Sensometrics: Thurstonian and Statistical Models

Rune Haubo Bojesen Christensen

Kongens Lyngby 2008
IMM-M.Sc-2008-62
The primary aim of this project has been to bridge psychologically anchored probabilistic Thurstonian models applied in sensometrics and statistical methodology. Sensometrics is the scientific area that applies mathematical and statistical methods to problems from sensory and consumer science. This aim has been approached in the six elements that constitute this thesis. The six elements comprise three manuscripts, a software add-on package for R (a statistical programming environment for data analysis and graphics) and two technical chapters. An introductory chapter provides some background on sensometrics and Thurstonian modelling as well as an introduction to and an overview of the six elements connecting their individual contributions.

The first element is the manuscript “Estimation and Inference in the Same Different Test” which is directed at sensory and consumer scientists and considered for submission to the sensometrics journal *Food Quality and Preference*. The manuscript describes how likelihood based inference can extend the scope of inference in a Thurstonian analysis of same different tests. Previously proposed methods for inference are shown to be inappropriate in a wide range of situations. It is shown how profile likelihoods provide important information and profile likelihood intervals are shown to be uniformly superior to the commonly used Wald confidence intervals. The main focus is on how to apply the proposed methods in practice, but derivations and technical details are provided in an appendix.

The second element is the manuscript “Estimation and Inference in the A-Not A test with sureness”. This manuscript is also directed at sensory and consumer scientists. It identifies a commonly used sensometric models based on Thurston-
nian theory and Signal Detection Theory as specific types of well known statistical models called generalized linear models, multivariate generalized linear models and extensions thereof notably the cumulative link location-scale model. This identification facilitates handling of explanatory variables and opens the door to powerful statistical tools such as profile likelihoods and likelihood ratio tests. Explanatory variables are currently not accounted for in much Thurstonian modelling, so the proposed methods extends the scope for inference and possible level of sophistication in the models. Especially important are concepts of learning and fatigue effects during testing which is easily modeled in the proposed framework.

The third element is the manuscript “Thurstonian Models for Sensory Discrimination Tests as Generalized Linear Models”, which is a joint work with Professor Per Bruun Brockhoff. This manuscript is also considered for submission to Food Quality and Preference. It identifies several discrimination tests defined by their so called psychometric functions as generalized linear models with a binomial distribution and the inverse psychometric functions as link functions. The identification opens the door to efficient computational methods already programmed in common statistical software. The identification also leads to easy computation and application of profile likelihoods and profile likelihood confidence intervals, which are shown to be important in many common settings because ordinary Wald based confidence intervals will be misleading.

The fourth element is the R-package sensR which implements functions to perform analyzes described in the previous three manuscripts and other functions for sensometric analyzes. The package is publicly available online via www.cran.r-project.org The package includes full documentation and help pages for all functions in an R-integrated fashion. The manual to the package is included in this thesis.

The fifth element is a technical chapter on basic likelihood theory, h-likelihood and mixed effect models. The chapter provides a selective review of theory and methods for these topics. The material provides the basis for extension of the models and methods presented in the manuscripts to allow for random effects. Such models can describe the covariance structure as well as the mean structure in data from discrimination tests applied to panels of judges or samples of consumers.

The sixth element is also a technical chapter dealing with extensions of the Thurstonian models for replicated discrimination tests applied to several judges or respondents. Existing models for these replicated settings aim at adjusting the standard errors of mean parameters for possible overdispersion in the data. The proposed models are able to model the covariance structure as well as the mean structure in the data and thereby provide a more complete mod-
elling of the data. The proposed models can be viewed as a synthesis of latent class mixture models and mixed effect models. This material is believed to be of general statistical interest, but further research on especially computational methods will be necessary before the material may condense into submittable manuscripts.
Det primære formål med dette projekt har været at bygge bro imellem de psykologisk forankrede probabilistiske Thurstoneske modeller anvendt i sensometri og statistiske metoder. Sensometri er det videnskabelige område, der anvender matematiske og statistiske metoder på problemer fra sensorik og forbrugervidenskab. Dette formål er søgt opnået med seks elementer der udgør denne afhandling. De seks elementer omfatter tre manuskripter, en software add-on pakke til det statistiske programmerings miljø R samt yderligere to tekniske kapitler. Et introducerende kapitel indeholder lidt baggrund om sensometri og Thurstoneske modeller samt en introduktion til og overblik over de seks separate elementer.

Det første element er manuskriptet “Estimation and Inference in the Same Different Test”, der er henvendt til sensorikere og tænkes indsendt til det sensometriske tidsskrift *Food Quality and Preference*. Manuskriptet beskriver hvordan likelihood baserede metoder udvider mulighederne for Thurstoneske analyser af same different eksperimenter. Tidligere foreslåede metoder for at analysere sådanne eksperimenter vises at være upassende i mange situationer. Det vises også hvordan profillikelihoodet bidrager med vigtig information og hvordan konfidensintervaller baseret på profillikelihoodet er de almindeligt brugte Wald-baserede konfidensintervaller overlegne. Fokus i manuskriptet er på hvordan metoderne bruges i praksis og udledninger samt tekniske detaljer er vedlagt i appendiks.

Det andet element er manuskriptet “Estimation and Inference in the A-Not A test with sureness”. Dette manuskript henvender sig også til sensorikere. Manuskriptet identificerer en ofte brugt sensometrisk model baseret på Thurstonesk teori og Signal Detection Theory som specifikke udgaver af velkendte statistiske modeller kaldet generaliserede lineære modeller, multivariate gen-
eraliserede lineære modeller og udvidelser heraf, specielt den kumulative link location-scale model. Denne identifikation muliggør håndteringen af forklarende variable og letter adgangen til stærke redskaber som profillikelihood og like-lihood ratio tests. Forklarende variable er almindeligtvis ikke inddraget i meget Thurstoneske modellering, så de foreslåede metoder udvider mulighederne for at udføre dækkende analyser og muliggør mere sofistikerede modeller. Særligt vigtigt er håndteringen af indlærings- og træthedstendenser under testsessioner, hvilket let håndteres i den foreslåede type af modeller.


Det fjerde element er R-pakken sensR, hvor en række funktioner er implementeret til at foretage analyser, som dem, der er beskrevet i de tre ovenstående manuskripter samt funktioner for yderligere sensometriske analyser. Pakken er gjort offentligt tilgængelig online via www.cran.r-project.org Pakken inkluderer dokumentation og hjælpsider for alle funktioner som for standard R-funktioner. Manualet til pakken er inkluderet i denne afhandling.


Det sjette og sidste element er også et teknisk kapitel omhandlende udvidelser af Thurstoneske modeller der kan håndtere diskriminations tests, hvor hvert individ har foretaget testen gentagne gange. Eksisterende modeller forsøger at tilrette standardfejlen af middelværdiparametrene for overdispersionen i data. De foreslåede modeller er i stand til at beskrive kovariansstrukturen såvel som middelværddstrukturen og giver derfor en mere komplet beskrivelse af data. De
foreslåede modeller kan ses som en syntese af latent klasse modeller og modeller med stokastiske effekter. Dette materiale forventes at have en general statistisk interesse men yderligere undersøgelser af specielt beregningsmæssige metoder vil være nødvendigt før materialet forventes at kondensere til et manuskript.
This thesis was prepared at DTU Informatics, Statistics Section, the Technical University of Denmark in fulfillment of the requirements for acquiring the Master of Science degree in engineering.

The project accounts for 30 ECTS points and has been carried out in the spring semester 2008 and was skilfully supervised by Professor Per Bruun Brockhoff.

The thesis deals with Thurstonian and statistical models in sensometrics. Sensometrics is the scientific area that applies mathematical and statistical methods to problems from sensory and consumer science. The main focus is on developing statistical methods, models and software tools for frequently used Thurstonian based discrimination methods applied in sensory and consumer science.

The thesis consists of an introductory chapter and a collection of: three manuscripts, a manual for a software add-on package and two additional chapters.

Lyngby, August 2008

Rune Haubo Bojesen Christensen
Papers and topics included in the thesis

A. Rune Haubo Bojesen Christensen. Estimation and Inference in the Same Different test. Manuscript for Food Quality and Preference

B. Rune Haubo Bojesen Christensen. Estimation and Inference in the A-not A test with Sureness. Manuscript for Food Quality and Preference

C. Per Bruun Brockhoff and Rune Haubo Bojesen Christensen. Thurstonian models for sensory discrimination tests as generalized linear models. Manuscript for Food Quality and Preference

D. Rune Haubo Bojesen Christensen and Per Bruun Brockhoff. Manual for R-package sensR. Package available via [www.cran.r-project.org](http://www.cran.r-project.org)


F. Rune Haubo Bojesen Christensen. Thurstonian models for replicated difference tests.
Foremost I wish to thank my supervisor Professor Per Bruun Brockhoff for many fruitfull discussions and your always helpful advise. Our meetings have always been held in a pleasant atmosphere and I have always left filled with enthusiasm about the project. Thank you for givning me the uppertunity to contribute to and attend the sensometrics conference in Canada this July and sending me to the UseR! conference in August in Germany. Thank you for believing in me, and for helping me to obtain a PhD stipendium, so I can continue studying and working under your supervision. I very much look forward to continue our fine collaboration in the next three years to come.

Graham Cleaver at Unilever, Vlaardingen, the Netherlands kindly provided the soup data with which I worked in the pre-masters project and which are used extensively in one of the manuscripts in the present thesis. Thank you also for answering all my questions about the data and study and not the least for your hospitality during Per’s and my visit in January 2008.

And thanks to Finn Kuno Christensen, the librarian at DTU Informatics for providing me with an abundance of books and articles without which I would without doubt have been able to accomplish only an insignificant fraction of what is now in this thesis.

Finally, but not the least, Rikke, my beautiful wife needs a special thanks. Thank you for your patience and your invaluable support during the project. Thank you for all your love and everything else.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Summary</strong></td>
<td>i</td>
</tr>
<tr>
<td><strong>Resumé</strong></td>
<td>v</td>
</tr>
<tr>
<td><strong>Preface</strong></td>
<td>ix</td>
</tr>
<tr>
<td><strong>Papers and topics included in the thesis</strong></td>
<td>xi</td>
</tr>
<tr>
<td><strong>Acknowledgements</strong></td>
<td>xiii</td>
</tr>
<tr>
<td><strong>1 Introduction and Overview</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background on Sensometrics and Thurstonian Modelling</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Overview of the Thesis</td>
<td>3</td>
</tr>
<tr>
<td><strong>A Estimation and Inference in the Same Different Test</strong></td>
<td>7</td>
</tr>
<tr>
<td>A.1 Abstract</td>
<td>8</td>
</tr>
<tr>
<td>A.2 Introduction</td>
<td>8</td>
</tr>
</tbody>
</table>
A.3 The Likelihood Function and ML Estimators .................. 10
A.4 Profile Likelihood and Measures of Uncertainty ................. 13
A.5 Inferential procedures for the same-different test ............... 15
A.6 Coverage Probability ........................................... 21
A.7 Difference and Equivalence Testing ............................. 22
A.8 Discussion ......................................................... 25
A.9 R-functions ......................................................... 27
A.10 Derivation of the MLEs of $\tau$ and $\delta$ ................. 28
A.11 Variance-covariance Matrix of the Parameters .................. 29
A.12 Likelihood root statistic and $p$-values ......................... 31
A.13 Simulation under the Same-Different Model .................... 31
A.14 R-code used in the Manuscript ................................ 32

B Estimation and Inference in the A-Not A test with Sureness 35
B.1 Abstract .......................................................... 36
B.2 Introduction ....................................................... 36
B.3 A-Not A without Sureness ........................................ 39
B.4 A-Not A with Sureness .......................................... 42
B.5 Analyzes with Explanatory Variables ............................ 50
B.6 ROC analysis ....................................................... 55
B.7 On the Assumption of Normality ................................ 57
B.8 Thresholds and Model Interpretation ............................ 59
B.9 Discussion and Summary ......................................... 60
CONTENTS

B.10 R-code used in the Manuscript .......................... 61

C Thurstonian models for sensory discrimination tests as generalized linear models ........................................ 71

C.1 Abstract ............................................................... 72
C.2 Introduction .......................................................... 72
C.3 General and Generalized Linear Models ......................... 73
C.4 The triangle, duo-trio, 2-AFC and 3-AFC as generalized linear models ....................................................... 76
C.5 The A-NOT A method ............................................... 81
C.6 Likelihood and Confidence Intervals ............................ 82
C.7 Summary and discussion ........................................... 86
C.8 Implementation in R ............................................... 87
C.9 GLM Attributes for Discrimination Tests ....................... 90

D Manual for R-package sensR ........................................ 101

E Likelihood, H-likelihood and Mixed Effect Models ............... 151

E.1 Introduction ......................................................... 152
E.2 Preliminary Likelihood Theory ................................... 153
E.3 Laplace’s Integral Approximation ................................ 156
E.4 Generalized Linear Models ....................................... 161
E.5 Extended Quasi-Likelihood ....................................... 166
E.6 Cumulative link Models ........................................... 169
E.7 Laplace Approximation in GLMMs .............................. 171
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>E.8 Extended Likelihood and h-loglikelihood</td>
<td>174</td>
</tr>
<tr>
<td>E.9 Normal Linear Mixed Models</td>
<td>177</td>
</tr>
<tr>
<td>F Thurstonian Models for Replicated Discrimination Tests</td>
<td>183</td>
</tr>
<tr>
<td>F.1 Introduction</td>
<td>184</td>
</tr>
<tr>
<td>F.2 The Conditional Model</td>
<td>184</td>
</tr>
<tr>
<td>F.3 Conditional Expectations</td>
<td>187</td>
</tr>
<tr>
<td>F.4 Mean and Variance of the random effects distribution</td>
<td>188</td>
</tr>
<tr>
<td>F.5 Attenuation Effects</td>
<td>192</td>
</tr>
<tr>
<td>F.6 Examples</td>
<td>193</td>
</tr>
<tr>
<td>F.7 Computational Issues</td>
<td>197</td>
</tr>
<tr>
<td>F.8 Additional R-functions</td>
<td>203</td>
</tr>
<tr>
<td>Bibliography</td>
<td>209</td>
</tr>
</tbody>
</table>
Introduction and Overview

The thesis deals with Thurstonian and statistical models in sensometrics. Sensometrics is the scientific area that applies mathematical and statistical methods to problems from sensory and consumer science. The main focus is on developing statistical methods, models and software tools for frequently used Thurstonian based discrimination methods applied in sensory and consumer science.

In sensometrics there exists two schools of thought for analyzing data from sensory and consumer science. A purely empirical school applies statistical methods without clear anchoring of the methods to psychological theories of sensory perception and cognitive decision processes. Thurstone (1927a,b,c) in a series of seminal papers strongly influenced these psychological theories. The approach is called probabilistic, which has caused the former approach to become known as deterministic even though it often applies statistical methods and therefore draws empirically based conclusions. Preachers of the latter school often regard the approaches of the former as close to meaningless.

The primary aim of this project has been to unify the two approaches via identification of the psychologically anchored probabilistic Thurstonian models as well known types of statistical models.

The remainder of this chapter provides some background on sensometrics and thurstonian modelling (section 1.1) and an introduction to and overview of the
Sensory and consumer data is frequently produced and applied as the basis for decision making in the food industry worldwide. Similarly for many other industries, for instance in the car, fragrance and high end hi-fi/TV industries the production and interpretation of such data is also an integral part of product development and quality control. Denmark is no different - all major food industries produce and use such data. Bang & Olufsen is a Danish example of a non-food company where sensory data and analysis play an important role. In food research, data is produced and used similar to the industrial use. Academic environments specifically for sensory and consumer sciences exists worldwide.

The development and application of statistics and data analysis in this area is called sensometrics. A common feature of sensory data is the use of human beings as measurement instruments. But humans are difficult to calibrate and different people have different perceptions of the same influences on their senses. This leads to several levels of variations and constitutes a special challenge to the sensometrician. Obtaining experimental data is often very expensive, so extracting the optimal amount of information is very important in product development and control as well as in research.

Sensometrics interfaces with many other fields. When the focus is on the perceptual aspects of individuals, several statistical methods reappear in experimental psychology and psychophysics. Medical decision making is another field in which humans (doctors) assess images from MR and CT scannings, hence the same challenges apply here.

The application of novel and advanced statistical methods has been and most likely continue to be conditional on the availability of software facilitating the application of the methodology. A significant part of this project has therefore focused on the development of the add-on package sensR for R (R Development Core Team 2008a).

In his seminal papers, Thurstone (1927a,b,c) developed his Law of Comparative Judgment. This law or model, has become widely acknowledged as a model for human perception of stimuli in its generality. The model describes the discrimination process for a single individual and is often applied under certain convenient simplifying assumptions. As such the model is applied in many re-
1.2 Overview of the Thesis

search areas and industries to judge differences between confusable stimuli (eg. products) or to investigate human perception. The model is however frequently applied to several individuals assuming that it holds in this case also without theoretical justification. One of the subjects in this project has been the development of Thurstonian models for replicated discrimination tests.

A discrimination method originally developed in the area of signal detection theory is frequently used also in sensory science, acoustical experiments and medical decision making and is identified as a Law of Categorical Judgment (Torgerson, 1958; Luce, 1994) supplementing Thurstone’s law. The analysis of data obtained using this method (such experiments are typically called ”A - Not A” or ”yes- no” potentially supplemented with ”with sureness”) is frequently based on ineffective estimates of key quantities.

Development of statistical methodology for modelling individual differences in discrete data has seen a break through in the last decade or so, but methodology to assess individual differences has yet to be developed for sensory data. Inference on individual differences is very important in assessing for instance whether two products are similar, in that some individuals may be able to detect a difference and others unable. In consumer studies some consumers may have a preference because they do detect product differences whereas other consumers do not detect the product difference and therefore have no preference. Characterization of the distribution of individuals is therefore of great importance and the statistical theory of such models calls for development. In the examples just mentioned, the primary objective can either be the products or the individuals as such - in both cases characterizing individual differences is mandatory for valid inference, emphasizing the importance of the development of the statistical models. One of the parts of this thesis is devoted to the development of such models. Although much development is still left it is a step in a promising direction.

1.2 Overview of the Thesis

The first manuscript (appendix A) on estimation and inference in the same different test is fairly self contained and needs only little introduction. I hope to submit the manuscript to the journal; Food Quality and Preference in the autumn of 2008. In the introduction of the manuscript, I state that code used in preparing the manuscript will be made available online. I expect the journal to provide a site for this. To make the code available for the readers of this thesis, I have included it in an appendix following the manuscript. The manuscript is written with the sensory analyst in mind, so the statistical level is held on a
moderate level.

The second manuscript (appendix B) on inference in the A-Not A test with sureness is also fairly self contained. Likewise I wish to submit the manuscript to *Food Quality and Preference* in the autumn of 2008. As for the previous manuscript, code used to prepare the manuscript appears in an appendix following the manuscript. A future paper is planned to follow up on the cumulative link location-scale models treated here and introduce random effects to cope naturally with replications over individuals.

The third manuscript (appendix C) is a joint work with Professor Per Bruun Brockhoff. When Per invited me to collaborate on the manuscript, a draft of the first three major sections were already on the desk. The contributions from my side include the section on likelihood and confidence intervals, the example for the A-Not A method and the appendices on R-functions and computational methods.

The manual for the *sensR* package for R ([R Development Core Team](http://www.cran.r-project.org)) presented as the fourth element in this thesis provides documentation for the *sensR* package. A significant amount of time during this project has gone into preparing this package. The is intended to be available via [www.cran.r-project.org](http://www.cran.r-project.org), but it can also be downloaded from [http://www2.imm.dtu.dk/~rhbc/sensR_1.0.0.zip](http://www2.imm.dtu.dk/~rhbc/sensR_1.0.0.zip) and [http://www2.imm.dtu.dk/~rhbc/sensR_1.0.0.tar.gz](http://www2.imm.dtu.dk/~rhbc/sensR_1.0.0.tar.gz) if CRAN fails. I started working on a handful of functions written by Per Bruun Brockhoff, which all have been extensively rewritten and enhanced.

The package implements Thurstonian based discrimination methods commonly used in the sensory society. Most of the functionality in the package is described in the previous three manuscripts. This includes methods for the same different method, the discrimination methods defined by their psychometric functions; the duo-trio, triangle, 2- and 3-AFC methods and the A-not A (or “yes-no”) method. Several extractor functions, such as summary, print and plot methods have been implemented supporting the fitting functions. The suggested profile likelihood methods and profile likelihood confidence intervals have also been implemented for these methods.

The most important function for the A-Not A with sureness method is the *clls* function implementing the cumulative link location-scale model ([Agresti](http://www.cran.r-project.org)). This function has to the best of my knowledge not been implemented in a publicly available function in R in its generality and may therefore be of more general interest for other areas of applied statistics. Areas where signal detection theory or ROC curve methods are applied such as medical diagnostics and radiology are obvious candidates, but other areas where ordinal and interval data are obtained are also potential target groups.
Much effort has been made to ensure that the functions in the package use numerical efficient and stable algorithms (Nocedal and Wright 2006; Bates and Watts 1988), and that user supplied arguments are checked for consistency. The intention is that the functions should provide sensible answers even if the functions are used in a non-sensible manner to some extent. Efforts have also been made to take advantage of the scoping rules in R (Venables and Ripley 2000; Chambers 1998). This mostly means that care is taken that actual arguments to functions are evaluated in the correct environments. This facilitates more advanced programming with the functions such as including functions from sensR in user written functions and supplying expressions as 'actual arguments' in the function calls. Advantage has also been taken of the object oriented structure in writing methods for the objects returned by the model-fitting functions (R Development Core Team 2008c). Additional references on the R language, and structure of various objects include (Chambers and Hastie 1992; Venables and Ripley 2002; R Development Core Team 2008b).

The fifth appendix, E includes a treatment of some basic likelihood theory, h-likelihood and mixed effect models. The review contains what I have considered important concepts and is not meant to be comprehensive in any way. The likelihood theory provides a very brief basis for the likelihood based methods developed in the manuscripts. The material on h-likelihood and mixed effect models is closely related to generalized linear models, which plays an important role in the manuscripts on methods for A-Not A with sureness and Thurstonian models as generalized linear models.

The theory on mixed effect models and the h-likelihood motivation of them provides the basis for the development of the mixture-mixed effect Thurstonian models for replicated discrimination tests. The h-likelihood approach to generalized linear mixed effect models was studied, but the considerations never condensed into writings and are therefore missing from the treatment in appendix E. Due to time constraints and a change of priority this gap was not filled. The h-likelihood framework contains many extensions of linear and generalized linear mixed effect models worthy of consideration. Time did however not allow further treatment of the model class.

The last element included in this thesis E is the work on Thurstonian models for replicated difference tests. These models can be viewed upon as a synthesis of latent class mixture models and generalized linear mixed effect models, where the random effects distribution is a mixture of a probability point mass and a continuous part. I have not been able to find mentioning of this type of models in the literature, so the treatment contains only few references. This part of the thesis is still in a premature phase and I expect to continue this work as part of the PhD study. One aspect of these models that need further work is that of estimation, which partly motivated the focus in this thesis on
computational methods for generalized linear mixed models. A few algorithms have been sketched and only the simplest; a direct integration and optimization has been implemented in the function \texttt{discrimR} (for Replicated discrimination method) included in \texttt{sensR}. 
Appendix A

Estimation and Inference in the Same Different Test

Manuscript for *Food Quality and Preference*
A.1 Abstract

Inference for the Thurstonian $\delta$ (also known as $d'$, "d-prime" in the signal detection theory literature) using its estimated variance is shown to be appropriate only in a limited range of situations. Situations where it is not appropriate occurs frequently. We propose a likelihood based inferential procedure, that is appropriate in all situations and we promote the use of profile likelihood curves.

The familiar Wald test and confidence interval rests on a variance estimate of the estimate of $\delta$ and is frequently misleading or not computable. Likelihood confidence intervals are shown to be meaningful and computable in all situations.

We discuss difference and equivalence testing for $\delta$ and show that the above points are especially important for equivalence tests. All methods discussed in the present paper are implemented in our free R-package, sensR.

A.2 Introduction

Miss Anna Sens, a sensory analyst performs an ordinary two-interval same-different experiment with 13 same and 13 different samples (elsewhere referred to as concordant and discordant sample pairs). The purpose of the experiment is to estimate the difference between the two products; $\delta$, the Thurstonian measure of discriminability. She obtains 8 same-answers to same-samples and 4 same answers to different-samples. The usual estimates of the parameters are $\hat{\tau} = 1.23$; $\hat{\delta} = 1.89$, with a 95% Wald confidence interval for $\delta$ being [0.51, 3.26] and corresponding $p$-value of 0.00369 for the test of $H_0: \delta = 0$ versus $H_1: \delta \neq 0$. This pleases her, as she now believes, she has demonstrated that the assessors were able to discriminate between the products. In this paper we justify the estimates by showing that they are in fact maximum likelihood estimates (MLEs). We also show that a better 95% profile likelihood confidence interval for $\delta$ is [0.00, 3.21] with a corresponding $p$-value of 0.0563 and that there is only weak evidence in the data of a positive $\delta$. Miss Anna Sens has not demonstrated that the assessors were able to discriminate between the two products on the 5% level.

\footnote{Anna began her public career with her appearance in \cite{Schlich1993} together with Mr. Mark Ting, whom we shall meet again shortly.}
Mr. Mark Ting also has two products and performs a same-different experiment with 13 same and 13 different samples. He is also interested in estimating the difference between the products, but on the contrary to Anna, he aims to perform the equivalence test $H_0 : \delta > \Delta$ versus $H_1 : \delta < \Delta$. He decides to declare equivalence, if there is reasonable evidence of $\delta$ being below $\Delta$ in the data. He chooses $\Delta = 1.5$, so if the upper confidence limit for $\delta$ is lower than $\Delta$, he has shown equivalence. Mark also obtains 8 same-answers to same-samples, but 11 same-answers to different-samples. Now the fraction of same-answers to same-samples (8/13) is lower than same-answers to different-samples (11/13), hence the only sensible estimate of $\delta$ is 0, and no previously published methods can guide Mark on how to obtain inference. We show that a 95% profile likelihood based confidence interval for $\delta$ is $[0, 1.34]$, hence Mark succeeds in showing equivalence of the two products.

The above two examples demonstrate the need for proper statistical methods for analysis of same-different experiments and this is the topic of this paper. We take a Thurstonian approach in that the methods we propose aim at inference for the Thurstonian $\delta$—a measure of discriminability or difference between sensations, eg. two products. Our treatment is confined to the ordinary two interval same-different test (Macmillan et al., 1977) assuming a $\tau$-strategy (see eg. O’Mahony and Rousseau, 2002, and references therein for a description of different strategies). The signal detection theory and Thurstonian representation of this test was developed by Macmillan et al. (1977) and will not be repeated here. The methods we discuss applies to both the short version and under mild assumptions also to the long version of the test. In the short version each individual receives either a same or a different pair. In the longer version each individual receives both a same and a different pair unaware that this is the case (see eg. O’Mahony and Rousseau, 2002). We follow general statistical practice and add a ‘hat’ to the estimate of a parameter, eg. $\hat{\delta}$, but note that $d’$ (d-prime) is elsewhere used instead of $\delta$.

The methods we propose are based on the assumptions that all observations are independent, that the distributions of sensory intensity have the same variance and are uncorrelated, ie. that Thurstone’s Case V (Thurstone, 1927a) applies. These assumptions are further discussed in section A.8.

This paper proceeds as follows. In section A.3 we discuss the likelihood function for the same-different test and show that previously proposed estimators of $\tau$ and $\delta$ coincide with maximum likelihood estimators (MLEs) in well-behaved situations. In section A.4 we show that the (profile) likelihood and corresponding confidence interval is a better measure of uncertainty and evidence of the parameter estimates than the variance and Wald confidence intervals. In section A.5 we discuss inferential procedures. We show how to use likelihood contour
and profile likelihood plots and discuss how to interpret confidence intervals. In section A.6 we show that the coverage probability of the likelihood confidence intervals is generally close to the nominal level. In section A.7 we discuss difference and equivalence testing using likelihood based methods. We end with a discussion and a summary in section A.8. For specific recommendations and illustration of the methods we suggest to read sections A.5 and A.7.

To improve the readability for the sensory scientist we have focused on the practical use of the methods in the main part of the paper. We document derivations and definitions in the appendices and hope that the theoretically or technically interested reader will find them illuminating.

An overview of functions related to the same-different protocol in our free R-package {R Development Core Team} (2008a) sensR (Christensen and Brockhoff 2008b; Brockhoff and Christensen 2008) is found in appendix A.9. The package is available from www.cran.r-project.org or by email from the corresponding author. All analyzes in this paper were performed using these functions. We have made available the code to produce the numbers and figures in the paper at www.???.dk.

### A.3 The Likelihood Function and ML Estimators

The data of a same-different test are distributed as the product of two binomial distributions and the data can fall in one of four distinct classes. The answer can be “same” or “different” when the pair of observations are either same or different. Let the probabilities of an answer falling in each of the categories be $p_j$, where $j = \{ss, sd, dd, sd\}$ ($s = \text{same}; d = \text{different}$) and the first subscript refers to the answer and the second refers to the true status. The probabilities are then given by the following relations (Macmillan et al. 1977)

\[
\begin{align*}
    p_{ss} &= 2\Phi\left(\frac{\tau}{\sqrt{2}}\right) - 1, \\
    p_{ds} &= 1 - p_{ss} = 2\{1 - \Phi\left(\frac{\tau}{\sqrt{2}}\right)\} \\
    p_{sd} &= \Phi\left(\frac{\tau - \delta}{\sqrt{2}}\right) - \Phi\left(\frac{-\tau - \delta}{\sqrt{2}}\right), \\
    p_{dd} &= 1 - p_{sd},
\end{align*}
\]

where $\Phi$ is the standard normal cumulative distribution function.

Assuming the data are independent and ignoring additive constants, the log-
The likelihood function for the same-different protocol is given by

\[ l(\tau, \delta; y) = \sum_j n_j \log p_j , \]  

(A.3.1)

where \( n_j \) is the number of observations in the \( j \)th category. Provided data are obtained in all four classes and the fraction of same-answers to same-samples is larger than the fraction of same-answers to different-samples, the MLEs of \( \tau \) and \( \delta \) are given as solutions to

\[ \hat{\tau} = \sqrt{2} \Phi^{-1} \left( \frac{2n_{ss} + n_{ds}}{2(n_{ds} + n_{ss})} \right) , \quad \frac{n_{sd}}{n_{dd} + n_{sd}} = \Phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) - \Phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) , \]  

(A.3.2)

where the estimate of \( \delta \) has to be obtained by iterative methods (appendix A.10). These estimators are identical to those derived in Macmillan et al. (1977); Kaplan et al. (1978), but they have to our knowledge not previously been justified as MLEs.

When these conditions are not satisfied, the estimates of \( \tau \) and \( \delta \) cannot be computed from (A.3.2), but in several cases MLEs exist nonetheless. We have summarized the 12 possible cases in table A.3. Note that there is no information on \( \delta \) when no same or different answers are obtained (case 7 and 8 respectively) or in the peculiar cases, when no same or different samples are provided (cases 11 and 12 respectively). In all other cases inference for \( \delta \) is possible. In section A.4 we will examine the frequency with which these cases occur for different \( \tau \), \( \delta \) and sample sizes.

As an example consider case 8, where no different answers have been obtained. In this case it is not possible to know, if the response “different” would be observed more often for different samples, than for same samples. There is therefore no information on the discriminable distance between the two stimuli. The same argumentation goes for case 7.

As another example consider case 5. The contour plot of the likelihood function, and the profile likelihoods for the parameters for some data, that falls in case 5 is shown in figure A.1. It is clear that the likelihood increases with \( \delta \) confirming that for case 5 the MLE of \( \tau \) is positive and finite and that the MLE of \( \delta \) is infinity.

In some cases, as indicated in the \( l(\hat{\theta}; y) \)-column, it is possible to show that the maximum log-likelihood is 0. In all other cases, the value depends on the specific set of data.

Even though \( \delta \) may be an unbiased measure of discriminability in contrast to the proportion of correct responses [Macmillan et al. 1977, Kaplan et al. 1978], the
Table A.1: MLEs and likelihood function under various conditions of the observed data.

<table>
<thead>
<tr>
<th>Case</th>
<th>Condition</th>
<th>$\tau^a$</th>
<th>$\delta^a$</th>
<th>$l(\hat{\theta}; y)^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$n_{ss}/N_s &gt; n_{sd}/N_d$</td>
<td>$\hat{\tau}$</td>
<td>$\delta$</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>$n_{ss}/N_s \leq n_{sd}/N_d$</td>
<td>$\hat{\tau}$</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>$n_{ss} = 0$</td>
<td>$\hat{\tau}$</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>$n_{ds} = 0$</td>
<td>$\hat{\tau}$</td>
<td>$\infty$</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>$n_{sd} = 0$</td>
<td>$\hat{\tau}$</td>
<td>$\infty$</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>$n_{dd} = 0$</td>
<td>$\hat{\tau}$</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>$n_{ss} = 0 &amp; n_{sd} = 0$</td>
<td>0</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>$n_{ds} = 0 &amp; n_{dd} = 0$</td>
<td>$\infty$</td>
<td>NA</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>$n_{ds} = 0 &amp; n_{sd} = 0$</td>
<td>$\infty$</td>
<td>$\infty$</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>$n_{ss} = 0 &amp; n_{dd} = 0$</td>
<td>$\hat{\tau}$</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>$n_{ss} = 0 &amp; n_{ds} = 0$</td>
<td>$\infty$</td>
<td>NA</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>$n_{sd} = 0 &amp; n_{dd} = 0$</td>
<td>$\hat{\tau}$</td>
<td>NA</td>
<td>-</td>
</tr>
</tbody>
</table>

$a$: Hat-notation indicates that the MLE is positive, finite and depends further on the data, NA stands for “Not Available”, i.e. no information on the parameter.

$b$: Value of the log-likelihood at the MLEs. "-" indicates further dependence on the data.

Figure A.1: Countour (left) and profile likelihood for $\delta$ (right), for data in case 5.
estimator of $\delta$ is not generally unbiased. The estimator may be approximately unbiased for intermediate $\delta$ around 2-3, but for low and for high $\delta$, the estimator is biased. Consider for instance the extreme case where $\delta = 0$. In this case some estimates will be 0, but some will also be positive, hence the estimator $\delta$ is positively biased.

### A.4 Profile Likelihood and Measures of Uncertainty

If we are interested in both $\tau$ and $\delta$, we can use the likelihood function directly to assess the uncertainty of the parameters and display it in a contour plot. To a good approximation, the likelihood ratio (also known as a normalized or relative likelihood) follows a $\chi^2_m$ distribution with $m$ degrees of freedom, where $m$ is the number of parameters. (see eg. Pawitan [2001], Boyles [2008], hence a 100$(1 - \alpha)$% highest likelihood confidence region is given by

$$\text{CR: } \Big\{ (\tau, \delta); 2[l(\hat{\tau}, \hat{\delta}; y) - l(\tau, \delta; y)] < \chi^2_{m, (1-\alpha)} \Big\}$$  \hspace{1cm} (A.4.1)

where incidentally $\chi^2_{m, (1-\alpha)} = \alpha$.

Often we are only interested in $\delta$ and wish to treat $\tau$ as a nuisance parameter. A general method to remove nuisance parameters is to profile the likelihood over these parameters. It is important to remove nuisance parameters properly and not merely substitute them with their MLEs to account for the uncertainty in estimating the nuisance parameters. The profile likelihood for $\delta$ is given by

$$l_{\tau}(\delta; y) = \arg \max_{\tau} l(\delta, \tau; y),$$ \hspace{1cm} (A.4.2)

where for each $\delta$ we wish to assess, we maximize the full likelihood over $\tau$. The profile likelihood for $\tau$ is given similarly. We explain the profile likelihood and how to use it more thoroughly in section [A.5]. A highest profile likelihood interval for $\delta$ is given by

$$\text{CI: } \Big\{ \delta; 2[l_{\tau}(\hat{\delta}; y) - l_{\tau}(\delta; y)] < \chi^2_{m, (1-\alpha)} \Big\}$$  \hspace{1cm} (A.4.3)

Profiling the likelihood often fares well, when the number of nuisance parameters is not too large and notably does not increase with the sample size. This is not a problem here and in section [A.6], we further verify that the profile likelihood confidence interval for $\delta$ has good coverage probability.

Traditionally the uncertainty of MLEs are assessed from the variance-covariance matrix of the parameters, which is often derived as the observed Fisher information matrix, by the less optimal expected Fisher information matrix, by some
approximation to one of these or by simulation. We derive the observed Fisher
information matrix for \((\tau, \delta)\) and show how to obtain standard errors for the
parameters from this in appendix A.11. Bi (2002) derived the variance of \(\hat{\delta}\)
using the bivariate delta method. This estimate coincides with that of the cor-
rect diagonal entry of the variance-covariance matrix as derived from observed
Fisher information matrix (see appendix A.11).

The Wald statistic is based on the asymptotic normality if \(\hat{\delta}\) and depends only
on \(\hat{\delta}\) and its estimated variance. The Wald CI is based on the Wald statistic
and is given by

\[
\text{CI: } \left\{ \hat{\delta}; \frac{|\hat{\delta} - \delta|}{\text{se}(\hat{\delta})} < z_{1-\alpha/2} \right\}, \tag{A.4.4}
\]

the limits of which, we can also write in the familiar fashion: \(\hat{\delta} \pm z_{1-\alpha/2}\text{se}(\hat{\delta})\).

The Wald CI would be exact, if the profile likelihood of \(\hat{\delta}\) was a Gaussian curve.
In other cases, the quality of the Wald CI depends on how close the profile
likelihood is to a Gaussian curve and especially how symmetric it is. This
is illustrated in figure A.3 where the solid curve is the profile likelihood and
the dotted curve is the Gaussian/Wald approximation for Anna’s experiment.
The 95% and 99% confidence intervals are given by the intersections with the
horizontal gray lines. It is clear that the profile likelihood is not symmetric and
that the Wald CI is an inappropriate approximation for \(\hat{\delta}\).

The variance of \(\hat{\delta}\) and therefore also the Wald CI cannot be a good measure of
the uncertainty of \(\hat{\delta}\), when data fall in other than case 1 (cf. table A.3) since only
in this case, is it possible for the profile likelihood to be reasonably symmetric.
This is also reflected by the fact that the observed Fisher information matrix
is only defined for positive and finite \(\hat{\delta}\) (appendix A.11). Throughout we will refer
to the Wald statistic, CI and the variance as “computable”, when data fall in
case 1. It is possible to obtain an estimate of the variance of \(\hat{\delta}\) by simulation
or approximate methods (eg. add a small number (eg. 0.5) to all empty cells),
but this will not render the variance estimate or the corresponding Wald CI as
a good measure of the uncertainty about \(\hat{\delta}\).

Now that we have established that data has to fall in case 1, if we want to use the
variance (ie. by using the Wald test or CI) to assess the uncertainty about \(\delta\),
we turn our attention to the frequency with which data fall in case 1. Recall that
the profile likelihood is often asymmetric when \(\delta\) is large or close to zero, hence
the frequency of case 1 serves as a lower bound for appropriate application of
the Wald statistic. In table A.4 we have summarized the frequency distribution
of the cases\(^2\) 1-10 for different \(\tau\), \(\delta\) and sample sizes.

\(^2\)We exclude from consideration those cases, where no same or different samples are pro-
vided.
A.5 Inferential procedures for the same-different test

The table shows that only for large sample sizes, $\delta$ around 2-3 and preferably moderate to large $\tau$, can we rely on the Wald statistic to be computable and a reasonable tool for inference. Not only will the Wald CI not reflect the evidence in data outside this interval, the power (for instance the probability that a CI does not include 0 in a discrimination test) will also be affected.

The power of a test using the Wald statistic depends on 1) the probability that the Wald statistic is computable and 2) the probability that we get a significant result given the alternative hypothesis is true. The power therefore decreases with the probability that the Wald statistic is computable. It is a discomforting feature that the probability that the Wald statistic is computable in general decreases with increasing $\delta$, when $\delta$ is above 2-3. It is natural to expect that the probability of obtaining a significant result, ie. the power, increases with the true $\delta$, but this is not the case for the Wald statistic. This tendency is especially pronounced for small $\tau$.

The likelihood ratio statistic does not share this feature, and the power does increase with the true $\delta$.

Inference based on the likelihood is accessible in all cases except those, where there is no information in the data about $\delta$, ie. cases 7 and 8. This occurs only for small sample sizes and with frequencies below 5%. This is not a criticism of likelihood inference, but merely a characteristic of the same different model. It warns us not to use small sample sizes for our studies.

A.5 Inferential procedures for the same-different test

How should Miss Anna Sens and Mr. Mark Ting report the results of their experiments? From the likelihood principle (Birnbaum, 1962) we know that the likelihood function contains all the available information in the sample about the parameters. It is therefore sensible to simply show a plot of the likelihood function for some values of the parameters around their MLE. We have included a contour plot of the likelihood surface for Anna’s experiment in figure A.2. We have also included a plot of the confidence regions for the parameters, cf. equation (A.4.1).

In the examples in the introduction, the main interest was in $\delta$ only and not in $\tau$, so it is natural to provide the profile likelihood curves for $\delta$. The parameter $\delta$ can also be regarded as a nuisance parameter if interest is in $\tau$. We show the normalized profile likelihood for both parameters in Anna’s experiment in
Table A.2: Simulated frequencies of cases in table A.3 based on $10^6$ samples in each case. When frequencies are zero or one up to three significant digits, decimals have been left out to enhance readability.

<table>
<thead>
<tr>
<th>$N^d$</th>
<th>$\tau$</th>
<th>Case</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.5</td>
<td>case 1</td>
<td>0.47</td>
<td>0.49</td>
<td>0.57</td>
<td>0.81</td>
<td>1</td>
<td>0.96</td>
<td>0.45</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.53</td>
<td>0.51</td>
<td>0.43</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.04</td>
<td>0.55</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>case 1</td>
<td>0.47</td>
<td>0.51</td>
<td>0.62</td>
<td>0.91</td>
<td>1</td>
<td>1</td>
<td>0.81</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.53</td>
<td>0.49</td>
<td>0.38</td>
<td>0.09</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0.79</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>case 1</td>
<td>0.46</td>
<td>0.51</td>
<td>0.65</td>
<td>0.95</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.54</td>
<td>0.49</td>
<td>0.35</td>
<td>0.05</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
</tr>
<tr>
<td>2a</td>
<td></td>
<td>case 1</td>
<td>0.46</td>
<td>0.47</td>
<td>0.53</td>
<td>0.71</td>
<td>0.98</td>
<td>0.80</td>
<td>0.26</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.54</td>
<td>0.53</td>
<td>0.47</td>
<td>0.28</td>
<td>0.02</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.20</td>
<td>0.74</td>
<td>0.97</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>case 1</td>
<td>0.46</td>
<td>0.49</td>
<td>0.57</td>
<td>0.81</td>
<td>1</td>
<td>0.98</td>
<td>0.57</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.54</td>
<td>0.51</td>
<td>0.43</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.43</td>
<td>0.89</td>
</tr>
<tr>
<td>2b</td>
<td></td>
<td>case 1</td>
<td>0.45</td>
<td>0.48</td>
<td>0.58</td>
<td>0.87</td>
<td>1</td>
<td>1</td>
<td>0.98</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.55</td>
<td>0.52</td>
<td>0.42</td>
<td>0.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.02</td>
<td>0.43</td>
</tr>
<tr>
<td>20</td>
<td>0.5b</td>
<td>case 1</td>
<td>0.36</td>
<td>0.37</td>
<td>0.38</td>
<td>0.44</td>
<td>0.45</td>
<td>0.21</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.56</td>
<td>0.55</td>
<td>0.53</td>
<td>0.44</td>
<td>0.20</td>
<td>0.05</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 3</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.08</td>
<td>0.31</td>
<td>0.69</td>
<td>0.91</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>1c</td>
<td></td>
<td>case 1</td>
<td>0.41</td>
<td>0.42</td>
<td>0.46</td>
<td>0.58</td>
<td>0.80</td>
<td>0.53</td>
<td>0.15</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.59</td>
<td>0.57</td>
<td>0.54</td>
<td>0.41</td>
<td>0.11</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.08</td>
<td>0.45</td>
<td>0.84</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>case 1</td>
<td>0.23</td>
<td>0.24</td>
<td>0.28</td>
<td>0.44</td>
<td>0.75</td>
<td>0.76</td>
<td>0.46</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 2</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
<td>0.34</td>
<td>0.07</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 4</td>
<td>0.15</td>
<td>0.15</td>
<td>0.16</td>
<td>0.17</td>
<td>0.18</td>
<td>0.17</td>
<td>0.10</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
<td>0.36</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 6</td>
<td>0.15</td>
<td>0.14</td>
<td>0.11</td>
<td>0.04</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 8</td>
<td>0.03</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>case 9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.01</td>
<td>0.08</td>
<td>0.15</td>
</tr>
</tbody>
</table>

$^a$: Few samples also ended up in case 4
$^b$: Few samples also ended up in cases 4, 6 and 9
$^c$: Few samples also ended up in cases 3, 4, 6, 7, 8, 9 and 10
$^d$: Number of same and different samples both equal to $N/2$
A.5 Inferential procedures for the same-different test

Figure A.2: Likelihood contour plot for Anna’s experiment. Left: Normalized likelihood with confidence regions. Right: Log likelihood contours.

Figure A.3: Profile likelihood of $\delta$ and $\tau$ for Anna’s experiment. Horizontal gray lines denotes 95% and 99% confidence limits. Dotted lines are normal approximations and the dashed line is the estimated likelihood.
Figure A.4: Profile likelihood of $\delta$ and $\tau$ for Mark’s experiment. Horizontal gray lines denote 95% and 99% confidence limits.
The horizontal grey lines denote the 95% and 99% confidence limits, cf. equation (A.4.3).

The normalized (profile) likelihood curves measure the evidence in the data about the parameters (Royall 1997; Blume 2002). The profile likelihood of \( \delta \) has its peak at the MLE, \( \hat{\delta} \) and for Anna’s experiment likely values for \( \delta \) are those between zero and around three. The profile likelihood of \( \delta \) is highly asymmetric and that of \( \tau \) only little so. The profile likelihood of \( \delta \) shows that values between 1 and 3 are generally more likely than those between 0 and 1—information which is not contained in the CI. We therefore propose to show the profile likelihood when reporting the results of a same-different experiment.

When the profile likelihood is close to symmetric as it is for \( \tau \), a confidence interval captures most of the information in the sample about the parameter. In case of approximate symmetry, the researcher may stick with the CI, but we strongly discourage reporting \( p \)-values instead. CIs and \( p \)-values are two sides of the same coin and the information in the CI is almost always more relevant than the that of the \( p \)-value (Nester 1996) (in appendix A.12 we show how the likelihood root statistic is defined for the same-different protocol and how it can be used to obtain \( p \)-values). The \( p \)-value measures extremeness in the data under a null hypothesis. It does not in general measure evidence in data about a parameter (Goodman 1999; Cornfield 1966; Blume and Peipert 2003). We consider hypothesis testing further in section A.7.

We will now explain the profile likelihood in more detail using the data from Anna’s experiment as an example. First, by expanding all terms in (A.3.1) and inserting the observed data, we find that the log likelihood function is given by

\[
\begin{align*}
\ell(\tau, \delta; y) &= 8 \log[2\Phi(\tau/\sqrt{2}) - 1] + 5 \log[2 - 2\Phi(\tau/\sqrt{2})] + \\
&4 \log[\Phi((\tau - \delta)/\sqrt{2}) - \Phi((-\tau - \delta)/\sqrt{2})] + \\
&9 \log[1 - \Phi((\tau - \delta)/\sqrt{2}) - \Phi((-\tau - \delta)/\sqrt{2})].
\end{align*}
\] (A.5.1)

The likelihood function is therefore simply a function of the parameters (\( \tau \) and \( \delta \)) given the observed data.

To illustrate the profile likelihood using Anna’s data, suppose that we are interested in inference for \( \delta \). The most obvious approach would perhaps be to merely replace \( \tau \) with the ML estimate, \( \hat{\tau} \). This would make the likelihood (A.5.1) a function of \( \delta \) only. This is called the estimated likelihood and is different from the profile likelihood, but it serves as a good reference. Recall that the MLE of
$\tau$ is $\hat{\tau} = 1.23$, the estimated log likelihood for Anna’s data is therefore given by

$$l(\hat{\tau}, \delta; y) = 8 \log[2\Phi(1.23/\sqrt{2}) - 1] + 5 \log[2 - 2\Phi(1.23/\sqrt{2})] +$$

$$4 \log[\Phi((1.23 - \delta)/\sqrt{2}) - \Phi((-1.23 - \delta)/\sqrt{2})] +$$

$$9 \log[1 - \Phi((1.23 - \delta)/\sqrt{2}) - \Phi((-1.23 - \delta)/\sqrt{2})].$$

(A.5.2)

To convert this from an estimated log likelihood to an estimated likelihood take the anti-log (or exp) function of the estimated log likelihood. The maximum of this likelihood is however not 1 as in the figure. To normalize so that the curve has a maximum at 1, we divide the estimated likelihood with its maximum value.

We have illustrated this estimated likelihood in figure A.3 for $\delta$ with the dashed curve. It is clear that this curve is more narrow than the profile likelihood reflecting that we do not account properly for the uncertainty in estimating $\tau$ in the estimated likelihood.

The profile likelihood does account for the uncertainty in the estimation of $\tau$ properly. The profile log likelihood is given by equation (A.4.2) which says that for each $\delta$, we want to know the value of profile likelihood, we have to maximize the likelihood for $\tau$. For instance we could compute the value of the log likelihood at the MLE of both $\tau$ and $\delta$; ($\hat{\tau} = 1.23$, $\hat{\delta} = 1.89$) with equation (A.5.1) and we would get $-16.7$. If we want the value of the profile likelihood at $\delta = 1$, we substitute $\delta$ with 1 in (A.5.1). Next, we maximize over $\tau$ and find that the maximum value of the log likelihood is $-17.3$. This is the value of the profile log likelihood for $\delta = 1$, which we can state explicitly as $l_\tau(\delta = 1; y) = -17.3$. This value occurs at $\tau = 0.99$, which is clearly different from $\hat{\tau} = 1.23$, hence the difference between the estimated and the profile log likelihoods.

To compute the whole profile likelihood curve in figure A.3 (right), we need to evaluate the profile log likelihood at a lot of values for $\delta$ between 0 and 4. To obtain the solid line in figure A.3 we also need to anti-log the profile log likelihood and normalize as explained for the estimated likelihood above. The profile likelihood is obviously hard to evaluate without a computer. Especially one will benefit from an automatic optimization routine. The profile-function in R-package sensR has automatized this task.

One very desirable property of the likelihood CI is that it respects the allowed values of the parameters, hence the likelihood CI for $\delta$ will not include values below zero. This property is not shared by the Wald CI.

A CI, say a 95% kind, is commonly justified by the frequency property: 95% of the 95% confidence intervals we make, will contain the true parameter. As such it tells us about what happens in the long run, but provides no informa-
A.6 Coverage Probability

In this section we examine the coverage probability of the profile likelihood CI for $\delta$. The coverage probability of a CI is the probability that the CI will contain the true parameter. A coverage probability at the nominal level (eg. 95%) is a desirable feature of a CI, but it is not the only relevant feature. At least as important is that the CI contains those values of the parameter most supported by data, as discussed in section A.5, and that the CI is frequently computable as discussed in section A.4. We have also computed the coverage probability of the Wald CI, but note that this is only relevant for situations in which the Wald statistic is frequently computable, ie. when $\delta$ is around 2-3, the sample size is large and preferably $\tau$ is moderate to large.

Under the scenarios in table A.6, we simulated 50,000 data sets from the same-different model (A.3.1) (See appendix A.13 on how to do this), computed the 95% profile likelihood CI and the 95% Wald CI for $\delta$ and recorded if the CIs contained the parameter with which we simulated. The coverage probability in table A.6 is expressed as the proportion of times, the true parameter is covered by the CI. Table A.6 summarizes the results. Note that the standard error of the estimated probabilities in the table is approximately $se(\hat{p}) = \sqrt{0.05 \cdot 0.95/50,000} = 0.001$ providing reasonable accuracy.

The coverage probabilities are based on those cases that were computable only. The frequencies with which the Wald and likelihood CIs are computable are shown in table A.4. The Wald CI is computable if data fall in case 1 and the likelihood CI is computable in all cases except 7 and 8, where no information on $\delta$ is available.
The coverage probability of the profile likelihood CI (denoted log-lik) is close to the nominal level (0.95) for medium sized $\delta$ and is in general higher for very low and very large $\delta$, hence the type I error is well controlled. For the profile likelihood CI there seems to be little dependence on the sample size and the value of $\tau$. The coverage probability does however get near 90% for $\delta$ around 3-4, small sample size and large $\tau$. This reminds us not to put too much emphasis on frequentist probability statements ($p$-values and the like) with small sample sizes.

The coverage probability of the Wald CI, where this is frequently computable (ie. for $\delta$ around 2-3 and large sample size) is close to the nominal level. This occurs because the Wald CI closely approximates the profile likelihood CI in these situations.

A peculiar situation occurs For small sample sizes and $\delta$ around 2-3. Here the coverage probability is 100% and it is certain, that if the Wald confidence interval is computable, then it will contain the true parameter. Since the Wald CI is only computable around half the time in these settings, it is not really helpful.

For small $\delta$ (especially 0.25 - 0.5) the coverage probability for the Wald CI is systematically very low. This means that not only is the Wald CI computable in around 50% of the time, the CI will also contain the true parameter far less frequently than it should. This evidently leaves the Wald statistic inappropriate for equivalence testing.

Based on the simulation studies summarized in tables [A.4] and [A.6] we recommend not to use the Wald CIs, but report profile likelihood CIs and preferably show a plot of the profile likelihood curve.

### A.7 Difference and Equivalence Testing

In difference testing we aim to show a difference and in equivalence testing we hope to show equivalence (also termed parity or similarity). Equivalence testing has been discussed numerous times in the sensometric literature (see eg. [Bi, 2005], [Meyners, 2007b], [Bi, 2007], [Ennis et al., 2007], [Ennis, 2007], [Bi, 2006a]), but none with specific reference to tests of $\delta$ or same-different experiments.

Equivalence is accepted, if the difference between the sensations is smaller than some pre-specified $\Delta$ as demonstrated in the introductory example with Mr. Mark Ting. We propose to take a Thurstonian likelihood based approach in
Table A.3: Coverage probability for Wald and log-likelihood based 95% confidence intervals and proportion of convergences (Conv.) for the Wald interval. All likelihood intervals converged. Coverage probability for the Wald intervals are based only on those cases which converged.

<table>
<thead>
<tr>
<th>N</th>
<th>τ</th>
<th>Type</th>
<th>0</th>
<th>0.25</th>
<th>0.5</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.5</td>
<td>log-lik</td>
<td>0.974</td>
<td>0.973</td>
<td>0.974</td>
<td>0.967</td>
<td>0.946</td>
<td>0.936</td>
<td>0.981</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.972</td>
<td>0.779</td>
<td>0.857</td>
<td>0.934</td>
<td>0.970</td>
<td>0.980</td>
<td>0.947</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>log-lik</td>
<td>0.975</td>
<td>0.975</td>
<td>0.974</td>
<td>0.955</td>
<td>0.946</td>
<td>0.946</td>
<td>0.946</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.975</td>
<td>0.794</td>
<td>0.889</td>
<td>0.951</td>
<td>0.957</td>
<td>0.955</td>
<td>0.969</td>
<td>0.894</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>log-lik</td>
<td>0.971</td>
<td>0.973</td>
<td>0.974</td>
<td>0.951</td>
<td>0.948</td>
<td>0.948</td>
<td>0.944</td>
<td>0.933</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.972</td>
<td>0.841</td>
<td>0.920</td>
<td>0.965</td>
<td>0.953</td>
<td>0.953</td>
<td>0.951</td>
<td>0.968</td>
</tr>
<tr>
<td>100</td>
<td>0.5</td>
<td>log-lik</td>
<td>0.959</td>
<td>0.974</td>
<td>0.974</td>
<td>0.973</td>
<td>0.942</td>
<td>0.938</td>
<td>0.978</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.973</td>
<td>0.750</td>
<td>0.833</td>
<td>0.924</td>
<td>0.971</td>
<td>0.988</td>
<td>0.889</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>log-lik</td>
<td>0.959</td>
<td>0.977</td>
<td>0.977</td>
<td>0.963</td>
<td>0.952</td>
<td>0.945</td>
<td>0.981</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.974</td>
<td>0.779</td>
<td>0.868</td>
<td>0.941</td>
<td>0.973</td>
<td>0.976</td>
<td>0.960</td>
<td>0.623</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>log-lik</td>
<td>0.972</td>
<td>0.973</td>
<td>0.971</td>
<td>0.958</td>
<td>0.951</td>
<td>0.943</td>
<td>0.942</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.975</td>
<td>0.834</td>
<td>0.911</td>
<td>0.963</td>
<td>0.962</td>
<td>0.947</td>
<td>0.966</td>
<td>0.962</td>
</tr>
<tr>
<td>20</td>
<td>0.5</td>
<td>log-lik</td>
<td>0.975</td>
<td>0.957</td>
<td>0.966</td>
<td>0.944</td>
<td>0.951</td>
<td>0.981</td>
<td>0.972</td>
<td>0.992</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.971</td>
<td>0.772</td>
<td>0.845</td>
<td>0.977</td>
<td>1.000</td>
<td>1.000</td>
<td>0.984</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>log-lik</td>
<td>0.972</td>
<td>0.971</td>
<td>0.973</td>
<td>0.973</td>
<td>0.923</td>
<td>0.972</td>
<td>0.984</td>
<td>0.981</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.959</td>
<td>0.779</td>
<td>0.846</td>
<td>0.919</td>
<td>0.996</td>
<td>1.000</td>
<td>0.944</td>
<td>0.199</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>log-lik</td>
<td>0.973</td>
<td>0.954</td>
<td>0.950</td>
<td>0.955</td>
<td>0.942</td>
<td>0.907</td>
<td>0.904</td>
<td>0.985</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>0.959</td>
<td>0.978</td>
<td>0.967</td>
<td>0.994</td>
<td>0.995</td>
<td>0.999</td>
<td>0.963</td>
<td>0.887</td>
</tr>
</tbody>
</table>

*a*: No. same and different samples both equal to \( N/2 \)
both situations. We believe the relevant parameter to consider in both difference and equivalence testing is $\delta$ and independently of whether interest is in the assessors or in the products. The null and alternative hypotheses for discrimination testing are

$$H_0 : \delta = 0 \quad H_1 : \delta \neq 0.$$  

The null and alternative hypotheses for equivalence testing are

$$H_0 : \delta > \Delta \quad H_1 : \delta < \Delta,$$

where $\Delta$ is a tolerance specified by the researcher. In the literature, equivalence tests are often represented by two one sided tests (TOST) Schuirmann (1981, 1987), Berger and Hsu (1996), but since $\delta$ is bounded above zero, the equivalence test for the same-different protocol involves only this single test (This holds more generally for other protocols also based on Thurstonian models such as alternative forced choice, the duo-trio, the triangle and the A-Not A protocols).

If Anna was not interested in the magnitude of $\delta$, but merely whether it was different from zero, she could have used Pearson’s $\chi^2$-test or Fisher’s exact test, to test if the frequency of same and different answers would differ between same and different samples in a $2 \times 2$ table. Both tests (using Monte Carlo simulation in the former (see eg. Meyners, 2007a)) both result in the two-sided $p$-value of $p = 0.24$ (the one-sided $p$-value of Fisher’s exact test is half this value) consistent with the likelihood result. This procedure differs substantially from those otherwise discussed in this paper in that it does not provide an estimate of the parameters nor the uncertainty about these, and that it is not useful in equivalence testing. We regard such testing as generally less informative than likelihood modelling in that it does not provide information on the location of $\delta$ and the uncertainty about that location. The simplicity of the testing does however have some appeal and in simple settings, it may provide sufficient information.

Likelihood modelling and pure hypothesis testing represents two different statistical approaches aiming at answering two different answers (Lehmann 1993, Royall 1997, Efron 1998). The likelihood approach aims a describing what the data say, and the latter tells us what decision to make while controlling the frequency with which the wrong decision is made (the $\alpha$ risk). As noted by Cox (1958) even when the goal is to make a decision it is helpful to clarify what the data say before that information is combined with additional information to provide a decision.
A.8 Discussion

The statistical theory of likelihood is well developed, but has only to a limited degree entered common practice in many areas including sensometrics and sensory science. We believe this is due mainly to two circumstances. First, the traditional treatments of the theory is accessible only to mathematical statisticians and second, the availability of software supporting inference based on likelihood is limited.

In this paper we have tried to make likelihood methods accessible and shown the advantage of such methods for sensory science. Regarding the second matter, we provide the R-package sensR.

Relying on the variance as a good measure of uncertainty in $\hat{\delta}$ and consequently using the Wald test and CI applies to some extend more generally to other test protocols. The main problem is that the profile likelihood of $\delta$ is highly asymmetric, when $\delta$ is not of intermediate size, and the sample size is large. This is also why Bi et al. (1997) recommend not to use the Wald test and CI, when $N \leq 1.65p_c(1-p_c)/(p_c - 1/m)^2$, where $N$ is the sample size, $m$ is the number of choice alternatives and $p_c$ is the proportion of correct responses. This rule obviously does not apply for the same-different experiment.

Miller (1996) investigated the behavior of the variance of $\delta$ (denoted by him $d'$) for the A-Not A protocol (called by him the “yes-no” protocol) and his table 2 clearly shows, that for $\delta$’s that are not intermediate and obtained from a large sample size, the variance of $\delta$ is an inappropriate measure of the uncertainty in $\delta$. This is a general problem for inference about $\delta$, since $\delta$ is bounded above 0 and the low information about the location of $\delta$, for large $\delta$. Miller also examined several approximate methods to arrive at CIs for $\delta$ in a A-Not A test. He concludes with the recommendation to compute the CI by several methods to ensure, that the CI is not too dependent on the computational method. He found no method to be universally satisfactory. The profile likelihood CI deals with this problem in all test protocols and provides the CI most supported by the data in any situation. Brockhoff and Christensen (2008) show how to obtain and use the (profile) likelihood and corresponding CIs for the duo-trio, triangle, 2 and 3-AFC and A-Not A test protocols.

The methods described in this paper are derived assuming that observations are independent. There is a need for models that allows modelling of the distributions of subjects with respect to their discriminabilities or preferences to cope with heterogeneous populations. We hope that future research will result in such models.
Other assumptions include those of equality and no correlation of the latent distributions—the most widely used simplifications of Thurstone’s law of comparative judgment (Thurstone, 1927a). These assumptions may be believed to be tenable in most applications, but formal models in which the assessment of these issues is possible still needs development.

Another issue that needs further addressing is that of power for both difference and equivalence testing. How does for instance the power depend on \( \tau \) and the ratio of same and different samples? Analytical power and sample size calculations requires the knowledge of the distribution of the test statistic under the alternative hypothesis. From the simulations in this paper it is clear that such calculations should not proceed with the Wald statistic, and deriving the distribution of the likelihood ratio or root statistics under the alternative probably requires some work. An appealing alternative is empirical calculations of power based on simulation of the \( p \)-value under the alternative. We provide a function `samediffPwr` in our package `sensR` that will do just that. The price for the simplicity of this method is computational time.

Lee et al. (2007) examined whether assessors can use a \( \beta \) as well as a \( \tau \) criterion in the same different test. It is natural to do a comparison of two models based on the two criteria respectively in a likelihood ratio test. We believe the present formulation of the likelihood is a step toward a more general comparison of models with the purpose of assessing the cognitive strategy of the assessors.

In conclusion we have shown the advantage of (profile) likelihood methods for appropriate estimation and inference in same different experiments. We have shown that situations in which the variance of \( \hat{\delta} \) and the Wald statistic is not computable and an inappropriate measure of uncertainty about \( \hat{\delta} \) occurs frequently. We have proposed to use and report the profile likelihood curve and corresponding CI, since these are aways computable provided there is evidence about \( \delta \) in the data and shown that the profile likelihood CI has good coverage probability. We suggest to use hypotheses about \( \delta \) for both equivalence and difference testing and show that the Wald statistic is never appropriate for the former and only appropriate for the latter in very limited situations. The profile likelihood however always provides valid and appropriate inference. We provide a free R-package that facilitates the methodology developed in this paper.
Appendix

A.9 R-functions

This section gives a short overview of the functions related to the same-different protocol implemented in the R-package sensR. The package is constantly expanding and additional functionality for the same-different protocol may be added with time. The authors are happy to receive contributions and suggestions for improvements. Further information on the package and comprehensive documentation of all functions is included in the package. The package can be downloaded from www.cran.r-project.org.

The main function is `samediff`, which fits a same-different model to data and estimates the maximum likelihood estimates of the parameters. The function returns an object of class "samediff" for which a number of methods exist.

Methods for samediff objects:

- **summary** will give a summary of the `samediff` fit including a table with parameter estimates, confidence intervals and p-values. Profile likelihood CIs are default, but Wald type CIs can be chosen.

- **profile** will compute the profile likelihood of the parameters. There exists a plot method for profile objects, that will make reasonable default plots of the profile likelihoods of the parameters as shown in this paper.

- **confint** will compute the confidence intervals based on profile likelihoods.

- **contour** will make a contour plot of the likelihood surface. By default the normalized likelihood is plotted with confidence limits, but this can be altered by the user.

- **plot** will make a plot of the distributions of sensory intensity provided the parameter estimates exists and are positive and finite.

- **samediffSim** will simulate data from a same-different models. The function requires the user to specify the parameters, the number of same and different samples and the number of set of observations to simulate.

A number of additional methods exist including `vcov` that will give the variance-covariance matrix if it exists, `coef` that will give the parameter estimates, `log-`
Lik will give the value of the log-likelihood at the MLE, AIC will give the value of AIC for the model fit and update will update a the model fit to eg. changes in the data.

A.10 Derivation of the MLEs of $\tau$ and $\delta$

The maximum likelihood estimates (MLEs) of $(\tau, \delta)$ satisfy the score equations

$$S(\hat{\tau}) = \frac{\partial l}{\partial \tau} = 0, \quad S(\hat{\delta}) = \frac{\partial l}{\partial \delta} = 0,$$

where the derivatives are given by

$$\frac{\partial l}{\partial \tau} = n_{ss} \left[ 2 \Phi \left( \frac{\tau}{\sqrt{2}} \right) - 1 \right]^{-1} \frac{2}{\sqrt{2}} \phi \left( \frac{\tau}{\sqrt{2}} \right) +$$

$$n_{ds} \left[ 2 - 2 \Phi \left( \frac{\tau}{\sqrt{2}} \right) \right]^{-1} \frac{-2}{\sqrt{2}} \phi \left( \frac{\tau}{\sqrt{2}} \right) +$$

$$n_{sd} \left[ \Phi \left( \frac{\tau - d}{\sqrt{2}} \right) - \Phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right]^{-1} \left\{ \frac{1}{\sqrt{2}} \phi \left( \frac{\tau - d}{\sqrt{2}} \right) + \frac{1}{\sqrt{2}} \phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right\} +$$

$$n_{dd} \left[ 1 - \Phi \left( \frac{\tau - d}{\sqrt{2}} \right) + \Phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right]^{-1} \left\{ \frac{-1}{\sqrt{2}} \phi \left( \frac{\tau - d}{\sqrt{2}} \right) - \frac{1}{\sqrt{2}} \phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right\},$$

$$\frac{\partial l}{\partial \delta} = n_{sd} \left[ \Phi \left( \frac{\tau - d}{\sqrt{2}} \right) - \Phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right]^{-1} \left\{ \frac{-1}{\sqrt{2}} \phi \left( \frac{\tau - d}{\sqrt{2}} \right) + \frac{1}{\sqrt{2}} \phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right\} +$$

$$n_{dd} \left[ 1 - \Phi \left( \frac{\tau - d}{\sqrt{2}} \right) + \Phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right]^{-1} \left\{ \frac{1}{\sqrt{2}} \phi \left( \frac{\tau - d}{\sqrt{2}} \right) - \frac{1}{\sqrt{2}} \phi \left( \frac{-\tau - d}{\sqrt{2}} \right) \right\},$$

where $\phi$ and $\Phi$ denotes the standard normal PDF and CDF respectively.

Starting with the latter score equation, we find that at the MLE, assuming they are positive and finite, we have the following relation

$$\frac{n_{sd}}{n_{dd} + n_{sd}} = \Phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) - \Phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right). \quad (A.10.1)$$

Using this relation, we find that the last two terms of the first score equation disappears, reducing it to

$$n_{ss} \left[ 2 \Phi \left( \frac{\hat{\tau}}{\sqrt{2}} \right) - 1 \right]^{-1} \frac{2}{\sqrt{2}} \phi \left( \frac{\hat{\tau}}{\sqrt{2}} \right) + n_{ds} \left[ 2 - 2 \Phi \left( \frac{\hat{\tau}}{\sqrt{2}} \right) \right]^{-1} \frac{-2}{\sqrt{2}} \phi \left( \frac{\hat{\tau}}{\sqrt{2}} \right) = 0,$$
from which we find that the MLE of \( \tau \) is

\[
\hat{\tau} = \sqrt{2} \Phi^{-1} \left( \frac{2n_{ss} + n_{ds}}{2(n_{ds} + n_{ss})} \right).
\]  

(A.10.2)

The MLE of \( \delta \) can be obtained by solving \( \text{(A.10.1)} \) for \( \hat{\delta} \), which has to be done iteratively.

## A.11 Variance-covariance Matrix of the Parameters

The observed Fisher information matrix is the Fisher information matrix evaluated at the MLE, \( \hat{\theta} = (\hat{\tau}, \hat{\delta}) \), which is given by

\[
I(\hat{\theta}) = - \frac{\partial^2 l}{\partial \theta^2} \bigg|_{\theta = \hat{\theta}}.
\]

Denote the elements of the observed Fisher information matrix by

\[
I(\hat{\tau}, \hat{\delta}) = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}
\]

Then its inverse,

\[
I^{-1}(\hat{\tau}, \hat{\delta}) = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}
\]

is the variance-covariance matrix of \( (\hat{\tau}, \hat{\delta}) \). The standard errors of \( (\hat{\tau}, \hat{\delta}) \) are therefore given as

\[
\text{se}(\hat{\tau}) = \sqrt{I^{11}} \quad \text{se}(\hat{\delta}) = \sqrt{I^{22}}.
\]

The values of \( I^{11} \) and \( I^{22} \) are given by

\[
I^{11} = I^* I_{22} \quad I^{22} = I^* I_{11},
\]  

(A.11.1)
where $I^* = [I_{11}I_{22} - I_{12}^2]^{-1}$, and the elements of the observed Fisher information matrix are given by

\[
I_{11} = -\frac{\partial^2 l}{\partial \tau^2} \bigg|_{\theta = \hat{\theta}} = \frac{(n_{dd} + n_{sd})^3}{2n_{dd}n_{sd}} \left[ \phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) + \phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) \right]^2 + \frac{2(n_{ds} + n_{ss})^3}{n_{ds}n_{ss}} \phi^2 \left( \frac{\hat{\tau}}{\sqrt{2}} \right)
\]

\[
I_{12} = I_{21} = -\frac{\partial^2 l}{\partial \tau \partial \delta} \bigg|_{\theta = \hat{\theta}} = \frac{-2(n_{dd} + n_{sd})^3}{2n_{dd}n_{sd}} \left[ \phi^2 \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) - \phi^2 \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) \right]
\]

\[
I_{22} = -\frac{\partial^2 l}{\partial \delta^2} \bigg|_{\theta = \hat{\theta}} = \frac{(n_{dd} + n_{sd})^3}{2n_{dd}n_{sd}} \left[ \phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) - \phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) \right]^2
\]

Using the relations [A.11.1], we find that

\[
I_{11}^{11} = \left\{ \frac{2(n_{ds} + n_{ss})^3}{n_{ds}n_{ss}} \phi^2 \left( \frac{\hat{\tau}}{\sqrt{2}} \right) \right\}^{-1}
\]

The variance of $\hat{\tau}$ is therefore seen to depend only on same-samples. The variance of $\hat{\delta}$ on the other hand, depends on both same and different samples

\[
I_{22}^{22} = w^{-2} \left\{ \frac{n_{sd}n_{dd}}{N_d^3} + \frac{n_{ss}n_{ds}}{N_s^3} \frac{v^2}{w^2u^2} \right\}
\]

where, following the notation of [Bi (2002)], we have that

\[
w = \frac{1}{\sqrt{2}} \left\{ -\phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) + \phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) \right\}
\]

\[
v = \frac{1}{\sqrt{2}} \left\{ \phi \left( \frac{\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) + \phi \left( \frac{-\hat{\tau} - \hat{\delta}}{\sqrt{2}} \right) \right\}
\]

\[
u = \sqrt{2} \phi \left( \frac{\hat{\tau}}{\sqrt{2}} \right)
\]

Another equivalent way to arrive at these variance and standard error estimates are via the delta method (eg. Pawitan, 2001) as utilized by [Bi (2002)]. Note that at $\hat{\delta} = 0$ and $\hat{\delta} = \infty$, $w = 0$ leaving $I_{22}^{22}$ undefined.
A.12 Likelihood root statistic and \( p \)-values

The well known Wald statistic is given as

\[
t(\theta) = \frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} .
\]

The assumption is that approximately \( t(\theta) \sim N(0,1) \), which under the null hypothesis \( H_0 : \theta = \theta_0 \) leads to the familiar formula for the \( p \)-value

\[
p = 1 - \Phi\{t(\theta_0)\}.
\]

We have previously seen Wilk’s likelihood ratio statistic

\[
w(\theta) = 2\left\{ l(\hat{\theta}) - l(\theta) \right\},
\]

which is asymptotically distributed as \( w(\theta) \sim \chi^2_m \), where \( m \) is the dimension of \( \theta \). For scalar \( \theta \), the likelihood root statistic (see eg. Severini 2000; Brazzale et al. 2007) is simply defined as the signed square root of \( w(\theta) \)

\[
r(\theta) = \text{sign}(\hat{\theta} - \theta) \left[ 2\left\{ l(\hat{\theta}) - l(\theta) \right\} \right]^{1/2},
\]

which is asymptotically distributed as \( r(\theta) \sim N(0,1) \), and the \( p \)-value is defined similarly as for the Wald statistic.

If interest is in \( \delta \), the likelihood root statistic is given by

\[
r(\theta) = \text{sign}(\hat{\delta} - \delta) \left[ 2\left\{ l(\hat{\delta}, \hat{\tau}; y) - l_\tau(\delta; y) \right\} \right]^{1/2},
\]

where \( l_\tau(\delta; y) \) is the profile likelihood of \( \delta \). The likelihood root statistic for \( \tau \) is defined similarly.

Often and traditionally we wish to take \( \theta_0 = 0 \), but in equivalence testing, we wish to take \( \theta_0 = \Delta \).

A.13 Simulation under the Same-Different Model

Given the information \( \delta, \tau, N_s, N_d \), where \( N_s \) and \( N_d \) are the number of same and different samples respectively, we have that

\[
n_{ss} \sim \text{Bin}(p_{ss}; N_s) , \quad n_{sd} \sim \text{Bin}(p_{sd}; N_d) .
\]
To simulate data from the same-different model, first sample $n_{ss}$ and $n_{sd}$ from the binomial distributions above where $p_{ss}$ and $p_{sd}$ are given in section A.3 and second, compute

$$n_{ds} = N_s - n_{ss}, \quad n_{dd} = N_d - n_{sd}.$$

### A.14 R-code used in the Manuscript

```{r}
### chunk number 1: contours
library(sensR)
(sadiC <- samediff(4, 8, 0, 12))
lim <- list(tau = c(1e-4, 2), delta = c(1e-4, 2), length = c(100, 100))
contour(sadiC, lim = lim)

### chunk number 2: contours2
plot(profile(sadiC), 2)

### chunk number 3: ContourAnna
require(sensR)
Anna <- samediff(8, 5, 4, 9)
lim <- list(tau = c(1e-4, 4), delta = c(1e-4, 4), length = c(100, 100))
contour(Anna, norm = TRUE, lim = lim)

### chunk number 4: ContourAnna2
contour(Anna, norm = FALSE, lim = lim)

### chunk number 5: ProfileAnna
par(mfrow = c(1, 2))
x <- plot(profile(Anna), fig = FALSE)
lim <- sapply(c(0.95, .99), function(x) exp(-qchisq(x, df=1)/2) )
head(x)

## tau:
plot(x$tau, x$nplTau, type = "l", las = 1,
    xlab = expression(tau),
    ylab = "Normalized Profile Likelihood",
    main = "")
abline(h = lim, col = "grey")

## Normal approximation:
Tau <- coef(Anna)[1]
lines(x$tau, exp(-((x$tau - Tau)^2/(2 * vcov(Anna)[1, 1]))) , lty = 3)

## delta:
plot(x$delta, x$nplDelta, type = "l", las = 1,
    xlab = expression(delta),
    ylab = "Normalized Profile Likelihood",
    main = "")
```
```r
xlab = expression(delta),
ylab = "Normalized Profile Likelihood",
main = "")
abline(h = lim, col = "grey")
## Normal approximation:
delta <- coef(Anna)[2]
lines(x[,3], exp(-((x[,3] - delta)^2/(2 * vcov(Anna)[2, 2]))), lty = 3)
## Estimated likelihood
e11 <- sapply(x[,3], function(delta)
  sensR:::llSameDiff(Tau, delta, ss=8, ds=5, sd=4, dd=9)
) nelTau <- exp(e11 - max(e11))
lines(x[,3], nelTau, lty = 2)
par(mfrow = c(1, 1))

###################################################
### chunk number 6: ProfileMark
###################################################
Mark <- samediff(8, 5, 11, 2)
plot(profile(Mark))

###################################################
### chunk number 7: misc eval=FALSE
###################################################
## Value of the log-likelihood for Anna's data:
## logLik(Anna)
## Profile likelihood for Anna's data at delta = 1:
## optimize(sensR:::llSameDiff, c(1e-4, 10), delta = 1, ss=8, ds=5, sd=4, dd=9, maximum = TRUE)

###################################################
### chunk number 8: misc2 eval=FALSE
###################################################
## (sadi <- samediff(8, 5, 4, 9))
## plot(profile(sadi))
## (mat <- matrix(c(8, 5, 4, 9), 2))
## chisq.test(mat, sim = TRUE, B = 1e6, alternative = "greater")
## fisher.test(mat, alternative = "greater")
## fisher.test(mat)
## (sadi2 <- samediff(8, 5, 11, 2))
## plot(profile(sadi2))
## (mat2 <- matrix(c(8, 5, 11, 2), 2))
## chisq.test(mat2, sim = TRUE)
## fisher.test(mat2, alternative = "less")
##
```
Appendix B

Estimation and Inference in the A-Not A test with Sureness

Manuscript for Food Quality and Preference
B.1 Abstract

In this paper we show that the A-Not A test with sureness can be identified as a multivariate generalized linear model known as a cumulative probit model in the statistical literature. A generalization of the model allows for different scales of the distributions of sensory intensity and is an extension of the binormal model known from signal detection theory.

We embed the A-Not A test with sureness in a regression framework which allows for general predictor variables in describing the location (d-prime) and scale of the distributions of sensory intensity. The framework is capable of handling several products, as well as explanatory variables such as gender and age as well as habits regarding the use of the product - all in the same model much like in usual linear models with normally distributed response. This facilitates assessment of the influence of these variables on the discriminability. We also show how it is possible within the proposed framework to assess learning and fatigue effects.

As an illustration, we analyse data from a study on discrimination of packaged soup conducted by Unilever Research.

As a last issue we discuss the relation of the proposed model to ROC (Receiver Operating Characteristic) curves and how changes in the model may change the ROC curve and AUC (Area Under the Curve).

We provide functions in our free R-package SensR that will fit the models and provide estimates, standard errors and nice plotting features etc.

B.2 Introduction

The A-Not A test, possibly with sureness is a commonly used discrimination and preference test. The test is also called the “yes-no” test in the signal detection theory (SDT) literature (eg. Macmillan and Creelman [2005]) and is a variant of the ratings method. Experiments with these tests are often analyzed with SDT-methods via receiver operating characteristic (ROC) curves. We show in this article how data from such tests can be analyzed efficiently and how learning and fatigue effects can be assessed. We also show how several test products can be
handled in the same model and how respondent and experiment characteristics can be included in the analysis. The extensions we propose greatly extends the scope of inferences possible from these types of data.

We develop the models based on a probabilistic and Thurstonian approach to sensory perception and are based on the law of categorical judgment (Torgerson, 1958; Luce, 1994)—a variant of the law of comparative judgment (Thurstone, 1927a,b,c). The models we propose are all fitted with the method of maximum likelihood, so the estimates of the Thurstonian $\delta$ (denoted $d'$; d-prime in the majority of the SDT literature) are optimal in a certain statistical sense. We illustrate all methods and models on a dataset from a study of packet soup conducted by Unilever Food & Health Research Institute, Vlaardingen, Netherlands. We describe the data and study in section B.2.1.

The A-Not A protocol has previously been considered in a sensometric context by O’Mahony (1992); Bi and Ennis (2001b,a). Brockhoff and Christensen (2008) showed how the identification of the duo-trio, triangle, 2-AFC and 3-AFC protocols as generalized linear models with different link functions facilitates modelling of explanatory variables in the Thurstonian framework.

To the best knowledge of the authors no previously proposed models have been proposed that can allow for several products or for explanatory variables in the assessment of discriminability. We propose models that allow for both and provide functions in our sensR package for R (R Development Core Team, 2008a) that will fit the models.

We will consider six classes of models in this paper. First we show how the basic A-Not A test with two response categories can be identified as an ordinary generalized linear model with a binomial distribution and a probit link. We compare this with the SDT methods for analyzing “yes-no” tests. We extend the generalized linear model to allow for response scales with more than two response categories and identify such models as cumulative link models. We show that an extension of this model is the binormal model (Dorfman and Alf, 1969; Luce, 1994) is a special case of the cumulative link location-scale model (CLLS) with a probit link. Again we compare with a SDT method to analyze rating experiments.

To quantify the effects of the explanatory variables and examine the differences between the test products, we shall need the most advanced of the six models; the cumulative link location-scale (CLLS) model with a probit link (CPLS). This model is not readily estimated by standard statistical software, so we provide the R function cplls in our package.

We contrast these models and describe the advantages of using the likelihood
based approach over the SDT-methods. We motivate all models with direct reference to Thurstones model assuming normal distributions of sensory intensity.

All analyzes in this paper are performed with functions in sensR or with standard R functions. We have made available the code used in the paper at www.url.com.

The paper proceeds as follows. In section B.3 we consider methods for analyzing A-Not A data without sureness. We also establish the notation we use throughout the paper. In section B.4 we consider simple methods and models for A-Not A data with sureness. In section B.5 we extend the methods and models of the previous section and allow for several test products and explanatory variables. In this section we also finish our analysis of the soup data. In section B.6 we describe and show how the methods here are related to ROC (Receiver Operating Characteristic) curves and AUC (Area Under the ROC Curve) and show how these measures can be obtained from the fitted models. We end with a discussion and summary in section B.9.

B.2.1 Description of Data

Unilever Research conducted a discrimination study of packet soup in a A-Not A test with sureness.

Five test products were compared against a reference product by 187 respondents. Prior to testing, respondents were familiarized with the reference product. In each trial the respondent was presented with a sample and asked to place an answer on the following response scale

<table>
<thead>
<tr>
<th>'Reference'</th>
<th>'Not Reference'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sure</td>
<td>Not Sure</td>
</tr>
</tbody>
</table>

Each respondent tasted a total of 10 products among which four were reference products. Each respondent therefore tasted one of the test products twice. Testing was split over two days to minimize sensory fatigue and performed on three different locations.

A range of explanatory variables were obtained describing the respondents and the experiment. Variables in the first category comprise the age of the respondents, gender, the type of soup regularly consumed (home made, canned or dry mix), the frequency with which soup is usually consumed and how easy the respondent found the testing. Variables specific to the experiment comprise the
sequence in which the testing was performed for each of the individuals (1–10), the location (one of three locations) and the day (one of two).

Some of the aims of the study were to quantify the effect of the explanatory variables and to examine whether discrimination was different for the five test products. Respondents are assumed to be approximately representative for the population of interest, so results from this study are expected to generalize well to the population of interest.

B.3 A-Not A without Sureness

In this section we consider models for A-Not A test data, where the response can fall in two categories. We give examples using likelihood based and SDT approaches.

![Simple Thurstonian model with distribution sensory intensity from reference (solid) and test (dashed) products.](image)

Figure B.1: Simple Thurstonian model with distribution sensory intensity from reference (solid) and test (dashed) products.

The Thurstonian model assumes that the sensory intensity of two products, say reference and test products, is normally distributed (cf. figure B.1). If the sensory intensity of the two products differ with respect to some attribute, then it may be possible to distinguish the two products.
In the A-Not A test, it is assumed that respondents adopt a decision rule defined by a threshold, $\theta$ as depicted in figure B.1. If the sensation of a sample presented to a respondent is of less magnitude than $\theta$, the respondent will respond “Reference product”. If the sensation is of larger magnitude than $\theta$, the respondent will respond “Test product”.

We can represent this strategy mathematically as follows: The sensory intensity, $S$ is a normal random variable, $S \sim N(\mu_k, 1^2)$, where $k = 1, 2$ indicates the reference and test products respectively and $\mu_k$ are the means of the distributions of sensory intensity. For identifiability we take $\mu_1 = 0$ and seek to estimate the parameter $\mu_2 = \delta$. Because the distributions themselves are unobservable, the magnitude of their standard deviation, or scale as we shall call it, is unknown. For simplicity we take it to be unity, $\sigma = 1$. Later we will allow for the scales of the distributions to differ. Note that the unit on the horizontal axis is $\sigma$, so we express the magnitude of $\delta$ in terms of the variability, $\sigma$—a signal to noise ratio. This formulation amounts to assuming the following linear model for the sensory intensity of the $i$th sample on the $k$th product

$$S_{ik} = \mu_k + \epsilon_{ik}, \quad \epsilon_{ik} \sim N(0, 1^2) \quad \text{(B.3.1)}$$

We do not observe $S$ directly, but only the dichotomized version; $Y$ of the variable. The random variable $Y$ can take two possible values; “Ref” or “Test” and follows a Bernoulli distribution. Suppose we have $i = 1, \ldots, n_k$ samples of the $k$th type, then we may write

$$Y_{ik} \sim \text{Bernoulli}(\pi_k)$$

where $\pi_k = P(Y = "Ref" | \mu_k)$ is the probability of observing $Y = "Ref$$, which depends on the type of product being tested through $\mu_k$.

The observed value of $Y$ depends on whether $S$ is above or below $\theta$. We observe

$$Y = "Ref" \quad \text{if} \quad -\infty < S < \theta$$
$$Y = "Test" \quad \text{if} \quad \theta < S < \infty$$

This means that the probability of observing “Ref” to a sample given we know whether it is a reference or test sample can be written

$$P(Y = "Ref" | \mu_k) = \pi_k = P(S \leq \theta | \mu_k) = \Phi(\theta - \mu_k)$$

where $\Phi$ denotes the standard normal CDF. Rewriting this slightly, leads to the following formulation

$$Y_{ik} \sim \text{Bernoulli}(\pi_k) \quad \Phi^{-1}(\pi_k) = \theta - \mu_k \quad \text{(B.3.2)}$$
where $\Phi^{-1}$ is the inverse standard normal CDF, also known as the probit function. This can be identified as a generalized linear model (GLM) with a Bernoulli distribution and a probit link \cite{McCullagh1989, Brockhoff2008}. Recall that we imposed a constraint on $\mu_k; \mu_1 = 0$. Also note the minus in the right-hand-side of the model \eqref{eqn:B3.2}. Most software that fits generalized linear models assume a plus. Consequently, the estimate of $\delta$ is a parameter estimate from the probit GLM with the sign reversed.

**Example** We now illustrate the analysis of the soup data with the generalized linear model. We also compare the analysis with the standard SDT method.

We observed responses in six response categories and for six different products for the soup data. We group the first three and the last three response categories and treat the 5 test products as the same product. We obtain the aggregated data in table \ref{tab:B.1}.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Response</th>
<th>&quot;Ref&quot;</th>
<th>&quot;Not Ref&quot;</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td></td>
<td>358</td>
<td>381</td>
<td>739</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>245</td>
<td>863</td>
<td>1108</td>
</tr>
</tbody>
</table>

Fitting the data in table \ref{tab:B.1} in the generalized linear model \eqref{eqn:B3.2} leads to the following estimates. $\hat{\theta} = -0.039$ and $\hat{\delta} = 0.729$. We used these estimates to make figure \ref{fig:B.1}. Their standard errors are standard output from generalized linear models and they are given as $\text{se}(\theta) = 0.046$ and $\text{se}(\delta) = 0.062$. The standard error of $\theta$ is not of much interest, but that of $\delta$ provides strong evidence, that $\delta$ is positive and quite certainly between 0.5 and 1. Note that the estimates of $\theta$ and $\delta$ are maximum likelihood estimates (MLEs) because generalized linear models are fitted with the method of maximum likelihood (ML).

Signal detection theory (SDT) also provides a way to analyze such data \cite{Macmillan2005}, which in fact leads to exactly the same estimates albeit no standard errors or other measures of uncertainty of the estimates. SDT defines the Hit rate as probability of correctly identifying a test product as “not reference”. The False Alarm rate is the probability of falsely identifying a reference product as “not reference”, which we can state as

$$\text{Hit rate} = 1 - \pi_2$$
$$\text{False Alarm rate} = 1 - \pi_1$$

The estimates of $\theta$ (which \cite{Macmillan2005} denotes $k$) and $\delta$ (denoted
The A-Not A with Sureness model
In the A-Not A test with sureness, respondents are required to supplement their classification with a sureness judgment. They could for instance be required to place their answer on the response scale indicated in table B.2.

With $j = 1, \ldots, J = 6$ response categories, there are $J - 1 = 5$ thresholds; $\theta_j$. Note that this is a direct extension of the case with only two response categories discussed above. The distribution of sensory intensity, $S$ remains $S \sim N(\mu_k, 1^2)$, but $Y$ is no longer Bernoulli distributed, as it can fall in each of $J$ categories. The response $Y$ can be represented by the vector $X = (X_1, \ldots, X_j, \ldots, X_{J-1})$, where $X_j = 1$ if $Y$ falls in the $j$th category. We therefore have the following relation

$$P(Y = j) = P(X_j = 1) = \pi_j,$$

and we say that $X$ is multinomially distributed;

$$X_{ik} \sim \text{Multinom}(\pi_k, m_{ik}),$$

where in our formulation, the index $m_{ik} = 1$. The density is the multivariate extension of the Bernoulli;

$$P(X = (x_1, \ldots, x_{J-1})) = \prod_{j=1}^{J} \pi_{j}^{x_j}$$

The parameter, $\pi$ is a vector of length $J - 1$ and $\pi_j$ is the probability that a response falls in category $j$. The probability of the last category is given by

$$\pi_J = 1 - \sum_{j=1}^{J-1} \pi_j.$$ 

Denote the thresholds

$$-\infty = \theta_0 < \theta_1 < \ldots < \theta_J = \infty$$

then we observe

$$Y = j \quad \text{if} \quad \theta_{j-1} < S < \theta_j$$

This means that the response $Y$ falls in category $j$ when the sensory intensity, $S$ is between the thresholds $\theta_{j-1}$ and $\theta_j$. We have shown the distributions of sensory intensity and the thresholds in figure B.2. It is convenient to express this via cumulative probabilities

$$P(Y \leq j | \mu_k) = P(S \leq \theta_j | \mu_k) = \Phi(\theta_j - \mu_k)$$

Note that this again is a direct extension of the case with two response categories. Now $\gamma_{jk}$ are cumulative probabilities defined by

$$\gamma_{jk} = P(Y \leq j | \mu_k) = \sum_{h=1}^{J} \pi_{hk}$$
Rewriting the model as in the previous section, we can identify the cumulative link model \citep{Agresti2002}, with a probit link function

\[
\Phi^{-1}(\gamma_{jk}) = \theta_j - \mu_k \tag{B.4.1}
\]

This is a multivariate generalized linear model \citep{Fahrmeir2001,McCullagh1989}, which most statistical software packages provides functions to fit. Note that some software assumes a plus rather than a minus in the linear predictor.

In the discussion of \citep{McCullagh1980}, \citep{Altham1980} noted the close relation between the cumulative models in (B.4.1) and the SDT model.

Pearson’s $\chi^2$-test provides a goodness of fit tests for the model (B.4.1). The test statistic is

\[
X^2 = \sum_{h=1}^{H} \frac{(o_h - e_h)^2}{e_h}, \tag{B.4.2}
\]

where $H$ is the number of data cells, $o_h$ are the observed cell counts and $e_h$ are the expected cell counts under the considered model. The statistic $X^2$ follows asymptotically a $\chi^2_m$-distribution with $m$ degrees of freedom, where $m = H - p$ and $p$ is the number of parameters in the model. The asymptotics are not good if some cell counts are small and a general rule is that all expected cell counts should be above 5 and preferably higher.

**Example** In this example we fit the model (B.4.1) to the soup data and evaluate the fit of the model. In the table B.2, we have pooled the data for different test products, but retained the full response scale.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Response</th>
<th>“Reference”</th>
<th>“Not Reference”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Sure</td>
<td>Not Sure</td>
</tr>
<tr>
<td>Ref</td>
<td></td>
<td>132</td>
<td>161</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>96</td>
<td>99</td>
</tr>
</tbody>
</table>

Note the very high proportion of responses in the last category for test products (cf. table B.2). Fitting the model (B.4.1) to the data in table B.2 gives the estimates in table B.3. We used these estimates to produce figure B.2. We defer comparison with the SDT approach to handle such data to after we have treated the scale issue.

The estimate of $\delta$ and its standard error is slightly lower than that from the probit GLM in the previous section and the $p$-value and confidence interval show
compelling evidence of a positive \( \delta \). The \( p \)-values of the thresholds is not of much use, since there is no reason they should be zero. The confidence intervals of the thresholds does however contain some information. The confidence intervals for the third and forth thresholds overlaps slightly reflecting that there is less evidence about their separation than for the remaining thresholds. This occurs because relatively few data were obtained in the third and fourth categories (cf table [B.2]) and because the distance between the thresholds is small. We could consider pooling data from these two categories and then fit the model with only four thresholds. For ease of comparison with other models, we retain the five intercepts. In situations with little information on the separability of response categories, it is a good idea pool data in these categories to improve identifiability of the model.

Table B.3: Cumulative link model fit to soup data with 95% Wald confidence interval.

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>( \delta )</td>
<td>0.669</td>
<td>0.053</td>
<td>0.564</td>
<td>0.773</td>
<td>1.14e-36</td>
</tr>
<tr>
<td>Thresholds</td>
<td>( \theta_1 )</td>
<td>-0.821</td>
<td>0.047</td>
<td>-0.912</td>
<td>-0.729</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_2 )</td>
<td>-0.270</td>
<td>0.043</td>
<td>-0.354</td>
<td>-0.186</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_3 )</td>
<td>-0.078</td>
<td>0.042</td>
<td>-0.161</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_4 )</td>
<td>0.074</td>
<td>0.043</td>
<td>-0.009</td>
<td>0.157</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_5 )</td>
<td>0.479</td>
<td>0.044</td>
<td>0.394</td>
<td>0.564</td>
<td></td>
</tr>
</tbody>
</table>

We can use Pearson’s \( \chi^2 \) test to evaluate the goodness of fit of this model to these data. The observed value of the test statistic is \( X^2 = 20.2 \) which is highly significant on \( 12 - 6 = 6 \) degrees of freedom. The model (B.4.1) therefore appears not to be appropriate for the soup data. The sureness rating provided information facilitating evaluation of goodness of fit of the model. This means that model (B.3.2) also is inappropriate for these data—an assumption we were not immediately able to assess without the sureness rating. In the following we shall see that allowing for differences in scale for the distributions of sensory intensity greatly improves the fit.

B.4.1 Allowing for Differences in Scale

In this section we extend the previous model (B.3.1) to allow for different scales, ie. different standard deviations for the distributions of sensory intensity for reference and test products.

The distributions of reference and test products can be viewed as distributions
of noise and signal + noise respectively \cite{O'Mahony1992}. If there is some (neural and/or physical) variation associated with the signal, then the variation of the sensory intensity of test products will be larger than that of reference products and the scale of the distribution of test products will be larger than that of reference products.

To accommodate this, we can allow the standard deviation or the scale of the distribution of \( S \) to depend on the product type as well, i.e. we assume \( S \sim N(\mu_k, \sigma_k^2) \), which we can write as a heterogeneous linear model

\[
S_{ik} = \mu_k + \epsilon_{ik}, \quad \epsilon_{ik} \sim N(0, \sigma_k^2)
\] (B.4.3)

Then the cdf of \( S \) is \( \Phi((s - \mu_k)/\sigma_k) \). The absolute magnitude of the scale is not estimable, but differences between scales are. We therefore take the scale of reference products as one, \( \sigma_1 = 1 \) and estimate \( \sigma_2 \) relative to that.

The resulting model can be written

\[
\Phi^{-1}(\gamma_{jk}) = \frac{\theta_j - \mu_k}{\sigma_k} \quad j = 1, \ldots, J - 1 = 5, \quad k = 1, 2.
\] (B.4.4)

This model is known as the binormal model, for which \cite{Dorfman1969, Grey1972} proposed algorithms to obtain ML estimates. This model is no longer in the class of (multivariate) generalized linear models, because the predictor is non-linear in the scale parameters. The binormal model (B.4.4) is a special case of the more general cumulative logit location-scale model (CLLS) with a probit link, which allows for a general predictor for location and scale parts. We consider this CLLS model in section B.5, where we expand the model to allow for several test products and explanatory variables. Not all software packages will fit the model (B.4.4), so we provide the function \texttt{clls} in \texttt{R} package \texttt{sensR} that will fit the model (B.4.4) and generalizations of it that we will consider in later sections.

For computational purposes, we work with \( \tau_k = \log \sigma_k \) to force the scale to be positive. We have found the profile likelihood of \( \tau \) to be closely approximated by a Gaussian curve in contrast to that of \( \sigma \), which is often skewed (cf. figure B.5). We therefore obtain appropriate confidence intervals for \( \sigma \) by constructing the normal-based Wald confidence interval for \( \tau \) and then afterwards transforming it to the \( \sigma \)-scale.

We can evaluate the evidence in the data about two competing and nested models via the likelihood ratio. We can for instance compare the models (B.4.4) and (B.4.1) with and without scale to examine the evidence that \( \sigma_2 \) is different from unity or equivalently that \( \tau_2 \) is different from zero. The likelihood ratio statistic is the difference in twice the log likelihoods, for the models. Twice
the negative log likelihood of a model is called the deviance possibly controlled for the maximum achievable likelihood (McCullagh and Nelder, 1989; Pawitan, 2001). The likelihood ratio statistic is therefore also the difference in deviance between the two models. Asymptotically the difference in deviance follows a $\chi^2$ distribution with $r$ degrees of freedom

$$\Delta\text{Dev} \sim \chi^2_r,$$

where $r$ is the difference in the number of parameters of the two models (see eg. Pawitan, 2001).

Example In this example we continue the analysis of the soup data. First we apply a generalization of the SDT method considered in the example in section B.3 for rating experiments. Thereafter we fit the binormal model (B.4.4) to the soup data.

The SDT approach is to calculate cumulative proportions (Macmillan and Creelman, 2005), calculate “z-scores”, ie. apply the inverse standard normal cdf also denoted “empirical probit transform” to obtain cumulative Hit and False Alarm rates. These rates for our soup data are shown in the table B.4. With several estimates of $\delta$, the problem arises of how to obtain a single adequate measure of discriminability. This is where the SDT method will no longer provide MLEs and therefore leads to less optimal estimates. One notable feature of table B.4 is that the estimates of $\delta$ are roughly increasing with category. This indicates that the scales of the distributions of sensory intensity no longer share the same scale.

Table B.4: Probit transformed cumulative Hit and False Alarm rates and SDT estimates of $\hat{\delta}$

<table>
<thead>
<tr>
<th>$\Phi($Hit rate$)$</th>
<th>$\Phi($False Alarm rate$)$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$-0.921$</td>
<td>$-1.362$</td>
<td>$0.441$</td>
</tr>
<tr>
<td>$-0.262$</td>
<td>$-0.931$</td>
<td>$0.668$</td>
</tr>
<tr>
<td>$-0.039$</td>
<td>$-0.768$</td>
<td>$0.729$</td>
</tr>
<tr>
<td>$0.100$</td>
<td>$-0.605$</td>
<td>$0.705$</td>
</tr>
<tr>
<td>$0.535$</td>
<td>$-0.219$</td>
<td>$0.754$</td>
</tr>
</tbody>
</table>

An ad-hoc solution is to obtain an estimate of the scale by regressing the transformed Hit rate on the transformed False Alarm rates;

$$
\Phi(\text{Hit rate}) = a + b\Phi(\text{False Alarm rate}) + \epsilon
$$

and use the estimated slope as an estimate of the ratio of scales, ie. $\hat{b} = \sigma_2/\sigma_1$. This regression is denoted the zROC curve and we have shown it for the soup.
data in figure B.3 (left). To see that \( \hat{b} \) is not an optimal estimate, suppose instead we had regressed the False Alarm rate on the Hit rate. That would have lead to another, but equally valid estimate, \( \tilde{b} \); it matters whether the errors are assumed on the Hit rate or on the False Alarm rate.

\[
\begin{array}{c}
0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
0.0 & 0.2 & 0.4 & 0.6 & 0.8 & 1.0 \\
\end{array}
\]

\[
\begin{array}{c}
\text{True positive ratio} \\
\text{False positive ratio} \\
\end{array}
\]

\[
\begin{array}{c}
z(\text{Hit Rate}) \\
z(\text{False Alarm rate}) \\
\end{array}
\]

Figure B.3: zROC and ROC curves for the soup data

The binormal model can be fitted with our function \texttt{cls} and we refer the reader to the online code to see how this can easily be done. Fitting the binormal model to the soup data, we obtain the results in table B.5. The distributions of sensory intensity based on these estimates are shown in figure B.4.

Table B.5: Binormal model fit to soup data with 95\% Wald confidence interval.

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>( \delta )</td>
<td>0.827</td>
<td>0.077</td>
<td>0.676</td>
<td>0.977</td>
<td>1.81e-27</td>
</tr>
<tr>
<td>Thresholds</td>
<td>( \theta_1 )</td>
<td>-0.899</td>
<td>0.052</td>
<td>-1.002</td>
<td>-0.796</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_2 )</td>
<td>-0.293</td>
<td>0.044</td>
<td>-0.379</td>
<td>-0.207</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_3 )</td>
<td>-0.080</td>
<td>0.043</td>
<td>-0.164</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_4 )</td>
<td>0.090</td>
<td>0.043</td>
<td>0.005</td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>( \theta_5 )</td>
<td>0.548</td>
<td>0.048</td>
<td>0.454</td>
<td>0.643</td>
<td></td>
</tr>
<tr>
<td>Scale</td>
<td>( \tau_2 )</td>
<td>0.217</td>
<td>0.061</td>
<td>0.096</td>
<td>0.337</td>
<td>2.07e-04</td>
</tr>
<tr>
<td></td>
<td>( \sigma_2 )</td>
<td>1.242</td>
<td>1.101</td>
<td>1.401</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The estimated scale clearly makes a notable difference in the appearance of the Thurstonian model.

The difference in deviance (cf. (B.4.5)) between the models (B.4.1) and (B.4.4)
Figure B.4: Thurstonian-binormal model with distribution sensory intensity from reference (solid) and test (dashed) products.

Figure B.5: Profile likelihoods for the $\tau = \log \sigma$ and $\sigma$ parameterizations.
without and with scale respectively is 12.5. On one degree of freedom, this is highly significant providing compelling evidence of a difference in scale. The confidence intervals in table \[B.5\] and the profile likelihoods of \(\tau\) and \(\sigma\) also provide clear evidence that the scale parameter is needed.

Using Pearson’s \(\chi^2\) test to evaluate the goodness of fit of this model, we find that the observed value of the test statistic is \(X^2 = 7.37\) which is not significant compared to a \(\chi^2\)-distribution with \(12 - 7 = 5\) degrees of freedom. The model \[B.4.4\] therefore appears to be appropriate for the soup data.

### B.5 Analyzes with Explanatory Variables

#### B.5.1 Allowing for Several Products

Until now, we have analyzed data as if only one reference and one test product were tested. Identification of the Thurstonian model for the A-not A test with sureness as a cumulative link location-scale model with a probit link allows generalization of the predictor (ie. the right hand side of model \[B.4.4\]). We first consider how to handle several products in the same model and later extend to allow for additional explanatory variables.

Suppose that the test comprised several, say \(l = 1, \ldots, L\) products, where \(l = 1\) refers to the reference product and \(l = 2, \ldots, L\) denotes different test product formulations. Let \(k = 1\) denote the reference product and \(k = 2\) for \(l > 1\) denote test products. Then we can allow for different locations of the distributions of sensory intensity of the test products in the following model

\[
\Phi^{-1}(\gamma_{jl}) = \theta_j - \mu_l \exp(\tau_k),
\]  

\[ \text{[B.5.1]} \]

With several test products, there may be no reason to expect some signals to vary more than other signals. This assertion can be assessed by comparing the deviance of model \[B.5.1\] and the following model where \(L - 1\) scales are estimated.

\[
\Phi^{-1}(\gamma_{jl}) = \frac{\theta_j - \mu_l}{\exp(\tau_k)}.
\]

**Example** In the soup experiment, five similar but different formulations of the test products were considered, hence \(L = 6\). The following analysis will
B.5 Analyzes with Explanatory Variables

consider whether there are any important differences between the test products. The analysis will be performed within a single model thereby utilizing all the information in the data rather than making five small models for each of the test products.

We begin by considering model (B.5.1) which assumes that the distributions of sensory intensity for the six products differ with respect to location of the mean, but that the five test products share the same scale. The fitted model is shown in table B.6.

Table B.6: Cumulative probit location-scale model fit to soup data with 95% Wald confidence interval.

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>$\delta_1$</td>
<td>0.642</td>
<td>0.091</td>
<td>0.464</td>
<td>0.821</td>
<td>8.91e-13</td>
</tr>
<tr>
<td></td>
<td>$\delta_2$</td>
<td>1.030</td>
<td>0.130</td>
<td>0.775</td>
<td>1.286</td>
<td>1.44e-15</td>
</tr>
<tr>
<td></td>
<td>$\delta_3$</td>
<td>0.601</td>
<td>0.115</td>
<td>0.376</td>
<td>0.827</td>
<td>8.76e-08</td>
</tr>
<tr>
<td></td>
<td>$\delta_4$</td>
<td>0.912</td>
<td>0.126</td>
<td>0.666</td>
<td>1.159</td>
<td>2.05e-13</td>
</tr>
<tr>
<td></td>
<td>$\delta_5$</td>
<td>1.138</td>
<td>0.135</td>
<td>0.875</td>
<td>1.402</td>
<td>1.32e-17</td>
</tr>
<tr>
<td>Thresholds</td>
<td>$\theta_1$</td>
<td>-0.899</td>
<td>0.052</td>
<td>-1.001</td>
<td>-0.796</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>-0.294</td>
<td>0.044</td>
<td>-0.380</td>
<td>-0.208</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>-0.080</td>
<td>0.043</td>
<td>-0.165</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>0.090</td>
<td>0.043</td>
<td>0.005</td>
<td>0.175</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_5$</td>
<td>0.549</td>
<td>0.048</td>
<td>0.454</td>
<td>0.643</td>
<td></td>
</tr>
<tr>
<td>Scale</td>
<td>$\tau_2$</td>
<td>0.202</td>
<td>0.061</td>
<td>0.082</td>
<td>0.322</td>
<td>4.89e-04</td>
</tr>
<tr>
<td></td>
<td>$\sigma_2$</td>
<td>1.224</td>
<td>1.085</td>
<td>1.380</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There seems to be some difference between the test products. To evaluate the evidence that there is a difference between test products we can perform a likelihood ratio test of the models (B.4.4) and (B.5.1). The likelihood ratio statistic on four degrees of freedom is 19, which provides strong evidence in favour of different means for the test products (p-value < .0001).

Other possible extensions of model (B.4.4) allows for a difference in scale among all the test products rather than in location, or difference in both location and scale among all test products. We found that the first of these extensions provided only minor and insignificant improvement to model (B.4.4) and that the latter extension provided only minor and insignificant improvement to model (B.5.1). This corresponds well with our expectation that different, but similar test products may produce different discriminabilities, but are less likely to produce different scales. In general more information (eg. more observations) are needed in order to provide evidence for difference in scale compared to that of
difference in mean.

We have shown the distributions of sensory intensity predicted from the model (B.5.1) in figure B.6. The test products seems to consist of roughly two groups. The first group contains two product with $\hat{\delta}$ around 0.6. The second group contains three products with $\hat{\delta}$ around 1.

![Figure B.6: Distributions of sensory intensity as predicted from model (B.5.1) with different discriminability for the five test products (dashed). The five test products share a common scale, which is 1.22 times that of the reference product.](image)

B.5.2 Analysis with General Predictor

A natural extension to the models considered above, is to allow other variables in the models. These variables may describe the products (eg. concentration of some ingredient) respondents (eg. gender or consumption habits) or the experimental conditions (eg. location, time or round) such as those described for the soup dataset in section B.2.1.

A general model for the $i$th observation can be written as

$$\Phi^{-1}(\gamma_{ij}) = \frac{\theta_j - x'_i \beta}{\exp(z'_i \zeta)} ,$$

(B.5.2)

where $x'_i$ is a $p$-vector of variables for the $i$th observation with associated $p$-vector
of location parameters \( \beta \), \( z_i' \) is a \( q \)-vector of variables with associated \( q \)-vector of scale parameters \( \zeta \). Notice that the log-linear model for the scale makes the effects multiplicative.

It is helpful to think of the variables as one of three types. First there is the response variable, second we have the indicator variable of product type and third we have what we shall call explanatory variables. The explanatory variables of the soup study contains the subject specific variables; gender, age, soup type etc. and the experiment specific variable; order, day etc. Note that we do not include the indicator variable of true product type among the explanatory variables.

Explanatory variables in the location part of the model can either influence the location of an answer on the response scale or they can influence the discriminability of a product. If an explanatory variable interacts with the indicator variable of product type, then the explanatory variable affects the discriminability of the product. If the variable does not interact with the indicator variable of product type, then the explanatory variable affects the location of the response on the response scale, but not the discriminability.

In the framework of model (B.5.2) it is possible to consider the effect of successive product evaluations provided subjects are required to assess several times. Such effects are often referred to as fatigue or learning effects if the discriminability of subjects decrease or increase with the assessment number. It is however also possible that the response tends to higher or lower categories on the response scale during the assessments. Such a drift need not be associated with a change in discriminability.

**Example**  One of the aims of the soup experiment was to assess the effect of a range of explanatory variables in the data. In the following, we complete the analysis of the soup data.

We find that an appropriate model for the soup data is that displayed in table B.7.

This model allows for different \( \delta \)s for each of the test products and an effect of day, type of soup regularly consumed and the order of the assessments for each day on the location of the responses on the response scale. We also found that discrimination improved on the second day. We found that discrimination was 0.25 standard deviations higher on the second day than on the first day, and that this difference was the same for all test products.
Table B.7: Cumulative probit location-scale model fit to soup data with 95% Wald confidence interval.

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>Location</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Test prod 1</td>
<td>0.513</td>
<td>0.122</td>
<td>0.275</td>
<td>0.752</td>
<td>1.21e-05</td>
</tr>
<tr>
<td></td>
<td>Test prod 2</td>
<td>0.911</td>
<td>0.135</td>
<td>0.648</td>
<td>1.175</td>
<td>6.40e-12</td>
</tr>
<tr>
<td></td>
<td>Test prod 3</td>
<td>0.472</td>
<td>0.130</td>
<td>0.218</td>
<td>0.726</td>
<td>1.33e-04</td>
</tr>
<tr>
<td></td>
<td>Test prod 4</td>
<td>0.779</td>
<td>0.139</td>
<td>0.506</td>
<td>1.051</td>
<td>1.11e-08</td>
</tr>
<tr>
<td></td>
<td>Test prod 5</td>
<td>1.017</td>
<td>0.146</td>
<td>0.731</td>
<td>1.303</td>
<td>1.52e-12</td>
</tr>
<tr>
<td></td>
<td>Day 2</td>
<td>-0.381</td>
<td>0.246</td>
<td>-0.863</td>
<td>0.100</td>
<td>6.03e-02</td>
</tr>
<tr>
<td></td>
<td>Canned soup</td>
<td>-0.138</td>
<td>0.065</td>
<td>-0.265</td>
<td>-0.012</td>
<td>1.61e-02</td>
</tr>
<tr>
<td></td>
<td>Dry mix soup</td>
<td>0.134</td>
<td>0.083</td>
<td>-0.029</td>
<td>0.297</td>
<td>5.40e-02</td>
</tr>
<tr>
<td></td>
<td>Test prod, day 2</td>
<td>0.248</td>
<td>0.125</td>
<td>0.002</td>
<td>0.493</td>
<td>2.41e-02</td>
</tr>
<tr>
<td></td>
<td>Order, day 1</td>
<td>-0.141</td>
<td>0.028</td>
<td>-0.196</td>
<td>-0.086</td>
<td>2.77e-07</td>
</tr>
<tr>
<td></td>
<td>Order, day 2</td>
<td>-0.037</td>
<td>0.027</td>
<td>-0.090</td>
<td>0.017</td>
<td>8.86e-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>Thresholds</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>-1.480</td>
<td>0.115</td>
<td>-1.705</td>
<td>-1.254</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_2$</td>
<td>-0.871</td>
<td>0.109</td>
<td>-1.085</td>
<td>-0.658</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_3$</td>
<td>-0.656</td>
<td>0.108</td>
<td>-0.867</td>
<td>-0.445</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_4$</td>
<td>-0.484</td>
<td>0.107</td>
<td>-0.694</td>
<td>-0.274</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\theta_5$</td>
<td>-0.021</td>
<td>0.107</td>
<td>-0.230</td>
<td>0.188</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Parameter</th>
<th>log(Scale)</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test prod</td>
<td>0.185</td>
<td>0.061</td>
<td>0.066</td>
<td>0.305</td>
<td>1.16e-03</td>
</tr>
<tr>
<td></td>
<td>Scale</td>
<td>1.204</td>
<td>1.068</td>
<td>1.356</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Five different $\delta$-values were found for each of the test products ranging from 0.51 for product 3 to 1.02 for product 6 on the first day of testing. All $\delta$-values are measured relative to the reference product and measured in units of the standard deviation of the distribution of reference products.

We found that consumers of canned soup voted 0.14 standard deviations below consumers of home made soup, and consumers of dry mix soup voted 0.13 standard deviation above. This means that consumers of dry mix soup generally voted highest and consumers of canned soup voted lowest. Consumers of home made soup were intermediate.

We also found an effect of the variables order and day. On the first day, the votes decreased with 0.14 standard deviations on each successive assessment. On the second day the votes also decreased albeit only with 0.04 standard deviations on each successive assessment. After five assessments the average vote had thus decreased with 0.7 standard errors on the first day, but only 0.2 standard errors on the second day. During testing on the first day, the votes decreased with an amount comparable to $\delta$ for several of the test products. The average
difference between the boundaries is 0.36, so the votes decreased about two response categories during the first day, but hardly one on the second day. The response criteria seemed to stabilize some between the two days. The votes on the beginning of the second day are 0.38 standard deviation lower than on the beginning of the first day. Already being lower than on the first day, it seems reasonable, that the votes does not decrease as much on the second day.

We did not find that discrimination changed within the days, so there is no there was no evidence of any learning of fatigue effects other than that between days. Neither did we find indications that discriminability was dependent on other subject or experimental specific variables.

**B.6 ROC analysis**

Recall that the Hit rate (HR) is probability of correctly identifying a test product as “not reference”. With several thresholds, and assuming the binormal model (B.4.4) holds, this can be expressed as

\[
\text{Hit Rate} = P(S > \theta_j | \text{test}) = 1 - P(S \leq \theta_j | \text{test}) = 1 - P(Y \leq j | \mu_2) = 1 - \gamma_{j2} .
\]

Similarly the False Alarm rate (FAR) can be expressed as

\[
\text{False Alarm Rate} = P(S > \theta_j | \text{ref}) = 1 - P(S \leq \theta_j | \text{ref}) = 1 - P(Y \leq j | \mu_1) = 1 - \gamma_{j1} .
\]

The parametric ROC curve is a plot of the Hit rate versus the False Alarm rate (Macmillan and Creelman 2005). This corresponds to varying the thresholds \( \theta_j \) and obtaining values of the two rates at each threshold. With a given set of data and model, the thresholds are fixed, but the ROC curve can also be obtained by expressing the Hit rate as a function of the False Alarm rate under the given model. Hit rates can then be obtained for any False Alarm rate. Under the binormal model, the thresholds can be expressed as

\[
\theta_j = \Phi^{-1}(1 - \text{FAR}) ,
\]

and putting this into the expression for the Hit rate yields

\[
HR = \Phi \left( \frac{\Phi^{-1}(\text{FAR}) + \delta}{\sigma_2} \right) .
\]

ROC curves are also defined for the more involved models with additional co-variates. Consider now a model, where the location depends on factor \( a_m \) with
$M$ levels and another factor $b_{ko}$ with $2 \times O$ levels, where $k$ indicates product type as before. The False Alarm rate is now given as

$$\text{FAR} = \Phi \left( \frac{\theta_j - \mu_1 - a_m - b_{1o}}{\sigma_1} \right),$$

and the Hit Rate can be expressed as

$$\text{HR} = \Phi \left( \frac{\Phi^{-1}(\text{FAR}) \sigma_1 - \mu_1 + \mu_2 - b_{1o} + b_{2o}}{\sigma_2} \right).$$

The factor $a_m$ cancels out in the expression for the Hit rate. This factor affects the location of the responses on the response scale and corresponds to moving the thresholds along the ROC curve; it does not affect the discriminability and therefore neither the shape of or area under the ROC curve. The factor $b_{ko}$ on the other hand changes the ROC curve, and different levels of $k$ corresponds to different curves, different discriminabilities and different areas under the curves. The effect of covariates in both location and scale parts was examined by Tosteson and Begg (1988). They also found that any variable in the scale whether or not it interacts with product type changes the ROC curve.

We have included the ROC curves for the 5 test products to illustrate the differences between products in figure B.7. Also shown is the difference in discrimination between the two days.

ROC curves are not only used in sensometrics and signal detection theory, but also in medical decision making, radiology and biostatistics. The pair (Hit rate, False Alarm rate) are terms seldom used in the latter areas. Instead they use the two pair terms (sensitivity, 1 - specificity) and (true positive ratio, false positive ratio) or simply as express the conditional probabilities directly. The motivation of and relations among these terms are explained by for instance Metz (1978).

The area under the ROC curve (AUC) is a commonly used measure of discrimination or sensitivity. Since the location parameter estimates are approximately Gaussian, we propose to compute the limits of confidence intervals on this scale and convert the confidence interval to the AUC scale afterwards. The AUC for the first test product is 0.658 with 95% confidence limits, (0.601, 0.711). When the AUC is closer to one, the asymmetry of the confidence interval will be more pronounced.

### B.7 On the Assumption of Normality

The assumption of normality of the latent distributions for the Thurstonian model has long since been a subject of discussion. In formulating his law of...
B.7 On the Assumption of Normality

Figure B.7: ROC curves for the 5 test products (left) and ROC curves for the fourth product (right) on the first (solid) and second (dashed) day.

comparative judgment, [Thurstone (1927a)] wrote that normality was a definition and the psychological scale an artificial construct. In [Thurstone (1927b)] he however expressed the view that the assumption of normality is experimental verifiable and that he has found the normal density to be applicable for some stimuli and not for others. [O’Mahony (1992)] describes normality as an assumption and discuss studies in which he claims the assumption was found not to hold. [Luce (1994)] express some reservation about the ability in practice to test the assumption.

As a definition, normality provides a common measure of discriminability and facilitates comparison of $\delta$-values for several unrelated experiments. The normal distribution is convenient because it is easy to work with, but there seems to be no other justification for assuming that the distributions of sensory intensity should be normal than the Central Limit Theorem.

Whether or not the Thurstonian interpretation of the cumulative link model is based on an assumption or a definition of normality, the cumulative link model does not rest on any such assumption. The cumulative probit model was motivated via the existence of normal latent distribution in section B.3 and B.4, but notice that any monotone transformation of $S$ would lead to the same stochastic ordering of the observations $Y$. Normality of $S$ is therefore not an assumption in fitting and interpreting the cumulative probit model. This holds for any other link function. In fact the assumption of existence of a latent distribution is not needed for appropriate application of the cumulative link model. It can be viewed merely as a statistical model for ordered categorical data. The likeli-
hood of the data is however dependent on the specific choice of link function and different links will lead to different fits and different predictions. A natural measure of fit is the residual deviance, which can be interpreted much like residual sums of squares in linear models.

Changing the link function amounts to assuming another distribution of sensory intensity. We have evaluated the residual deviance and Pearson’s $\chi^2$ goodness of fit statistic assuming four distributions of sensory intensity as shown in table B.8. The normal, logistic and Cauchy distributions are all symmetric, but the Gumbel (or type I extreme value distribution), which corresponds to the complementary log-log (c-log-log) link is skewed. The densities corresponding to these distributions are shown in figure B.8. The logistic and normal densities are much alike, the Cauchy has a thicker tail and the Gumbel is skewed. The cauchit link provides a bad fit compared to the remaining and is clearly inappropriate. Interestingly, the c-log-log link, assuming a skewed distribution of sensory intensity provides a fit close to that of the logistic, and slightly better than with the probit link. There is no formal statistical test, but the table does show, that there is little information on the precise shape of the latent distributions in the soup data.

Table B.8: Residual deviance and Pearson $\chi^2$ goodness of fit test assuming different latent distributions for the Binormal model fitted to the soup data.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Link Function</th>
<th>Residual Deviance</th>
<th>Pearson $X^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Probit</td>
<td>5377.55</td>
<td>7.37</td>
<td>0.39</td>
</tr>
<tr>
<td>Logistic</td>
<td>Logit</td>
<td>5375.49</td>
<td>5.33</td>
<td>0.62</td>
</tr>
<tr>
<td>Gumbel</td>
<td>c-log-log</td>
<td>5375.65</td>
<td>5.49</td>
<td>0.60</td>
</tr>
<tr>
<td>Cauchy</td>
<td>Cauchit</td>
<td>5401.26</td>
<td>11.57</td>
<td>0.12</td>
</tr>
</tbody>
</table>

B.8 Thresholds and Model Interpretation

When the link function in (B.4.1) is the logistic rather than the probit, the model is known as the proportional odds model (McCullagh [1980]). When the link is the complementary log-log link, the model is known as the grouped proportional hazard model (Cox [1972]).

In the context of the proportional odds model, (Peterson and Harrell Jr. [1990] proposed partial proportional odds. This is another generalization of the cumulative link model (cf. model [B.3.2]) which has a natural interpretation if the model is considered merely as a threshold model rather than a model for latent distributions. (Cox [1995] compared models with partial proportional odds and
Figure B.8: Densities for distributions corresponding to the four link function used in table B.8; Gumbel (solid), normal (dashed), logistic (dotted), Cauchy (dash-dot). All densities are in their standard versions, but the logistic has scale $\sqrt{3}/\pi$, so its standard deviation is unity.

Table B.9: Models considered in the paper and their properties.

<table>
<thead>
<tr>
<th>Model$^a$</th>
<th>−Sureness</th>
<th>+Sureness</th>
<th>Scale</th>
<th>Expl. var.$^b$</th>
<th>ML$^c$</th>
<th>Mult. Prod$^d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDT 2 cat.</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDT &gt; 2 cat.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Binormal</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLM</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>CPLS</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

$^a$: SDT = Signal Detection Theory method with two or more than two (with sureness) response categories. GLM: Generalized linear model. CP: Cumulative probit model. CPLS: Cumulative probit location-scale model.

$^b$: Indicates if the model can handle explanatory variables.

$^c$: Indicates if the model is estimated with the method of maximum likelihood.

$^d$: Indicates if the model can handle multiple products/samples.
the location-scale models and found little difference in fit although a difference in interpretation.

The model (B.4.1) specifies that the latent distributions (ie. the distributions of sensory intensity) have the same scales independent of the value of $\mu_k$. An equivalent way to express this is to say that the distance between the thresholds $\theta$ remains constant irrespectively of the value of $\mu_k$. The first interpretation makes a specific reference to the latent distributions and the latter does not. The model need not have an interpretation in terms of latent distributions, but can merely be a model for an ordinal variable. In such cases the latter formulation can be advantageous.

In the Thurstonian interpretation of the latent distributions, the thresholds are fixed, and we aim to draw conclusions regarding the shape of the distributions. In other applications it is the other way around; the distributions are fixed and the thresholds are variable. In the latter case, we model the difference between the thresholds $\theta$ and the model (B.4.1) says that the distances between the thresholds are constant for different $\mu$. The model (B.4.1) on the other hand says that the distances between the thresholds change in a systematic fashion.

Consider for example the case, where $k = 1, 2$ corresponding to two products in the setting of this paper. The thresholds for $k = 1$ (reference products) and $j = 1, \ldots, J - 1$ are $\theta_j / \sigma_1 = \theta_j$. The thresholds for $k = 2$ (test products) are $\theta / \sigma_2 = \theta^*_j$, where $|\theta^*_j| < |\theta_j|$ if $\sigma_2 > \sigma_1$ (ie. the scale for test products is larger than that for reference products). The thresholds $\theta^*_j$ are therefore shrunken towards zero relative to $\theta_j$.

The proportional odds model (ie. (B.4.1) with a logit link) results in proportional odds because the distances between the thresholds $\theta$ are constant. In the model (B.4.4), the distances between the thresholds are no longer constant, but changes with $\mu_k$, and the odds are no longer proportional. The location scale model (B.4.4) therefore provides a test for the assumption of proportional odds.

### B.9 Discussion and Summary

In this paper we have considered a series of different models for binary and ordinal data. These can originate from A-Not A (yes-no) tests and from rating experiments such as the A-Not A with sureness experiments.

A closely related type of experiment is the ranking experiment. This differs from the rating experiment in that response categories are defined by numbers, for
instance from 1 to the number of categories. A ranking scale amounts to assuming that the distance between the thresholds is known. Such data are known as interval data in contrast to the ordinal data considered in this paper. Often however, the assumption, that the distance between the thresholds is known is inappropriate. There may for example be some disagreement among the judges or respondents about these distances. In this case it can be informative to model the data as ordinal rather than interval and apply the methods treated in this paper.

The models considered in this paper are summarized with respect to their properties in table B.9. The models for models allowing for sureness are only applicable with three or more response categories to make any scale differences identifiable.

The identification of the Thurstonian models for the A-Not A protocol as generalized linear models or extensions thereof facilitates modelling of various explanatory variables much like in usual linear models (eg. ANOVA, Multiple Linear Regression and their synthesis, ANCOVA). A fuller description of the data is possible within the probabilistic Thurstonian framework.

The allowance for explanatory variables is a general route to assessment of important effects such as learning and fatigue effects as well as memory effects. These effects can now be assessed in formal statistical models and powerful tools such as likelihood ratio tests are readily applicable.

The identification of these Thurstonian models as certain types of statistical regression models is also a promising route to appropriate models to handle replications and this paper provides a basis for understanding these more involved models. Attempts of coping with replications for rating experiments (Ennis and Bi, 1999; Bi, 2006b) have focused on adjusting standard errors of parameters for overdispersion effects. GLMs and CLLS-models prepare the way for proper modelling of the covariance structure in the data in the face of replications via mixed effect versions of these models. This is an area of current research.

Appendix

B.10 R-code used in the Manuscript

---

```r
###################################################
### chunk number 1: init
```
### chunk number 1: data loading and preparation

```r
# Load data
load('D:/DTU/EksProjPBB/Analysis/tdat.RData')

# Load libraries
library(sensR)
library(MASS)
library(xtable)

# Preprocess data
tdatr <- tdat[tdat$RESP!=183 & tdat$RESP!=222,]
tdatr$order <- as.factor(tdatr$ORDER)
tdatr$DayProd <- with(tdatr, 1*(PROD == "Test" & DAY == 4))
head(tdatr, 10)

# Create contingency tables
(tab26 <- with(tdatr, table("Product" = PROD, "Response" = SURENESS)))
dimnames(tab26)[[2]] <- c("Sure", "Not Sure", "Guess", "Not Sure", "Sure")
dat26 <- expand.grid(sureness = as.factor(1:6), prod = c("Ref", "Test"))
dat26$wghts <- c(t(tab26))

(tab22 <- with(tdatr, table("Product" = PROD, "Response" = SUREBI)))
dimnames(tab22)[[2]] <- c("Ref", "Not Ref")
dat22 <- expand.grid(Resp = c("Ref", "Test"), prod = c("Ref", "Test"))
dat22$wghts <- c(t(tab22))
```

### chunk number 2: GLM models

```r
# Fit GLM model
glm1 <- glm(Resp == "Ref" ~ prod, data = dat22, weights = wghts,
            family = binomial(probit))
(sum1 <- summary(glm1))
fitted(glm1)
```

### chunk number 3: Thurstone Simple

```r
delta <- seq(-4, 5, len = 200)
plot(delta, dnorm(delta), ylab = "",
     xlab = "Sensory magnitude",
     type = "l", las = 1)
lines(delta, dnorm(delta, -coef(glm1)[2]), lty = 2)
abline(v = coef(glm1)[1])
axis(side = 3, at = coef(glm1)[1], labels = expression(theta))
text(-2.5, .2, "Reference
products")
text(3, .2, "Test
products")
axis(side = 1, at = -coef(glm1)[2],
     labels = expression(delta / sigma))
segments(-coef(glm1)[2], -.2, -coef(glm1)[2],
         max(dnorm(delta, -coef(glm1)[2])), lty = 3)
```

### chunk number 4: xtab22

```r
xtab22 <- cbind(tab22, "Total" = rowSums(tab22))
print(xtable(xtab22, digits = 0), only.contents = TRUE)
```

### chunk number 5: CPModel

```r
```
```r
# R-code used in the Manuscript

# chunk number 6: ThurstoneCP

cp <- clls(sureness ~ prod, scale = ~1, data = dat26,
weights = wghts, Hess = TRUE, method = "probit")

(sum.cp <- summary(cp))

# chunk number 7: xtab26

xtab26 <- cbind(tab26, "Total" = rowSums(tab26))

print(xtable(xtab26, digits = 0), only.contents = TRUE)

# chunk number 8: CLModelEstimate

### cp <- clls(sureness ~ prod, scale = ~1, data = dat26,
### weights = wghts, Hess = TRUE, method = "probit")

b <- coef(summary(cp))
pvals <- format(pnorm(abs(b[,3]), lower = FALSE), digits = 3)
pvals[-1] <- rep("", 5)
rnam <- c("$\delta$", "$\theta_1$", "$\theta_2$", "$\theta_3$",
"$\theta_4$", "$\theta_5$")
rnam2 <- c("Location", "Thresholds", NA, NA, NA, NA)
b2 <- data.frame(rnam2, rnam, b[-3], confint(cp), pvals)
colnames(b2) <- c("Type", "Parameter", "Estimate",
"Std. Error", "Lower", "Upper", "$p$-value")
b2

# Pearson's GOF statistic:
e <- cp$fitted.values[c(1, 7), ] * rowSums(tab26)
(Pearson <- sum((tab26 - e)^2 / e))

# chunk number 9: CLModel
```
cap <- "Cumulative link model fit to soup data with 95\% Wald confidence interval."
lab <- "tab:CLModel"
print(xtable(b2, caption = cap, label = lab, digits = 3),
      include.rownames = FALSE,
      sanitize.text.function = function(x){x},
      caption.placement = "top")

### chunk number 10: SDTcumTable

xtab <- SDT(tab26)
cap <- "Probit transformed cumulative Hit and False Alarm rates and SDT estimates of $\hat \delta$"
lab <- "tab:SDTcumTable"
colnames(xtab) <- c("$\Phi$(Hit rate)", "$\Phi$(False Alarm rate)", "$\hat \delta$"
print(xtable(xtab, digits = 3, caption = cap, label = lab),
      include.rownames = FALSE,
      sanitize.text.function = function(x){x},
      caption.placement = "top")

### chunk number 11: ROCzROCsoup1

SDT.tab <- SDT(tab26)
plot(SDT.tab[,2], SDT.tab[,1], ylim = c(-1.5, 1),
     xlim = c(-1.5, 1),
     xlab = "z(False Alarm rate)",
     ylab = "z(Hit Rate)"
abline(a = 0, b = 1)
(co <- coef(lm(SDT.tab[,1] ~ SDT.tab[,2])))
abline(co, lty = 2)

### chunk number 12: ROCzROCsoup2

## mean(SDT.tab[,1] - co[2] * SDT.tab[,2])
d1 <- mean((1/co[2]) * SDT.tab[,1] - SDT.tab[,2])
da <- sqrt(2 / (1 + co[2]^2)) * d2
s <- co[2]
ROC(d2, scale = co[2])

### chunk number 13: BinormalModel

Binormal <- clls(sureness ~ prod, ~ prod, data = dat26,
                 weights = wghts, Hess = TRUE, method = "probit")
(sum.Bin <- summary(Binormal))
b <- coef(sum.Bin)
fac <- qnorm(c(.025, .975))
CI <- b[,1] + b[,2] %o% fac
CI <- rbind(CI, exp(CI[7 ,]))
b <- rbind(b, exp(b[7 ,]))
(b2 <- format(cbind(b[-3], CI), digits = 3))
(pvals <- format(pnorm(abs(b[,3]), lower = FALSE), digits = 3))
pvals[1:7] <- rep("", 7)
b[8, 2] <- NA
rownames(b) <- 1:8
rnam <- c("\delta", "\theta_1", "\theta_2", "\theta_3",
"\theta_4", "\theta_5", "\tau_2", "\sigma_2")
rnam2 <- c("Location", "Thresholds", NA, NA, NA, NA, "Scale", NA)
b3 <- data.frame("Type" = rnam2, "Parameter" = rnam, b[-3,], CI,
pvals)
colnames(b3)[c(3:7)] <- c("Estimate", "Std. Error", "Lower", "Upper",
"$p$-value")
b3

## Pearson's GOF statistic:
e <- Binormal$fitted.values[ c(1,7),] * rowSums(tab26)
(PearsonBin <- sum((tab26 - e)^2 / e))
### chunk number 16: ProfileTau

```r
### chunk number 16: ProfileTau

```r
data26$Prod <- unclass(data26$prod) - 1

cp.sc <- clls(sureness ~ prod, scale = ~ prod, data = data26,
weights = wghts, method = "probit", Hess = TRUE)

summary(cp.sc)

anova.scale <- anova(cp, cp.sc)

N <- 20
dev <- double(N)

for(i in 1:N)

dev[i] <- clls(location = sureness ~ prod,
scale = ~ offset(int[i] * Prod),
method = "probit",
data = data26, weights = wghts)$deviance

spl <- spline(int, dev, n = 300, "natural")

plot(spl$x, exp(-(spl$y - min(spl$y))/2), type = "l",
xlab = expression(tau[2]), las = 1,
ylab = "Normalized Profile Likelihood")

## Add Normal approximation on tau-representation:

int2 <- seq(-.2, .6, len=1000)

lines(int2, exp(-(int2 - log(coef(cp.sc)[7]))^2 / (2 * vcov(cp.sc)[7, 7])),
lty = 2, col = "blue")

## Add confidence limits:

level <- c(0.95, 0.99)

lim <- sapply(level, function(x)

exp(-qchisq(x, df=1)/2) )

abline(h = lim, col = "grey")

### chunk number 17: ProfileSigma

```r
### chunk number 17: ProfileSigma

```r
### sigma-parameterization profile likelihood:

spl <- spline(int, dev, n = 300, "natural")

plot(spl$x, exp(-(spl$y - min(spl$y))/2), type = "l",
xlab = expression(sigma[2]), las = 1, ylim = c(.96, 1.6))

## Add Normal approximation:

var.sigma <- coef(cp.sc)[7]^2 * vcov(cp.sc)[7, 7]

lines(int, exp(-(int - log(coef(cp.sc)[7]))^2 / (2 * var.sigma)),
lty = 2, col = "blue")

## Add confidence limits:

abline(h = lim, col = "grey")

### chunk number 18: ProfileDelta eval=FALSE

```r
### chunk number 18: ProfileDelta eval=FALSE

```r
N <- 20
dev <- double(N)

int <- seq(-.05, .5, len=N)

for(i in 1:N)

dev[i] <- clls(sureness ~ offset(int[i] * Prod),
scale = ~ prod, data = dat, weights = wghts)$deviance

spl <- spline(int, dev, n = 300, "natural")

```
```r
## plot(spl$x, exp(-(spl$y - min(spl$y))/2), type = "1",
## xlab = expression(delta), las = 1,
## ylab = "Normalized Profile Likelihood")
## ## Add Normal approximation:
## int2 <- seq(.5, 1.2, len=1000)
## lines(int2, exp(-(int2 - coef(cp.sc)[1])^2 /
## (2 * vcov(cp.sc)[1, 1] )), lty = 2,
## col = "black")
## ## Add confidence limits:
## abline(h = lim, col = "grey")
## ## Normal approximation based on difference in deviance:
## cp3 <- cpls(sureness ~ 1, scale = ~ prod, data = dat,
## weights = wghts, hess = TRUE)
## (var.beta <- coef(cp)^2 / anova(cp, cp3)[[6]][2])
## lines(int2, exp(-(int2 - coef(cp)[1])^2 /
## (2 * var.beta)), lty = 3,
## col = "black")
##
##--------------------------------------------------
##### chunk number 19: SevProdMod
##--------------------------------------------------

cp.a <- clls(sureness ~ PRODID, ~ PROD, data = tdatr,
method = "probit", Hess = TRUE)
cp.a.sum <- summary(cp.a)
b <- coef(cp.a.sum)
fac <- qnorm(c(.025, .975))
CI <- b[,1] + b[,2] %o% fac
CI <- rbind(CI, exp(CI[11, ]))

(pvals <- format(pnorm(abs(b[,3]), lower = FALSE), digits = 3))
pvals[-c(1:5, 11)] <- rep("", 6)
b[12, 2] <- NA
rownames(b) <- 1:12
rnam <- c("$\delta_1$", "$\delta_2$", "$\delta_3$", "$\delta_4$",
"$\delta_5$", "$\theta_1$", "$\theta_2$", "$\theta_3$",
"$\theta_4$", "$\theta_5$", "$\tau_2$", "$\sigma_2$"
)
rnam2 <- c("Location", rep(NA, 4), "Thresholds", rep(NA, 4), "Scale", NA)
b3 <- data.frame("Type" = rnam2, "Parameter" = rnam, b[,-3], CI,
pvals)
colnames(b3)[c(3:7)] <- c("Estimate", "Std. Error", "Lower", "Upper",
"$p$-value")

## names(anova.a1a)
## anova.a1a[[6]][2]

cp.b <- clls(sureness ~ PROD, ~ PRODID, data = tdatr, method =
"probit", Hess = TRUE)
### cp.c <- clls(sureness ~ PRODID, ~ PRODID, data = tdatr, method =
### "probit", Hess = TRUE)
anova.a1a <- anova(cp.b, cp.a)
### anova(cp.a1, cp.b, cp.a, cp.c)

## names(anova.a1a)
## anova.a1a[[6]][2]

cp.c <- clls(sureness ~ PRODID, ~ PRODID, data = tdatr, method =
"probit", Hess = TRUE)
anova.a1a <- anova(cp.b, cp.a)

## names(anova.a1a)
## anova.a1a[[6]][2]

## names(anova.a1a)
## anova.a1a[[6]][2]

cap <- "Cumulative probit location-scale model fit to soup data
with 95\% Wald confidence interval."
lab <- "tab:SevProd"
print(xtable(b3, caption = cap, label = lab, digits = 3),
```

include.rownames = FALSE, #only.contents = TRUE,
sanitize.text.function = function(x){x},
caption.placement="top")

###################################################
### chunk number 21: ScalePRODID
###################################################
plot(delta, dnorm(delta), ylab = "", xlab = "Sensory magnitude",
type = "l", las = 1)
lines(delta, dnorm(delta, coef(cp.a)[1], cp.a$sigma), lty = 2)
lines(delta, dnorm(delta, coef(cp.a)[2], cp.a$sigma), lty = 2)
lines(delta, dnorm(delta, coef(cp.a)[3], cp.a$sigma), lty = 2)
lines(delta, dnorm(delta, coef(cp.a)[4], cp.a$sigma), lty = 2)
lines(delta, dnorm(delta, coef(cp.a)[5], cp.a$sigma), lty = 2)
abline(v = cp.a$theta)
lab <- c(expression(theta[1]), expression(theta[2]),
expression(theta[3]), expression(theta[4]), expression(theta[5]))
axis(side = 3, at = cp.a$theta, labels = lab)
text(-2.5, .2, "Reference\nproducts")
text(3.5, .2, "Test\nproducts")

###################################################
### chunk number 22: ModelInvolved
###################################################
### strt <-
c(0.5135, 0.9113, 0.4721, 0.7786, 1.0169, -0.3814, -0.1383, 0.1337,
0.2475, -0.1405, -0.0365, -1.4795, -0.8714, -0.6562, -0.484,
-0.0207, 1.2036)
cp.inv <- clls(location = sureness ~ PRODID + DAY +
SOUPTYPE + DAY:ORDER + DayProd,
scale = ~ PROD, method = "probit",
data = tdatr, Hess = TRUE)
(sum.inv <- summary(cp.inv))
b <- coef(sum.inv)
fac <- qnorm(c(.025, .975))
CI <- b[,1] + b[,2] %o% fac
nr <- nrow(b)
CI <- rbind(CI, exp(CI[nr,]))
b <- rbind(b, exp(b[nr,]))
pvals <- format(pnorm(abs(b[,3]), lower = FALSE), digits = 3)
pvals[-c(1:11, nr)] <- rep("", 6)
b[1:nr+1, 2] <- NA
rownames(b) <- 1:(nr+1)
rname <- c(paste("Test prod", 1:5), "Day 2", "Canned soup",
"Dry mix soup",
"Test prod, day 2", "Order, day 1", "Order, day 2",
"$\theta_{1}$", "$\theta_{2}$", "$\theta_{3}$", "$\theta_{4}$", "$\theta_{5}$", "Test prod", "Test prod")
rmam2 <- c("Location", rep(NA, 10), "Thresholds", rep(NA, 4),
"log(Scale)", "Scale")
b3 <- data.frame("Type" = rname2, "Parameter" = rname, b[,-3], CI, pvals)
colnames(b3)[c(3:7)] <- c("Estimate", "Std. Error", "Lower", "Upper",
"$p$-value")
b3
B.10 R-code used in the Manuscript

### chunk number 23: TabInv

cap <- "Cumulative probit location-scale model fit to soup data with 95\% Wald confidence interval."
lab <- "tab:Inv"
print(xtable(b3, caption = cap, label = lab, digits = 3),
include.rownames = FALSE, #only.contents = TRUE,
sanitize.text.function = function(x){x},
caption.placement = "top")

### chunk number 24: ROC5Tests

roc1 <- ROC(coef(cp.inv)[1], scale = cp.inv$sigma, fig = TRUE,
las = 1)
roc2 <- ROC(coef(cp.inv)[2], scale = cp.inv$sigma, fig = FALSE)
roc3 <- ROC(coef(cp.inv)[3], scale = cp.inv$sigma, fig = FALSE)
roc4 <- ROC(coef(cp.inv)[4], scale = cp.inv$sigma, fig = FALSE)
roc5 <- ROC(coef(cp.inv)[5], scale = cp.inv$sigma, fig = FALSE)
lines(roc2[[1]], roc2[[2]], lty = 2)
lines(roc3[[1]], roc3[[2]], lty = 3)
lines(roc4[[1]], roc4[[2]], lty = 4)
lines(roc5[[1]], roc5[[2]], lty = 5)

### chunk number 25: ROCDay

rocd1 <- ROC(coef(cp.inv)[4], scale = cp.inv$sigma, fig = TRUE,
las = 1)
rocd2 <- ROC(coef(cp.inv)[4] - coef(cp.inv)[9],
scale = cp.inv$sigma, fig = FALSE)
lines(rocd2[[1]], rocd2[[2]], lty = 2, col = "blue")

### chunk number 26: AUC

tmp <- AUC(coef(cp.a)[1], scale = cp.a$sigma,
se.d = summary(cp.inv)$coefficients[1, 2])
auc <- round(tmp$val, 3)
low <- round(tmp$lower, 3)
upp <- round(tmp$upp, 3)

### chunk number 27: res.dev

models <- list(Binormal,
update(Binormal, method = "logistic"),
update(Binormal, method = "cloglog"),
update(Binormal, method = "cauchit"),
start = coef(Binormal)))

Pears.mods <- sapply(models, function(x) {
  e <- x$fitted.values[ c(1,7),] * rowSums(tab26)
  sum(((tab26 - e)^2 / e)
})
P.val <- pchisq(Pears.mods, df = 12 - 5, lower = FALSE)
rdev <- sapply(models, deviance)

link <- c("Probit", "Logit", "c-log-log", "Cauchit")
dist <- c("Normal", "Logistic", "Gumbel", "Cauchy")
c.nam <- c("Distribution", "Link Function", "Residual Deviance", "Pearson $\chi^2$", "$p$-value")
xtab <- data.frame(dist, link, rdev, Pears.mods, p.val)
colnames(xtab) <- c.nam
cap <- "Residual deviance and Pearson $\chi^2$ goodness of fit test assuming different latent distributions for the Binormal model fitted to the soup data."
lab <- "tab:res.dev"
print(xtable(xtab, caption = cap, label = lab),
  include.rownames = FALSE,
  caption.placement = "top",
  sanitize.text.function = function(x){x})

###################################################
### chunk number 28: LinkDist
###################################################
x <- seq(-4, 4, len = 500)
y <- sensR:::dgumbel(x)
plot(x, y, type = "l", ylim = c(0, .5),
xlab = "", ylab = "Density")
lines(x, dnorm(x), lty = 2, col = "red")
#lines(x, dlogis(x), lty = 3, col = "blue")
lines(x, dlogis(x, , sqrt(3)/pi), lty = 3, col = "blue")
lines(x, dcauchy(x), lty = 4, col = "green")
Appendix C

Thurstonian models for sensory discrimination tests as generalized linear models

Manuscript for Food Quality and Preference
C.1 Abstract

Sensory discrimination tests such as the triangle, duo-trio, 2-AFC and 3-AFC tests produce binary data and the thurstonian decision rule models link the underlying sensory difference $\delta$ to the observed number of correct responses. In this paper it is shown how each of these four situations can be viewed as a so-called generalized linear model with the inverse of the psychometric function in each case as the corresponding statistical link function. The underlying sensory difference $\delta$ becomes directly a parameter of the statistical model and the estimate $d'$ and its standard error becomes the “usual” output of the statistical analysis. The freely available statistical software R [http://www.r-project.org] was used in the present paper and an R-package with the additional routines developed in the paper will be available.

C.2 Introduction

Sensory discrimination tests such as the triangle, duo-trio, 2-AFC and 3-AFC tests together with the A-not A and Same-Different tests produce binary data leading to count data when aggregated over replications and/or assessors/consumers. Hence, the basic statistical methods needed for planning such experiments and analysing such data as they come, can be found among methods based on the binomial distribution and standard methods for analysing tables of counts. The former include basic binomial based power and sample size calculations and exact binomial hypothesis testing, see eg. Schlich (1993). The latter could be using the Pearson’s $\chi^2$-test or McNemar’s test for paired 2-by-2 tables, see eg. Bi (2006b). The weakness of this approach working on the count scale is that it is test protocol dependent: the number of expected correct answers for the same products depend heavily on which test that is carried out. This has been pointed out by several authors, see eg. Ennis (1993), and amounts to the lack of a common framework for comparing the underlying sensory differences across different testing paradigms. The thurstonian approach of transforming the number of correct answers into an estimate, called d-prime ($d'$), of the underlying (relative) sensory difference, is the solution of this deficiency of the count data statistical approach.

Generalized linear models, McCulloch and Searle (2001), are extensions of linear models (regression, analysis of variance and combinations thereof) designed to
cope with non-normal data including binary and count (e.g. Poisson) data. The observed data is linked to a linear model structure by a non-linear function. The benefit of the approach is that the whole statistical apparatus of linear modelling carries over to the more general non-normal/non-linear setting. So apart from a theoretical framework, it provides a practical approach, through existing software, to analyse binary/count data of various kinds with the same level of detail (ANOVA, Regression, ANCOVA etc) as more commonly known to be carried out for quantitative data with normally distributed errors.

It becomes quite obvious below that the thurstonian approach to handling sensory difference testing will, at least in some occasions, amount to a generalized linear model approach but with the so-called links defined by the thurstonian psychometric functions rather than the traditionally used links. There exists several statistical packages with generalized linear model features that allows the user to specify special link functions hence needed. It is therefore possible to explore this intersection between statistical theory and practice and the thurstonian approach of sensory discrimination testing. This has not been done before, and we show that it does contribute non-trivially to the ability to analyze sensory difference data.

In the next section the generalized linear models are introduced and motivated in the context of sensory data. Then two sections devoted to each their type of testing protocols is given: One on the four classical protocols: 2-AFC, 3-AFC, duo-trio and triangle and one on the A-not A protocol Each section contains theoretical expressions of the models for the protocols, the relation to the generalized linear models is clarified and a simple example is given. The R-package is described in the appendix together with a short description of how to get started with the R-programme.

C.3 General and Generalized Linear Models

In this section the generalized linear models are introduced and motivated by first introducing the general linear model. The general linear model is a theoretical framework that in a unified way covers many basic kinds of statistical analyses: (fixed effects) analysis of variance (ANOVA), (multiple) linear regression (MLR) and analysis of covariance (ANCOVA), (see e.g. McCulloch and Searle, 2001; Crawley, 2005). Any such classical analysis can be seen as a result of fitting a linear model to the data $y_i, i = 1, \ldots, n$ given by

$$y_i = x_i^T \beta + e_i$$
where $x_t^i$ is a vector of design variables for the $i$th observation and $\beta$ is the corresponding vector of unknown parameters. The components of $x_t^i$ are either quantitative values (as in MLR) or zero-one dummy codings of qualitative information (as factors in ANOVA). Components of $\beta$ are either real regressions coefficients or an effect of a qualitative treatment level. Both type of components may be present (as in ANCOVA). The “linear” name comes from the fact that the expected structure in the data is expressed as a linear function of the parameters in the model:

$$E(y_i) = x_t^i \beta$$

Note that this does not exclude the possibility of modelling non-linear relations between quantitative $x$ and $y$ by the use of eg. polynomial regression models.

It is common to supplement with an assumption of observations being independent and homoscedastic normally distributed:

$$e_i \sim N(0, \sigma^2)$$

and the theory is build around this assumption, although most practical statistical procedures will be valid at least approximately in much greater generality due to the Central Limit Theorem.

In short, the outcome of an application of a version of the general linear model is always essentially two things:

1. The estimated parameter values $\hat{\beta}$.
2. The uncertainty information about these estimates $\text{Cov}(\hat{\beta})$.

And there are direct and simple matrix-based formulas for computing these two things, which may be found in textbook covering the general linear model, see eg. McCulloch and Searle (2001).

From this, t-test’s for the significance of individual or simple combinations of parameters can be found. Also F-test for composite hypotheses often employed in ANOVA can be deduced from this. The estimates of the parameters $\hat{\beta}$ are least squares estimates and under the assumption of normal homoscedastic errors, they are also maximum likelihood estimates (MLE).

However, for clearly non-normal data, for instance categorical/binary data, the methods based on the general linear model cannot in general be trusted. Fortunately, there is an even more general family of methods, the generalized linear models, that includes the normal linear model as a special case and also allows for non-normality, categorical responses, non-homogeneous variances and some
kind of non-linearity in the data structure. Let us assume that we observe a binary response $y_i$, for which we can assume that the basic observations come from a binomial distribution:

$$y_i \sim \text{bin}(1, p_i)$$

This could be the responses from a number of triangle tests $i = 1, \ldots, n$. Also assume that different stimulus concentrations $x_i$ were applied and that the consumer panel background information, e.g. gender, is of particular interest. Had the response been quantitative, e.g. sweetness intensity, we could have used a general linear model with gender as a qualitative descriptor and the stimulus concentration as a quantitative descriptor, and express a model where the response could depend (linearly) on the stimulus concentration, possibly depending on the gender, that is, a model with the two "independent" variables gender (qualitative) and stimulus concentration (quantitative) and their interaction:

$$E(y_i) = \beta_{\text{Gender}(i)} + \beta_{\text{Gender}(i)} x_i$$  \hspace{1cm} (C.3.1)

As it seems definitetly questionable to set up such a quantitative model directly for the binary zero-one outcomes $y_i$, it could be somewhat less artifical, perhaps, to do so for the expected value of the outcome, as actually expressed in (C.3.1). Recall, that the expected value for a binomial $\text{bin}(1, p_i)$ outcome is given by $p_i$, that can take any value between zero and one, so theoretically $p_i$ is a quantitative and continuous construct and a model relating $p_i$ to the stimulus concentration and gender factor is what is seeked here. However, there is still a theoretical flaw in (C.3.1): The right hand side quantitative expression can in principle attain any value as well negative values as values larger than 1, whereas the left hand side expected value only can attain values between 0 and 1. This theoretical conflict can be avoided by linking the linear model to a suitable non-linear function of the expected values instead, for instance the logit-function:

$$\log\left(\frac{E(y_i)}{1 - E(y_i)}\right) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_{\text{Gender}(i)} + \beta_{\text{Gender}(i)} x_i$$  \hspace{1cm} (C.3.2)

The logit-function maps values between 0 and 1 into the full set of real valued numbers. This model is known as a logistic regression and/or log-odds model and has a history of its own, cf. Cramer (2002). The generalized linear model (GLIM), see McCullagh and Nelder (1989) is a theoretical and methodological framework that includes the logistic regression model, the classical linear normal model, and many other models depending on the specific choice of the two concepts:

- The distribution of $y_i$. 
- The link between $E(y_i)$ and the linear model structure.

The logistic regression model is a GLIM with the binomial distribution and the logistic link. A probit regression model, cf. [Cramer (2002)], is a GLIM with the binomial distribution and the probit (inverse standard normal) link. A classical linear normal model is a GLIM with the normal distribution and the identity link. The most well-known GLIMs, that is, those that are commonly available in software, includes the normal, binomial, poisson and gamma distributions and the identity, logit, probit, log, complementary loglog, square-root and inverse links.

The benefit of such a framework is that it offers an approach for the analysis of data that can be used for all these situations together with the theoretical foundation for the analysis. The results of a GLIM is essentially exactly as for the linear model:

1. The estimated parameter values $\hat{\beta}$.
2. The uncertainty information about these estimates $\text{Cov} (\hat{\beta})$.

There are generally no closed form solutions for these two pieces of information. However, maximum likelihood estimates can be obtained by a simple algorithm, the Iterative Weighted Least Squares (IWLS) algorithm. Essentially, the algorithm repeatedly apply a weighted linear model to a transformed version of the data. However, this is an integrated part of GLIM software and the end user does not have to know the details of this. The main message here is that we can handle the analysis of non-normal data following exactly the same principles as we would use for a classical normal based analysis, at least whenever the relevant model can be expressed as a GLIM.

### C.4 The triangle, duo-trio, 2-AFC and 3-AFC as generalized linear models

A common feature for the four sensory difference tests mentioned above is that the observation consists of a single number of correct answers, $y$, that can be assumed to follow a binomial distribution $\text{bin}(n, p)$. For the no difference hypothesis test nothing more is needed, the classical binomial based statistical methods can be used. In thurstonian modelling the focus is on quantifying/estimating the underlying sensory difference $\delta$ between the two products that are compared.
C.4 The triangle, duo-trio, 2-AFC and 3-AFC as generalized linear models

in the difference test. This is done by setting up mathematical models for the
cognitive decision processes that is used by assessors in each sensory test proto-
col, see e.g. [Frijters (1979)]. In this way the observed number of correct answers
is directly linked to the underlying sensory difference \( \delta \) through a protocol specific non-linear function, the so-called psychometric function. The psychometric
functions used in this section is found in [Ennis (1993)], but all of them have
origins going much further back in time, e.g. [Thurstone (1927a)].

For the K-AFC method the psychometric function is given by:

\[
f_{K-AFC}(\delta) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-(z - \delta)^2/2\right] \Phi(z)^{K-1} dz
\]

where \( \Phi \) is the standard normal distribution function. So for \( K = 3 \) it becomes:

\[
f_{3AFC}(\delta) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-(z - \delta)^2/2\right] \Phi(z)^2 dz
\]

And for \( K = 2 \) it simplifies to:

\[
f_{2AFC}(\delta) = \Phi\left(\frac{\delta}{\sqrt{2}}\right)
\]

For the triangle method the psychometric function is:

\[
f_{tri}(\delta) = 2 \int_{0}^{\infty} \left\{ \Phi\left[-z\sqrt{3} + \delta\sqrt{2/3}\right]
+ \Phi\left[-z\sqrt{3} - \delta\sqrt{2/3}\right] \right\} \phi(z) dz
\]

And the duo-trio psychometric function is given by:

\[
f_{d-t}(\delta) = 1 - \Phi(\delta/\sqrt{2}) - \Phi(\delta/\sqrt{6})
+ 2\Phi(\delta/\sqrt{2})\Phi(\delta/\sqrt{6})
\]

Operationally these functions are used in their inverse versions, since given an
observed fraction of correct answers \( \hat{p} = \frac{y}{n} \) the value of delta, called \( d' \) (d-prime), that fits with the psychometric function:

\[
f_{ps}(d') = \hat{p}
\]

is determined. The psychometric functions for the four protocols are shown in
figure C.4. Also shown is the function corresponding to the logistic link (dashed).
The theoretical model version of this operation together with the basic binomial
assumption can be expressed in the following way, where \( y_i \) is the answer given in the \( i \)'th test:

\[
y_i \sim \binom{1, p_i}, \quad g(p_i) = \delta
\]  

(C.4.6)

It is now evident that this is a GLIM with the binomial distribution and the inverse psychometric function \( f_{ps}^{-1}(\cdot) = g(\cdot) \) as the link—compare with equation (C.3.2). The linear structure of this model is the simplest possible: a model with only an “intercept” parameter: \( \delta \).

Some statistical software for GLIMs have the option of allowing for user defined link-functions, so the practical consequences of identifying the difference test situation as a GLIM is that the computation of \( d' \) becomes the estimation of the intercept in a GLIM. Hence, the variance of the \( d' \) is given for free since it is given by the standard error of the intercept estimate, which is automatically provided by the GLIM approach. Also, when \( d' \)'s for different product versions are found with the purpose of comparing these, the comparison can be carried out as a standard linear contrast investigation in a GLIM in analogy with a classical two-sample t-test. Moreover frameworks where more than one or two experimental setting are carried out, calling for several different d-prime calculations and comparisons or subsequent modelling/investigation of these, can be incorporated into a single joint analysis of the data in a usual t-test, ANOVA or regression manner. All this will be exemplified in the following. Note that this is no different from what is usually recommended in these situations, (see eg. Bi et al., 1997). The novelty lies in the fact that the situation is embedded into a classical (generalized) linear model setting which means that roughly all that is covered in (Bi et al., 1997) becomes standard (generalized) linear model stuff.

We have included the definition of the link functions and other GLM attributes in appendix [C.9](#) for a complete specification.

Another novelty is the fact that an implementation of everything covered in this paper is available for everyone to download from [http://www.cran.r-project.org](http://www.cran.r-project.org) in terms of the R-package sensR, see the Appendix. And since R is Open Source software everyone can obtain it for free at [http://www.r-project.org](http://www.r-project.org).

### C.4.1 Examples

We start with a simple example. Assume that we have 10 out of 15 correct responses in one of the four discrimination test settings covered here. For the triangle and 3-AFC setting this would amount to a one-tailed exact binomial P-value for the hypothesis of no difference of 0.0085, cf. eg. the tables of Schlich.
For the duotrio and 2-AFC settings the P-value is 0.1588. Using the generalized linear model procedure of the R-software together with the four psychometric families as given in the R-package sensR, we obtain standard linear model statistical output as shown in table C.1.

It is not the intention here to teach the reader the syntax of R nor the specifics of the GLIM procedure of R. It should be clear, however, that the output as presented in table C.1 is of standard linear model type as would be given by any statistical programme: The estimate, its standard error, a statistic for the hypothesis that the parameter is zero (parameter value divided by its standard error) and finally a P-value for this hypothesis test. For comparison, the results from a standard logistic regression is given also.

The P-values coming from this standard GLIM analysis should not be used, however. Primarily, because they as default in linear and generalized linear model software are given as two-tailed P-values. For the discrimination testing situation here we should use one-tailed P-values and there is no reason not to use the exact binomial based P-values. For this reason a procedure was constructed such that only the proper and relevant information for the simple discrimination test analysis is given and such that the GLIM syntax of R is hidden for the user. Due to its simplicity the exact R-command is given here: At the R command prompt simply write: \texttt{discrim(10, 15, "triangle")} and the \(d\)' standard error and the proper P-value is given together with confidence limits of \(d\)'.

The package can also be used for power and sample size computations, (Schlich [1993], Ennis [1993]). For instance the power (probability of detecting) \(d' = 1\) with \(n = 30\) tests using a level 0.05 hypothesis test is found as \texttt{discrimPwr(1, 30, .05, "triangle")}. Similarly the number of samples needed for obtaining a power of 90\% for \(d' = 1\) using a level 0.05 hypothesis test is found as \texttt{discrimSS(1, .9 , .05, "triangle")}. The result for the triangle test as well as for three other tests are given in table C.2.

These functions can easily be used also by users new to R. Experienced R-users may use these as background for various plotting of the basic psychometric functions. A simple discrimination test simulation device is also provided.

We now turn to a more involved example, showing how we may combine a traditional statistical approach with the psychometric foundation of thurstonian modelling. Assume that four product samples of increasing concentrations each were tested by two groups of subjects, say, males and females, yielding the data seen in table C.3.

So for this study, 160 subjects (consumers) were used. The most natural choice of analysis for a statistically minded person would be a binomial based mod-
elling, typically the logistic regression within the GLIM framework. The structure of the experiment and the samples enables a decomposition of the information into the following effects: main effects of gender (1 degree of freedom (DF)), linear trend effect of Sample concentrations (1 DF), remaining sample differences (2 DF), interaction between linear trends and gender, i.e. different slopes for genders (1 DF) and the remaining sample-gender interaction (2 DF). Formally, the full model could be expressed as

$$g_{\text{logistic}}(p_{ij}) = \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \alpha_j + \beta_j \text{Conc}_i + \gamma_{ij} \quad (C.4.7)$$

where \( p_{ij} \) is the probability of success for the \( n_{ij} \) subjects of gender \( j \) for sample type \( i \). The results of such an analysis can be put into a so-called analysis of deviance (ANODE) table similar to an ANOVA table. In table C.4 this is found under the Logistic heading.

Each deviance value expresses the difference between having the effect in the model or not, sequentially from the top. The difference is measured by minus twice the difference in log-likelihoods and standard statistical theory calls for using a \( \chi^2 \)-distribution for approximate P-values. In table C.4 the results of doing a triangle and 3AFC based analysis of the data corresponding to the following thurstonian generalized linear models:

$$g_{\text{Triangle}}(p_{ij}) = \alpha_j + \beta_j \text{Conc}_i + \gamma_{ij} \quad (C.4.8)$$

and

$$g_{\text{ThreeAFC}}(p_{ij}) = \alpha_j + \beta_j \text{Conc}_i + \gamma_{ij} \quad (C.4.9)$$

where \( g_{\text{Triangle}} \) and \( g_{\text{ThreeAFC}} \) are the proper link functions, that is, the inverted psychometric functions. Based on the implementations of these in the R-package sensR, these additional analyses can be handled similar to the standard logit-based analysis, see the appendix for details. It is clear from the results in table C.4 that when it comes to the conclusions regarding the structure of the underlying pattern of effects it makes hardly any difference what link function is used in this case. The conclusion is rather clear: There is a main effect of gender and a linear increase of sensory difference related to the sample concentration. All other possible effects are clearly non-significant. It is also clear that the full structured model in each case is really just the same as having eight different binomial distributions, and hence exactly the same model, no matter what link function is specified. The final (significant) model is then given by 3 parameters:

$$g(p_{ij}) = \alpha_j + \beta \text{Conc}_i, \quad j = 1, 2.$$  

As mentioned in the introductory section, the standard outcome of a model like this, are the estimates of these parameters together with their standard errors. These are shown in table C.5 for all three models.
Whereas parameter estimates in the logistic regression have log-odds ratio interpretations suitable for many purposes, for instance within medical statistics, they do not as such have a thurstonian interpretation. If the data of table C.3 were stemming from triangle tests the parameters are directly interpretable as d-primes: males have a d-prime one less than females, no matter what the concentration level is (within the range of the data). And a concentration increase of 1 ($1.686 - 0.658 \approx 1$) induces an increase of d-prime of 0.5 (for both genders). Standard GLIM post hoc analysis can be used to provide predictions for any concentration value together with confidence intervals for these, see figure C.5 for the triangle d-prime results. Note that the confidence bands for the two lines are heavily overlapping, expressing the fact that for each single concentration we cannot significantly determine a gender difference. The structure and the results of the analysis in table C.5 however shows clearly that the gender difference definitely is significant.

C.5 The A-NOT A method

The A-Not A (or “Yes-No”) method has been used as one of several possible sensory techniques for product testing (Bi and Ennis, 2001b, a; Bi, 2006b). It has been a commonly used procedure within the theory of signal detection (Macmillan and Creelman, 2005). We consider for now the monadic A-not A design, where each subject is presented with only one test sample - A or Not-A. The following formula is used to compute $d'$ from responses on a number of A and not-A samples.

$$d' = \Phi^{-1} \left( \frac{\text{No of not-A responses}}{\text{No of not-A samples}} \right) - \Phi^{-1} \left( \frac{\text{No of not-A responses}}{\text{No of A samples}} \right) \tag{C.5.1}$$

This can be identified as the difference between two means (contrast) in an independent two-sample setup, corresponding to the standard two-sample Student’s t-test setting. Let $y_1, \ldots, y_{n_N}$ be the responses for the $n_N$ subjects presented with the Not-A samples and $y_{n_N+1}, \ldots, y_{n_N+n_A}$ be the responses for the $n_A$ subjects presented with the A samples. Then a proper statistical model for these data, allowing for different response probabilities for the two groups of data can be expressed as the following generalized linear model:

$$y_{ij} \sim \text{bin}(1, p_j), \quad \Phi^{-1}(p_j) = z_j \tag{C.5.2}$$

where $\Phi^{-1}(\cdot)$ is the probit link, $i = 1, \ldots, n_j$ and $j = \{A, \text{Not-A}\}$. With this
formulation, the A-not A \( d' \) becomes
\[
d' = z_{\text{Not-A}} - z_A
\]
Hence, the A-not A test can be handled by arranging the data properly and using the probit link function—one of the standard links in any GLIM software.

C.5.1 Example

Suppose we conducted a A-Not A test and obtained the data shown in table C.6. If we assume the data are independent, we may estimate \( \delta \) using the above mentioned A-Not A method of the R-package sensR. We obtain the usual estimate \( d' = .378 \) along with the standard error \( \text{se}(d') = .178 \) and the P-value = .0237. This estimate of \( d' \) coincides with that given by Bi and Ennis (2001b). The estimates of the standard errors are close (Bi and Ennis (2001b) obtain \( \text{se}(d') = \sqrt{.032} \approx .179 \)) whereas the p-values differ slightly (Bi and Ennis (2001b) obtain P-value = 0.034 based on Pearson’s \( \chi^2 \) statistic), hence their estimate is slightly conservative.

C.6 Likelihood and Confidence Intervals

Another advantage of the identification of the discrimination models as generalized linear models is that likelihood based confidence intervals (CIs) are easy to compute. The main advantages of the likelihood CIs over the classical symmetric Wald CIs are that they need not be symmetric reflecting more properly the evidence about the parameters, and that they are defined in extreme situations such as when all samples are answered correctly or when the observed proportion of correct answers is at or below the guessing probability of the test.

As noted by Bi et al. (1997), computing Wald based CIs for \( \delta \), when the estimate, \( \hat{\delta} \) is low is problematic. This happens because the likelihood is highly asymmetric as we shall show shortly.

The theory of likelihood, likelihood and likelihood CIs will not be explained here. The topics are introduced in the context of same-different tests in (Christensen and Brockhoff, 2008a) and the concepts carry over to the test protocols considered here. For the statistically interested reader we recommend the book by Pawitan (2001) as a very good introduction to likelihood inference.

\(^1\)This dataset is also considered in (Bi and Ennis, 2001b)
Consider first the situation that we in a triangle test obtain 10 correct and 10 incorrect answers. The likelihood for this test is shown in figure C.1 as the solid line. The 95% and 99% likelihood CIs are given by the intersection with grey lines horizontal lines. These lines follows from the assymptotic $\chi^2$-distribution of the relative or normalized likelihood of one parameter. The level of the lines are given by $\exp(-F(1-\alpha)/2)$, where $F(\cdot)$ is the cdf of a $\chi^2$-density with one degree of freedom. For $\alpha = (0.05, 0.01)$, the level of the lines are at $(0.1465, 0.0362)$.

In package MASS [Venables and Ripley 2002] for R, a function confint to compute likelihood confidence intervals for generalized linear models is implemented. We can make use of this function via the implementation of the link functions for the discrimination test protocols in package sensR to obtain the likelihood CI for $\delta$ in the triangle test; $(0, 2.55)$. This is clearly another advantage of the identification of the discrimination test protocols as generalized linear models.

The Wald confidence interval is based on a normal approximation to the likelihood curve. We have shown this symmetric approximation to the likelihood in figure C.1 with the dashed line. The Wald CI is in this case $(0.317, 2.62)$ and leads to the wrong conclusion that $\delta$ is significantly different from 0 on the 95% level. The truth is that the likelihood of a true underlying $\delta = 0$ is reasonably high and supported by data.

The Wald CI is obvious only reliable, when the likelihood curve is symmetric. We have found that the likelihood is often asymmetric for the duo-trio and triangle tests and reasonably symmetric for 2 and 3-AFC tests and for A-Not A tests. The symmetry is generally better for $\delta$ of intermediate size (ie. around 2-3) and increases with the sample size. It is clear that for equivalence testing [Schuirmann 1987; Bi 2005; 2007; Ennis et al. 2007; Meyners 2007b], where $\delta$ is often close to 0, the use of likelihood is mandatory for valid inference.

In figure C.2 the likelihood is shown for a duo-trio test with 10 correct and 12 incorrect answers. In such a situation, where the fraction of correct answers is lower than the guessing probability the estimate of $\delta$ is 0 and symmetric Wald based CIs are inappropriate. An appropriate 95% likelihood confidence interval is however given by $(0, 1.5)$ as seen in the figure. In equivalence testing the true $\delta$ is often near 0 and the situation just described is not uncommon.

Another common situation, when the sample size is small to moderate and the true $\delta$ is high occurs when all answers are correct. The MLE of $\delta$ is now $\infty$, but we can still produce a likelihood and corresponding CI. We have shown a likelihood for duo-trio test with 10 out of 10 answers were correct in figure C.3. Note that even with this small sample size, we find that the 95% likelihood CI is $(2.54, \infty)$. 

In figure C.2 the likelihood is shown for a duo-trio test with 10 correct and 12 incorrect answers. In such a situation, where the fraction of correct answers is lower than the guessing probability the estimate of $\delta$ is 0 and symmetric Wald based CIs are inappropriate. An appropriate 95% likelihood confidence interval is however given by $(0, 1.5)$ as seen in the figure. In equivalence testing the true $\delta$ is often near 0 and the situation just described is not uncommon.

Another common situation, when the sample size is small to moderate and the true $\delta$ is high occurs when all answers are correct. The MLE of $\delta$ is now $\infty$, but we can still produce a likelihood and corresponding CI. We have shown a likelihood for duo-trio test with 10 out of 10 answers were correct in figure C.3. Note that even with this small sample size, we find that the 95% likelihood CI is $(2.54, \infty)$. 

In figure C.2 the likelihood is shown for a duo-trio test with 10 correct and 12 incorrect answers. In such a situation, where the fraction of correct answers is lower than the guessing probability the estimate of $\delta$ is 0 and symmetric Wald based CIs are inappropriate. An appropriate 95% likelihood confidence interval is however given by $(0, 1.5)$ as seen in the figure. In equivalence testing the true $\delta$ is often near 0 and the situation just described is not uncommon.

Another common situation, when the sample size is small to moderate and the true $\delta$ is high occurs when all answers are correct. The MLE of $\delta$ is now $\infty$, but we can still produce a likelihood and corresponding CI. We have shown a likelihood for duo-trio test with 10 out of 10 answers were correct in figure C.3. Note that even with this small sample size, we find that the 95% likelihood CI is $(2.54, \infty)$.
Figure C.1: Likelihood (solid) and the normal approximation (dashed) to this for a triangle test with 10 correct and 10 incorrect answers. Intersection with horizontal grey lines denote the 95% and 99% likelihood confidence intervals (solid) and Wald confidence intervals (dashed).
Figure C.2: Likelihood for a duo-trio test with 10 correct and 12 incorrect answers. Intersection with horizontal grey lines denote the 95% and 99% likelihood confidence interval.
In our package, sensR, we have included confint and profile methods facilitating the use of (profile) likelihood methods and corresponding CIs.

![Figure C.3: Likelihood for a duo-trio test with 5 correct and 0 incorrect answers. Intersection with horizontal grey lines denote the 95% and 99% likelihood confidence interval.](image)

C.7 Summary and discussion

We have clarified the practical, technical and scientifically relevant connection between certain thurstonian models used in sensory and signal detection science and the statistical concept of generalized linear models. This covers the duo-trio, triangle, 2-AFC, 3-AFC and A-Not A test protocols. We have used this connection to implement the basic d-prime calculations for these protocols in the R-package sensR by implementing the corresponding psychometric functions and their inverses. For the basic d-prime calculation this provides nothing new in the sense that it gives the proper d-prime results well known previously from the literature. The novelty here lies in providing a free and open source software package to do the discrimination test analysis including power, sample size and simulation procedures. And a package embedded in the internationally expanding R-environment. A real benefit is that the d-prime uncertainty computation
becomes an integral part of the statistical analysis and even more so the fact that it is now possible to directly combine standard statistical analysis such as ANOVA and regression analysis with thurstonian modelling in a natural and optimal way.

The generalized linear models have previously been brought up as a tool for analysing sensory data, (Hunter et al., 2000; Brockhoff and Müller, 1997) but only in their classical versions not specifically taking the thurstonian approach. A discussion of classical GLIM links versus thurstonian links was brought up in (Brockhoff, 1995). In (Brockhoff, 2003) ordinary and corrected GLIMs were discussed as a tool to handle the replicated situation, but still in their classical versions. The developments of the current paper stand as the basis for formulating formal thurstonian versions of models for replicated discrimination tests as alternative to the commonly used beta-binomial approach, cf. (Ennis and Bi, 1998; Bi and Ennis, 1998). This is ongoing work. Also ongoing is the further developments of this approach to handle A-Not A protocols with sureness leading to ordinal data with and without replications.

Appendix

C.8 Implementation in R

R is an Open Source implementation of the well-known S language. R is a language and environment for statistical computing and graphics freely available for anyone and easily downloaded from the R-project home page [http://www.r-project.org/](http://www.r-project.org/). It is not the intention to introduce the use of R here. Introductory material can be found on the website and Dalgaard (2002) gives a good introduction with focus on statistical issues, while (Venables and Ripley, 2002) is a comprehensive reference. A so-called R-package named sensR was constructed by the authors and this package can be downloaded from [http://cran.r-project.org](http://cran.r-project.org) or will be emailed by the authors if requested. The package includes full documentation, help-files and examples in an R-integrated fashion.

The remaining part of this appendix will briefly describe the R-code used in this paper.

R-code for the examples in section 3.1

We assume that you have R up and running and that you have installed the
package `sensR` on your computer. To load the package and make the tools presented in this paper available, simply at the R-prompt type

```
> library(sensR)
```

With the package loaded, we obtain the simple discrimination, power and sample size calculations with

```
> discrim(success = 10, total = 15, method = "triangle")
```

Call: discrim(success = 10, total = 15, method = "triangle")

Results for the triangle test:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-prime</td>
<td>2.321377</td>
<td>0.6510397</td>
<td>1.045363</td>
<td>3.597391</td>
<td>0.008504271</td>
</tr>
</tbody>
</table>

```
> discrimPwr(delta = 1, sample.size = 30, alpha = 0.05, method = "triangle")
```

[1] 0.2330314

```
> discrimSS(delta = 1, power = 0.9, alpha = 0.05, method = "triangle")
```

[1] 283

The naming of the arguments, is not necessary as the examples in the text shows. To lean more about the functions, use the ? operator, to get help on say the `discrim` function, type `?discrim`.

To obtain a plot of the normal distributions of sensory intensity, simply type eg. `plot(discrim(10, 15, "triangle"))` (not shown).

Next we show how to fit the model with gender and concentration as explanatory variables. We only show the final model, but the model is easily extended:

```
> data <- expand.grid(conc = 1:4, gender = c("Males", "Females"))
> data$correct <- c(9, 11, 13, 14, 13, 14, 16, 18)
> data$total <- rep(20, 8)
```
> model <- glm(cbind(correct, total - correct) ~ gender + conc, + data, family = triangle)
> summary(model)$coefficients

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| Intercept| 0.6580368  | 0.5865198 | 1.121934 | 0.26189031 |
| genderFemales | 1.0278006 | 0.4312735 | 2.383176 | 0.01716399 |
| conc      | 0.5022051  | 0.1973866 | 2.544271 | 0.01095060 |

Fits of models with other links are obtained by exchanging triangle with eg. threeAFC for the 3-AFC method and binomial for the logit link.

**R-code for examples in section 4.1**

To fit the A-Not A model, we use the AnotA function

> AnotA(x1 = 57, n1 = 100, x2 = 42, n2 = 100)

Call: AnotA(x1 = 57, n1 = 100, x2 = 42, n2 = 100)

Results for the A-Not A test:

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std. Error</th>
<th>Lower</th>
<th>Upper</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>d-prime</td>
<td>0.3782676</td>
<td>0.1784076</td>
<td>0.02859508</td>
<td>0.02371745</td>
</tr>
</tbody>
</table>

**The psychometric functions**

The psychometric functions are implemented as inverse link-functions in so-called family objects. One can (as we did) use these for plotting purposes and we will show how to obtain a simple plot (which is not shown due to space requirements) of the psychometric function for the duo-trio test:

> x <- seq(0, 5, length = 100)
> y <- duotrio()$linkinv(x)
> plot(x, y, type = "l", ylim = c(0.5, 1), ylab = "probability",
+     xlab = "d-prime")

**Profile likelihoods**

The profile likelihood for the triangle experiment in figure C.1 can be made with the following command
Further information on how to compute and plot the profile likelihoods in the end situations is found in the help pages (e.g. ?duotrio) and examples are given, which can be run with for instance example(duotrio).

**Additional features**

The package *sensR* currently also contains functions to analyze same-different tests as well as well as perform ROC (Receiver Operating Characteristic) curve analysis of A-Not A type experiments including AUC (area under the ROC curve) computations.

The package is continuously evolving with more features and will include functions to support the developments mentioned in the summary and discussion. The authors are happy to receive comments and suggestions for changes or additional features and functions.

### C.9 GLM Attributes for Discrimination Tests

#### C.9.1 Model and Likelihood

The model for a discrimination test assuming independent data is given in equation ([C.4.6](#)). If other effects might influence discrimination, they can be included in the model and we may define the more general model

\[ g(\pi) = \eta = X\beta, \]

where \( \eta \) is the general linear predictor, \( X \) is a \( n \times p \) design matrix and \( \beta \) is a \( p \) vector of parameters.

The log likelihood function equals the log density function and for binomial data it is given by

\[
l(\delta; y_i) = \log f(y_i) = w_im_i \left\{ y_i^{\delta_i} \log \frac{\pi_{\delta_i}}{1 - \pi_{\delta_i}} + \log(1 - \pi_{\delta_i}) \right\} + \log \left( \frac{m_i}{m_i y_i^{\delta_i}} \right),
\]

---

90 Appendix C
where $y_i^* = y_i/m_i$ is the scaled response and $w_i$ are prior weights. The additive constant is typically excluded from the definition of the log likelihood, since the log likelihood is only meaningful up to an additive constant anyway.

The deviance components are defined as
\[
d_i = 2w_i\{l(y_i; \hat{y}_i) - l(\hat{\pi}_i; y_i)\} = 2w_im_i\left\{y_i^* \log \frac{y_i^*}{\hat{\pi}_i} + (1 - y_i^*) \log \frac{1 - y_i^*}{1 - \hat{\pi}_i}\right\},
\]
which leads to the deviance residuals
\[
r_{d_i} = \text{sign}(\pi_i - \hat{\pi}_i) \sqrt{d_i}.
\]
The deviance residuals are in general more nearly normal than other types of residuals. The standardized deviance residual is defined analogously to standardized residuals in normal linear regression
\[
r_{d_i}^* = \text{sign}(\pi_i - \hat{\pi}_i) \sqrt{\frac{d_i}{\phi(1 - h_{ii})}},
\]
where $\phi$ is a possible variance inflation factor and $h_{ii}$ are the so called leverage elements; the diagonal elements of the projection (or hat) matrix
\[
H = X(X^tWX)^{-1}X^tW.
\]
The matrix $W$ is the GLM weight matrix to be defined in the next section.

The standardized deviance residuals can be compared to the standard normal distribution in eg. a quantile-quantile plot or a histogram. They can also be plotted versus fitted values and potential predictors.

### C.9.2 GLM Attributes

The parameter estimates $\hat{\beta}$ can be computed using iterative weighted least squares (IWLS) or a more direct optimization of the likelihood function with respect to the parameters. Nevertheless we need the elements of the IWLS algorithm to compute variance estimates of the parameters and the leverage components.

The weight matrix is diagonal with elements given by
\[
W_{ii} = w_i \left(\frac{\partial \pi_i}{\partial \eta}\right)^2 V(\pi_i)^{-1}
\]
where $V(\pi_i) = \pi_i(1 - \pi_i)$ is the variance function for binomial data, ie. $\text{var}(y_i) = \phi V(\pi_i)$ and the derivative depends on the specific discrimination test.

The variance of $\hat{\beta}$ can be approximated with the expected Fisher information matrix, $\mathbf{I}(\hat{\beta}) = \mathbf{X}^tW\mathbf{X}$, hence

$$\text{var}(\hat{\beta}) = \phi (\mathbf{X}^tW\mathbf{X})^{-1}.$$  

The inverse link functions which maps from the $\delta$ scale to the probability scale and are given by the psychometric functions in equations (C.4.1)–(C.4.6). The derivatives \( \frac{\partial \pi_i}{\partial \eta} = \frac{\partial}{\partial \delta} f_{ps}(\delta) \) are 0 for non-positive $\delta$, whereas for positive $\delta$ they are given by

\[
\frac{\partial}{\partial \delta} f_{\text{tri}}(\delta) = 2\sqrt{2/3} \int_0^\infty \left\{ \phi \left( -z\sqrt{3} + \delta \sqrt{2/3} \right) 
- \phi \left( -z\sqrt{3} - \delta \sqrt{2/3} \right) \right\} \phi(z) \, dz \\
\frac{\partial}{\partial \delta} f_{\text{d-t}}(\delta) = -\phi(\delta/\sqrt{2})/\sqrt{2} - \phi(\delta/\sqrt{6})/\sqrt{6} + 2\left\{ \phi(\delta/\sqrt{2})\Phi(\delta/\sqrt{6})/\sqrt{2} + \Phi(\delta/\sqrt{2})\phi(\delta/\sqrt{6})/\sqrt{6} \right\} \\
\frac{\partial}{\partial \delta} f_{2\text{AFC}}(\delta) = \phi(\delta/\sqrt{2})/\sqrt{2} \\
\frac{\partial}{\partial \delta} f_{3\text{AFC}}(\delta) = \int_{-\infty}^\infty (z - \delta) \phi(z - \delta) \Phi^2(z) \, dz
\]

For general psychometric function, $f_{ps}()$, its inverse, the link function maps from the probability scale to the $\delta$ scale and is given by

$$\hat{\delta} = f_{ps}^{-1}(\pi) = \begin{cases} \text{root of} \{ f(\delta) - \pi \} , & p_0 < \pi < 1 \\ 0, & 0 < \pi \leq p_0 \end{cases},$$

to which there is seldom an analytical solution. Here $p_0$ is the chance probability for $f_{ps}()$, ie. $p_0 = 1/2$ for the duo-trio and 2AFC and $p_0 = 1/3$ for the triangle and 3AFC. For the 2AFC function, the inverse can be found analytically and is

$$\hat{\delta} = f_{2\text{AFC}}^{-1}(\pi) = \sqrt{2}\Phi^{-1}(\pi) \quad p_0 < \pi < 1.$$
Figure C.4: Psychometric functions for the four discrimination tests. The logistic link function is also shown (dashed).
Figure C.5: The effect of concentration on d-prime for each gender. 95% confidence intervals shown for males (dashed) and females (dotted).
### Table C.1: Standard output from 10 correct out of a total of 15 responses.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SE</th>
<th>z-value</th>
<th>P-value (2-tailed normal)</th>
<th>P-value (1-tailed exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle</td>
<td>2.3214</td>
<td>0.6510</td>
<td>3.566</td>
<td>0.0004</td>
<td>0.0085</td>
</tr>
<tr>
<td>3-AFC</td>
<td>1.1159</td>
<td>0.4359</td>
<td>2.560</td>
<td>0.0105</td>
<td>0.0085</td>
</tr>
<tr>
<td>Duotrio</td>
<td>1.5189</td>
<td>0.7159</td>
<td>2.122</td>
<td>0.0339</td>
<td>0.1509</td>
</tr>
<tr>
<td>2-AFC</td>
<td>0.6091</td>
<td>0.4734</td>
<td>1.287</td>
<td>0.198</td>
<td>0.1509</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.6931</td>
<td>0.5477</td>
<td>1.266</td>
<td>0.206</td>
<td></td>
</tr>
</tbody>
</table>
Table C.2: Power and Sample size computations for four discrimination tests—see the text for details.

<table>
<thead>
<tr>
<th>Test</th>
<th>Power</th>
<th>Sample size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Triangle</td>
<td>.2330</td>
<td>283</td>
</tr>
<tr>
<td>3-AFC</td>
<td>.9542</td>
<td>23</td>
</tr>
<tr>
<td>2-AFC</td>
<td>.2283</td>
<td>319</td>
</tr>
<tr>
<td>Duotrio</td>
<td>.9173</td>
<td>30</td>
</tr>
</tbody>
</table>
Table C.3: Data for involved example

<table>
<thead>
<tr>
<th>Sample</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correct</td>
<td>Total</td>
</tr>
<tr>
<td>A (Conc=1)</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>B (Conc=2)</td>
<td>11</td>
<td>20</td>
</tr>
<tr>
<td>C (Conc=3)</td>
<td>13</td>
<td>20</td>
</tr>
<tr>
<td>D (Conc=4)</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Effect</td>
<td>DF</td>
<td>Logistic Deviance</td>
</tr>
<tr>
<td>--------------</td>
<td>----</td>
<td>-------------------</td>
</tr>
<tr>
<td>Conc</td>
<td>1</td>
<td>6.681</td>
</tr>
<tr>
<td>Gend</td>
<td>1</td>
<td>5.886</td>
</tr>
<tr>
<td>Sample</td>
<td>2</td>
<td>0.0427</td>
</tr>
<tr>
<td>Gend:Conc</td>
<td>1</td>
<td>0.177</td>
</tr>
<tr>
<td>Gend:Sample</td>
<td>2</td>
<td>0.263</td>
</tr>
<tr>
<td>Model</td>
<td>7</td>
<td>13.0505</td>
</tr>
<tr>
<td>Residual</td>
<td>152</td>
<td>188.735</td>
</tr>
</tbody>
</table>
Table C.5: Parameter estimates for final model fits.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Logistic</th>
<th>Triangle</th>
<th>ThreeAFC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>Estimate</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.663</td>
<td>0.453</td>
<td>0.658</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.187</td>
<td>0.448</td>
<td>1.686</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.414</td>
<td>0.161</td>
<td>0.502</td>
</tr>
</tbody>
</table>
Table C.6: A-Not A data.

<table>
<thead>
<tr>
<th>Sample Response</th>
<th>“A”</th>
<th>“Not-A”</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57</td>
<td>43</td>
<td>100</td>
</tr>
<tr>
<td>Not-A</td>
<td>42</td>
<td>58</td>
<td>100</td>
</tr>
</tbody>
</table>
Appendix D

Manual for R-package sensR

Software add-on package for R downloadable from www.cran.r-project.org
Package ‘sensR’

July 17, 2008

**Type** Package  
**Title** Thurstonian models for sensory discrimination  
**Version** 1.0.0  
**Date** 2008-05-07  
**Author** Rune Haubo B Christensen and Per Bruun Brockhoff  
**Maintainer** Rune Haubo B Christensen <rhbc@imm.dtu.dk>  
**Suggests** MASS  

**Description** Provides methods for sensory discrimination

**methods** duotrio, triangle, 2-AFC and 3-AFC. This enables the calculation of d-primes, variances for d-primes and comparisons of different d-primes.

**License** GPL-2 | GPL-3

**R topics documented:**

---

<table>
<thead>
<tr>
<th>AUC.default</th>
<th>AUC computation</th>
</tr>
</thead>
</table>

**Description**

This is the default AUC function for scalar d-primes, which will compute Area Under the ROC curve (ROC is an acronym for receiver operating characteristic).
Usage

## Default S3 method:
AUC(d, scale = 1, se.d, CI.alpha = 0.05, ...)
## S3 method for class 'discrim':
AUC(d, CI.alpha = 0.05, ...)

Arguments

d a unit length vector with the value of d-prime for which AUC is to be computed or a discrim object from the fitting of a A-not A test with AnotA

scale a unit length vector giving the ratio of scale (ie. standard deviation) of the latent distribution for the no-class items relative to that of the yes-class items

se.d standard error of d (d-prime). If provided, the function will compute confidence limits of value of AUC—cf. in section value.

CI.alpha the type I level of the confidence interval of AUC

... additional arguments passed integrate

Details

The function calls integrate to obtain the area under the ROC curve implied by d and scale.

Confidence limits are based on a normal approximation of d and not of AUC. The limits are computed, if an estimate of the standard error of d is provided. Note that the limits does not take the uncertainty in estimating the scale nor that of estimating the standard error of d into account.

Value

A list with components. If se.d is supplied to the default method or if a discrim object is supplied, the object contains the latter three additional elements.

value the estimated value of AUC

res.int the result from the call to integrate
lower  the lower confidence limit
upper  the upper confidence limit
CI.alpha  echoes the provided CI.alpha

Author(s)

Rune Haubo B Christensen

Examples

(odor <- matrix(c(112, 112, 72, 53, 22, 4, 7, 38, 50, 117, 101, 62), 2,
  byrow = TRUE))
(d.primes <- SDT(odor)[,3])
for(i in 1:5) print(AUC(d.primes[i]))
## Provide standard error of d-prime and compute CI:
fm1 <- AnotA(8, 25, 1, 25)
AUC(fm1$coef, , fm1$se)
AUC(fm1)

AnotA  Analysis of A-not-A tests

Description

Computation of dprime and it’s uncertainty for the monadic A-not-A test
together with the one-tailed exact P-value of the difference test (Fisher’s
Exact test).

Usage

AnotA(x1, n1, x2, n2, …)
Arguments

\begin{itemize}
\item \texttt{x1} The number of (correct) A-answers on A-samples
\item \texttt{n1} The total number of A-samples
\item \texttt{x2} The number of A-answers on not-A-samples
\item \texttt{n2} The number of not-A-samples
\item \ldots Additional arguments passed to \texttt{glm}
\end{itemize}

Details

The function uses the \texttt{glm} and \texttt{fisher.test} functions of the \texttt{stats} package. Note that all arguments have to be positive integers.

Value

An object of class "\texttt{discrim}" (which has a print method). This is a list with elements

\begin{itemize}
\item \texttt{coef} named vector of coefficients (d-prime)
\item \texttt{res.glm} the glm-object from the fitting process
\item \texttt{vcov} variance-covariance matrix of the coefficients
\item \texttt{se} named vector with standard error of the coefficients (standard error of d-prime)
\item \texttt{data} a named vector with the data supplied to the function
\item \texttt{p.value} one-sided p-value from Fisher’s exact test (\texttt{fisher.test})
\item \texttt{test} a string with the name of the test (A-Not A) for the print method
\item \texttt{call} the matched call
\end{itemize}

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References

See Also

print.discrim, discrim, discrimPwr, discrimSim, discrimSS, findcr

Examples

# data: 10 of the A-samples were judged to be A
# 20 A-samples in total
# 3 of the not-A samples were judged to be A
# 20 not-A-samples in total

AnotA(10, 20, 3, 20)

## Extended example plotting the profile likelihood
xt <- cbind(c(3, 10), c(20 - 3, 20 - 10))
lev <- gl(2, 1)
summary(res <- glm(xt ~ lev,
       family = binomial(link = probit)))
N <- 100
dev <- double(N)
level <- c(0.95, 0.99)
delta <- seq(1e-4, 5, length = N)
for(i in 1:N)
  dev[i] <- glm(xt ~ 1 + offset(c(0, delta[i])),
     family = binomial(probit))$deviance
plot(delta, exp(-dev/2), type = "l",
   xlab = expression(delta),
   ylab = "Normalized Profile Likelihood")
## Add Normal approximation:
lines(delta, exp(-(delta - coef(res)[2])^2 /
   (2 * vcov(res)[2,2])), lty = 2)
## Add confidence limits:
lim <- sapply(level, function(x)
  exp(-qchisq(x, df=1)/2 )
abline(h = lim, col = "grey")

ROC.default

Plot the Receiver Operating Characteristic Curve
Description

The function computes and plots the empirical ROC (receiver operating characteristic) curve.

Usage

ROC(object, ...)

## Default S3 method:
ROC(object, se.d, scale = 1, length = 1000,
    fig = TRUE, se.type = c("CI", "SE"), CI.alpha = 0.05, ...)

## S3 method for class 'discrim':
ROC(object, length = 1000, fig = TRUE,
    se.type = c("CI", "SE"), CI.alpha = 0.05, ...)

Arguments

object the class of the object defines, which of the methods is invoked. If object is a single element numeric vector it is taken as a d-prime value and the default method is invoked. If the object is of class discrim (works for AnotA objects), the method for discrim objects is invoked.

se.d a unit length vector with the standard error of d-prime. If supplied confidence intervals or standard errors are plotted

scale a unit length vector giving the ratio of scale (ie. standard deviation) of the latent distribution for the no-class items relative to that of the yes-class items

length the length of the vectors to be plotted. Longer vectors gives more smooth curves.

fig Should a plot be produced?

se.type The type of band for the ROC curve, "CI" for confidence interval and "SE" for standard error.

CI.alpha the type I level of the confidence interval of AUC

... additional arguments to plot and lines
Details

The function currently ignores the variance of the scale in the computation of the uncertainty of the ROC curve.

Value

The function makes a plot of the ROC curve, and if se.d is supplied, standard errors or confidence intervals for the curve are added to the plot.

The function also (invisibly) returns a list with the following components

- ROCx: x-coordinates to the ROC curve
- ROCy: y-coordinates to the ROC curve
- lower: y-coordinates to the lower limit
- upper: y-coordinates to the upper limit

Author(s)

Rune Haubo B Christensen

Examples

```r
## ROC.default:
(mat <- matrix(c(8, 17, 1, 24), 2, byrow = TRUE))
(d.prime <- SDT(mat, "probit")[3])
ROC(d.prime)
## ROC.discrim:
fm1 <- AnotA(8, 25, 1, 25)
ROC(fm1)
```
Description

The function computes d-prime for any 2 x J table where J >= 2 for the "yes–no" or "A-Not A" experiment using the Signal Detection Theory (SDT) algorithm to compute J-1 d-prime's. The algorithm is also called the "empirical probit transform". The function also provides the "logit" counterpart.

Usage

```r
SDT(tab, method = c("probit", "logit"))
```

Arguments

- `tab`: A 2 x J table with true class relation in rows (only two true classes) and the J-class response in columns
- `method`: should the empirical probit or logit transform be computed?

Value

A (J-1) x 3 matrix. The first two columns contains the z-transform of the Hit rate and the False Alarm rate respectively—ready to plot along with the empirical ROC curve. The third column contains the estimated d-primes.

Author(s)

Rune Haubo B Christensen

References


Examples

```r
### Design table:
## 8 "yes"-responses to no-samples
```
## Betabin

### Description

Fits the beta binomial model to data.

### Usage

```r
betabin(data, start = c(.5,.5), method = c("mu-gamma", "alpha-beta"),
        vcov = TRUE, 
        
        ## S3 method for class 'betabin':
        summary(object, alpha=.05, ...)
```

### Arguments

- **object**: an object of class "betabin", ie. the result of `betabin()`.
- **alpha**: the allowed type I error for confidence intervals
data matrix or data.frame with two columns; first column contains the number of success and the second the total number of cases. The number of rows should correspond to the number of observations.

start starting values to be used in the optimization

vcov logical, should the variance-covariance matrix of the parameters be computed?

method The desired representation. Note that while the "mu-gamma" is often the most natural and easiest to interpret, it can cause convergence problems when overdispersion is close to non-existence.

... additional arguments passed to optim in betabin. Not used in summary.betabin.

Details

The following additional methods are implemented objects of class betabin: print, vcov, logLik and coef.

Value

An object of class betabin with elements

coeff named vector of coefficients

vcov variance-covariance matrix of the parameter estimates
data a named vector with the data supplied to the function
call the matched call

logLik the value of the log-likelihood at the MLEs

method the method used for the fit

correction 0 indicates convergence. For other error messages, see ?optim.

message possible error message - see ?optim for details

counts the number of iterations used in the optimization - see ?optim for details

Author(s)

Rune Haubo B Christensen
References


See Also

triangle, twoAFC, threeAFC, duotrio, discrimPwr, discrimSim, discrimSS, samediff, AnotA, findcr

Examples

```r
## Create data:
x <- c(3, 2, 6, 8, 3, 4, 6, 0, 9, 9, 0, 2, 1, 2, 8, 9, 5, 7)
n <- c(10, 9, 8, 9, 8, 6, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 9, 10)
dat <- data.frame(x, n)

(bb <- betabin(dat, method = "mu-gamma"))
summary(bb)
vcov(bb)
logLik(bb)
AIC(bb)
coef(bb)
```

Description

Fits a cumulative link location-scale model to an ordered response variable. When the scale part is left unspecified, the model reduces to a cumulative link model assuming a constant scale. With the default logistic link function, the model reduces to the famous Proportional Odds Model. With the probit link and a single two-level factor in both location and scale parts, the model is known as the Binormal model in the Signal Detection Theory and the Psychometric literature.
Usage

```r
clls(location, scale, data, weights, start, ..., subset,
     na.action, contrasts = NULL, Hess = FALSE, model = TRUE,
     method = c("logistic", "probit", "cloglog", "cauchit"))
```

Arguments

- **location**: a formula expression as for regression models, of the form `response ~ predictors`. The response should be a factor (preferably an ordered factor), which will be interpreted as an ordinal response, with levels ordered as in the factor. The model must have an intercept: attempts to remove one will lead to a warning and be ignored. An offset may be used. See the documentation of `formula` for other details.

- **scale**: a optional formula expression as for the location part, of the form `~ predictors`, i.e. with an empty left hand side. If left unspecified, the model assumes a constant scale and reduces to the cumulative link model. An offset may be used. See the documentation of `formula` for other details.

- **data**: an optional data frame in which to interpret the variables occurring in `formula`.

- **weights**: optional case weights in fitting. Default to 1.

- **start**: initial values for the parameters. This is in the format `c(beta, theta, sigma)`: see the Values section.

- **...**: additional arguments to be passed to `optim`, most often a `control` argument.

- **subset**: expression saying which subset of the rows of the data should be used in the fit. All observations are included by default.

- **na.action**: a function to filter missing data.

- **contrasts**: a list of contrasts to be used for some or all of the factors appearing as variables in the model formula.

- **Hess**: logical for whether the Hessian (the observed information matrix) should be returned. Use this if you intend to call `summary` or `vcov` on the fit.

- **model**: logical for whether the model matrix should be returned.
method

logistic or probit or complementary log-log or cauchit (cor-
responding to a Cauchy latent variable).

Details

The implementation is highly inspired by polr in package MASS and should
give compatible results, if scale is left unspecified.

Note that standard errors are appropriate for tau = log sigma and not
for sigma, because the profile likelihood is usually more symmetric for tau
than for sigma. Therefore vcov will give the variance-covariance matrix of
the parameters with tau rather than sigma and summary.clls will report
standard errors for log sigma. Notice also that a relevant test for sigma is
H_0: sigma = 1, so the relevant test for log sigma is H_0: log(sigma) = 0.
This is reflected in the z value for sigma returned by summary.clls.

There are methods for the standard model-fitting functions, including summary,
vcov, anova, and an extractAIC method.

Value

A object of class "clls". This has components

coefficients the coefficients of the location (beta), the intercepts (theta)
and the scale (sigma).

beta the parameter estimates of the location part.

theta the intercepts/thresholds for the class boundaries.

sigma the parameter estimates of the scale part.

tau parameter estimates of the scale part on the log scale; ie.
tau = log sigma.

deviance the residual deviance.

fitted.values a matrix, with a column for each level of the response with
the fitted probabilities.

fitted.case a vector of same length as response, with the fitted prob-
abilities on a case-by-case basis.

lev the names of the response levels.

terms.location a terms structure describing the location part.

terms.scale a terms structure describing the scale part.
df.residual  the number of residual degrees of freedoms, calculated using the weights.
edf           the (effective) number of degrees of freedom used by the model
n, nobs      the (effective) number of observations, calculated using the weights.
call          the matched call.
method        the matched method used.
convergence   the convergence code returned by optim.
niter         the number of function and gradient evaluations used by optim.
Hessian       if Hess is true, the observed Fisher information matrix.
location      if model is true, the model.frame for the location part.
scale         if model is true, the model.frame for the scale part.

References


See Also

polr, optim, glm, multinom.

Examples

```r
options(contrasts = c("contr.treatment", "contr.poly"))
## Extend example from polr in package MASS:
## Fit model from polr example:
data(housing, package = "MASS")
f1 <- clls(Sat ~ Infl + Type + Cont, weights = Freq, data = housing)
f1
summary(f1)
```
## With probit link:
summary(update(fm1, method = "probit"))

## Allow scale to depend on Cont-variable
summary(fm2 <- update(fm1, scale =~ Cont))
anova(fm1, fm2)
## which seems to improve the fit

discrim

**Sensory discrimination analysis**

**Description**

Computes the d-prime, its standard error and P-value for the hypothesis of no sensory difference for any of four methods: 2-AFC, 3-AFC, duotrio and triangle tests.

**Usage**

```r
discrim(success, total,
       method = c("duotrio", "threeAFC", "twoAFC", "triangle"), ...)
```

**Arguments**

- `success`: the number of correct answers
- `total`: the total number of answers (sample size)
- `method`: the discrimination protocol. Four allowed values: "twoAFC", "threeAFC", "duotrio", "triangle"
- `...`: additional arguments passed to `glm`

**Details**

The function uses the one of the dedicated binomial families and the base function `glm`
Value

An object of class `discrim` with elements

- **coef**: named vector of coefficients (d-prime and tau)
- **res.glm**: the glm-object from the fitting process
- **vcov**: variance-covariance matrix of the coefficients
- **se**: named vector with standard error of the coefficients (standard error of d-prime)
- **data**: a named vector with the data supplied to the function
- **p.value**: one-sided p-value from Fisher's exact test (`fisher.test`)
- **test**: a string with the name of the test (A-Not A) for the print method
- **call**: the matched call

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References


See Also

`triangle`, `twoAFC`, `threeAFC`, `duotrio`, `discrimPwr`, `discrimSim`, `discrimSS`, `samediff`, `AnotA`, `findcr`

Examples

```R
## Running the simple discrimination test analyses:
discrim(10, 15, "twoAFC")
discrim(10, 15, "threeAFC")
discrim(10, 15, "duotrio")
discrim(10, 15, "triangle")
```
Description

Computes the power of the hypothesis test of no sensory difference for any one of four methods: 2-AFC, 3-AFC, duotrio and triangle tests given the underlying sensory difference delta, the type I test level and the sample size.

Usage

discrimPwr(delta, sample.size, alpha = .05,
             method = c("duotrio", "threeAFC", "twoAFC", "triangle"))

Arguments

delta            The underlying sensory difference (non-negative)
sample.size      The sample size (a positive integer)
alpha            The type I level of the test (must be between zero and one)
method           The discrimination test protocol. Four allowed values: "twoAFC", "threeAFC", "duotrio", "triangle"

Details

The function uses one of the dedicated binomial families.

Value

The power.

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff


References


See Also

triangle, twoAFC, threeAFC, duotrio, discrim, discrimSim, AnotA, discrimSS, samediff, findcr

Examples

```r
## Finding the power of a discrimination test with a sensory delta of 1, 
## a sample of size 30 and a type I level of .05:
discrimPwr(1, 30, 0.05, "twoAFC")
discrimPwr(1, 30, 0.05, "threeAFC")
discrimPwr(1, 30, 0.05, "duotrio")
discrimPwr(1, 30, 0.05, "triangle")
```

Description

The model is a synthesis of a mixture and a mixed effect model. The random effect distribution for the cluster term (often individuals) is a point mass for delta = 0 and a continuous distribution for delta > 0.

The function fits the model and computes d-prime for an average subject, 2) the variance among subjects, 3) the "posterior" probability of a subject being a discriminator (with delta > 0), 4) the "posterior" expectation on the random effect (ie. the subject-specific delta) and 5) the probability that a randomly chosen individual is a discriminator (ie. the probability mass at delta = 0 in the random effects distribution)

Warning: This function is preliminary; see the details for further information.
Usage

discrimR(formula, data, weights, cluster, start, subset, na.action,
        contrasts = NULL, hess = FALSE, ranef = FALSE, zi = FALSE,
        method = c("duotrio", "probit", "threeAFC", "triangle",
                     "twoAFC"), ...)

Arguments

formula A formula where the lhs is the binomial response. An
         indicator vector or a matrix with two column; successes
         and failures like in a call to glm with a binomial family.
         The rhs should be 1; no other predictors are currently
         allowed, but extending this is ongoing work.

data The data.frame in which to look for variables.

weights Possible weights

cluster The clustering variable; should be a factor.

start Optional starting values; recommended in the current im-
        plementation

subset ...

na.action ...

contrasts ...

hess Should the hessian of the parameters be computed?

ranef Should the random effect estimates be computed?

zi Should the posterior probabilities of a subject being a dis-
      criminator be computed?

method Should correspond to the actual test applied.

... Additional arguments to optim. control=list(trace=TRUE,
                                           REPORT=1) is recommended, so the reduction in deviance
      and convergence can be followed.

Details

This function is preliminary and improving it is ongoing work. The com-
putational methods are expected to change completely. This will hopefully
facilitate methods for more general rhs-formulae with additional predictors.
Currently no methods or extractor functions have been written, so the user
will have to select the relevant elements from the fitted object (see below).
Implementation of methods and extractor functions will occur in due course.
Value

A list with the following elements:

- **fpar**
  The fixed effect parameter, i.e. delta (for an average individual)

- **rpar**
  A vector with two elements: The first element is the variance component (standard deviation) on the log-scale, where optimization is performed. The second element is the variance component (standard deviation) on the original scale.

- **deviance**
  Deviance for the model

- **se**
  Standard errors for 1) the fixed effect parameter and 2) the variance component on the log-scale

- **convergence**
  Convergence message from optim

- **lli**
  Log-likelihood contributions from each of the observations.

- **ranef**
  The random effect estimates for the levels of the clustering factor (often individual)

- **zi**
  Posterior probabilities of a subject being a discriminator

- **p**
  The probability that a randomly chosen individual is a discriminator (i.e. the probability mass for delta > 0 in the random effects distribution)

- **fitted**
  Fitted values

- **Y**
  The scaled response vector on which optimization is performed.

- **call**
  The matched call

Author(s)

Rune Haubo B Christensen

See Also

triangle, twoAFC, threeAFC, duotrio, discrimPwr, discrimSim, discrimSS, samediff, AnotA, findcr
Examples

```r
## Not run:
freq <- c(10, 8, 10, 9, 8, 9, 1, 10, 10, 8, 2, 6, 7, 6, 4, 5, 5, 3, 9, 9, 5, 5, 8, 8, 9, 9)
tmp <- data.frame(id = factor(1:30), n = rep(10, 30), freq = freq)
head(tmp)
str(tmp)

fm <- discrimR(cbind(freq, n - freq) ~ 1, tmp, cluster = id,
               start = c(.5, .5), method = "twoAFC",
               ranef = TRUE, zi = TRUE, hess = TRUE,
               control=list(trace=TRUE, REPORT=1))

names(fm)
fm[1:4]
## End(Not run)
```

discrimSS

Sensory discrimination sample size calculation

Description

Computes the sample size for any one of four methods: 2-AFC, 3-AFC, duotrio and triangle tests given the underlying sensory difference $d$, the type 1 test level and the required power.

Usage

```r
discrimSS(delta, power, alpha,
          method = c("duotrio", "threeAFC", "twoAFC", "triangle"))
```

Arguments

- **delta**: the underlying sensory difference (larger than zero)
- **power**: the wanted power (between zero and one)
- **alpha**: the type 1 level of the test (between zero and one)
- **method**: the discrimination protocol. Four allowed values: "twoAFC", "threeAFC", "duotrio", "triangle"
Details

The function uses one of the dedicated binomial families and the `discrimPwr` function.

Value

The sample size

Author(s)

Per Bruun Brockhoff and Rune Haubo B Christensen

References


See Also

`triangle, twoAFC, threeAFC, duotrio, discrim, discrimSim, AnotA, discrimPwr, samediff, findcr`

Examples

```r
## Finding the necessary sample size:
discrimSS(1, 0.9, 0.05, "twoAFC")
discrimSS(1, 0.9, 0.05, "threeAFC")
discrimSS(1, 0.9, 0.05, "duotrio")
discrimSS(1, 0.9, 0.05, "triangle")
```
discrimSim

Simulates replicated difference tests

Description

Simulates the outcome of sample.size replicated sensory difference tests (for any one of four protocols: 2-AFC, 3-AFC, duotrio and triangle tests) for a given d-prime value and a given overdispersion (default 0).

Usage

```r
discrimSim(sample.size, replicates, delta, sd.indiv = 0, 
            method = c("duotrio", "halfprobit", "probit", "triangle", 
                        "twoAFC", "threeAFC"))
```

Arguments

- `sample.size` The sample size - number of subjects
- `replicates` Number of replications per subject
- `delta` The delta (d-prime) value
- `method` The discrimination protocol
- `sd.indiv` the individual variability in d-prime values. A value of 0 (default) is default and corresponds to complete independence

Details

The function uses the four thurstonian binomial families.

Value

A vector of length sample.size with the number of correct answers for each subject.
duotrio

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References


See Also

triangle, twoAFC, threeAFC, duotrio, discrimPwr, discrim, AnotA, discrimSS, samediff, findcr

Examples

```r
## Running simulations:
discrimSim(sample.size = 10, replicates = 3, delta = 2, 
method = "triangle", sd.indiv = 1)
```

duotrio

Create duotrio binomial family

Description

Creates a copy of the binomial family with the inverse link function changed to equal the duotrio psychometric function and correspondingly changed link function and derivative of the inverse link function.

Usage

duotrio()
Value

A binomial family object for models. Among other things it includes the psychometric function as \texttt{linkinv} and the inverse psychometric function (for direct dprime computation) as \texttt{linkfun}

Note

Several functions in this package makes use of the function, but it may also be used on its own—see the example below.

Author(s)

Per Bruun Brockhoff

References


See Also

triangle, twoAFC, threeAFC, discrim, discrimPwr, discrimSim, AnotA, discrimSS, samediff, findcr

Examples

xt <- matrix(c(10, 5), ncol = 2) ## data: 10 correct answers, 5 incorrect
res <- glm(xt ~ 1, family = duotrio)
summary(res)

## Extended example plotting the profile likelihood
## data: 10 correct answers, 5 incorrect
xt <- matrix(c(10, 5), ncol = 2)
summary(res <- glm(xt ~ 1, family = duotrio))
N <- 100
dev <- double(N)
delta <- seq(1e-4, 5, length = N)
for(i in 1:N)
    dev[i] <- glm(xt ~ -1 + offset(delta[i]),
        family = duotrio)$deviance
plot(delta, exp(-dev/2), type = "l",
    xlab = expression(delta),
    ylab = "Normalized Profile Likelihood")
## Add Normal approximation:
lines(delta, exp(-(delta - coef(res))^2 /
    (2 * vcov(res))), lty = 2)
## Add confidence limits:
level <- c(0.95, 0.99)
lim <- sapply(level, function(x)
    exp(-qchisq(x, df=1)/2) )
abline(h = lim, col = "grey")
points(confint(res), rep(lim[1], 2), pch = 4)

findcr

Find the critical value of a one-tailed binomial test

Description

Finds the critical value in a one-tailed binomial test

Usage

findcr(sample.size, alpha = .05, p0 = .5)

Arguments

sample.size the sample size of the binomial test (must be a positive integer)
alpha the type I error-level of the test (must be between zero and one)
p0 the binomial probability under the null-hypothesis (must be between zero and one); 1/2 for the duotrio and twoAFC tests and 1/3 for the triangle and threeAFC tests
Value

The critical value in a one-tailed binomial test, that is, the smallest integer such that the null hypothesis binomial probability of being larger than or equal to this number is smaller than or equal to the type I error-level of the test.

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

See Also

triangle, twoAFC, threeAFC, duotrio, discrim, discrimPwr, discrimSim, AnotA discrimSS, samediff

Examples

```r
## Find the critical value for a triangle test for the level 0.05 test
## with 25 subjects:
findcr(sample.size = 25, , p0 = 1/3)
```

plot.discrim

Plot function for discrim objects

Description

This function plots the latent distributions of sensory intensity corresponding to the items or products tested in the discrimination test.

Usage

```r
## S3 method for class 'discrim':
plot(x, main = TRUE, length = 1000, ...)
```
Arguments

- **x**: The `discrim` object whose latent distributions are to be plotted.
- **main**: Whether to include an automatically generated title on the plot? Default is `TRUE`.
- **length**: The length of the vectors to be plotted. Longer vectors give more smooth curves.
- **...**: Additional arguments to `plot` and `lines`.

Author(s)

Rune Haubo B Christensen

Examples

```r
## Generate discrim objects to be plotted:
fm1 <- discrim(10, 15, "threeAFC")
fm2 <- discrim(10, 15, "triangle")
par(mfrow=c(2,1))  ## Split plotting window in two
## Plot the distributions of sensory intensity for the two objects
## and increase the line width
plot(fm1, lwd=2)
plot(fm2, lwd=2)
```

Description

This function plots the latent distributions of sensory intensity corresponding to the items or products tested in the discrimination test.
Usage

```r
## S3 method for class 'samediff':
plot(x, main = TRUE, length = 1000,
     limits, fig = TRUE, ...)
```

Arguments

- **x**: The `samediff` object whose latent distributions are to be plotted
- **main**: include an automatically generated title on the plot? Default is `TRUE`
- **length**: the length of the vectors to be plotted. Longer vectors gives more smooth curves, but can take a little time.
- **limits**: optional limits on the x-axis; vector of length two.
- **fig**: logical: Should the function create the plot? Defaults to `TRUE`.
- **...**: additional arguments to `plot` and `lines`

Value

If `fig = TRUE`, the function will produce the plot. The function invisibly returns a data.frame with elements

- **z**: values for the x-axis of length `length`.
- **base.dist**: y-values for the base distribution of same-samples, ie. a standard normal distribution
- **delta.dist**: y-values for the distribution of different-samples, ie. a normal distribution centred at `delta` with unit variance.

This facilitates later plotting and changing the appearance of the plot.

Author(s)

Rune Haubo B Christensen
Examples

```r
## Make same-diff object:
sadi <- samediff(8, 5, 4, 9)
## Plot distributions of sensory intensity:
plot(sadi)
```

Description

The function prints the estimated AUC of an AUC object as well as confidence intervals, they are part of the AUC object.

Usage

```r
## S3 method for class 'AUC':
print(x, digits = getOption("digits"), ...)
```

Arguments

- `x`: The AUC object to be printed
- `digits`: the number of digits desired in printing
- `...`: currently not used

Author(s)

Rune Haubo B Christensen

Examples

```r
## Obtain d-prime and standard error for an arbitrary test:
fm1 <- AnotA(8, 25, 1, 25)
```
## Print AUC with and without confidence limits:

```r
AUC(fm1$coef, , fm1$se)
AUC(fm1)
```

---

### print.discrim

*Print method for objects of class "discrim"*

#### Description

The function will print results from a discrimination test stored in an object of class "discrim" made with one of the functions `AnnotA`, `discrim`.

#### Usage

```r
## S3 method for class 'discrim':
print(x, digits = getOption("digits"), alpha = 0.05, ...)
```

#### Arguments

- **x**: an object of class "discrim"
- **digits**: number of digits in resulting table of results
- **alpha**: the allowed type I level of the Wald based confidence interval
- **...**: currently not used

#### Value

A table of results of the discrimination test is printed including estimates, standard errors, lower confidence limit, upper confidence limit and a p-value for a test of the null-hypothesis of the parameter being zero.

#### Author(s)

Rune Haubo B Christensen
References


See Also

AnotA, discrim, discrimPwr, discrimSim, discrimSS, findcr

Examples

```r
## Print results of an A-Not A test:
AnotA(10, 20, 3, 20)
```

---

**profile.discrim**

*Profile likelihood methods for discrim objects.*

**Description**

Computes the (normalized or relative) profile likelihood for the parameters of a discrimination test, plots the normalized profile likelihood and computes profile likelihood confidence intervals.

**Usage**

```r
## S3 method for class 'discrim':
profile(fitted, min = 0, max = 3, numpts = 50, ...)

## S3 method for class 'profile.discrim':
plot(x, level = c(0.99, 0.95), fig = TRUE,
     method = "natural", n = 500, ...)

## S3 method for class 'discrim':
confint(object, parm, level = 0.95, ...)
```
Arguments

- `fitted` a `discrim` object.
- `x` a `profile.discrim` object.
- `object` a `discrim` object.
- `parm` currently not used.
- `min` the minimum delta for which to do the profiling. By default set to 0, which for numerical stability is change internally to 1e-4.
- `max` the maximum delta beyond the MLE for which to do the profiling.
- `numpts` control parameter: At how many points should the profile likelihood be evaluated?
- `method` the type of spline to be used in approximating the profile likelihood curve (trace)—see `spline` for details.
- `n` the number of spline interpolations to use in plotting the profile likelihood curve (trace).
- `level` for `plot`: At which levels to include horizontal lines to indicate confidence levels in plots of the normalized profile likelihoods. For `confint`: at which level to compute the confidence interval.
- `fig` logical: Should the normalized profile likelihoods be plotted?
- `...` For `plot`: additional arguments to `plot`. For `confint`: additional arguments to `confint.glm` in package MASS. For `profile`: additional arguments to `glm`.

Value

For `profile`: An object of class "profile.discrim", "data.frame"—a `data.frame` with two columns giving the value of the parameter and the corresponding value of the profile likelihood.

For `plot`: An object of class "nProfile.discrim", "data.frame"—the `data.frame` from the `profile`-object with an extra columns containing the normalized profile likelihood.

For `confint`:
A 2x2 matrix with columns named "lower", "upper" giving the lower and upper (1 - alpha)% confidence interval for the parameters named in the rows.
Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References


Examples

```r
## 7 success out of 10 samples in a duo-trio experiment:
dd <- discrim(7, 10, "duotrio")
plot(profile(dd))
confint(dd)
points(confint(dd), rep(.1465, 2), pch = 3)
```

profile.samediff

Profile likelihood methods for samediff objects.

Description

Computes the (normalized or relative) profile likelihood for the parameters of a same-different test, plots the normalized profile likelihood and computes profile likelihood confidence intervals.

Usage

```r
## S3 method for class 'samediff':
profile(fitted, which = 1:2, max = 2, numpts = 100,
       max.delta = 10, max.tau = 10, ...)
## S3 method for class 'profile.samediff':
plot(x, which = 1:nc, level = c(0.99, 0.95),
     fig = TRUE, ...)
```
## S3 method for class 'samediff':
confint(object, parm = c("tau", "delta"), level = 0.95, max = c(10, 10), ...)

Arguments

fitted a `samediff` object

x a `profile.samediff` object

object a `samediff` object

which numeric: which parameters to profile or plot; either "1" or "2" or "1:2" to mean "tau", "delta" or both respectively.

parm the parameter(s) to compute the confidence interval for

max for `profile`: control parameter to specify how many units beyond the MLE, the profiling should proceed. For `confint`: control parameter, that can control the convergence for especially very large `delta`

numpts control parameter: At how many points should the profile likelihood be evaluated?

max.delta control parameter: The maximum point at which to evaluate the profile likelihood for `delta`

max.tau same as `max.delta` for "tau".

level for `plot`: At which levels to include horizontal lines to indicate confidence levels in plots of the normalized profile likelihoods. For `confint`: at which level to compute the confidence interval.

fig logical: Should the normalized profile likelihoods be plotted?

... not currently used.

Value

For `profile`: An object of class "`profile.samediff"", "data.frame"”—a `data.frame` with two columns for each parameter profiled giving the value of the parameter and the corresponding value of the profile likelihood.

For `plot`: An object of class "`nProfile.samediff"", "data.frame"”—the `data.frame` from the `profile`-object with extra columns corresponding to the `which` parameter containing the normalized profile likelihood.
For `confint`: A 2x2 matrix with columns named "lower", "upper" giving the lower and upper (1 - alpha)% confidence interval for the parameters named in the rows.

Author(s)

Rune Haubo B Christensen

See Also

`summary.samediff`

Examples

```r
# data: 8 of the same samples were judged to be same
#      5 of the same samples were judged to be different
#      4 of the different samples were judged to be same
#      9 of the different samples were judged to be different

sadi <- samediff(8, 5, 4, 9)
confint(sadi)
plot(profile(sadi))
```

Description

Computation of tau and dprime and their uncertainties for the same different test using maximum likelihood.

Usage

```r
samediff(nsamesame, ndiffsame, nsamediff, ndiffdiff, VCOV = TRUE)
```
Arguments

- **nsamesame**: The number of same-answers on same-samples
- **ndiffsame**: The number of different-answers on same-samples
- **nsamediff**: The number of same-answers on different-samples
- **ndiffdiff**: The number of different-answers on different-samples
- **VCOV**: Should the variance-covariance matrix of the parameters be computed. Defaults to TRUE.

Details

The function computes the maximum likelihood estimates of tau and delta.

Value

An object of class `samediff` with elements

- **coef**: named vector of coefficients (d-prime and tau)
- **vcov**: variance-covariance matrix of the coefficients
- **se**: named vector with standard error of the coefficients (standard error of d-prime)
- **data**: a named vector with the data supplied to the function
- **test**: a string with the name of the test (same-different)
- **call**: the matched call
- **convergence**: convergence indicator. 0 indicates convergence. For error codes see optim.
- **logLik**: Value of the log-likelihood at the MLE of the parameters.
- **case**: A case indicator for internal use

Author(s)

Rune Haubo B Christensen
Examples

    # data: 8 of the same samples were judged to be same
    # 5 of the same samples were judged to be different
    # 4 of the different samples were judged to be same
    # 9 of the different samples were judged to be different

    samediff(8, 5, 4, 9)

Description

Computes the power for a same-different discrimination experiment with a no-difference null hypothesis via simulation.

Usage

    samediffPwr(n = 1000, tau, delta, Ns, Nd, alpha = 0.05)

Arguments

    n    the number of samples to use in the simulation. More samples means higher precision, but takes longer to compute.
    tau    the value of tau
    delta    the underlying sensory difference under the alternative hypothesis (non-negative)
    Ns    the number of same-samples (a positive integer)
    Nd    the number of different-samples (a positive integer)
    alpha    the type I level of the test (must be between zero and one)
Details

The power is computed using simulations. n datasets is simulated from the Same Different model with specified parameters. The power is the fraction of times the p-value is lower than alpha.

Under some parameter combinations, there is a non-significant probability that data will fall, so that the MLE of delta is not defined and the p-value is not defined. All such undefined p-values are silently ignored.

The estimated power may change between runs and especially if the power is either very large or very small (ie. close to 0 or 1). Using more simulations will provide higher accuracy.

It is often a good idea to run the power simulation a couple of times to ensure that the variation in the result is acceptable.

Value

A single numeric value giving the power of the specified test.

Author(s)

Rune Haubo B Christensen

References


See Also

samediff, samediffSim

Examples

## Finding the power of a discrimination test with a sensory delta of 2
## (alternative hypothesis) versus a null hypothesis of delta = 0 with
## a sample of size 2 x 10 and a type I level of .05. n should be higher
## for a reasonable precision:
samediffSim

samediffPwr(n = 100, tau = 1, delta = 2, Ns = 10, Nd = 10)

samediffSim **Simulates data from a samediff test**

Description

Simulates the outcome of n same-different experiments.

Usage

samediffSim(n, tau, delta, Ns, Nd)

Arguments

- **n**: the number of experiments to simulate.
- **tau**: the value of "tau".
- **delta**: the value of delta (d-prime).
- **Ns**: number of same-samples
- **Nd**: number of different-samples

Details

The function makes two calls to `rbinom`.

Value

A matrix of with n rows and four columns named `ss`, `ds`, `sd`, `dd` with the number of same-answers to same-samples, different-answers to same-samples, same-answers to different-samples and different-answers to different-samples respectively.
### Examples

```r
## Running simulations:
samediffSim(n = 10, tau = 1, delta = 1, Ns = 10, Nd = 10)
```

### Description

Makes a summary of a `samediff` object with option to use profile likelihood for confidence intervals and p-values or the assymptotic variance-covariance matrix.

### Usage

```r
## S3 method for class 'samediff':
summary(object, profile = TRUE, ...)
```

### Arguments

- `object`: a `samediff` object
- `profile`: logical: Should the profile likelihood be used for confidence intervals and p-values for the parameters? Defaults to TRUE. If FALSE the assymptotic variance-covariance matrix
derived from the observed Fisher information matrix will be used. See Details for more information.

... can be level, eg 0.95 to specify the confidence level of the intervals.

Details

Note that the variance-covariance matrix does not always exist in contrast to the profile likelihood. profile = FALSE may therefore cause confidence intervals etc. to be NA.

Value

An object of class summary.samediff inheriting elements from the samediff object and with the following additional elements

- **table**
  - matrix with parameter estimates, standard errors, confidence intervals and p-values.
- **AIC**
  - the AIC of the object.

Author(s)

Rune Haubo B Christensen

See Also

confint.samediff, profile.samediff

Examples

```r
# data: 8 of the same samples were judged to be same
# 5 of the same samples were judged to be different
# 4 of the different samples were judged to be same
# 9 of the different samples were judged to be different

sadi <- samediff(8, 5, 4, 9)
summary(sadi)
summary(sadi, FALSE)
```
threeAFC

Create 3-AFC binomial family

Description

Creates a copy of the binomial family with the inverse link function changed to equal the 3-AFC psychometric function and correspondingly changed link function and derivative of the inverse link function.

Usage

threeAFC()

Value

A binomial family object for models. Among other things it includes the psychometric function as linkinv and the inverse psychometric function (for direct dprime computation) as linkfun.

Note

Several functions in this package makes use of the function, but it may also be used on its own—see the example below.

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References

triangle

See Also

triangle, twoAFC, duotrio, discrim, discrimPwr, discrimSim, AnotA, discrimSS, samediff, findcr

Examples

```r
txt <- matrix(c(10, 5), ncol = 2) # data: 10 correct answers, 5 incorrect
res <- glm(txt ~ 1, family=threeAFC)
summary(res)

## Extended example plotting the profile likelihood
## data: 10 correct answers, 5 incorrect
xt <- matrix(c(10, 2), ncol=2)
summary(res <- glm(xt ~ 1, family = threeAFC)#, etastart = etastart))
N <- 100
dev <- double(N)
level <- c(0.95, 0.99)
delta <- seq(1e-4, 5, length = N)
for(i in 1:N)
  dev[i] <- glm(xt ~ -1 + offset(delta[i]),
                family = threeAFC)$deviance
plot(delta, exp(-dev/2), type = "l",
     xlab = expression(delta),
     ylab = "Normalized Profile Likelihood")

## Add Normal approximation:
lines(delta, exp(-(delta - coef(res))^2 /
          (2 * vcov(res))), lty = 2)
## Add confidence limits:
lim <- sapply(level, function(x)
      exp(-qchisq(x, df=1)/2 )
abline(h = lim, col = "grey")
```

triangle

Create triangle binomial family

Description

Creates a copy of the binomial family with the inverse link function changed to equal the triangle psychometric function and correspondingly changed link function and derivative of the inverse link function.
Usage

triangle()

Value

A binomial family object for models. Among other things it includes the psychometric function as linkinv and the inverse psychometric function (for direct dprime computation) as linkfun.

Note

Several functions in this package make use of the function, but it may also be used on its own—see the example below.

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References


See Also

duotrio, twoAFC, threeAFC, discrim, discrimPwr, discrimSim, AnotA, discrimSS, samediff, findcr

Examples

xt <- matrix(c(10, 5), ncol = 2) ## data: 10 correct answers, 5 incorrect
res <- glm(xt ~ 1, family = triangle)
summary(res)

## Extended example plotting the profile likelihood
## twoAFC

**Create 2-AFC binomial family**

**Description**

Creates a copy of the binomial family with the inverse link function changed to equal the 2-AFC psychometric function and correspondingly changed link function and derivative of the inverse link function.

**Usage**

`twoAFC()`

**Value**

A binomial family object for models. Among other things it includes the psychometric function as `linkinv` and the inverse psychometric function (for direct dprime computation) as `linkfun`. 

```r
## data: 10 correct answers, 9 incorrect
xt <- matrix(c(10, 9), ncol = 2)
summary(res <- glm(xt ~ 1, family = triangle))
N <- 100
dev <- double(N)
delta <- seq(1e-4, 3, length = N)
for(i in 1:N)
  dev[i] <- glm(xt ~ -1 + offset(delta[i]),
                family = triangle)$deviance
plot(delta, exp(-dev/2), type = "l",
     xlab = expression(delta),
     ylab = "Normalized Profile Likelihood")

## Add Normal approximation:
lines(delta, exp(-(delta - coef(res))^2 /
          (2 * vcov(res))), lty = 2)

## Add confidence limits:
level <- c(0.95, 0.99)
lim <- sapply(level, function(x) exp(-qchisq(x, df=1)/2) )
abline(h = lim, col = "grey")
```
Note

Several functions in this package makes use of the function, but it may also be used on its own—see the example below.

Author(s)

Rune Haubo B Christensen and Per Bruun Brockhoff

References


See Also

triangle, threeAFC, duotrio, discrim, discrimPwr, discrimSim, AnotA, discrimSS, samediff, findcr

Examples

xt <- matrix(c(10, 5), ncol = 2) ## data: 10 correct answers, 5 incorrect
res <- glm(xt ~ 1, family = twoAFC)
summary(res)

## Extended example plotting the profile likelihood
## data: 10 correct and 8 incorrect:
xt <- matrix(c(10, 8), ncol = 2)
summary(res <- glm(xt ~ 1, family = twoAFC))
N <- 100
dev <- double(N)
level <- c(0.95, 0.99)
delta <- seq(1e-4, 3, length = N)
for(i in 1:N)
  dev[i] <- glm(xt ~ -1 + offset(delta[i]),
                family = twoAFC)$deviance
plot(delta, exp(-dev/2), type = "l",
xlab = expression(delta),
ylab = "Normalized Profile Likelihood")
## Add Normal approximation:
lines(delta, exp(-(delta - coef(res))^2 / (2 * vcov(res))), lty = 2)
## Add confidence limits:
lim <- sapply(level, function(x)
    exp(-qchisq(x, df=1)/2) )
abline(h = lim, col = "grey")
Appendix E

Likelihood, H-likelihood and Mixed Effect Models
E.1 Introduction

The chapter begins with some preliminary likelihood theory. The h-likelihood is a special kind of an extended likelihood. The extended likelihood is, as the name suggests, an extension of the usual likelihood of observed quantities to include quantities not directly observed, called random effects. The h-likelihood is closely connected to generalized linear models (GLMs) and make extensive use of the laplace approximation to integrals. These two issues therefore play a central role in the following treatment. Multivariate GLMs (MGLMs) for ordinal data are an extension of binomial GLMs. They are treated here because such models can be applied to fit Thurstonian models for the A-not A test with sureness, but they have not yet been considered in the h-likelihood framework. Models using the h-likelihood embrace the important linear mixed effect models (LMM) and generalized linear mixed effect models (GLMM). A more traditional approach for fitting GLMMs via the laplace integral approximation and a penalized version of the IWLS-algorithm is briefly reviewed.

The material in this chapter is kept short with focus on definitions and relations rather than examples and interpretation. An exception is however the treatment of the laplace integral approximation, where a concrete example is given. That example is considered again in a modification of the laplace approximation in the chapter on Thurstonian models for replicated difference tests.

Most of the material in this chapter is primarily based on (Pawitan, 2001) and (Lee et al., 2006), so these sources will not be referenced but in special cases. Other sources will be referenced when appropriate. The material on MGLMs is however primarily based on (Fahrmeir and Tutz, 2001). Although the authors of (Lee and Nelder, 2006a) have tried and largely succeeded in giving a comprehensive treatment of the h-likelihood framework, several aspects and important points have been left out of the book. Especially the definition of and computation of different levels of profiling of the h-likelihood for especially binary data is not particular clear from (Lee and Nelder, 2006a). Details of the fitting algorithm for GLMMs is also much clearer from the original papers. Studying the following original papers has therefore been very illuminating (Nelder and Lee, 1991; Lee and Nelder, 1998; 1996; 2000; 2001; 2002; 2003; Yun and Lee, 2004; Lee and Nelder, 2005; Lee et al., 2006; 2006b; Lee et al., 2007; Noh and Lee, 2007). The following papers, primarily on generalized linear models and quasi-likelihood were also very helpful in providing more background on the h-likelihood (Pierce and Schafer, 1986; Nelder and Pregibon, 1987; Cox...
E.2 Preliminary Likelihood Theory

The likelihood function

The likelihood function $L(\theta; y)$ assuming a stochastic model $f_\theta(y)$ is the probability of the observed data $y$ taken as a function of $\theta$.

The connection between the density $f_\theta(y)$ and the likelihood $L(\theta; y)$ is

$$L(\theta; y) = f_\theta(y).$$

On the left hand side the observed data $y$ is fixed and $\theta$ varies, while on the right hand side $\theta$ is fixed and $y$ is random. The function $f_\theta(y)$ describes where $y$ will occur as a function of the parameter, $\theta$, while $L(\theta; y)$ describes the distribution of information of where $\theta$ might be, given the data $y$.

The likelihood from several data sets comprising independent pieces of information is

$$L(\theta; y_1, y_2) = f_\theta(y_1)f_\theta(y_2) = L(\theta; y_1)L(\theta; y_2),$$

which on the log-scale becomes additive and the log likelihood (short for log-likelihood) is given by

$$l(\theta) = \log L(\theta) = \log L(\theta; y_1) + \log L(\theta; y_2).$$

Invariance

Suppose a transformation of the observed data $x$ is $y = g(x)$. Then the likelihood based on $y$ is

$$L(\theta; y) = L(\theta; x) \left| \frac{\partial x}{\partial y} \right|. $$

The likelihood ratio, however is transformation invariant

$$\frac{L(\theta_2; y)}{L(\theta_1; y)} = \frac{L(\theta_2; x)}{L(\theta_1; x)},$$

since the Jacobian $J = \left| \frac{\partial x}{\partial y} \right|$ cancels out and the proportional likelihoods carry the same information. In fact, due to the likelihood principle (Birnbaum 1962), the likelihood function carries all the information in data about a parameter $\theta$. Further the likelihood function is minimal sufficient (Pawitan 2001).
Functions of the Likelihood  

The score function is the first derivative of the log likelihood

\[ S(\theta) = \frac{\partial}{\partial \theta} l(\theta) , \]

and the MLE \( \hat{\theta} \) is the solution to the score equation

\[ S(\theta) = 0 , \]

such that it satisfies \( S(\hat{\theta}) = 0 \).

The Fisher information is defined as

\[ I(\theta) = -\frac{\partial^2}{\partial \theta^2} l(\theta) . \]

The observed Fisher information \( I(\hat{\theta}) \) is the Fisher information evaluated at the MLE \( \hat{\theta} \), and it measures the curvature of the log likelihood at \( \hat{\theta} \) which may be interpreted as the information about \( \hat{\theta} \) in the data.

Under some regularity conditions, the score function has the properties

\[ E_\theta S(\theta) = 0 \]
\[ \text{var}_\theta S(\theta) = E_\theta I(\theta) = I(\theta) . \]

The last quantity is the expected Fisher information, which evaluated at the MLE is \( I(\hat{\theta}) \) and could be called the estimated expected Fisher information.

Profile likelihood  

Given a joint likelihood \( L(\theta, \eta) \), the profile likelihood is \( L(\theta, \hat{\eta}_\theta) \), where it is emphasized that in general, \( \eta \) is a function of \( \theta \).

The profile likelihood is not a proper likelihood, since it is not based on a probability of some observed data. Typically the score function does not have zero expectation and variance matching the expected Fisher information and may lead to over-precision and biased estimates.

Approximations and modified profile likelihoods  

The log likelihood can be approximated with a second order Taylor expansion around the MLE \( \hat{\theta} \)

\[ l(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^t I(\hat{\theta})(\theta - \hat{\theta}) \]
\[ = \log L(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^t I(\hat{\theta})(\theta - \hat{\theta}) , \]
E.2 Preliminary Likelihood Theory

from which we can get

\[
\log \frac{L(\theta)}{L(\hat{\theta})} = -\frac{1}{2} (\theta - \hat{\theta})^t I(\hat{\theta})(\theta - \hat{\theta}) ,
\]  

(E.2.1)

so the quadratic approximation of the loglihood is exactly treating \( \hat{\theta} \) as normally distributed, which holds approximately in many cases.

Asymptotically we therefore have that \( \hat{\theta} \sim N(\theta, I^{-1}(\hat{\theta})) \) and

\[
2 \log \frac{L(\hat{\theta})}{L(\theta)} = (\theta - \hat{\theta})^t I(\hat{\theta})(\theta - \hat{\theta}) \sim \chi^2_p ,
\]

where \( p \) is the dimension of \( \theta \). An approximate density of \( \hat{\theta} \) is therefore

\[
f(\hat{\theta}) = \frac{|I(\hat{\theta})/(2\pi)|^{1/2}}{L(\hat{\theta})} \exp \left\{ -\frac{1}{2} (\theta - \hat{\theta})^t I(\hat{\theta})(\theta - \hat{\theta}) \right\} ,
\]

which immediately leads to the highly accurate magic formula

\[
f(\hat{\theta}) = \frac{1}{L(\hat{\theta})} \left| \frac{\partial \hat{\eta}}{\partial \theta} \right| = \frac{1}{L(\hat{\theta})} L(\hat{\theta}, \hat{\eta}) ,
\]

(E.2.2)

The multi-parameter equivalent is

\[
f(\hat{\theta}, \hat{\eta}) = \frac{|I(\hat{\theta}, \hat{\eta})/(2\pi)|^{1/2}}{L(\hat{\theta}, \hat{\eta})} \cdot
\]

Let \( \hat{\eta}_\theta \) be the MLE of \( \eta \) at fixed \( \theta \) and \( I(\hat{\eta}_\theta) \) the corresponding Fisher information. At fixed \( \theta \) the approximate density of \( \hat{\eta} \) is

\[
f(\hat{\eta}_\theta) = \frac{|I(\hat{\eta}_\theta)/(2\pi)|^{1/2}}{L(\theta, \hat{\eta}_\theta)} \left| \frac{\partial \hat{\eta}}{\partial \theta} \right| ,
\]

where \( L(\theta, \hat{\eta}_\theta) \) is the profile likelihood of \( \theta \). The marginal distribution of \( \hat{\eta} \) is

\[
f(\hat{\eta}) = f(\hat{\eta}_\theta) \left| \frac{\partial \hat{\eta}}{\partial \theta} \right| \approx \frac{1}{L(\theta, \hat{\eta}_\theta)} \left| \frac{\partial \hat{\eta}}{\partial \theta} \right| ,
\]

hence the conditional distribution of \( \hat{\theta} \) given \( \hat{\eta} \) is

\[
f(\hat{\theta}|\hat{\eta}) = \frac{f(\hat{\theta}, \hat{\eta})}{f(\hat{\eta})} \approx \frac{1}{L(\theta, \hat{\eta})} \left| \frac{\partial \hat{\eta}}{\partial \theta} \right| .
\]
where the constant $c$ is approximately $(2\pi)^{1/2}$.

The modified loglihood resembling an approximate conditional likelihood is

$$l_m(\theta) = l(\theta, \hat{\eta}_\theta) - \frac{1}{2} \log |I(\hat{\eta}_\theta)/(2\pi)| + \log \left| \frac{\partial \hat{\eta}}{\partial \hat{\eta}_\theta} \right|,$$

where $l(\theta, \hat{\eta}_\theta)$ is the ordinary profile loglihood of $\theta$.

In lucky cases the parameters $(\theta, \eta)$ are information orthogonal such that $\left| \frac{\partial \hat{\eta}}{\partial \hat{\eta}_\theta} \right| = 1$ i.e. $E(\partial^2 l/\partial \theta \partial \eta) = 0$, which leads to an adjusted profile loglihood (Cox and Reid 1987)

$$p_\eta(l|\theta) = l(\theta, \hat{\eta}_\theta) - \frac{1}{2} \log |I(\hat{\eta}_\theta)/(2\pi)|,$$  

which we may be abbreviate $p_\eta(l)$. Further, for (computational) convenience the expected Fisher information may be used instead of the observed.

**Integrated likelihoods**  
For scalar parameters the integrated likelihood to (E.2.1) is

$$\int L(\theta) \, d\theta \approx \int \exp(\log L(\hat{\theta}) - I(\hat{\theta})(\theta - \hat{\theta})^2/2) \, d\theta$$

$$= L(\hat{\theta})|I(\hat{\theta})/(2\pi)|^{-1/2},$$

where the last expression is the Laplace integral approximation. It is highly accurate, when the integrand is close to quadratic. For a two-parameter model, we have

$$L_{int}(\theta) = \int L(\theta, \eta) \, d\eta \approx L(\theta, \hat{\eta}_\theta)|I(\hat{\eta}_\theta)/(2\pi)|^{-1/2},$$

which on the log scale becomes

$$l_{int}(\theta) \approx l(\theta, \hat{\eta}_\theta) - \frac{1}{2} \log |I(\hat{\eta}_\theta)/(2\pi)|,$$

hence, given orthogonal parameters, the integrated likelihood is approximately the adjusted profile loglihood (E.2.3).

**E.3 Laplace’s Integral Approximation**

This section illustrates the laplace approximation to integrals with a concrete example and place more focus on interpretation than do other sections in this chapter.
In essence the Laplace approximation approximates the integrand by a unnormalized Gaussian distribution. This is done by a second order Taylor expansion of the logarithm of the integrand around the mode of the integrand. As an example consider the psychometric function for the 3AFC method (see eg. [Frijters 1979], [Ennis 1993], [Bi 2006b])

\[ f(\delta) = \int_{-\infty}^{\infty} \Phi^2(z) \phi(z - \delta) \, dz. \]  

(E.3.1)

Here the integrand is \( h(\delta, z) = \Phi^2(z) \phi(z - \delta) \). We now approximate \( h(\delta, z) \) by a Gaussian function \( h^*(\delta, z) \) ie. by a second order Taylor expansion around the mode of the logarithm of integrand.

\[ \log h(\delta, z) \approx \log h^*(\delta, z) = \log h(\delta, z_0) - \frac{1}{2} D(z - z_0)^2, \]

where

\[ D = -\frac{\partial^2}{\partial z^2} \log h(\delta, z) \bigg|_{z = z_0} \]

and \( z_0 \) is the mode - the solution to the score equation

\[ S(z) = \frac{\partial}{\partial z} \log h(\delta, z) = 0 \]

The integrand \( h() \) and the Gaussian approximation \( h^(*) \) is shown in figure E.1. The corresponding quadratic approximation on the log-scale is shown in figure E.2. At \( \delta = 1 \) we find

\[ z_0 = 1.35, \quad h(1, z_0) = 0.3118, \quad D = 1.536 \]

We then approximate the integral (E.3.1) as

\[ f(\delta) \approx \int_{-\infty}^{\infty} \exp \left( \log h(\delta, z_0) - \frac{1}{2} D(z - z_0)^2 \right) \, dz \]

\[ = h(\delta, z_0) \sqrt{\frac{2\pi}{D}} \]

At \( \delta = 1 \) we find that the Laplace approximation to \( f(1) = .6337 \) is \( \tilde{f}(1) = 0.6305 \), which corresponds to a relative error of 0.51% and an absolute error of 0.32%. The psychometric function and its Laplace approximation is shown for a range of values of \( d' \) in figure E.3.
Figure E.1: The integrand $h(\delta = 1, z)$ (black solid) and the approximation $h^*(\delta = 1, z)$ (red dashed).
Figure E.2: The logarithm of the integrand $\log h(\delta = 1, z)$ (Black solid) and the quadratic approximation to it (red dashed).
Figure E.3: The 3AFC psychophysics function (black solid) and the approximation using Laplace’s integral approximation (red dashed).
E.4 Generalized Linear Models

Generalized linear models are regression models, which are characterized by assuming that the response follows a distribution from the exponential (dispersion) family and a link function linking the expected response to the linear predictor. The choice of distribution specifies a particular mean-variance relationship.

The exponential dispersion family has log density for the $i$th observation $y_i$ with known prior weight $w_i$
\[
\log f(y_i) = w_i \left( \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) .
\] (E.4.1)

This family embrace the important distributions; Binomial, Poisson, Gaussian and Gamma.

The mean and variance are
\[
E(y_i) = b'(\theta_i) = \mu_i \\
\text{var}(y_i) = \frac{\phi}{w_i} b''(\theta_i) = \frac{\phi}{w_i} V(\mu_i) .
\]

Since the mean only depends on $\theta$, the canonical parameter, the term $c(y, \phi)$ may be left unspecified. The variance depends on $\theta$ through $V(\mu)$ and $\phi$, which does not depend on the mean $\mu$ and may account for variance, not captured by the mean-variance relationship.

The systematic part of the model is described by the linear predictor and the link function
\[
g(\mu) = \eta = X \beta ,
\]
where $g()$ is the link function, $\eta$ is the linear predictor, $X$ is a $N \times p$ matrix of predictors and $\beta$ is a $p$-vector of (fixed) parameters. If $\theta = g()$, we have the canonical link function, which simplifies estimation in some cases, but need not be relevant for a particular application.

Iterative Weighted Least Squares  The MLE of $\beta$ can (but need not) be obtained with IWLS, which is also called iterative re-weighted least squares (IRLS), since the weights $W$ are recalculated in each iteration. The IWLS can be derived as follows. We seek the set of parameters, which maximize the loglikelihood, which is mathematically equivalent to (E.4.1). The MLEs satisfy the score equations
\[
S(\theta_i) = \frac{\partial l(\theta_i, y_i)}{\partial \theta_i} = \frac{w_i}{\phi} \{ y_i - b'(\theta_i) \} ,
\] (E.4.2)
which is also the derivative of the likelihood kernel

\[ w_i y_i \theta_i - b(\theta_i) / \phi. \]

We seek the parameter estimates on the \( \beta \) scale, hence we write

\[
S(\beta) = \frac{\partial l(\beta, \phi; y)}{\partial \beta} = \sum_i \frac{\partial l(\theta_i, \phi; y_i)}{\partial \theta_i} \frac{\partial \theta_i}{\partial \beta} = w_i \phi^{-1} \sum_i \frac{\partial \theta_i}{\partial \beta} (y_i - \mu_i)
\]

\[
= w_i \phi^{-1} \sum_i \frac{\partial \mu_i}{\partial \beta} V(\mu_i)^{-1}(y_i - \mu_i)
\]

First note that

\[
\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \eta_i} x_i.
\]

Then we linearize \( \mu_i \) around an initial estimate \( \mu_i^0 \)

\[
\mu_i \approx \mu_i^0 + \frac{\partial \mu_i}{\partial \beta} (\beta - \beta^0)
\]

\[
= \mu_i^0 + \frac{\partial \mu_i}{\partial \eta_i} x_i (\beta - \beta^0),
\]

hence

\[
y_i - \mu_i \approx y_i - \mu_i^0 - \frac{\partial \mu_i}{\partial \eta_i} x_i (\beta - \beta^0).
\]

The next step is to put this into the score equation \((E.4.2)\). We obtain

\[
\sum_i \frac{\partial \mu_i}{\partial \eta_i} V(\mu_i)^{-1} x_i \{y_i - \mu_i^0 - \frac{\partial \mu_i}{\partial \eta_i} x_i (\beta - \beta^0)\} = 0.
\]

We may write this as

\[
(X^T \Sigma^{-1} X) \beta = X^T \Sigma^{-1} z,
\]

where \( \Sigma^{-1} \) is a diagonal matrix with elements

\[
\Sigma_i^{-1} = w_i \phi^{-1} \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^{-2} V(\mu_i)^{-1}
\]

and \( z \) is a vector with elements

\[
z_i = \eta_i^0 + \frac{\partial \eta_i}{\partial \mu_i} (y_i - \mu_i^0).
\]
We may write $\Sigma^{-1} = W/\phi$, where $W$ can be used as weights in the algorithm, since the estimates of $\beta$ do not depend on $\phi$. The elements of $W$ and $z$ depend on the current estimate of $\beta$ and are updated after each iteration.

Note that the IWLS algorithm simplifies to ordinary least squares (OLS) when the distribution is normal and the link is the identity link in which case only one iteration is needed to obtain the MLE $\hat{\beta}$.

**Deviance** In GLMs the likelihood ratio statistic is called the (unscaled) deviance and is defined as twice the difference in the maximal obtainable loglikelihood and that under the current model

$$D = 2\{l(y; y) - l(\hat{\mu}; y)\}.$$

In normal models the scaled deviance $D/\phi$ is the scaled residual sum of squares, i.e. the observed mean square, having an exact $\chi^2_{N-p}$ distribution with $N-p$ degrees of freedom. In general the distribution of the scaled deviance is only approximate and in special cases not good at all (such as when data are binary).

The deviance can be decomposed into a sum of deviance components

$$D = \sum d_i,$$

which leads to the deviance residuals

$$r_{di} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i} \quad (E.4.4)$$

In non-normal GLMs, the deviance residuals are typically closer to normal than alternatives such as the Pearson residuals (Pierce and Schafer 1986)

$$r_{pi} = \frac{y_i - \hat{\mu}_i}{(V(\hat{\mu}_i)/w_i)^{1/2}}.$$

The leverage for GLMs can be defined as the diagonal elements of

$$X(X^tWX)^{-1}X^tW \quad (E.4.5)$$

**Example: Gamma GLM** The log-density of the gamma distribution reads

$$\log f(y) = \alpha(-y/\mu - \log \mu) + \alpha \log y + \alpha \log \alpha - \log y - \log \Gamma(\alpha), \quad (E.4.6)$$

which can be identified as an instant of (E.4.1) with elements

$$\theta = -1/\mu \quad b(\theta) = -\log(-\theta) = \log \mu \quad b'(\theta) = 1/\theta \quad b''(\theta) = -1/\theta^2$$

$$\phi = 1/\alpha \quad c(y, \phi) = \alpha \log y + \alpha \log \alpha - \log y - \log \Gamma(\alpha)$$

$$E(y) = \mu \quad \text{var}(y) = \phi V(\mu) = \phi \mu^2$$
Now assume the following link between the linear predictor \( \eta \) and \( \mu \) is given by

\[
\eta = g(\mu) = X\beta,
\]

where we take \( g() \) to be the logarithm. The IWLS algorithm (E.4.3) then has the elements

\[
\frac{\partial \eta}{\partial \mu} = \frac{\partial}{\partial \mu} \log \mu = 1/\mu
\]

\[
W_i = w_i(\partial \eta_i/\partial \mu_i)^{-2}V(\mu_i)^{-1} = w_i\mu_i^{-2} = w_i
\]

\[
z = \eta + (y - \mu)(\partial \eta/\partial \mu) = X\beta + (y - \mu)1/\mu = X\beta + y/\mu - 1,
\]

and

\[
\text{var}(\hat{\beta}) = \phi(X^tWX)^{-1} = \phi(X^tX)^{-1},
\]

where the last equality holds if \( w_i = 1 \) for all \( i \). The deviance components are

\[
d_i = 2\left\{ l(y_i; y_i) - l(\hat{\mu}_i; y_i) \right\}
\]

\[
= 2\left\{ -y_i/y_i - \log(y_i) + \hat{\mu}_i/y_i - \log(\hat{\mu}_i) \right\}
\]

\[
= 2\left\{ - \log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i \right\}.
\]

**Example: Poisson GLM**  The log density of the Poisson distribution reads

\[
\log f(y) = \exp\{y \log(\mu) - \mu - \log(y!)\}
\]

which can be identified as an instant of (E.4.1) with elements

\[
\theta = \log(\mu) \quad b(\theta) = \exp(\theta) \quad b'(\theta) = \exp(\theta) \quad b''(\theta) = \exp(\theta)
\]

\[
\phi = 1 \quad c(y, \phi) = -\log(y!)
\]

\[
E(y) = \mu \quad \text{var}(y) = \phi V(\mu) = \mu
\]

Now assume the following link between the linear predictor \( \eta \) and \( \mu \) is

\[
\eta = g(\mu) = X\beta,
\]

where we take \( g() \) to be the logarithm. The IWLS algorithm (E.4.3) then has the elements

\[
\frac{\partial \eta}{\partial \mu} = \frac{\partial}{\partial \mu} \log \mu = 1/\mu
\]

\[
W_i = w_i(\partial \eta_i/\partial \mu_i)^{-2}V(\mu_i)^{-1} = w_i\mu_i^{-2}\mu_i^{-1} = \mu_i
\]

\[
z = \eta + (y - \mu)(\partial \eta/\partial \mu) = X\beta + (y - \mu)1/\mu = X\beta + y/\mu - 1.
\]

The deviance components are

\[
d_i = 2\left\{ l(y_i; y_i) - l(\hat{\mu}_i; y_i) \right\}
\]

\[
= 2\left\{ y_i \log(y_i) - y_i - y_i \log(\hat{\mu}_i) + \hat{\mu}_i \right\}
\]

\[
= 2\left\{ y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i) \right\}.
\]
E.4 Generalized Linear Models

Example Binomial GLM  The log density of the binomial distribution reads

\[
\log f(y_i) = y_i \log \frac{p}{1-p} + m_i \log(1-p) + \log \left( \frac{m_i}{y_i} \right),
\]

\[
\log f(y_i^*) = m_i \left\{ y_i^* \log \frac{p}{1-p} + \log(1-p) \right\} + \log \left( \frac{m_i}{m_i y_i^*} \right),
\]

where \( y_i^* = y_i / m_i \) is the scaled response and \( y_i \) is the unscaled response; the number of successes out of \( m_i \) trials. For Bernoulli data the two formulations are identical and \( y_i^* = y_i \). The latter formulation allows estimation for grouped data to proceed identical to that of ungrouped data with a modified response \( y_i^* \) and prior weights \( w_i = m_i \).

The log density can be identified as an instant of (E.4.1) with elements

\[
\theta = \log \left( \frac{p}{1-p} \right), \quad p = \frac{\exp(\theta)}{1 + \exp(\theta)},
\]

\[
b(\theta) = -m \log(1-p) = m \log(1 + \exp(\theta)), \quad b^*(\theta) = \log(1 + \exp(\theta)),
\]

\[
b'(\theta) = \frac{m \exp(\theta)}{1 + \exp(\theta)} = mp \quad b''(\theta) = mp(1-p) \quad b'^*(\theta) = p \quad b''^*(\theta) = p(1-p)
\]

\[
\phi = 1 \quad c(y_i, \phi) = \log \left( \frac{m_i}{y_i} \right), \quad c(y_i^*, \phi) = \log \left( \frac{m_i}{m_i y_i^*} \right)
\]

\[
E(y_i) = \mu_i = m_i p \quad \text{var}(y_i) = \phi V(\mu_i) = \phi m_i p_i (1 - p_i) = \phi \mu_i (1 - \frac{\mu_i}{m_i})
\]

\[
E(y^*) = \mu^* = p \quad \text{var}(y^*) = \phi V(\mu^*) = \phi p(1-p),
\]

where starred versions relate to the distribution of \( y_i^* \) rather than \( y_i \). The deviance components are

\[
d_i = 2w_i \{ l(y_i; y_i) - l(\hat{\mu}_i; y_i) \}
\]

\[
= 2w_i \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right\}
\]

\[
= 2w_i m_i \left\{ y_i^* \log \frac{y_i^*}{\hat{p}_i} + (1 - y_i^*) \log \frac{1 - y_i^*}{1 - \hat{p}_i} \right\},
\]

The elements of IWLS are much simpler for \( \mu^* \) than for \( \mu \), hence we will work on that scale in the following. Now assume the (canonical) logit link

\[
\eta = g(\mu^*) = \log \left( \frac{p}{1-p} \right) = X\beta.
\]
The IWLS algorithm (E.4.3) then has the elements
\[
\frac{\partial \eta}{\partial \mu^*} = \frac{\partial}{\partial \log p} \log \frac{p}{1-p} = \frac{1}{p(1-p)} = V^{-1}(\mu^*)
\]

\[
W_i = w_i m_i (\partial \eta_i / \partial \mu^*_i)^{-2} V(\mu^*_i)^{-1} = w_i m_i V(\mu^*_i) = w_i m_i p_i (1 - p_i)
\]

\[
z = \eta + (y - \mu^*)(\partial \eta / \partial \mu^*) = X\beta + \frac{y - p}{p(1-p)}.
\]

If instead we assume the probit link, we have
\[
\eta = g(\mu^*) = \Phi^{-1}(\mu^*) = \Phi^{-1}(p) = X\beta,
\]
where \(\Phi()\) is the Gaussian cdf. The IWLS algorithm (E.4.3) then has the elements
\[
\frac{\partial \mu^*}{\partial \eta} = \frac{\partial}{\partial \eta} \Phi(\eta) = \varphi(\eta)
\]

\[
W_i = w_i m_i (\partial \mu^*_i / \partial \eta_i)^2 V(\mu^*_i)^{-1} = w_i m_i \varphi^2(\eta_i) \{p_i(1 - p_i)\}^{-1}
\]

\[
z = \eta + (y - \mu^*)(\partial \eta / \partial \mu^*) = X\beta + (y - p)\varphi^{-1}(\eta),
\]
where \(\varphi()\) is Gaussian PDF.

### E.5 Extended Quasi-Likelihood

**Quasi likelihood** In 1974 Wedderburn coined the concept of quasi-likelihood. He defined it as a function \(q(\mu_i; y_i)\) satisfying
\[
\frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)},
\]
which resembles the score equations for a GLM
\[
\sum_i \frac{\partial q(\mu_i; y_i)}{\partial \mu_i} = \sum_i \frac{\partial \mu_i (y_i - \mu_i)}{\phi V(\mu_i)},
\]
but one may treat it merely as an estimating equation.

Wedderburn’s theory assumed that \(\phi\) is known and in itself it does not provide a means of estimating \(\phi\). It does however allow the variance to depend on \(\phi\), thereby extending the standard GLM to allow for eg. Poisson and binomial models with \(\phi \neq 1\).

Several variance estimators of the parameters are possible for quasi likelihood models; Hessian based estimates and *sandwich* based method-of-moments estimates.
Extended quasi likelihood  Wedderburn proposed to use a method of moments estimate for $\phi$

$$\tilde{\phi} = \frac{1}{N - p} \sum_i \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

with $\mu$ evaluated at the MLE of $\beta$.

Using the magic formula (E.2.2) for fixed $\phi$, an approximate distribution of $b'(\hat{\theta}) = \hat{\mu} = y$ is

$$f(y) \approx f(\hat{\theta}) \left| \frac{\partial \hat{\theta}}{\partial \hat{\mu}} \right|^{-1} = \left\{ 2\pi \phi b''(\hat{\theta}) \right\}^{-1/2} \frac{L(\theta, \phi)}{L(\hat{\theta}, \phi)}.$$ 

With the deviance being

$$D(y, \mu) = 2 \log \frac{L(\hat{\theta}, \phi = 1)}{L(\theta, \phi = 1)},$$

an approximate log density of $y$ is

$$\log f(y) \approx -\frac{1}{2} \log \left\{ 2\pi \phi V(y) \right\} - \frac{1}{2\phi} D(y, \mu),$$

where $V(y) = b''(\hat{\theta})$. This leads to the extended quasi likelihood for observation $y_i$, which really is a quasi loglikelihood

$$q_i(\mu_i, \phi; y_i) = -\frac{1}{2} \log \left\{ 2\pi \phi V(y_i) \right\} - \frac{1}{2\phi} d(y_i, \mu_i). \quad (E.5.1)$$

The pseudo likelihood can be obtained by using the Pearson deviate $X^2 = \sum_i (y_i - \mu_i)^2 / V(\mu_i)$ instead of the deviance

$$PL(\mu_i, \phi; y_i) = -\frac{1}{2} \log \left\{ 2\pi \phi V(y_i) \right\} - \frac{1}{2\phi} \frac{(y_i - \mu_i)^2}{V(\mu_i)}.$$ 

An approximate profile likelihood of $\phi$ is obtained by using $\hat{\mu}$ rather than $\mu$ in the deviance, hence

$$l(\phi, \hat{\mu}) \approx -\frac{1}{2} \log \left\{ 2\pi \phi V(y) \right\} - \frac{1}{2\phi} D(y, \hat{\mu}). \quad (E.5.2)$$

In the normal case, this is exactly the profile loglikelihood of the variance $\sigma^2$ and the deviance statistic is a $\phi \chi^2$-variate, ie. a gamma-variate with mean $\phi$ and
variance $2\phi^2$. In the non-normal case, the deviance statistic is approximately $\phi \chi^2$ distributed. This amounts to assuming that the deviance residuals (E.4.4) are normally distributed.

### Joint GLM of mean and dispersion

The quasi loglikelihood (E.5.1) may be rewritten in the form

$$q_i(\mu_i, \phi_i; y_i) = \frac{1}{2} \{-d_i/\phi_i - \log \phi_i\} - \frac{1}{2} \log(V(y_i)),$$

which takes the form of gamma distribution (E.4.6) with independent variable $d_i$, dispersion parameter $\alpha = 1/2$, mean $E(d_i) = \phi_i$ and variance $\text{var}(d_i) = 1/\alpha \phi_i^2 = 2\phi_i^2$.

This identification suggests an alternating algorithm, where we estimate the $\hat{\mu}_i$ in an appropriate GLM and estimate $\phi_i$ (resembling dispersion elements or inverse prior weights) with a gamma GLM. We therefore have the two interlinked GLMs

$$E(y_i) = \mu_i \hspace{1em} \eta_i = g(\mu_i) = x_i^T \beta \hspace{1em} \text{var}(y_i) = \phi_i V(\mu_i)$$

$$E(d_i) = \phi_i \hspace{1em} \zeta_i = h(\phi_i) = g_i^T \gamma \hspace{1em} \text{var}(d_i) = 2\phi_i^2.$$

We can optimize the parameters using IWLS for each of the models by fixing $\phi_i$ in the estimation of $\hat{\mu}_i$ in the mean-model and using the deviance components $d_i$ from the mean-model in the estimation of $\hat{w}_i$ in the dispersion-model.

### REML procedure for joint GLMs

The MLEs of dispersion parameters tend to be biased and we should strive to use a REML procedure instead. We can obtain an approximate REML procedure by using the adjusted profile loglikelihood for the dispersion parameters

$$p_{\beta}(q) = [q - \{\log \det(\mathcal{I}(\hat{\beta}_\gamma)/(2\pi))\}/2]|_{\beta = \hat{\beta}},$$

where $\mathcal{I}(\hat{\beta}_\gamma) = \frac{1}{\sigma}(X^tWX)$ is the observed Fisher information in case of the canonical link function and the expected otherwise. Since this adjusted profile loglikelihood involves slow convergence, Lee and Nelder (1998) proposed to alter the response of the dispersion-model to $d_i^* = d_i/(1 - h_i)$, where $h_i$ are diagonal elements of the projection or hat matrix from the mean model—the leverage (E.4.5).
E.6 Cumulative link Models

Cumulative link models are models for ordinal data, where the ordered nature of the data is used in the definition and estimation of model and parameters. Basic instances of the cumulative link models, can be seen as multivariate generalized linear models (MGLM). Cumulative link models does however embrace more involved models, which are not contained the MGLM class. Cumulative link models are often motivated as a model for latent distributions, where the link function defines the cdf of the latent distribution. The class of MGLMs allows modelling of change in location of these latent distributions, but does not include modelling of change in scale of the latent distributions.

The multivariate exponential family  A random vector $y$ follows a multivariate exponential family, if its log density is on the form

$$\log f(y|\theta, \phi) = \frac{y'\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

here $y = (y_1, \ldots, y_q)^t$ and $\theta = (\theta_1, \ldots, \theta_q)^t$. The first two moments are

$$E(y) = \frac{\partial b(\theta)}{\partial \theta} = b'(\theta) \quad \text{var}(y) = a(\phi) \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta^t}$$

The multinomial distribution  Let $Y$ be a categorical variable, that can fall in each of $J$ distinct categories. Then $Y$ can be represented by the vector

$$y = (y_1, \ldots, y_j, \ldots, y_q), \quad j = 1, \ldots, q = J - 1$$

where $y_j = 1$ if $Y$ falls in the $j$th category and zero otherwise. We therefore have the following relation.

$$P(Y = j) = P(y_j = 1) = \pi_j$$

The PDF of $y$ is

$$P(y) = \prod_{j=1}^{J} \pi_j^{y_j} ,$$

where $\pi_j = 1 - \sum_{j=1}^{q} \pi_j$ and $y_j = 1 - \sum_{j=1}^{q} y_j$. The first two moments are

$$E(y) = \pi = (\pi_1, \ldots, \pi_q)^t \quad \text{var}(y) = \text{diag}(\pi) - \pi \pi^t$$

where the former is a $q$-vector and the latter is a $q \times q$-matrix.
Let \( \{y\}^m \) be an IID sample and let \( y. = \sum_{i=1}^m y_i \). Now the distribution of \( y. \) is called the (grouped, but unscaled) multinomial distribution, \( M(m, \pi) \) with index \( m \) and parameter \( \pi \). The distribution is given as

\[
P(y. = (m_1, \ldots, m_q)) = \frac{m!}{\prod_{j=1}^q m_i!} \prod_{j=1}^k \pi_{m_j}^j,
\]

and the first two moments are

\[
E(y.) = m\pi \quad \text{var}(y.) = m\{\text{diag}(\pi) - \pi\pi^t\}.
\]

**IWLS for multivariate response** The IWLS algorithm is again

\[
X^t WX\beta = XWz,
\]

where, for observations \( i = 1, \ldots, n \) the entries of the matrices are

\[
X = (X^t_1, \ldots, X^t_n)^t, \quad W = \text{diag}(W_i), \quad \eta = (\eta^t_1, \ldots, \eta^t_n)^t, \quad z = \eta + \left\{ \frac{\partial \eta^t_1(y_1 - \mu_1)}{\partial \mu_1}, \ldots, \frac{\partial \eta^t_n(y_n - \mu_n)}{\partial \mu_n} \right\}^t
\]

and

\[
W_i = \frac{m_i w_i}{\phi} D_i V_i^{-1} D_i^t = \frac{m_i w_i}{\phi} D_i (\partial \pi_i^t)^{-1} D_i^t = \frac{m_i w_i}{\phi} (\partial \pi_i^t)^{-1} D_i (\partial \pi_i^t)^{-1} = (D_i^{-1})^t
\]

where we have assumed that data are in grouped format, such that \( y_i \sim M(m_i, \pi_i) \), which generalize to the ungrouped situation if all \( m_i = 1 \). The dimension of \( W_i, D_i \) and \( V_i \) are \( q \times q \), \( w_i \) are prior weights, \( m \) is the multinomial index and \( \phi \) is the dispersion parameter. Note that \( W \) is block diagonal.

Suppose we want to fit the cumulative probit model

\[
\eta_i = X_i \beta = g_j(\pi_i) = \Phi^{-1}(\gamma_{ij}),
\]

and that \( J = 4 \), hence \( j = 1, 2, q = 3 \), then the model reads

\[
\begin{pmatrix}
\eta_{i1} \\
\eta_{i2} \\
\eta_{i3}
\end{pmatrix} =
\begin{pmatrix}
\Phi^{-1}(\gamma_{i1}) \\
\Phi^{-1}(\gamma_{i2}) \\
\Phi^{-1}(\gamma_{i3})
\end{pmatrix} =
\begin{pmatrix}
\Phi^{-1}(\pi_{i1}) \\
\Phi^{-1}(\pi_{i1} + \pi_{i2}) \\
\Phi^{-1}(\pi_{i1} + \pi_{i2} + \pi_{i3})
\end{pmatrix},
\begin{pmatrix}
\pi_{i1} \\
\pi_{i2} \\
\pi_{i3}
\end{pmatrix} =
\begin{pmatrix}
\Phi(\eta_{i1}) \\
\Phi(\eta_{i2}) - \Phi(\eta_{i1}) \\
\Phi(\eta_{i3}) - \Phi(\eta_{i2})
\end{pmatrix}.
\]
The elements of the IWLS algorithm therefore reads

\[
D_i = \frac{\partial \pi_t^i}{\partial \eta_i} = \left( \frac{\partial \pi_{i1}}{\partial \eta_i} \frac{\partial \pi_{i2}}{\partial \eta_i} \frac{\partial \pi_{i3}}{\partial \eta_i} \right) = \begin{pmatrix}
\varphi(\eta_{i1}) & -\varphi(\eta_{i1}) & 0 \\
0 & \varphi(\eta_{i2}) & -\varphi(\eta_{i2}) \\
0 & 0 & \varphi(\eta_{i3})
\end{pmatrix}
\]

\[
V_i = \begin{pmatrix}
\pi_{i1} (1 - \pi_{i1}) & -\pi_{i2} \pi_{i1} & -\pi_{i3} \pi_{i1} \\
-\pi_{i1} \pi_{i2} & \pi_{i2} (1 - \pi_{i2}) & -\pi_{i2} \pi_{i3} \\
-\pi_{i1} \pi_{i3} & -\pi_{i2} \pi_{i3} & \pi_{i3} (1 - \pi_{i3})
\end{pmatrix}
\]

\[
\frac{\partial \eta_i}{\partial \pi_t^i} = (D_i^{-1})^t = \begin{pmatrix}
\varphi^{-1}(\eta_{i1}) & 0 & 0 \\
\varphi^{-1}(\eta_{i2}) & 0 & 0 \\
\varphi^{-1}(\eta_{i3}) & 0 & 0
\end{pmatrix}
\]

\[
X_i = \begin{pmatrix}
1 & 0 & 0 & z_i^t \\
0 & 1 & 0 & z_i^t \\
0 & 0 & 1 & z_i^t
\end{pmatrix}
\]

\[
\beta = (\theta_1, \theta_2, \theta_3, \zeta^t)^t
\]

The structure on the matrices is a consequence of the model assumption of cumulative probabilities. Had we assumed a baseline category type model (which would be more appropriate for nominal rather than ordinal data), the matrices would have had a different structure. If we chose another link function, the elements of some of the matrices will be different. For instance if we chose the logit link, \( \varphi(\eta_{ij}) \) would be substituted with

\[
\frac{\exp(\eta_{ij})}{(1 + \exp(\eta_{ij}))^2} = \frac{\pi_{ij}}{(1 + \pi_{ij})^2}
\]

and \( \varphi^{-1}(\eta_{ij}) \) with

\[
\frac{(1 + \exp(\eta_{ij}))^2}{\exp(\eta_{ij})} = \frac{(1 + \pi_{ij})^2}{\pi_{ij}}.
\]

### E.7 Laplace Approximation in GLMMs

This section describes the extension of GLMs to GLMMs including normally distributed random effects. The approach described here is more traditional than that of the h-likelihood and is primarily based on the papers [Bates and DebRoy 2004](https://doi.org/10.1111/j.1467-9868.2004.00458.x) [Doran et al. 2007](https://doi.org/10.1093/occenv/7.1.27) and the vignette [Bates 2008](https://cran.r-project.org/web/packages/lme4/lme4.pdf) for the `[lme4](https://cran.r-project.org/web/packages/lme4/lme4.pdf)` R-package [Bates et al. 2008](https://cran.r-project.org/web/packages/lme4/lme4.pdf). Additional related literature includes [Liu and Pierce 1993, 1994](https://doi.org/10.1007/BF02290110) [Wolfinger 1993](https://doi.org/10.1007/BF02293783).

A GLMM is a GLM in which the predictor contains normally distributed random
effects
\[ \eta = X\beta + Zb , \]
where \( X \) is a \( n \times p \) matrix of predictors corresponding to the \( p \)-vector of fixed effect parameters \( \beta \), \( Z \) is a \( n \times q \) model matrix for the \( q \)-vector of random effect \( b \). It is assumed that \( b \) follows a multivariate normal distribution
\[ b \sim N(0, \Sigma(\theta)) , \]
where \( \Sigma \) is \( q \times q \), which may be large, but \( \theta \) is typically low dimensional.

The likelihood of the parameters \( \beta \) and \( \theta \) is mathematically equivalent to the marginal distribution of the data \( y \) given the parameters, hence the log-likelihood is
\[ l_b(\beta, \theta; y) = \log f(y; \beta, \theta) = \log \int p(y|\beta, b) f(b|\Sigma(\theta)) \, db . \tag{E.7.1} \]
Denote the integrand by \( h(y, b; \beta, b, \Sigma(\theta)) = p(y|\beta, b) f(b|\Sigma(\theta)) \). The Laplace approximation to the integral proceeds by approximating the log-integrand by its second order Taylor expansion. We perform the expansion around the maximizer of the integrand with respect to \( b \) and denote it by \( \tilde{b} \). This is also known as the posterior mode.

\[ \log h(y, b; \beta, b, \Sigma) \approx \log h^*(y, b; \beta, b, \Sigma) \]
\[ = \log h(y, \tilde{b}; \beta, b, \Sigma) - \frac{1}{2} (b - \tilde{b}) D (b - \tilde{b}) , \]
where
\[ D = -\frac{\partial^2}{\partial b^t \partial b} \log h(y, b; \beta, b, \Sigma) \Bigg|_{b=\tilde{b}} . \]

The approximation to the integral is
\[ l_b(\beta, \theta; y) \approx \log \int \exp(\log h^*) \, db \]
\[ = \log \int \exp(\log h(\tilde{b}) - \frac{1}{2} (b - \tilde{b}) D (b - \tilde{b})) \, db \]
\[ = \log h(\tilde{b}) + \log |D/(2\pi)|^{-1/2} \]
\[ = \log h(\tilde{b}) - \frac{1}{2} \log |D/(2\pi)| \]

The matrix of the negative second order derivatives, \( D \) corresponds to the observed Fisher information matrix of \( \tilde{b} \). We may approximate this by its
expectation—the expected Fisher information matrix evaluated at $\tilde{b}$. We can manipulate the matrix by the following

$$D = -\frac{\partial^2}{\partial \beta^t \partial \beta} l(\beta, b; y) - \frac{\partial^2}{\partial \beta^t \partial \beta} \log f(b; \Sigma)$$

$$= D_1 + \Sigma^{-1}$$

At fixed $\beta$ we may view the estimation of $\tilde{b}$ from $l(\beta, b; y)$ as an IRLS problem with $D_1 = Z^tWZ$. The estimation of $\tilde{b}$ from $h(b)$ may be seen as penalized IRLS (PIRLS) problem

$$(Z^tWZ + \Sigma^{-1})b = Z^t z ,$$

and a variance estimate of $\tilde{b}$ is $(Z^tWZ + \Sigma^{-1})^{-1}$. Here $W$ is a diagonal matrix of weights and $z$ is a vector of adjusted dependent variables

$$z = \eta + G(y - \mu) ,$$

where $G$ is a vector of derivatives $d\eta/d\mu$. The PIRLS algorithm works by incorporating the contribution from the normal distribution by adding $q$ “pseudo-observations” with unit weight and observed values of 0.

On the deviance scale, the Laplace approximation to the log likelihood is

$$-2l(\beta, \theta; y) = -2 \log \left\{ \int p(y|\beta, b)f(b|\Sigma(\theta)) \, db \right\}$$

$$\approx 2 \log \left\{ \int \exp \left\{ -\frac{1}{2} \left[ d(\beta, \tilde{b}, y) + \tilde{b}^t \Sigma^{-1} \tilde{b} + \log |\Sigma| + b^t D^{-1} b \right] \right\} \, db \right\}$$

$$= d(\beta, \tilde{b}, y) + \tilde{b}^t \Sigma^{-1} \tilde{b} + \log |\Sigma| + \log |D| ,$$

where $\tilde{b}$ are the maximizers of the integrand in (E.7.1), $d(\beta, \tilde{b}, y)$ is the deviance from the linear predictor only and $\log |D| = -\log |Z^tWZ + \Sigma^{-1}|$. A clever way to evaluate $\log |D|$ is to obtain the Cholesky decomposition as

$$\log |D| = -\log |Z^tWZ + \Sigma^{-1}| = -2 \log |L| ,$$

where $L$ is a lower triangular matrix from the decomposition $LL^t = Z^tWZ + \Sigma^{-1}$.

This approach of “adding pseudo observations” and determining $\tilde{b}$ using IWLS clearly has strong similarities with the h-loglikelihood approach.
E.8 Extended Likelihood and h-loglihood

The extended likelihood of unknown parameters $\theta$ and unobservable random quantities is

$$L(\theta, v; y, v) = f_\theta(y, v).$$

On the right-hand side is the joint distribution of $(y, v)$; $f_\theta(y, v) = f_\theta(v)f_\theta(y|v)$, which we may think of as the process of first sampling $v$ from $f_\theta(v)$ and then for fixed $v$ sampling $y$ from $f_\theta(y|v)$. On the left hand side is the joint likelihood of $(\theta, v)$;

$$L(\theta, v; y, v) = L(\theta; y)L(\theta, v; v|y),$$

ie. the marginal likelihood of $\theta$ based on the $y$ and the conditional likelihood of $\theta, v$ based on $v$ given $y$. In data generation $v$ are random instances and in likelihood inference, $v$ are unknowns to be estimated.

The classical likelihood principle states that the marginal likelihood $L(\theta; y)$ carries all the information in the data $y$ about the parameter $\theta$. The extended likelihood principle (Bjørnstad, 1996) states that the extended likelihood $L(\theta, v; y, v)$ carries all the information about $\theta, v$, hence the $L(\theta, v; v|y)$ does not carry any information about $\theta$ and it must carry all the information about $v$.

The marginal loglihood of $\theta$ can be obtained from the extended loglihood $l_e(\theta, v)$, by integrating out $v$. This integration is in general complicated and a Laplace approximation to the integral can be considered

$$l(\theta) = \int l_e(\theta, v) \, dv \approx p_v(l_e) = \left[ l_e - \frac{1}{2} \log |D(l_e, v)/(2\pi)| \right]_{v=\hat{v}_\theta},$$

where $D(l_e, v) = -\partial^2 l_e/\partial v^2$ and $\hat{v}_\theta$ solves $\partial l_e/\partial v = 0$ for fixed $\theta$. This approximation coincides with the adjusted profile likelihood (E.2.3) of $\theta$. In some cases when there is very little information in the data on the random effects such as binary matched pairs, some bias may persist for $\hat{\theta}$. Here a higher order Laplace approximation or (adaptive) Gauss-Hermite quadrature may be in place. In case $(y, v)$ both are normal, the first order Laplace approximation is exact.

Since $L(\theta, v; v|y)$ carries all information about $v$, we may use for instance the mean $E_\theta(v|y)$ or the mode of this likelihood as an estimator of $\hat{v}$. The latter resembles the ML estimator, and is in Bayesian terms called the maximum a posteriori (MAP) estimate. Since $L(\theta; y)$ does not carry information about $v$, estimating $v$ from $L(\theta, v; v|y)$ is equivalent to maximizing the entire extended likelihood for fixed $\theta$.

The simplest inference for $v$ may be obtained by plugging in the fixed parameter estimate $\hat{\theta}$ in the likelihood for the random effects; $L(\hat{\theta}, v; v|y) = f_{\hat{\theta}}(v|y)$ which mathematically is equivalent to a Bayesian posterior density. The vari-
ance of the estimate \( \hat{v} \) from this estimated likelihood does however not take into consideration the uncertainty from \( \hat{\theta} \).

In general we may use an adjusted profile likelihood as an approximation to the conditional likelihood of \( v \) given \( \theta \)

\[
p_\theta(l_e|v) = l_e(\hat{\theta}_v, v) - \frac{1}{2} \log |I(\hat{\theta}_v)|
\]

to obtain the correct variance of \( \hat{v} \).

**Canonical scale and h-likelihood** The estimate \( \hat{\theta}_\theta \) is said to be information neutral with respect to \( \theta \) if for two values of the fixed parameter \((\theta_1, \theta_2)\)

\[
\frac{L(\theta_1, \hat{\theta}_{\theta_1}; y, v)}{L(\theta_2, \hat{\theta}_{\theta_2}; y, v)} = \frac{L(\theta_1; y)}{L(\theta_2; y)},
\]

ie. the profile likelihood ratio equals the marginal likelihood ratio of \( \theta \). If a scale for \( v \) exists for which this property holds that scale is called the canonical scale. Equivalently the property can be stated

\[
\frac{L(\theta_1, \hat{\theta}_{\theta_1}; v|y)}{L(\theta_2, \hat{\theta}_{\theta_2}; v|y)} = 1
\]

which means that the ratio does not carry any information about \( \theta \) thereby meeting the classical likelihood principle.

The h-likelihood \( H(\theta, v) \) is defined as the extended likelihood \( L(\theta, v; y, v) \) for which \( v \) is canonical.

The canonical scale has the following important properties

- The MLE \( \hat{\theta} \) from the marginal likelihood \( L(\theta; y) \) coincides with the HMLE from the h-likelihood \( H(\theta, v) \).
- The information matrices from the two likelihoods are equivalent; \( I^{-1}_m = I^{-1}_h \).
- An estimate of \( \text{var}(\hat{v} - v) \) can be obtained from \( I^{22}_h \)

where

\[
I_h^{-1}(\hat{\theta}, \hat{v}) = \begin{pmatrix} I^{11}_h & I^{12}_h \\ I^{21}_h & I^{22}_h \end{pmatrix}
\]
The scale \( v \) may be canonical for only a subset of the fixed parameters. Suppose the fixed parameters consists of the subsets \((\theta, \phi)\), then we may have that

\[
\frac{L(\theta_1, \phi, \hat{v}_{\theta_1, \phi}; y, v)}{L(\theta_2, \phi, \hat{v}_{\theta_2, \phi}; y, v)} = \frac{L(\theta_1, \phi; y)}{L(\theta_2, \phi; y)}
\]

but

\[
\frac{L(\theta, \phi_1, \hat{v}_{\theta, \phi_1}; y, v)}{L(\theta, \phi_2, \hat{v}_{\theta, \phi_2}; y, v)} \neq \frac{L(\theta, \phi_1; y)}{L(\theta, \phi_2; y)},
\]

hence \( v \) is canonical for \( \theta \), but not for \( \phi \) and joint inference from the h-likelihood is possible only for \((\theta, v)\), while the inference for \( \phi \) requires a marginal likelihood or an approximation thereof. An approximate marginal likelihood for \( \phi \) is given by the adjusted profile likelihood

\[
p_v(h) = \left[ h - \frac{1}{2} \log |D(h, v)/(2\pi)| \right]_{v=\hat{v}}, \tag{E.8.1}
\]

where \( D(h, v) \) is a function of \( \phi \), but not of \( \theta \) reflecting that \( \theta \) and \( v \) are information orthogonal.

Since the h-likelihood is defined with \( v \) on a particular scale, we can obtain invariant inference for monotone nonlinear transformations \( u = u(v) \).

\[
H(\theta, v) = H(\theta, v(u)) = L(\theta, v(u); y, v) = f_\theta(y|v(u))f_\theta(v(u))
\]

which differs from the raw extended likelihood

\[
L(\theta, u; y, u) = f_\theta(y|v(u))f_\theta(v(u))|J(u)| = H(\theta, u)|J(u)|
\]

Here \( v \) is, but \( u \) is not canonical for the extended likelihood of \((\theta, u)\). Using the h-likelihood ML estimates are invariant with respect to transformation of both fixed and random parameters. If the monotone transformation \( u = v(u) \) is linear, the Jacobian \( J(u) = 1 \) and \( H(\theta, v) = H(\theta, u) \). Up to linear transformations of \( v \) the canonical scale is unique.

In general \( v \) is canonical if \( I(\hat{v}_\theta) \) in \( p_v(l_v) \) is free of \( \theta \), see (E.8.1).
The normal linear mixed model can be stated
\[ y = X\beta + Zv + e, \]  
(E.9.1)
where \( X \) and \( Z \) are a \( N \times p \) and \( N \times q \) model matrices for the \( p \)-vector of fixed effects \( \beta \) and the \( q \)-vector of random effects \( v \) and \( e \sim MVN(0, \Sigma) \), \( v \sim MVN(0, D) \) and \( e \) and \( v \) are assumed independent. Now assume that \( D = \sigma_v^2 I_q \) and \( \Sigma = \sigma_e^2 I_N \), where \( I_k \) is an \( k \times k \) identity matrix. Then the fixed variance component parameter is \( \tau = (\sigma_v^2, \sigma_e^2) \).

The implied marginal distribution of \( y \) is
\[ y \sim MVN(X\beta, V), \]
where \( V = ZDZ^t + \Sigma \).

The marginal loglihood of the fixed parameters \((\beta, \tau)\) is
\[ l(\beta, \tau) = -\frac{1}{2} \log |2\pi V| - \frac{1}{2} (y - X\beta)^t V^{-1} (y - X\beta), \]  
(E.9.2)
Taking the derivative of the loglihood with respect to \( \beta \)
\[ \partial l / \partial \beta = X^t V^{-1} (y - X\beta) \]
and equating to zero yields the estimate
\[ \hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y. \]

Using this estimate, the profile likelihood of \( \tau \) is
\[ l_p(\tau) = -\frac{1}{2} \log |2\pi V| - \frac{1}{2} (y - X\hat{\beta})^t V^{-1} (y - X\hat{\beta}), \]  
(E.9.3)
The Fisher information of \( \beta \) is obtained as
\[ I(\hat{\beta}) = -\partial^2 l / \partial \beta^2 = -\partial / \partial \beta (X^t V^{-1} (y - X\beta)) = X^t V^{-1} X. \]

The uncertainty about \( \hat{\beta} \) from this formula does not take into account the uncertainty in the estimation of \( \hat{\tau} \). Although the influence is small in most cases (\( E(\partial^2 l / \partial \beta \partial \tau) = 0 \)), we can adjust the profile loglihood and obtain the REML adjustment
\[ p_\beta(l|\tau) = l(\hat{\beta}_\tau, \tau) - \frac{1}{2} \log |X^t V^{-1} X/(2\pi)|, \]  
(E.9.4)
which in normal linear mixed models coincide with an exact conditional and marginal loglihood. Optimization of this loglihood is however rather slow. An efficient approach will be considered later.
Traditional estimation of random effects  The joint log-density function of \( y \) and \( v \) may be written as
\[
\log f(y, v) = c - \frac{1}{2} (y - X\beta - Zv)^t \Sigma^{-1} (y - X\beta - Zv) - \frac{1}{2} v^t D^{-1} v.
\]
The corresponding estimates of \( \beta \) and \( \tau \) are MLE. Joint estimation of \( \beta \) and \( v \) work when \( v \) is on the canonical scale for \( \beta \), which is the case here. The derivative of the density with respect to \( \beta \) and \( \tau \) are
\[
\frac{\partial \log f}{\partial \beta} = X^t \Sigma^{-1} (y - X\beta - Zv)
\]
\[
\frac{\partial \log f}{\partial v} = Z^t \Sigma^{-1} (y - X\beta - Zv) - v^t D^{-1},
\]
hence we arrive at the formulas
\[
(X^t \Sigma^{-1} X)\hat{\beta} = X^t \Sigma^{-1} (y - Zv)
\]
\[
(Z^t \Sigma^{-1} Z + D^{-1})\hat{v} = Z^t \Sigma^{-1} (y - X\beta),
\]
which can be written as Henderson’s mixed model equation
\[
\begin{pmatrix}
X^t \Sigma^{-1} X & X^t \Sigma^{-1} Z \\
Z^t \Sigma^{-1} X & Z^t \Sigma^{-1} Z + D^{-1}
\end{pmatrix}
\begin{pmatrix}
\hat{\beta} \\
\hat{v}
\end{pmatrix}
= \begin{pmatrix}
X^t \Sigma^{-1} y \\
Z^t \Sigma^{-1} y
\end{pmatrix}.
\]
Note that if we set \( D^{-1} = 0 \), we obtain the fixed estimates of \( v \), hence \( D^{-1} \) is responsible for the shrinkage estimates of \( v \), when they are treated as random.

Estimates of \( (\beta, v) \) can be obtained by solving (E.9.7), given an estimate of \( \tau \). It is however more efficient to solve (E.9.6). An update of \( \hat{\tau} \) can be obtained from the profile likelihood (E.9.3) or the the adjusted profile likelihood (E.9.4) giving the ML and the REML estimates respectively.

**H-likelihood for normal linear mixed models**  The extended loglihood of all parameters is
\[
l_v(\beta, \tau, v) = \log f(y, v) = \log f(y|v) + \log f(v)
\]
\[
= - \frac{1}{2} \log |2\pi \Sigma| - \frac{1}{2} (y - X\beta - Zv)^t \Sigma^{-1} (y - X\beta - Zv) - \frac{1}{2} v^t D^{-1} v.
\]
For this to be a h-loglihood, we need to verify, that \( v \) are on the canonical scale. The first derivative of (E.9.8) is given by (E.9.5), hence Fisher information with respect to \( v \) is
\[
I(\hat{v}) = Z^t \Sigma^{-1} Z + D^{-1}.
\]
This means that $v$ can be canonical for $\beta$ but not for $\tau$, since $I(\hat{v})$ depends in $\tau$ but not on $\beta$. It can be shown (Lee and Nelder, 2006a) that it is the canonical scale, hence (E.9.8) is a h-loglihood.

The matrix on the left hand side of (E.9.7) can in light of the h-loglihood be seen to be $I_h(\hat{\beta}, \hat{v})$ at the ML estimates of $(\beta, v)$. A variance estimate, that properly takes into account the uncertainty that comes from estimating $\beta$ is given as $I_{h22}^{-1}$. This estimate is larger than that obtained if $\beta$ is assumed known at the MLE; $I_{h22}^{-1} = Z^t \Sigma^{-1} Z + D^{-1}$.

The marginal loglihood of $(\beta, \tau)$ (E.9.2) and the adjusted profile loglihood (E.9.4) of the variance parameters both involve the marginal variance-covariance matrix $V^{-1}$ or $|V|$ and since $V$ is potentially very large and ill conditioned, we would prefer to work with something else. The marginal loglihood can be obtained in terms of the adjusted profile likelihood

$$ l(\beta, \tau) = h(\beta, \tau, \hat{v}_{\beta, \tau}) - \frac{1}{2} \log |I(\hat{v}_{\beta, \tau})/(2\pi)| $$

$$ = p_v(h|\beta, \tau) , $$

where $I(\hat{v}_{\beta, \tau})$ is given above. The marginal loglihood can also be obtained as

$$ l(\beta, \tau) = \int_{-\infty}^{\infty} f_{\beta, \tau, v}(y, v) \, dv $$

$$ = -\frac{1}{2} \log |2\pi \Sigma| - \frac{1}{2} \left( y - X \beta - Z \hat{v}_{\beta, \tau} \right)^t \Sigma^{-1} \left( y - X \beta - Z \hat{v}_{\beta, \tau} \right) $$

$$ - \frac{1}{2} \log |2\pi D| - \frac{1}{2} \hat{v}_{\beta, \tau}^t D^{-1} \hat{v}_{\beta, \tau} - \frac{1}{2} \log |I(\hat{v}_{\beta, \tau})/(2\pi)| . $$

This is the Laplace approximation to the integral, which in the normal case is exact.

We can further marginalize the loglihood above, or profile over $\beta$ to obtain a loglihood of $\tau$ only

$$ l(\tau) = l(\hat{\beta}_r, \tau) = h(\hat{\beta}_r, \tau, \hat{v}_r) - \frac{1}{2} \log |I(\hat{v}_r)/(2\pi)| . $$

We may also adjust the profile loglihood for the estimation of $\beta$, ie. include the REML adjustment

$$ p_\beta = (\hat{\beta}_r, \tau) - \frac{1}{2} \log |X^t V^{-1} X/(2\pi)| $$

$$ = h(\hat{\beta}_r, \tau, \hat{v}_r) - \frac{1}{2} \log |I(\hat{v}_r)/(2\pi)| - \frac{1}{2} \log |X^t V^{-1} X/(2\pi)| $$

$$ = p_v(h|\tau) $$
Estimation  We can write the Henderson equations (E.9.7) as a weighted least squares problem
\[ (T^t \Sigma_a^{-1} T) \hat{\delta} = T^t \Sigma_a^{-1} y_a , \] (E.9.9)
where we have taken
\[ y_a = \begin{pmatrix} y \\ \psi_M \end{pmatrix} , \quad T = \begin{pmatrix} X & Z' \\ 0 & I \end{pmatrix} , \quad \delta = \begin{pmatrix} \beta \\ v \end{pmatrix} , \quad \Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix} , \]

Where \( \psi_M = 0 \) are pseudo or quasi data assumed to be normal with expectation \( E(\psi_M) = v \), variance \( D \) and independent of \( y \). Also we may define the augmented linear model
\[ y_a = T \delta + e_a , \quad e_a = \begin{pmatrix} e \\ e_M \end{pmatrix} . \]
Note that \( \Sigma \) and \( D \) are diagonal.

The estimation of \((\beta, \tau, v)\) proceeds as follows. Given the current estimate of \( \tau \), update \( \hat{\delta} \) by solving (E.9.9). With an update of \( \hat{\delta} \), a REML update of \( \tau \) is given by two gamma GLMs with responses \( d^*_i = d/(1 - q) \) and \( d^*_M = d_M/(1 - q_M) \) respectively, log links, only intercepts in the linear predictor and prior weights \((1 - q)/2 \) and \((1 - q_M)/2 \) respectively. Iterate until convergence.

It holds that
\[
E(d^*_i) = \sigma^2 , \quad \text{var}(d^*_i) = 2\sigma^2/(1 - q_i) , \quad E(d^*_M) = \sigma^2_v , \quad \text{var}(d^*_M) = 2\sigma^2_v/(1 - q_{Mi}) .
\]

The elements \( q_i \) and \( q_{Mi} \) are the leverage elements—the diagonal elements of
\[
T (T^t \Sigma_a^{-1} T)^{-1} T^t \Sigma_a^{-1} ,
\]
and the deviance elements \( d_i \) and \( d^*_i \) are
\[
d_i = (y_i - x_i^t \hat{\beta} - z_i^t \hat{\psi})^2 , \quad d^*_M = (\psi_M - \hat{v}_i)^2 = \hat{v}_i^2 .
\]
The ML estimates are obtained by taking the leverage elements to be zero.

Multi-component models  Multi-component models contain several random effects
\[ y = X \beta + Z_1 v_1 + \cdots + Z_m v_m + e , \]
where \( Z_i \) are \( N \times q_i \) model matrices and \( v_i \) are independent \( MVN_{q_i}(0, D_i) \). This model may be written as (E.9.1) by adjoining the elements as
\[
Z = [Z_1 \cdots Z_m] , \quad v = (v_1 \cdots v_m) .
\]
now with $(m + 1)$ variance components parameters.

If the random effects $v$ are not IID, but

$$D = \sigma^2_v R$$

with a known matrix $R$, the simple IID form can derived as

$$y = X\beta + ZR^{1/2}R^{1/2}v + e$$
$$= X\beta + Z^*v^* + e ,$$

where $Z^* = ZR^{1/2}$ and $v^* = R^{1/2}v$ and $R^{1/2}$ is the square root matrix of $R$. 
Appendix F

Thurstonian Models for Replicated Discrimination Tests
F.1 Introduction

When [Thurstone (1927a)] formulated his law of comparative judgment, he wrote the following:

It [the law] applies fundamentally to the judgments of a single observer [Thurstone's emphasis] who compares a series of stimuli by the method of paired comparison when no 'equal' judgments are allowed.

He emphasized that the law applies to a single observer, which of course makes good sense, since the law is a model for “the discriminative process”—which by definition is a single-subject process.

Often however we will be interested in evaluating the difference between several stimuli, which have been assessed by several assessors. We propose next a statistical model, that applies Thurstone’s law to each subject while handling all observations.

R functions to fit the models are partly included in the sensR package (discrimR will fit what we will call the two-parameter model). The rest are included in the appendix.

F.2 The Conditional Model

Suppose several individuals participated in a test and that it is reasonable to assume that they differ with respect to ability to discriminate. Then we may assume that data follows the following sampling scheme:

1. Sample $\delta_i^*$ from $N(\delta, \sigma_{\delta}^2)$ for each of $m$ individuals; $i = 1, \ldots, m$.

2. Define

$$
\delta_i = \begin{cases} 
0, & \delta_i^* \leq 0 \\
\delta_i^*, & \delta_i^* > 0 
\end{cases}
$$
3. Sample the response \( y_i \) conditional on \( \delta_i \) from a binomial distribution; 
\[ y_i | \delta_i \sim \text{Bin}(\pi_i; m_i), \] 
where \( \pi_i = f_{ps}(\delta_i) \) and \( f_{ps}() \) is the psychometric function for the given test \cite{Brockhoff2008}.

This sampling scheme can also be stated as

1. Sample \( Z_i \sim \text{Bernoulli}(p) \) for \( i = 1, \ldots, m \).
2. Define \( (\delta_i | Z_i = 0) := 0 \) and sample \( \delta_i | Z_i = 1 \sim f(\delta_i | Z_i = 1) \), where the latter density is given in \cite{F.2.1} below.
3. Execute step 3. given above.

The conditional distribution \( F(\delta_i | Z_i = 1) \) is a truncated normal distribution with density
\[
f(\delta_i | Z_i = 1) = c \frac{1}{\sigma_\delta} \phi \left( \frac{\delta_i - \delta}{\sigma_\delta} \right), \quad \delta_i > 0 \tag{F.2.1}
\]
where \( c = [1 - \Phi(-\delta/\sigma_\delta)]^{-1} \) is the normalizing constant.

With \( p = c^{-1} = 1 - \Phi(-\delta/\sigma_\delta) \), the density of \( \delta_i \) can be stated as
\[
f(\delta_i) = \begin{cases} 
0, & \delta_i < 0 \\
\Phi(-\delta/\sigma_\delta), & \delta_i = 0 \\
\frac{1}{\sigma_\delta} \phi \left( \frac{\delta_i - \delta}{\sigma_\delta} \right), & \delta_i > 0
\end{cases} \tag{F.2.2}
\]
where \( \Phi() \) denotes the normal CDF and \( \phi() \) denotes the normal PDF.

We aim to estimate the conditional model
\[
y_i | \delta_i \sim \text{Bin}(\pi_i; m_i), \quad g(\pi_i) = \delta_i = \delta + b_i, \tag{F.2.3}
\]
where \( g() \) is the inverse psychometric function and the distribution of \( \delta_i \) is described by \( f(\delta_i) \).

By expressing the Bernoulli probability, \( p \) in terms of the parameters of the conditional distribution of \( \delta_i | Z_i = 1 \), we assume that the probability mass at \( Z_i = 0 \) for \( f(\delta_i) \) in some sense corresponds to that not covered by the density for \( Z_i = 1 \). As such, this model is a special case of a more general model, where \( p \) need not equal \( 1 - \Phi(-\delta/\sigma_\delta) \). We will consider this model shortly.

Denote the probability of sampling an individual with a \( \delta_i > 0 \) by \( p^o \), such that now \( Z_i \sim \text{Bernoulli}(p^o) \), where we have assumed that \( p^o := p \) until now. The
distribution of $\delta_i$ is now given by
\[
f(\delta_i) = \begin{cases} 
0, & \delta_i < 0 \\
1 - p^o, & \delta_i = 0 \\
c_2 \frac{1}{\sigma_\delta} \phi \left( \frac{\delta_i - \delta}{\sigma_\delta} \right), & \delta_i > 0 
\end{cases}, \quad (F.2.4)
\]
where the normalizing constant is $c_2 = p^o/p$. There are three possible situations regarding the value of $p^o$. First, as mentioned above, it can with reasonable confidence equal $p$ in which case the model reduces to the two parameter model discussed above. If on the other hand $p^o > p$, we have a three parameter model; $\theta = (\delta, \sigma_\delta, p^o)$, which is able to account for an extra probability mass at $Z_i = 0$. In this case we may write $p^o = p + p' - 1$ (ie. $1-p^o = (1-p)+(1-p')$), and $1-p'$ is the extra probability of $Z_i = 0$ compared to that of the two parameter model. If, in the third situation, we are in the rather unlikely situation that $p^o < p$, less probability of $Z_i = 0$ remains then that predicted by the two parameter model and it is questionable whether this type of model is appropriate at all. (This peculiar situation could however occur if for instance a panel of judges are selected with generally low sensitivities, but some of those with zero sensitivity are excluded based on prior experience.)

To estimate the model $\{F.2.3\}$, we need the marginal distribution of $y_i$, since the likelihood is mathematically equivalent to this quantity. The marginal distribution of $y_i$—the contribution of the $i$th individual to the likelihood is
\[
f(y_i) = \int f(y_i|\delta_i, Z_i) \, dF(\delta_i)
= f_{\delta_i=0}(y_i|\delta_i) f(\delta_i) + \int_0^\infty f_{\delta_i>0}(y_i|\delta_i) f(\delta_i) \, d\delta_i \quad (F.2.5)
\]
where $p_0$ is the guessing probability, and
\[
f_{p_0}(y_i) = \binom{n_i}{y_i} p_0^{y_i} (1-p_0)^{n_i-y_i} \quad \delta_i = 0 \quad \pi_i = p_0 \\
f_{\pi}(y_i|\delta_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i-y_i} \quad \delta_i > 0 \quad \pi_i > p_0
\]
The distribution thus has the components
\[
f(y_i) = (1-p^o) f_1(y_i) + p^o f_2(y_i),
\]
where $f_1(y_i) = f_{p_0}(y_i)$ and $f_2(y_i) = \frac{1}{p} \int_0^\infty f_{\pi}(y_i|\delta_i) \phi((\delta_i - \delta)/\sigma_\delta)/\sigma_\delta \, d\delta_i$.

The likelihood of the parameters $\theta = (\delta, \sigma_\delta, p^o)$ given all data is therefore
\[
L(\theta; y) = \prod_i f(y_i).
\]
F.3 Conditional Expectations

The estimates \( \hat{\delta}_i \) can be found as conditional expectations. They are called best predictors by McCulloch and Searle (2001) and in a Bayesian language we would call them posterior means. Another estimate of \( \delta_i \) would be posterior modes also called empirical Bayes estimates, which equal the posterior means only in well behaved normal linear models.

\[
\hat{\delta}_i = E(\delta_i | y_i) = \int_{-\infty}^{\infty} \delta_i f(\delta_i | y_i) \, d\delta_i \\
= \int_{-\infty}^{\infty} \frac{\delta_i f(y_i | \delta_i) f(\delta_i)}{f(y_i)} \, d\delta_i \\
= \frac{\int_{-\infty}^{\infty} \delta_i f(y_i | \delta_i) f(\delta_i) \, d\delta_i}{f(y_i)}
\]

(F.3.1)

where \( f(y_i) \) is the likelihood contribution from the \( i \)th individual. Denote the numerator in the last expression in (F.3.1) by \( h(y_i) \). Then this quantity is

\[
h(y_i) = \int_{-\infty}^{\infty} \delta_i f(y_i | \delta_i) f(\delta_i) \, d\delta_i \\
= h_{\delta_i=0}(y_i) + h_{\delta_i>0}(y_i) \\
= 0 + c_2 \int_{0}^{\infty} \delta_i f(y_i | \delta_i) \phi((\delta_i - \delta)/\sigma_\delta) / \sigma_\delta \, d\delta_i
\]

The \( \hat{Z}_i \)s can also be estimated as conditional expectations similarly to the \( \hat{\delta}_i \)s:

\[
\tilde{Z}_i = E(Z_i | y_i) = \int Z_i \, dF(Z_i | y_i) = \sum_{Z_i \in (0,1)} Z_i f(Z_i | y_i) \\
= \sum_{Z_i \in (0,1)} Z_i f(y_i | Z_i) f(Z_i) / f(y_i) \\
= 0 + f_{Z_i=1}(y_i | Z_i) p^o / f(y_i) \\
= \frac{p^o f_2(y_i)}{f(y_i)} = \frac{p^o f_2(y_i)}{(1 - p^o) f_1(y_i) + p^o f_2(y_i)}
\]

(F.3.2)

where all components have already been computed in the evaluation of the likelihood. We expect \( \tilde{Z}_i \in [0,1] \) and interpret them as the propensity (or probability) of the \( i \)th individual being a discriminator given the information in the data.
F.4 Mean and Variance of the random effects distribution

The parameters $\delta$ and $\sigma_\delta$ are parameters of the distribution of $\delta_i$, $f(\delta_i)$ in (F.2.2), but they do not have the meaning of mean and variance of that distribution. The same parameters are the parameters of the conditional distribution $f(\delta_i|Z_i = 1)$ in (F.2.1), and also in this case, they do not have the meaning of mean and variance. In the following we will derive those two new sets of first and second moments.

Consider first the mean of a truncated normal distribution, truncated below some point $t$ (the following is closely based on [Barr and Sherrill 1999]). The density of a truncated standard normal random variable, $Z$ truncated below at $t$ is

$$f(z) = c(t)e^{-z^2/2}, \quad z \geq t,$$

and zero for other $z$. The normalizing constant depends on the truncation point and is

$$c(t) = \left\{ \sqrt{\frac{2\pi}{3}} \right\}^{-1}.$$

The mean of $Z$ is

$$E(Z) = c(t) \int_t^\infty ze^{-z^2/2} \, dz = -c(t)e^{-z^2/2}\bigg|_t^\infty = c(t)e^{-t^2/2}.$$

To compute the variance $V(Z) = E(Z^2) - E^2(Z)$, we need (for positive $t$)

$$E(Z^2) = c(t) \int_t^\infty z^2e^{-z^2/2} \, dz = \frac{c(t)}{2} \int_t^\infty ze^{-z^2/2}2z \, dz$$

$$= \frac{c(t)}{2} \sqrt{\frac{2\pi}{3}} \int_{t^2}^\infty \frac{1}{2^{3/2}\Gamma(3/2)}u^{1/2}e^{-u/2} \, du,$$

where the last integral is that of a $\chi^2$ density with 3 degrees of freedom. We may therefore write

$$E(Z^2) = \begin{cases} c(t)\sqrt{\pi/2}\frac{1-C_3(t^2)}{1-C_3(t^2)} & t \geq 0 \\ c(t)\sqrt{\pi/2}\frac{C_3(t^2)}{1+C_3(t^2)} & t < 0 \end{cases},$$

where $C_3(t^2)$ is the area under the $\chi^2$-density with three degrees of freedom up to the $t$th quantile. The variance of $Z$ is therefore given by

$$V(Z) = \begin{cases} c(t)\{\sqrt{\pi/2}(1-C_3(t^2)) - c(t)e^{-t^2}\} & t \geq 0 \\ c(t)\{\sqrt{\pi/2}(1+C_3(t^2)) - c(t)e^{-t^2}\} & t < 0 \end{cases}.$$
Suppose now that $X$ is a nonstandard truncated normal random variable following $\mathcal{N}(\mu, \sigma)$, truncated at $t'$. Then $(X - \mu)/\sigma = Z$ is the standardized equivalent, truncated at $t$, where $t$ solves $\sigma t + \mu = t'$, i.e., $t = (t' - \mu)/\sigma$. The mean and variance of $X$ is therefore given by $E(X) = \sigma E(Z) + \mu$ and $V(X) = \sigma^2 V(Z)$.

Application of these formulæ shows that the mean and variance of $f(\delta_i|Z_i = 1)$ (F.2.1) are

$$
\delta_i'' = E(\delta_i|Z_i = 1) = \sigma \delta c(t)e^{-t^2/2} + \delta
$$

$$
\sigma_{\delta''}^2 = V(\delta_i|Z_i = 1) = \sigma^2 c(t)\{\sqrt{\pi/2}[1 + C_3(t^2)] - c(t)e^{-t^2}\},
$$

where $t = -\delta/\sigma\delta$, hence $c(t) = (\sqrt{2\pi p})^{-1}$.

Consider now instead $f(\delta_i)$ (F.2.2), for which we will derive the first two moments. A standardized version of the density truncated at $t$ will have the following form:

$$
f(z) = \begin{cases} 
\Phi(t) & z = t \\
\phi(z) & z > t 
\end{cases}.
$$

The mean is

$$
E(Z) = \int z \, dF(z) = t\Phi(t) + \int_t^\infty zf_{z>t}(z) \, dz
$$

$$
= t\Phi(t) + c_1 \int_t^\infty ze^{-z^2/2} \, dz
$$

$$
= t\Phi(t) + c_1 e^{-t^2/2},
$$

where $c_1 = (2\pi)^{-1/2}$. The non-standard density has the form

$$
f(x) = \begin{cases} 
\Phi\left(\frac{x'-\mu}{\sigma}\right) & x = t' \\
\frac{1}{\sigma} \phi\left(\frac{x'-\mu}{\sigma}\right) & x > t'
\end{cases},
$$

with mean

$$
E(X) = \sigma[t\Phi(t) + c_1 e^{-t^2/2}] + \mu,
$$

where $t = (t' - \mu)/\sigma$.

The variance is given by

$$
V(X) = \int [X - E(X)]^2 \, dF(X)
$$

$$
= [t' - E(X)]^2 \Phi(t) + \int_{t'}^\infty [X - E(X)]^2 \frac{1}{\sigma} \phi\left(\frac{X - \mu}{\sigma}\right) \, dX
$$
The mean and variance of the distribution of $f(\delta_i)$ are therefore
\[
\delta' = E(\delta_i) = \sigma_\delta t \Phi(t) + \frac{1}{\sqrt{2\pi}} e^{-t^2/2} + \delta
\]
\[
\sigma_{\delta'}^2 = V(\delta_i) = \delta'^2 \Phi(t) + \int_0^\infty [\delta_i - \delta']^2 \frac{1}{\sigma_\delta} \phi \left( \frac{\delta_i - \delta}{\sigma_\delta} \right) \, d\delta_i,
\]
where $t = -\delta/\sigma_\delta$.

The standard error of $\hat{\delta}'$, $\text{se}(\hat{\delta}')$ can be estimated by the delta method
\[
\text{se}(\hat{\delta}') = \text{se}(\hat{\delta}) f'(\delta, \sigma_\delta) |_{\delta = \hat{\delta}, \sigma_\delta = \hat{\sigma}_\delta},
\]
where $f(\delta) = E(X)$ and
\[
f'(\delta, \sigma_\delta) = (1 - \frac{1}{\sigma_\delta}) \frac{\delta}{\sigma_\delta} \phi(-\delta/\sigma_\delta) + \Phi(\delta/\sigma_\delta)
\]  
(F.4.1)
is the partial derivative of $f(\delta, \sigma_\delta)$ with respect to $\delta$.

For the distribution in $[F.2.4]$, the mean and variance is
\[
E'(\delta_i) = \int \delta_i \, d\delta_i
\]
\[
= 0 + c_2 \int_0^\infty \delta_i \phi((\delta_i - \delta)/\sigma_\delta)/\sigma_\delta \, d\delta_i
\]
\[
= c_2 E(\delta_i)
\]
\[
V'(\delta_i) = \int (\delta_i - \delta)^2 \, dF(\delta_i)
\]
\[
= \delta^2(1 - p_o) + \frac{c_2}{\sigma_\delta} \int_0^\infty (\delta_i - \delta)^2 \phi((\delta_i - \delta)/\sigma_\delta)/\sigma_\delta \, d\delta_i
\]

The standard error of $\hat{\delta}' = E'(\delta_i)$ is now as before, only
\[
f'(\delta, \sigma_\delta) = c_2 [-\Phi(-\delta/\sigma_\delta) + \left\{ 1 - \frac{1}{\sigma_\delta} \right\} \frac{\delta}{\sigma_\delta} \phi(-\delta/\sigma_\delta) + 1],
\]  
which does reduce to the expression in $[F.4.1]$, when $c_2 = 1$, which happens when the three parameter model reduce to the two parameter model.

We would prefer to use profile likelihoods for assessing the information in each parameter, but this is not readily available information for these models, hence we have to do with simple approximate standard errors.
The behaviour of $E(\delta_i)$ and $V(\delta_i)$ (blue) as well as $E(\delta_i|Z_i = 1)$ and $V(\delta_i|Z_i = 1)$ (red) as a function of the truncation point $t$ is illustrated in figure F.1. It is seen that the mean approaches that of the original distribution ($\delta^* \sim N(-1, 1)$, black solid), when $t$ approaches minus infinity and that it approaches $t$ as $t$ approaches plus infinity although $E(\delta_i)$ does this at a much faster rate than $E(\delta_i|Z_i = 1)$. The variance also approaches that of the original distribution ($\delta^* \sim N(0, 2)$, black solid) as $t$ gets small and it approaches zero as $t$ gets large. Again $V(\delta_i)$ approaches at a much faster rate than $V(\delta_i|Z_i = 1)$ in correspondence with our expectation.

Figure F.1: Mean (left) and variance (right) for the distributions of $\delta_i$ (blue) and $\delta_i|Z_i = 1$ (red) versus the truncation point $t$. See the text for details.
Figure F.2: Left: Marginal (dashed, for $\sigma_\delta = 1$) and condition (solid) prediction of $\hat{\pi}$ for the four psychometric tests (top to bottom); 2-AFC (blue), 3-AFC (black), duo-trio (red) and triangle (green). Right: Stationary point $\delta_0$ versus heterogeneity $\sigma_\delta$ in $\delta_i$.

**F.5 Attenuation Effects**

The marginal mean, ie. unconditional expectation, assuming that the conditional model holds is given by

$$
\pi_m(\delta, \sigma_\delta) = E_{\delta_i}(\pi_i) \\
= E_{\delta_i}(E(y_i|\delta_i)) \\
= E_{\delta_i}(g^{-1}(\delta_i)) = E_{\delta_i}(f_{ps}(\delta_i)) \\
= \int f_{ps}(\delta_i) \, dF(\delta_i) \\
= f_{ps}(\delta_i|\delta_i = 0) f(\delta_i|\delta_i = 0) + \int_0^\infty f_{ps}(\delta_i|\delta_i > 0) f(\delta_i|\delta_i > 0) \, d\delta_i \\
= p_0(1 - p^0) + c_2 \int_0^\infty f_{ps}(\delta_i) \phi((\delta_i - \delta)/\sigma_\delta)/\sigma_\delta \, d\delta_i.
$$

(F.5.1)

This relation between $\delta$ and the marginal prediction of $\pi$—the probability of a correct answer is shown in the left panel of figure F.2 (dashed lines) for a moderate heterogeneity of $\sigma_\delta = 1$ along with the ordinary psychometric functions, $f_{ps}(\delta_i)$ (solid lines) for the four psychometric tests (from top to bottom); 2-AFC (blue), 3-AFC (black), duo-trio (red) and triangle (green). The function $\pi_{\text{Marg}}$
included in the appendix estimates $\pi_m$ and takes as arguments $\delta$, $\sigma_\delta$ and the type of test.

It is clear that “stationary points” (denote these $(\delta_0, \pi_0)$) exist, where the marginal and conditional curves cross. For a $\delta$ below $\delta_0$, the marginal model will predict a larger $\hat{\pi}_m$ than the conditional model $\hat{\pi}_c$ and above the points, the marginal predictions will be smaller than that of the conditional model. An equivalent interpretation is that for a given success probability $\pi$ and heterogeneity $\sigma_\delta$, the conditional model will predict the $d'_c$ proposed by the psychometric function (ie. Thurstones model, when applied to any single individual), where as the marginal model will predict $d'_m$, which will be closer to the stationary point than will $d'_c$.

The stationary points $(\delta_0, \pi_0)$ are given as the $\delta$, which solves

$$\pi_m(\delta, \sigma_\delta) - f_{\text{ps}}(\delta) = 0.$$ 

It is worth to note, that the stationary points depend on the heterogeneity $\sigma_\delta$. This is shown in the right panel of figure F.2, where we have plotted the stationary point expressed as $\delta_0$ versus $\sigma_\delta$ for the 2-AFC test. The function $\text{delta0}$ estimates the stationary point $(\delta_0, \pi_0)$ and takes as arguments $\sigma_\delta$ and the name of the psychometric test.

## F.6 Examples

### F.6.1 Example 1

[Duineveld and Meyners (2008)] report on a study in which the triangle test was used. A concentration of flavour close to the detection limit was tested against tab water. 30 individuals performed between 3 and 10 assessments each in a total of 185 tests in which the correct sample was chosen in 100 cases. [Duineveld and Meyners (2008)] maintain focus on the probability of a correct answer—we will analyse the data in terms of the Thurstonian model for the triangle test and assume that this model applies to each individual. We will also report our results in terms of $d'$s, that is, we will report the $d'$ for an average individual, the variance of this, the estimated $d'$s for each individual as well as the estimated probability of each individual being a discriminator, ie. having a $d' > 0$. In estimating these quantities we will assume that the model (F.2.3) is appropriate, especially we assume that the $\delta_i$s follow the distribution $f(\delta_i)$ in (F.2.2).

If we ignore that several assessors were used in this study, and that they might have different sensitivities, we obtain a marginal $d'_m$ of 1.67 and an estimated
standard error of 0.186. We know that this estimate of the standard error is optimistic, since it ignores the variance between individuals.

The estimate of $\hat{\pi}_m$ and $d'_{\text{m}}$ is the same, when using the beta-binomial model, but the standard error of the estimate is larger.

![Figure F.3: Random effects for the Duineveld and Meiners data set. The propensity of being a discriminator is shown in the plot.](image)

If we fit the model (F.2.3), we obtain the parameter estimates $(\hat{\delta}_c, \hat{\sigma}_\delta) = (1.62, 1.08)$, where the standard errors of $\hat{\delta}_0$ and $\log \hat{\sigma}_\delta = 0.0763$ are 0.234, 0.296. Comparing these estimates with the marginal estimates above, we see that the marginal estimate $d'_{\text{m}}$ is slightly larger than the conditional estimate $\hat{\delta}_c$. This is exactly, what we expect cf. figure [F.2] where (since the estimated standard deviation (1.08) is close to that used in the figure (1.00)) we see that the stationary point for the triangle test is around two, hence $d'_{\text{m}}$ should—and indeed is, closer to two than $d'_c$. Using the function delta0 enclosed in the appendix, we find that the stationary point for the triangle test with $\sigma_\delta = 1.08$ actually occurs at $(\hat{\delta}_0, \pi_0) = (1.99, 0.603)$.

The conditional parameter estimates given above are the estimated mean and standard deviation of the distribution of $\delta^*$, and are not the mean and standard deviation of the distribution of $\delta_i$. The mean and standard deviation of $\delta_i$ are
instead $\hat{\delta}' = 1.65$ and $\hat{\sigma}_{\delta'} = 1.02$ respectively. It is worth to note that the standard error of the mean $d'$ is larger in the conditional model than in the marginal naive model reflecting the variability among assessors.

A plot of the $\hat{\delta}$s is given in figure E.3 (vertical axis) with normal quantiles on the horizontal axis. In the plot we have shown the probabilities of each individual being an individual, i.e. having $d' > 0$. According to the model, the probability of a randomly selected assessor being a discriminator is 0.933.

A better plot would be obtained by using the theoretical quantiles of the density $f(\delta_i)$, rather than simply the normal, but as the probabilities of all individuals being discriminators are rather high (all are above .70), we do not expect this to disturb the picture. The plot is seen to be fairly linear supporting the assumed model. The deviation from linearity for the lower quantiles are expected, since the plot will have a horizontal asymptote at zero reflecting that no individual can have a $d'$ below zero. It is seen that none of the examined assessors are expected to have a $d'$ below one half and that the better assessors have $d'$s close to three. From looking at the data we see that a few assessors had very low success rates (0 out of 7, 2 out 10 and 1 out 6), hence a better fit may be obtained by including an additional mixture probability. Such a model would assume that assessors ($\delta_i$s) were drawn from a distribution being a mixture of the one proposed above and an additional mass in zero.

F.6.2 Example 2

Bi and Ennis (1998) reports on an experiment using the 2-AFC test with 30 panelists and 10 replications. The pooled estimate of $d'$ and its standard error (naive estimate) are 0.715 and 0.107. Fitting the proposed conditional model to these data give the estimates $\hat{\delta} = 0.78$ with $\hat{\sigma}_{\delta} = 0.205$ and $\hat{\sigma}_\delta = 1.19$. The corresponding discriminal ability for an average individual is $\hat{\delta}' = 0.964$. The standard deviation $\hat{\sigma}_{\delta'}$ is 0.938.

The random effects plot for the Bi and Ennis data are show in figure E.4 Notice the asymptotic behaviour towards zero. In this plot the discrete nature of the data is more obvious and also now the range of individual propensities of being a discriminator is larger than in the previous example. The probability of a randomly sampled individual being a discriminator is 0.744.
Figure F.4: Random effects for the Bi and Ennis data set. The propensity of being a discriminator is shown in the plot.
F.7 Computational Issues

The conceptually simplest and most direct approach is to estimate $\theta$ in the likelihood (F.2.5), possibly including $p^o$ in the three-parameter model is using a general optimization routine such as the Gauss-Newton or BFGS procedure (Nocedal and Wright, 2006). The integral can also be handled using a general integration routine such as adaptive quadrature. Care need to be taken in scaling the parameters appropriately. A reasonable possibility is to scale $p^o$ by one tenth to ensure comparability in information/curvature with $\delta$ and to perform optimization of $\sigma_\delta$ on the log-scale to make it approximately quadratic around the mode and allow optimization on the entire real line. If lower and upper bounds are not allowed on the parameters, the optimizer should return infinity (or a large value), if values for $p^o$ outside (0, 1) are tried. This simple approach is implemented in function `discrimR` in package `sensR`. The likelihood in (F.2.5) includes the parameter $p^o$ and a direct optimization of this parameter as well as the remaining parameters is possible (as is implemented in `discrimR3`—R for Replicated and 3 for 3-parameter model included in appendix.)

To lower the computational burden, the constant terms in the likelihood and expressions for the conditional expectations are ignored. These constant terms cancel out in the expressions for the conditional expectations, hence they do not change. The absolute value of the likelihood does however change, hence the reported deviance might be incomparable to that of other models.

The integrals in the expressions for the likelihood and conditional expectations are not evaluated exactly as they are written above. It is preferable to perform a change of variables to $b_i = \delta_i - \delta$. For instance the integral in the likelihood, i.e. the last integral in (F.2.5) then becomes

$$
\int_{-\delta}^{\infty} f_\pi(y_i|b_i)\phi(b_i/\sigma_\delta)/\sigma_\delta \, db_i .
$$

(F.7.1)

F.7.1 A Truncated Laplace Approximation

Another possible although approximate way to evaluate the integral, is to consider the laplace approximation. This method approximates an integral on the whole real line, so it has to be “truncated” to apply to integrals such as the one in (F.7.1). We will consider such a truncation in this section.

We may consider a laplace approximation of the second term in the likelihood.
function $f_2(y_i)$ in (F.2.5). Denote the integrand
\[ h(y_i, b_i) = f_\pi(y_i|b_i) f_b(b_i; \mu, \sigma_\delta), \]
where we have performed a change of variables; $b_i = \delta_i - \delta$ and $f_b()$ is the normal density of $b_i$ with parameters $\mu$ and $\sigma_\delta$. We approximate the integrand by a Gaussian, i.e. a bell-shaped curve corresponding to a second order Taylor expansion on the log scale
\[ \log h(y_i, b_i) \approx \log h^*(y_i, b_i) \]
\[ = \log h(y_i, \tilde{b_i}) - \frac{1}{2} D(b_i - \tilde{b_i})^2 + O(), \]
where $\tilde{b_i}$ is the mode of $h(y_i, b_i)$ wrt. $b_i$ and
\[ D = -\frac{\partial^2}{\partial b_i^2} \log h(y_i, b_i)|_{b_i=\tilde{b_i}}, \]
hence $D$ is the observed Fisher information.

The approximation is then
\[ f_2(y_i) = \frac{1}{p} \int_{-\delta}^{\infty} h(y_i, b_i) \, db_i \]
\[ \approx \frac{1}{p} \int_{-\delta}^{\infty} h^*(y_i, b_i) \, db_i \]
\[ = \frac{1}{p} h(y_i, \tilde{b_i}) \sqrt{2\pi\hat{\sigma}^2} \left[ 1 - \Phi \left( -\frac{\delta - \hat{\mu}}{\hat{\sigma}} \right) \right], \]
where $\hat{\mu} = h(y_i, \tilde{b_i})$ and $\hat{\sigma}^2 = D^{-1}$.

Excluding the term in square brackets gives the ordinary Laplace approximation on $\mathbb{R}$, hence the term in square brackets truncates the region of integration to $b_i > -\delta$. The approximation (F.7.2) may therefore be denoted a “truncated laplace approximation”.

As an example consider the psychometric function for the 3AFC method
\[ f(\delta) = \int_{-\infty}^{\infty} \Phi^2(z) \phi(z - \delta) \, dz, \]
where the integrand is $h(\delta, z) = \Phi^2(z) \phi(z - \delta)$. This function is shown as the solid line in figure F.5 for $\delta = 1$. The Gaussian approximation, $h^*(\cdot)$ is also shown as the red, dashed curve. The ordinary laplace approximation to $f(\delta)$ is $f(1) \approx h(\delta, \tilde{z}) \sqrt{2\pi\hat{\sigma}^2} = 0.6305$, whereas the exact (i.e. a much better but computational infeasible approximation) integral is $f(1) = 0.6337$. 

\[ \]
Figure F.5: Gaussian approximation (red, dashed curve), approximated psychometric function for the 3AFC test (black, solid curve) and truncation point $Z = 1$ (red, dashed vertical line).
Suppose now we need the integral from $Z = 1$ to $\infty$ instead. We can then approximate the integral by $f_Z(\delta) \approx h(\delta, \hat{z}) \sqrt{2\pi \hat{\sigma}^2} \left[ 1 - \Phi\left( \frac{-\delta - \hat{\mu}}{\hat{\sigma}} \right) \right]$, hence $f_Z(1) = 0.421$ whereas the exact integral is $f_Z(1) = 0.445$.

The integrand is a bit skewed to the right whereas the Gaussian approximation (naturally) is symmetric. The ordinary laplace approximation is however quite accurate because the error to the left of the center almost cancels out with the error to the right of the center as is evident from figure [F.5]. With the truncated laplace approximation, the error on one side of the center is potentially much larger (relatively) than on the other side. For instance, the error to the right of the center in the approximation in figure [F.5] is much larger than to the left, if the integrand is truncated at $Z = 1$. We therefore in general expect a larger error with the truncated laplace approximation than with the ordinary laplace approximation.

The absolute and relative error of the ordinary laplace approximation are 0.32% and 0.51% respectively whereas the errors of the truncated laplace approximation with truncation point $Z = 1$ are 2.4% and 5.4% respectively, hence about 10 times larger. The accuracy of the approximation depends on how close the log-integrand is to being quadratic and where the truncation point is located. For small truncation points, the truncated and ordinary laplace approximations almost coincide, whereas for truncation points around the center of the density, the error increase the further the log-integrand is from being quadratic. For large truncation points, the approximation as well as the exact solution tends to zero, hence the absolute error also tends to zero.

### F.7.2 Estimation via the EM algorithm

The parameters of the conditional model can also be estimated via the EM algorithm: Denote the full data $x_i = (y_i, Z_i)$. Assuming $Z_i$ is known, we can derive the weighted log likelihood based on the this data as

$$l_i^w(\theta; x_i) = (1 - Z_i) \log f_1(y_i) + Z_i \log f_2(y_i) .$$

The M-step consists of optimizing this objective function and the E-step consists of estimating $Z_i$ as above. The weighted log-likelihood can be written

$$l^w(\theta; y_i) = l_1(\theta; y_i) + l_2(\theta; y_i) ,$$

(F.7.3)
where the two terms are
\[ l_1(\theta; y_i) = \log f_{w_1}^w(y_i) \]
\[ = w_1^i m_i \left\{ y_i^* \log \frac{p_0}{1-p_0} + \log(1-p_0) \right\} + \log \left( \frac{m_i}{m_i y_i^*} \right) \]
\[ l_2(\theta; y_i) = \log f_{w_2}^w(y_i) \]
\[ = \log \int_{-\delta}^{\infty} f_{\pi}^w(y_i | b_i) f_b(b_i; 0, \sigma_\delta) \, db_i, \]
where
\[ \log f_{\pi}^w(y_i | b_i) = w_2^i m_i \left\{ y_i^* \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i) \right\} + \log \left( \frac{m_i}{m_i y_i^*} \right) \]
\[ f_b(b_i; 0, \sigma_\delta) = \phi(b_i / \sigma_\delta) / \sigma_\delta, \]
and \( w_1^i = 1 - Z_i \) and \( w_2^i = Z_i \) are prior weights, \( m_i \) is the binomial denominator and \( y_i^* = y_i / m_i \) is the scaled response. The additive constants in the expressions of \( \log f_{w_1}^w(y_i) \) and \( f_{\pi}^w(y_i | b_i) \) may be ignored in the estimation as they do not change the shape of log likelihood function or the location of its maximum.

Note that when \( p_0 = 1/2 \) as it occurs in the duo-trio and the 2AFC tests, we have that \( \log \frac{p_0}{1-p_0} = 0 \), hence \( l_1(\theta; y_i) = w_1^i m_i \log 1/2 \), which depends on the data and the parameters only through the weights \( w_1^i \) and \( m_i \).

This formulation of the likelihood has the advantage, that the two separate models can be assessed individually. They each contain parameter estimates, standard errors, residuals and fitted values.

The EM algorithm iterates between the following two steps

1. With fixed weights, estimate \( \theta = (\delta, \sigma_\delta) \) using the likelihood (F.7.3)
2. With fixed \( \theta \), estimate the weights \( Z_i \) using either of
\[ \tilde{Z}_i = \frac{p_0 f_2(y_i)}{f(y_i)} \]
\[ \tilde{Z}_i = 1 - \frac{(1-p_0) f_1(y_i)}{f(y_i)}, \]
The estimator of \( Z_i \) coincides with that given in (F.3.2). Here \( p^o = p = 1 - \Phi(-\delta / \sigma_\delta) \) in the two-parameter model and \( p^o = \frac{1}{n} \sum_i Z_i \) in the three-parameter model.
F.7.2.1 Extended likelihood formulation

The extended likelihood (Pawitan, 2001; Lee and Nelder, 2006a) is concerned with the fixed parameters $\delta, \sigma_\delta$ and potentially $p_\theta$, the random parameters $Z, b$ based on the observed data $y$. The data are generated by sampling from the distributions of random effects $f_{p_\theta}(Z)$ and $f_{\sigma_\delta}(b)$ and conditional on these values, sampling from $f_{\delta}(y|Z,b)$. From an inference point of view, we wish to estimate $\delta, \sigma_\delta, p_\theta$ from the marginal likelihood $L(\delta, \sigma_\delta, p_\theta; y)$ and estimate $Z_i, b_i$ from $f(Z, b|y)$. The extended likelihood is

$$L_e(\delta, \sigma_\delta, p_\theta, b, Z; y, b, Z) = f(b) f(Z) f(y|b, Z) = f(b, Z|y) f(y) = f(Z|b, y) f(b) f(y) ,$$

where

$$f_{p_\theta}(Z) = p_\theta^Z (1 - p_\theta)^{(1-Z)}$$

$$f_{\sigma_\delta}(b) = \begin{cases} 
1 & b = -\delta \\
\frac{1}{p} (2\pi \sigma_\delta)^{-1/2} \exp \left\{ \frac{b^2}{2\sigma_\delta^2} \right\} & b > -\delta 
\end{cases}$$

$$f_{\delta}(y_i|b, Z) = f_1(y_i|b_i)^{(1-Z_i)} f_2(y_i|b_i)^{Z_i}$$

$$f_1(y_i|b) = k_{p_\theta}^y/(1 - p_\theta)^{(y_i-n_i)}$$

$$f_2(y_i|b) = k_{\pi} y_i/(1 - \pi_i)^{(y_i-n_i)}$$

This extended loglikelihood is also a h-loglikelihood (Lee et al., 2006), since the random effects are on the weak canonical scale, i.e., occurs linearly in the predictor. We may use the h-loglikelihood directly for estimation of all parameters, but it may give biased estimates if the information per parameter is low.

The likelihood integrated with respect to. $b$ also can be obtained as

$$L_{Z,b}(\delta, \sigma_\delta, p_\theta; y|b, Z) = f_1(y)(1 - p_\theta) + p_\theta \int f_{2,\delta}(y|b) f_{\sigma_\delta}(b) \, db$$

$$\approx f_1(y)(1 - p_\theta) + p_\theta \left\{ f_{2,\delta}(y|\tilde{b}) f_{\sigma_\delta}(\tilde{b}) \sqrt{2\pi D^{-1}} \left[ 1 - \Phi \left( \frac{-\delta - \tilde{b}}{D^{-1/2}} \right) \right] \right\} \bigg|_{b=\tilde{b}} ,$$

and the (approximated) log-likelihood can be written as

$$p_{Z,b}(\delta, \sigma_\delta, p_\theta; y|b, Z) = \log f_1^w(y) + \left\{ f_{2,\delta}^w(y|b) + \log f_{\sigma_\delta}(b) + \frac{1}{2} \log |2\pi D^{-1}| + \log \left[ 1 - \Phi \left( \frac{-\delta - \tilde{b}}{D^{-1/2}} \right) \right] \right\} \bigg|_{b=\tilde{b}} ,$$
where the second part of the extended log likelihood (first term in curled braces) is penalized by the second term in curled braces. This far the likelihood is the ordinary laplace approximation, which also resembles the adjusted profile likelihood of Cox and Reid (1987) (hence the notation with \( p_v() \); the profile log likelihood, where \( v \) is profiled out). The last term in curled braces is the truncation adjustment. All three terms are evaluated at \( \hat{b}_i \); the maximizer of \( l_{2e}(\theta,b_i) \) with respect to \( b_i \).

It is interesting to note, that when there is much information per (random) parameter, the maximizer of \( l_{2e}(\theta,b) \) is a close approximation to \( l(\theta;y) \) and therefore also to \( p_v(\theta;y_i) \), hence the mode of the integrand is a close approximation to the integral itself. This does however not hold, when the per parameter information is low, where we expect the penalizing term in \( p_v(\theta;y_i) \) to be of substantial magnitude. Also the truncation adjustment will be substantial with little per parameter information and more so if \( \delta \) is small and \( \sigma_\delta \) is not small.

Note that \((Z_i,b_i)\) can be seen as one random effect with two attributes; \( Z_i = P(b_i > 0) \) and \( b_i \) is the value of the random effect, if \( b_i > 0 \). We therefore have two measures of uncertainty in \( b_i \); the variance of the estimate \( \hat{b}_i \) and the posterior probability that \( b_i \) is not zero; \( P(b_i > 0) = P(b_i \neq 0) \). We can approximate the variance of \( \hat{b}_i \) with

\[
\text{var}(\hat{b}_i) = (Z^t W Z + \Sigma^{-1})^{-1},
\]

where \( Z \) is the indicator matrix for the random effects, \( W \) is the diagonal GLM weight matrix and \( \Sigma \) is the covariance matrix from \( f(b) \). This estimate resembles the inverse of the expected rather than the observed Fisher information matrix. This variance estimate is then conditional on \( b_i > 0 \). If we use \( Z_i \) as weights in the GLM and therefore scale the elements of \( W \) with \( 1/Z_i \), we obtain another unconditional variance estimate, which includes the uncertainty as to whether \( b_i \) is actually positive so it will be larger.

Appendix

F.8 Additional R-functions

The additional R-functions include functions to compute the mean and variance of the truncated normal distribution as well as the random effect distribution of the two-parameter model, functions to compute the marginal prediction of \( \pi; \pi_m, \delta_0 \) at the stationary point for some \( \pi_0 \) and dispersion, the truncated normal density and a function that will fit the three-parameter model.
## Auxiliary function

```r
ct <-
## Expectation of the truncated normal distribution, truncated
function(t) 1/(sqrt(2*pi) * (1 - pnorm(t)))

E.Zi <- function(tc.point, mean, sd) {
## at tc.point with parameters mean and sd.
  t <- (tc.point - mean)/sd
  sd * ct(t) * exp(-t^2/2) + mean
}

E <- function(tc.point, mean = 0, sd = 1) {
## Expectation of the truncated normal mixture distribution,
## truncated at tc.point with parameters mean and delta.
  t <- (tc.point - mean)/sd
  c1 <- 1/sqrt(2 * pi)
  sd * (t * pnorm(t) + c1 * exp(-t^2/2)) + mean
}

E <- function(tc.point, mean = 0, sd = 1) {
## Expectation of the truncated normal mixture distribution,
## truncated at tc.point with parameters mean and sd.
  t <- (tc.point - mean)/sd
  t <- (tc.point - mean)/sd
  c1 <- 1/sqrt(2 * pi)
  sd * (t * pnorm(t) + c1 * exp(-t^2/2)) + mean
}

V <- function(tc.point, mean = 0, sd = 1) {
## Variance of the truncated normal mixture distribution, truncated
## at tc.point with parameters mean and sd.
  t <- (tc.point - mean)/sd
  Eest <- E(tc.point, mean, sd)
  fun <- function(x) (x - Mean)^2/sd * dnorm((x - mean)/sd)
  val <- double(length(tc.point))
  for(i in 1:length(tc.point)) {
    Mean <- Eest[i]
    val[i] <- integrate(Vectorize(fun), tc.point[i], Inf)$value
  }
  cbind((tc.point - Eest)^2 * pnorm(t) + val, (tc.point - Eest)^2 * pnorm(t), val)
}

V.Zi <- function(tc.point, mean = 0, sd = 1) {
## Variance in the truncated normal distribution, truncated at
## tc.point with parameters mean and sd.
  t <- (tc.point - mean)/sd
  c.t <- ct(t)
  c3 <- pchisq(t^2, df = 3)
  c3 <- ifelse(t < 0, -c3, c3)
  v <- c.t * (sqrt(pi/2) * (1 - c3) - c.t * exp(-t^2) )
  sd^2 * v
}

seE <- function(se, m, sd) {
## Standard error of an estimate delta', based on the parameters
## delta and sigma_delta and the standard error of delta, se.
  grad <- (1 - 1/sd) * m * pnorm(-m/sd) / sd + pnorm(m/sd)
  se * grad
}

piMarg <-
## Estimate the marginal probability; \pi_m given the delta
## parameter, the standard deviation parameter in random effects
## distribution and the name of the discrimination protocol.
function(delta = 0, sd = 1, method = c("duotrio", "threeAFC", "triangle", "twoAFC", "probit"), p0, ...) {
  fun <- function(di, d) link.inv(di) * dnorm((di - d)/sd) / sd
  Call <- match.call()
  m <- match.call(expand.dots = FALSE)
  m$... <- NULL
  m[[1]] <- as.name("c")
  m <- eval.parent(m)
```
method <- match.arg(method)
link.inv <- switch(method,
  duotrio = duotrio()$linkinv,
  threeAFC = threeAFC()$linkinv,
  triangle = triangle()$linkinv,
  twoAFC = twoAFC()$linkinv,
  probit = pnorm
)

if(missing(p0))
  p0 <- ifelse(method %in% c("triangle", "threeAFC"), 1/3, .5)
else
  stopifnot(p0 > 0 && p0 < 1)
  stopifnot(all(delta >= 0))
  stopifnot(length(sd) == 1 && sd > 0)
  p <- pnorm(delta / sd)
  val <- double(length(delta))
  for(i in 1:length(delta))
    val[i] <- integrate(fun, 0, Inf, d = delta[i])$value
  PI <- p0 * (1 - p) + val
  list(pi.marg = PI, Call = Call, method = method, delta = delta, sd = sd)
}
delta0 <-
## Estimate the stationary point (delta_0, pi_0). Calculates delta_0
## given some probability, p0 and standard deviation, sd and name of
## the discrimination method of interest.
function(sd = 1, method = c("duotrio", "threeAFC", "triangle", "twoAFC",
  "probit"), p0, ...)
{
  rootFun <- function(delta, sd, p0) {
    link.inv(delta) - piMarg(delta = delta, sd = sd, method = method,
      p0 = p0)[[1]]
  }
  Call <- match.call()
  m <- match.call(expand.dots = FALSE)
  m$method <- m$... <- NULL
  m[[1]] <- as.name("c")
  m <- eval.parent(m)
  method <- match.arg(method)
  link.inv <- switch(method,
    duotrio = createduotrio()$linkinv,
    threeAFC = createthreeAFC()$linkinv,
    triangle = createtriangle()$linkinv,
    twoAFC = createtwoAFC()$linkinv,
    probit = pnorm
  )
  if(missing(p0))
    p0 <- ifelse(method %in% c("triangle", "threeAFC"), 1/3, .5)
  else
    stopifnot(p0 > 0 && p0 < 1)
    stopifnot(all(sd > 0))
    d0 <- pi0 <- double(length(sd))
    for(i in 1:length(sd)) {
      d0[i] <- uniroot(rootFun, c(.01, 6), sd = sd[i], p0 = p0)$root
    }
    pi0 <- link.inv(d0)
    list(d0 = d0, pi0 = pi0, Call = Call, method = method, SD = sd)
}
## Estimates the parameters of the three-parameter thurstonian model for replicated discrimination tests. See function 'discrimR' in package 'sensR' for further details about arguments, output etc.

```
function(formula, data, weights, cluster, start, subset, na.action, 
  contrasts = NULL, hess = FALSE, ranef = FALSE, zi = FALSE, 
  method = c("duotrio", "probit", "threeAFC", "triangle", 
  "twoAFC"), ...) 
```

```{r}
## ...: arguments to optim, such as control=list(trace=TRUE)
## start: optional starting values

### Function to minimized:

```{r}
fminNM <- function(beta) {
    return(Inf)
  for(j in 1:nl) {
    func1 <- function(u) {
      x <- X[ind==j, ,drop=FALSE] ## within cluster
      y <- Y[ind==j]
      w <- wt[ind==j] 
      p <- link.inv(eta)
      b <- exp(beta[3] * na) ## exp(sigma)
      d <- dnorm(u/b)/b ## normal density
      p0 <- p0*(w * y) * (1 - p0)^(w * (1 - y))
      fp0 <<- p0^(w * y) * (1 - p0)^(w * (1 - y))
    }
    h[j] <<- integrate(Vectorize(func1), -beta[1], Inf)$value
    f[j] <<- (1 - beta[3]) * fp0 +
  }
  if (all(f > 0))
    -2*sum(log(f)) # negative log likelihood
  else Inf
}
```

```{r}
m <- match.call(expand.dots = FALSE)
m$start <- m$hess <- m$method <- m$... <- m$ranef <- m$zi <- NULL
m[[1]] <- as.name("model.frame")
m <- eval.parent(m)
Terms <- attr(m, "terms")
X <- model.matrix(Terms, m, contrasts)
con <- attr(X, "contrasts")
wt <- model.weights(m)
if (!length(wt))
  wt <- rep(1, n)
offset <- model.offset(m)
if (length(offset) <= 1)
  offset <- rep(0, n)
Y <- model.response(m)
if (NCOL(Y) == 2) {
  n <- Y[, 1] + Y[, 2]
  Y <- ifelse(n == 0, 0, Y[, 1]/n)
  wt <- wt * n
}stopifnot(all(wt >= 0))
method <- match.arg(method)
link.inv <- switch(method, 
  duotrio = duotrio()$linkinv, 
  probit = pnorm, 
  threeAFC = threeAFC()$linkinv, 
  triangle = triangle()$linkinv, 
  twoAFC = twoAFC()$linkinv, 
)
F.8 Additional R-functions 207

## Parameters to fminNM:

nb <- NCOL(X); nb
ns <- 1; ns ## 1
nl <- nlevels(m$n"(cluster)")); nl
ind <- c(unclass(m$n"(cluster)"); ind ## indcator for clusters
f <- h <- g <- fp0 <- double(nl)
p0 <- ifelse(method %in% c("triangle", "threeAFC"), 1/3, .5)
Call <- match.call()

## Starting values:

suppressWarnings{
  if(missing(start)) {
    gf <- glm.fit(X, Y, family=binomial(probit))
    b1 <- as.vector(gf$coef); b1
    start <- c(b1, 1)
  }
}
if(length(start) != nb + ns + 1)
  stop("start is not of correct length")
if(start[nb + ns] <= 0)
  stop("starting value for sigma needs to possitive")
start <- c(start[1:nb], log(start[nb + ns]), start[3])

## Optimization

fit <- optim(par=start, fn=fminNM, method="BFGS",
              hessian=hess, ...)

## Output:

fpar <- fit$par[c(1, 3)]; fpar
rpar <- c(fit$par[nb+ns], exp(fit$par[nb+ns])); rpar
se <- if(hess) sqrt(diag(solve(fit$hessian)))
else NULL
deviance <- fit$value; deviance
p <- 1 - pnorm(-fit$par[1]/exp(fit$par[nb + ns]))
c2 <- fpar[2] / p

data {  
   for(j in 1:nl) {  
     func2 <- function(u) {  
       x <- X[ind==j, ,drop=FALSE] ## within cluster
       y <- Y[ind==j]
       w <- wt[ind==j]
       p <- link.inv(eta)
       b <- exp(fit$par[nb + ns]) ## exp(sigma)
       d <- dnorm(u/b)/b ## normal density
       prod (p^(w * y) * (1 - p)^(w * (1 - y)) ) * (fit$par[1] + u) * d
     }
     g[j] <- integrate(Vectorize(func2), -fit$par[1], Inf)$value
   }
   ranef <- c2 * g / f
   fitted <- link.inv(ranef)
   resid <- sign(Y - fitted) * sqrt(-2*log(f))
   V <- fitted * (n - fitted)/n
   resid.p <- (Y - fitted)/sqrt(V)
  }
else
  {  
   ranef <- NULL
   g <- NULL
  }
if(zi) {
zi <- fpar[2] * (h / p) / f

else {
  zi <- NULL
}

list(fpar=fpar, rpar=rpar, deviance=deviance, se=se, Call = Call,
     convergence=fit$convergence, lli = log(f), ranef = ranef, g = g, zi
     = zi, p = p, c2 = c2, fitted = fitted, resid = resid,
     resid.p = resid.p, Y = Y)

tnorm <- function(t, z, mean = 0, sd = 1) {
  ## Density function for a truncated normal distribution, truncated
  ## at 't' with parameters 'mean' and 'sd' (of the un-truncated
  ## density).
  ct <- 1 - pnorm(t, mean, sd)
  f <- double(length(z))
  f <- ifelse(z >= t, dnorm((z - mean)/sd)/sd/ct , t)
  f
}


discrimination rates in replicated triangle tests. *Food Quality and Preference*
19, pp. 292–305.

Science* 13(2), pp. 95–122.

80–91.

Ennis, D., J. Bi, and M. Meyners (2007). Letter to the editor: Discussion on

Sensory Studies* 8(53-70).

Ennis, D. M. and J. Bi (1998). The beta-binomial model: Accountin for inter-
trial variation in replicated difference and preference tests. *Journal of Sensory

for inter-trial variation in replicated ratings. *Journal of Sensory Studies* 14,

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on
Generalized Linear Models* (Second ed.). Springer series in statistics. Springer-
Verlag New York, Inc.

Frijters, J. E. R. (1979). Variations of the triangular method and the relation-
ship of its unidimensional probabilistic models to three-alternative forced-
choice signal detection theory methods. *British Journal of Mathematical and
Statistical Psychology* 32(229-241).


Normal and logistic models. *Journal of Mathematical Psychology* 9, pp. 128–
139.

New York.

January, Pau, France.


R Development Core Team (2008c, June). *Writing R Extensions*. Version: 2.7.1.


Schuirmann, D. J. (1981). On hypothesis testing to determine if the mean of a normal distribution is contained in a known interval. *Biometrics* 37, pp. 617.


