model the interaction. Although models were specified in these works, a formal maximum likelihood approach was not taken.

Crowder (1980) suggested the use of *proportional linear models* with a parameter of proportionality for each individual. That model allowed for different variances, but differed from (3.2) by again requiring 'known product values' and identical linear parameters for each individual. The latter requirement, though, could be omitted.

The β-parameters in model (3.2) are closely related to the procrustes-like 'stretching and shrinking' constants used to eliminate scale differences in Næs & Solheim (1991). In the Appendix we show, using a slight reformulation of the definitions that they are related as the slope in a regression of y on x is related to the slope in the inverse regression of x on y.

3.2.2 Assessor precision

Naively we could use the estimates of the individual standard deviations, σ_a, $a = 1, \ldots, A$ as a measure of 'how good' an assessor is. This would, however, favour an assessor scoring consistently within a narrow interval, whether the assessor is able to separate the products or not. By a weighted analysis this assessor might inappropriately be given a heavier weight.

In the model (3.2) such a situation would lead to a low β_a-value for this assessor. Intuitively an appropriate measure of 'assessor precision' could be β_a/σ_a. In the Appendix it is shown that

$$\frac{\beta_a^2}{\sigma_a^2} \propto \frac{E\left(\mathrm{MS}^a_{prod} - \mathrm{MS}^a_{error}\right)}{E\left(\mathrm{MS}^a_{error}\right)}, \tag{3.6}$$

where MS^a_{prod} and MS^a_{error} are the mean squares from a one-way analysis of variance on the results from assessor a, *cf.* the appendix for details. Thus by 'ignoring the expectations' the right-hand side becomes $\mathrm{F}_a - 1$, where F_a is the F test statistic for the product effect in the analysis of variance for assessor a. Examples of application have confirmed that the proportionality

$$c\frac{\hat{\beta}_a^2}{\hat{\sigma}_a^2} \approx F_a - 1 \tag{3.7}$$

is a good approximation for some constant c. This agrees well with the use of the fraction β_a/σ_a as an assessor precision measure, since the measure apparently expresses the assessor's ability to distinguish between the products.

3.3 Estimation

Consider a set $\{y_{apr}\}$ of observations of the form described in the Introduction. We seek maximum likelihood estimates in model (3.2) with the identifiability restrictions (3.3) imposed on the parameters. From the normality assumption the log-likelihood function becomes

$$\log\left\{\prod_{a=1}^{A}\prod_{p=1}^{P}\prod_{r=1}^{R}\frac{1}{\sqrt{2\pi}\sigma_a}\exp\left[-\frac{1}{2\sigma_a^2}\left(y_{apr}-\alpha_a-\beta_a\nu_p\right)^2\right]\right\} = \tag{3.8}$$
$$-\frac{1}{2}APR\log(2\pi) - \frac{1}{2}PR\sum_{a=1}^{A}\log(\sigma_a^2) - \frac{1}{2}\sum_{a=1}^{A}\sum_{p=1}^{P}\sum_{r=1}^{R}\frac{1}{\sigma_a^2}\left(y_{apr}-\alpha_a-\beta_a\nu_p\right)^2$$

with $\bar{\nu} = 0$ and $MS_\nu = 1$.

In general it is not possible to solve the resulting equations analytically and we use instead an iterative approach to obtain the parameter estimates. The iteration scheme derives from the interpretation of the model as regressions on the ν_ps. For given values of $\nu_1, \ldots \nu_p$ satisfying (3.3) the expression (3.8) splits up into A terms, the ath of which corresponds to a usual regression of $y_{a11}, \ldots, y_{aPR}$ on $\nu_1, \ldots, \nu_P$. The corresponding well-known maximum likelihood estimates for the remaining parameters are

$$\begin{aligned}\hat{\beta}_a &= \sum_{p=1}^{P} \bar{y}_{ap\cdot}\nu_p, \\ \hat{\alpha}_a &= \bar{y}_{a\cdot\cdot} \\ \hat{\sigma}_a^2 &= \frac{1}{PR}\sum_{p=1}^{P}\sum_{r=1}^{R}\left(y_{apr} - \hat{\alpha}_a - \hat{\beta}_a\nu_p\right)^2,\end{aligned} \tag{3.9}$$

for $a = 1, \ldots, A$, where, for example, $\bar{y}_{a\cdot\cdot}$ denotes the average of the y_{apr}s over the indices replaced by dots.

Assuming now that the parameters $(\alpha_a, \beta_a, \sigma_a)$ for $a = 1, \ldots, A$ are known we may obtain the maximum likelihood estimates of the ν_ps by noting that we have a set of independent random variables

$$\tilde{Y}_{apr} = \frac{Y_{apr} - \alpha_a}{\beta_a} \sim \mathrm{N}\left(\nu_p, \frac{\sigma_a^2}{\beta_a^2}\right), \tag{3.10}$$

which corresponds to the model of a one-way analysis of variance with different, but known, variances. The maximum likelihood estimate of ν_p is obtained as the weighted mean

$$\hat{\nu}_p = \sum_{a=1}^{A}\frac{\beta_a^2}{\sigma_a^2}\left(\frac{\bar{y}_{ap\cdot} - \alpha_a}{\beta_a}\right) \Bigg/ \sum_{a=1}^{A}\frac{\beta_a^2}{\sigma_a^2}. \tag{3.11}$$

If some of the β_as are zero it is easily seen that the scores from the corresponding assessors do not contribute to the estimates of the ν_ps, and (3.11) still holds, cancelling out the β_as in the denominators, of course. Note that the weights in (3.11) are the squared assessor precisions. Moreover, note that this 'one-way ANOVA step' does not guarantee that the restrictions $\bar{\nu} = 0$ and $MS_\nu = 1$ are fulfilled. This is achieved by a subsequent reparametrization as shown in the iteration scheme below. Missing observations cause no problems here — all sums should just be considered as sums over indices corresponding to non-missing observations.

We may now set up an iteration scheme alternating between the regression step from (3.9) and the estimation of the ν_ps from (3.11), thus arriving at the procedure involving the following steps

1. Initialize the ν_ps by letting $\hat{\nu}_p^{(0)} = (\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})/\sqrt{\sum_{p=1}^{P}(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})^2}$ for $p = 1, \ldots, P$.

2. Estimate $(\alpha_a, \beta_a, \sigma_a^2)$ for $a = 1, \ldots, A$ according to (3.9).

3. Estimate ν_p according to (3.11), for $p = 1, \ldots, P$.

4. Modify the ν_p estimates by first subtracting $\bar{\nu}$ from the νs, and then dividing the result by $\sqrt{\sum_{p=1}^{P}(\nu_p - \bar{\nu})^2}$

   Repeat steps 2–4 until convergence is achieved.

Step 4 is the reparametrization required for the restrictions (3.3) to be fulfilled. If this algorithm converges we have found a local, possible global, maximum of the likelihood function. It does not seem possible to prove the existence of a unique maximum of (3.8) in general, but it can be shown that by use of this algorithm we obtain a consistent and asymptotically efficient estimator of the entire parameter vector, despite the fact this has not been proved to correspond to a global maximum of the likelihood function for fixed R. The proof of this, however, will not be given here.

In Crowder (1980) the parameter estimation is approached in a slightly different way, that here would correspond to consideration of the likelihood function as a function of the triplets $(\alpha_a, \beta_a, \sigma_a)$, $a = 1, \ldots$,A, obtained by substitution of (3.11) in the likelihood function. Then an optimization algorithm is applied to the reduced function. Our approach offers a simple algorithm, and by the explicit description of this, consistency and asymptotic efficiency of the solutions can be proved, as mentioned above.

In the case of constant variances, $\sigma_1^2 = \cdots = \sigma_A^2$, the steps 1 and 2 of the algorithm correspond to the estimation method originally suggested by Mandel (1961) with the modification that we have that $MS_\nu = 1$.

3.4 Successive hypothesis testing

As mentioned, model (3.2) suggested in this paper is a submodel of the general interaction model (3.5) with different assessor variances. For inferential purposes it is useful to consider some further specializations as given in the following nested sequence of models for the expectations:

$$\begin{aligned}
\mathbf{M}_0 :&\quad \mathbf{E}Y_{apr} = \gamma_{ap}\\
\mathbf{M}_1 :&\quad \mathbf{E}Y_{apr} = \alpha_a + \beta_a\nu_p\\
\mathbf{M}_2 :&\quad \mathbf{E}Y_{apr} = \alpha_a + \nu_p\\
\mathbf{M}_3 :&\quad \mathbf{E}Y_{apr} = \alpha_a
\end{aligned}$$

where $\mathbf{M}_0$ is model (3.5) and $\mathbf{M}_1$ is model (3.2). For each of the models we assume that the Y_{apr}s are independent, normally distributed random variables with variances, σ_a^2, depending on the assessor. The hypothesis of equal variances could be invoked at any level; though a natural starting point for reducing model complexity would be to test this hypothesis by a Bartlett test under $\mathbf{M}_0$. In the following we allow for different variances throughout but the analysis with variance homogeneity follows the same path.

A logical way to proceed with the analysis is to test the various hypotheses from M_0 successively down to M_3. For example, to test M_1 against M_0 we use minus twice the log likelihood ratio test statistic:

$$-2\log \mathrm{Q} = PR\sum_{a=1}^{A} \log\left(\frac{\hat{\sigma}^2_{(1),a}}{\hat{\sigma}^2_{(0),a}}\right) \tag{3.12}$$

where

$$\hat{\sigma}^2_{(\ell),a} = \frac{1}{PR}\sum_{p=1}^{P}\sum_{r=1}^{R}\left(y_{apr} - \hat{y}^{(\ell)}_{apr}\right)^2 \tag{3.13}$$

for $\ell = 0, 1$ is the maximum likelihood estimate of the variance for assessor a in model M_ℓ based on the predicted values $\hat{y}^{(\ell)}_{apr}$ from this model. Asymptotically, if M_1 holds, this test statistic follows a χ^2 distribution with $AP - (2A + P - 2)$ degrees of freedom. Because all of the models considered belong to the so-called class of curved exponential families (see Barndorff-Nielsen, 1978), the proof of this assertion follows standard methods once the consistency and efficiency of the estimator has been established.

In a similar way, the hypotheses M_2 and M_3 may be tested. However, special care must be taken with regard to the case when the products are alike. In that case the slopes, β_a, are not identifiable in M_1 and the usual asymptotic results, leading to approximate χ^2 tests for the likelihood tests, do not hold. As soon as some of the ν_ps differ this problem disappears theoretically, but the approximation may be suspected to be poor when we are close to the situation of equal products.

A practically feasible way of avoiding this problem is first to test for product differences by testing M_3 directly against M_0. If product differences are not convincing the entire modelling becomes less relevant, and otherwise we may proceed by testing the various models successively by use of the asymptotic results for likelihood inference.

Since the identifiability problem only occurs when all the products are alike, we are able to perform pairwise comparisons of the products based on model M_1. A pairwise comparison of products p_1 and p_2 is a test of the model

$$\tilde{\mathbf{M}}_2 : \mathbf{E}Y_{apr} = \alpha_a + \beta_a \nu_p \ , \nu_{p_1} = \nu_{p_2}$$

against the model $\mathbf{M}_1$. Estimation of parameters in $\tilde{\mathbf{M}}_2$ proceeds as for $\mathbf{M}_1$ with the only modification that in step 3 of the algorithm the estimates for ν_{p_1} and ν_{p_2} are obtained by substitution of $\bar{y}_{ap\cdot}$ in (3.11) by the average score over both of the products. As test statistic we use minus twice the log-likelihood ratio statistic as above. If $\tilde{\mathbf{M}}_2$ holds this test statistic follows asymptotically a χ^2-distribution with one degree of freedom.

3.5 Example

In a sensory evaluation, five varieties ($P = 5$) of the same food product were compared. Different recipes were used to produce the five products and they were compared by a trained sensory panel consisting of eight assessors ($A = 8$) giving scores for a number of different sensory properties on an intensity scale from 0 to 1. Only one of these properties is considered here.

The data were kindly provided by Dansk Sensorik Center[1]. In the Appendix a raw data table is given containing the 160 original figures. For reasons of confidentiality we can not reveal the type of food and the properties being assessed.

Each product was assessed by each assessor four times by means of four replicates (R=4). Evaluations were 'blind' and the order of presentation of products randomized within each replication. Although the replicates were organized as four different sessions (blocks), we assume for now that no block effect is present. In the Extension section below we outline how to handle the case with block effects. The observations were continuous in the sense that the scores were given as marks on a line segment.

In Figure 3.1 the 160 observations are presented. Each vertical line resembles the line segment, scaled from 0 to 1, used for scoring by each assessor. There seems to be a general tendency that the products 1 and 4 are judged to have a high intensity of the property in question, while products 2, 3 and 5 have low intensities. There is, however, a considerable variation between the assessors in the way they evaluate the products. To investigate how model (3.2) reflects these variations, note for now the following assessor characteristics:

1. Assessor 3 and 4 are similar with respect to use of scale and individual variation, but have quite different levels.

2. Assessor 4 and 6 are similar with respect to individual variation, but use the scale differently.

3. Assessor 7 seems to use the scale similar to assessor 6, but clearly has a larger variation.

4. Assessor 8 seems to have turned the scale around compared to the others.

This is clearly not a complete list of information on assessor differences seen from Figure 3.1, but it exemplifies the different types of assessor effects mentioned in the Introduction, including the 'scale reversal' by assessor 8.

[1] Dansk Sensorik Center is a danish company specialized in performing and selling sensory evaluations to the food industry

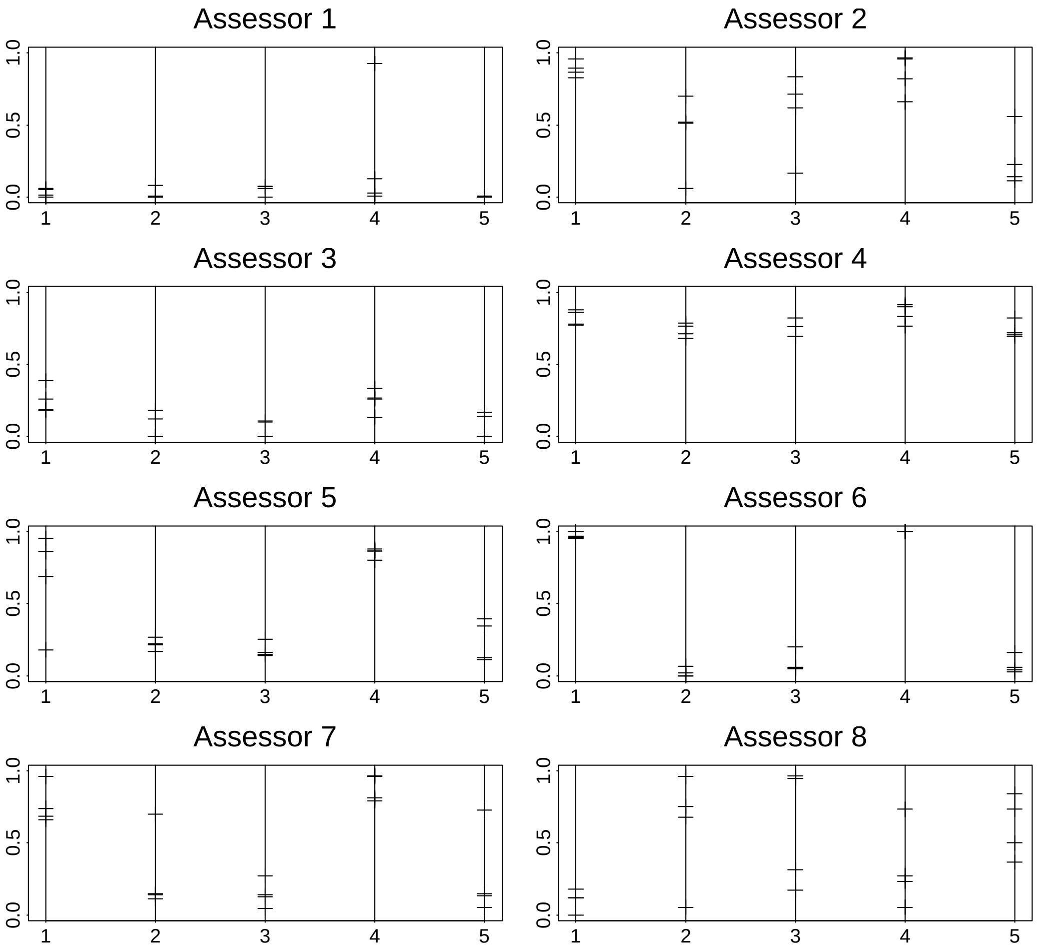

Figure 3.1: Original scores of the five products for each assessor. The five products are represented by positions on the abscissae and the ordinates label the actual scores. Each horizontal mark is one replicate.

Prior to the further analysis we transform the data. Figure 3.2 shows the mean-variance relationship for the data, resembling that of a binomial. We therefore consider new observations, $\tilde{y}_{apr}$ given by

$$\tilde{y}_{apr} = \arcsin(\sqrt{y_{apr}}),$$

as the function $f(x) = \arcsin(\sqrt{x})$ is the variance stabilizing function for the binomial distribution, see, e.g. Weisberg (1985).

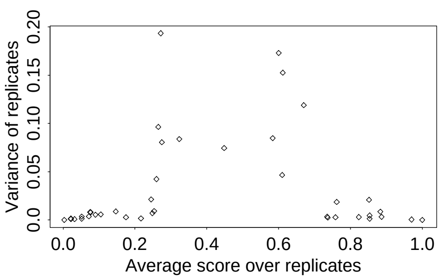

Figure 3.2: Sample variance plotted against average for each combination of assessor and product.

We now proceed as outlined in the previous section. The model $\mathbf{M}_0$ is fitted to the data, i.e. the eight individual one-sided analyses of variance are performed. These results are presented in Table 3.1. Note that all but assessors 1 and 8 distinguish the five products significantly at the 5 % level. As indicated earlier, the individual variances, $\mathrm{MS}_e^{(a)}$, also seem to differ considerably. This is confirmed by a Bartlett test giving a χ^2 test statistic of 45.3 on 7 degrees of freedom, which is extremely significant. Here we have used the corrected test statistic due to Bartlett (1937), that in our case takes the form

$$c\left\{AR(P-1)\log\left(\frac{1}{A}\sum_{a=1}^{A}\mathrm{MS}_e^{(a)}\right) - R(P-1)\sum_{a=1}^{A}\log\left(\mathrm{MS}_e^{(a)}\right)\right\} \tag{3.14}$$

where

$$c = \left\{1 + \frac{1}{3(A-1)}\left(\frac{A}{R(P-1)} - \frac{1}{AR(P-1)}\right)\right\}^{-1} \tag{3.15}$$

is the Bartlett correction factor. Next the model $\mathbf{M}_3$ is tested against $\mathbf{M}_0$, yielding

$$-2\log Q = 199.0$$

Table 3.1: Results from the eight individual one-way analyses of variance. F is the F test statistic for equality of the five products and P the corresponding p value, i.e. $F = \mathrm{MS}_{prod}/\mathrm{MS}_e$

Assessor	$\sqrt{\mathrm{MS}_{prod}}$	$\sqrt{\mathrm{MS}_e}$	F	P
1	0.34	0.27	1.6	0.23
2	0.63	0.25	6.3	0.003
3	0.37	0.18	4.0	0.020
4	0.14	0.072	3.7	0.027
5	0.69	0.20	11.9	0.000
6	1.41	0.10	194	0.000
7	0.78	0.24	10.1	0.000
8	0.57	0.37	2.4	0.097

Table 3.2: Individual parameter and precision estimates under model $\mathbf{M}_1$

Assessor	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}$	$\hat{\beta}/\hat{\sigma}$
1	0.19	0.24	0.26	0.94
2	0.91	0.57	0.24	2.34
3	0.32	0.36	0.16	5.25
4	1.09	0.14	0.063	2.15
5	0.72	0.67	0.18	3.73
6	0.73	1.41	0.089	15.9
7	0.75	0.76	0.23	3.35
8	0.72	-0.51	0.34	-1.50

which is highly significant on $48 - 16 = 32$ degrees of freedom. Hence the data show clear product differences, and we can proceed with the analysis as described in the preceding section.

We now carry out the estimation algorithm. The algorithm was implemented in Turbo Pascal and run on a personal computer. After 10 iterations convergence was obtained in the sense that no parameter estimate changed on the first eight decimals between the last two iterations. The estimates are listed in Tables 3.2 and 3.3. In comparing Table 3.1 and Table 3.2 it should be noted that MS_e in Table 3.1 is the unbiased variance estimate, i.e. with 15 degrees of freedom, whereas $\hat{\sigma}$ in Table 3.2 is the maximum likelihood estimate.

The test for $\mathbf{M}_1$ under $\mathbf{M}_0$ gave

$$-2\log Q = 17.4$$

Table 3.3: Product parameter estimates under model $\mathbf{M}_1$

Product	$\hat{\nu}$
1	0.49
2	-0.43
3	-0.32
4	0.60
5	-0.34

on $8 \cdot 3 - 5 + 2 = 21$ degrees of freedom, yielding a p-value of 0.69. The model $\mathbf{M}_1$ thus seems to offer a satisfactory description of the data, and specifically of the assessor dissimilarities. Finally the differences in use of scale are tested, i.e. model $\mathbf{M}_2$ is tested under $\mathbf{M}_1$, yielding

$$-2 \log Q = 130.1$$

on 7 degrees of freedom. Thus, there certainly is a highly significant difference between the assessors' use of the scale. We therefore use model $\mathbf{M}_1$ as our final model, and turn to the parameter estimates of Tables 3.2 and 3.3. To see how well the model did in catching the four enumerated effects above, the relevant parameter estimates are listed in Table 3.4. We see that the estimates indeed reflect the observations from Figure 3.1. Even less obvious observations are consistent with parameter estimates, for instance that scores from assessor 3 are slightly more variable than those from assessor 4, and that those from assessor 6 are even more precise.

A graphical presentation of the fitted model is given in Figure 3.3. Note that the arcsin-transformation changes the intensity scale from $(0,1)$ to $(0,\pi/2)$. For each assessor (a) predicted values from the model are represented by the individual regression line $(\nu, \hat{\alpha}_a + \hat{\beta}_a \nu)$. This can be interpreted as a representation of the individual scale. Moreover the individual standard deviation is pictured by lines, which are the prediction line plus or minus the standard deviations. Finally the observed means, $(\hat{\nu}_p, \bar{y}_{\cdot p \cdot})$, of the four replicates of each product are plotted. It is our opinion that Figure 3.3 offers a clear and easily interpretable visualization of the different assessor effects in the data.

Each assessor could also be evaluated by a single number, namely the assessor precision estimate from The Model section,

$$\hat{\beta}_a / \hat{\sigma}_a$$

These numbers are listed in Table 3.2, and we see the close relationship with the individual F values from Table 3.1.

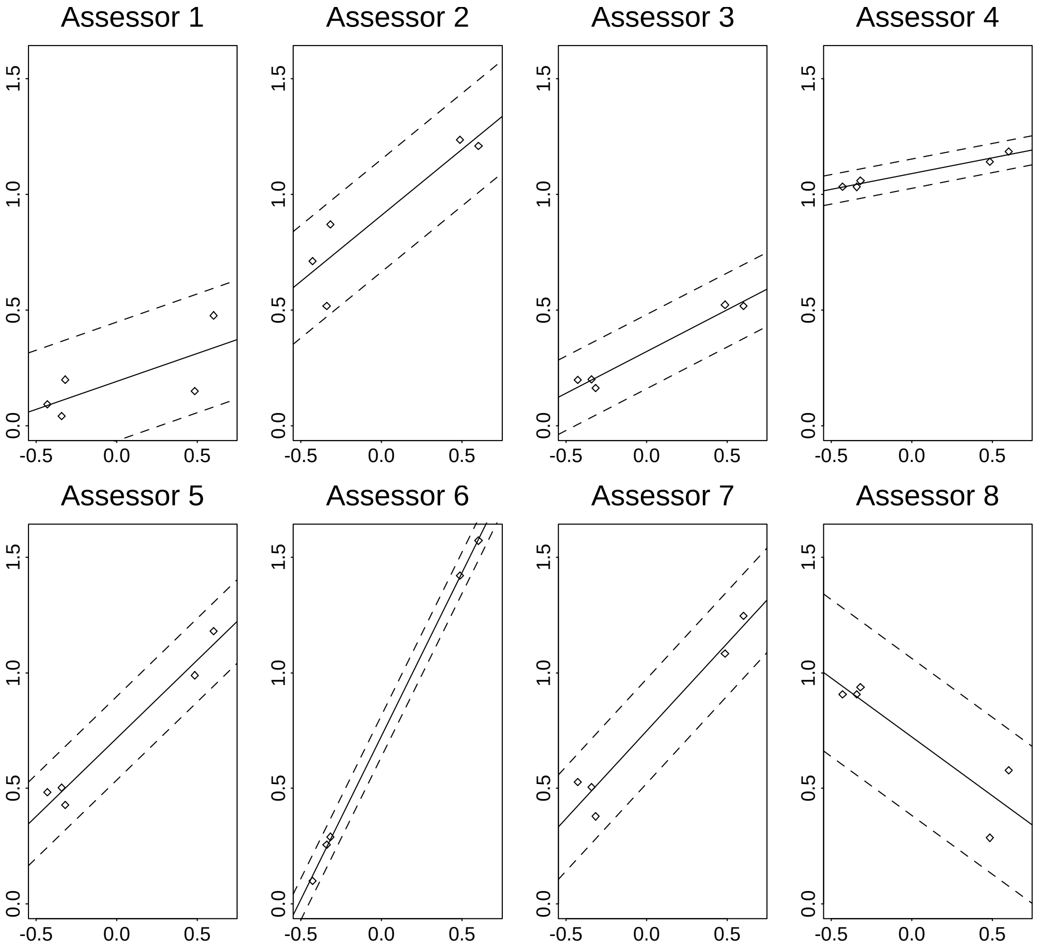

Figure 3.3: Fitted regression lines based on transformed data for each assessor. The ordinates represents the scores given by the assessors and the abscissae are the estimated span of product values. The dotted lines are the fitted lines shifted with plus and minus the individual standard deviations, and the plotted points are the actual observed averages over replicates versus estimated product levels from Table 3.3.

Table 3.4: The four examples of individual differences from Figure 3.1 with corresponding parameter estimates.

	Observations from Figure 1			Estimates from Table 2		
Case no	Level	Scale	Std.	Level	Scale	Std.
1	$\alpha_3 \neq \alpha_4$	$\beta_3 \approx \beta_4$	$\sigma_3 \approx \sigma_4$	$\hat{\alpha}_3 = 0.32$	$\hat{\beta}_3 = 0.36$	$\hat{\sigma}_3 = 0.16$
				$\hat{\alpha}_4 = 1.09$	$\hat{\beta}_4 = 0.14$	$\hat{\sigma}_4 = 0.063$
2		$\beta_4 \neq \beta_6$	$\sigma_4 \approx \sigma_6$		$\hat{\beta}_4 = 0.14$	$\hat{\sigma}_4 = 0.063$
					$\hat{\beta}_6 = 1.41$	$\hat{\sigma}_6 = 0.089$
3		$\beta_6 \approx \beta_7$	$\sigma_6 \neq \sigma_7$		$\hat{\beta}_6 = 1.41$	$\hat{\sigma}_6 = 0.089$
					$\hat{\beta}_7 = 0.76$	$\hat{\sigma}_7 = 0.23$
4		$\beta_8 \neq \beta_a$			$\hat{\beta}_8 = -0.51$	
		$a = 1, \ldots, 7$			$\hat{\beta}_a > 0$	
					$a = 1, \ldots, 7$	

Table 3.5: P values corresponding to χ^2 test statistics from the pairwise comparisons of products under model $\mathbf{M}_1$.

Product	1	2	3	4	5
1	—	0.000	0.000	0.008	0.000
2		—	0.018	0.000	0.053
3			—	0.000	0.573
4				—	0.000
5					—

A particularly nice feature of the precision estimates is that the ‘scale reversal’ by assessor 8 is revealed. In this particular case, however, this effect will not worry us too much, as assessor 8 is separating the five products at the 8 % level of significance only.

Finally we perform the pairwise comparisons for each combination of products, see Table 3.5. We see that all but one pairwise comparison turns up significantly on 5.3 % level, only products 3 and 5 can be said to be equal with respect to the property in question.

3.6 Extensions

Model (3.2) can be extended to be applicable in more general situations. In the present section we sketch possibilities for handling block effects and factorial experiments. Finally, we briefly comment on the case of multivariate observations.

3.6.1 Block effects

Consider the set-up from the previous sections with the additional complexity of allowing a block effect representing the replications. This gives rise to the following nested sequence of models for the expectations:

$$\begin{aligned}
\mathbf{M}_0^* :&\ \ \mathbf{E}Y_{apr} = \gamma_{apr}\\
\mathbf{M}_1^* :&\ \ \mathbf{E}Y_{apr} = \alpha_a + \beta_a\eta_{pr}\\
\mathbf{M}_2^* :&\ \ \mathbf{E}Y_{apr} = \alpha_a + \beta_a(\nu_p + \delta_r)\\
\mathbf{M}_3^* :&\ \ \mathbf{E}Y_{apr} = \alpha_a + \beta_a\nu_p
\end{aligned}$$

For each of the models we still assume that the Y_{apr}s are independent, normally distributed random variables with variances, σ_a^2, depending on the assessor. Model $\mathbf{M}_0^*$ is a saturated model and model $\mathbf{M}_1^*$ may be reformulated as a version of the model (3.2) with no replicates within 'product' levels, and estimation breaks down for both (see Discussion). Model $\mathbf{M}_2^*$ could, however, be tested against the model with only first-order interaction terms:

$$\mathbf{E}Y_{apr} = \tilde{\alpha}_a + \tilde{\nu}_p + \tilde{\delta}_r + \tilde{\gamma}_{ap} + \tilde{\eta}_{pr} + \tilde{\xi}_{ar}$$

Model $\mathbf{M}_3^*$ can be used to test whether the block effect is significant, If not, we are back in the situation of the previous sections.

Model $\mathbf{M}_2^*$ is the model of primary interest here, as it represents a logical extension incorporating block effect in the presence of individual differences in the use of scale. The general idea is to invoke the linear structure 'inside the β_as'. Since any differences in the products, including block effects, are measured through the evaluations of the assessors only, all such effects ought to be affected by the assessors' different use of the scale. This consideration should not prevent us from checking the model in light of the data, of course.

Estimation of parameters in model $\mathbf{M}_2^*$ is done as for model (3.2) in the Estimation Section; the only difference being that the weighted one-way ANOVA step used to estimate the ν_ps, becomes a weighted two-way additive ANOVA instead.

3.6.2 Factorial experiments

Consider again the basic set-up of the previous sections, but assume that $P = F_1 \cdots F_K$, where F_k is the number of factor levels for factor number $k = 1, \ldots, K$

from an $A \times (F_1 \times \cdots \times F_K)$ factorial design. In analogy with the handling of the block effects we consider, for example, the following nested sequence of models for the expectations:

$$\begin{aligned}
\mathbf{M}_0^{**}: &\quad \mathbf{E}Y_{af_1 \cdots f_K} = \gamma_{af_1 \cdots f_K} \\
\mathbf{M}_1^{**}: &\quad \mathbf{E}Y_{af_1 \cdots f_K} = \alpha_a + \beta_a \eta_{f_1 \cdots f_K} \\
\mathbf{M}_2^{**}: &\quad \mathbf{E}Y_{af_1 \cdots f_K} = \alpha_a + \beta_a(\nu_{f_1}^{(1)} + \cdots + \nu_{f_K}^{(K)})
\end{aligned}$$

For $R > 1$, i.e. with 'true replicates', we can treat $\mathbf{M}_0^{**}$ and $\mathbf{M}_1^{**}$ as the earlier defined $\mathbf{M}_0$ and $\mathbf{M}_1$, since the product factor can be considered as a factor itself. From $\mathbf{M}_1^{**}$ through the 'additive' model $\mathbf{M}_2^{**}$ towards further simplifications, various models could be relevant, depending on the experiment and the data, just as in a usual multi-way analysis of variance.

Parameter estimates in any such model can be found as described in the Estimation Section, with the appropriate modification of the weighted ANOVA step.

In a factorial design without replications the analysis may be based on a model without some of the higher order interactions, analogously to the method for block effects. If, for example, an $A \times (F_1 \cdots F_K)$ factorial design is used, the model with interaction terms including at most $K - 1$ of the F_Ks can be taken as the starting point.

3.6.3 Multivariate observations

It is straightforward to extend the estimation procedure of this paper, including the design extensions from above, to perform estimation in a multivariate version of model (3.2):

$$Y_{apr} \sim \mathrm{N}_q(\mu_{ap}, \Sigma_a) \ , \ \mu_{apl} = \alpha_{al} + \beta_{al}\nu_{pl} \ , \ l = 1, \ldots, q$$

with Y_{apr} independent and Σ_a symmetric positive definite, $a = 1, \ldots, A$. This is highly relevant for sensory data, as such typically consist of evaluations of several closely related properties of the products.

For practical purposes, however, the large number of parameters in the A different $q \times q$ covariance matrices $\Sigma_1, \ldots, \Sigma_A$ weakens the applicability and possible results of the model considerably. Restrictions must be made on $\Sigma_1, \ldots, \Sigma_A$ to make this approach feasible. This is, however, beyond the scope of the present paper.

3.7 Discussion

By use of model (3.2) we try to kill two birds with one stone: modelling assessor differences and estimating product differences. The first seems to succeed fairly well.

The estimation of the assessor parameters could be invoked in the training of the assessors, increasing the possible power of subsequent product difference tests, and a plot like Figure 3.3 is useful as feed-back to the panel leader.

Whether an improved estimation of the product values is obtained is more difficult to judge. The desire to weight the results from different assessors according to their abilities, or even to omit the scores from some assessors before averaging over the rest, is frequently expressed by sensory scientists. If this is done on the basis of the F tests from each assessor, for example, as shown in Table 3.1, a severe selection bias must be accounted for. The analysis based on model (3.2) may be seen as a way of doing that.

So far we have been less successful in providing a test for the overall product main effect based on the model. The test of $\mathbf{M}_3$ against $\mathbf{M}_0$ in the Successive Hypothesis Testing section is a test of both main and interaction effects of the product. In this test we do not take the possibly differences among the β values into account. What we would like to do, is to test equality of ν_js in model (3.2). While this is certainly possible, in principle, the un-identifiability of the β_as makes it difficult to obtain even an asymptotic distribution of the test statistic. One possibility is to use a permutation test based on a randomization of the scores within assessors using, for example, the likelihood ratio test statistic. However, the pairwise comparisons performed actually accounts for the differences in use of scale. The main concern here is the accuracy of the χ^2 distributional approximation, which also is a subject of ongoing work.

Although we believe that the approach presented in this paper, possibly after a transformation of the scores like the one applied here, is an improvement over the frequently employed analyses of variance, a number of problems still remain, some of which are discussed below.

The question of regarding assessor effects as fixed or random arises. For use in training of assessors, the fixed effect model presented in this paper seems appropriate. For the subsequent analysis of the products in question, a random assessor effect model can be argued to be more relevant. Lundahl & McDaniel (1988) discuss this issue and conclude that assessor effects in sensory evaluations usually should be considered random. The random effect equivalent of model (3.2) would be a random coefficient regression model with 'unknown regressors' and heteroscedastic variance, and could be an issue of future research.

The phenomenon of assessors using different ranges of the scale is sometimes called the 'rubber-yardstick' effect. Gay and Mead (1992) use a maximum likelihood approach to data with such effects very much like ours, though assuming homogeneity of variances. They mention the complexity of the implementation and possible non-convergence of the iterative estimation scheme as drawbacks to this approach. The simplicity of the algorithm in the Estimation Section and the stated convergence results refute these objections to some extent, though we acknowledge the fact that zero-probability events like 'identical replicate observations' can occur in these kind of data, weakening the practical value of theoretical convergence results slightly. The chance of such 'singular' observations increases with number of parameters and decreases with number of observations. For the example in this paper, no convergence problems occurred.

The assumption of normality is, of course, never strictly fulfilled, since the scores are limited to a fixed range and sometimes confined to a discrete scale. Least squares methods are often used for such problems and our asymptotic results still hold for a 'nice' parametric family, but the possibility of 'singular' observations, as mentioned in the previous paragraph, should be kept in mind.

Another possible violation of assumptions is heteroscedasticity due to products rather than assessors. For fixed scale scoring one typically observes that the results are 'squeezed' at the extremes of the scale, see Figure 3.2. Both this and the possible non-normality could be approached by data transformation prior to analysis.

In the case of only one observation for each combination of product and assessor, we can obtain a perfect regressional fit for one assessor maintaining finite residuals for the remaining ones, by stretching the ν_p estimates appropriately. The likelihood function thus tends to infinity as the variance for that assessor tends to zero, and the maximum likelihood equation is therefore useless. This is in agreement with the point made by Snee (1982), that in a two-way table with one observation per cell it is not possible to distinguish interaction from row- (or column-) related variance inhomogeneity.

When the number of replications is small one might expect some effect approaching the behaviour from the case without replications. Thus, the quality of the asymptotic χ^2 approximations may well be questioned for small R. This leads to the consideration of another type of asymptotics corresponding to an increasing number of assessors. Then, however, the number of parameters increases with the number of assessors and it would be more natural to use a model with random assessor effects. A non-asymptotic approach, based on simulations or permutation tests, might also be considered.

3.8 References

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families.* Wiley, Chichester.

Bartlett, M. S. (1937). Properties of suffiency and statistical test. *Proceedings of the Royal Society London Series A*, **160**, 268–282.

Crowder, M. J. (1980). Proportional linear models. *Applied Statistics*, **29**(3), 299–303.

Gay, C. and Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *Journal of Sensory Studies*, **7**, 205–228.

Lundahl, D. S. and McDaniel, M. R. (1988). The panelist effect — fixed or random? *Journal of Sensory Studies*, **3**, 113–121.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *Journal of American Statistical Association*, **56**, 878–888.

Mandel, J. (1969). The partitioning of interaction in analysis of variance. *Journal of Research—National Bureau of Standards, Section B, Mathematical Sciences*, **73B**(4), 309–328.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics*, **13**(1), 1–18.

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference*, **2**, 187–199.

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.

Snee, R. D. (1982). Nonadditivity in a two-way classification: Is it interaction or nonhomogeneous variance? *Journal of the American Statistical Association*, **77**(379), 515–519.

Ten Berge, J.M.F. (1977). Orthogonal Procrustes rotation to maximal agreement for two or more matrices. *Psychometrika*, **42** (2), 267–276.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics*, **5**, 232–242.

Weisberg S. (1985). *Applied Linear Regression*. Wiley, New York, pp. 134.

Yates, F. and Cochran, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556–580.

3.9 Appendix

3.9.1 The relation between scale parameters and Procrustes constants

In Næs & Solheim (1991), constants g_a that minimize the function T_1,

$$T_1 = \sum_{\tilde{a},a} \sum_{p=1}^{P} \left(g_a(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot}) - g_{\tilde{a}}(\bar{y}_{\tilde{a}p\cdot} - \bar{y}_{\tilde{a}\cdot\cdot})\right)^2 \tag{3.16}$$

are found. This is a one-dimensional version of general Procrustes analysis (Ten Berge, 1977). The g_as are chosen to make the scaled individual product profiles 'close' to each other. An almost identical problem is to choose scaling constants $\tilde{g}_a$ such that the scaled individual product profiles are all 'close' to the average product profile, i.e. to find constants $\tilde{g}_a$ that minimize the function, $\tilde{T}_1$,

$$\tilde{T}_1 = \sum_{a=1}^{A} \sum_{p=1}^{P} \left((\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot}) - \tilde{g}_a(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})\right)^2 . \tag{3.17}$$

The solution to the minimization of $\tilde{T}_1$ is the well-known least squares estimate,

$$\tilde{g}_a = \frac{\sum_{p=1}^{P}(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})}{\sum_{p=1}^{P}(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})^2}. \tag{3.18}$$

If we take the unstandardized initial ν_p estimate, $\nu_p = \bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot}$ in model (3.2) in the present paper, we get from (3.9):

$$\begin{aligned} \tilde{\beta}_a &= \frac{\sum_{p=1}^{P} \bar{y}_{ap\cdot}(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})}{\sum_{p=1}^{P}(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})^2} \\ &= \frac{\sum_{p=1}^{P}(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})}{\sum_{p=1}^{P}(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})^2} \end{aligned} \tag{3.19}$$

Examining expressions (3.18) and (3.19) we see that the $\tilde{g}_a$s are the slopes in the regressions of the $(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})$s on the $(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})$s, whereas the $\tilde{\beta}_a$s are the slopes in the regressions of the $(\bar{y}_{ap\cdot} - \bar{y}_{a\cdot\cdot})$s on the $(\bar{y}_{\cdot p\cdot} - \bar{y}_{\cdot\cdot\cdot})$s.

3.9.2 Assessor precision parameters and F-statistics

The mean squares from the one-way analyses of variance for each assessor are

$$\mathrm{MS}^{(a)}_{prod} = \frac{1}{P-1} \sum_{p=1}^{P} R(\bar{Y}_{ap\cdot} - \bar{Y}_{a\cdot\cdot})^2$$

and

$$\mathrm{MS}^{(a)}_{error} = \frac{1}{P(R-1)} \sum_{p=1}^{P} \sum_{r=1}^{R} (Y_{apr} - \bar{Y}_{ap\cdot})^2$$

It is well known that

$$E\left(\mathrm{MS}^{(a)}_{error}\right) = \sigma_a^2$$

and since

$$E(\bar{Y}_{ap\cdot} - \alpha_a) = \beta_a \nu_p \text{ and } \sum_{p=1}^{P} (\nu_p - \bar{\nu})^2 = 1$$

we get

$$E\left(\mathrm{MS}^{(a)}_{prod}\right) = \sigma_a^2 + \beta_a^2 \frac{R}{P-1}$$

Thus, we have

$$\frac{E\left(\mathrm{MS}^{(a)}_{prod}\right) - E\left(\mathrm{MS}^{(a)}_{error}\right)}{E\left(\mathrm{MS}^{(a)}_{error}\right)} = \frac{R}{P-1} \frac{\beta_a^2}{\sigma_a^2}$$

confirming the proportionality claimed in (3.6).

3.9.3 Data table

Originally the scores were measured on a line segment represented by the interval $(0, 15)$, although we have rescaled them to the interval $(0, 1)$ in Figure 3.1. The original values are given in Table 3.6.

Table 3.6: Original observations measured on a $(0, 15)$ scale.

Assessor	1	2	3	4	5	6	7	8
Product	0.8	12.4	3.9	11.7	12.9	14.4	9.9	2.7
1	0.2	14.4	2.7	12.9	2.7	14.3	14.4	0.0
	0.0	13.4	5.8	13.2	14.3	14.5	11.1	1.8
	0.9	13.0	2.8	11.6	10.3	15.0	10.3	1.8
Product	0.0	7.8	2.7	10.2	2.5	1.0	1.7	0.8
2	0.0	10.5	0.0	11.5	3.2	0.3	2.1	11.3
	1.2	7.7	1.8	10.7	3.3	0.0	10.5	10.2
	0.1	0.9	0.0	11.8	4.0	0.0	2.2	14.4
Product	0.0	12.5	1.6	10.4	2.2	3.0	2.1	2.6
3	1.1	10.7	0.0	12.3	2.1	0.8	1.9	14.5
	1.1	9.3	0.0	11.4	3.8	0.7	0.7	4.7
	0.9	2.5	1.5	11.4	2.4	0.9	4.1	14.2
Product	0.1	9.9	4.0	11.5	13.0	15.0	12.2	4.1
4	13.9	14.4	2.0	12.5	12.0	15.0	14.5	3.5
	0.4	14.5	3.9	13.5	13.2	15.0	11.9	0.8
	1.9	12.3	5.0	13.7	13.0	15.0	14.4	11.0
Product	0.0	1.7	2.5	10.8	1.7	0.9	2.2	11.0
5	0.1	2.1	0.0	10.6	1.9	0.6	10.9	7.5
	0.1	3.4	2.1	10.4	5.9	0.4	2.0	5.5
	0.0	8.4	0.0	12.3	5.2	2.4	0.8	12.6

3.10 Add: Asymptotics and algorithm convergence

3.10.1 Asymptotics

Presented here are results and sketched proofs of the asserted asymptotic properties for model (3.2) (or $\mathbf{M}_1$). It may be seen as a formal verification of asymptotic results for curved exponential families. And although maybe generally accepted and straightforward consequences of well-known theories, these results are difficult to find explicitly in the literature. Ghosh (1994), p. 16, gives a general construction of an mle in a curved family that resembles some of the following. Consider the basic model $\mathbf{M}_0$. This is a regular exponential family (Barndorff-Nielsen, 1978) of order $A(P+1)$ with canonical parameter

$$\theta = \left(-\frac{1}{2\sigma_1^2}, \ldots, -\frac{1}{2\sigma_A^2}, \frac{\mu_{11}}{\sigma_1^2}, \ldots, \frac{\mu_{1P}}{\sigma_1^2}, \ldots, \frac{\mu_{A1}}{\sigma_A^2}, \ldots, \frac{\mu_{AP}}{\sigma_A^2}\right) \tag{3.20}$$

and canonical statistic

$$\mathrm{T} = \left(\sum_{p=1}^{P} y_{1p}^2, \ldots, \sum_{p=1}^{P} y_{Ap}^2, y_{11}, \ldots, y_{1P}, \ldots, y_{A1}, \ldots, y_{AP}\right)$$

The canonical parameter θ varies in the open set $\Theta = \mathrm{I\!R}_-^A \times \mathrm{I\!R}^{AP}$. Let us consider the model of interest, $\mathbf{M}_1$. For convenience we choose the parametrization of $\mathbf{M}_1$ with restrictions

$$\beta_1 = 1 \quad , \quad \alpha_1 = 0.$$

Model $\mathbf{M}_1$ can be specified then, as a sub-model of $\mathbf{M}_0$ by a mapping, η, of

$$\theta = \left(\sigma_1^2, \ldots, \sigma_A^2, \alpha_2, \ldots, \alpha_A, \beta_2, \ldots, \beta_A, \nu_1, \ldots, \nu_P\right) \in \Theta_1$$

with $\Theta_1 = \mathrm{I\!R}_+^A \times \mathrm{I\!R}^{2A+P-2} \setminus \{\theta \in \Theta | \nu_1 = \cdots = \nu_P\}$ into Θ defined as

$$\eta(\theta) = \left(-\frac{1}{2\sigma_1^2}, \ldots, -\frac{1}{2\sigma_d^2}, \right.$$
$$\left. \frac{\nu_1}{\sigma_1^2}, \ldots, \frac{\nu_P}{\sigma_1^2}, \ldots, \frac{\alpha_2 + \beta_2\nu_1}{\sigma_2^2}, \ldots, \frac{\alpha_2 + \beta_2\nu_P}{\sigma_2^2}, \frac{\alpha_A + \beta_A\nu_1}{\sigma_A^2}, \ldots, \frac{\alpha_A + \beta_A\nu_P}{\sigma_A^2}\right)$$

Proposition 1 *M_1 is a curved exponential family in the sense that*

(i) *η is three times continuously differentiable.*

(ii) *The matrix of partial derivatives of η has full rank $3A + P - 2$.*

(iii) *η is homeomorphic, i.e. one-to-one and both ways continuous.*

Proof: (i) is clearly true and (ii) can be checked by explicitly writing up the matrix and showing linearly independence of the columns. That η is one-to-one can also be verified by basic arguments. The corresponding inverse is given by simple expressions, that clearly are continuous and (iii) holds as well.

Proposition 2 *There is a neighborhood $U(\theta_0)$ around the true parameter point $\theta_0 \in \Theta_1$, independent of R, such that the restriction of model $\boldsymbol{M}_1 = \{P_\theta | \theta \in \Theta_1\}$ to $U(\theta_0) \subset \Theta_1$ satisfies the following with P_{θ_0}-probability tending to 1 as $R \to \infty$:*

(i) *A unique maximum likelihood estimate, $\hat{\theta}_R \in U(\theta_0)$ exists.*

(ii) *$\hat{\theta}_R$ is consistent for estimating θ_0.*

(iii) *Let $I(\theta)$ be the Fisher information matrix. We then have*

$$\sqrt{RI(\theta_0)}\left(\hat{\theta}_R - \theta_0\right) \xrightarrow{D} N(0, I)$$

Proof: From (i) and (ii) in Proposition 1 and properties of the regular exponential family $\mathbf{M}_0$, the assumptions of Theorem 4.1 in Lehmann (1983) are obvious or straightforward to verify. The assertions of the proposition thus hold.

Note that Proposition 2 does not ensure the asymptotic existence of a unique global maximum for the likelihood function. However again using Proposition 1 and a property of the regular exponential family we can obtain:

Proposition 3 *With P_{θ_0}-probability tending to 1 as $R \to \infty$ the likelihood function (3.8) has a unique maximum $\hat{\theta}_R$ satisfying (ii) and (iii) of Proposition 2.*

Proof: Choose $U(\theta_0) \subset \Theta_1$ according to Proposition 2. Denote the likelihood function L. We must show that

$$\theta \notin U(\theta_0) \Rightarrow L(\theta) < L(\theta_0) \tag{3.21}$$

where the dependence on R of the likelihood function is suppressed in the notation. Due to the continuity property of the inverse η-mapping, Proposition 1 (iii), there is a neighborhood $V(\eta_0)$ around $\eta_0 = \eta(\theta_0)$, such that

$$\theta \notin U(\theta_0) \Rightarrow \eta(\theta) \notin V(\eta_0) \tag{3.22}$$

In the regular exponential family $\mathbf{M}_0$ we have P_{η_0}-almost surely, that the likelihood function is strictly concave, the maximum likelihood estimator $\hat{\eta}_R$ exists and satisfies $\hat{\eta}_R \to \eta_0$ (Barndorff-Nielsen, 1978). Thus there exists an $\varepsilon > 0$ and an $N > 0$ such that for $R \geq N$

$$\hat{\eta_R} \in V(\eta_0) \text{ and } [\eta \notin V(\eta_0) \Rightarrow L(\eta) < L(\hat{\eta}_R) - \varepsilon] \tag{3.23}$$

The wanted implication (3.21) is then a direct consequence of (3.22) and (3.23).

3.10.2 On the estimation algorithm

The conjecture in Brockhoff & Skovgaard (1994) that the estimation algorithm will converge to a local maximum of the likelihood function is not proved. I could have chosen to stop the discussion here but as the problem seems quite general, I have decided to present the framework in which such a convergence result may be obtained, and the 'result' is expressed as a Conjecture with a subsequent discussion of validity.

The idea of cyclically fixing some parameters and maximizing the likelihood function with respect to the remaining parameters is quite common and applied in several cases, see Jensen *et al.* (1991) and references therein. Indeed in the present thesis the alternating principle is used a number of times. Jensen *et al.* (1991) proves the following general convergence result: Let $\Theta \subset \mathbb{R}^k$ be the parameter space of a statistical model, and let L denote a continuous function on Θ. Assume the following:

(i) Let θ_0 be a starting value, such that

$$D_0 = \{\theta \in \Theta | L(\theta) \geq L(\theta_0)\}$$

is compact.

(ii) The function L is uniquely maximized over D_0 for $\theta = \hat{\theta}$.

(iii) Suppose that we have given parameter functions

$$\psi_i : \; D_0 \rightarrow \Theta_i \;\; (i = 1, \ldots, k)$$

and let $M_i(\theta)$, $\theta \in D_0$ be the corresponding sections:

$$M_i(\theta) = \{\eta \in D_0 | \psi_i(\eta) = \psi_i(\theta)\} \;\; (i = 1, \ldots, k).$$

Then we assume that, for $i = 1, \ldots, k$ and $\theta \in D_0$, L is maximized uniquely by $T_i(\theta)$ on the section $M_i(\theta)$ and that $T_i(\theta)$ is continuous on D_0.

(iv) Assume we have enough sections, or more precisely

$$\sup_{\eta \in M_i(\theta)} L(\eta) = L(\theta) \;\; (i = 1, \ldots, k)$$

implies $\theta = \hat{\theta}$, or equivalently $T_i(\theta) = \theta$ $(i = 1, \ldots, k)$ implies $\theta = \hat{\theta}$.

Then the algorithm

$$\theta_{n+1} = T_1 \circ \cdots \circ T_k(\theta_n)$$

converges to $\hat{\theta}$ for any starting value in D_0.

We cannot, though, establish the unique existence of the maximum likelihood estimator in model (3.2) for any fixed R. The following extension of the result of Jensen *et al.* (1991) to cover the case of the likelihood function not being uniquely maximized is straightforward to obtain:

Proposition 4 *Under the assumptions (i), (iii) above and*

(ii*) *The set $\hat{D}_0$ of local maxima in D_0 for the function L is finite.*

(iv*) *Assume we have enough sections, or more precisely*

$$\sup_{\eta \in M_i(\theta)} L(\eta) = L(\theta) \quad (i = 1, \ldots, k)$$

implies $\theta \in \hat{D}_0$, or equivalently $T_i(\theta) = \theta$ $(i = 1, \ldots, k)$ implies $\theta \in \hat{D}_0$

the algorithm

$$\theta_{n+1} = T_1 \circ \cdots \circ T_k(\theta_n)$$

converges to a point in $\hat{D}_0$ for any starting value in D_0.

Proof: The arguments of the proof in Jensen *et al.* (1991) may be copied directly to obtain that (θ_n) has a convergent subsequence (θ_{n_l}) with limit θ^*, say, such that

$$\theta^* = T_k(\theta^*) = T_{k-1}(\theta^*) = \cdots T_1(\theta^*) \tag{3.24}$$

end hence from (iv∗) that $\theta^* \in \hat{D}_0$. If other subsequences with other limits exist, say, $\theta_1^*, \ldots, \theta_m^*$ then

$$L(\theta^*) = L(\theta_1^*) = \cdots = L(\theta_m^*)$$

for otherwise the points could not be limit points. But this is in contradiction with the uniqueness property (iii), and the convergence result is proved.

The following conjecture expresses the aimed result.

Conjecture 1 *Under the basic assumption (3.4) the following is true for the algorithm in Section 3.3:*

(i) *If $R > 1$ we have P_{θ_0}-almost surely that a maximum of the likelihood function exists and the algorithm converges to a such.*

(ii) *With P_{θ_0}-probability tending to 1 as $R \to \infty$ the algorithm will converge to the unique maximum likelihood estimate $\hat{\theta}_R$.*

On the proof: First note that the algorithm in Section 3.3 equivalently can be viewed as fixing all but one parameter at a time, and as such is on the form of the above. We must verify the assumptions of Proposition 4. Although assumption (i) is satisfied for a regular exponential family (Barndorff-Nielsen, 1978) it is not automatically satisfied for the curved subfamily Θ_1, as Θ_1 is an open set. However explicit examination of the log-likelihood function (3.8) shows that letting any parameter tend to either of its two possible extremes, while fixing the remaining parameters at arbitrary points, makes the log-likelihood tend to $-\infty$; at least if not "too many replicate observations" are equal. But with P_{θ_0}-probability 1 no replicate observations are equal, and assumption (i) holds true. Assumption (iii) is the uniqueness in each maximizing step earlier pointed out to be the fullfilled. Assumption (ii$*$) (and even Assumption (ii)) will hold asymptotically since the starting value for the algorithm is chosen as the maximum likelihood estimates under the regular exponential family $\mathbf{M}_0$, see the argument in the proof of Proposition 3. This is relevant for Conjecture 1(ii). For fixed R, Assumption (ii$*$) may be verified by some compactness argument or may not be crucial.

It is Assumption (iv$*$) that represents the difficulty. This is the assumption that should ensure that a limiting point is actually a local maximum and not, for example, the 'bottom' of a saddlepoint in a diagonal direction nor the top of a diagonal ridge. I will have to leave this point open.

If the second derivatives of the log-likelihood function are calculated it will, of course, always be possible to check whether or not the convergence point of the algorithm (for it is bound to converge to something) is a local maximum or not.

3.11 Add: Simulation

A small simulation study was carried out to investigate the validity of the asymptotic approximations used in the paper. Now and then in this and the next section, model $\mathbf{M}_1$ will be referred to as the 'assessor model'. Three scenarios were exploited:

1. Parameters set to estimates from the paper, from Table 3.2 and 3.3. The test for $\mathbf{M}_1$ under $\mathbf{M}_0$ is examined.

2. $\beta_1 = \cdots = \beta_A = 1$, and the remaining parameters set to estimates from the paper. The test for $\mathbf{M}_2$ under $\mathbf{M}_1$ is examined.

3. $\nu_3 = \nu_5 = -0.33$, and the remaining parameters set to estimates from the paper. The test for $\nu_3 = \nu_5$ under $\mathbf{M}_1$ is examined.

Together with an investigation of the χ^2 statistic in each case we study the distribution of an equivalent approximate F statistic. To me this idea was presented by Ib Skovgaard, that again had it from Bent Jørgensen. Consider a general linear

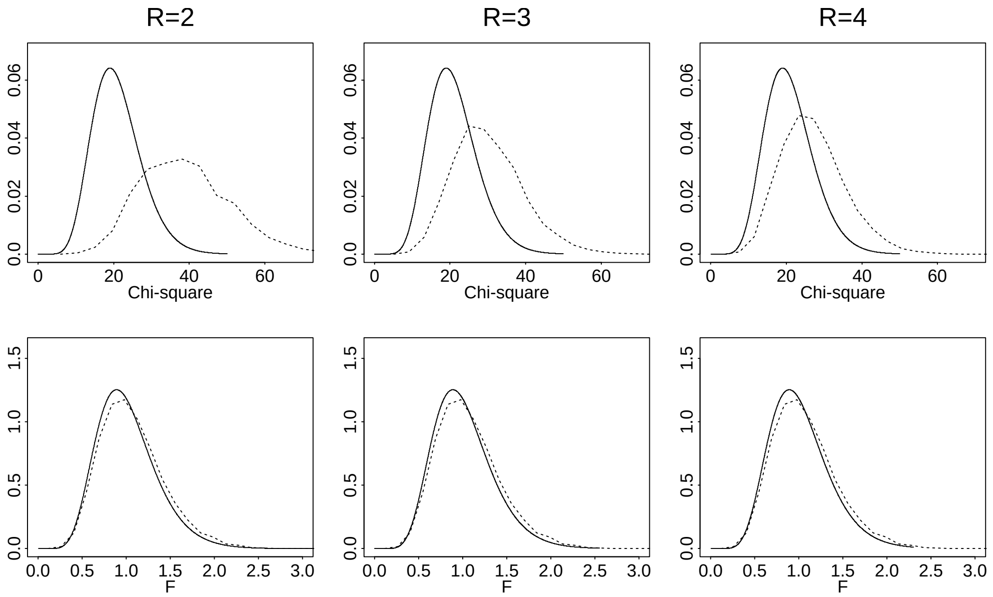

Figure 3.4: Observed (dotted) based on 10000 simulations and true densities (solid) for the χ^2 $(DF = 21)$ and F statistics $(DF = (21, 120))$ for the test of $\mathbf{M}_1$ under $\mathbf{M}_0$.

normal model, where the mean vector μ is tested to belong to L_1 under the model that $\mu \in L_0$, $L_1 \subset L_0 \subset \mathbb{R}^N$, $\dim(L_i) = d_i$, $i = 1, 2$. The relation between the $-2\log Q$ statistic and the F statistic is given by

$$-2\log Q = N \log\left(\frac{d_0 - d_1}{N - d_0} F + 1\right). \tag{3.25}$$

The conjecture by Bent Jørgensen was that in non-linear models the approximate F statistics may in some cases have better small sample properties than the χ^2 statistic. We investigate this in the present setup by calculating the derived F statistics in the three scenarios as

$$\begin{aligned} F_1 &= \left(e^{C_1/N} - 1\right) \frac{N - AP}{AP - 2A - P + 2} \\ F_2 &= \left(e^{C_2/N} - 1\right) \frac{N - 2A - P + 2}{A - 1} \\ F_3 &= \left(e^{C_3/N} - 1\right) (N - 2A - P + 2), \end{aligned}$$

where $N = APR$ and $C_i = -2\log Q_i$ is the χ^2 statistic in scenario i. In each case 10000 simulations were performed for $R = 2$, $R = 3$ and $R = 4$ respectively, and the two test statistics were calculated.

Table 3.7: Observed 5% test levels (in %) for the χ^2 and F statistics. The standard error for an estimated level based on 10000 simulations is at most 0.5, and for a true 5% level, 0.218

	Scenario 1		Scenario 2		Scenario 3	
R	χ^2	F	χ^2	F	χ^2	F
2	66.6	17.6	17.8	8.73	17.7	12.5
3	35.9	9.01	11.9	6.65	11.7	8.78
4	23.4	7.06	9.83	6.09	9.18	7.08

In the Figures 3.4-3.6 the observed distributions are compared to the approximation ones, and in Table 3.7 the observed 5% test levels are listed. The plots of observed densities are all smoothed estimates. It is seen that throughout, the F test provide better results than the χ^2. And by using the F test not too unreliable results are obtained, although improvements in the $R = 2$ case would be desirable. For Scenario 1 the ν_p estimates were also investigated and compared to plain product averages for $R = 4$. The average (of the five estimates) variance of the ν_p estimates in the assessor model was 0.000673, whereas the corresponding value for the plain averages was 0.00141, i.e. the model estimates are more efficient.

Since the distribution of $-2 \log Q$ is independent of the orthogonal projection of the observation vector $x \in \mathbb{R}^N$ onto L_1, it may be argued that the 'sufficient reduction' should be considered instead. This is then the projection onto $L_1^\perp$ and amounts to having $N - d_1$ observations instead of N. This will increase the derived F statistics, and as the general problem in our application is observed distributions laying too far to the right, we did not explore the reduction approach further in this case.

3.12 Add: Random model

In this section a random assessor effect version of model (3.2) is presented, a mean and variance based estimation algorithm is suggested and finally the method is applied to the data of the paper of this chapter, Brockhoff and Skovgaard (1994). The section does not give an exhaustive treatment of the problem, and the suggested approach should only be seen as an indication of possible directions the work with this type of models may lead to.

Let the notation be as in the paper. Assume that we have sampled A individuals from the 'assessor population', and let $(A_a, B_a, \log S_a^2)$ denote the random triplet of level, scale and log-variance for assessor a, $a = 1, \ldots, A$. The model is now specified

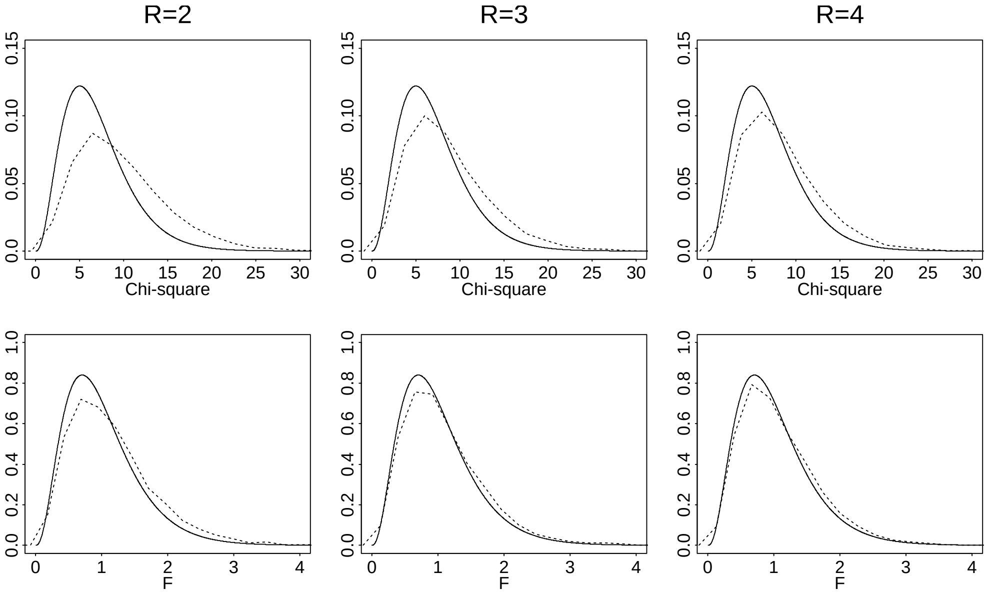

Figure 3.5: Observed (dotted) based on 10000 simulations and true densities (solid) for the χ^2 ($DF = 7$)and F statistics ($DF = (7, 141)$) for the test of $\mathbf{M}_2$ under $\mathbf{M}_1$.

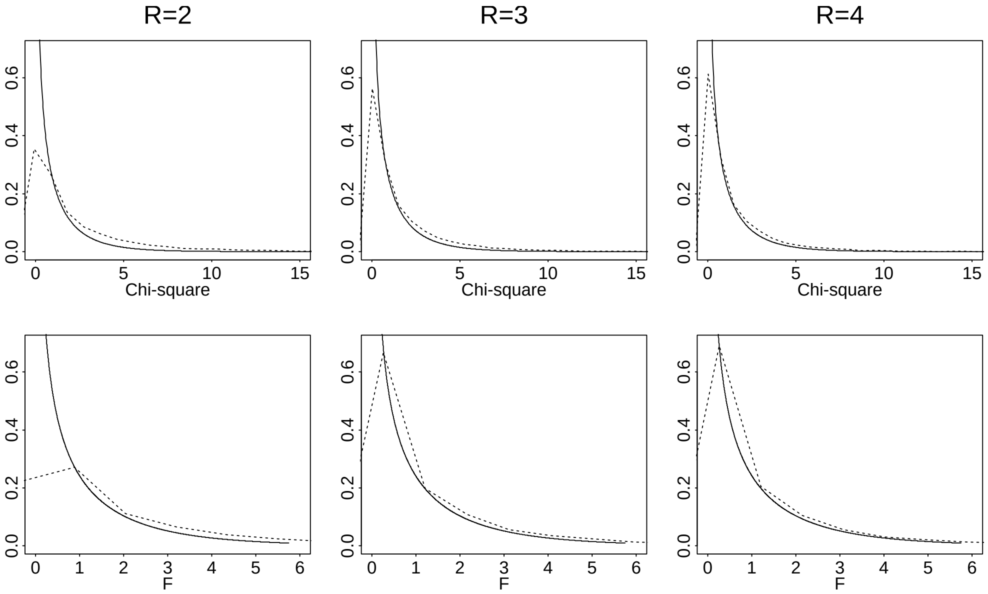

Figure 3.6: Observed (dotted) based on 10000 simulations and true densities (solid) for the χ^2 ($DF = 1$)and F statistics ($DF = (1, 141)$) for the test of $\nu_3 = \nu_5$ under $\mathbf{M}_1$

by assuming that conditional on the sampled assessors, i.e. given

$$\begin{aligned} Z &= ((A_1, B_1, \log S_1^2), \ldots, (A_A, B_A, \log S_A^2)) \\ &= ((\alpha_1, \beta_1, \log \sigma_1^2), \ldots, (\alpha_A, \beta_A, \log \sigma_A^2)) = z \end{aligned}$$

the model (3.2) hold and by specifying the distribution of Z. The model is therefore fully specified by,

$$\begin{array}{rcl} Y_{apr} & = & A_a + B_a \nu_p + \varepsilon_{apr} \\ \varepsilon_{apr} \mid Z = z & \sim & N(0, \sigma_a^2) \\ (A_a, B_a, \log S_a^2) & \sim & N\left((0, 1, \sigma_{\log}), \Sigma_{pop}\right) \end{array} \tag{3.26}$$

where

$$\Sigma_{pop} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} & \sigma_{AS} \\ \sigma_{AB} & \sigma_B^2 & \sigma_{BS} \\ \sigma_{AS} & \sigma_{BS} & \sigma_S^2 \end{pmatrix}$$

is a symmetric, positive definite matrix, $(A_a, B_a, \log S_a^2)$, $a = 1, \ldots, A$ are independent and the ε_{apr}'s are conditionally independent given $Z = z$. The choice of population mean structure, $\mathrm{E}\, A_a = 0$ and $\mathrm{E}\, B_a = 1$ merely makes the model parametrization identifiable. Inference about the parameters $\nu_1, \ldots, \nu_P$ is based on the following three-step principle, given 'current' values for ν_p, $p = 1, \ldots, P$,

1. Estimate the population parameters $\sigma_{\log}$ and Σ_{pop}.

2. 'Predict' the assessor random variables $(A_a, B_a, \log S_a^2)$, $a = 1, \ldots, A$.

3. Estimate $\nu_1, \ldots, \nu_P$ conditional on the predicted values.

Add 1

Given $Z = z$ we can obtain estimates $(\hat\alpha_a, \hat\beta_a, \hat\sigma_a^2)$, $a = 1, \ldots, A$ from the individual regressions together with conditional variance estimates for the parameter estimates, that is, we have

$$\mathrm{E}\left((\hat\alpha_a, \hat\beta_a, \hat\sigma_a^2) \mid Z\right) = (A_a, B_a, S_a^2) \tag{3.27}$$

$$\mathrm{Cov}\left((\hat\alpha_a, \hat\beta_a, \hat\sigma_a^2) \mid Z\right) = \begin{pmatrix} (X^T X)^{-1} S_a^2 & \begin{matrix} 0 \\ 0 \end{matrix} \\ 0 \quad 0 & \frac{2}{PR-2} S_a^4 \end{pmatrix} \tag{3.28}$$

where X is the $PR \times 2$ design matrix for each individual. The estimation of the population parameters is now based on the first two marginal moments of $(\hat\alpha_a, \hat\beta_a, \hat\sigma_a^2)$,

$a = 1, \ldots, A$. Using the laws of iterated conditional expectation and covariance it is easily found, since S_a^2 has the Log-Normal distribution, that

$$
\begin{aligned}
\mathrm{E}\left((\hat{\alpha}_a, \hat{\beta}_a, \hat{\sigma}_a^2)\right) &= (0, 1, \mathrm{e}^{\sigma_{\log} + \sigma_S^2/2}) && (3.29)\\
\mathrm{Var}\,(\hat{\alpha}_a) &= \sigma_A^2 + \frac{\sum_{p=1}^P \nu_p^2}{PR\sum_{p=1}^P (\nu_p - \bar{\nu})^2} \mathrm{e}^{\sigma_{\log} + \sigma_S^2/2} && (3.30)\\
\mathrm{Var}\,(\hat{\beta}_a) &= \sigma_B^2 + \frac{1}{R\sum_{p=1}^P (\nu_p - \bar{\nu})^2} \mathrm{e}^{\sigma_{\log} + \sigma_S^2/2} && (3.31)\\
\mathrm{Cov}\,(\hat{\alpha}_a, \hat{\beta}_a) &= \sigma_{AB} - \frac{\bar{\nu}}{R\sum_{p=1}^P (\nu_p - \bar{\nu})^2} \mathrm{e}^{\sigma_{\log} + \sigma_S^2/2} && (3.32)\\
\mathrm{Var}\,(\hat{\sigma}_a^2) &= \mathrm{e}^{2\sigma_{\log} + \sigma_S^2} \left(\mathrm{e}^{\sigma_S^2} \frac{PR}{PR-2} - 1\right) && (3.33)\\
\mathrm{Cov}\,\left(\hat{\alpha}_a, \log \hat{\sigma}_a^2\right) &= \sigma_{AS} && (3.34)\\
\mathrm{Cov}\,\left(\hat{\beta}_a, \log \hat{\sigma}_a^2\right) &= \sigma_{BS} && (3.35)
\end{aligned}
$$

First (3.29) and (3.33) are solved to give $\hat{\sigma}_{\log}$ and σ_S^2 by inserting on the left hand sides the observed equivalents,

$$
\begin{aligned}
\hat{\mathrm{E}}\,(\hat{\sigma}_a^2) &= \frac{1}{A} \sum_{a=1}^{A} \hat{\sigma}_a^2 = \bar{\hat{\sigma}}^2\\
\hat{\mathrm{Var}}\,(\hat{\sigma}_a^2) &= \frac{1}{A-1} \sum_{a=1}^{A} \left(\hat{\sigma}_a^2 - \bar{\hat{\sigma}}^2\right)^2
\end{aligned}
$$

Next the estimates $\hat{\sigma}_{\log}$ and σ_S^2 are substituted in the remaining equations and the observed equivalents for any of the remaining left hand sides are inserted in an analogous way. Note that by this procedure the expected individual variance $\mathrm{E}\,S_a^2$ is estimated by the average of the estimated individual variances $\bar{\hat{\sigma}}^2$.

Add 2

For the conditional expectation of $(A_a, B_a, \log S_a^2)$, $a = 1, \ldots, A$ given the observations $y = (y_1, \ldots, y_A)$ we use an approximation given by the Normal distribution. This leads to predictions given by

$$
\left(\hat{A}_a, \hat{B}_a, \log \hat{S}_a^2\right) = (0, 1, \hat{\sigma}_{\log}) + (y_a - \hat{\nu})\, \hat{\Sigma}_{mar}^{-1} \hat{\Sigma}_{21}, \qquad (3.36)
$$

where $\hat{\nu}$ is the PR-vector of ν_p-estimates corresponding to one assessor, Σ_{mar} is the marginal covariance matrix for the individual observation set Y_a and Σ_{21} is the covariance between Y_a and $(A_a, B_a, \log S_a^2)$. As notation indicates they can be found not to depend on a, and they are specified by the following equalities,

$$\begin{aligned}
\mathrm{Var}\ (Y_{apr}) &= \sigma_A^2 + \nu_p^2\sigma_B^2 + 2\sigma_{AB}\nu_p + \mathrm{e}^{\sigma_{\log}+\sigma_S^2/2} && (3.37)\\
\mathrm{Cov}\ (Y_{ap_1r_1}, Y_{ap_2r_2}) &= \sigma_A^2 + \nu_{p_1}\nu_{p_2}\sigma_B^2 + \sigma_{AB}(\nu_{p_1} + \nu_{p_2}) && (3.38)\\
\mathrm{Cov}\ (A_a, Y_{apr}) &= \sigma_A^2 + \sigma_{AB}\nu_p && (3.39)\\
\mathrm{Cov}\ (B_a, Y_{apr}) &= \sigma_B^2\nu_p + \sigma_{AB} && (3.40)\\
\mathrm{Cov}\ \left(\log S_a^2, Y_{apr}\right) &= \sigma_{AS} + \nu_p\sigma_{BS} && (3.41)
\end{aligned}$$

Add 3

Given the predicted values the ν_p-estimates are calculated as in step 3 of the fixed effects procedure in Section 3.3.

After step 2 in each iteration a modification of the predicted A_a's and B_a's is inserted to make the algorithm converge to a solution corresponding to $\mathrm{E}\,A_a = 0$ and $\mathrm{E}\,B_a = 1$. Inference about $\nu_1, \ldots, \nu_P$ is now based on the final estimate of the marginal covariance structure,

$$\hat{\mathrm{Cov}}\,(\nu_1, \ldots, \nu_P) = \left(X^T\hat{\Sigma}^{-1}X\right)^{-1}, \tag{3.42}$$

where now X denotes the $APR \times P$ overall design matrix and Σ is the block diagonal $APR \times APR$ covariance matrix with A identical blocks of Σ_{mar}.

3.12.1 Example

Using the average product levels as starting values the random model was fitted to the data used earlier in this chapter, i.e. the arcsin($\sqrt{\cdot}$)-transformed rescaled figures of Table 3.6. For the sake of comparison the estimates in the fixed effects model, Tables 3.2 and 3.3, were reparametrized to satisfy $\sum_{a=1}^{A}\hat{\alpha}_a = 0$ and $\sum_{a=1}^{A}\hat{\beta}_a = A$. We will go through the results of the approach and compare those with the earlier obtained results.

First we note that by using the predicted values $(\hat{A}_a, \hat{B}_a, \hat{S}_a^2)$ we obtain information about the individuals exactly of the same form as for the fixed assessor model. We could thus make an equivalent to Figure 3.3. For the comparison here we will consider Table 3.8 only. The estimates in the two models look quite similar with one major exception: the individual standard deviations have been 'drawn towards the mean' in the random model. This implies the similar tendency for the precision predictions $\hat{B}/\hat{S}$, and has the interesting consequence that the assessors will be weighted more equally for the ν_p estimation in the random model than in the fixed model. One may also note that a few changes in the assessors relative positions occur; this is not studied in further detail.

Table 3.8: Individual parameter and precision estimates under model $\mathbf{M}_1$ and predicted dittos under the random model

Assessor	$\hat{\alpha}$	$\hat{A}$	$\hat{\beta}$	$\hat{B}$	$\hat{\sigma}$	$\hat{S}$	$\hat{\beta}/\hat{\sigma}$	$\hat{B}/\hat{S}$
1	-0.17	-0.18	0.53	0.55	0.26	0.32	2.07	0.80
2	0.059	0.062	1.25	1.25	0.24	0.16	5.14	3.62
3	-0.21	-0.21	0.79	0.79	0.16	0.27	4.93	1.35
4	0.88	0.83	0.30	0.33	0.063	0.17	4.72	0.93
5	-0.29	-0.26	1.48	1.46	0.18	0.18	8.19	3.80
6	-1.38	-1.25	3.10	3.01	0.089	0.14	34.9	10.4
7	-0.38	-0.34	1.66	1.64	0.23	0.17	7.35	4.49
8	1.48	1.35	-1.11	-1.03	0.34	0.28	-3.29	-1.71

Table 3.9: Conventional fixed model ANOVA table

Effect	SS	DF	F	P
Assessor	11.96	7	32.7	0.000
Product	6.82	4	32.6	0.000
Ass⋆Pro	9.40	28	6.42	0.000
Error	6.27	120		

The ANOVA table is the usual way of reporting the results of an analysis of variance, see Table 3.9. In Table 3.10 we report the results of the fixed assessor model in a similar way. The effects have been named in an ANOVA-like manner, but refers in that order to the Bartlett test, the assessor fit test, the equality of β_as test and the equality of products 'test'. The latter in quotation marks as we do not actually test this hypothesis owing to the lack of asymptotic results due to identifiability problems. For the two interaction effects we report both the χ^2 and the derived F statistics, see Section 3.11. The fixed assessor model table is, of course, not unique in the sense that every test statistic depends on what other effects are included in the model.

Finally let us consider inference about the product parameters $\nu_1, \ldots, \nu_P$. Table 3.11 lists the estimates under the two models together with the product averages. The latter are the estimates under both the fixed and mixed ANOVA models. No large differences can be observed. With both the test statistics from the fixed assessor model, the fixed and random ANOVA models and the random assessor model we have all together five possible approaches for performing paired comparisons of the product levels. These are all listed in Table 3.12 For the tests in the random assessor model the

Table 3.10: Fixed assessor model 'ANOVA table'

Effect	$-2\log Q$	DF	P	F	DF	P
Reproducibility	45.3	7	0.000			
Excess interaction	17.4	21	0.69	0.66	$(21, 120)$	0.86
Scale interaction	130.1	7	0.000	25.3	$(7, 141)$	0.000
Product	181.6					

Table 3.11: Product parameter estimates under model $\mathbf{M}_1$ and the random assessor model

Product	$\hat{\nu}$	$\hat{\nu}$	
	fixed effect	random effect	$\bar{y}_{\cdot p \cdot}$
	model	model	
1	0.90	0.89	0.85
2	0.48	0.47	0.51
3	0.53	0.51	0.54
4	0.95	0.94	1.00
5	0.52	0.49	0.50

normal distribution was used with a variance calculated from (3.42). The population parameters needed for the variance estimation were estimated to

$$\begin{aligned} \hat{\sigma}_{\log} &= -3.19 \\ \hat{\Sigma}_{pop} &= \begin{pmatrix} 0.65 & -0.87 & 0.21 \\ -0.87 & 1.35 & -0.53 \\ 0.21 & -0.53 & 0.39 \end{pmatrix} \end{aligned}$$

The random assessor model seems in general to provide results similar to those of the mixed ANOVA, whereas the fixed assessor model resembles the fixed ANOVA. In one case, however, the ν_2–ν_5 comparison the fixed assessor model gives a test probability much smaller than the fixed ANOVA. This may be due to poor approximation of the distribution in this case.

Table 3.12: Pairwise product comparisons

Product	2	3	4	5	Method
1	.000	.000	.008	.000	Fix. Ass.(χ^2)
1	.000	.000	.012	.000	Fix. Ass.(F)
1	.000	.000	.015	.000	Fix. ANOVA
1	.000	.000	.149	.000	Rand. Ass.
1	.024	.040	.336	.020	Rand. ANOVA
2		.018	.000	.053	Fix. Ass.(χ^2)
2		.026	.000	.069	Fix. Ass.(F)
2		.547	.000	.846	Fix. ANOVA
2		.539	.000	.774	Rand. Ass.
2		.813	.002	.939	Rand. ANOVA
3			.000	.573	Fix. Ass.(χ^2)
3			.000	.597	Fix. Ass.(F)
3			.000	.426	Fix. ANOVA
3			.000	.888	Rand. Ass.
3			.004	.755	Rand. ANOVA
4				.000	Fix. Ass.(χ^2)
4				.000	Fix. Ass.(F)
4				.000	Fix. ANOVA
4				.000	Rand. Ass.
4				.002	Rand. ANOVA

Chapter 4

Random effect threshold models for dose-response relations with repeated measurements

The paper is a revised version under consideration by *Journal of the Royal Statistical Society Series B.* Series B is the methodological oriented part of this journal published by The Royal Statistical Society in Britain. It "aims to publish papers on the theoretical and methodological aspects of statistics."

The problem of estimating sensory thresholds triggered the work of this Chapter, and as indicated this directs us into the world of generalized linear models (GLIM), McCullagh and Nelder (1989). We thus give a brief introduction to generalized linear modelling below, together with a sensory methodological review of threshold determination. This is based on Brockhoff (1993) and illustrates how statistical modelling in some instances may offer a unifying approach to a topic otherwise approached in various ad hoc ways. One might raise the question: what may GLIM's in general offer to the field of sensometrics? In the additional section after the paper in this Chapter we address this question.

The generalized linear model

Let us briefly define the GLIM; the reader is referred to McCullagh and Nelder (1989) for details. Let Y be a random n-vector of independent observations and X an $n \times p$ design matrix. If $\mu = \mathrm{E}\,Y$ the GLIM is defined by the density $f(y;\theta)$, $\theta \in \Theta \subset \mathbb{R}^q$ of Y_i, and by a real function g, that links the mean with a linear model,

$$g(\mu) = (g(\mu_1), \ldots, g(\mu_n)) = \beta X^t, \qquad \beta \in \mathbb{R}^p \tag{4.1}$$

The function g is called the *link function.* The density $f(y;\theta)$ defines the error distribution and is usually assumed to be in the class of exponential families. The classical linear Normal model is a GLIM with the identity function as link and $f(y;\theta)$ as the Normal density. A basic feature of the GLIM is the allowance for a non-trivial mean-variance relationship, usually denoted by a variance function $V(\mu)$. The variance function is specified through the specification of the error distribution $f(y;\theta)$. An example is the binomial distribution that has the parabolic variance function $V(\mu) = \mu(1-\mu)$. Maximum likelihood estimates for the parameter vector θ are found numerically by an algorithm of various names and interpretations, e.g. Fisher Scoring, Newton-Raphson and Iterative Weighted Least Squares (IWLS). A central point to make here is that this algorithm depends on the first two moments of $f(y;\theta)$ only. The method of estimation can thus be applied to any situation, where the first two moments have been specified, whether or not a parametric model is underlying. In these cases the approach is referred to as a quasi-likelihood method.

Sensory thresholds

In the sensory literature, see for instance Amerine *et al.* (1965), a sensory threshold is usually defined in the following way:

> A threshold is the minimum concentration of a stimulus that can de detected/discriminated/recognized 50% of the time.

The phrase '50% of the time' leaves room for various interpretations, which certainly can be seen in the sensory literature. The typical experiment for determining thresholds is a K-alternative forced choice experiment, see for instance Frijters (1988), with repeated observations on each assessor and a baseline guessing probability of $\alpha = 1/K$. At this point we consider also the basic test methods, the duo, duo-trio and triangle tests, as 2- and 3-AFC methods, although not commonly done so. One suggested approach, acknowledging the difference of 'between' and 'within' assessor variability, see Meilgaard (1987), is to perform 'individual experiments' to determine individual thresholds and then, subsequently, average these individual thresholds. Lea (1988) suggest a similar two-step procedure in the basic setup of a single difference test: count the number of assessors that individually is significant, and then make a significance test on that number.

But even in a situation where replications and individuals are not mixed, the '50% of the time' has been given different interpretations. Guadagni *et al.* (1972) used the 'significance' method, whereas Meilgaard (1987) points out that a direct 'proportion' approach is the proper one to use. The former amounts to defining a dose to be above threshold level if the number of correct responses makes the statistical test of $\alpha+(1-\alpha)/2$ significant, and the latter if more than a proportion of $\alpha+(1-\alpha)/2$ of the responses are correct. None of these concepts are easily generalized to situations with varying number of assessors from session to session, nor varying baseline probabilities, of which at least the former situation will occur frequently.

Most of such definition controversies could be avoided if one agreed on the basic model (1.1) (or a simplified version of it) in the Thesis Introduction for the unobservable thresholds. Introducing the notation of this Chapter we let T_{ij} denote the threshold of individual i when presented to dose d_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. The doses d_{ij} are assumed to be on log-scale and will often be the same across individuals and some of them also within individuals, but this is not essential. A sub-model of model (1.1) without the interaction effect, may now be written as

$$T_{ij} = Z_i + \varepsilon_{ij} \tag{4.2}$$

where the ε_{ij}s are independent zero-mean Normals $N(0, \sigma_\varepsilon^2)$ and the Z_is are independent Normals with the population mean threshold (on log-scale) as expected value, $Z_i \sim N(\beta, \sigma_Z^2)$. The probability p_{ij} that the i'th individual responds correctly when given dose d_{ij} is then

$$p_{ij} = \alpha + (1-\alpha)\Phi\left(\frac{d_{ij} - \beta}{\sigma_T}\right) \tag{4.3}$$

where $\sigma_T^2 = \sigma_Z^2 + \sigma_\varepsilon^2$. Defining so-called chance-corrected probabilities as $p_{ij}^{corr} = (p_{ij} - \alpha)/(1 - \alpha)$, (4.3) may be written as

$$\Phi^{-1}\left(p_{ij}^{corr}\right) = -\frac{\beta}{\sigma_T} + \frac{d_{ij}}{\sigma_T} \tag{4.4}$$

The mean population threshold β now automatically becomes the 'median effective dose', ED_{50}, in the corrected probabilities, and it can be estimated through a classical dose-response approach, see for instance Finney (1971). In GLIM terminology (4.3), or (4.4), defines the link function, and a formal maximum likelihood estimation of the parameters is obtained by the IWLS algorithm mentioned above.

However, (4.2) and (4.3) in fact define a GLIM with random effects, that in this case amounts to observations within individuals being correlated with a correlation depending on the doses d_{ij}. This is not accounted for in the classical IWLS algorithm. The handling of such models is the topic of the paper in this Chapter, and as this area is relatively new in statistical research, it has, not surprisingly, not found its way into the sensory applications yet. For the rest of this presentation therefore, whenever we refer to the dose-reponse setup and IWLS algorithm we consider the classical independent case, bearing in mind that the random effect version should in fact be employed, if repeated observations on different individuals are at hand.

Pangborn (1981) noted that the derivation of the threshold from the observed frequencies of correct reponses could be done by interpolation or by statistical analysis (regression). Early applications resemble the 'interpolation attitude', see for instance Guadagni *et al.* (1963). Patton and Johnson (1957) and Meilgaard (1975) use linear regression on the intermediate concentrations. All these techniques correspond in the light of the dose-response setup above to assuming linearity of Φ for the intermediate concentrations. Salo (1970) introduced the corrected probabilities and acknowledged the non-linear relationship by the use of log-probability plots based on values of Φ. This was also used by Mulders (1973), but in both cases the determination of the threshold was done by reading off the plot (based on some definition of '50% of the time'). Punter (1983) performed 'sigmoid-curve' fitting on corrected probabilities, referring to a Fortran-routine due to Drake (1975), which may be similar to the IWLS algorithm.

Guadagni *et al* . (1973) allready viewed the situation as a classical dose-reponse setup and they referred to Litchfield and Wilcoxon (1949) for the computational technique. And as late as in Tekin and Karaman (1992) the same approach was used. These techniques may be seen as one step of the general IWLS algorithm. But with todays knowledge in GLIM's in general, and available soft- and hardward, it should, however, be quite straightforward to employ (a version of) the full IWLS algorithm. MacRae (1987) notes that a computer program has been developed that fits a formal model to triangle data, but the connection to GLIM's was not made.

The GLIM approach opens up for generalizations in various ways: working on original instead of corrected probabilities, allowing varying number of individuals and baseline probabilities, and finally to include in a proper way the random assessor effects in both estimation and evaluation of uncertainty in estimation. These generalizations are all included in the approach of the following paper together with the generalization that the distributions of Z_i and ε_{ij} above need not be Normals.

Random effect threshold models for dose-response relations with repeated measurements

Per M. Brockhoff
Dept. of Mathematics and Physics
The Royal Veterinary and
Agricultural University at Copenhagen
Thorvaldsenvej 40
1871 Frederiksberg, Denmark
and
Hans-Georg Müller
Division of Statistics
University of California
Davis, Ca 95616 USA

Summary

A random threshold model for dose-response designs with repeated measurements is introduced. Inference procedures for this model are discussed. The framework of generalized linear models is used to propose a quasi-likelihood inference procedure. The feasibility of this approach is illustrated with the analysis of food science data concerning a forced choice sensory experiment.

Key words and phrases: Binomial regression, forced choice experiment, generalized linear model, iterated weighted least squares, likelihood, overdispersion, quasi-likelihood, sensory experiment, variance component.

4.1 Introduction

We consider a dose-response study involving m individuals for which repeated binary measurements are obtained by applying several stimuli to the same individual and observing the responses. An example would be a sensory experiment in the area of food science and food technology. The response of interest often is detection (yes or no) of odour or flavour from a solution of a chemical substance which is presented to an individual at various concentrations. Odour and flavour detection thresholds across the population are then of primary interest. Other examples occur in Phase II Clinical Studies where a number of subjects are given various doses of a drug, some of them are treated repeatedly with varying doses, and success/failure of the therapy or, alternatively, the occurrence of severe side effects (yes or no) is assessed, again with interest focussed on population average behaviour . Further examples for repeated binary measurement designs are mentioned in Anderson and Aitkin (1985).

As compared to other repeated measurements dose-response designs considered previously by Pierce and Sands (1975), Elashoff (1981) and Anderson and Aitkin (1985), Zeger *et al.* (1988), or more recently Breslow and Clayton (1993), the novelty of our approach lies in the model and the proposed algorithm: In the model it is assumed that the observed responses are determined by underlying unobservable subject specific thresholds which are randomly distributed between individuals. The random threshold determines an individual's capability for reacting to a given stimulus. A reaction will occur for a given subject, if the stimulus exceeds the subject's threshold. Such an approach makes sense biologically as dose-response relations are usually inferred from distributions of such thresholds between individuals, see for instance Im and Gianola (1988) and Morgan (1992), Section 1.5.

Since repeated responses obtained from the same individual are dependent, inference drawn from a standard dose-response approach which ignores those dependencies is incorrect. Our random threshold based approach takes repeated measurements into account by assuming that the random threshold effect is the same for repeated measurements on the same subject. Besides its biological plausibility, this approach has the attractive feature that it leads to a natural separation of individual threshold random effect and measurement random effect. We show in Section 4.4 and 4.5 that this model can be implemented with a quasi-likelihood based iterated weighted least squares algorithm and is computationally very feasible.

Formally, let Y_{ij} be a binary observation (response yes or no) made on the i-th individual under administration of dose level d_{ij} of the substance of interest, where $i = 1, ..., m$, $j = 1, ..., n_i$. Let x_{ij}^T be a p-vector of individual covariates of which the first typically will be 1, the corresponding parameter playing the role of an intercept.

Let T_{ij} be the random threshold for individual i when given dose d_{ij}, and let $\beta \in \mathbb{R}^p$ be a parameter vector. Our basic random threshold model is then

$$T_{ij} = x_{ij}^T\beta + Z_i + \epsilon_{ij}, \quad i = 1, ..., m; \quad j = 1, ..., n_i, \tag{4.5}$$

where $Z_1, ..., Z_m$ are zero mean i.i.d. random variables independent of the errors $\{\epsilon_{ij}\}$, which are also i.i.d. zero mean random variables. The variances

$$\sigma_Z^2 = \text{Var}(Z_i) \text{ and } \sigma_\epsilon^2 = \text{Var}(\epsilon_{ij})$$

are assumed to exist.

The "random thresholds" T_{ij} thus are composed of a systematic component $x_{ij}^T\beta$, which describes the covariate effects, assumed to be non-random, the random subject effects Z_i and the random error effects ϵ_{ij} which vary across measurements. The actual observations are realizations of binary random variables Y_{ij} defined as the indicators of T_{ij} being less than the doses d_{ij},

$$Y_{ij} = 1_{\left\{x_{ij}^T\beta + Z_i + \epsilon_{ij} \leq d_{ij}\right\}}. \tag{4.6}$$

Given observations

$$\left(x_{ij}^T, Y_{ij}\right), \ 1 \leq i \leq m, \ 1 \leq j \leq n_i,$$

one seeks inference procedures for the covariate parameter vector β and the variance components σ_Z^2 and σ_ϵ^2.

We note that in the area of sensory experiments as well as in the other examples mentioned above, interest is focussed on the distribution of the individual random thresholds T_{ij} across the population in dependency on the covariates, not so much on the subject random effects Z_i. Therefore, it is not of major interest here to consider conditional likelihood and inference, given the $Z_i's$. For this reason, the approach developed in the following is a marginal quasi-likelihood approach in the sense of Breslow and Clayton (1993), respectively a population-average model in the sense of Zeger *et al.* (1988). In this regard it is distinct from penalized quasi-likelihood respectively subject-specific models which describe an individual's response to changing covariates rather than the population response. Such subject-specific approaches for binomial regression with repeated measurements have been developed by Schall (1991), who uses linearization methods to develop generalizations of an estimation procedure for classical linear models with normal random effects (Harville, 1977).

Various population average respectively marginal quasi-likelihood methods which are related to our proposal have been discussed in the literature. Prominent among them is the work by Anderson and Aitkin (1985), who assume that the subject random effects are normally distributed and that the error random effects have a logistic distribution. Then a numerical integration method by Gaussian quadrature is invoked to maximize the likelihood. In our approach, the numerical evaluation of a complex likelihood function is avoided. As we propose a quasi-likelihood procedure, only second moments need to be evaluated.

Another well known but somewhat limited alternative approach to deal with repeated measurements in Generalized Linear Models is to assume that the measurements are independent, but that there is heterogeneity of the response probabilities due to the different threholds for different individuals. This approach leads to overdispersion in the binomial regression model (see Wedderburn, 1974, or McCullagh and Nelder, 1989); an example would be the beta-binomial model. We will discuss these alternative modelling approaches below in the context of data analysis for a sensory tasting experiment in Section 4.4.

The likelihood approach is discussed in Section 4.2, and a more feasible quasi-likelihood method is developed in Section 4.3. Section 4.4 contains an application to a forced choice sensory experiment and Section 4.5 the results of some simulations. One complicating feature of such experiments to be discussed in more detail below is that they often involve nonzero baseline probabilities.

4.2 Likelihood approach and forced choice experiments

We use the notation F_ϵ, F_Z to denote cumulative distribution functions of random variables ϵ and Z. In model (4.5), (4.6) the likelihood is

$$L\left(\beta, \sigma_\epsilon^2, \sigma_Z^2; \, y\right) = \prod_{i=1}^{m} P\left(\epsilon_{ij} \leq d_{ij} - Z_i - x_{ij}^T\beta, \; j = 1, ..., n_i\right),$$

and the conditional likelihood, conditioning on $Z_i = z$, becomes

$$\begin{aligned} &L^{(z)}\left(\beta, \sigma_\epsilon^2, \sigma_Z^2; \, y\right) \\ &= \prod_{i=1}^{m} \prod_{j=1}^{n_i} \left[P\left(\epsilon_{ij} < d_{ij} - z - x_{ij}^T\beta\right)\right]^{Y_{ij}} \left[1 - P\left(\epsilon_{ij} < d_{ij} - z - x_{ij}^T\beta\right)\right]^{1-Y_{ij}}. \end{aligned}$$

Hence,

$$L\left(\beta, \sigma_Z^2, \sigma_\epsilon^2; \, y\right) = \prod_{i=1}^{m} \int_{-\infty}^{\infty} \left(\prod_{j=1}^{n_i} U_{ij}\right) F_Z(dz), \tag{4.7}$$

where

$$U_{ij} = \left[F_\epsilon\left(d_{ij} - z - x_{ij}^T\beta\right)\right]^{Y_{ij}} \left[1 - F_\epsilon\left(d_{ij} - z - x_{ij}^T\beta\right)\right]^{1-Y_{ij}}. \tag{4.8}$$

If θ is one of the components of the parameter of interest, the corresponding likelihood equations $\frac{\partial}{\partial\theta}(\log L) = 0$ become

$$\sum_{i=1}^{m} \left[\int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \prod_{j=1}^{n_i} U_{ij} \right) F_Z(dz) \right] \Big/ \left[\int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{n_i} U_{ij} \right\} F_Z(dz) \right] = 0, \qquad (4.9)$$

which even for common cases (normal or logistic distributions for F_Z or F_ϵ) cannot be simplified further and must be solved numerically.

A modification of the likelihood is necessary for forced choice experiments. Typically, in such experiments a subject chooses between a fixed number of alternatives, and the response is recorded as a success if the choice is correct, a failure otherwise. The distinguishing feature of such forced choice experiments is that there is a baseline response probability α with $\alpha > 0$, which corresponds to the probability of success for a random choice. If for instance in a forced choice sensory experiment as described in more detail in Section 4.4 each subject is presented with three samples to choose from, only one of which actually contains the substance to be detected, the baseline probability will be $\alpha = 1/3$.

Observing that in such a situation $P(Y_{ij} = 1) = \alpha + (1-\alpha)P(\epsilon_{ij} \leq d_{ij} - Z_i - x_{ij}^T\beta)$, the obvious modification in equations (4.8) and (4.9) is to replace $U_{ij} = U_{ij}^{(0)}$ by

$$U_{ij}^{(\alpha)} = \left[\alpha + (1-\alpha)F_\epsilon\left(d_{ij} - z - x_{ij}^T\beta\right)\right]^{Y_{ij}} \left[(1-\alpha)\left(1 - F_\epsilon\left(d_{ij} - z - x_{ij}^T\beta\right)\right)\right]^{1-Y_{ij}}. \qquad (4.10)$$

One problem with incorporating a baseline into the likelihood approach is that Silvapulle's (1980) results on the existence of maximum likelihood estimates may not apply: One crucial assumption is the convexity of $-\log G$ and $-\log(1-G)$, where G is the dose-response curve. Let Φ denote the Gaussian distribution function and φ the Gaussian density. Taking G to be the probit curve (see Finney, 1978, or Morgan, 1992) with baseline, $G(x) = \alpha + (1-\alpha)\Phi(x)$, a simple calculation shows that

$$\frac{d^2}{dx^2}\left[-(\log G(x))\right] \geq 0$$

is equivalent to

$$g(x) = x\left(\alpha + (1-\alpha)\Phi(x)\right) + (1-\alpha)\varphi(x) \geq 0,$$

which for $\alpha > 0$ is not satisfied for sufficiently negative x.

This problem and the considerable numerical difficulties when attempting to solve the likelihood equations (4.9), (4.10), naturally lead to the consideration of a quasi-likelihood approach (Wedderburn, 1972; McCullagh and Nelder, 1989).

4.3 Quasi-likelihood approach and iterated weighted least squares algorithm

We consider a quasi-likelihood approach in the sense that we obtain the second moment structure of the observations given current parameter values. Our focus is on estimation with iteratively reweighted least squares, and asymptotic normality is assumed to hold under certain regularity conditions. This can be justified rigourously via the one-step iteration argument developed for instance in McCullagh and Nelder (1989), Section 2.5.

From the outset, we will consider known baseline probabilities α_{ij} for each response probability. Usually, if the baseline probabilities are non-zero due to a forced choice experiment, $\alpha_{ij} = 1/(\text{number of choices offered at observation } y_{ij})$. In the classical case without baselines, where no choices are made, set all $\alpha_{ij} = 0$. The basic model (4.5) for the thresholds remains the same, but the actual outcomes of the experiment are now the observations

$$Y_{ij} = 1_{\left\{x_{ij}^T\beta + Z_i + \epsilon_{ij} < d_{ij}\right\} \cup \{U_{ij} < \alpha_{ij}\}}, \tag{4.11}$$

where the U_{ij}*'s* are independent uniform on $[0,1]$ and independent of all other random variables ϵ and Z. Thus defining the convolved distribution

$$F_T(v) = \int_{-\infty}^{\infty} F_\epsilon\,(v - z)\,F_Z(dz), \tag{4.12}$$

and setting

$$\tilde{p}_{ij} = F_T\left(d_{ij} - x_{ij}^T\beta\right),$$

we have that

$$E(Y_{ij}) = p_{ij} = \alpha_{ij} + (1 - \alpha_{ij})\tilde{p}_{ij} \tag{4.13}$$

and

$$\text{Var } Y_{ij} = p_{ij}(1 - p_{ij}), \tag{4.14}$$

$$\text{Cov}\,(Y_{ij}, Y_{i'k}) = 0 \quad \text{for all } i \neq i', \quad \text{Cov}\,(Y_{ij}, Y_{ik}) = p_{ijk} - p_{ij}p_{ik} \quad \text{for } j \neq k, \tag{4.15}$$

where

$$\begin{aligned} p_{ijk} &= E\,(Y_{ij}Y_{ik}) \\ &= P\,([T_{ij} < d_{ij} \text{ or } U_{ij} < \alpha_{ij}] \text{ and } [T_{ik} < d_{ij} \text{ or } U_{ik} < \alpha_{ik}])\,. \end{aligned} \tag{4.16}$$

If there is no baseline, i.e., all $\alpha_{ij} = 0$, we would have accordingly

$$p_{ijk} = R\left(d_{ij} - x_{ij}^T\beta,\, d_{ik} - x_{ik}^T\beta\right),$$

where

$$R(u,v) = \int_{-\infty}^{\infty} F_\epsilon\,(u-z)\,F_\epsilon\,(v-z)\,F_Z(dz). \tag{4.17}$$

For given parameter values β and given F_ϵ and F_Z, the $p_{ij}\,'s$ and $p_{ijk}\,'s$ can be calculated by numerical integration. Alternatively, one may use Monte Carlo integration as follows. Let $z^{(1)},...,z^{(N)}$ be an independent sample from F_Z for a given large N, and let $\epsilon_1^{(1)},...,\epsilon_1^{(N)}$ and $\epsilon_2^{(1)},...,\epsilon_2^{(N)}$ be two independent samples from F_ϵ. Further let $U_1^{(1)},...,U_1^{(N)}$ and $U_2^{(1)},...,U_2^{(N)}$ be independent uniform (0,1) samples. Now set

$$\hat{p}_{ij} = \frac{1}{N}\sum_{q=1}^{N} 1_{\left\{x_{ij}^T\beta+z^{(q)}+\epsilon_1^{(q)}<d_{ij}\right\}} \tag{4.18}$$

and

$$\hat{p}_{ijk} = \frac{1}{N}\sum_{q=1}^{N} 1_{\{A\cap B\}}, \tag{4.19}$$

where

$$\begin{aligned} A &= \left\{\left(x_{ij}^T\beta + z^{(q)} + \epsilon_1^{(q)} < d_{ij}\right) \text{ or } (U_1 < \alpha_{ij})\right\}, \\ B &= \left\{\left(x_{ik}^T\beta + z^{(q)} + \epsilon_2^{(q)} < d_{ik}\right) \text{ or } (U_2 < \alpha_{ik})\right\}. \end{aligned}$$

If F_ϵ , F_Z are parametrically specified, (4.12)-(4.19) imply that for given parameter values one can calculate the mean and covariance structure of the observations to any degree of accuracy. This enables us to develop a version of the iteratively reweighted least squares algorithm for the current model.

Assume for the following that F_ϵ , F_Z and F_T depend on scale parameters σ_ϵ, σ_Z and σ_T (as ϵ, Z and $T = \epsilon + Z$ are zero mean random variables) such that

$$F_\epsilon(x) = \bar{F}_\epsilon(\frac{x}{\sigma_\epsilon}),\ F_Z(x) = \bar{F}_Z(\frac{x}{\sigma_Z}),\ F_T(x) = \bar{F}_T(\frac{x}{\sigma_T}), \tag{4.20}$$

$$\sigma_T^2 = \sigma_\varepsilon^2 + \sigma_Z^2, \tag{4.21}$$

where $\bar{F}_\epsilon$, $\bar{F}_Z$, and $\bar{F}_T$ are standardized versions of these distributions with unit variance. Moreover, assume that $\bar{F}_T$ has a density $\bar{f}_T$. One then obtains for the means

$$\mu_{ij} = EY_{ij} = \alpha_{ij} + (1-\alpha_{ij})\bar{F}_T\left(\sigma_T^{-1}(d_{ij} - x_{ij}^T\beta)\right). \tag{4.22}$$

The link function is therefore

$$g(\mu) = \bar{F}_T^{-1}\left(\frac{\mu-\alpha}{1-\alpha}\right). \tag{4.23}$$

Assume current values of all parameters are $\hat{\sigma}_{T,\ell}$, $\hat{\sigma}_{Z,\ell}$, $\hat{\sigma}_{\epsilon,\ell}$ and $\hat{\beta}_\ell$. Defining a $(p+1)$-vector of linear parameters $\gamma = (\gamma_1, ..., \gamma_{p+1})^T$ with

$$\gamma_1 = \sigma_T^{-1},\ \gamma_i = -\beta_{i-1}/\sigma_T,\ i = 2, ..., p+1, \tag{4.24}$$

and current estimates $\hat{\gamma}_\ell = (\hat{\gamma}_{\ell 1}, ..., \hat{\gamma}_{\ell p+1})^T$, current linear predictor and mean are then given by

$$\hat{\eta}_{\ell,ij} = \hat{\gamma}_{\ell 1} d_{ij} + x_{ij}^T (\hat{\gamma}_{\ell 2}, ..., \hat{\gamma}_{\ell p+1})^T \text{ and } \hat{\mu}_{\ell,ij} = \alpha_{ij} + (1-\alpha_{ij})\bar{F}_T(\hat{\eta}_{\ell,ij}). \tag{4.25}$$

Define

$$z_{\ell,ij} = \hat{\eta}_{\ell,ij} + (y_{ij} - \hat{\mu}_{\ell,ij}) \left[(1-\alpha_{ij})\bar{f}_T(\hat{\eta}_{\ell,ij})\right]^{-1}. \tag{4.26}$$

We then carry out a generalized linear regression of the $z_{\ell,ij}{}'s$ on the $(d_{ij}, x_{ij}^T)'s$, employing the covariance structure

$$\begin{aligned} \text{Cov}\ (z_{\ell,ij}, z_{\ell,i'j'}) &\approx [(g'(\hat{\mu}_{\ell,ij}) g'(\hat{\mu}_{\ell,i'j'}) \text{Cov}\ (y_{ij}, y_{i'j'}))] \qquad (4.27) \\ &= \left[(1-\alpha_{ij})(1-\alpha_{i'j'})\bar{f}_T(\hat{\eta}_{\ell,ij})\, \bar{f}_T(\hat{\eta}_{\ell,i'j'})\right]^{-1} \text{Cov}\,(y_{ij}, y_{i'j'}). \end{aligned}$$

The covariances $\text{Cov}(y_{ij}, y_{i'j'})$ require calculation of p_{ij} (4.13) and p_{ijk} (4.16), which can be done by approximating (4.12) and (4.17) with (4.18) and (4.19) for sufficiently large N.

Let now $M = \sum_{i=1}^{m} n_i$ denote the total number of available measurements. We order index pairs (i,j) lexicographically and then rewrite the observations as $(d_k,\ x_k^T,\ y_k),\ k = 1, ..., M$ with single index. Define $z_\ell = (z_{\ell 1}, ..., z_{\ell M})^T$, where the components are written with single index. Define matrices

$$\begin{aligned} W_\ell &= [(W_{\ell,rs})]_{1 \le r,s \le M}\,,\ \text{with elements} \\ W_{\ell,rs} &= \text{Cov}\ (z_{\ell r}, z_{\ell s}),\ \text{as given in (4.27) and} \\ X_\ell &= [X_{\ell,rs}]_{1 \le r \le M,\, 1 \le s \le p+1}\,,\ \text{with elements} \end{aligned}$$

$$X_{\ell,rs} = \begin{cases} d_r & s = 1 \\ 1 & s = 2 \quad , \quad 1 \le r \le M. \\ x_{r,s-1} & 3 \le s \le p+1 \end{cases}$$

where $x_r = (x_{r,1}, ..., x_{r,p})^T$ and d_r, x_r are r-th dose respectively r-th covariate vector in single index notation.

Then the generalized linear regression iteration step yields the updated linear $(p+1)$ dimensional parameter vector

$$\hat{\gamma}^{(\ell+1)} = \left(X_\ell^T W_\ell^{-1} X_\ell\right)^{-1} X_\ell^T W_\ell^{-1} z_\ell \tag{4.28}$$

with

$$\overset{\wedge}{\text{Cov}}\,(\hat{\gamma}^{(\ell+1)}) = \left(X_\ell^T W_\ell^{-1} X_\ell\right)^{-1}. \tag{4.29}$$

At convergence, assuming $\hat{\gamma}^{(\ell)} \to \hat{\gamma}$, $\overset{\wedge}{\text{Cov}}\,(\hat{\gamma}^{(\ell)}) \to \text{Cov}(\hat{\gamma}) = \hat{\Sigma}$ as $\ell \to \infty$ and applying the asymptotic normality assumption, this yields approximate large sample $100(1-\alpha)\%$ confidence regions

$$C_{1-\alpha} = \left\{\hat{\gamma} \in \mathrm{I\!R}^{p+1} : (\hat{\gamma}-\gamma)\hat{\Sigma}^{-1}(\hat{\gamma}-\gamma)^T \le c_{1-\alpha}\right\}. \tag{4.30}$$

Here, $c_{1-\alpha}$ is the $100(1-\alpha)\%$ quantile of the χ^2_{p+1} distribution, and $\hat{\Sigma}$ is estimated by $\hat{\Sigma} = \overset{\wedge}{\text{Cov}}\,(\hat{\gamma}^{(\ell+1)})$ as given in (4.29), with estimates (4.18) and (4.19) substituted into (4.27). Confidence intervals for linear combinations of individual parameters are determined by corresponding projections, as usual.

Updating of $\hat{\sigma}_Z$ can be done in at least two ways.

Method A. If one has repeated measurements for the subjects, first find an estimate of $\hat{\sigma}_\epsilon$ by averaging corresponding individual estimates $\hat{\sigma}_{\epsilon,i}$, $i = 1, ..., M$, where $\hat{\sigma}_{\epsilon,i}$ is obtained from the n_i repeated measurements made on the i-th individual; individuals with $n_i = 1$ do not contribute. Estimates $\hat{\sigma}_{\varepsilon,i}$ are obtained as ordinary maximum likelihood estimators for the scale parameter which do not require knowledge of the random effects Z_i. Then $\hat{\sigma}_\epsilon$ is estimated by means of a weighted average,

$$\hat{\sigma}_\epsilon = \sum_{i=1}^{M} w_i \hat{\sigma}_{\epsilon,i}, \tag{4.31}$$

where the weights w_i are determined in an obvious way from the inverse information matrices for the M (or less) individual fits. Alternatively, the estimate $\hat{\sigma}_\epsilon$ can be obtained from an analysis conditional on the random effects corresponding to an independent probit regression with different but fixed levels for individuals. This is done by a single probit analysis, entering all data.

Then in the ℓ-th iteration, this estimate $\hat{\sigma}_\epsilon$ is used to determine $\hat{F}_{\epsilon,\ell+1} \equiv \hat{F}_\epsilon \equiv \bar{F}_\epsilon(\cdot/\hat{\sigma}_\epsilon)$. Since the iteration produces updated estimates for $\hat{\sigma}_{T,\ell+1}$ according to (4.24), (4.28) and thus of $\hat{F}_{T,\ell+1} = \bar{F}_T(\cdot/\sigma_{T,\ell+1})$, $\hat{F}_{Z,\ell+1}$ and thus the updated $\hat{\sigma}_{Z,\ell+1}$ can be determined by deconvolution according to (4.12). Starting values may be obtained from an initial (unweighted) linear regression, assuming independent data.

Method B. Updating $\hat{\sigma}^2_Z$ is also possible by first obtaining predicted values $\hat{Z}_i$ for the random subject effects Z_i. Here, let ζ_i be the random location parameter of F_ε for the i-th subject, i.e.,

$$P\left(Y_{ij} = 1\right) = \alpha_{ij} + \left(1 - \alpha_{ij}\right) F_\varepsilon\left(\left(d_{ij} - \zeta_i\right)/\sigma_\varepsilon\right),$$

where $\zeta_i = Z_i + x_i^T\beta$.

We may now obtain "Restricted Maximum Likelihood Estimators" $\hat{\sigma}_{\varepsilon,i}$, ζ_i such that

$$\begin{aligned}\hat{\sigma}^2_{Z,\ell} &= \text{empirical (weighted) variance of } \zeta_i - x_i^T\hat{\beta}_\ell \\ \hat{\sigma}^2_{\varepsilon,\ell} &= \text{empirical (weighted) mean of } \hat{\sigma}^2_{\varepsilon,i}\end{aligned}$$

satisfy $\hat{\sigma}^2_{T,\ell} = \hat{\sigma}^2_{\varepsilon,\ell} + \hat{\sigma}^2_{Z,\ell}$, in accordance with (17). One possibility is to rescale estimators $\hat{\sigma}^2_{\varepsilon,\ell}$, $\hat{\sigma}^2_{Z,\ell}$ accordingly. Analogous to Method A, an alternative is to perform the analysis with different, fixed levels for the individuals providing the estimate $\hat{\sigma}_\epsilon$ together with $\hat{\zeta}_1, \ldots, \hat{\zeta}_m$ and their standard errors of estimation.

A special case of particular interest is the probit model (Finney, 1971). Letting Φ and φ be standard Gaussian distribution function respectively density, we obtain this model by setting

$$\bar{F}_\epsilon \equiv \Phi,\ \bar{F}_z \equiv \Phi,\ \bar{F}_T \equiv \Phi,\ \bar{f}_T \equiv \varphi \tag{4.32}$$

in (4.20), (4.21), (4.23)-(4.25). In this case, the link function $\bar{F}_T^{-1}$ is the classical probit link Φ^{-1} and the proposed method then corresponds to probit analysis with repeated measurements.

4.4 Data application: A forced choice sensory experiment

Throughout this section we consider the special case of a probit model (4.32) as described above. A sensory experiment was performed by one of the authors (P.B.) to determine the odour detection thresholds for certain esters important for the flavour of apples. It was performed as a forced choice experiment, see Frijters (1988), as follows: Each of 10 individuals was given pairs or triplets of glasses with water or a watery solution of a chemical substance in a given concentration. The subjects knew that exactly one glass contained a substance, and they were asked to point out the "substance glass" among the three glasses. The order of presentation of the various concentrations of the different chemical substances was randomized.

In this particular experiment, the covariate vector $x_{ij}^T = 1$ was one-dimensional and inference for the corresponding one-dimensional parameter $\hat{\beta}$ is of main interest, as this parameter then corresponds to the mean threshold in the population, ET_{ij} in model (1).

We report here the results obtained with the proposed method for the two substances propyl acetate and isopenthyl acetate and contrast them with results obtained

Table 4.1: Design of two sensory experiments, each conducted with $m = 10$ subjects. The n_i's are the number of repeated measurements made on the i-th subject out of the array of different concentrations. The same dose was given at most twice to the same subject.

	Propyl acetate	Isopenthyl acetate
$(n_1, ..., n_{10})$	(12,9,12,12,12,12,12,6,12,9)	(13,9,13,13,13,13,13,10,13,10)
log10–concentrations	-3,-2,-1,-0.6,-0.3,0	-5,-4,-3,-2.3,-2,-1.3,-1
α_{ij}'s	1/2 or 1/3	1/2

with a classical generalized linear model with independence assumption as well as results from the maximum likelihood approach of Anderson and Aitkin (1985), properly modified for the baselines α_{ij}. Some information on the designs of these experiments is given in Table 4.1.

The approach of Anderson-Aitkin corresponds to the case where $\bar{F}_Z = \Phi$ and $\bar{F}_\epsilon$ is the standardized logistic distribution function. This leads to the approximate log-likelihood,

$$\sum_{i=1}^{m} \log \left\{ \sum_q \left(\prod_{j=1}^{n_i} U_{ij} \right) A_q \right\}$$

where q and A_q are the normal quadrature points and weights respectively, and

$$U_{ij} = \frac{\exp\{y_{ij}\eta_{ij}\}}{1 + \exp\{\eta_{ij}\}} - \frac{(-1)^{y_{ij}}\alpha_{ij}}{1 + \exp\{\eta_{ij}\}} \tag{4.33}$$

with

$$\eta_{ij} = \frac{\pi}{\sqrt{3}\sigma_\epsilon} \left(d_{ij} - q\sigma_Z - x_{ij}^T \beta \right)$$

Reparametrizing and differentiating (4.33), the resulting estimating equations become those of a weighted logistic regression with baselines. Different from Anderson and Aitkin is the presence of the dose covariate d_{ij}. As a consequence, both variance parameters are identifiable. Due to the baseline it is not possible to apply the standard package EGRET, and the algorithm as described in Anderson and Aitkin (1985), extended to cover baselines, as well as the algorithm of the present paper, were implemented in the S-PLUS package. The results for the procedures and the two experiments are listed in Table 4.2. Note that according to (4.24), $\hat{\sigma}_T = 1/\hat{\gamma}_1$, $\hat{\beta} = -\hat{\gamma}_2/\hat{\gamma}_1$.

Table 4.2: Results for the proposed Repeated Measures Approach as compared to the Anderson-Aitkin method and to assuming independent data (Independence Model) for two sensory experiments. $\hat{\Sigma}$ is the estimated covariance matrix for the linear parameters $(\hat{\gamma}_1, \hat{\gamma}_2)^T$.

Method	Estimates	Propyl acetate	Isopenthyl acetate
Repeated Measures Model (Method A)	$\hat{\sigma}^2_\epsilon$	0.68	1.55
	$\hat{\sigma}^2_Z$	0.00	0.17
	$\hat{\gamma}_1$	1.27	0.76
	$\hat{\gamma}_2$	0.89	2.57
	$\hat{\beta}$	-0.70	-3.38
	$\hat{\Sigma}$	0.30 0.15 0.15 0.13	0.068 0.18 0.18 0.56
	$1.96\sqrt{\widehat{\text{Var}}\ (\hat{\beta})}$	0.39	0.82
Anderson and Aitkin Model	$\hat{\sigma}^2_\epsilon$	0.76	1.74
	$\hat{\sigma}^2_Z$	0.00	0.10
	$\hat{\gamma}_1$	2.07	1.38
	$\hat{\gamma}_2$	1.48	4.57
	$\hat{\beta}$	-0.71	-3.32
	$\hat{\Sigma}$	0.56 0.32 0.32 0.32	0.46 1.22 1.22 3.44
	$1.96\sqrt{\widehat{\text{Var}}\ (\hat{\beta})}$	0.37	0.93
Independence Model (Probit)	$\hat{\gamma}_1$	1.27	0.77
	$\hat{\gamma}_2$	0.89	2.58
	$\hat{\beta}$	-0.70	-3.36
	$1.96\sqrt{\widehat{\text{Var}}\ (\hat{\beta})}$	0.39	0.78

The estimates of the linear parameters and of β do not depend much on whether the correlation structure is taken into account. The confidence intervals, crucial for inference, however, do widen considerably if the dependencies coming from the repeated measures structure are taken into account. We note that for Propyl acetate the variance component $\hat{\sigma}_Z^2$ turns negative at an iteration step. When this happens, $\hat{\sigma}_Z^2$ is set to zero for the next iteration step, and if it happens again, the algorithm terminates. Thus, in the case of propyl acetate, the random effect is found to be negligible. It is noteworthy that this finding is coming out also of the Anderson and Aitkin approach. The standard errors in the Anderson and Aitkin approach are found by calculating the second derivatives of the log-likelihood function. Note that the standard errors arising from the weighted logistic regression in the estimation algorithm are not the correct ones, as they correspond to regarding the weights as known. And the weights do themselves depend on the parameters. For these computations and the simulations below we used 5 quadrature points.

Calculating overdispersion factors, as described below, leads in both cases to the independent model, as they were both smaller than 1, 5.83/8 and 8.68/9 respectively.

4.5 Simulation

A simulation study was performed to investigate the tendencies from the data application above. Using $\beta = -2.5$, $\sigma_\epsilon^2 = 3$ and $\sigma_Z^2 = 3$, data sets with assumed doses corresponding to seven different log 10 dose concentrations as for the isopenthyl acetate experiment, see Table 4.1, were generated according to (2). All baselines were assumed to be zero. Each of 10 individuals was assumed to be presented with 14 doses, including two replicates at each dose level, i.e., $n_i = 14$, $i = 1, ..., 10$.

Three alternative methods were compared to two versions of our proposed method. The five methods applied were:

Method 1 The independent probit model,

$$\Phi^{-1}(p_{ij}) = \gamma_2 + \gamma_1 d_{ij}$$

Method 2 Overdispersion correction by means of correcting the variance of the parameter estimates by the overdispersion factor, h, which is obtained by Pearson's χ^2(McCullagh and Nelder, 1989):

$$h = \frac{1}{m-1} \sum_{i=1}^{m} \frac{n_i \left(\mathrm{OBS}_i - \mathrm{EXP}_i\right)^2}{\left(n_i - \mathrm{EXP}_i\right) \mathrm{EXP}_i},$$

where EXP_i is the expected number of responses for individual i in the independent probit model,

$$\mathrm{EXP}_i = \sum_{j=1}^{n_i} \Phi(\hat{\gamma}_2 + \hat{\gamma}_1 d_{ij}),$$

and OBS_i is the total observed number of responses for individual i.

Method 3 The Anderson and Aitkin maximum likelihood approach.

Method 4 The marginal approach of this paper with a version of Method B for updating variance parameters. Since the covariate x_{ij}^T is just a constant, Method B is equivalent to pre-estimating the intra-class correlation $\sigma_Z^2/(\sigma_Z^2 + \sigma_\epsilon^2)$, and using this fraction to deconvolve the estimated total variance, $\hat{\sigma}_{\ell,T}^2$ in each iteration step. Special attention is required at the pre-estimation step due to the considerable probability of extreme observations for an individual. The following solution which proved to be feasible was adopted:

(i) Estimate ζ_i and σ_ϵ^2 from the model $\Phi^{-1} = \delta_i + \gamma d_{ij}$, as $\hat{\zeta}_i = -\hat{\delta}_i/\hat{\gamma}$ and $\hat{\sigma}_\epsilon^2 = 1/\hat{\gamma}^2$, omitting those individuals for which either no response or response at all levels is recorded.

(ii) Set for the "extreme" individuals,

$$\zeta_i = \begin{cases} -5 & , \text{if } \sum_{j=1}^{n_i} y_{ij} = n_i \quad \text{(response at all levels)} \\ -1 & , \text{if } \sum_{j=1}^{n_i} y_{ij} = 0 \quad \text{(no response)} \end{cases}$$

This choice comes from an interpretation of ζ_i as the threshold for individual i: if an individual can detect any given dose, then the threshold can be interpreted to be smaller than the smallest dose, in this case -5, and vice versa.

(iii) Calculate the empirical variance of $\hat{\zeta}_1, \ldots, \hat{\zeta}_m$,

$$\sigma_Z^2 = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\zeta}_i - \bar{\hat{\zeta}})^2$$

and the estimate of the intra class correlation.

(iv) At each step of the algorithm the intra class correlation fraction is used to split the estimated total variance $\sigma_{T,\ell}^2$ into the two parts.

As starting values for the linear parameters the estimates from fitting the independent model are used.

Table 4.3: Results for 600 simulations of a repeated measures design. True value is $\beta = -2.5$.

	Average β	Average $1.96\sqrt{\hat{\text{Var}}(\hat{\beta})}$	Non-coverage percentage
Method 1	-2.50	0.58	33.2
Method 2	-2.50	1.20	4.67
Method 3	-2.51	0.84	27.5
Method 4	-2.51	1.18	5.17
Method 5	-2.51	1.24	4.83

Method 5 The marginal approach of this paper with a version of Method B for updating variance parameters: σ_ϵ^2 is pre-estimated as in Method 4 above, and kept fixed throughout. In each step of the algorithm this fixed estimated is used for an additive updating, as described in Section 4.4 above:

$$\hat{\sigma}^2_{Z,\ell} = \max\left\{\hat{\sigma}^2_{T,\ell} - \hat{\sigma}^2_\epsilon, 0\right\}$$

Table 4.3 shows the results for the 600 simulations.

It is clear that the analysis which ignores the dependencies in the data will lead to actual confidence levels which are unacceptably off the nominal levels so that inference is invalid when using this method for a repeated measurements design. This finding is not unexpected. More unexpected is the poor behaviour of the Anderson and Aitkin approach for determining confidence bands. The difference in the lengths of corresponding confidence intervals is striking. The overdispersion approach and the two versions of the present paper work comparably well and in accordance with the nominal level. The feasibility of the overdispersion method should be seen in the light of the relative simple experimental setup assumed for the simulation, notably the assumption of vanishing baseline probabilities. It is apparent that in the data analysis reported above only the results coming from our proposed repeated measures analyses should be used.

4.6 Discussion

We proposed an approach for dose-response designs with repeated measurements based on a random effects model for unobservable thresholds. Several authors have investigated generalized linear models with random effects using various approaches. In some of these works (cf. Zeger and Karim (1991) and references therein), a conditional approach is adopted: The GLM setting is expressed conditionally on random effects with specified distributions and the parameters of these distributions are assumed to follow certain priors in a Bayesian sense. In contrast, our approach is unconditional and we estimate the unknown parameters directly, including the variance components. Neither the distributions of the random effects nor of the observed data are assumed to belong to the exponential family.

Furthermore, our approach allows to incorporate a known background response rate, laedning to nonzero baseline probabilities which are allowed to vary from observation to observation. This amounts to a change of link function, in fact the resulting "link function" is not the same from observation to observation, and therefore these models cannot be fitted with standard packages like EGRET. Assuming no baseline, i.e., $\alpha_{ij} = 0$, a conditional approach would lead us to the link function $\bar{F}_\epsilon^{-1}$(4.20) in contrast to the unconditional link function $\bar{F}_T^{-1}$(4.20),(4.22). Thus the treatment of the random effects in our model leads to a new link function to be used in the weighted iterated least squares algorithm as compared to the conditional approach.

An important feature of our proposed approach is the Monte Carlo calculation of the covariance of the binary observations, (4.14)-(4.19). In other generalized estimating equations approaches this covariance is approximated to first order, Zeger *et al.* (1988) and Breslow and Clayton (1993).

Both simulations and data analysis show that the proposed method is feasible for dose-response designs with repeated measurements, is easy to implement, not too computationally intensive and provides reasonable results. It is vastly superior to the alternative which treats all the obtained data as independent.

In particular, when interest focuses on the behaviour of reaction random thresholds across a population, the proposed statistical model for the random thresholds is an attractive option. It leads to feasible inference procedures, and allows straightforward interpretation of the parameters as biological thresholds.

Acknowledgements

The research of Per M. Brockhoff was supported by the Danish Research Academy and the Danish Ministry of Education as part of the FØTEK Programme. The research of Hans-Georg Müller was supported in part by NSF Grant DMS 9305484. A number of helpful comments from the two referees are gratefully acknowledged.

References

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *J. Royal Statist. Soc.* **B47**, 203-210.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. American Statist. Assoc.* **88**, 9-25.

Collet, D. (1991). *Modelling Binary Data*, Chapman & Hall, London.

Elashoff, J.D. (1981). Repeated-measures bioassay with correlated measures and heterogeneous variances: A Monte Carlo study. *Biometrics* **37**, 475-482.

Finney, D.J. (1971). *Probit Analysis.* Cambridge University Press, Cambridge.

Frijters, E.R. (1988). Sensory difference testing and the measurement of sensory discriminability. In: *Sensory Analysis of Foods,* Ed. J.R. Piggott, Elsevier Applied Science, London.

Goldstein, H. (1991). Non-linear multilevel models, with an application to discrete response data. *Biometrika* **58,** 45-51.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. American Statist. Assoc.* **72,** 320-340.

Im, S. and Gianola, D. (1988). Mixed models for binary data with an application to lamb mortality data. *Appl. Statist.* **37**, 196-204.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data.* Chapman and Hall, London.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78,** 719-727.

Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response model. *J. Royal Statist. Soc.* **B43**, 310-313.

Stiratelli, R., Laird N. and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-971.

Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61,** 439-447.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. American Statist. Assoc.* **86,** 79-86.

4.7 Add: GLIMs in sensometrics

For the interval scale data in the typical multivariate sensory data set, a first thing to notice is that GLIM's were originally developed for univariate observations. This excludes the situations in the subsequent Chapters 5 and 6 from being put in the GLIM frame. The conventional univariate fixed effects ANOVA is, as mentioned, a GLIM, and the mixed model ANOVA is a special case of a GLIM with random effect, which we have discussed in further detail in the preceding paper. The assessor models of Chapter 3 are not GLIM's, but the estimation principles employed are in essence of the quasi-likelihood type. And it is worth noting that adopting the variance function feature of the GLIM, could lead to a direct modelling of the product dependent variance heterogeneity, compensated for by transformation in Chapter 3, by a suitable function. This may provide a more flexible and data driven way of stabilizing variance.

The application of GLIM's for ordinal scale data might be the sleeping beauty in sensometrics. I have seen no examples of this in the food science literature but both McCullagh and Nelder (1989) and Fahrmeir and Tutz (1994) include sensory examples of this kind. In particular in the light of the scale discussion in the Introduction, this approach seems intriguing, although the models might be difficult to grasp for non-statistically minded experimenters.

The nominal scale data and in particular the binary data should offer a comprehensible motivation and explanation for the entrance into the GLIM world. Let us consider the basic difference test situation, So, m individuals were 'exposed' to the same 'dose' a number of times, say n_i, $i = 1, \ldots, m$. There is a huge amount of literature on the method of paired comparisons and I will and cannot give an exhaustive review of it. Historically, these methods date back to Fechner in the nineteenth century and Thurstone in the 1920's, see David (1969) and references therein. I would like, however, to try to relate the GLIM frame to the 'signal detection' approach of estimating sensory differences, see Frijters (1988).

Given a 'stimulus' product and a number of 'placebo' products, two questions arise: Has the stimulus product a detectable difference, and what/how much is the difference. The former question may be answered by classical binomial testing, with the duo, duo-trio and triangular methods being widely used, see for instance Lea (1988), where the relative powers of these approaches are also discussed. One problem in power discussions is that the underlying psychophysical processes are not the same for different test methods. Ennis (1990, 1993) showed that the power of, for example, the 3-alternative forced choice (3-AFC) method is superior to the triangular method. In the 3-AFC, subjects are asked to point out the product with the most (or least) of some attribute, whereas in the triangular method, subjects are asked to point out the most different product. The presentation here is not a contribution to these power discussions but rather an attempt generally to relate a classical statistical approach to the traditional psychophysical approach, as the latter forms the basis of the estimation of sensory differences.

So for simplicity let us consider a K-AFC experiment, where subjects are asked to point out the product with the most of some attribute. The results in Ennis (1990, 1993) are based on the basic psychometric relationships that comes from assuming that the $K-1$ placebo's are realizations of independent standard Normal variates, $X_i \sim N(0,1)$, $i=1,\ldots,K-1$ and the stimulus is a realization of a $N(d,1)$-variate, $d>0$. By a straightforward conditioning argument the probability of the stimulus outcome being larger than all the others are, see for instance Ennis (1993),

$$g(d) = \int_{-\infty}^{\infty} \Phi^{K-1}(x)\phi(x-d)dx \tag{4.34}$$

where Φ and ϕ denotes the standard Normal distribution and density functions respectively. The 'signal detection' method of sensory difference estimation is now simply based on explicit calculations of the function g, relating an observed fraction of correct responses to a sensory difference value d.

A statistical approach, as indicated in the Thesis Introduction and employed in the preceding paper is as follows. We let Z_i denote the individual threshold and T_{ij} the threshold for individual i at the jth replication, $j=1,\ldots,n_i$, $i=1,\ldots,m$. The model (1.1) of the Thesis Introduction, without interaction effect, becomes

$$T_{ij} = Z_i + \varepsilon_{ij} \tag{4.35}$$

where the Z_is and ε_{ij}s are independent Normals and the overall mean μ is assumed zero owing to unidentifiability. For the same reason all the variables may be assumed to be standard Normals, and the probability of a threshold T_{ij} being less than a difference d is then

$$h(d) = \Phi\left(\frac{d}{\sqrt{2}}\right) \tag{4.36}$$

From a GLIM viewpoint the functions h and g are inverse link functions. The two approaches differ in that the 'signal detection' approach assumes randomness on the samples whereas the 'statistical' approach assumes randomness on individuals. However, the statistical random error ε_{ij} can also be said to include randomness on the samples. For $K=2$ the link functions, $h(d)$ and $g(d)$, indeed coincide, and the two approaches are in principle the same.

For $K>2$ the two link functions differ, and if one believes in the psychophysical relationships, the function $g(d)$ should be chosen. The GLIM approach, though, with any link function, may offer a way to calculate uncertainties in the estimation of d. Since the linear model specification in this case is just the one-parameter model of constant level, employing the IWLS estimation algorithm with g^{-1} as the link function leads to the same estimation of d as the deterministic 'signal detection' approach. But at the same time it provides an estimate of standard error of estimation, that is based on a first order Taylor approximation to the link function g^{-1}.

The model specified (4.35) and (4.36) is really a random effect GLIM, as discussed in the paper. Thus a GLIM approach could also provide standard errors of estimation for d, that properly takes the random assessor effects into account. Combining in this way the psychophysical functions with formal random effects GLIM modelling, may offer a nice common framework for the work with these methods, but future research on this will have to show the reach of this approach.

Chapter 5

Two-dimensional covariance component models applied to sensory data

This is a working paper intended for submission to *Journal of Sensory Studies* early 1995. This American journal has as purpose "to promote technical and practical advancements of sensory science by publishing papers of broad coverage to include observational and experimental studies in the application of sensory evaluation to the food, medical, agricultural, biological, pharmaceutical, cosmetic, consumer and material sciences. This includes research work dealing with new developments in sensory methods, consumer testing, experimental design, statistical analysis, scaling, psychophysics and computer applications."

Two-dimensional covariance components models applied to sensory data.

Per M. Brockhoff
Centre of Food Research, Dept. of Mathematics and Physics,
Royal Veterinary and Agricultural University,
Thorvaldsensvej 40, DK-1871 Frederiksberg C, Denmark.

Barbara Guggenbühl
Dept. of Enology and Viticulture,
University of California,
1023 Wickson Hall, Davis, CA 95616, USA.

Running title: Covariance components

Abstract

The two-dimensional generalization of the analysis of variance model with random judge effects is presented as a way to analyse the correlation between a sensory attribute and a chemical measurement, or any two paired observations, in a designed experiment. The model is formulated and discussed in detail and emphasis is put on interpretations. The approach is illustrated by two time intensity sensory experiments on the effect of salivary flow rates on sensory perception. An illustrative technique as aid for ANOVA modelling, the factor structure diagram, is presented and used throughout the paper. A test for significance of an overall relationship is presented and applied in the two examples.

5.1 Introduction

Correlating instrumental measurements with sensory evaluations is an important tool for research in human perception. As a result the subject has been given much attention in the sensory literature, see for example Kjølstad et al. (1990), Martens and Martens (1986) or Brockhoff et al. (1993). A typical setup is that of calibrating panel profile mean scores with near infrared(NIR) or gas chromatographic(GC) measurements, imbedding the setup into the chemometrical field of multivariate calibration, see Næs and Martens (1988).

Often in these studies a designed experiment was performed, and prior to the calibration step analyses of variance were carried out to investigate the significances of the designed effects on sensory as well as instrumental measurements. But in the subsequent calibration analysis the design is often ignored; the partial least-squares or principal components regression analysis is achieved by aggregating all observations. This is perfectly allright for an explorative analysis, but one must be careful to quantify the strength of association, for example by a correlation coefficient, as it depends heavily on the variability spanned by the observations, which is controlled by the designer of the experiment. Such a correlation coefficient expresses mainly a sample–to–sample relationship, but includes also a within-sample (residual) relationship. It is the study of such associations on different levels, that is the main objective of this paper.

Experimental studies, as referred to above, are usually performed such that for each single instrumental measurement there are observations for all judges. This makes the averaging over judges a reasonable approach, although other approaches exist, see Næs and Kowalski (1989). In the experiments of the present paper there are for each judge a sensory evaluation as well as an instrumental measurement, that is related to both the sample and the judge. This adds the further complexity, that some of the observed correlation is induced by judge effects. We will give a systematic approach to the detection and interpretation of the various sources of correlation. For simplicity we consider only a two-dimensional setup, but the principles apply in analogous way to the general multivariate setup.

Based on a one-dimensional formulation of the random judge effect model, we present the two-dimensional equivalent. Quite some emphasis is put upon the basic factor structure of an experiment, as this is important for the interpretations, which is described in detail. Finally the method is applied to two experiments.

5.2 Variance component models

Consider a balanced sensory experiment with I judges (factor J), P samples (factor S) and K replicates (factor R), i.e. we have observations, x_{ipk}, from random variables X_{ipk} , $i = 1, \ldots, I$, $p = 1, \ldots, P$, $k = 1, \ldots, K$. The most general fixed effects analysis of variance model for this situation includes the three main effects and the three interaction terms,

$$X_{ipk} = \mu + \alpha_i + \beta_p + \delta_k + (\alpha\beta)_{ip} + (\alpha\delta)_{ik} + (\beta\delta)_{pk} + \varepsilon_{ipk} \qquad (5.1)$$

where the errors, ε_{ipk}s, are independent and Normal distributed. We have included all effects with the R factor as this often represents a real effect, for instance day of sensory session.

For inferential purposes going beyond the panel in use, it is more relevant to consider a model where effects due to judges are random, see for instance Lundahl and McDaniel (1988),

$$X_{ipk} = \mu + A_i + \beta_p + \delta_k + (A\beta)_{ip} + (A\delta)_{ik} + (\beta\delta)_{pk} + \varepsilon_{ipk} \qquad (5.2)$$

where all additional random variables on the right hand side, indicated by the use of a capital A, are independent Normals:

$$A_i \sim N(0, \sigma_J^2) \ , \ (A\beta)_{ip} \sim N(0, \sigma_{J\times S}^2) \ , \ (A\delta)_{ik} \sim N(0, \sigma_{J\times R}^2) \ , \ \varepsilon_{ipk} \sim N(0, \sigma_E^2) \quad (5.3)$$

This *random effect model* is also frequently called a *mixed model* — as both fixed and random effects are present. It is also called a *repeated measures model* with no *between factors* and the two *within factors* R and S. In line with Searle et al. (1992) we will denote the model (5.2)-(5.3) a *variance component model.* Each observation is modelled to consist of a component of variation for each random effect (apart from the fixed effects). Theoretically this can be seen from (5.2) and (5.3) as the variance of the sum of independent random variables is the sum of the variances:

$$\mathrm{Var}\, X_{ipk} = \sigma_J^2 + \sigma_{J\times S}^2 + \sigma_{J\times R}^2 + \sigma_E^2 \qquad (5.4)$$

The analysis of data based on either (5.1) or (5.2)-(5.3), or similar models, to detect and interpret significant effects is called analysis of variance (ANOVA), and can be performed by many statistical software package on the market. Estimation and interpretation of the variance components (5.4) themselves are less common, but can be done for example with the Procedure Mixed of SAS.

5.3 Two-dimensional covariance component models

Assume now that in addition to x_{ipk} we also observe y_{ipk}, that is, we observe two-dimensional random variables

$$\begin{pmatrix} X_{ipk} \\ Y_{ipk} \end{pmatrix}, \; i=1,\dots,I, \; p=1,\dots,P, \; k=1,\dots,K$$

Two kinds of inferential purposes emerge: (i) What and how are the treatment effects on the two attributes in question and (ii) how are the two attributes related ? Purpose (i) can be achieved by performing an ANOVA as described above for both attributes or by doing a single multivariate analysis of variance (MANOVA). The multivariate equivalent of both (5.1) and (5.2)-(5.3) can be handled by Procedure GLM in SAS by means of Wilks statistics, the multivariate equivalents of the F statistics, see for example Anderson (1958). It is the purpose (ii) that attracts our attention in the present paper. It turns out that a deeper understanding of the multivariate (two-dimensional) equivalent of the variance component model (5.2)-(5.3), the *two-dimensional covariance component model*, is the key issue for this purpose.

Therefore we will formulate this model in detail:

$$\begin{aligned}\begin{pmatrix} X_{ipk} \\ Y_{ipk} \end{pmatrix} &= \begin{pmatrix} \mu^x \\ \mu^y \end{pmatrix} + \begin{pmatrix} A_i^x \\ A_i^y \end{pmatrix} + \begin{pmatrix} \beta_p^x \\ \beta_p^y \end{pmatrix} + \begin{pmatrix} \delta_k^x \\ \delta_k^y \end{pmatrix} + \begin{pmatrix} (A\beta)_{ip}^x \\ (A\beta)_{ip}^y \end{pmatrix} \\ &\quad + \begin{pmatrix} (A\delta)_{ik}^x \\ (A\delta)_{ik}^y \end{pmatrix} + \begin{pmatrix} (\beta\delta)_{pk}^x \\ (\beta\delta)_{pk}^y \end{pmatrix} + \begin{pmatrix} \varepsilon_{ipk}^x \\ \varepsilon_{ipk}^y \end{pmatrix} \qquad (5.5)\end{aligned}$$

where all random variables on the right hand side are independent and bivariate Normals:

$$\begin{aligned} \begin{pmatrix} A_i^x \\ A_i^y \end{pmatrix} &\sim N_2(0, \Sigma_J), & \begin{pmatrix} (A\beta)_{ip}^x \\ (A\beta)_{ip}^y \end{pmatrix} &\sim N_2(0, \Sigma_{J\times S}), \\ \begin{pmatrix} (A\delta)_{ik}^x \\ (A\delta)_{ik}^y \end{pmatrix} &\sim N_2(0, \Sigma_{J\times R}), & \begin{pmatrix} \varepsilon_{ipk}^x \\ \varepsilon_{ipk}^y \end{pmatrix} &\sim N_2(0, \Sigma_E) \end{aligned} \qquad (5.6)$$

with

$$\begin{aligned} \Sigma_J &= \begin{pmatrix} \sigma_{x,J}^2 & \sigma_{xy}^J \\ \sigma_{xy}^J & \sigma_{y,J}^2 \end{pmatrix}, & \Sigma_{J\times S} &= \begin{pmatrix} \sigma_{x,J\times S}^2 & \sigma_{xy}^{J\times S} \\ \sigma_{xy}^{J\times S} & \sigma_{y,J\times S}^2 \end{pmatrix}, \\ \Sigma_{J\times R} &= \begin{pmatrix} \sigma_{x,J\times R}^2 & \sigma_{xy}^{J\times R} \\ \sigma_{xy}^{J\times R} & \sigma_{y,J\times R}^2 \end{pmatrix}, & \Sigma_E &= \begin{pmatrix} \sigma_{x,E}^2 & \sigma_{xy}^E \\ \sigma_{xy}^E & \sigma_{y,E}^2 \end{pmatrix} \end{aligned} \qquad (5.7)$$

The model given by (5.5), (5.6) and (5.7) is basically the model (5.2)-(5.3) expressed for both X_{ipk} and Y_{ipk} together with a corresponding model for the covariance between X_{ipk} and Y_{ipk}, as indicated by the two-dimensional equivalent of (5.4),

$$\mathrm{Cov}\,(X_{ipk}, Y_{ipk}) = \Sigma_J + \Sigma_{J\times S} + \Sigma_{J\times R} + \Sigma_E \tag{5.8}$$

where the four components of covariance come from the four random effects of the model. Mean squares in this model become 2×2 matrices, with diagonal elements that equal the mean squares from the two univariate ANOVA's, and off-diagonal elements that are mean cross products, the empirical covariances.

The two-dimensional covariance component model is not very often used. Standard textbooks on multivariate analysis do not explicitly formulate this model, but Searle *et al.* (1992), p. 378-380 state that it is straightforward to generalize the ANOVA method of variance components estimation to the two-dimensional (multivariate) setup.

5.4 Factor structure

A basic feature of the ANOVA modelling is the partitioning of the total variation in the data into parts due to each effect in play. This partitioning is determined by the design of the experiment, and as such has nothing to do with dimension of the observations. If the design is *orthogonal*, see Tjur (1984), this partioning is unique, meaning that there is no confounding/ambiguity in the ascription of variations to each effect. The basic design of the present paper, with no missing values, is orthogonal.

As the design of an experiment is fundamental to the subsequent analysis, it makes sense to adopt the approach of Tjur (1984, 1991) of visualizing the design by a *factor structure diagram*, see Figure 5.1. In the Appendix we have given a short description of the construction and possible use of such diagrams.

The calculated mean squares for a design, whether one- or multi-dimensional, are also fundamental. It is in the interpretation of these mean squares the covariance component model differs from the fixed effect model: some are modelled/interpreted as variances(random effects) and some are modelled/interpreted as the variation between different but fixed levels(fixed effects).

5.5 Interpretations

5.5.1 Parameters of the covariance components model

The covariance component model (5.5)-(5.7) has as an underlying interpretation, that the judges ‘has been sampled at random’ from ‘the population of judges’, and from

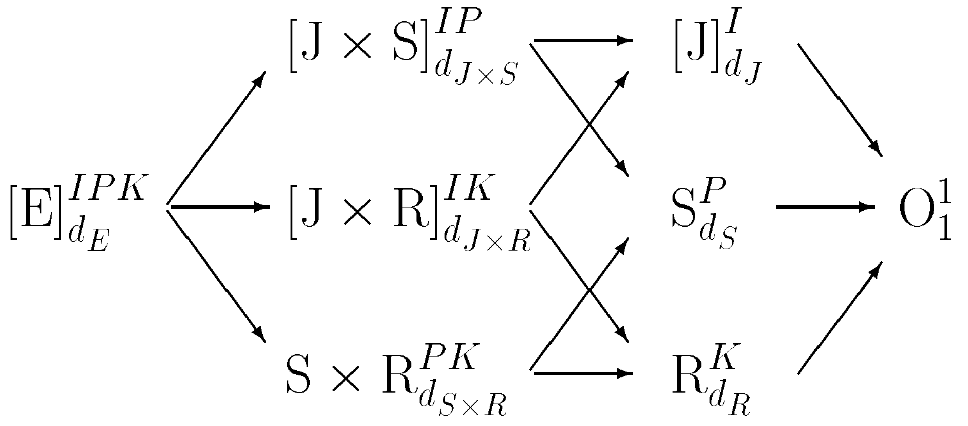
$[\mathrm{E}]_{d_E}^{IPK}$
$[\mathrm{J} \times \mathrm{S}]_{d_{J\times S}}^{IP}$
$[\mathrm{J} \times \mathrm{R}]_{d_{J\times R}}^{IK}$
$\mathrm{S} \times \mathrm{R}_{d_{S\times R}}^{PK}$
$[\mathrm{J}]_{d_J}^{I}$
$\mathrm{S}_{d_S}^{P}$
$\mathrm{R}_{d_R}^{K}$
O_1^1

Figure 5.1: Factor structure diagram for the variance component model (5.5)-(5.7). Superscripts denote the number of levels for a factor, and subscripts denote degrees of freedom. Brackets indicate that effects are random.

this the following interpretations come forth:

Σ_E:

$\sigma_{x,E}^2$ ($\sigma_{y,E}^2$): The variation in attribute X (Y) when all effects have been accounted for, i.e. variation within judges, samples and replicates.

σ_{xy}^E: The covariation between attribute X and Y when all effects have been accounted for, i.e. the relationship between X and Y that would be observed if one had repeated observations for a single judge, sample and replicate.

Σ_J:

$\sigma_{x,J}^2$ ($\sigma_{y,J}^2$): The variation in attribute X (Y) from judge to judge within samples and replicates.

σ_{xy}^J: The covariation between attribute X and Y from judge to judge within samples and replicates, i.e. the relationship between X and Y seen from judge to judge for the same sample and replicate.

$\Sigma_{J\times S}$:

$\sigma_{x,J\times S}^2$ ($\sigma_{y,J\times S}^2$): The variation in attribute X (Y) induced by the judges' differences in sample differences

$\sigma_{xy}^{J\times S}$: The covariation between attribute X and Y induced by the judges' differences in sample differences, i.e. a measure of the relationship between the interaction effects in X and Y, for example it will be positive if judges have the same tendencies in both X and Y to evaluate, say, the difference between sample 1 and sample 2 differently (a judge with a large positive difference in X also has a large positive difference in Y)

$\Sigma_{J\times R}$: Analogous to $\Sigma_{J\times S}$.

For each covariance component there is also a corresponding correlation given by the fundamental relationship between correlation and covariance. For the error component, for example, the component of correlation, ρ_E, becomes

$$\rho_E = \frac{\sigma_{xy}^E}{\sqrt{\sigma_{x,E}^2}\sqrt{\sigma_{y,E}^2}} \tag{5.9}$$

The basic problem is that we do not observe these nice interpretable entities directly, we 'observe' the mean squares instead.

5.5.2 Mean squares

Each mean square for the random effects is a combination of different covariance components. Exactly which combination can be seen from the expectations of the mean squares, that has the same structure as in the univariate case, see Anderson (1958),

$$\begin{aligned}
\mathrm{E}\left(\mathrm{MS}_E\right) &= \Sigma_E && (5.10)\\
\mathrm{E}\left(\mathrm{MS}_{J\times S}\right) &= \Sigma_E + K\Sigma_{J\times S} && (5.11)\\
\mathrm{E}\left(\mathrm{MS}_{J\times R}\right) &= \Sigma_E + P\Sigma_{J\times R} && (5.12)\\
\mathrm{E}\left(\mathrm{MS}_J\right) &= \Sigma_E + K\Sigma_{J\times S} + P\Sigma_{J\times R} + KP\Sigma_J && (5.13)
\end{aligned}$$

This means that the off-diagonal element of, say, MS_J, is the observed covariance over judges, that as above can be transformed to a correlation, which in fact precisely is the correlation between X and Y observations averaged over samples and replicates for each judge. But this covariance/correlation not only expresses judge covariance/correlation; it includes some of the other three components as well, as seen in (5.13). Similar comments apply to $\mathrm{MS}_{J\times S}$ and $\mathrm{MS}_{J\times R}$.

The mean squares for the fixed effects consists of components from the random effects as well as the fixed effects

$$\begin{aligned}
\mathrm{E}\left(\mathrm{MS}_{S\times R}\right) &= \Sigma_E + IQ_{S\times R} && (5.14)\\
\mathrm{E}\left(\mathrm{MS}_S\right) &= \Sigma_E + K\Sigma_{J\times S} + IQ_{S\times R} + IKQ_S && (5.15)\\
\mathrm{E}\left(\mathrm{MS}_R\right) &= \Sigma_E + P\Sigma_{J\times R} + IQ_{S\times R} + IPQ_R, && (5.16)
\end{aligned}$$

where the Qs are the 'variational components' stemming from the variation between the levels of a fixed effect. Conceptually these expected values have the same structure as for the random effects, and similar calculations and interpretations can be performed although the Q components and mean squares are not real covariances, they are just measuring the (co)variability of the levels of a fixed effect.

5.5.3 Regressional equivalents

To begin with, consider a typical regressional setup with paired observations (x_i, y_i). There are two possible approaches to investigating the relationship between x and y, the 'correlation' approach and the 'regression' approach. The point is that although the two approaches are conceptually different the test for 'no relationship' becomes the same in the two approaches: the tests for zero correlation respectively zero slope are the same t tests. The nice thing about the observed correlations in the mean squares above, although they are combinations of different effects, is that they have direct interpretations in terms regressions of certain averages of the data.

These related regressions serve as a good tool for the understanding of the various levels of association the mean squares really measure. Consider the observed mean square for factor J, MS_J. The off-diagonal element is the observed covariance between x_{ipk} and y_{ipk} seen over judges, and the corresponding correlation is the correlation that comes from a linear regression of $\bar{y}_{i..}$ on $\bar{x}_{i..}$, i.e. a linear regression on averages for each judge.

The off-diagonal element of $\mathrm{MS}_{J\times S}$ is the empirical covariance corresponding to the regression of estimated $J\times S$ interaction effects, i.e. the regression of $\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...}$ on $\bar{x}_{ij.}-\bar{x}_{i..}-\bar{x}_{.j.}+\bar{x}_{...}$.

In an analogous way the off-diagonal element of $\mathrm{MS}_{J\times R}$ is the empirical covariance corresponding to the regression of estimated $J \times R$ interaction effects, i.e. the regression of $\bar{y}_{i.k}-\bar{y}_{i..}-\bar{y}_{..k}+\bar{y}_{...}$ on $\bar{x}_{i.k}-\bar{x}_{i..}-\bar{x}_{..k}+\bar{x}_{...}$

Finally the off-diagonal of $\mathrm{MS}_E = \hat{\Sigma}_E$ is the error covariance and corresponds to a regression of the y residuals from the fixed effects model (5.1) on the x residuals from the same model. The regression coefficient from this regression is the same that emerges by including x_{ipk} as an additional covariate in the fixed effects model (5.1) for y_{ipk}.

5.6 Estimation

Estimates for the 'clean' variance components, Σ_E, $\Sigma_{J\times S}$, $\Sigma_{J\times R}$ and Σ_J can be found by solving the equations (5.10)-(5.13). This is known as the ANOVA method of estimation, see Searle et al. (1992), and amounts in the present case to solving three sets of four linear equations with four unknowns. The issue of estimating variance components is a huge research area of its own and various methods exist. For orthogonal designs, the ANOVA method is equivalent to the generally accepted method of restricted maximum likelihood (REML), see Searle et al. (1992).

If there are missing values, leading to a non-orthogonal design, things get a little complicated. The partitioning of variability is no longer unique and the expected mean squares become complicated to calculate. A better solution would be a formal REML approach, but this goes beyond the scope of Searle et al. (1992) and certainly beyond what we want to address in this context. In Dembster *et al.* (1981) a version of the EM-algorithm was employed for estimation in covariance components models. The way we proceede is to use the Random statement of Procedure Mixed in SAS in a univariate setting, but with observations where either one or both observations are missing, set to 'missing'. This will make SAS calculate the expected mean squares for each effect listed in the Random statement. If 'variational components' should be calculated for fixed effects also, one merely lists these effects in the Random statement as well, alone for the purpose of having SAS to calculate the equations to be solved.

As mentioned, the Procedure Mixed of SAS can handle the univariate case, but it cannot estimate multivariate components of covariance in a mixed model. For this reason it will be necessary to solve the equations (5.10)-(5.13) by other means in each case. The factor structure diagram provides, in the orthogonal case, an easy way of deducing the equations to be solved, as will be clear in the following.

Solving the equations (5.10)-(5.13) does not guarantee that the resulting variational components are non-negative nor that derived components of correlations are within the required -1 to 1. The way to do this is to set 'negative components' to zero and correlations to either -1 or 1.

5.7 The hypothesis of independence

The main objective is to investigate the relationship between the two variables X and Y. This may lead to the hypothesis of independence, that is, of no association between x and y on any (random) level. This hypothesis is therefore equivalent to

$$\sigma_{xy}^{E} = \sigma_{xy}^{J\times S} = \sigma_{xy}^{J\times R} = \sigma_{xy}^{J} = 0 \tag{5.17}$$

which is seen from the equations (5.10)-(5.13) to be equivalent to the case, where the four off-diagonal elements of $\mathrm{E}(\mathrm{MS}_E)$, $\mathrm{E}(\mathrm{MS}_{J\times S})$, $\mathrm{E}(\mathrm{MS}_{J\times R})$ and $\mathrm{E}(\mathrm{MS}_J)$ are zero. As these mean squares are independent random variables with different distributions defined by different parameters, the hypothesis splits up into four independent hypotheses, one for each random effect (stratum, see the Appendix). In other words the test for independence really constitutes four independent tests, that each is a test for zero correlation, or equivalently, a test for zero slope in a linear regression, i.e. four standard t-tests. One uses the phrase *to test the effect in each stratum.* The t test statistic for independence in, say, J stratum is:

$$t = \sqrt{DF_J - 1}\frac{c_J}{\sqrt{1 - c_J^2}}, \tag{5.18}$$

where c_J is the 'observed' judge correlation, that is, the correlation calculated from MS_J.

However, a single test for no correlation overall can be deduced by classical maximum likelihood theory, see the Appendix, and this test is applied in the present paper as well.

5.8 Materials and sensory methods

5.8.1 Samples

D-Glucose (certified A.C.S., Fisher Scientific, Pittsburgh PA) and medium viscosity Na-Carboxymethylcellulose (CMC; Sigma Chemical Co., St. Louis MO) were used for the sample preparation. Aqueous solutions of three different levels of glucose and three different levels of CMC as well as all combinations of the glucose and CMC concentration levels were examined. A distilled water sample was included. In Table 5.1 the composition of the sixteen samples presented in the experiment is given. For samples containing CMC, glucose stock solutions of 40 g/l, 90 g/l and 140 g/l were prepared. Since glucose contributes to the viscosity of a solution, the amount of gum added to the samples containing no sugar, 4%, and 9% of glucose was different for each sample to achieve the same viscosity for all samples within a food thickener level, see Table 5.1. Gum was added slowly while stirring the sugar solutions on a magnetic stirrer until the gum was completely dissolved (10 to 20 minutes). For the unsweetened samples the gum was added directly to de-ionized water. Fresh samples were prepared every week and kept at room temperature until testing. The physical viscosity of all samples was measured using a Carrimed viscometer (CS 500).

5.8.2 Sensory evaluation

Prior to the eight sessions of data collection the panel was trained using scalar rating of sweetness and viscosity intensity and then in the use of the time intensity method. In each of the scalar training sessions the judges rated eight samples on a 10 cm line scale for either sweetness or viscosity. In the time intensity training the judges rated

Table 5.1: Composition of experimental samples

Sample	Glucose	CMC
	(g/L)	(g/L)
1	0	0.00
2	40	0.00
3	90	0.00
4	140	0.00
5	0	4.41
6	0	6.93
7	0	9.91
8	40	4.00
9	90	3.24
10	140	2.50
11	40	6.29
12	90	5.65
13	140	5.00
14	40	9.20
15	90	8.28
16	140	7.50

the same samples for sweetness and viscosity intensity using the procedure described in the following.

A computerized time intensity system was used for recording the temporal sweetness and viscosity characteristic of the sixteen samples. The sensory booths used for testing were equipped with Apple Macintosh Plus computers which were connected via a TOPS network to an Apple Macintosh II. A time intensity program, see Borton (1990), allowed the judge to rate the intensity of sweetness and viscosity by manipulating a mouse connected to the computer in the booth. Moving the mouse to the right indicated an increase in intensity whereas moving the mouse to the left indicated a decrease in intensity. The judges could see the intensity rating on a line scale displayed on the computer screen. At the time a sample was ingested, rating of sweetness or viscosity intensity was initiated by clicking on the 'Go' icon shown on the screen below the line scale. After ten seconds a spit icon and an acoustic signal prompted the judges to expectorate the sample. Intensity rating was continued until the sensation was extinguished. In all tasting sessions (training sessions and formal data collection sessions) two reference samples for each of the sensory attributes were given to anchor the line scale. For the sweetness as well as for the viscosity ratings, the intensity of the less intense sample corresponded to 10% of the scale whereas the more intensive sample corresponded to 90% of the scale. The judges were allowed to taste the reference samples as many times as needed.

The sixteen samples were presented in a balanced design in duplicate and the order of presentation was completely randomized over the sixteen samples and the twenty judges. After evaluation of the low and high standards for the attribute being rated in that session, eight samples were evaluated; thus, two sessions were necessary to complete one replication. The same sensory attribute was evaluated four times in a row in order to complete the two replications. All the samples were evaluated in sensory test booths under red light. 20 ml aliquots were presented at room temperature in 40 ml plastic cups coded with three digit random numbers. Since the perceived viscosity of non Newtonian solutions is affected by shear rate, judges were told to move the tongue as consistently as possible during evaluation of the samples. Judges rinsed with de-ionized water between samples.

5.8.3 Saliva collection

Saliva was collected from the right parotid gland of the 20 judges who participated in the time intensity part of the experiment. A modified Carlson-Crittenden vacuum cap which was placed over the orifice of Stenson's duct in the right cheek of the subjects, see Shannon *et al.* (1962). The outer ring chamber of the cap was connected with Tygon tubing (3.2 mm outer diameter, 1.6 mm inner diameter) to the lab vacuum (approximately 80 kPa) to hold the cap in place. Saliva flowed by gravitational force to the inner chamber and then via Tygon tubing to the sialometer, see Pangborn *et al.* (1971). The increasing weight of accumulating saliva caused a vertical displacement of a spring which was converted in a voltage reading between 0 and 2 mV.

The sialometer was connected to a variable power supply and was interfaced to a Macintosh II by an analog to digital signal Lab NB board (National Instruments Co., Austin TX). An amplifier (gain of 1000) was placed between the sialometer and the board which was configured for a bipolar (=B15V) input, see Bonnans (1991). A LabView II computer program recorded saliva readings twice a second. Furthermore, the program extracted the saliva readings every 15 seconds starting at 0 seconds in a separate file, see Bonnans (1991). The sialometer system was calibrated with water from 0 ml up to 1.5 ml in steps of 0.05 ml. The mean of ten readings for every step was used to calculate the calibration curve. For the calculation of the saliva weight, the computer readings of all samples were corrected such that the value of the samples corresponded to the value of the calibration curve at time 0.

The saliva flow was measured over a period of two minutes in response to thirteen of the samples used in the time intensity experiment. Because of the extremely time-consuming nature of the saliva collection, samples 11-13 (Table 5.1) were excluded. Previous to any saliva measurements the salivary flow was stimulated by 20 ml aqueous citric acid solution (2g/L) to fill the Tygon tubing from the cap to the sialometer. After two minutes, parotid saliva flow (without ingestion of a sample) was collected for two minutes to get an unstimulated or 'baseline' measurement for each judge. The saliva flow elicited by each of the thirteen samples was collected in duplicate in two separate sessions. To have the same experimental conditions as in the sensory part of the experiment, the judges rated the intensity of sweetness in the first, and viscosity in the second session of saliva collection using the time intensity procedure described above. Between the samples the judges rinsed with de-ionized water and waited for one minute before starting with the next sample.

5.8.4 Two data sets

A second experiment similar to the one described above, although with only three samples, was carried out, where the glucose concentration of the saliva were measured. In this paper we only consider a small part of the total data material. To summarize, we consider two data sets, referred to as the 'sweetness data' and the 'glucose data' in the following. The sweetness data are bivariate observations for 20 judges, 13 samples and 2 replicates, the two components being the time to maximum sweetness intensity (Tmax) and the salivary flow rate (Sal). For this data there are 10 missing values of which 6 were single replicates and the remaining 4 were both replicates for two combinations of judge and sample. The glucose data are bivariate observations for 19 judges, 3 samples and 4 time points. One of the original 20 judges was left out owing to too many missing observations. For simplicity the two original replicates were averaged, and the final data set has no missing values. The two components of the glucose data set are the glucose concentration (Gluc) and the salivary flow rate (Sal).

5.9 Results and Discussion

For the sweetness data the effect of the $S \times R$ interaction was clearly non-significant for both Tmax and Sal, and was left out of the subsequent modelling. The factor structure diagram for the model used is shown in Figure 5.2 and the mean squares and 'variational components' are listed in Table 5.2. As mentioned, 10 observations were missing leading to a superscript to $[E]$ of 510 instead of 520 in the factor structure diagram. Moreover, since for two combinations of judge and sample both replicates were missing, the superscript of $[J \times S]$ is 58 instead of 60, leading to correspondingly adjusted degrees of freedom. As the partitioning of variance now is non-unique, a choice has to made about which mean squares to use. The ones in Table 5.2 are the 'sequentially' calculated mean squares, in SAS terminology the `Type I` mean squares.

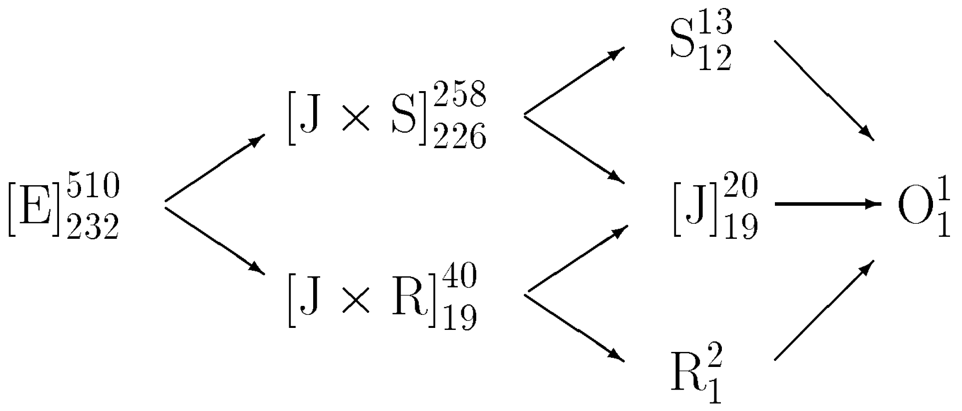
$[\mathrm{E}]_{232}^{510}$
$[\mathrm{J}\times\mathrm{S}]_{226}^{258}$
$[\mathrm{J}\times\mathrm{R}]_{19}^{40}$
S_{12}^{13}
$[\mathrm{J}]_{19}^{20}$
R_{1}^{2}
O_{1}^{1}

Figure 5.2: Factor structure diagram for the final model of the sweetness data.

Also the expected mean squares from which the 'variational components' are to be calculated cannot be deduced explicitly from the factor structure diagram. Instead the Random statement is used with Proc GLM of SAS is used to get the expected mean squares for random as well as fixed effects:

$$\begin{aligned}
\mathrm{E}\,(\mathrm{MS}_E) &= \Sigma_E \\
\mathrm{E}\,(\mathrm{MS}_{J\times S}) &= \Sigma_E + 1.97\Sigma_{J\times S} \\
\mathrm{E}\,(\mathrm{MS}_{J\times R}) &= \Sigma_E + 0.014\Sigma_{J\times S} + 12.74\Sigma_{J\times R} \\
\mathrm{E}\,(\mathrm{MS}_J) &= \Sigma_E + 1.99\Sigma_{J\times S} + 12.75\Sigma_{J\times R} + 25.48\Sigma_J \\
\mathrm{E}\,(\mathrm{MS}_S) &= \Sigma_E + 1.99\Sigma_{J\times S} + 0.016\Sigma_{J\times R} + 0.021\Sigma_J + 39.23Q_S \\
\mathrm{E}\,(\mathrm{MS}_R) &= \Sigma_E + 0.012\Sigma_{J\times S} + 12.8\Sigma_{J\times R} + 0.0074\Sigma_J + 0.015Q_S + 255Q_R
\end{aligned}$$

If these equations are compared to the equations derived from the factor structure diagram, as described in the Appendix, one would see that the influence of the 10 missing values are quite inconsiderable.

Note again that we calculate '(co)variational components' for random effects as well as for fixed effects in a similar way, although interpretations in a strict sense are different. From a data analysis point of view, however, it is convenient, and sensible, just to observe where the (co)variations are, whether random or fixed, and then subsequently make the appropriate interpretation.

The two leftmost columns in Table 5.2 are basically the usual (Type I) MANOVA table. Without presenting the exact test etc. we note that the listed effects are all of significance.

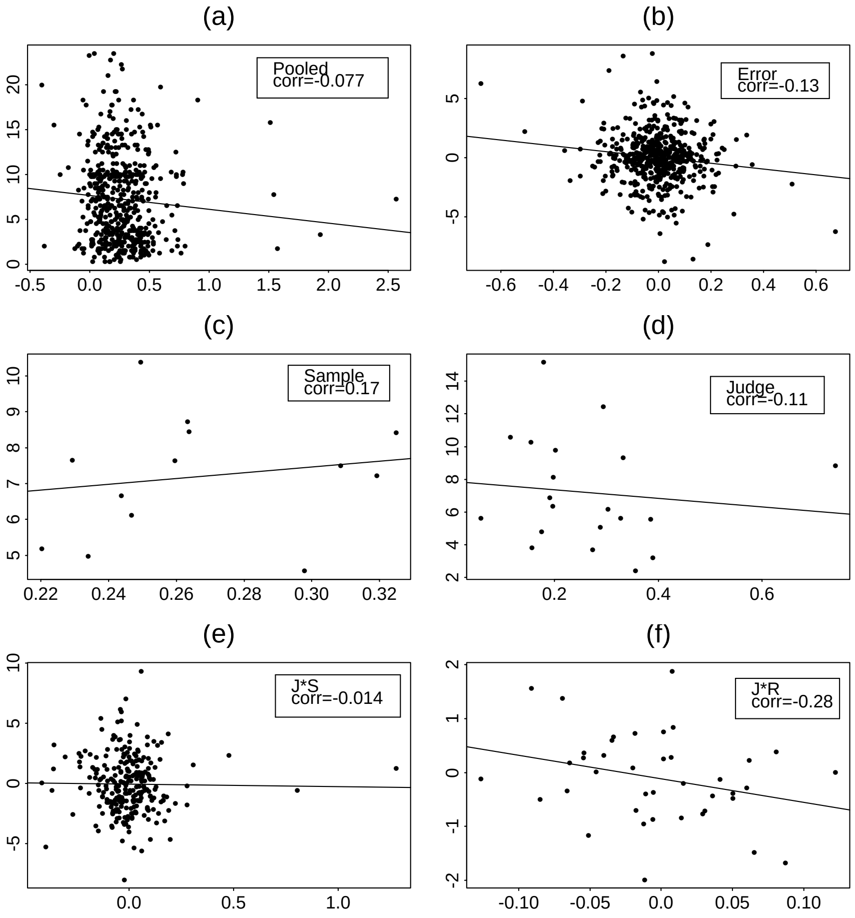

Figure 5.3: Time to maximum sweetness intensity plotted versus salivary flow rate over the various effects. (a) Raw data, (b) residuals from the model (5.1), (c) sample averages, (d) judge averages, (e) $J \times S$ interaction effects and (f) $J \times R$ interaction effects. Correlations in legends are the correlation between the plotted points

Table 5.2: Observed mean squares, estimated 'variational components', and 'cleaned' correlations for Tmax and Sal in the sweetness data set.

Factor	Mean squares		Var. components		Correlations
	x	y	x	y	
R	0.66	-0.11	0.0$^{(a)}$	0.00077	
	-0.11	0.019	0.00077	0.0$^{(a)}$	
S	103.5	0.44	2.31	0.011	
	0.44	0.055	0.011	0.00012	0.69
J	284.4	-1.45	10.4	-0.047	
	-1.45	0.55	-0.047	0.018	-0.11
$J \times R$	17.0	-0.30	0.51	-0.018	
	-0.30	0.073	-0.018	0.0034	-0.44
$J \times S$	12.6	-0.0095	1.15	0.032	
	-0.0095	0.050	0.032	0.010	0.30
E	10.4	-0.073	10.4	-0.073	
	-0.073	0.030	-0.073	0.030	-0.13

$^{(a)}$ Negatively estimated, set to 0.0

To get a better impression of the observed correlation structure expressed in the mean squares we have in Figure 5.3 made scatterplots of the data in accordance with the regressional interpretations described above. Figure 5.3(a) shows that ignoring the design totally indicates an overall negative correlation between Tmax and Sal; the t test for independence on 508 degrees of freedom gives a p value of 0.13. Figures 5.3(b)-3(f) show that this overall tendency covers over a more pronounced negative error- and $J \times R$ correlation together with a positive sample correlation. Table 5.3 shows that of the observed 'real' (random effects) correlations, only the error correlation is significant. Note that the t tests of Table 5.3 relates to the correlations in the legends of Figure 5.3, apart from minor discrepancies due to the missing values.

Table 5.3: T test statistics and P values based on (5.18) for independence between Tmax and Sal in each of the four strata.

Effect	$DF - 1$	t	P
Error	231	-2.03	0.044
$J \times S$	225	-0.18	0.86
$J \times R$	18	-1.20	0.24
J	18	-0.49	0.63

Remember that these observed mean squares do not express 'clean' effects. These should be sought in the 'variational components'. The derived 'clean' correlations, as indicated by equation (5.9), are also listed in Table 5.2. We see that there is a strong positive correlation between Tmax and Sal over samples, and a positive $J \times S$ correlation. The latter means that there is a tendency, not that clear though, that judges with a higher salivary flow rate for sample 1 as compared to sample 2 use longer time to reach maximum sweetness intensity for sample 1 as compared to sample 2. More pronounced is the negative $J \times R$-correlation. This has the interpretation, that judges with a higher salivary flow rate at the second session than at the first session tend to use less time to reach maximum sweetness intensity at the second session.

The test for independence overall, across the random effects, yielded a $-2 \log Q$ test statistic, see the Appendix, of 5.89. This gives a p value of 0.21 based on the χ^2_4 distribution and 0.23 based on the more accurate type of approximation in Brockhoff (1994). Assuming for a moment that the sample effect also is random and doing the same test yields a test statistic of 6.30 on 5 degrees of freedom leading to $p = 0.28$ respectivly $p = 0.31$. This is less significant; not surprising as the overall correlation was negative, whereas the sample correlation was positive.

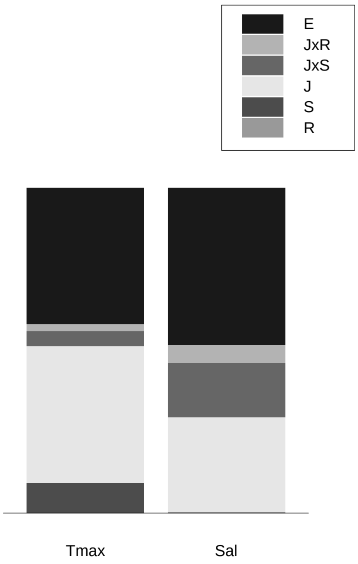

Figure 5.4: Relative sources of variation for sweetness data.

Finally the calculation of the 'clean' variational components makes it possible to investigate the relative importance of the various effects. This is done univariately and has nothing to do with analysis of covariance/correlation, other than the importance of a correlation could be seen in the light of the relative importance of an effect in

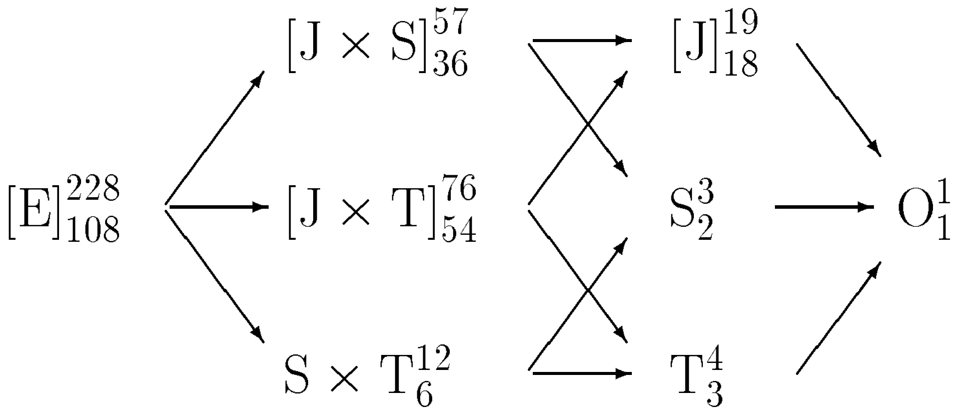
$[\mathrm{E}]_{108}^{228}$
$[\mathrm{J} \times \mathrm{S}]_{36}^{57}$
$[\mathrm{J} \times \mathrm{T}]_{54}^{76}$
$\mathrm{S} \times \mathrm{T}_{6}^{12}$
$[\mathrm{J}]_{18}^{19}$
S_{2}^{3}
T_{3}^{4}
O_{1}^{1}

Figure 5.5: Factor structure diagram for the final model of the glucose data.

general. For each variable all variational components are added to give the total variation in the variable, and then the relative size of each component is investigated. These are illustrated in Figure 5.4, and we see that the error and judge variations are the largest for both variables, although the $J \times S$-interaction variation also is considerable for the salivary flow rate.

For the glucose data all effects were significant as illustrated in the factor structure diagram in Figure 5.5. Figures 5.6 and 5.7 and Tables 5.4 and 5.5 show the results for the glucose data in a similar way as above for the sweetness data.

Figure 5.6 illustrates how a pooled (ignoring design) moderately positive correlation covers over an almost 100 % 'designed' association between Gluc and Sal together with some less strong components. Figures 5.6(c), (e) and (g) show that from sample to sample, from time to time and even in the way each combination of time and sample deviates from the additive level of the combination, there is a strong positive association between the glucose concentration and salivary flow rate. Looking at the four plots to the right in Figure 5.6 and Table 5.5, we see that there is no significant judge- or $J \times S$-correlation. There is indeed a significant error correlation, which means that whenever within a single person, time point and sample for some reason the salivary flow rate becomes larger, the same is the case for the glucose concentration. There is a quite strong positive $J \times T$ correlation, which means that whenever a judge deviates from the additive level on a sample, he/she tends to deviate in the same direction for both Gluc and Sal. The test statistic for overall independence over the four random effects is 21.3 that based on either the χ_4^2 distribution or the approximation in Brockhoff (1994) is extremely significant.

Investigating the 'cleaned' components, see the rightmost column of Table 5.4,

Table 5.4: Observed mean squares, estimated 'variational components', and 'cleaned' correlations for Gluc and Sal in the glucose data set.

Factor	Mean squares		Var. components		Correlations
	x	y	x	y	
T	64.7	8.47	0.94	0.14	
	8.47	1.20	0.14	0.020	1.00
S	210.6	9.22	2.61	0.11	
	9.22	0.41	0.11	0.0048	1.0[(a)]
J	9.04	-0.13	0.57	-0.019	
	-0.13	0.31	-0.019	0.022	-0.17
$J \times T$	0.48	0.062	0.088	0.017	
	0.062	0.032	0.017	0.0069	0.71
$J \times S$	1.92	0.041	0.43	0.0078	
	0.041	0.022	0.0078	0.0028	0.23
$S \times T$	10.6	0.58	0.55	0.030	
	0.58	0.036	0.030	0.0013	1.0[(a)]
E	0.22	0.0094	0.22	0.0094	
	0.0094	0.011	0.0094	0.011	0.19

[(a)] Estimated larger than 1, set to 1.0

Table 5.5: T test statistics and P values based on (5.18) for independence between Gluc and Sal in each of the four strata.

Effect	$DF - 1$	t	P
Error	107	2.05	0.043
$J \times S$	35	1.20	0.24
$J \times T$	53	4.21	0.0001
J	17	-0.33	0.75

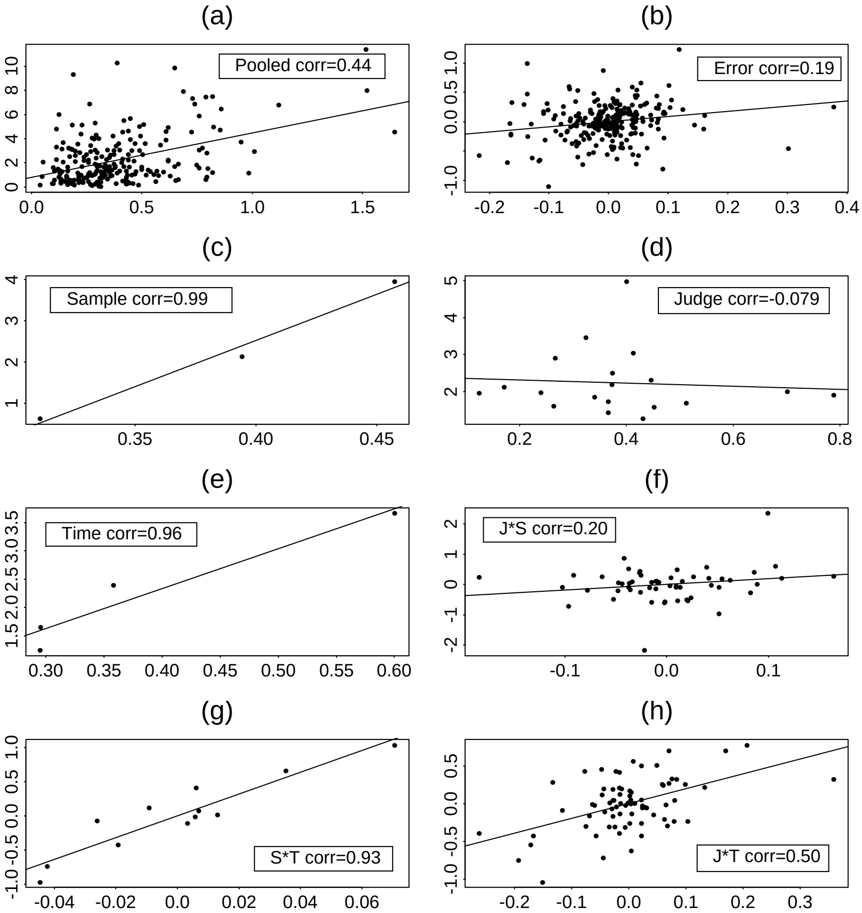

Figure 5.6: Glucose concentration plotted versus salivary flow rate over the various effects. (a) Raw data, (b) residuals from the model (5.1), (c) sample averages, (d) judge averages, (e) time averages, (f) $J \times S$ interaction effects, (g) $S \times T$ interaction effects and (h) $J \times T$ interaction effects.

we get interpretations that support the MS-interpretations above. In Figure 5.7 the relative importance of the 'cleaned' variational components shows a quite different behaviour as compared to Figure 5.4. The error-component is much less here, partly because the two replicates were averaged out. The judge component is much smaller for Gluc, instead the sample component is important. And the time component is quite important for both variables.

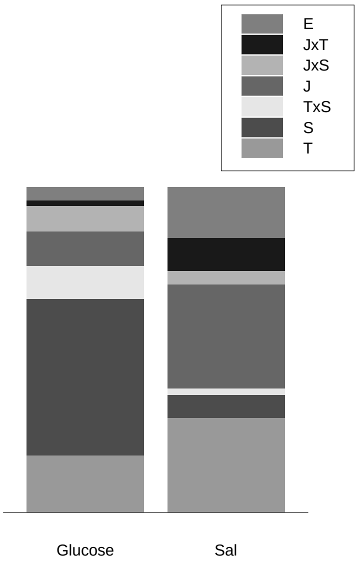

Figure 5.7: Relative sources of variation for glucose data.

5.10 Conclusion

We have presented a unified approach to the analysis of the relationship between two (or more) variables, when a designed experiment is performed. It should be clear, that the principles outlined can be generalized to any design structure of an experiment. We have illustrated the importance of acknowledging the fact that correlation really exists on different levels; the use of a single correlation coefficient to summarize a relationship does not make sense. We suggested ways of graphically illustrating the different components.

The basic design structure of an experiment is the key issue in understanding which levels of correlation/covariance are present in a data set. We believe that the factor structure diagrams are of great help to the analyst in handling complex structures, also in the non-orthogonal case, even though the rules of the appendix do not apply explicitly.

The test for overall independence is a new development, and the approximation results of Brockhoff (1994) make the use of the test very reliable.

We stated that SAS cannot handle the multivariate covariance component models. This may not be quite true. If only the covariance structure is considered, that is, all effects are assumed random, or the fixed effects are 'removed' prior to further analysis, the model falls under the category treated by Jöreskog (1978). In SAS the Proc Calis can handle such models. This may in some cases be a reasonable way to proceed, especially if more complex covariance structures are expected, but the flexibility in our pragmatic approach to the handling of fixed and random effects would be lost. We also believe, that 'sticking to a close connection' with the MANOVA models enhances the understanding of these quite complicated models.

5.11 References

Anderson, T.W. (1958). *An Introduction to Multivariate Statistics.* Wiley, New York.

Bonnans, S. (1991). *Effect of sample composition and salivary flow on temporal perception of sweetness.* Masters Thesis, University of California, Davis.

Borton, C.B. (1990). Weinschmecker. A computerized data collection system for sensory testing. Unpublished.

Brockhoff, P.M. (1994). *Statistical Analysis of Sensory Data.* Ph.D.-thesis, Dept. of Mathematics and Physics and Center of Food Research, Royal Veterinary and Agricultural University, Copenhagen, Denmark.

Brockhoff, P.M., Skovgaard, I.M., Poll, L. and Hansen K. (1993). A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements. *Food Quality and Preference* **4**, 215-222.

Dembster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in covariance components models. *J. Am. Stat. Ass.* **76**(374), 341–353.

Jensen, J.L. (1988). Uniform saddlepoint approximations. *Adv. Appl. Prob.* **3**, 622-634.

Jensen, J.L. (1991). A large deviation-type approximation for the "Box class" of likelihood ratio criteria. *J. Am. Stat. Ass.* **86**(414), 437-440.

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* **43**(4), 443-477.

Lundahl, D.S. and McDaniel, M.R. (1988). The panelist effect — fixed or random? *Journal of Sensory Studies* **3**, 113-121.

Kjølstad, L., Isaksson, T. and Rosenfeld, H.J. (1990). Prediction of sensory quality by near infrared reflectance analysis of frozen and freeze dried green peas (Pisum sativum). *J. Sci. Food Agric.* **51**, 247-260.

Martens, M. and Martens H. (1986). Near-infrared reflectance determination of sensory quality of peas. *Applied Spectroscopy* **40**(3), 303-310.

Næs, T. and Kowalski, B. (1989). Predicting sensory profiles from external instrumental measurements. *Food Quality and Preference*, 135-147.

Næs, T. and Martens, H. (1989). *Multivariate Calibration.* Wiley, Chichester.

Pangborn, J., Eriksson, F. and Remi, K. (1971). Simplified sialometer for continuous weight monitoring of salivary secretion. *J. Dent. Res.* **50**, 1689.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components.* Wiley, New York.

Shannon, I.L., Prigmore, J.R. and Chauncey, H.H. (1962). Modified Carlson-Crittenden device for the collection of parotid fluid. *J. Dent. Res.* **41**, 778–783.

Tjur, T. (1984). Analysis of variance models in orthogonal designs. *International Statistical Review* **52**, 33-81.

Tjur, T. (1991). Analysis of variance and design of experiments. *Scandinavian Journal of Statistics* **18**, 273-369.

5.12 Appendix

5.12.1 Factor structure diagrams

This presentation is entirely based on Tjur(1984, 1991), and the rules described apply only to orthogonal designs. The factor structure diagram for a model is a listing of all factors/effects in the model with arrows from *finer* factors to *coarser* factors, see Figure 5.1. A factor is finer than another if it corresponds to a subpartioning of the data. For example, $J \times S$ is finer than both S and J, but not finer than R. These concepts are mathematically defined in the two given references and we shall not go into this here. The E and O are 'pseudo-factors' that are always present in an experiment. The 'factor' E stands for the finest of all factors, the factor with a level for each experimental unit, corresponding to the error effect, and the 'factor' O is the coarsest of all factors, the factor with only one level, corresponding to the total variability.

The square-bracketed factors are the random effects, and the remaining factors the fixed effects. The diagram can help the analyst to get hold of all factors/effects in play in a given experiment, and it can be explicitly helpful for calculating degrees of freedom, expected mean squares, finding the appropriate F statistics etc. In the following we will need the two conventions about a factor G in a diagram:

Factors to the right of G

= all factors in the diagram that can be 'hit' by going to the right following arrows in all possible directions starting from G.

Factors to the left of G

= all factors in the diagram with G to the right of them.

Degrees of freedom

Having constructed the raw diagram, we add as a superscript to each factor the number of levels for the factor, e.g. I for the factor J. Moreover we add a subscript 1 for the factor O. The degrees of freedom for any factor G in the diagram can now be found recursively, starting from the right and writing degrees of freedom as subscripts, by the following rule:

Subtract from the superscript of G the sum of all subscripts of factors to the right of G.

F statistics

The four random effects splits the factors into four disjoint *strata*. One might say that the four random effects splits the fixed effects among them. These strata can be mathematically defined, Tjur (1984, 1991) and can be read from the diagram:

The fixed effect G belongs to the first random (bracketed) effect that is met, by going leftwards in the diagram from G.

Table 5.6: The partitioning of factors into disjoint strata.

Stratum	Factors
Error	E, $S \times R$
Judge×Sample	$J \times S$, S
Judge×Replication	$J \times R$, R
Judge	J, O

The four strata are listed in Table 5.6.

The F test for each fixed effect is now to be performed in the stratum in which it belongs, e.g. the S effect must be tested versus the $J \times S$ interaction mean square term.

Expected mean squares

For both fixed and random effects the expected mean squares can be found from the diagram based on the following rule, that applies to orthogonal and non-nested designs:

> The expected mean square for factor G is a linear combination of 'variational components' of G itself and all factors to the left of G with coefficients that are the total number of observations, IPK, divided with the superscript of the factor in question.

The phrase 'variational component' refers to real variance components for the random effects and variation between the fixed levels for the fixed effects.

5.12.2 Test for overall independence

Let $\mathrm{MS}_1, \ldots, \mathrm{MS}_K$ be the 2×2 mean squares corresponding to the collection of effects, over which we test overall independence. From Brockhoff (1994) we get that the test statistic $-2\log Q$ is given by:

$$-2\log Q = -\sum_{k=1}^{K} d_k \log\left\{1 - \frac{\mathrm{MS}_{k12}^{2}}{\mathrm{MS}_{k11}\mathrm{MS}_{k22}}\right\} \tag{5.19}$$

where d_k are the corresponding degrees of freedom and the notation given by

$$\mathrm{MS}_k = \begin{pmatrix} \mathrm{MS}_{k11} & \mathrm{MS}_{k12} \\ \mathrm{MS}_{k12} & \mathrm{MS}_{k22} \end{pmatrix}$$

is used. This test-statistic has asymptotically a χ^2 distribution with K degrees of freedom. For small samples this approximation will be poor, and in Brockhoff (1994) a vastly superior approximation was developed based on so-called uniform saddlepoint approximation methods, due to Jensen (1988, 1991). We refer to Brockhoff (1994) for an exact computational description of this approximation, as this is rather technical, including a numerical solution of a nonlinear equation and several iterative-based function evaluations.

5.13 Add: Test for overall independence

5.13.1 Introduction

In this section we present the likelihood ratio test for overall independence between sets of variables in a multivariate covariance component model. This can be seen as a supplement to a merely stratified approach. We show that this test is in the Box class and we outline how to apply gamma saddlepoint methods for approximating the distribution of $-2\log Q$. Both of these issues become straightforward consequences of the fact, that the likelihood ratio, Q, is a product of independent terms. The performance of the approximation is illustrated by a small example. Finally the assumptions necessary for the design are discussed and some relaxations are given.

5.13.2 Likelihood ratio test

Assume we have n independent p-dimensional observations. Let $\mathcal{B}$ be the set of random effects, $\{d_B\}_{B\in\mathcal{B}}$ the corresponding degrees of freedom. and assume that $\mathcal{B}$ satisfies the four conditions of Tjur (1991) set up for a random effects design, see Section 5.13.5 below.

We then have that the collection of random effect sums of squared deviations, $\{\mathrm{SSD}_B \mid B \in \mathcal{B}\}$ is G-sufficient for the collection of canonical covariance components $\{\Lambda_B \mid B \in \mathcal{B}\}$, see Møller (1984), p. 5.1. The marginal distribution is the product of the independent p-dimensional Wishart distributions,

$$\mathrm{SSD}_B \sim W_p(\Lambda_B, d_B)\ ,\ B \in \mathcal{B}. \tag{5.20}$$

The maximum likelihood estimator, $\hat{\Lambda}_B$, of Λ_B becomes

$$\hat{\Lambda}_B = \frac{1}{d_B}\mathrm{SSD}_B,$$

and the total likelihood is, see Anderson (1958), p. 154:

$$L(\{\Lambda \mid B \in \mathcal{B}\}) = \prod_{B\in\mathcal{B}} \frac{\mid \mathrm{SSD}_B \mid^{\frac{1}{2}(d_B-p-1)} \exp\left(-\frac{1}{2}\mathrm{tr}(\Lambda_B^{-1}\mathrm{SSD}_B)\right)}{2^{pd_B/2}\pi^{p(p-1)/2}|\Lambda_B|^{d_B/2}\prod_{i=1}^{p}\Gamma(\frac{1}{2}(d_B+1-i))} \tag{5.21}$$

We want to test the hypothesis of overall independence between sets of variates. For simplicity we consider only two sets with dimensions q_1 and q_2, with $q_1 + q_2 = p$. It is clear, however, that the following can be formulated similarly for arbitrary sets of variables. The null hypothesis is that the Λ_B's are block diagonal,

$$\mathrm{H}_0: \ \ \Lambda_B = \begin{bmatrix} \Lambda_B^{(1)} & 0 \\ 0 & \Lambda_B^{(2)} \end{bmatrix}\ ,\ B \in \mathcal{B}$$

where $\Lambda_B^{(i)}$ is a q_i-dimensional square matrix corresponding to variate set i, $i = 1, 2$. Under H_0 the MLE, $\hat{\Lambda}_B^0$, of Λ_B is

$$\hat{\Lambda}_B^0 = \frac{1}{d_B}\begin{bmatrix} \text{SSD}_B^{(1)} & 0 \\ 0 & \text{SSD}_B^{(2)} \end{bmatrix},$$

where $\text{SSD}_B^{(1)}$ and $\text{SSD}_B^{(2)}$ are the block diagonal elements of $\hat{\Lambda}_B$. The likelihood ratio becomes

$$Q = \frac{L(\hat{\Lambda}_B^0)}{L(\hat{\Lambda}_B)} = \prod_{B \in \mathcal{B}} \left(\frac{\mid \text{SSD}_B \mid}{\mid \text{SSD}_B^{(1)} \mid\mid \text{SSD}_B^{(2)} \mid} \right)^{d_B/2} \tag{5.22}$$

or

$$Q = \prod_{B \in \mathcal{B}} Q_B$$

where each component

$$Q_B^{2/d_B} = \frac{\mid \text{SSD}_B \mid}{\mid \text{SSD}_B^{(1)} \mid\mid \text{SSD}_B^{(2)} \mid} \tag{5.23}$$

is the well-known likelihood ratio for independence in a single stratum, see Anderson (1958), p. 232-233 and Møller (1984), p. 6.5. It should be noted, however, that in Anderson (1958) the G-sufficient reduction is not performed, and the likelihood ratio is deduced based on the profile likelihood function leading to the total number of observations instead of d_B in the expression above.

In the case of testing single stratum independence between two one-dimensional variables the likelihood ratio test (5.23) becomes equivalent to a t-test on $d_B - 1$ degrees of freedom. For general dimensions the distribution of the single stratum test statistic $\log Q_B^{2/d_B}$ has been approximated in various ways. In Anderson (1958) and Møller (1984) approximations based on an expansion due to Box (1949) are given. In Jensen (1991) three alternative approximations are compared to the Box approximations. One of the approximations in Jensen (1991) is a Gamma based saddlepoint approximation. A Gamma based saddlepoint approximation is what we will present for the overall independence test. For this we need the Laplace transform, ϕ, of $-2\log Q$. From (5.22) we get

$$-2\log Q = -\sum_{B \in \mathcal{B}} d_B \log\left\{ \frac{\mid \text{SSD}_B \mid}{\mid \text{SSD}_B^{(1)} \mid\mid \text{SSD}_B^{(2)} \mid} \right\} \tag{5.24}$$

In Anderson (1958) the moments of

$$\lambda_B = \frac{\mid \text{SSD}_B \mid}{\mid \text{SSD}_B^{(1)} \mid\mid \text{SSD}_B^{(2)} \mid}$$

are derived in the case with unconstrained error structure, i.e. B represents the error component and the sums of squared deviations entering λ_B are Wishart distributed with $n-1$ degrees of freedom. Only the degrees of freedom differ in the present case (as the derived moments do not depend on the parameters), and we can 'translate' formula (3) of Anderson (1958), p. 235 directly to obtain the Laplace transform of $-2\log Q$,

$$\phi(t) = \mathrm{E}\left\{ \mathrm{e}^{t(-2\log Q)} \right\} = \tag{5.25}$$
$$\left[\prod_{B\in\mathcal{B}} \prod_{i=1}^{p} \left(\frac{\Gamma\left((1-2t)\frac{d_B}{2} + \frac{1-i}{2} \right)}{\Gamma\left(\frac{d_B}{2} + \frac{1-i}{2} \right)} \right) \right] \left[\prod_{B\in\mathcal{B}} \prod_{i=1}^{2} \prod_{j=1}^{q_i} \left(\frac{\Gamma\left(\frac{d_B}{2} + \frac{1-j}{2} \right)}{\Gamma\left((1-2t)\frac{d_B}{2} + \frac{1-j}{2} \right)} \right) \right],$$

that is finite for $d_B > p-1$. This expression shows that the likelihood ratio test is of the Box class, as it can be recognized as the class defining expression in Jensen (1991). We can thus apply the gamma based saddlepoint approximation method of Jensen (1991). Instead we will use the more general formulated approximation of Jensen (1988), see Chapter 7. In Jensen (1994) the conjecture is made that this will provide limiting relative exactness in the tails, though this is not proved.

5.13.3 Saddlepoint approximation

Let ϕ denote the Laplace transform of $W = -2\log Q$ and κ the cumulant generating function, $\kappa(t) = \log\phi(t)$. As W is a positive random variable, the approximation from Jensen (1988) becomes, see Section 7.5,

$$P(W \geq w) \approx \frac{\mathrm{e}^{\kappa(t)-tw}}{t\sigma_t} \frac{t\sigma_t \mathrm{e}^{t\sigma_t\sqrt{\lambda_t}} \lambda_t^{\lambda_t/2}}{(t\sigma_t + \sqrt{\lambda_t})^{\lambda_t}} \int_{\sqrt{\lambda_t}(t\sigma_t+\sqrt{\lambda_t})}^{\infty} \frac{v^{\lambda_t-1}}{\Gamma(\lambda_t)} \mathrm{e}^{-v} dv, \tag{5.26}$$

where

$$\lambda_t = 4\frac{\kappa''(t)^3}{\kappa'''(t)^2},$$

and $t = t_w$ is the saddlepoint, that is, the solution to $\kappa'(t) = w$. For calculation of (5.26) we need the Gamma, Incomplete Gamma, Digamma, Trigamma and Tetragamma functions, see Abramowitz and Stegun (1964).

In the special case of $q_1 = q_2 = 1$ we have that

$$\phi(t) = \left(\prod_{B\in\mathcal{B}} \frac{\Gamma\left(\frac{d_B-1}{2} - td_B \right)}{\Gamma\left(\frac{d_B-1}{2} \right)} \right) \left(\prod_{B\in\mathcal{B}} \frac{\Gamma\left(\frac{d_B}{2} \right)}{\Gamma\left(\frac{d_B}{2} - td_B \right)} \right)$$

And if we let $D(x) = \log\Gamma(x)$, $x_1 = \frac{d_B-1}{2} - td_B$ and $x_2 = \frac{d_B}{2} - td_B$, we can write the cumulants as

$$\begin{aligned}
\kappa(t) &= \sum_{B\in\mathcal{B}} \left\{ D(\frac{d_B}{2}) - D(\frac{d_B-1}{2}) + D(x_1) - D(x_2) \right\} \\
\kappa^{'}(t) &= \sum_{B\in\mathcal{B}} d_B \left\{ \psi(x_2) - \psi(x_1) \right\} \\
\kappa^{''}(t) &= \sum_{B\in\mathcal{B}} d_B^2 \left\{ \psi^{'}(x_1) - \psi^{'}(x_2) \right\} \\
\kappa^{'''}(t) &= \sum_{B\in\mathcal{B}} d_B^3 \left\{ \psi^{''}(x_2) - \psi^{''}(x_1) \right\}
\end{aligned}$$

where $\psi(x) = D^{'}(x)$ is the Digamma, $\psi'(x)$ the Trigamma and $\psi^{''}(x)$ the Tetragamma function.

For the calculations in the paper of this chapter and the example below the Splus software were used offering the Splus versions of the Gamma and Incomplete Gamma functions. The Di- Tri- and Tetragamma functions were implemented in line with Jensen(1995), that again is based on Bernardo (1976), Schneider (1978) and Abramowitz and Stegun (1964):

$$\psi(x) = \begin{cases} -\gamma - \frac{1}{x}, & x < 10^{-5} \\ \tilde{\psi}_1(x), & x > 8.5 \\ -\sum_{l=0}^{k} \frac{1}{x+l} + \tilde{\psi}_1(x+k+1), & 10^{-5} \le x \le 8.5 \end{cases}$$

$$\psi^{'}(x) = \begin{cases} \frac{1}{x^2}, & x < 10^{-5} \\ \tilde{\psi}_2(x), & x > 8.5 \\ \sum_{l=0}^{k} \frac{1}{(x+l)^2} + \tilde{\psi}_2(x+k+1), & 10^{-5} \le x \le 8.5 \end{cases}$$

$$\psi^{''}(x) = \begin{cases} -\frac{2}{x^3}, & x < 10^{-4} \\ \tilde{\psi}_3(x), & x > 8.5 \\ -2\sum_{l=0}^{k} \frac{1}{(x+l)^3} + \tilde{\psi}_3(x+k+1), & 10^{-4} \le x \le 8.5 \end{cases}$$

where $\gamma = 0.5772156649$, $k = \max_m\{x + m \le 8.5\}$ and

$$\tilde{\psi}_1(x) \quad = \quad \log x - \frac{1}{2x} - \frac{1}{12x^2} + \frac{1}{120x^4} - \frac{1}{256x^6}$$

$$\begin{aligned}\tilde{\psi}_2(x) &= \frac{1}{x}+\frac{1}{2x^2}+\frac{1}{6x^3}-\frac{1}{30x^5}+\frac{1}{42x^7}\\ \tilde{\psi}_3(x) &= -\frac{1}{x^2}-\frac{1}{x^3}-\frac{1}{2x^4}+\frac{1}{6x^6}+\frac{1}{6x^8}.\end{aligned}$$

5.13.4 Example

Inspired by a real data example, we consider a somewhat reduced typical sensory profile data set. Assume that 4 judges have assessed 2 products in 2 replications and assume the model (5.2)-(5.3). To compare the actual distribution of $-2\log Q$ under the null hypothesis, observations were generated based on the following parameter values: $\mu^x = \mu^y = 10$, $\beta_1^x = \beta_1^y = (\beta\delta)_{11} = (\beta\delta)_{12} = (\beta\delta)_{21} = 0$, $\beta_2 = 4$, $\delta_2 = 2$, $(\beta\delta)_{22} = -8$, and

$$\Lambda_E = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \quad , \quad \Lambda_J = \begin{pmatrix} 15 & 0 \\ 0 & 27 \end{pmatrix}$$
$$\Lambda_{J\times S} = \begin{pmatrix} 6 & 0 \\ 0 & 13 \end{pmatrix} \quad , \quad \Lambda_{J\times R} = \begin{pmatrix} 3 & 0 \\ 0 & 4 \end{pmatrix}$$

The four random effect SSD's are all on 3 degrees of freedom. 500 simulations were performed, and Figure 5.8 shows that the saddlepoint approximation is extremely superior to the basic χ^2-approximation on 4 degrees of freedom.

5.13.5 Design assumptions

In the following discussion some concepts may not be explicitly defined. The meaning should, however, be clear from the context and the reader is referred to Tjur (1991) for detailed definitions. The four conditions in Tjur (1991) set up for the collection of random effects $\mathcal{B}$ are:

($\mathcal{B}1$) The error effect is included in $\mathcal{B}$.

($\mathcal{B}2$) Any factor $B \in \mathcal{B}$ is balanced.

($\mathcal{B}3$) $\mathcal{B}$ is closed under formation of infima.

($\mathcal{B}4$) If X_B denotes the $n \times |B|$ design matrix for n observations corresponding to the factor $B \in \mathcal{B}$, then the matrices(vectors) $\left(X_B X_B^T\right)_{B\in\mathcal{B}}$ are linearly independent.

Condition ($\mathcal{B}3$) will be violated when the replication (block) effect, and thus also the $R \times S$-interaction, in the example above is considered random. This situation, or something similar, will often be relevant in sensory experiments. The unique

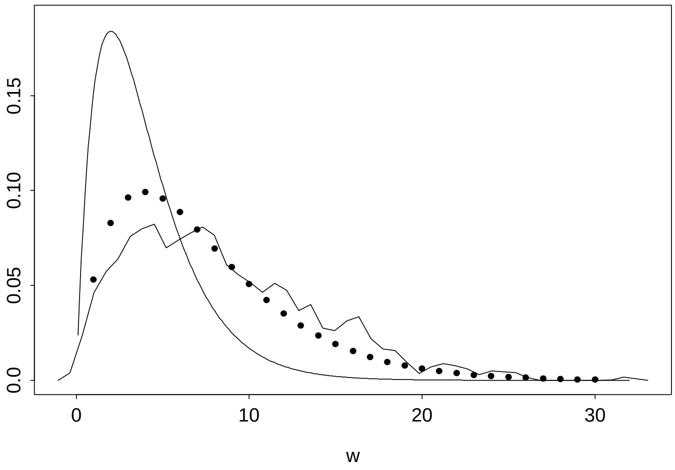

Figure 5.8: Empirical density of $-2\log Q$ based on 500 simulations, the $\chi^2(4)$-density (the smooth curve) and the saddlepoint approximated density corresponding to the approximation used for the tail probability (points), see Chapter 7

partitioning of factors into strata is then lost, but for the mentioned example the test for overall independence can still be performed as outlined. This is so, since the distribution of the collection of now five mean squares is still specified by (5.20), and similarly under the hypothesis.

If missing values are present in a data set, the condition ($\mathcal{B}2$) will typically be violated, but if the missing data do not destroy the estimability of the parameters of the fixed part of the model nor the variance component estimability condition ($\mathcal{B}4$), a test for overall independence as described can still be performed in the following sequentially defined way:

Method: First pull out the fixed effects, next sequentially calculate the mean squares for each of the random effects in some chosen order, and for these the test can be performed as described earlier.

This corresponds to making a choice of one out of many possible orthogonal partitionings of the observation space and basing the testing on that choice. The important point is that the hypothesis of overall independence will still be equivalent to independence in each expected mean square.

It is therefore clear that we can formulate the following proposition about the validity of the test procedure just given.

Proposition 5 *Let $\mathcal{F}$ and $\mathcal{B}$ denote the fixed resp. random effects of a mixed linear normal model. If $\mathcal{B}$ satisfies ($\mathcal{B}1$), ($\mathcal{B}2$) and the following*

($\mathcal{B}5$) *No factor in $\mathcal{F}$ is finer than a factor in $\mathcal{B}$,*

then the test method just given is valid.

The condition ($\mathcal{B}5$) ensures that each expected mean square corresponding to random effects are linear combinations of variance components.

Chapter 6

Three-way factor methods in sensory analysis

This paper is in its final version as it will appear as a chapter in the book *Multivariate Statistics for Sensory Data* in 1995. The book will have sensory scientists and experimenters as target group together with, of course, sensometricians.

Three-way Factor Methods in Sensory Analysis

Per M. Brockhoff
Royal Veterinary and Agricultural University
Thorvaldsensvej 40, DK-1871 Frederiksberg C
Denmark,

David Hirst
Scottish Agricultural Statistics Service
Rowett Research Institute
Bucksburn, Aberdeen AB2 9SB
Scotland

and

Tormod Næs
MATFORSK
Oslovegen 1, 1430 Ås,
Norway.

6.1 Introduction

6.1.1 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.

6.1.2 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}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and $k = 1, \ldots, p$. Replicates will here be denoted by $l = 1, \ldots, q$. The handling of replicates is discussed in Section 6.7. They can either be averaged over or treated separately, in which case each of the $m \times n \times p$ cells of the three-way matrix of data consists of q elements. This type of data can always be described by an analysis of variance model, see Searle (1971),

$$x_{ijkl} = \mu_k + \alpha_{ik} + \beta_{jk} + \delta_{ijk} + \varepsilon_{ijkl} \tag{6.1}$$

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_{jk} 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 α_{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.

6.2 Different TWFA models

6.2.1 TWFA as a generalization of PCA

As discussed in detail elsewhere in this book, standard PCA of an $n \times p$ matrix X is based on the following 'model'

$$X = TP' + E \tag{6.2}$$

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

$$\left\| X - TP' \right\|^2 \tag{6.3}$$

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 6.1. 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.

6.2.2 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 $i = 1, \ldots, m$, then one possible way to analyse

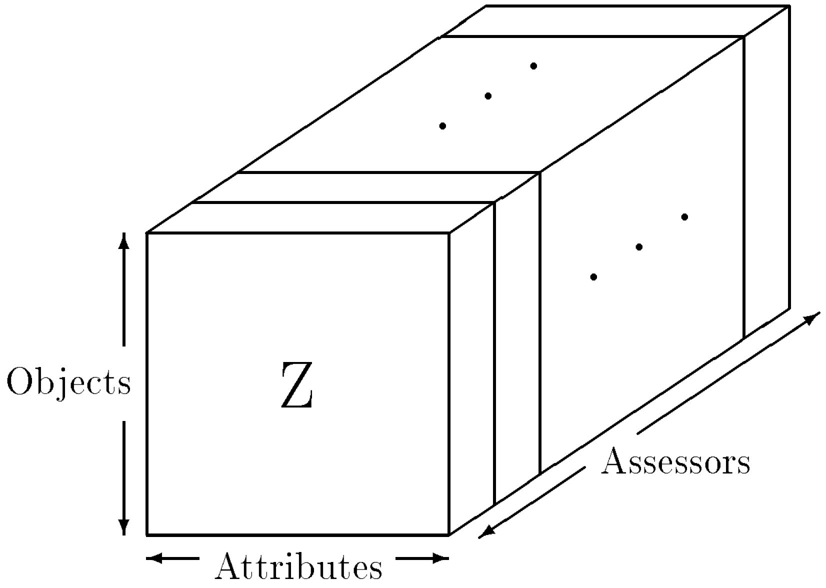

Figure 6.1: Three-way data matrix

the data is to model X_i as

$$X_i = T_iP' + E_i, \tag{6.4}$$

where P has dimension $p \times a$ $(a < p)$. 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

$$\sum_{i=1}^{m} \left\| X_i - T_iP' \right\|^2 . \tag{6.5}$$

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 $m \times p$ slice corresponding to the assessors-by-variables matrix for object j be Y_j, then we can write

$$Y_j = U_jP' + E_j. \tag{6.6}$$

Then minimizing $\sum_{j=1}^{n} \left\| Y_j - U_jP' \right\|^2$ will give exactly the same common loadings matrix P (and the same fit).

Alternatively, TWFA models can be based on the model

$$X_i = TP_i' + E_i,\ i = 1, \ldots, m \tag{6.7}$$

where now the loadings P_i differ from assessor to assessor. T has dimension $n \times b$ 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 $n \times m$ slice of objects-by-assessors for variable k.

There is also a third Tucker-1 model formed by writing

$$Z_k' = QW_k + E_k \tag{6.8}$$

or $Y_j = QR_j + E_j$. Here Q is the 'assessor scores' matrix with dimension $m \times c$, 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 $m \times na$ 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.

6.2.3 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, $i = 1, \ldots, m$. These W_i relate T and P through a different linear transformation for each assessor. This model is written as

$$X_i = TW_iP' \tag{6.9}$$

where the W_i have dimension $b \times a$. T $(n \times b)$ and P $(p \times a)$ are found to minimize the least squares criterion

$$\sum_{i=1}^{m} \left\| X_i - TW_iP' \right\|^2 \tag{6.10}$$

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

$$Y_j = QO_jP' \tag{6.11}$$

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_j's 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.

6.2.4 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 (6.9), 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 (6.11) 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

$$Z = TC(Q' \otimes P'), \tag{6.12}$$

where $\otimes$ is the kronecker or direct product. Z is the data unfolded to form an $n \times mp$ matrix of objects-by-(assessors $\times$ 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.

6.2.5 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

$$\begin{pmatrix} 10 & 2 \\ 4 & 8 \end{pmatrix}$$

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 $C_j P'$. 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.

6.2.6 The PARAFAC model

The other three mode method is the PARAFAC model which is defined by equation (6.9) 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 (6.9) the object and attribute dimensions would have to be equal. This is not the case with the Tucker-3 model.

6.2.7 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 (6.4). This results in an $a \times p$ common loadings matrix P, and m individual $n \times a$ scores matrices T_i, $i = 1, \ldots, m$. 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 (6.10). For these reasons this method is not to be recommended if Tucker-3 programs are available.

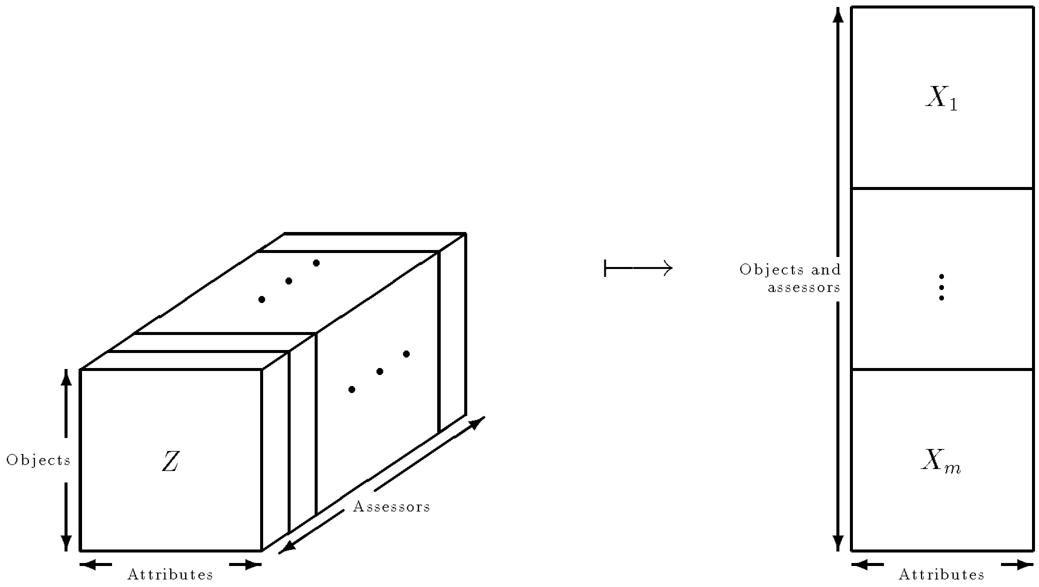

Figure 6.2: Unfolding of three-way data matrix.

6.3 Fitting the TWFA models

6.3.1 Tucker-1

The Tucker-1 model is based on unrestricted minimization over T_i and P of the quantity

$$\sum_{i=1}^{m} \left\| X_i - T_i P' \right\|^2, \quad P'P = I \tag{6.13}$$

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 6.2, 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, $i = 1, \ldots, m$, where S_i is the covariance matrix for assessor i.

6.3.2 Tucker-2

For this case the minimization is over W_i, T and P of the equation

$$\sum_{i=1}^{m} \left\| X_i - T W_i P' \right\|^2 \tag{6.14}$$

(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 (6.14) 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 $D = \sum_{i=1}^{m} X_i P P' X_i'$.

3. Put the eigenvectors associated with the b largest eigenvalues of D into the columns of a matrix T.

4. Compute $Q = \sum_{i=1}^{m} X_i' T T' X_i$.

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 $W_i = T' X_i P$.

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 (6.14) 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.

6.3.3 PARAFAC-CANDECOMP

In this case the optimization criterion is the same as above, namely

$$\sum_{i=1}^{m} \left\| X_i - TW_iP' \right\|^2 \tag{6.15}$$

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.

6.3.4 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,

$$W_i = \sum_{j=1}^{n} c_{ij} C_j \tag{6.16}$$

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 6.2.4. 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 $Z_1 Z_1'$.
   - P is formed from the first a eigenvectors of $Z_3 Z_3'$.
3. Q is formed from the first c eigenvectors of $Z_2 \left(TT' \otimes PP'\right) Z_2'$.
4. P is formed from the first a eigenvectors of $Z_3 \left(QQ' \otimes TT'\right) Z_3'$.
5. T is formed from the first b eigenvectors of $Z_1 \left(QQ' \otimes PP'\right) Z_1'$.
6. Repeat steps 3 to 5 until convergence
7. Put $C = T'Z\left(Q \otimes P\right)$

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.

6.4 Relationships to other work

6.4.1 Generalized Procrustes Analysis

The Procrustes rotation method discussed elsewhere in 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 6.2.2 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 $\sum_{i=1}^{m} \left\| X_i - TP_i' \right\|^2$. In this case P_i is a general matrix, but if it is forced to be orthogonal, then we can write

$$X_i P_i = T + E_i P_i, \qquad (6.17)$$

i.e. the common scores are found by rotating the original 'configurations' X_i to minimize

$$\sum_{i=1}^{m} \left\| X_i P_i - T \right\|^2 \qquad (6.18)$$

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.

6.4.2 Individual differences MDS versus TWFA

Consider the common scores and common loadings Tucker-2 model (6.9). 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

$$\mathrm{f}_{ij} = t_j W_i P' \tag{6.19}$$

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

$$\begin{aligned} D_{ij_1j_2} &= (t_{j_1} - t_{j_2})W_i P' P W_i' (t_{j_1} - t_{j_2})' \\ &= (t_{j_1} - t_{j_2})W_i W_i' (t_{j_1} - t_{j_2})' \\ &= (t_{j_1} - t_{j_2})V_i (t_{j_1} - t_{j_2})' \end{aligned} \tag{6.20}$$

where V_i is a general symmetric matrix. Hence we can write the Tucker-2 model as

$$(x_{ij_1} - x_{ij_2})(x_{ij_1} - x_{ij_2})' = (t_{j_1} - t_{j_2})V_i (t_{j_1} - t_{j_2})' \tag{6.21}$$

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

$$D_{ij_1j_2} = (t_{j_1} - t_{j_2})W_i P' P W_i' (t_{j_1} - t_{j_2}) = (t_{j_1} - t_{j_2})V_i (t_{j_1} - t_{j_2})' \tag{6.22}$$

where now V_i is diagonal with nonnegative diagonal elements. Therefore we have

$$D_{ij_1j_2} = \sum_{k=1}^{b} (t_{j_1k} - t_{j_2k})^2 v_{ki} \tag{6.23}$$

which is exactly the INDSCAL model used for individual differences MDS.

The individual differences MDS models are treated elsewhere in the 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.

6.4.3 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.

6.4.4 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.

6.5 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.

6.5.1 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 (6.1).

6.5.2 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.

6.6 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 6.3. 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.

$$T = BY + E. \tag{6.24}$$

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

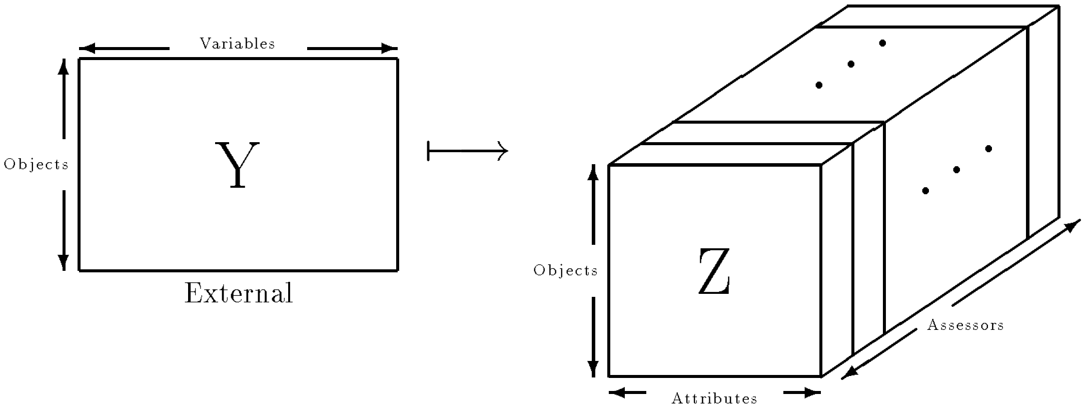

Figure 6.3: Data setup with external information.

parameters W, B and P are optimized by for instance the least squares criterion

$$\sum_{i=1}^{m} \left\| X_i - TW_iP' \right\|^2 \tag{6.25}$$

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.

CANDELINC, Carrol et al. (1980) is a method that is designed for optimization of equations like (6.25). 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_i's 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.

6.7 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.

6.8 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.

6.9 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 (6.1) 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

$$\bar{x}_{\cdot\cdot k} + (\bar{x}_{i\cdot k} - \bar{x}_{\cdot\cdot k}) + (\bar{x}_{\cdot jk} - \bar{x}_{\cdot\cdot k}) \tag{6.26}$$

This is identical to taking the sum of the mean over the assessors and the mean over the samples and subtracting the grand mean.

6.10 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 $X_i = T_iP' + E_i$, 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 up, 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.

6.11 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 the 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.

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

Table 6.1: The 12 cheeses with the names used in plots.

6.12 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 6.1. 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 6.2. 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.

6.12.1 PCA of the cheese data

In order to compare TWFA with more conventional methods a principal components analysis was performed on the object-by-attribute 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 6.4(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

	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

Table 6.2: The 14 common attributes with the names used in plots.

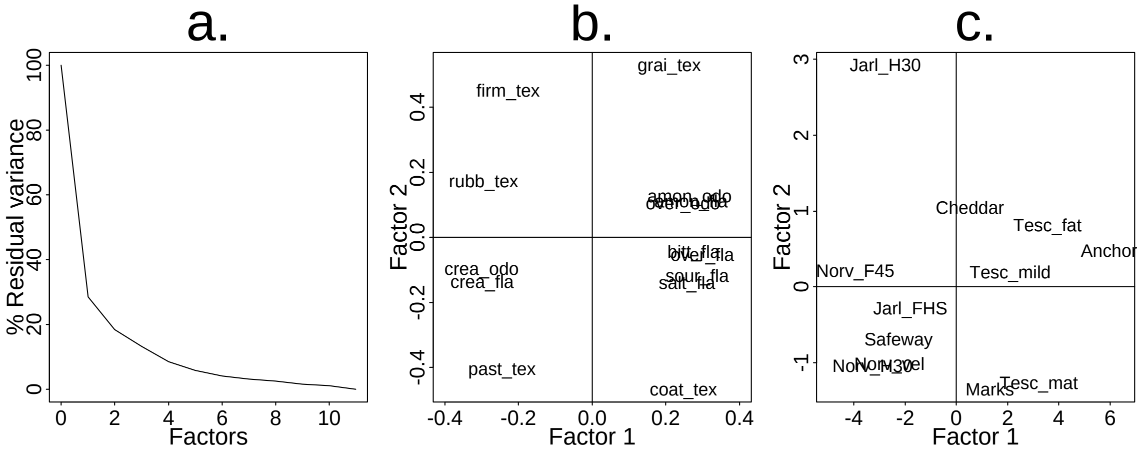

Figure 6.4: Results from PCA on assessor mean scores. (a) Residual variance, (b) loadings and (c) scores for the first two factors.

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.

6.12.2 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.

Scores and loadings for the first two factors are plotted in Figures 6.5(a) and (b) (common scores model) and Figures 6.6(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

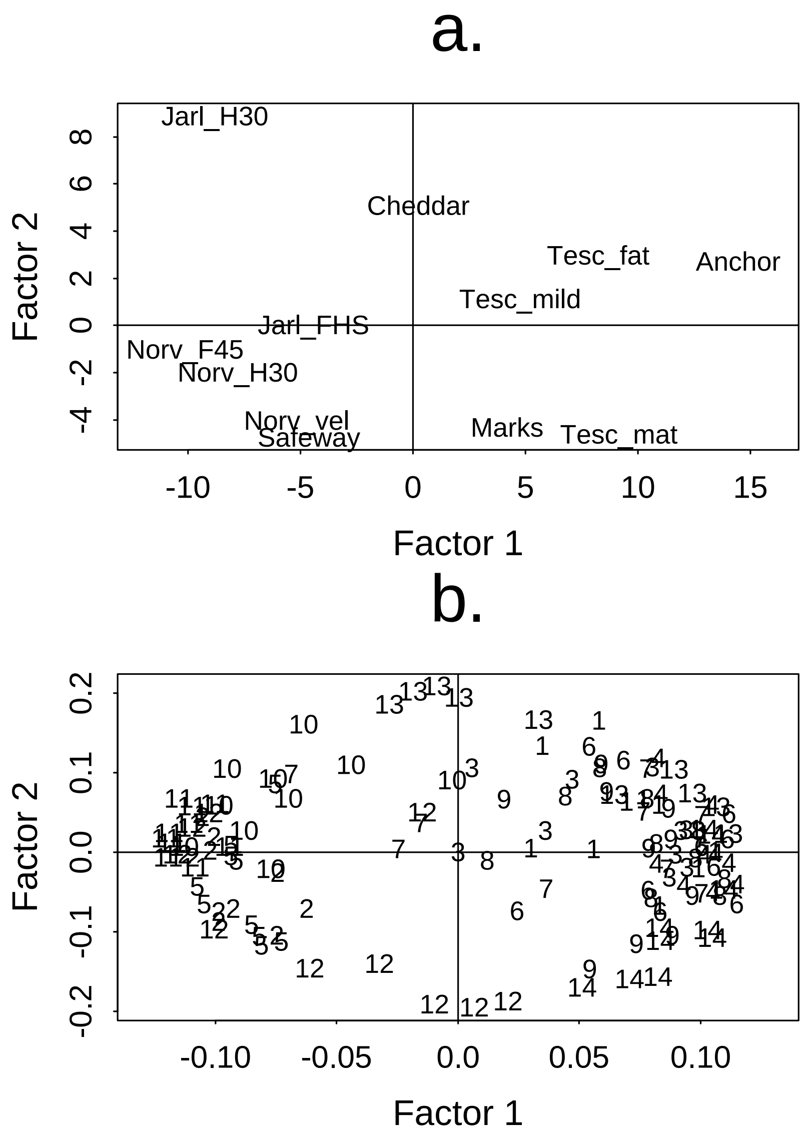

Figure 6.5: Scores (a) and loadings (b) for the first two factors in the 'common scores' version of the Tucker-1 model. The numbers in the loadings plot (b) refers to the attributes, cf. Table 6.2.

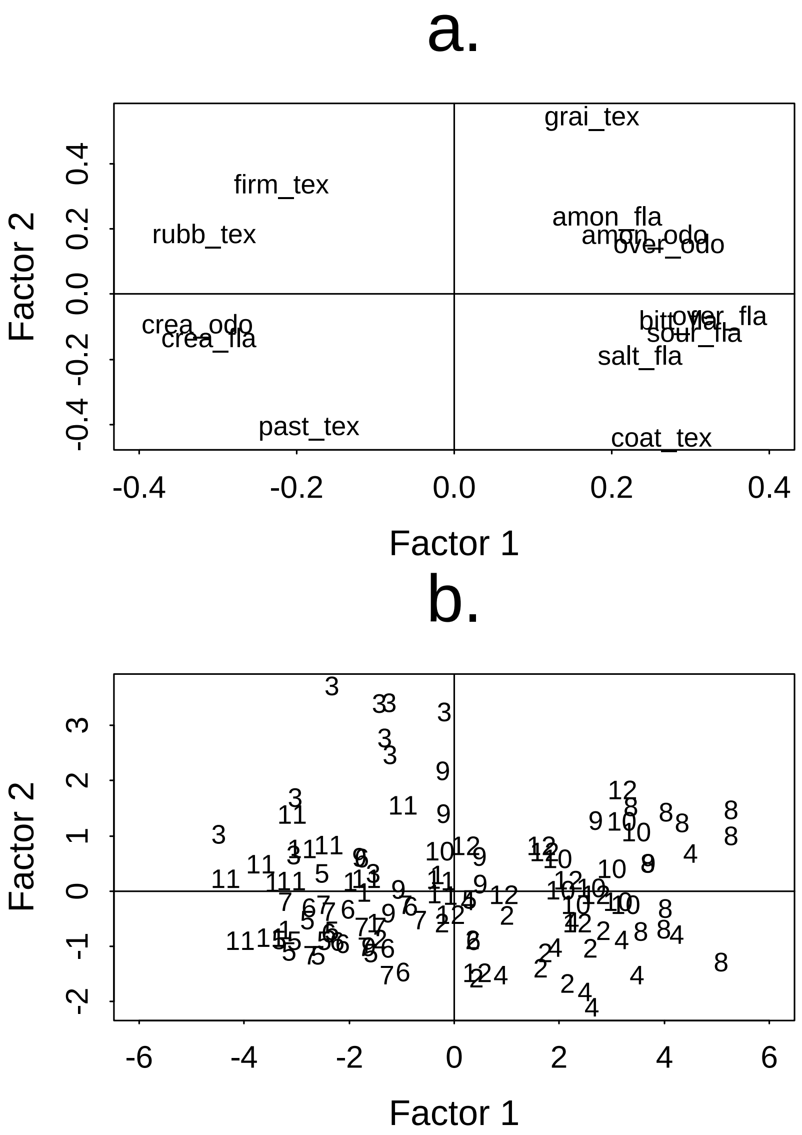

Figure 6.6: Loadings (a) and scores (b) for the first two factors in the 'common loadings' version of the Tucker-1 model. The numbers in the loading plot (b) refers to the cheeses, cf. Table 6.1.

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.

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 6.5(b)) (or all 120 assessor scores on the common loadings plot, Figure 6.6(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.

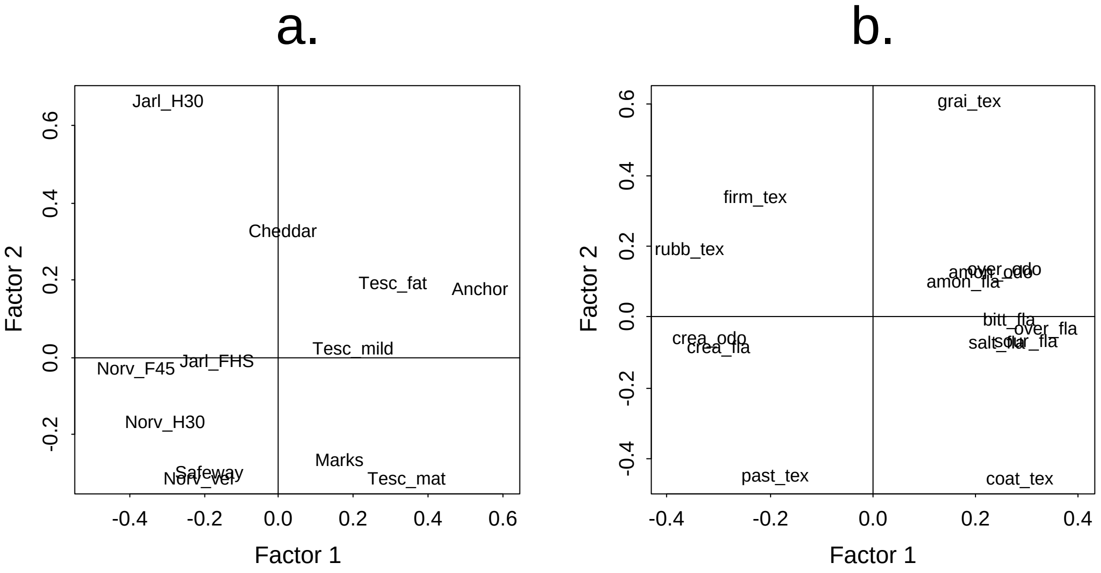

Figure 6.7: Common scores (a) and common loadings (b) for the two factors in the Tucker-2 model $a = b = 2$.

6.12.3 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 6.11, was fitted by performing the algorithm of section 6.3.2. 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 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 6.7(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 6.8(a)–(j) show these individual scores plots, which can be directly interpreted together with the common loadings plot, Figure 6.7(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 6.4(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.

6.12.4 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 6.9(a), (b) and (c).

In the assessor plot Figure 6.9(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

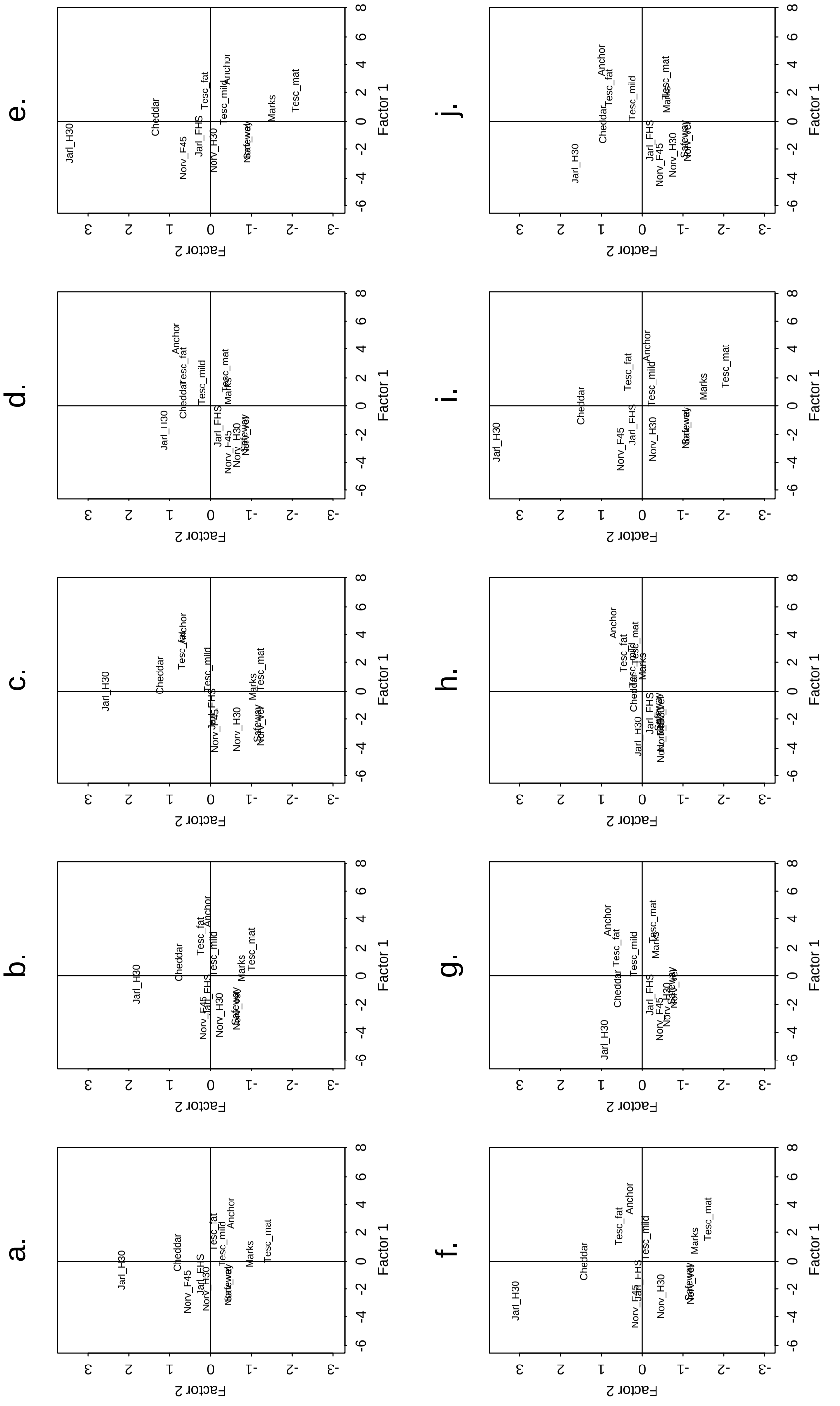

Figure 6.8: Individual scores for the 10 assessors for the two factors in the Tucker-2 model with $a = b = 2$.

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 6.9(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.

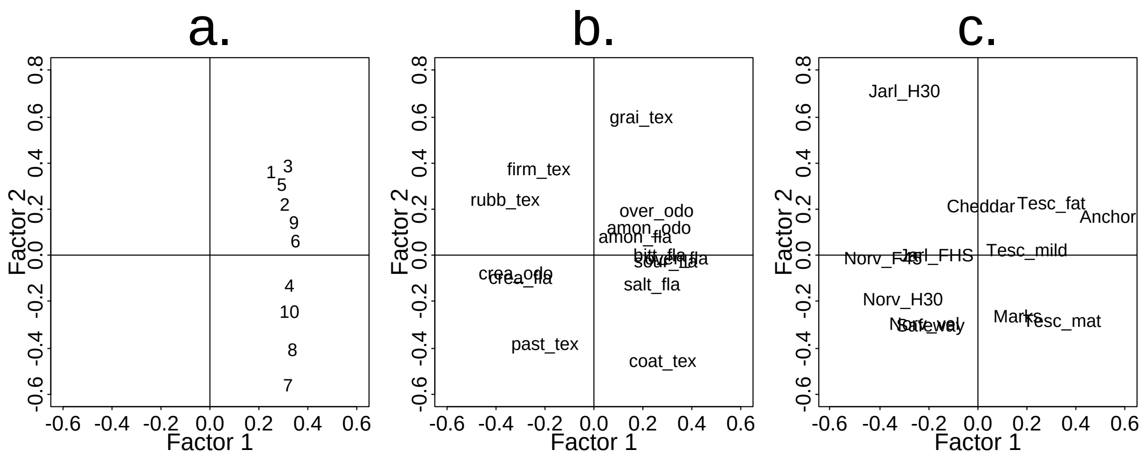

Figure 6.9: Results from the two-factor Tucker-3 model. (a) Assessor scores, (b) loadings and (c) cheese scores.

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 6.9(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

$$\begin{pmatrix} 6.8 & -1.5 \\ 2.1 & 4.1 \end{pmatrix} \text{ to } \begin{pmatrix} 8.3 & 1.4 \\ -2.4 & 1.0 \end{pmatrix}$$

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.

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, ie that they are sticky and doughy, represented by the attributes past_tex and coat_tex in Figure 6.9, whereas assessor 7 would describe them as grainy and hard, corresponding to the positions of rubb_tex, firm_tex and grai_tex in Figure 6.9. 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

$$\sum_{k=1}^{14}\sum_{i=1}^{10}\sum_{j=1}^{12}\frac{(x_{ijk}-\bar{x}_{...})^2}{SD_{ik}^2}=\sum_{k=1}^{14}\sum_{i=1}^{10}\frac{11SD_{ik}^2}{SD_{ik}}=140\cdot 11=1540.$$

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.

6.12.5 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.

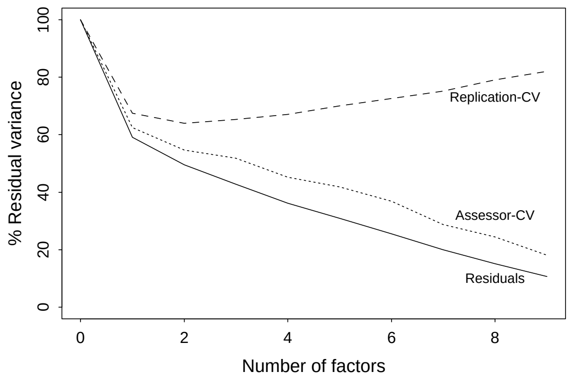

Figure 6.10: Regular and cross validated scree plots for Tucker-2 models with $a=b$.

Consider the Tucker-2 model first. In Figure 6.10 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

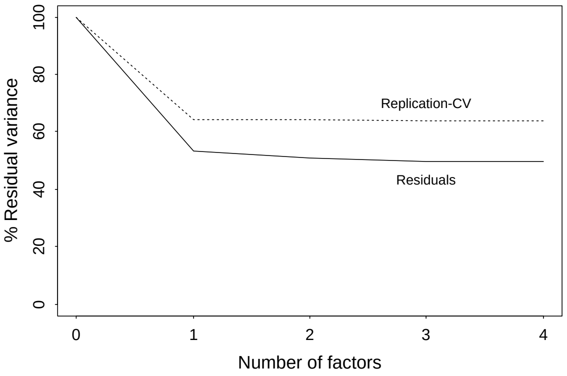

Figure 6.11: Regular and cross validated scree plots for Tucker-3 models with $a = b = 2$.

validated residual variance also seems to have leveled off at factor 2, maybe even at factor 1.

Fixing $a = b = 2$ we now turn to choice of dimension in the Tucker-3 model. In Figure 6.11 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 6.8. The Figures 6.12(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 6.8(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.

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.

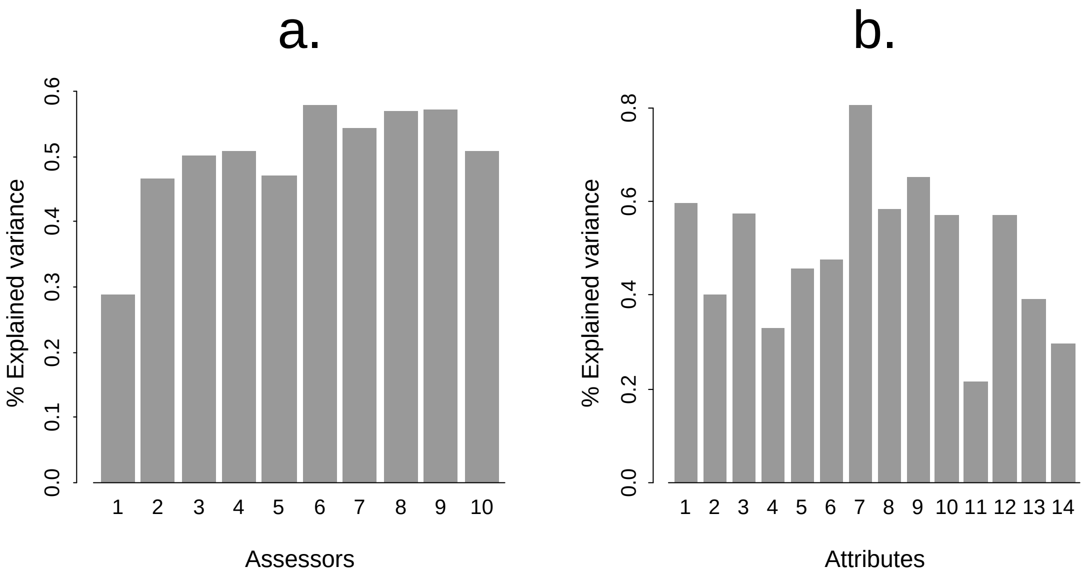

Figure 6.12: Assessor-wise and attribute-wise relative explained variance by the Tucker-2 model with $a = b = 2$.

6.12.6 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 6.5.2 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

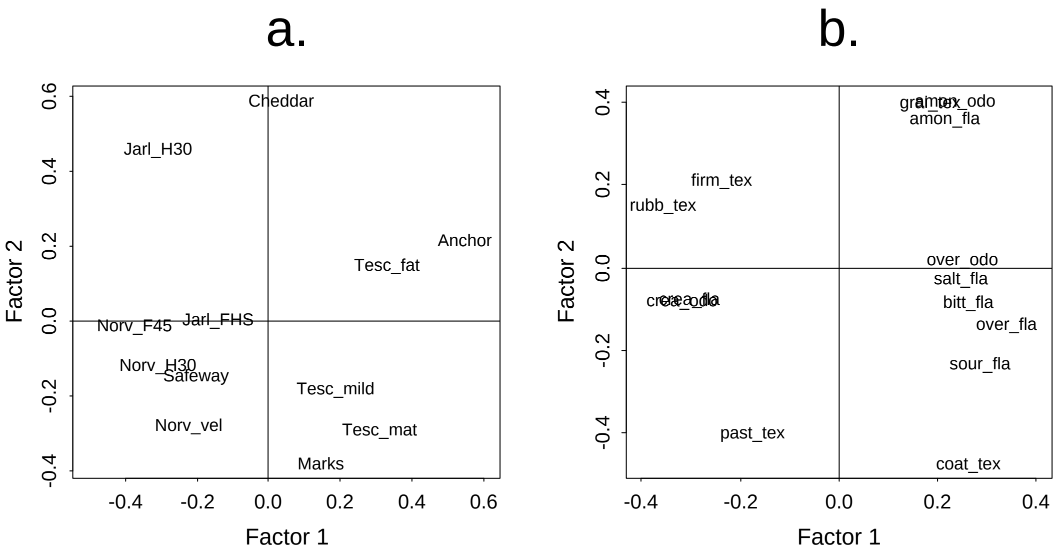

Figure 6.13: Common scores (a) and common loadings (b) for the two factors in the Tucker-2 model with $a = b = 2$ with the alternative pretreatment of the data.

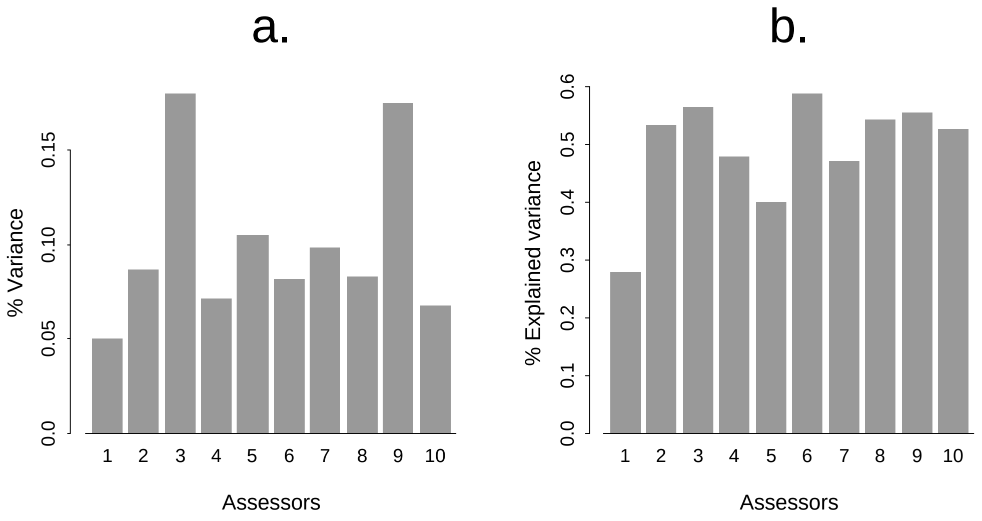

Figure 6.14: Variability in the alternative pretreated data. (a) Total variability for each assessor, (b) relative explained variance for each assessor by the Tucker-2 model with $a = b = 2$.

model with $a = b = 2$. Figures 6.13(a) and (b) show the common scores and loadings.

Comparing with the Tucker-2 model for the former pretreatment, Figures 6.7(a) and (b), and mean score PCA, Figures 6.4(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 6.15(a)–(j), show the same patterns as do Figures 6.8(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 6.14(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 6.14(a) are directly related to the spread of the cheeses in the individual score plots, Figures 6.15(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 6.13(a), they span the direction of the individual scores of assessor 6. This tendency is observed for all the individuals.

6.13 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

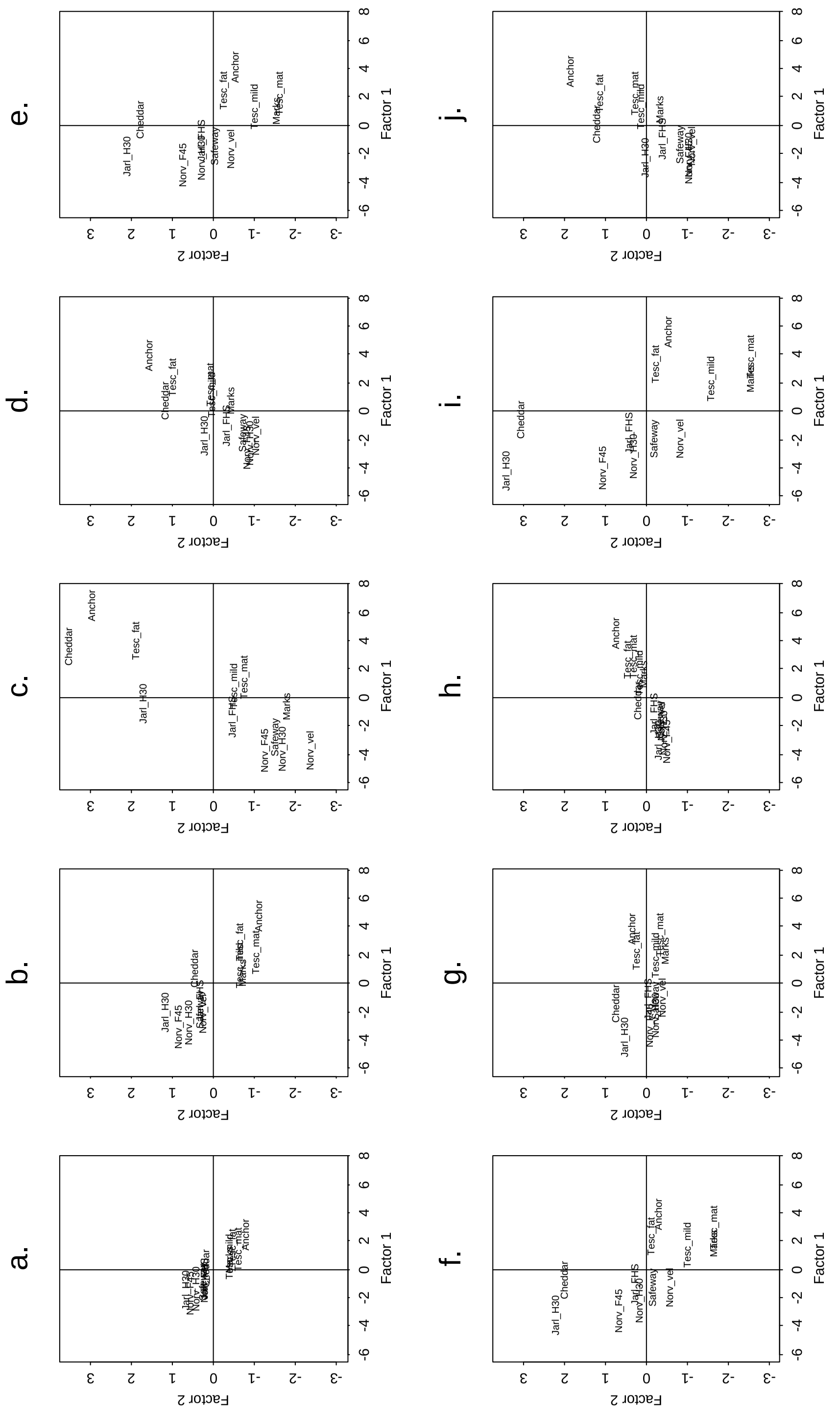

Figure 6.15: Individual scores for the 10 assessors in the Tucker-2 model with $a = b = 2$ for the alternatively pretreated data

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.

6.14 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., 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.

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.

Chapter 7

Application of saddlepoint approximations

7.1 Introduction

In this chapter a brief review of the saddlepoint approximations of densities and tail probabilities are given. Saddlepoint approximations based on the Gamma distribution are discussed. The density approximations will be derived though their properties not proven, and tail probability approximations will just be given. This review is partly based on Skovgaard (1994). The special case of linear combinations of chi-squares will be given specific treatment and it will be shown how this can be used for a general applicable approach to the handling of non-standard F-tests in mixed linear models.

7.2 Normal-based saddlepoint approximations of densities

Let f be the density function of a real random variable X, let ϕ denote the Laplace transform of X,

$$\phi(t) = \mathrm{E}\left(\mathrm{e}^{tX}\right)$$

and let T be the set where the Laplace transform is finite,

$$T = \{t \in \mathbb{R} \mid \phi(t) < \infty\}.$$

The saddlepoint of the density at $x \in \mathbb{R}$, $f(x)$, is based on the Laplace transform, and two conceptual different ways of deriving the same approximation exist, see for instance Reid (1988). We prefer the method of 'exponential tilting' that uses a rewriting of the density as

$$f(x) = f_t(x)\mathrm{e}^{-xt+\kappa(t)},\ t \in T, \tag{7.1}$$

where $\kappa(t)$ is the cumulant generating function (c.g.f.) of X,

$$\kappa(t) = \log\phi(t)$$

and $f_t(x)$ is a member of the conjugate exponential family,

$$f_t(x) = f(x)\mathrm{e}^{xt-\kappa(t)},$$

with mean $\kappa^{'}(t)$ and variance $\kappa^{''}(t)$. The basic idea is simply to use a Normal approximation, corresponding to the one term Edgeworth expansion, of the density f_t instead of the original f. In general, a Normal approximation is only expected to be reasonable in points close to the mean of the distribution to be approximated. Now t can be chosen such that $f_t(x)$ has mean x, say t_x, i.e. t_x is the solution to the (saddlepoint) equation

$$\kappa^{'}(t) = x, \tag{7.2}$$

and instead the density $f_{t_x}(x)$ is approximated by a Normal density with mean $\kappa^{'}(t_x) = x$ and variance $\kappa^{''}(t_x)$, i.e.

$$f(x) \approx \left(2\pi\kappa^{''}(t_x)\right)^{-1/2} \mathrm{e}^{-xt_x+\kappa(t_x)}. \tag{7.3}$$

The solution $t_x \in T$ exists uniquely for a regular family, see Barndorff-Nielsen (1978), when x is in the interior of the support of X, and as $\kappa^{'}(t)$ is increasing, t_x can always easily be found by a numerical line search.

7.3 Convergence properties

Often the interest is in the density, f_n, of the mean, $\bar{x}$, of n independent identical distributed observations, in which case (7.3) becomes

$$f_n(x) \approx \left(\frac{n}{2\pi\kappa^{''}(t_x)}\right)^{1/2} \mathrm{e}^{n(-xt_x+\kappa(t_x))}, \tag{7.4}$$

where κ is the c.g.f. corresponding to a single observation. This is an improvement compared to 1. and 2. order Edgeworth expansions as can be seen from Table 7.1. A main point is that Edgeworth approximations have unbounded relative error whereas the saddlepoint approximations have bounded (or vanishing) relative error in the tails.

Case number four and five in Table 7.1 refer to uniformity results of Jensen (1988). Under certain conditions the order of relative approximation error holds uniformly

Table 7.1: Convergence properties for density approximations $\hat{f}_n$. $R_n(x)$ is the absolute error $f_n(x) - \hat{f}_(x)$ and $RR_n(x)$ the relative error $R_n(x)/f_n(x)$. The statements with c_1 and c_2 should be understood as an existence of constants c_1 and c_2 such that the statement holds.

	As $n \to \infty$
1. order Edgeworth	$\sup_{-c_1 \leq x \leq c_2} \mid R_n(x) \mid = O\left(\frac{1}{\sqrt{n}}\right)$, $RR_n(x)$ unbounded
2. order Edgeworth	$\sup_{-c_1 \leq x \leq c_2} \mid R_n(x) \mid = O\left(\frac{1}{n}\right)$, $RR_n(x)$ unbounded
Normal-based saddlepoint	$\sup_{-c_1\sqrt{n} \leq x \leq c_2\sqrt{n}} \mid RR_n(x) \mid = O\left(\frac{1}{n}\right)$
Normal-based saddlepoint with additional assumptions	$\sup_{-c_1\sqrt{n} \leq x} \mid RR_n(x) \mid = O\left(\frac{1}{n}\right)$
Normal/Gamma-based saddlepoint with additional assumptions	$\sup_{-c_1\sqrt{n} \leq x} \mid RR_n(x) \mid = O\left(\frac{1}{n}\right)$ For n fixed: $\mid RR_n(x) \mid \to 0$ for $x \to \infty$

as x tends to the tail of the distribution, and even in some cases the approximation becomes 'exact in the limit'. This is a strong result that states, that for any n the relative error tends to zero in the tail. These results also form the basis of using saddlepoint approximations in a 'non-asymptotic' way as indicated by (7.3).

7.4 Gamma-based saddlepoint approximations of densities

We now go back to the notation without n. The difference between case number four and case number five in Table 7.1 really lies in the choice of distribution to use for the approximation of $f_t(x)$. The heuristics are, that if a distribution corresponding to the limiting distribution of $f_t(x)$ is chosen, the approximation becomes exact in the limit. For some classes of densities this choice is the Gamma distribution.

If X is a positive random variable, $f_t(x)$ can be approximated by the Gamma distribution, $\Gamma(\alpha, \beta)$, with the 'correct' mean and variance, i.e. α and β is chosen according to

$$\kappa^{'}(t_x) = \alpha\beta \text{ and } \kappa^{''}(t_x) = \alpha\beta^2.$$

This is the approach of Jensen (1991), although in that paper applied for tail probabilities. We will return to this below.

More general if X is not known to be positive, the approximating distribution can be chosen as a shifted (minus) Gamma distribution, i.e. among the class of densities $\{g(x;\mu,\alpha,\beta) \mid \beta \neq 0, \mu \in \mathbb{R}, \alpha > 0\}$, where

$$g(x;\mu,\alpha,\beta) = \frac{1}{\Gamma(\alpha)|\beta|}\left(\frac{x-\mu}{\beta}\right)^{\alpha-1} \mathrm{e}^{-(x-\mu)/\beta} \ , \ \frac{x-\mu}{\beta} > 0.$$

In Jensen (1988) the shifted Gamma density matching the first three cumulants of $f_{t_x}(x)$ is chosen, that is, μ, α and β are chosen according to the relations

$$\mu + \alpha\beta = x \ , \ \alpha\beta^2 = \kappa''(t_x) \ , \ 2\alpha\beta^3 = \kappa'''(t_x),$$

which leads to the following estimates,

$$\hat{\mu} = x - 2\frac{\kappa''(t_x)^2}{\kappa'''(t_x)} \ , \ \hat{\alpha} = 4\frac{\kappa''(t_x)^3}{\kappa'''(t_x)^2} \ , \ \hat{\beta} = \frac{\kappa'''(t_x)}{2\kappa''(t_x)}.$$

Thus the approximation can be written as

$$f(x) \approx \mathrm{e}^{-xt_x+\kappa(t_x)}.\frac{\hat{\alpha}^{\hat{\alpha}-1}}{\Gamma(\hat{\alpha})|\hat{\beta}|\mathrm{e}^{\hat{\alpha}}} \tag{7.5}$$

This is exactly the approximation given in equation (3.7) in Jensen (1988), although there derived in a slightly different way.

7.5 Approximation of tail probabilities

Often the entity of interest is a tail probability rather than just the density of X, for instance a test-probability. We thus search for an approximation of

$$P(X \geq x) = \int_x^\infty f(x)dx,$$

which may be found as

$$P(X \geq x) \approx \int_x^\infty \hat{f}(x)dx,$$

where $\hat{f}(x)$ is a saddlepoint approximation of the density. This direct integral approach may not be the easiest way to derive approximations for the tail probability, see Daniels (1987). We will very briefly discuss different methods and interpret them as integral versions (if possible) of density approximations. Lugannani and Rice (1980) derived at an integral version of the Normal-based saddlepoint approximation (7.3), of which probably the latter is more widespread. The Lugannani-Rice formula is the following,

$$P(X \geq x) = 1 - \Phi(\hat{w}) + \phi(\hat{u}) \left\{ \frac{1}{\hat{u}} - \frac{1}{\hat{w}} \right\}, \tag{7.6}$$

where Φ and ϕ now denote the standard Normal distribution and density functions,

$$\begin{aligned} \hat{w} &= \sqrt{2(t_x x - \kappa(t_x))} \mathrm{sign}(t_x) \\ \hat{u} &= t_x \sqrt{\kappa''(t_x)}, \end{aligned}$$

and t_x is the saddlepoint as before. Wood *et al.* (1993) deduces a generalized Lugannani-Rice formula, that in its appearance has Φ and ϕ substituted by the distribution and density functions for a general base distribution, which may be the Gamma. In Jensen (1988) a Gamma-approximation corresponding to (7.5) was worked out and has the form, here taken from Jensen (1995), where for simplicity $t = t_x$ and $\sigma_t = \sqrt{\kappa(t)^2}$ is used,

$$P(X \geq x) \approx \frac{\mathrm{e}^{\kappa(t) - tx}}{t\sigma_t} A_0 \left(t\sigma_t, \lambda_t; \mathrm{sgn} \left[\kappa'''(t) \right] \right), \tag{7.7}$$

where

$$\lambda_t = 4 \frac{\kappa''(t)^3}{\kappa'''(t)^2},$$

$$A_0 (u, \lambda; \delta) = \begin{cases} \frac{u \mathrm{e}^{u\sqrt{\lambda}} \lambda^{\lambda/2}}{(u+\sqrt{\lambda})^{\lambda}} \bar{\Gamma} \left(\lambda, \sqrt{\lambda}(u + \sqrt{\lambda}) \right) & \text{if } \delta = 1 \\ u \lambda^{\lambda} \mathrm{e}^{-u\sqrt{\lambda}} A_1 \left(\lambda, \sqrt{\lambda}(u - \sqrt{\lambda}) \right) & \text{if } \delta = -1, \end{cases} \tag{7.8}$$

$$A_1 (\lambda, z) = \int_0^1 \frac{v^{\lambda-1}}{\Gamma(\lambda)} \mathrm{e}^{zv} dv = \begin{cases} \frac{1 - \bar{\Gamma}(\lambda, -z)}{(-z)^{\lambda}} & z < 0 \\ \int_0^1 \frac{\mathrm{e}^{zv}}{\Gamma(\lambda)} v^{\lambda-1} dv & z \geq 0 \end{cases} \tag{7.9}$$

and

$$\bar{\Gamma}(\lambda, z) = \int_z^{\infty} \frac{v^{\lambda-1}}{\Gamma(\lambda)} \mathrm{e}^{-v} dv. \tag{7.10}$$

The strength of the Jensen-approximations is that they fall in the last category of Table 7.1 above, as the final Theorem 5 of Jensen (1988) states (among other things): If $\int f(x)^p dx < \infty$ for some $p > 1$ and $f(x)$ belongs to one of the classes I or III in Daniels (1954) then the approximations (7.5) and (7.7) are uniformly valid and become exact in the tail (for $n \geq p$).

7.6 Linear combinations of chi-squares

In this section we will study the case where the variable of interest, X, is a linear combination of independent chi-squared distributed random variables,

$$X = \sum_{i=1}^{K} a_i X_i \ , \ X_i \sim \chi^2(f_i) = \Gamma\left(\frac{f_i}{2}, 2\right) . \tag{7.11}$$

The Laplace transform of X is a product of Gamma Laplace transforms,

$$\phi(t) = \prod_{i=1}^{K} (1 - 2a_i t)^{-f_i/2},$$

which is finite for

$$t_- = \max_{a_i<0} \left\{\frac{1}{2a_i}\right\} < t < \min_{a_i>0} \left\{\frac{1}{2a_i}\right\} = t_+,$$

where the usual conventions that $\min\{\emptyset\} = \infty$ and $\max\{\emptyset\} = -\infty$ are used.

In Wood *et al.* (1993) this situation is taken as a numerical example with a comment that 'using arguments similar to those of Jensen (1988, 1991), it can be shown, that in this example, the relative error of the Normal-based Lugannani-Rice approximation stays bounded as x varies over $]0, \infty[$'.

The situation of the present section is also covered by the results of Jensen (1992), which can be seen from the expression for the Laplace transform of $X - \mathrm{E}\, X$,

$$\begin{aligned} \phi_{X-\mathrm{E}\,X}(t) &= \mathrm{e}^{-t\sum_{i=1}^{K} a_i f_i} \phi(t) \\ &= \prod_{i=1}^{K} \frac{\mathrm{e}^{-t a_i f_i}}{(1 - 2a_i t)^{f_i/2}}. \end{aligned}$$

With $\alpha_i = f_i/2$ and $c_i = 1/(2a_i)$ this becomes

$$\phi_{X-\mathrm{E}\,X}(t) = \prod_{i=1}^{K} \frac{\mathrm{e}^{-t\alpha_i/c_i}}{(1 - t/c_i)^{\alpha_i}}, \tag{7.12}$$

which is the class defining expression of Jensen (1992). In that paper uniformity results for an Esscher approximation are shown and explicit expressions for the non-asymptotic tail-behaviour are derived. Our study is also based on a direct approximation of the distribution of X. If asymptotic statements were to be made, two kinds could be relevant in our setup: $f_i \to \infty$, $i = 1, \ldots, K$ or $K \to \infty$.

Based on the results in Jensen (1992) we will (partly) show that the approximation (7.7) in this case provide limiting relative exactness in the tails. Table 7.2 shows the tail situations that can occur depending on the signs of the entering coefficients

Table 7.2: Tail situations for X

	$a_i > 0, \ \forall\, i$ $\mathrm{supp}(Z) = \mathbb{R}_+$	$a_i < 0, \ \forall\, i$ $\mathrm{supp}(Z) = \mathbb{R}_-$	$\exists\, i, j : \ a_i < 0, \ a_j > 0$ $\mathrm{supp}(Z) = \mathbb{R}$
Right tail	$0 < t_+ < \infty$	$t_+ = \infty$	$0 < t_+ < \infty$
$t \to t_+$	$y \to \infty$	$y \to 0$	$y \to \infty$
Left tail	$t_- = -\infty$	$-\infty < t_- < 0$	$-\infty < t_- < 0$
$t \to t_-$	$y \to 0$	$y \to -\infty$	$y \to -\infty$

$a_1, \ldots, a_K$. Viewing any left tail case as a right tail case for $-X$, we are essentially left with two right tail cases: **(i):** $0 < t_+ < \infty$, corresponding to the presence of at least one positive coefficient and **(ii):** $t_+ = \infty$, corresponding to the case where all coefficients are negative. The result is expressed in the following proposition:

Proposition 6 *Let X be given as in (7.11), let $\hat{p}(x)$ be the tail probability approximation (7.7) and assume that $\sum_{i=1}^{K} f_i > 2$. Then the relative error of approximation tends to zero as x tends to infinity, that is*

$$\lim_{x \to \infty} \frac{\hat{p}(x)}{P\,(X \geq x)} = 1. \tag{7.13}$$

Proof in case **(i)**: In the following the notation '$\sim$' is used to express asymptotic equivalence in the sense that

$$f(x) \sim g(x) \Leftrightarrow \lim_{x \to \infty} \frac{f(x)}{g(x)} = 1$$

or equivalently as $x \to \infty \Leftrightarrow t = t_x \uparrow t_+$

$$f(t) \sim g(t) \Leftrightarrow \lim_{t \uparrow t_+} \frac{f(t)}{g(t)} = 1.$$

The first three cumulants of X are given by

$$x = \kappa'(t) = \sum_{i=1}^{K} \frac{a_i f_i}{1 - 2a_i t} \tag{7.14}$$

$$\sigma_t^2 = \kappa''(t) = 2 \sum_{i=1}^{K} \frac{a_i^2 f_i}{(1 - 2a_i t)^2} \tag{7.15}$$

$$\kappa'''(t) = 8 \sum_{i=1}^{K} \frac{a_i^3 f_i}{(1 - 2a_i t)^3}. \tag{7.16}$$

From this it may be seen that $\kappa'''(t) \to \infty$ as $t \uparrow t_+$, i.e. for t large enough we have that

$$\hat{p}(x) = \frac{\phi(t)\mathrm{e}^{-tx}}{t\sigma_t} \frac{t\sigma_t \mathrm{e}^{t\sigma_t\sqrt{\lambda_t}}\lambda_t^{\lambda_t/2}}{\left(t\sigma_t + \sqrt{\lambda_t}\right)^{\lambda_t}} \bar{\Gamma}\left(\lambda_t, \left(t\sigma_t + \sqrt{\lambda_t}\right)\sqrt{\lambda_t}\right). \tag{7.17}$$

From Jensen (1992) we get, since $\sum_{i=1}^{K} f_i > 2$, that

$$\lim_{t \uparrow t_+} d_t = \frac{\alpha_1^{\alpha_1 - 1/2}}{\Gamma(\alpha_1)} \mathrm{e}^{-\alpha_1}, \tag{7.18}$$

where $\alpha_1 = f_1/2$ with the numbering of the coefficients chosen such that $t_+ = 1/(2a_1)$ and where d_t is the inversion integral stemming from

$$P(X \geq x) = \frac{\phi(t)\mathrm{e}^{-tx}}{t\sigma_t} d_t. \tag{7.19}$$

From (7.17), (7.18) and (7.19) it follows that to complete the proof we must verify that

$$\bar{\Gamma}\left(\lambda_t, \left(t\sigma_t + \sqrt{\lambda_t}\right)\sqrt{\lambda_t}\right) \sim \frac{\alpha_1^{\alpha_1 - 1/2}}{\Gamma(\alpha_1)} \mathrm{e}^{-\alpha_1} \frac{\left(t\sigma_t + \sqrt{\lambda_t}\right)^{\lambda_t}}{t\sigma_t \mathrm{e}^{t\sigma_t\sqrt{\lambda_t}}\lambda_t^{\lambda_t/2}} \tag{7.20}$$

The Incomplete Gamma function $\Gamma(\alpha, x) = \int_x^\infty x^{\alpha-1}\mathrm{e}^{-x}dx$ can be asymptotically expanded based on repeated integrations by parts to give

$$\Gamma(\alpha, x) = \mathrm{e}^{-x}x^{\alpha-1} \sum_{p=0}^{n} \frac{(\alpha-1)(\alpha-2)\cdots(\alpha-p)}{x^p} + R_n(\alpha, x) \tag{7.21}$$

where uniformly for α in a compact interval, say $[\alpha_1 - \varepsilon, \alpha_1 + \varepsilon]$,

$$R_n(\alpha, x) = O\left(\mathrm{e}^{-x}x^{\alpha_1 - \varepsilon - n - 1}\right), \quad \text{as } x \to \infty,\ n \geq \alpha_1 + \varepsilon - 1, \tag{7.22}$$

see Olver (1974), page 66–67. This implies that for $x_t = (t\sigma_t + \sqrt{\lambda_t})\sqrt{\lambda_t}$ and $\bar{x}_t = (t\sigma_t + \sqrt{\alpha_1})\sqrt{\alpha_1}$ we get the equivalence

$$\Gamma(\lambda_t, x_t) \sim \mathrm{e}^{-\bar{x}_t}\bar{x}_t^{\alpha_1 - 1} \tag{7.23}$$

since for t large enough λ_t is an ε-interval around α_1, and for n chosen such that $n \geq \alpha_1 + \varepsilon - 1$,

$$\frac{\Gamma(\lambda_t, x_t)}{\mathrm{e}^{-\bar{x}_t}\bar{x}_t^{\alpha_1 - 1}} \sim \frac{\Gamma(\lambda_t, x_t)}{\mathrm{e}^{-x_t}x_t^{la_t - 1}} = \sum_{p=0}^{n} \frac{(\alpha-1)(\alpha-2)\cdots(\alpha-p)}{x^p} + R_n(t),$$

where the finite sum will tend to 1 as $t \uparrow t_+$ and the remainder is

$$\begin{aligned} R_n(\alpha, x) &= \frac{O\left(\mathrm{e}^{-x} x^{\alpha_1-\varepsilon-n-1}\right)}{\mathrm{e}^{-x_t} x_t^{\lambda_t-1}} \\ &= O\left(x_t^{-\varepsilon-n}\right) \\ &\to 0. \end{aligned}$$

To complete the proof we must check that (7.20) and (7.23) 'match', i.e. that

$$\mathrm{e}^{-\sqrt{\alpha_1}(t\sigma_t+\sqrt{\alpha_1})} \left(\sqrt{\alpha_1}\,(t\sigma_t + \sqrt{\alpha_1})\right)^{\alpha_1-1} \sim \alpha_1^{\alpha_1-1/2}\mathrm{e}^{-\alpha_1} \frac{\left(t\sigma_t + \sqrt{\lambda_t}\right)^{\alpha_1}}{t\sigma_t \mathrm{e}^{t\sigma_t\sqrt{\alpha_1}} \alpha_1^{\alpha_1/2}}.$$

Using that $t\sigma_t \to \infty$ as $t \uparrow t_+$ this can be checked by direct inspection.
Proof in case **(ii)**: As above write the actual tail probability as

$$P(X \geq x) = \frac{\phi(t)\mathrm{e}^{-tx}}{t\sigma_t} d_t$$

with d_t the inversion integral

$$d_t = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_t(u) \frac{du}{1 + iu/(t\sigma_t)},$$

where $\phi_t(u)$ is the characteristic function of the renormalized tilted random variable, that is, see for instance Jensen (1988),

$$\phi_t(u) = \frac{\phi(t + iu/\sigma_t)}{\phi(t)} \mathrm{e}^{-iux/\sigma_t}.$$

From (7.14) and (7.15) it follows that $\lim_{t\to\infty} x/\sigma_t = -\sqrt{\alpha}$ and $\sigma_t \sim \sqrt{\alpha}/t$ from which it is straightforward to show, that the limiting characteristic function $\psi(u)$ as $t \to \infty$ corresponds to minus a renormalized gamma distribution

$$\psi(u) = \lim_{t\to\infty} \phi_t(u) = \left(1 + iu/\sqrt{\alpha}\right)^{-\alpha} \mathrm{e}^{iu\sqrt{\alpha}}, \tag{7.24}$$

where $\alpha = \sum_{i=1}^{K} f_i/2$. Let us for now assume that we can interchange the $\lim_{t\to\infty}$ and integration, then since $\lim_{t\to\infty} t\sigma_t = \sqrt{\alpha}$,

$$\begin{aligned} \lim_{t\to\infty} d_t &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \psi(u) \frac{du}{1 + iu/\sqrt{\alpha})} \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left(1 + iu/\sqrt{\alpha}\right)^{-\alpha-1} \mathrm{e}^{iu\sqrt{\alpha}} du \\ &= \frac{\sqrt{\alpha}}{2\pi i} \int_{-i\infty}^{i\infty} (1 - v)^{-\alpha-1} \mathrm{e}^{-\alpha v} dv. \end{aligned}$$

This can be recognized as the inversion integral for a $\Gamma(\alpha+1,1)$-density evaluated at α, thus

$$\lim_{t\to\infty} d_t = \frac{\sqrt{\alpha}}{\Gamma(\alpha+1)}\alpha^{\alpha}\mathrm{e}^{-\alpha} = \frac{\alpha^{\alpha-1/2}}{\Gamma(\alpha)}\mathrm{e}^{-\alpha}. \tag{7.25}$$

As $\kappa'''(t) \to -\infty$ for $t \to \infty$ it follows from (7.7)-(7.9) that to finish the argument we must show the equivalence

$$\frac{\alpha^{\alpha-1/2}}{\Gamma(\alpha)}\mathrm{e}^{-\alpha} \sim t\sigma_t\lambda_t^{\lambda_t}\mathrm{e}^{-t\sigma_t\sqrt{\lambda_t}}A_1\left(\lambda_t, \sqrt{\lambda_t}(t\sigma_t - \sqrt{\lambda_t})\right) \tag{7.26}$$

From (7.15) and (7.16) it follows easily that $\lim_{t\to\infty}\lambda_t = \alpha$ and thus

$$\lim_{t\to\infty} = \sqrt{\lambda_t}(t\sigma_t - \lambda_t) = 0.$$

Leipniz' rule now gives that the function A_1 is continuous in t and the limit can be taken directly, using $\alpha > 1$

$$\lim_{t\to\infty} A_1\left(\lambda_t, \sqrt{\lambda_t}(t\sigma_t - \sqrt{\lambda_t})\right) = \int_0^1 \frac{v^{\alpha-1}}{\Gamma(\alpha)}dv = \frac{1}{\alpha\Gamma(\alpha)}, \tag{7.27}$$

and (7.26) follows directly.

The allowance for the crucial interchanging of integration and going to the limit will be left unproven. One way to go would be to show that the convergence of $\phi_t(u)$ is uniform for $u \in \mathbb{R}$. Another approach would be to 'copy' the argument of Jensen (1992) used for the case (i), where suitable bounds for certain decompositions of the integral are obtained and dominated convergence is applied.

The proof apply directly to the more general class of distributions treated in Jensen (1992), namely those with Laplace transform given by (7.12), where also $K = \infty$ is allowed.

7.7 Ratios of linear chi-squared combinations

In this section we consider ratio's of the form

$$\frac{X}{Y} = \frac{\sum_{i\in K_1} a_i X_i}{\sum_{i\in K_2} b_i X_i}\ ,\ X_i \sim \chi^2(f_i), i = 1,\ldots,K\ ,\ \{1,\ldots,K\} = K_1\bigcup K_2,$$

and $X_i, i = 1,\ldots,K$ independent. We will discuss the approximation of the tail probability $P(X/Y > z)$, and subsequently in the following section how this can be utilized for generalized F-testing. Daniels (1954) derived at an expansion for the density of a ratio of two independent random variables. Instead we will base the approximation on the results of the previous sections. Also note that $K_1 \bigcap K_2$ may be non-empty, in which case the numerator and denominator are not independent.

If the denominator Y is a positive random variable the tail probability can be expressed directly in the form treated in the previous sections,

$$\begin{aligned} P\left(X/Y > z\right) &= P\left(X - zY > 0\right) \\ &= P\left(\sum_{i \in K} c_i X_i > 0\right), \end{aligned} \qquad (7.28)$$

with $c_i = a_i, i \in K_1 \setminus K_2$, $b_i = -zb_i, i \in K_2 \setminus K_1$ and $c_i = a_i - zb_i, i \in K_1 \bigcap K_2$.

If instead the numerator is a positive random variable, the expression for the tail probability depends on the sign of z,

$$P\left(X/Y > z\right) = \begin{cases} P\left(X - zY > 0\right) - 1 + P(Y > 0), & z > 0 \\ 1 - P\left(X - zY > 0\right) + P(Y > 0), & z < 0 \end{cases}, \qquad (7.29)$$

where both entering tail probabilities can be approximated. These expressions follow from simple set considerations: For $z > 0$ the we have the identities apart from null sets

$$\begin{aligned} \{X > zY\} &= \{X > zY, Y > 0\} \bigcup \{Y \leq 0\} \\ \{X/Y > z\} &= \{X > zY, Y > 0\}, \end{aligned} \qquad (7.30)$$

and for $z < 0$

$$\begin{aligned} \{X < zY\} &= \{X < zY, Y > 0\} \\ \{X/Y > z\} &= \{X < zY, Y < 0\} \bigcup \{Y > 0\}. \end{aligned} \qquad (7.31)$$

The probability $P(Y \leq 0)$ will often be very small, and this can be used in the general case to get bounds for the probability of interest. For if both X and Y can take negative as well as positive values it is not possible to get an explicit expression for the tail probability as above. Assume for now that $z > 0$. The (X, Y)-plane is partitioned into six disjoint sets, see Figure 7.1. The event of interest can be written as

$$\{X/Y > z\} = B \bigcup E.$$

Using the notation $P_{abc} = P(A \bigcup B \bigcup C)$ it is possible to saddlepoint approximate the following probabilities

$$\begin{aligned} P_{abc} &= P(X > 0) \\ P_{abf} &= P(Y > 0) \\ P_{bcd} &= P(X - zY > 0) \end{aligned}$$

and thus also the complementary probabilities P_{def}, P_{cde} and P_{aef}. Since $P_{abf} - P_{aef} = P(B) - P(E)$ we can write up bounds for the tail probability:

$$P\left(X/Y > z\right) \in \begin{cases} \left[P_{abf} - P_{aef}, P_{abf} - P_{aef} + 2\min\left\{P_{def}, P_{cde}, P_{aef}\right\}\right], & \text{if } P_{abf} > P_{aef} \\ \left[P_{aef} - P_{abf}, P_{aef} - P_{abf} + 2\min\left\{P_{abc}, P_{abf}, P_{bcd}\right\}\right], & \text{if } P_{abf} < P_{aef} \end{cases}.$$

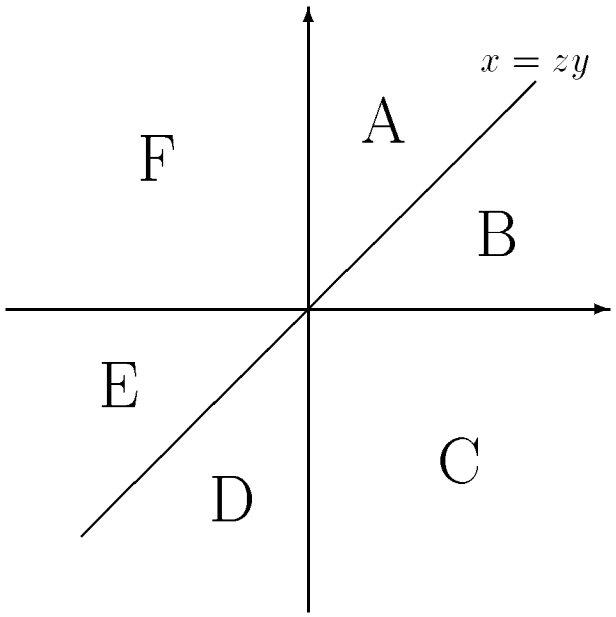

Figure 7.1: The partitioning of (X, Y)-space for $z > 0$

All the entering probabilities can now be saddlepoint approximated. If just one of the three half-planes has negligible probability mass this will give narrow bounds, and in any case this procedure will provide the information about the size of the bounds.

7.8 Non-standard F-tests

By non-standard F-tests we understand ratios of the form

$$F = \frac{\sum_{i\in K_1} \tilde{a}_i \tilde{X}_i}{\sum_{i\in K_2} \tilde{b}_i \tilde{X}_i} = \frac{\sum_{i\in K_1} a_i X_i}{\sum_{i\in K_2} b_i X_i} \tag{7.32}$$

where $\tilde{X}_i$, $i \in K_1 \cup K_2$ are independent and

$$\tilde{X}_i \sim \lambda_i \chi^2(f_i)/f_i \ , \ a_i = \frac{\tilde{a}_i \lambda_i}{f_i}, i \in K_1 \ , \ b_i = \frac{\tilde{b}_i \lambda_i}{f_i}, i \in K_2.$$

Tests of this form may occur in mixed model analyses of variance with non-balanced data or when the collection of random effects are not closed under formation of infima, see Section 5.13.5 for a design discussion. As also exemplified in Section 5.13.4 the latter is relevant in the typical 3-way sensory experiment with judge- and replication

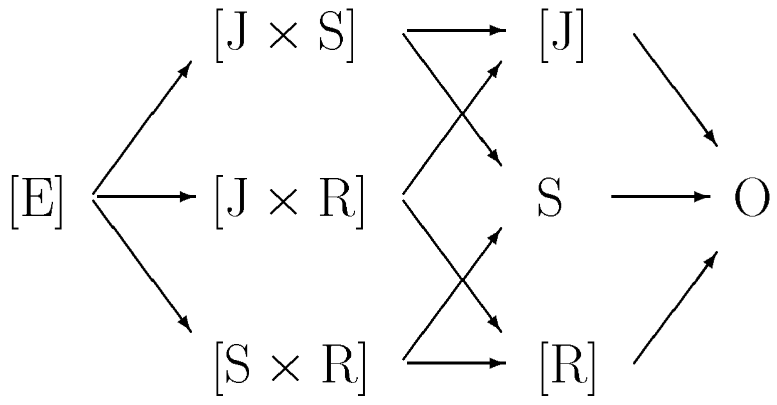

Figure 7.2: Factor structure diagram, see Chapter 5, for the model used for exemplifying non-standard F-testing

effects random. For simplicity we will throughout this section assume such a setup: I judges (J) have assessed P samples (S) in K replicates (R), and as model for a univariate sensory response we use the mentioned mixed model as illustrated by Figure 7.2

The issue of interest is to test for sample effect. A classical approach is to construct a test statistic based on the expected mean squares with approximately an F-distribution. Assuming no missing data the relevant expected mean squares in the present setup are, using the same notation as in Chapter 5:

$$\begin{aligned} \lambda_E = \mathrm{E}\,(\mathrm{MS}_E) &= \sigma_E^2 \\ \lambda_{J\times S} = \mathrm{E}\,(\mathrm{MS}_{J\times S}) &= \sigma_E^2 + K\sigma_{J\times S}^2 \\ \lambda_{S\times R} = \mathrm{E}\,(\mathrm{MS}_{S\times R}) &= \sigma_E^2 + P\sigma_{S\times R}^2 \\ \lambda_S = \mathrm{E}\,(\mathrm{MS}_S) &= \sigma_E^2 + K\sigma_{J\times S}^2 + P\sigma_{S\times R}^2 + Q_S \end{aligned}$$

In Lea *et al.* (1991) an approach leading to the test-statistic

$$F_1 = \frac{\mathrm{MS}_S}{\mathrm{MS}_{J\times S} + \mathrm{MS}_{S\times R} - \mathrm{MS}_E} \tag{7.33}$$

is suggested. This approach will break down whenever the denominator comes up negative , and an alternative is, see for instance Cochran and Cox (1957), to use F_2 given by

$$F_2 = \frac{\mathrm{MS}_S + \mathrm{MS}_E}{\mathrm{MS}_{J\times S} + \mathrm{MS}_{S\times R}}. \tag{7.34}$$

In both cases the numerator and denominator are independent and have identical expected values under the null hypothesis. Their approximate F-distributions can be understood in the light of the following observation due to Satterthwaite (1946): A linear combination of independent mean squares, or more general, say, a random variable as the numerator in (7.32), has a distribution that can be well approximated by a χ^2-distribution with the 'correct' variance, which for the numerator in (7.32) amounts to a $\chi^2(f)$-distribution with

$$f = \frac{\left[\sum_{i \in K_1} \tilde{a}_i \lambda_i\right]^2}{\sum_{i \in K_1} \frac{(\tilde{a}_i \lambda_i)^2}{f_i}}. \tag{7.35}$$

As the parameters in practice will be unknown, an estimated version of this will have to be used,

$$\hat{f} = \frac{\left[\sum_{i \in K_1} \tilde{a}_i \tilde{X}_i\right]^2}{\sum_{i \in K_1} \frac{\left(\tilde{a}_i \tilde{X}_i\right)^2}{f_i}} \tag{7.36}$$

or the smallest integer larger than $\hat{f}$.

It is clear that the distribution under the null hypothesis of F_1 and F_2 can be expressed as in (7.28) and (7.29) respectively, and thus saddlepoint approximated by (7.7). It is also clear that the saddlepoint approximation will depend on the parameters $\lambda = (\lambda_1, \ldots, \lambda_4) = (\lambda_E, \lambda_S, \lambda_{S\times R}, \lambda_{J\times S})$. As for the Satterthwaite approach estimated parameters will have to be used. The tempting straightforward choice for λ as the observations $(\mathrm{MS}_E, \mathrm{MS}_S, \mathrm{MS}_{S\times R}, \mathrm{MS}_{J\times S})$ is not meaningful, as we search for the distribution under the null hypothesis $(\lambda_S = \lambda_{S\times R} + \lambda_{J\times S} - \lambda_E)$ and the observed value of λ will fall outside the null hypothesis with probability 1, and possibly far away from it. The effect of doing this would be to move the distribution close to the observed value and then asking whether this observed value is extreme, which of course it would not be.

Instead we suggest to take the maximum likelihood estimates of λ under the null hypothesis. The Gamma log-likelihood for λ is proportional to a simple function,

$$l(\lambda) \propto -\sum_{i=1}^{4} \left\{ \frac{f_i}{2} \log \lambda_i + \frac{f_i}{2\lambda_i} x_i \right\}, \tag{7.37}$$

where $(x_1, \ldots, x_4) = (\mathrm{MS}_E, \mathrm{MS}_S, \mathrm{MS}_{S\times R}, \mathrm{MS}_{J\times S})$, and the ML-estimates are easily found by numerical maximization of (7.37) under the constraint that $\lambda_S = \lambda_{S\times R} + \lambda_{J\times S} - \lambda_E$.

The performance of this approximation was investigated in two situations by simulation and compared to the Satterthwaite methods. The two situations represent an extreme situation (A), corresponding to an unrealistic small experiment with 2 judges, 2 products and 2 replicates and (B) a typical experiment with 8 judges 4 products and 2 replicates:

(A)

$$I = P = K = 2,\ \sigma_E^2 = \sigma_{S\times R}^2 = \sigma_{J\times S}^2 = 1,$$

and hence

$$\lambda = (1, 5, 3, 3),\ f = (f_1, \ldots, f_4) = (1, 1, 1, 1)$$

(B)

$$I = 8, P = 4, K = 2,\ \sigma_E^2 = 1, \sigma_{S\times R}^2 = 2, \sigma_{J\times S}^2 = 4,$$

and hence

$$\lambda = (1, 25, 17, 9),\ f = (21, 3, 3, 21)$$

For the use of (7.7) in practice some care must be taken due to numerical instability, that occurs when the third cumulant $\kappa'''(t)$ gets 'close' to zero. What close is depends on the size of the second cumulant. In these cases the Lugannani-Rice formula (7.6) was used. The Lugannani-Rice formula also 'breaks down' numerically, when the observation is very close to the mean of the distribution in question. An alternative formula exists, see Daniels (1987), but for simplicity we set the probability to 0.5 in these cases. Rather arbitrary choices were made to ensure the numerical stability, and if we let $\hat{p}(x)$ denote a calculated approximation, it was chosen as, again with $t = t_x$,

$$\hat{p}(x) = \begin{cases} (7.7), & \text{if } \begin{array}{l} \lambda_t < 10^6, |\kappa'''(t)| > 0.1, \\ \kappa(t) - tx + \operatorname{sgn}(\kappa'''(t)) t\sigma_t\sqrt{\lambda_t} < 200 \end{array} \\ \tilde{p}(x), & \text{otherwise,} \end{cases} \tag{7.38}$$

where $\tilde{p}(x)$ is the 'continued' Lugannani-Rice formula

$$\tilde{p}(x) = \begin{cases} (7.6), & \text{if } |t\sigma_t| > 10^{-7}, \sqrt{2(tx - \kappa(t))} > 10^{-7} \\ 1/2 & \text{otherwise.} \end{cases}$$

Another important point when applying (7.7) is that the formulas are only right-tail formulas, as opposed to the Lugannani-Rice formula. This means that the point of interest, x, should always be seen in the light of the mean of the original distribution, and the choice of approximation of X or $-X$ taken accordingly. In fact computations have indicated that (7.7) becomes not only inaccurate but directly useless when the tail-probability gets close to 1.

The numerical investigation fell naturally into two parts: a basic distributional study based on true parameter values and a testing problem study. In the latter the actual F-testing performance was studied for 'real' simulated data. In the former the accuracy of the of the true saddlepoint approximation to the distributions of F_1 and F_2 were investigated. For F_1 the distribution/testing distinction is important due to the positive negativity probability and this problem is given special attention.

Throughout the section the simulations were done by simulating Gamma-distributed random variables with the Gamma-generator 'pgamma' in Splus. For the first part of the study a sample in each situation of 100000 were taken to estimate the true distribution of F_1 and F_2. Thus for the comparisons in the following the uncertainty due to this approach should be kept in mind. Figure 7.3 shows the results for the situation (A). The calculated degrees of freedom based on (7.35) were $(1, 1.32)$ and $(1.38, 2)$, which means that for the Satterthwaite approximations were used the $F(1,2)$ and $F(2,2)$-distributions respectively. The observed relative frequency of negative F_1's were 0.13279 and the saddlepoint approximated $P_0 = P(\text{MS}_{J\times S} + \text{MS}_{J\times R} - \text{MS}_E < 0)$ were $\hat{P}_0 = 0.115896$. From Figure 7.3(a) we see that for the 5% quantile the saddlepoint approximation of F_1 is a little better than the Satterthwaite method but further out in the tail both approximations become poor. The F_2-approximations seem both to work reasonable well.

For situation (B) we see in Figure 7.4 that the saddlepoint approximation is vastly superior in the tails. This is probably an effect of the quite different degrees of freedom for the entering χ^2-components. The approximating F-distribution is an 'average' F-distribution, whereas the tail behaviour really gets dominated by one of the components. The calculated degrees of freedom based on (7.35) were $(3, 6.24)$ and $(3.24, 6.75)$, which means that for the Satterthwaite approximations were used the $F(3,7)$ and $F(4,7)$-distributions respectively. The reason for the better behaviour of the saddlepoint approximation of F_1 in situation (B) as opposed to situation (A) should be sought in the negativity probability. In situation (B) there were observed no negative F_1-values out of 100000 samples, and the saddlepoint approximated probability was $\hat{p}_{(0)} = 4.75 \cdot 10^{-9}$.

The accuracy in the estimation of P_0 sets a limit for the accuracy of the tail probability approximation. Related to this is a consideration of what kind of limit is actually taken when we go to the tails of F_1 and F_2 respectively. It is clear that we have that

$$P(\text{MS}_S - x\text{MS}_{S\times R} - x\text{MS}_{J\times S} + x\text{MS}_E) \to \begin{cases} P_0 & \text{for } x \to \infty \\ 1 - P_0 & \text{for } x \to -\infty \end{cases} \tag{7.39}$$

and

$$P(\text{MS}_S + \text{MS}_E - x\text{MS}_{S\times R} - x\text{MS}_{J\times S}) \to \begin{cases} 0 & \text{for } x \to \infty \\ 1 & \text{for } x \to -\infty \end{cases} \tag{7.40}$$

But going to the limit in the approximations needs some additional considerations, and cannot be described directly as a tail of a distribution as the true distribution in

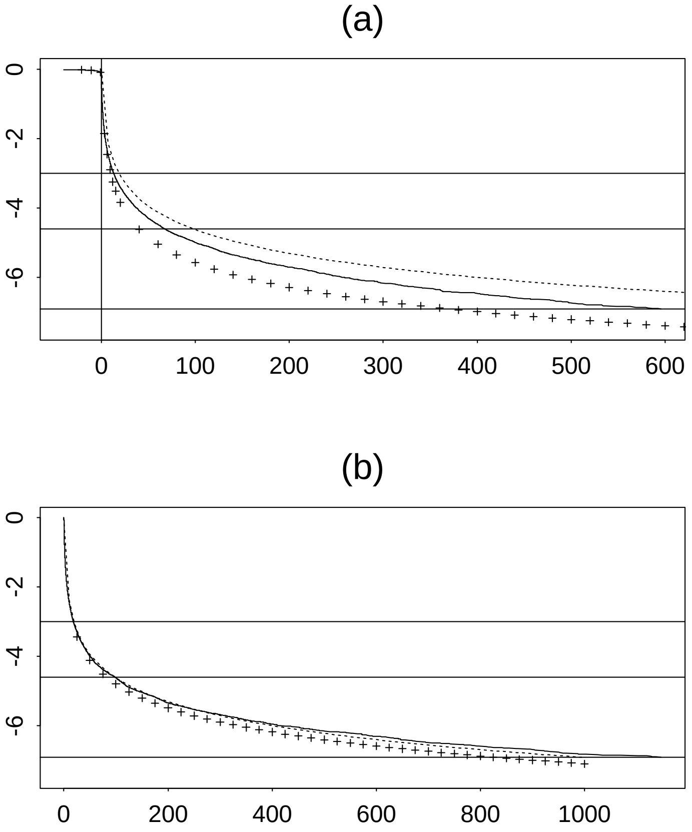

Figure 7.3: The logarithm of the tail probability in situation (A) for F_1 (a) and F_2 (b). The solid curve is the empirical log-tail-frequencies based on a random sample of 100000, the dotted curve is the 'true' Satterthwaite approximation and the marked points are the 'true' saddlepoint approximation. The three horizontal lines indicate the 5%, 1% and 0.1% quantiles.

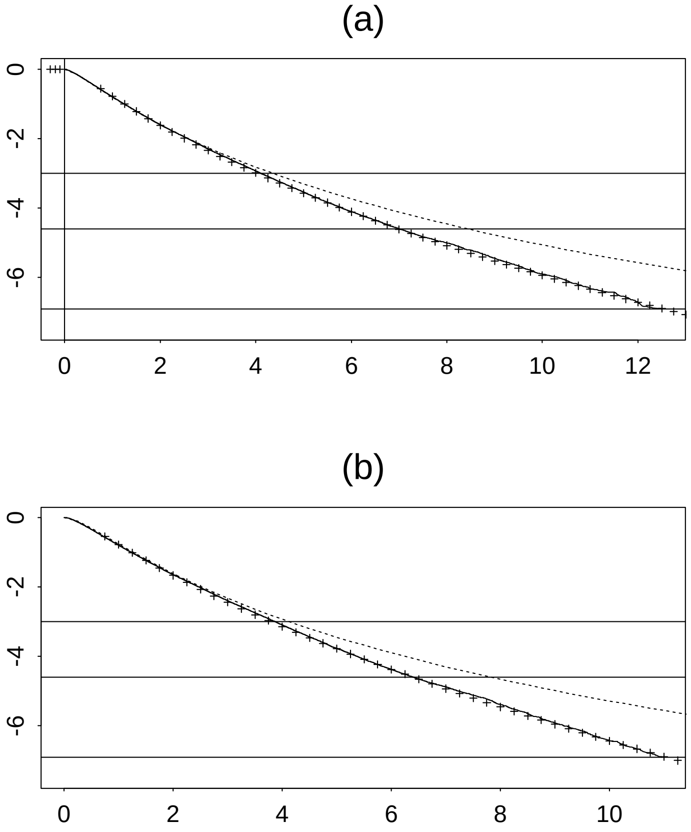

Figure 7.4: The logarithm of the tail probability in situation (B) for F_1 (a) and F_2 (b). The solid curve is the empirical log-tail-frequencies based on a random sample of 100000, the dotted curve is the 'true' Satterthwaite approximation and the marked points are the 'true' saddlepoint approximation. The three horizontal lines indicates the 5%, 1% and 0.1% quantiles.

question changes with x not only in position but also in shape. This is reflected in the set with finite Laplace transform, which for the F_1-case is

$$T_1(x) = \{t \in \mathrm{I\!R} \mid \phi_{x1}(t) < \infty\}$$

and for the F_2-case

$$T_2(x) = \{t \in \mathrm{I\!R} \mid \phi_{x2}(t) < \infty\}$$

where ϕ_{x1} is the Laplace transform of $\mathrm{MS}_S - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S} + x\mathrm{MS}_E$ and ϕ_{x2} is the Laplace transform of $\mathrm{MS}_S + \mathrm{MS}_E - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S}$. The upper endpoint of $T_2(x)$ does not depend on x, and the lower goes to 0 from the left, so for the F_2-case the right tail behaviour can in fact in the limit be described as a right tail behaviour of a single distribution, and the limiting exactness result applies.

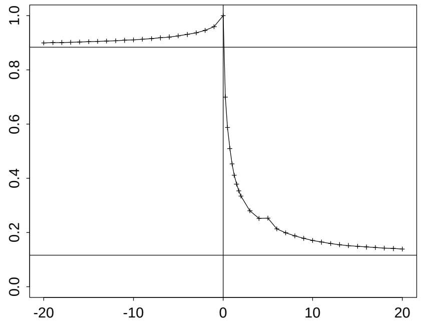

Figure 7.5: Values of the saddlepoint approximation of $P(\mathrm{MS}_S - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S} + x\mathrm{MS}_E > 0)$ for different values of x. The horizontal lines indicate $\hat{P}_0$ and $1-\hat{P}_0$. The 'discontinuity' in $x = 5$ is a numerical artifact due to the fact that $\kappa'''(5) = 0$, *cf.* (7.38)

For the F_1-case it is not so obvious, since the set $T_1(x)$ collapses in 0 as $x \to \infty$. Figure 7.5 indicates the asymptotic behaviour of the saddlepoint approximation of $P(\mathrm{MS}_S - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S} + x\mathrm{MS}_E > 0)$. We note that Figure 7.5 indicates that the saddlepoint approximation $\hat{P}(x)$ of $P(\mathrm{MS}_S - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S} + x\mathrm{MS}_E > 0)$ has the property

$$\hat{P}(x) \to \begin{cases} \hat{P}_0 & \text{for } x \to \infty \\ 1 - \hat{P}_0 & \text{for } x \to -\infty \end{cases}, \tag{7.41}$$

which was confirmed by computations for numerical very large xs.

We now turn to the actual testing problem. Whenever F_1 comes up negative it is not possible to define in a reasonable way the event of an outcome being more extreme. Conditioning on a positive outcome will lead to the test probability

$$\begin{aligned} & P\left(F_1 > x \mid \mathrm{MS}_{S\times R} + \mathrm{MS}_{J\times S} - \mathrm{MS}_E > 0\right) \\ = & \ \frac{P\left(\mathrm{MS}_S - x\mathrm{MS}_{S\times R} - x\mathrm{MS}_{J\times S} + x\mathrm{MS}_E > 0\right) - P_0}{1 - P_0}, \end{aligned} \qquad (7.42)$$

that can be saddlepoint approximated. Below we have made a comparison in 1500 random samples of the 'real' versions of Satterthwaite and saddlepoint approximations, that is, the observed mean squares are inserted in the Satterthwaite approximation and the ML-estimates under the null hypothesis are inserted in the saddlepoint approximation. For reference we also included the versions based on the true parameters. The Figures 7.6-7.9 show that these two methods are roughly working equally well in both situations. Based on 1500 samples, though, we do not have sufficient information about the far out tail behaviour, where as earlier indicated at least the 'true' saddlepoint method were superior to the 'true' Satterthwaite method. We will leave this point open for future investigation.

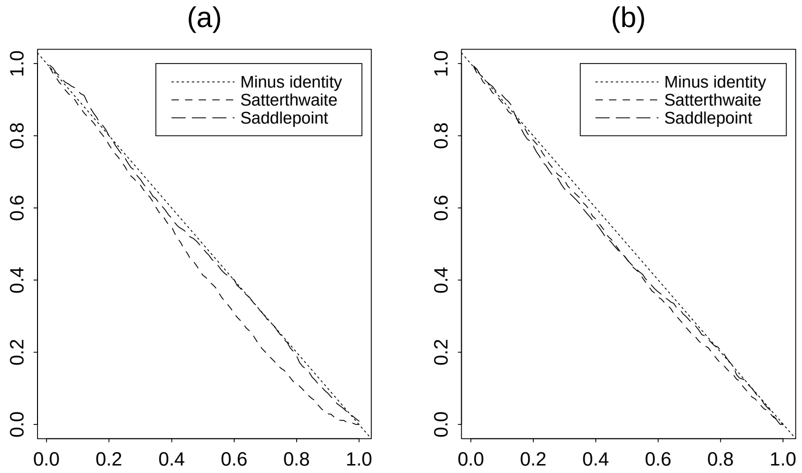

Figure 7.6: Quantiles for Satterthwaite and saddlepoint approximations of F_1-test in situation (A)

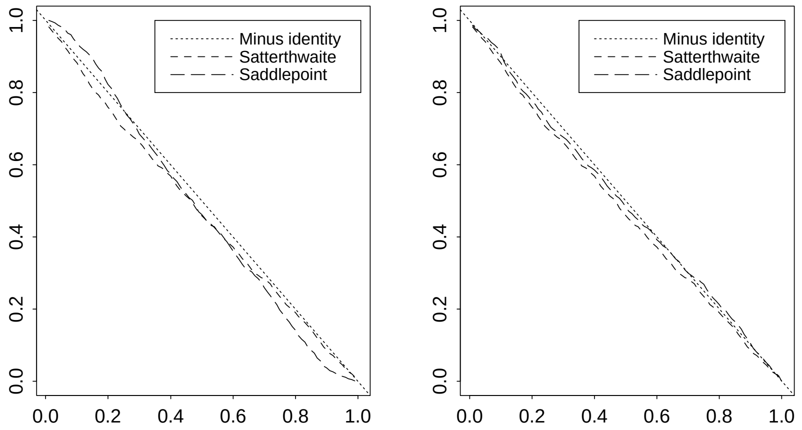

Figure 7.7: Quantiles for Satterthwaite and saddlepoint approximations of F_2-test in situation (A)

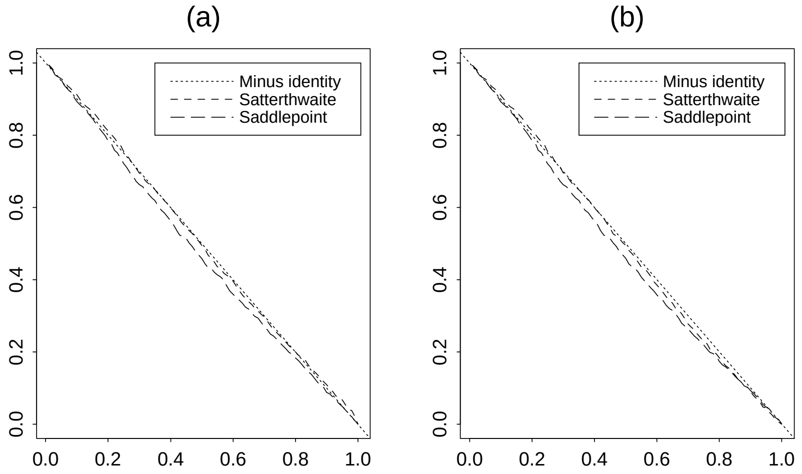

Figure 7.8: Quantiles for Satterthwaite and saddlepoint approximations of F_1-test in situation (B)

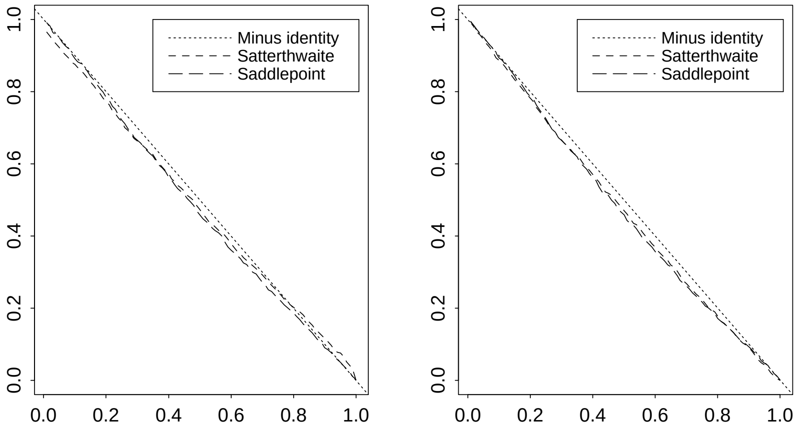

Figure 7.9: Quantiles for Satterthwaite and saddlepoint approximations of F_2-test in situation (B)

7.9 Remarks

We have shown that the use of Gamma saddlepoint approximation techniques is a potentially very accurate tool in non-standard F-testing. It should be noted, however, that we do not have limiting exactness results for the actual applied approximations, as we insert parameter estimates that themselves have error. We have also illustrated that the saddlepoint approach does not overcome the problem of negativity, as this is a more conceptual problem. The distributional problem of linear combinations of χ^2-distributed random variables and ratios of such is closely related to efforts on constructing confidence bands for the population versions of the same objects, see Burdick and Graybill (1992) and references therein. Many of the results on this subject are restricted to certain situations, for example positive coefficients $a_1, \ldots, a_K$. The saddlepoint approximations might offer a general applicable method for construction of such confidence bands, but whether this is possible is an open point.

Chapter 8

Concluding remarks

In this conclusion an exhaustive review of the thesis will not be given. The reader is referred to the Summary for a brief listing of contributions of the work. The scope of the various methods and results are discussed throughout the thesis. It may, however, be useful to consider the initiatives of the work in relation to the following grouping of potential beneficiary fields:

1. Promotion of known statistical methods rarely used in sensometrics.
2. Sensometric developments
   (a) Methodological
   (b) Theoretical
3. Statistical developments
   (a) Methodological
   (b) Theoretical

Although non-exhaustive, overlapping and to some extent subjective, the following grouping of contributions, on the next page, into these five categories gives an impression of the overall contribution of this thesis. Apart from this it has been a general objective to gain and communicate insight in basic sensometric and statistical issues relevant in general and in particular for sensory analysis experiments.

Conclusively, an important contribution of the thesis is its role as starting point for future research. Considering the five categories above, effort is needed in future to put any method from 2(a) and 3(a) into category 1. Of particular future importance, I believe, is the generalized linear models, as well in their classical formulation as in the still developing mixed model versions. The latter is certainly also an area still open for innovations qualifying for categorization into any of the five groups.

Contribution list:

1

1. Continuum regression
2. Generalized linear models
3. Two-dimensional covariance components models
4. Three-way factor methods

2(a)

1. Panel size prediction ability interpretation
2. Assessor model — fixed and mixed
3. Threshold determination via mixed GLIM's

2(b)

1. Relation between range estimates in assessor model and Procrustes-like stretching and shrinking values
2. Relation between three-way factor methods and various other multivariate techniques

3(a)

1. Two-step cross validation principle
2. The general assessor model (with arbitrary linear treatment structure) — fixed and random
3. A marginal algorithm for mixed GLIM's in a dose-response setup with repeated measurements
4. Application of saddlepoint approximations for non-standard F-testing.

3(b)

1. A convergence result for alternating algorithmic optimization of likelihood functions with non-unique maximum.
2. Likelihood ratio test for overall independence across strata between sets of variates in a multivariate mixed analysis of variance model with a Gamma based saddlepoint approximation.
3. Relative limiting exactness for a Gamma based saddlepoint approximation of the tail probability for Gamma convolutions.

To comply with the continual objection to the consideration of sensory descriptive data as interval scale data, the GLIM approach seems feasible. The study of experimental designs based on these models could be relevant in order to adopt to the GLIM frame, for instance, linear model designs taking order and carry-over effects into account. The problem of existence of maximum likelihood estimates in a binary regression model with non-zero baseline probabilities was pointed out in Brockhoff and Müller (1994).

The statistical and sensometric methodological scopes of the assessor model approach likewise need further research, especially the extension with a general linear model expressed for the products involved. This may be seen as a general extension of the linear normal models. And the random version of this is only given a very brief treatment in this Thesis and obviously needs additional work as well theoretical as methodological. Small sample improvements of approximating test statistic distributions, maybe based on saddlepoint methods, are also of future relevance.

For the saddlepoint approximation methods a number of theoretical problems was raised and left unsolved, some of which are: The limiting exactness in the tails of the applied Gamma approximation for $n = 1$ for the Box class of test statistics, an asymptotic result as the degrees of freedom tend to infinity for the applied Gamma approximation of Gamma convolution tail probabilities, the use of saddlepoint methods for construction of confidence bands in general and in particular for the Gamma convolutions.

All in all, the Thesis has raised at least as many questions as was answered. It reflects quite well the learning process I have been going through working my way from an unexperienced statistician to a little more experienced sensometrician. And it may still serve, I believe, as a general introduction for the statistically interested reader to the field of sensometrics, with a lot of unsolved problems and non-optimal solutions to work with throughout.

Complete Reference List

Abramowitz, M. and Stegun, I.A. (eds) (1964). *Handbook of Mathematical Functions.* New York: Dover.

Amerine, A.M., Pangborn, R.S. and Roessler, E.B. (1965). *Principles of Sensory Evaluation of Food.* Academic Press, New York and London.

Anderson, T.W. (1958). *An Introduction to Multivariate Statistics.* Wiley, New York.

Anderson, D.A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *J. Royal Statist. Soc.* **B47**, 203–210.

Arnold, G.M. and Williams, A.A. (1986). The use of Generalised Procrustes techniques in sensory analysis. In: *Sensory Analysis of Foods.* Ed. J.R. Piggott, Elsevier Applied Science, London.

Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families.* Wiley, Chichester.

Barndorff-Nielsen, O.E. (1983). On a formula for the conditional distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Bartlett, M. S. (1937). Properties of suffiency and statistical test. *Proceedings of the Royal Society London Series A*, **160**, 268–282.

Bernado, J.M. (1976). Algorithm AS 103. Psi (digamma) function. *Appl. Statist.* **25**, 315–317.

Bonnans, S. (1991). *Effect of sample composition and salivary flow on temporal perception of sweetness.* Masters Thesis, University of California, Davis.

Borton, C.B. (1990). Weinschmecker. A computerized data collection system for sensory testing. Unpublished.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Brockhoff, P.M. (1993). Sensoriske tærskelværdier — et metodestudium. *Praktisk-skriftlig opgave i støttefaget Sensorik.* Center of Food Research, Royal Veterinary and Agricultural University, Copenhagen, Denmark.

Brockhoff, P.M. (1994). *Statistical Analysis of Sensory Data.* Ph.D.-thesis, Dept. of Mathematics and Physics and Center of Food Research, Royal Veterinary and Agricultural University, Copenhagen, Denmark.

Brockhoff, P.M. and Guggenbühl, B. (1995). Two-dimensional covariance component models applied to sensory data. To be submitted to: *Journal of Sensory Studies.*

Brockhoff, P.M., Hirst D. and Næs, T. (1995). Three-way factor methods in sensory analysis. In: *Multivariate statistics for Sensory Data.* Ed. T. Næs and E. Risvik, Elsevier Science Publishers, Amsterdam, The Netherlands.

Brockhoff, P.M. and Müller, H.G. (1994). Random effect threshold models for dose-response relations with repeated measurements. Submitted to: *J. Royal Stat. Soc. Ser. B.*

Brockhoff, P.M. and Skovgaard, I.M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference* **5**, 215–224.

Brockhoff, P.M., Skovgaard, I.M., Poll, L. and Hansen K. (1993). A comparison of methods for linear prediction of apple flavour from gas chromatographic measurements. *Food Quality and Preference* **4**, 215–222.

Burdick, R.K. and Graybill (1992). *Confidence intervals on variance components.* Marcel Dekker, New York.

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., 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.

Cochran, W.G. and Cox, G.M. (1957). *Experimental Designs.* Wiley, New York.

Collet, D. (1991). *Modelling Binary Data.* Chapman & Hall, London.

Crowder, M. J. (1980). Proportional linear models. *Applied Statistics* **29**(3), 299–303.

Cruciani, G. , Baroni, M. , Clementi, S. , Costantino, G. , Riganelli, D. and Skagerberg, B. (1992). Predictive ability of regression models. Part I: Standard deviations of prediction errors (SDEP). *J. Chemometrics* **6**, 335–346.

Daniels, H.E. (1954). Saddlepoint approximations in statistics. *Ann. Math. Statist.* **25**, 631–649.

Daniels, H.E. (1987). Tail probability approximations. *Int. Statist. Rev.* **55**(1), 37–48.

David, H.A. (1969). *The Method of Paired Comparisons.* Charles Griffin, London.

Dembster, A.P., Rubin, D.B. and Tsutakawa, R.K. (1981). Estimation in covariance components models. *J. Am. Stat. Ass.* **76**(374), 341–353.

Drake, B. (1975). A fortran program 'SIGMPLOT' for fitting sigmoid curves to threshold data and for plotting the results. *Chem. Sens. Flav.* **1**, 519–533.

Elashoff, J.D. (1981). Repeated-measures bioassay with correlated measures and heterogeneous variances: A Monte Carlo study. *Biometrics* **37**, 475–482.

Ennis, D.M. (1990). Relative power of difference testing methods in sensory evaluation. *Food Technology* **44**, 114–117.

Ennis, D.M. (1993). The power of sensory discrimination methods. *J. Sens. Stud.* **8**, 353–370.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on generalized Linear Models.* Springer-Verlag, New York, Inc.

Fechner, G.Th. (1860). *Elemente der Psychophysik.* Brietkopf and Härtel, Leipzig. (English translation by H.E. Adler, Elements of Psychophysics. Holt, Rinehart and Winston, Inc., New York, 1966).

Finney, D.J. (1971). *Probit Analysis.* Cambridge University Press, Cambridge.

Flury, B.F. (1988). *Common Principal Components & related Multivariate Models.* Wiley, New York.

Frijters, E.R. (1988). Sensory difference testing and the measurement of sensory discriminability. In: *Sensory Analysis of Foods.* Ed. J.R. Piggott, Elsevier Applied Science, London.

Gay, C. and Mead, R. (1992). A statistical appraisal of the problem of sensory measurement. *J. of Sens. Stud.* **7**, 205–228.

Goldstein, H. (1991). Non-linear multilevel models, with an application to discrete response data. *Biometrika* **58**, 45–51.

Gower, J.C. (1975). Generalized procrustes analysis. *Psychometrika,* **40**(1),33–51.

Ghosh, J.K. (1978). *Higher Order Asymptotics.* NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, California.

Guadagni, D.G., Buttery, R.G. and Okano, S. (1963). Odor thresholds of some organic compounds associated with food flavours. *J. Sci. Fd. Agric.* **14**, 761–765.

Guadagni, D.G., Maier, V.P. and Turnbaugh, J.G. (1973). Effect of soe citrus juice constituents on taste thresholds for limonin and naringin bitterness. *J. Sci. Fd. Agric.* **24**, 1277–1288.

Hansen, K. , Poll, L. , Olsen, C.E. , and Lewis, M.J. (1992). The influence of oxygen concentration in storage asmospheres on the post storage of 'Jonagold' apples. *Lebensm.-Wiss. u.-Technol.* **25**, 457–461.

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.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Am. Stat. Ass.* **72**, 320–340.

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.

Hirst, D. and Næs, T. (1994). A graphical technique for assessing differences among a set of ranking. *J. Chemometrics* (In press).

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.

Hoerl, A.E. and Kennard, W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.

Im, S. and Gianola, D. (1988). Mixed models for binary data with an application to lamb mortality data. *Applied Statistics* **37**, 196–204.

Jensen, J.L. (1988). Uniform saddlepoint approximations. *Adv. Appl. Prob.* **3**, 622–634.

Jensen, J.L. (1991). A large deviation-type approximation for the "Box class" of likelihood ratio criteria. *J. Am. Stat. Ass.* **86**(414), 437–440.

Jensen, J.L. (1992). A note on a conjecture of H.E. Daniels. *Rev. Bras. Prob. Estatist.* **6**, 85–95.

Jensen, J.L. (1995). *Saddlepoint Approximations.* Clarendon Press, Oxford.

Jensen, S.T., Johansen, S. and Lauritzen, S.L. (1991). Globally convergent algorithms for maximizing a likelihood function. *Biometrika* **78**(4), 867–877.

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* **43**(4), 443–477.

Kjølstad, L., Isaksson, T. and Rosenfeld, H.J. (1990). Prediction of sensory quality by near infrared reflectance analysis of frozen and freeze dried green peas (Pisum sativum). *J. Sci. Food Agric.* **51**, 247–260.

Kloot, W.A. van der and Kroonenberg, P.M. (1985). External analysis with three-mode principal component models. *Psychometrika* **50**(4), 479–494.

Kowalski, K.G. (1990). On the predictive performance of biased regression methods and multiple linear regression. *Chemometrics and Intelligent Laboratory Systems* **9**, 177–184.

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.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis: A User's Perspective.* Clarendon Press, Oxford.

Larsen, M. and Poll, L. (1992). Odour thresholds of some important aroma compounds in strawberries, *Z. Lebensm. Unters. Forsch.* **195**, 120–123.

Lea, P. (1988). Triangeltester, partester, duo-trio-tester. Statistiske betragtninger, *NINF-Rapport*, **1**.

Lea, P., Næs, T. and Rødbotten, M. (1991). *Variansanalyse for sensoriske Data.* Ås Tryk, Ås.

Lehmann, E.L. (1983). *Theory of Point estimation.* John Wiley & Sons, New York.

Leurgans, S. and Ross, T. (1992). Multilinear models: Applications in spectroscopy. *Statistical Science* **7**(3), 289–319.

Litchfield, J.T. and Wilcoxon, F. (1949). A simplified method of evaluating dose-effect experiments. *J. Pharm. Experim. Therap.* **96**, 99–113.

Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables, *Adv. in Appl. Probab.* **12**, 475–495.

Lundahl, D. S. and McDaniel, M. R. (1988). The panelist effect — fixed or random? *Journal of Sensory Studies* **3**, 113–121.

MacRae, S. (1987). The psychological basis of sensory perception. *Chemistry and Industry*, 5 January, 7–12.

Mandel, J. (1961). Non-additivity in two-way analysis of variance. *J. of Am. Stat. Ass.* **56**, 878–888.

Mandel, J. (1969). The partitioning of interaction in analysis of variance. *Journal of Research—National Bureau of Standards, Section B, Mathematical Sciences* **73B**(4), 309–328.

Mandel, J. (1971). A new analysis of variance model for non-additive data. *Technometrics* **13**(1), 1–18.

Martens, M. and Martens H. (1986). Near-infrared reflectance determination of sensory quality of peas. *Applied Spectroscopy* **40**(3), 303–310.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

McEwan, J.A. and Schlich, P. (1991). Correspondence analysis in sensory evaluations. *Food Quality and Preference* **3**, 23–36.

Meilgaard, M.C. (1975). Flavor chemistry of beer part I: Flavor interaction between principal volatiles. *MBAA Technical Quarterly* **12**(2), 107–117.

Meilgaard, M.C. (1987). *Sensory Evaluation Techniques.* CRC Press, Inc., Boca Raton, Florida.

Morgan, B.J.T. (1992). *Analysis of Quantal Response Data.* Chapman and Hall, London.

Moskowitz, H.R. (1993). Sensory analysis procedures and viewpoints: Intellectual history, current debates, future outlooks. *J. Sens. Stud.* **8**, 241–256.

Mulders, E.J. (1973). The odour of white bread. *Z. Lebensm. Unters. Forsch.* **151**, 310–317.

Møller, J. (1984). Flerdimensionale kovarianskomponentmodeller. *Statistisk interna* nr. 38, Afdeling for Teoretisk Statistik, Aarhus Universitet.

Norwich, K.H. (1992). In: Ed. H.T. Lawless and B.P. Klein, *Sensory Science Theory and Applications an Foods.* Marcel Dekker, Inc., New York.

Næs, T. (1990). Handling individual differences between assessors in sensory profiling. *Food Quality and Preference* **2**, 187–199.

Næs, T. and Solheim, R. (1991). Detection and interpretation of variation within and between assessors in sensory profiling. *J. of Sens. Stud.* **6**, 159–177.

Næs, T. , Irgens, C. and Martens, H. (1986). Comparison of linear statistical methods for calibration of NIR instruments. *J. R. Statist. Soc. B* **35**(2), 195–206.

Næs, T. and Kowalski, B. (1989). Predicting sensory profiles from external instrumental measurements. *Food Quality and Preference* **4**, 135–147.

Næs, T. and Martens, H. (1989). *Multivariate Calibration.* Wiley, Chichester.

Næs, T. and Martens, H. (1985). Comparison of prediction methods for multicollinear data. *Commun. Statist.-Simula. Computa.* **14**(3), 545–576.

Olver, F.W.J. (1984). *Asymptotics and special Functions.* Academic Press, New York.

O'Mahony, M. (1986). *Sensory Evaluation of Food.* Marcel Dekker, Inc., New York.

Pangborn, R.M. (1981). A critical review of threshold, intensity and descriptive analyses in flavor research. *Flavor*, 3–33.

Pangborn, R.M. (1987) Sensory science in flavour research: Achievements, needs and perspectives. In: Ed. M. Martens, G.A. Dalen and H. Russwurm Jr., *Flavour Science and Technology*, John Wiley & Sons Ltd.

Pangborn, J., Eriksson, F. and Remi, K. (1971). Simplified sialometer for continuous weight monitoring of salivary secretion. *J. Dent. Res.* **50**, 1689.

Patton, S. and Josephson, D.V. (1957). A method for determining significance of volatile flavour compounds in foods. *Food. Res.* **22**, 316–318.

Piggott, J.R. (1986)(Editor). *Statistical Procedures in Food Research.* Elsevier Applied Science, London.

Piggott, J.R. and Sharman, K. (1986). Methods to aid interpretation of multidimensional data. In: *Statistical Procedures in Food Research.* Ed. J.R. Piggott, Elsevier Applied Science, London.

Poll, L. and Hansen, K. (1990). Reproducibility of headspace analysis of apples and apple juice. *Lebensm.-Wiss. u.-Technol.* **23**, 481–483.

Powers, J.J. and Ware, G.O. (1986). Discriminant analysis. In: *Statistical Procedures in Food Research.* Ed. J.R. Piggott, Elsevier Applied Science, London.

Punter, P.H. (1983). Measurements of human olfactory thresholds for several groups of structurally related compounds. *Chemical Senses* **7**(3), 215–235.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science* **3**(2), 213–238.

Refsgaard, H.H.F., Brockhoff, P.M., Poll, L., Olsen, C.E., Rasmussen, M. and Skibsted, L.H. (1994). Light induced sensory and chemical changes in aquavit. Submitted to: *Lebensm.-Wiss. und Techn.*

Salo, P. (1970). Determining the odor thresholds for some compounds in alcoholic beverages. *Journal of Food Science* **35**, 95–99.

Sanchez, E. and Kowalski, B.R. (1990). Tensorial resolution: a direct trilinear decomposition. *J. Chemometrics* **4**, 24–45.

Satterthwaite, F.E. (1946). An approximate distribution of estimates of variance components. *Biom. Bull.* **2**, 110–114.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719–727.

Schlich, P. (1993). Uses of change-over designs and repeated measurements in sensory and consumer studies. *Food Quality and Preference* **4**(4), 223–235.

Schneider, B.E. (1978). Algorithm AS 121. Trigamma Function, *Applied Statistics* **27**, 97–99.

Searle, S.R. (1971). *Linear Models.* John Wiley, New York.

Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components.* Wiley, New York.

Shannon, I.L., Prigmore, J.R. and Chauncey, H.H. (1962). Modified Carlson-Crittenden device for the collection of parotid fluid. *J. Dent. Res.* **41**, 778–783.

Silvapulle, M.J. (1981). On the existence of maximum likelihood estimators for the binomial response model. *J. Royal Statist. Soc.* **B43**, 310–313.

Skovgaard, I.M. (1994). Basic saddlepoint approximations, *Personal communication.*

Snee, R. D. (1982). Nonadditivity in a two-way classification: Is it interaction or nonhomogeneous variance? *J. of Am. Stat. Ass.* **77**(379), 515–519.

Stiratelli, R., Laird N. and Ware, J.H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961–971.

Stone, M. (1974). Cross-validatory choice and assesment of statistical prediction. *J. Roy. Stat. Soc. ser. B*, 111–133.

Stone, M. and Brooks, R.J. (1990). Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *J. R. Statist. Soc. B* **52** (2), 237–269.

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

Sundberg, R. (1993). Continuum regression and ridge regression. *J. R. Statist. Soc. B* **55** (3), 653–659.

Sutter, J.M., Kalivas, J.H. and Lang, P.M. (1992). Which principal components to utilize for principal components regression. *J. Chemometrics* **6**, 217–225.

Tekin, A.R. and Karaman, H. (1992). Odor thresholds of some derivatives of strawberry aldehyde. *Chemical Senses* **17**(6), 795–799.

Ten Berge, J.M.F. (1977). Orthogonal Procrustes rotation to maximal agreement for two or more matrices. *Psychometrika* **42**(2), 267–276.

Teranishi, R., Buttery, R.G., Stern, D.J. and Takeoka, G. (1991). Use of odor thresholds in aroma research. *Lebensm.-Wiss. u.-Technol.* **24**, 1–5.

Tjur, T. (1984). Analysis of variance models in orthogonal designs. *International Statistical Review* **52**, 33–81.

Tjur, T. (1991). Analysis of variance and design of experiments. *Scandinavian Journal of Statistics* **18**, 273–369.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis *Psychometrika* **31**, 279–311.

Tukey, J. W. (1949). One degree of freedom for non-additivity. *Biometrics* **5**, 232–242.

Wedderburn, R.W.M. (1974). Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Weisberg S. (1985). *Applied Linear Regression.* Wiley, New York.

Wood, A.T.A., Booth J.G. and Butler, W. (1993). Saddlepoint approximations to the CDF of some statistics with nonnormal limit distributions. *J. Am. Stat. Ass.* **88**(422), Theory and Methods, 680–686.

Vuataz, L. (1986). Sensory difference testing and the measurement of sensory discriminability. In: *Statistical Procedures in Food Research,* Ed. J.R. Piggott, Elsevier Applied Science, London.

Yates, F. and Cochran, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28**, 556–580.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049–1060.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Stat. Ass.* **86**, 79–86.