NOVEL STATISTICAL METHODS IN CONJOINT ANALYSIS AND PADÉ APPROXIMATION OF THE PROFILE LIKELIHOOD

by

Thomas James Prior

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

June, 2017

Abstract

This thesis addresses two topics in statistics that rely on the likelihood function for statistical inference. These topics include novel statistical research in conjoint analysis and a method to approximate the profile likelihood and create approximate profile likelihood confidence intervals.

After introducing choice-based conjoint analysis, we introduce a method to empirically assess and compare different conjoint analysis design strategies. A key feature of this method of comparison is that it makes very few assumptions about the data-generating process (essentially just that the respondents answer the surveys independently of one another) while remaining statistically valid; in particular, the respondents are not assumed to use the multinomial logit model.

We then turn to the statistical analysis of conjoint analysis survey results. We introduce a method to plot in two-dimensional space the heterogeneity in preferences across the respondents as inferred from the conjoint analysis survey results that can be used for visualization as well as mixture model assessment. We also introduce a novel method to accurately infer the number of natural clusters in preferences across the respondents. This method is shown in simulation studies to give more accurate results than latent class segmentation with AIC and BIC. We additionally suggest regressing estimated respondent preferences on their demographic variables as a strategy for hypothesis generation and model building.

We show that the profile likelihood can be well approximated near the maximum

likelihood estimate by the [2,2] Padé approximant. This approximation is shown to be better than that provided by a second-order Taylor series approximation. However, like a second-order Taylor series approximation, it can be used to construct a confidence interval with endpoints found using the quadratic formula. The resulting confidence interval is similar to a profile likelihood interval but with computation time similar to that of a Wald interval.

Thesis Readers: Charles A. Rohde (Advisor), Gary L. Rosner (Committee Chair), John F.P. Bridges, and Brian S. Caffo.

Acknowledgments

I thank my advisor, Dr. Charles Rohde, for his insight, direction, mentoring, and friendship during the course of this dissertation, as well as throughout my entire stay at the Johns Hopkins Bloomberg School of Public Health. I have worked closely with Dr. Rohde for my entire graduate career, and his critical feedback and excellent advice have been crucial throughout the process. I owe Dr. Rohde a debt of gratitude that cannot be summarized via acknowledgements.

Thanks go to Dr. John Bridges who sparked my interest in discrete choice experiments. Dr. Bridges kindly shared his expertise in this area with me. He also shared his data with me and allowed me to use these data to develop these statistical analyses. Dr. Bridges has been a great mentor and friend to me over the past several years.

Thanks go to Dr. Brian Caffo and Dr. Gary Rosner, committee chair, for serving on my dissertation committee. Both took time out of their busy schedules to provide exceptional feedback and to advise me through the process of writing my dissertation. My dissertation benefited a great deal from their feedback and suggestions.

I would like to thank Dr. Karen Bandeen-Roche for her excellent guidance and her kind and caring manner throughout my stay in the Biostatistics department. Dr. Bandeen-Roche contributed greatly to my development in the department.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Summary

This thesis addresses two topics in statistics. Novel statistical research in conjoint analysis is discussed in Chapters 2 and 3, and a method to approximate the profile likelihood and create approximate profile likelihood confidence intervals is described in Chapter 4. These topics are related in their reliance on the likelihood function for statistical inference.

There have been numerous applications of likelihood methods in public health. Likelihood methods are often used in public health to infer parameters that equal or represent quantities of significant public health interest, such as the risk of tuberculosis (Clark and Vynnycky 2004), breast cancer (Consortium and others 1999), and heart disease (Anderson et al. 1991). Such quantities can be termed "parameters of interest" and analyzed using the methods of Chapter 4. In applications of conjoint analysis in public health, the parameters of the likelihood represent the patient's preferences regarding medical choices, such as among Medicare Part D plans (Heiss, McFadden, and Winter 2010).

The profile likelihood is definable for any likelihood with more than one parameter, and the statistical methods in conjoint analysis model a person's choices using a multinomial logistic likelihood or a mixture model of such likelihoods. Because of this, methods involving the profile likelihood are usefully employed in conjunction with the statistical methods described for conjoint analysis. For example, in Swait's work on the role of the scale parameter in the estimation and comparison of multinomial logit models in conjoint analysis, a profile likelihood is used to estimate and create a confidence interval for the scale parameter (Swait and Louviere 1993).

1.2 Overview

In Chapters 2 and 3, novel statistical methods in conjoint analysis developed to solve specific real-world problems are described. Chapter 2 introduces conjoint analysis as well as a novel statistical method related to evaluating the design of discrete choice experiments, whereas Chapter 3 introduces novel statistical methods related to the analysis of discrete choice experiments.

Chapter 2 presents a novel statistical test to compare empirically the statistical efficiency of different design techniques. This test has been used to demonstrate that a design without attribute overlap outperforms a design with attribute overlap, without biasing the estimates of respondent preferences. This method avoids making assumptions about the data-generating process, and instead uses resampling of model-robust standard errors to create a confidence interval for the relative statistical efficiency of two designs.

Chapter 3 presents a novel statistical method for conjoint analysis for making correct inferences about the distribution of preferences across respondents, especially regarding the inference of the number of natural clusters of preferences. This method

demonstrates that inspection of individual respondents' preferences provides insight into the mixing distribution even when the parametric form that is hypothesized for it during the model fitting process is false. Finally, techniques for visualizing the estimates of individual respondents' preferences are presented which provide greater insight into respondents' preferences than the values of inferred model parameters or goodness-of-fit statistics alone.

Chapter 4 presents a method for obtaining confidence intervals when the sample size is small or there are many nuisance parameters. This method involves the Padé approximation of the profile likelihood function. The approximation involves no additional numerical maximizations of the the full likelihood after the maximum likelihood estimate is found. Instead, this method uses derivatives of the profile likelihood function at its maximum in order to obtain an approximation, and so in that way it is conceptually similar to methods described by Viveros (1987) and DiCiccio (2001).

Chapter 5 concludes with a summary of the thesis and of the limitations of the methods discussed.

Chapter 2

Conjoint Analysis and a Novel Method for Experimental Design Assessment

2.1 Overview

This chapter presents a novel statistical method developed to solve a specific real-world problem related to the design of conjoint analysis surveys. Specifically, a statistical test to compare empirically the statistical efficiency of different design techniques is presented. This method avoids making assumptions about the data-generating process, and instead uses resampling of model-robust standard errors to create a confidence interval for the relative statistical efficiency of two designs. Additionally, this chapter provides background and context with which to understand the problem and its solution.

Section 2.2 provides background on conjoint analysis. Section 2.3 describes the multinomial logit model, a probabilistic statistical model that is often used to make

inferences from conjoint analysis data. Section 2.4 provides an overview of areas of research in the design of conjoint analysis surveys. Section 2.5 presents a research problem in conjoint analysis survey design that was addressed using statistics, i.e., the comparison of the statistical efficiency of conjoint analysis survey design strategies.

2.2 Background

2.2.1 What is conjoint analysis?

Conjoint analysis began with a theoretical paper by Luce and Tukey (1964) in the form of "simultaneous conjoint measurement," a concept that was later introduced in an applied context in market research on consumers of goods (Green and Rao 1971), where it eventually became known as "conjoint analysis" (Green and Srinivasan 1978). In typical applications of conjoint analysis, a person (e.g., a consumer or a patient) is presented with a choice between two or more "alternatives," where each alternative is composed of two or more "attributes," the "levels" of which differ between the alternatives. The person must choose exactly one of the alternatives; their "response" to the "choice task" reveals which attributes are most impactful on the choice and the manner in which specific levels of those attributes impact the choice. Figure 4.1 displays the terminology using as an example a choice of preferred credit card.

As another example, a person could be asked to choose between a car that costs $20,000 with a gas mileage of 25 miles per gallon and another car that costs $35,000 with a gas mileage of 35 miles per gallon. That person's response shows what is more important to him when considering purchasing a car – gas mileage or car price. This particular kind of conjoint analysis is sometimes referred to as "choice-based conjoint analysis" or a "discrete choice experiment," since the respondent is asked to make a *choice* between two or more items. This is in contrast to metric conjoint

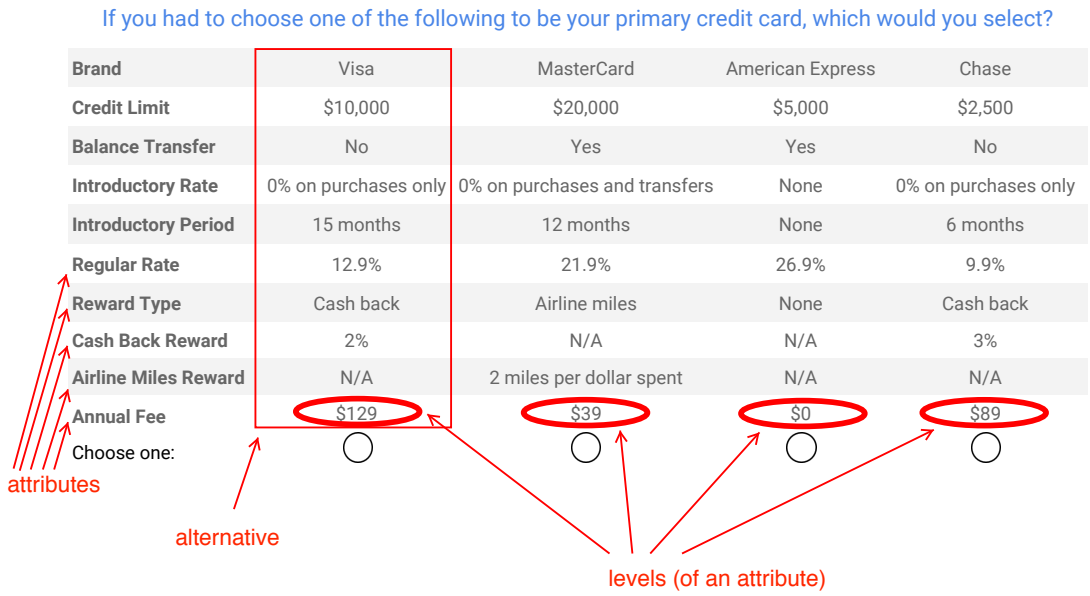

Figure 2.1: An example of a choice task from an online survey

analysis, where, for example, a respondent might be asked to assign, on a rating scale, how much he values gas mileage and car price, in general, as attributes of a car. In choice-based conjoint analysis, the importance of the attributes to the respondent is inferred from their response to the choice task as opposed to being elicited directly from the respondent.

If the respondent makes a choice under the conditions of a carefully engineered and calibrated survey between hypothetical alternatives (as opposed to a choice in a real economic or healthcare market, such as deciding whether or not to use a health savings account), then the technique is referred to as a "stated preference" method (as opposed to a "revealed preference" method).

This chapter focuses on choice-based conjoint analysis in stated preference surveys. Broadly, conjoint analysis is concerned with the identification of important attributes and the measuring of their importance, and the term encompasses all aspects of that process, from the formation of focus groups of experts and potential respondents to identify attributes important to the choice at hand, to the experimental design and

validation of choice tasks and surveys, to the statistical analysis of those surveys and the interpretation of the results (Bridges, Hauber, et al. 2011). This chapter discusses statistical methods used both in the design of such surveys and in the analysis of the responses obtained.

2.2.2 When is conjoint analysis used and why?

Green and Rao (1971) described their excitement for the possibility of shaping public policy by using conjoint analysis to assess trade-offs between various attributes of potential policies. Stated preference methods can be used to ascertain the value of goods that are not on the market since they do not yet exist or are not usually assigned value in traditional economic markets, such as attributes of the environment. Because of this, conjoint analysis is especially useful for evaluating the potential of new goods (e.g., as in the development of new consumer products) and potential policies (e.g., health policies, environmental policies, and economic policies).

Conjoint analysis has been employed successfully in myriad settings. Conjoint analysis has been used to investigate people's willingness to pay to preserve culturally significant marble monuments in Washington, DC (Morey and Rossmann 2003). Morey and Rossman found that most people would be willing to pay to preserve the historical monuments, but that a significant proportion of the young and non-white would either not be willing to pay anything to preserve them or would actually pay to ensure their deterioration. Eisen-Hecht et al. (2005) used conjoint analysis to determine which attributes of potential wetland restoration policies were most appealing to North Carolina landowners. Whitty et al. (2011) used conjoint analysis to ascertain the public's preferences regarding subsidies of pharmaceutical companies. Conjoint analysis has been used to understand what factors would most influence parents to use booster seats for their children (Cunningham et al. 2011). Conjoint analysis

has also been used to ascertain patient preferences for cost, severity of side effects, and life expectancy in men with prostate cancer (Sculpher et al. 2004). By asking respondents to make choices among realistic alternatives, it is possible to obtain patients' preferences for attributes like life expectancy that they might otherwise refuse to directly place a value on. Even if patients are willing to verbalize directly their value for an attribute like life expectancy, the monetary values obtained in this way can be grossly biased with respect to their true preferences; hence, conjoint analysis becomes a useful inferential tool.

The use of conjoint analysis in health care settings has increased in recent years. Heiss et al. (2010) used conjoint analysis to analyze consumer choices of Medicare Part D prescription drug plans. Geerts et al. (2013) used conjoint analysis to understand physician and nurses' attitudes toward long-acting injectable antipsychotics. Schellings et al. (2012) used conjoint analysis in the development of quality indicators for mental health care by surveying inspectors at psychiatric institutes. Bridges et al. (2011) used conjoint analaysis to obtain preferences for schizophrenia treatments directly from patients diagnosed with schizophrenia, and Hiligsmann et al. (2013) used conjoint analysis similarly for osteoporosis treatments. Conjoint analysis takes on special usefulness in mental health care settings because there are fewer objective measures of patient improvement (e.g., laboratory tests, biomarkers) than in other patient settings; typically, the patient reports his recent experiences or is observed by a physician. Johnson and Hauber et al. (2010) use conjoint analysis to examine discrepancies between patients' and physicians' perceptions of serious adverse events associated with Crohn's disease treatments. Marshall et al. [@-marshall2010conjoint] review systematically an increase in popularity of conjoint analysis applications in health care.

Conjoint analysis is especially useful for measuring the importance of what might

otherwise be considered subjective attributes of a patient experience. Conjoint analysis is also increasingly thought of as providing representations of people's preferences that are more accurate than those gained from other kinds of surveys, since the act of making a realistic choice is what people do on a daily basis, while being asked directly the monetary loss associated with, e.g., back pain, is more alien. Conjoint analysis can be used both for hypothetical choice scenarios or applied to real choices that patients have made; historically, it has been often applied to real choices of transportation among commuters (Boyd and Mellman 1980).

2.3 The multinomial logit model

2.3.1 Methodological origins

Statistical methods used in conjoint analysis have evolved greatly since Luce and Tukey's "simultaneous conjoint measurement" (1964). In fact, Luce and Tukey's paper contained no statistical methods whatsoever, but rather a set of axioms under which an ordinal response can be transformed into interval scales for the attributes composing the alternatives. Inspired by Luce and Tukey, Green and Rao (1971) explored using various different algorithms to transform ordinal responses to interval scales for attributes. Most of these algorithms involved using a mathematical transformation (e.g., logarithm square root, etc.) on ordinal responses and then using ANOVA on the transformed responses to explain them with respect to the varying attributes. Soon thereafter, McFadden (1973) detailed an approach using a probabilistic model involving the logistic transformation to infer attribute importances from consumer choices. While techniques such as least squares with monotone transformations of the data have fallen out of favor, analyses based on the framework of McFadden's "conditional logit" (now referred to as "multinomial logit") remain popular.

2.3.2 Description

The multinomial logit (MNL) model is easily interpretable, since it actually involves a statistical model that could reasonably explain the responses. This is in contrast to using least squares, which does not automatically provide predictions of choice responses, but rather maximizes an ad hoc objective function. The MNL model is generally better at predicting responses, and especially outperforms a linear regression when the respondents choose items lexicographically as opposed to in a purely compensatory way (Green and Srinivasan 1978). A disadvantage of using the MNL model has been that it requires more computer time to fit, but that disadvantage has essentially disappeared with increasing computer speeds.

The MNL model has an intuitively appealing connection to random utility theory, which is essentially that, if the "value" of an alternative is given by the product of a vector of attribute preference coefficients and a vector of the levels of the attributes of the alternative, and if respondents pick the alternative with the highest value, then the MNL model results when the additive error term attached to the value of each alternative has a Gumbel distribution (McFadden 1973). We describe this connection below, after some notation.

2.3.3 Notation

Suppose the respondent is asked to choose one of J alternatives and that there are I attributes, which take on different levels across the alternatives. Denote the level taken by the i-th attribute for the j-th alternative by x_{ij}. Denote the parameters of the model, which in practice are unknown and typically are estimated, by β_i. The parameter β_i is interpreted as describing the extent to which respondents favor alternatives with larger values of the i-th attribute. The collection of β_i is sometimes

referred to as "partworth utilities", "partworths", "taste coefficients", or "preference coefficients" because of this interpretation. Denote the response to the choice task as follows: let $y_j = 1$ if the j-th alternative is chosen and 0 otherwise. Then the MNL model can be expressed as follows:

$$\begin{aligned} V_j &= \sum_{i=1}^{I} \beta_i x_{ij} \\ p_j &= e^{V_j} / \sum_{k=1}^{J} e^{V_k} \\ [y_1, ..., y_J] &\sim multinomial\,(p_1, ..., p_J) \end{aligned} \tag{2.1}$$

Note that $\sum_{j=1}^{J} p_j = 1$; that is, the probabilities of choosing each individual alternative sum to 1 across the alternatives. Additionally, the above model is sometimes also used in a sense where the word "attribute" is replaced by the word "characteristic", where a "characteristic" is any characteristic of the alternative in that choice situation, including any known function of the attributes, the time of day, the gender of the respondent, etc., including "interaction terms" given by the multiplicative products of any of those quantities. Categorical attributes are typical in this setting, and they are typically dummy-coded in the statistical model (Bech and Gyrd-Hansen 2005).

2.3.4 Random utility theoretic formulation

McFadden (1973) showed that the multinomial logit model (2.1) results from the following choice situation:

$$
\begin{aligned}
V_j &= \sum_{i=1}^{I} \beta_i x_{ij} \\
\epsilon_j &\sim i.i.d.\ Gumbel() \\
U_j &= V_j + \epsilon_j \\
j^* &= argmax_j U_j \\
y_j &= \begin{cases} 1 & j = j^* \\ 0 & j \neq j^* \end{cases}
\end{aligned}
\tag{2.2}
$$

Note that the Gumbel distribution has cumulative distribution function $F(x) = e^{-e^{-x}}$, and its mode is zero. In (2.2), j^* represents the alternative chosen by the respondent. The values U_j are called "utilities", and the formulation is called "random utility theoretic" because the utilities have an additive random error term. In random utility theory, respondents choose the alternative that maximizes their utility. When this model is applied across multiple respondents, the Gumbel error terms mostly represent differences in preferences among the respondents. When this model is applied to a single respondent, the Gumbel error terms represent quixotic error and misspecification of the functional form of the dependence of V_j on the attributes. Train (2003) provides more discussion of the interpretation of these error terms.

2.3.5 Estimation

Consider a scenario in which N respondents each answer a conjoint survey, which will be allowed to differ in its number of choice tasks across respondents as well as the attribute levels and number of alternatives across choice tasks, but which has the same I attributes for all choice tasks. Denote the number of choice tasks for the n-th respondent by M_n. Denote the number of alternatives for the m-th choice task for the n-th respondent by J_{mn}, and denote that choice task's attribute levels by $x_{(mn)ij}$,

where i indexes attributes and j indexes alternatives.

The MNL model is typically estimated by maximum likelihood using Newton-Raphson. Let $j^*_{(mn)}$ denote the alternative chosen by the n-th respondent for his m-th choice task, so that it encapsulates all the observed data. Then the full multinomial likelihood for the observed data is written as follows:

$$L(\beta) = \prod_n \prod_m p_{(mn)j^*_{(mn)}} \tag{2.3}$$

where $p_{(mn)j^*_{(mn)}}$ is as p_j in (2.1) except with x_{ij} replaced by $x_{(mn)ij}$ and evaluated at the alternative that was chosen, $j^*_{(mn)}$. In words, this is the product of the probabilities assigned by the model (which, of course, depend on the value of its model parameters) to the alternatives that were chosen, across all choice tasks.

The log-likelihood $l(\beta) = \log L(\beta)$ is given in (2.4):

$$l(\beta) = \sum_n \sum_m \log p_{(mn)j^*_{(mn)}} \tag{2.4}$$

The first derivative (column vector) of the log-likelihood is given by

$$\frac{\partial l}{\partial \beta} = \sum_n \sum_m \left(x_{(mn)\cdot j^*_{(mn)}} - \sum_j p_{(mn)j} x_{(mn)\cdot j} \right) \tag{2.5}$$

where $x_{(mn)\cdot j}$ is the column vector of the attribute levels for the j-th alternative of the n-th respondent's m-th choice task. The second derivative (Hessian matrix) of the log-likelihood is given by

$$\frac{\partial l^2}{\partial \beta^t \partial \beta} = \sum_n \sum_m \sum_j \left(x_{(mn)\cdot j} - \bar{x}_{(mn)} \right) p_{(mn)j} \left(x_{(mn)\cdot j} - \bar{x}_{(mn)} \right)^t \tag{2.6}$$

where $\bar{x}_{(mn)} = \sum_j x_{(mn)\cdot j} p_{(mn)j}$ and t denotes transpose.

2.3.6 Inference

In the literature and applications of conjoint analysis, the matrix inverse of (2.6), evaluated at the MLE, is often used to estimate the variance-covariance matrix of the MLE. In the comparison of design performance (e.g., as in Section 2.5), this technique gives inaccurate and misleading results, and a different estimate for the variance-covariance matrix must be used instead. Using the matrix inverse of (2.6) directly to estimate variance requires the following assumptions:

i) Respondents are sampled randomly in a representative way from the population of potential respondents.

ii) Respondents all share the same preferences β, and they all respond to conjoint choice tasks using that same value of β, and they do so according to the multinomial logit formula (2.1).

An adjustment described below gives an estimate for the variance that requires only the following assumptions:

i) Respondents are sampled randomly in a representative way from the population of potential respondents.

In particular, with this adjustment it is not required that the respondents actually respond to choice tasks using a multinomial logit model (which is simply not a tenable assumption), and it is also not required that the respondents all have the same preferences (which is usually also not a tenable assumption). The formula for the adjustment used here is described by Royall (1986).

Let $\hat{\beta}$ denote the MLE obtained from estimating the MNL model. An estimate for the variance-covariance matrix of the MLE that only requires assumption i) is given by

$$\hat{V}_{\hat{\beta}} = \left(\frac{\partial l^2}{\partial \beta^t \partial \beta}\right)^{-1} \sum_n \left(\left(\frac{\partial l_n}{\partial \beta}\right)\left(\frac{\partial l_n}{\partial \beta}\right)^t\right)\left(\frac{\partial l^2}{\partial \beta^t \partial \beta}\right)^{-1}\Bigg|_{\beta=\hat{\beta}} \tag{2.7}$$

where l_n is the log-likelihood of the n-th respondent's responses. In particular, the n-th respondent has M_n responses, and the derivative of their log-likelihood is given by the inner sum in (2.5), as follows:

$$\frac{\partial l_n}{\partial \beta} = \sum_m \left(x_{(mn)\cdot j^*_{(mn)}} - \sum_j p_{(mn)j} x_{(mn)\cdot j}\right) \tag{2.8}$$

Combining (2.6), (2.7), and (2.8), we can make correct inferences regarding the respondents' preferences. Additionally, we are enabled to compare different design strategies by inspecting the overall size and other differences in the variance-covariance matrices of the estimators that they produce.

2.3.7 A note about the scale of errors

Consider the following data-generating processes, which is slightly generalized from (2.2) by the addition of a "scale factor" $\sigma > 0$ to the random error.

$$\begin{aligned} V_j &= \sum_{i=1}^{I} \beta_i x_{ij} \\ \epsilon_j &\sim i.i.d.\ Gumbel() \\ U_j &= V_j + \sigma\epsilon_j \\ j^* &= argmax_j U_j \\ y_j &= \begin{cases} 1 & j = j^* \\ 0 & j \neq j^* \end{cases} \end{aligned} \tag{2.9}$$

It can be shown that the data-generating process resulting from (2.9) is the same as the

data-generating process from (2.2) where the taste parameters are replaced by (β_i/σ). So, when comparing data from different data sources, it is possible that a difference in estimated taste parameters is caused by a difference in the scale of the random error between the two sources. The methodology for assessing experimental design and for statistical analysis in conjoint analysis studies described in this thesis largely ignores the issue of the scale of respondent's preferences. The issue of respondent scale arises especially when comparing multinomial logit models estimated from different data sources (Swait and Louviere 1993; Hensher, Louviere, and Swait 1998). Specifically, Swait suggests testing for differences in estimated taste parameters that arise purely from differences in scale before testing for more general differences between estimated taste parameters.

2.4 Research in experimental design

2.4.1 Introduction

Research in conjoint analysis is multifaceted, with areas of research varying in importance across domains of its application. Broadly, there are two main areas of research in conjoint analysis: experimental design and statistical analysis. This section gives an overview of research areas in the experimental design of conjoint surveys. Aspects of statistical analyses are discussed in Section 3.2.

2.4.2 Experimental design

In conjoint analysis, experimental design refers to the way in which the alternatives are constructed and displayed, the selection of the number of alternatives, the number of attributes per alternative, the number of questions asked per respondent, and

similar design issues that can have a large impact on the statistical results. The chief statistical concerns in experimental design are usually the estimability of the parameters of the statistical model that is likely to be used to analyze the data and the smallness of standard errors for parameters relevant to the investigator.

2.4.3 Attribute balance and statistical efficiency

Several guidelines have been proposed for generally good experimental design such as orthogonality, level balance, minimum overlap, and utility balance (Huber and Zwerina 1996). Research in experimental design in conjoint analysis gained renewed interest in the 2000s; Louviere et al. (2011) provide an overview of this literature. An example of this more recent statistical research in experimental designs is Scarpa and Rose's (2008) statistically efficient designs: these designs are made with the intention of obtaining small standard errors specifically, for example, for the estimates of willingness-to-pay derived from conjoint analysis.

2.4.4 Adaptive designs

The design of adaptive surveys, in which new choice tasks are generated "on the fly" in order to maximally provide further information about the respondent given his previous responses, is another important area of research in experimental designs. These designs also naturally entail statistical research into estimating respondents' parameters with very small sample sizes. Adaptive survey designs have been explored by Dahan et al. (2002) and Cunningham et al. (2010).

2.4.5 Respondent fatigue

Another important issue in experimental design is how to deal with respondent fatigue, the empirically observed phenomenon that survey respondents will tend to become fatigued and answer randomly, irrationally, or in some way that is inconsistent with their true preferences (e.g., selecting only the leftmost alternative displayed) when taking too long a survey or when asked to complete too many choice tasks especially when not incentivized. Empirical research, including statistical analyses, have been used to investigate the number of questions that a respondent can answer reliably under different conditions and incentives. For example, Savage and Waldman (2008) compare respondent fatigue in surveys conducted online and through mail.

2.4.6 Attribute selection

Other aspects of the design of conjoint analysis surveys include the process by which the attributes and attribute levels are selected. Hiligsmann (2013) details a method to select attributes in health care using a structured discussion technique in a focus group of potential respondents, where he argues that "a discrete choice experiment requires a rigorous and transparent approach to select attributes [to be included in the experiment]." It is similarly important to use a structured technique of some kind to pilot the initial version of the conjoint survey in small groups of respondents.

2.4.7 Opt-out and follow-up

Choice tasks in conjoint questionnaires often have an alternative or follow-up question allowing the respondent to choose "none of the above," "I would not use this treatment," "I would not buy this product," or more generally "opt-out." For example, in a health care setting, it could be the option not to undergo any presented procedure, and

in a marketing setting, the "opt-out" can be the option to express disinterest in all products. It can be important to include such an option to ensure that parameter estimates for preference attributes are not overly optimistic. Statistical methods related to the treatment of opt-out are important for identifying the factors in the choice task that lead people to opt out (e.g., was the decision too difficult), as well as simply for allowing correct inferences in the variety of different ways "opt-out" can be allowed in the survey design. The statistical treatment of opt-out has been explored in (Ryan and Skåtun 2004).

2.5 Research problem: Judging statistical efficiency of conjoint analysis survey designs

A wide variety of methods exist for the design of experiments of a conjoint questionnaire that purport to provide best results in terms of statistical efficiency (Olsen and Meyerhoff 2017). For example, there are designs that maximize "D-efficiency", "Bayesian D-efficiency", "S-efficiency", "C-efficiency", "G-efficiency", and "A-efficiency" (Scarpa and Rose 2008). There are also orthogonal designs and "optimal" orthogonal designs. Thus, in practice, an investigator using conjoint analysis has a wide variety of methods to choose from when designing the questionnaire, all of which claim to be efficient. The natural question, "Which method should be used?" is still an open one, as there have been few comparisons of the results obtained from the various design techniques (Olsen and Meyerhoff 2017). This article's novel contribution to this research question is a method to empirically compare the statistical efficiency of two designs. That method is described in this section.

2.5.1 Motivation

Suppose that each respondent is to be asked to complete n choice tasks and that, for simplicity, each attribute has only two possible levels, each choice task has only two alternatives, and all respondents are given the same survey. Then there are $2^I \times 2^I = 4^I$ possible choice tasks, from which n are selected to create a survey. Ideally, we would like to know what the respondent's response would be for all possible choice tasks. However, it has been shown that respondents can only reliably complete up to about ten choice tasks before they become fatigued and no longer answer in a way that is consistent with their true preferences. Even for a small number of attributes, such as $I = 3$, there are 64 possible choice tasks, which is much greater than ten. Therefore, in most situations, the designer is forced to design the survey using a subset of the possible choice tasks that is significantly smaller than the set of all possible choice tasks. Supposing that n has been decided upon, then there are $(4^I)^n = 4^{In}$ possible survey designs. For example, if there are $I = 4$ attributes, which is quite typical, and $n = 10$ questions per respondent, which is also typical, then there are $4^{4 \cdot 10} = 4^{40} > 1 \times 10^{24}$ possible survey designs.

This is to say that the design of the survey is an important and non-trivial exercise. There is typically a very large number of possible designs that could be used, none of which is clearly superior to all others. Thus, the designer should formulate choice tasks that best elicit information about preferences for attributes that are of interest to the investigator, under the constraint of limiting the survey to roughly ten questions. This section concerns a method to compare two designs to determine which is better at eliciting information.

2.5.2 Design strategies

There are various strategies employed to select a design from the large number of all possible designs. The most recent of these is a wave of statistically motivated "efficient" designs. These techniques hypothesize a true model, usually an MNL model, then choose a design that gives estimators from an MNL model fit a small variance about their true value.

This formula resulted in a large swath of design techniques, using different true models and different notions of "small variance." Because the actual data generating process of a respondent is not the true model that is used to generate the design, there is no assurance that a design obtained from one of these techniques is in any way more "efficient" than any other potential design. Hence, these techniques must be empirically evaluated. In this chapter, we present a method for empirically comparing the results of different designs.

2.5.3 Comparing parameter estimates from two designs

There is no guarantee that estimators from different designs converge to the same value. Thus, the first step in empirically comparing two designs is to compare the parameter estimates yielded by them. Suppose that you have administered two designs to two independent groups of respondents, respectively. Suppose that the responses to your first design yield a vector of coefficients $\hat{\beta}_1$ with an estimated variance-covariance matrix V_1 using (2.7), and similarly $\hat{\beta}_2$ and V_2 for the second design. A reasonable test of the difference between parameters is to examine $(\hat{\beta}_2 - \hat{\beta}_1)^t (V_1 + V_2)^{-1} (\hat{\beta}_2 - \hat{\beta}_1)$, comparing it to the values taken by a chi-square distribution with $dim(\beta_1)$ degrees of freedom. If the parameters are significantly different, it is less sensible to compare the variance of the estimators, since the primary difference between the designs is that they

result in different coefficients. Even if the parameter estimates are not significantly different, it can still be the case that the variance-covariance of the estimators are significantly different.

2.5.4 Comparing the statistical efficiency of two designs

It is possible to have two designs for which the two estimators approach the same value, but one of the estimators approaches it faster. In this case, it can be said that one of the designs is more statistically efficient than the other. In conjoint analysis, it is desirable to use a design that is maximally statistically efficient; this provides better estimates of respondents' preferences, or, alternatively stated, allows fewer respondents to be interviewed while achieving the same quality of estimation results.

Suppose that there are two groups of respondents; the first group receives questionnaires with design 1, resulting in a vector of coefficients $\hat{\beta}_1$ with an estimated variance-covariance matrix V_1 and similarly $\hat{\beta}_2$ and V_2 for design 2. Ideally, we would prefer a design that results in a variance-covariance matrix, say V_1, that has $|v_{1ij}| < |v_{2ij}|$ for all i and j. However, this comparison is too gross; there are many pairs of designs for which no such inequality holds but the investigator may still vastly prefer one design over the other. For example, consider these two variance-covariance matrices:

$$V_1 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \qquad V_2 = \begin{pmatrix} 0.1 & 0 \\ 0 & 3.1 \end{pmatrix} \tag{2.10}$$

In this situation, if the investigator is interested relatively equally in the values of the first and second parameters, then the investigator would probably vastly prefer V_2 over V_1, since it tells him almost exactly the value of the first parameter, while having a variance for the second parameter that is only slightly greater. Nevertheless, V_2 does

in fact have greater variance for the second parameter than V_1 does. Because there is some ambiguity in deciding which variance-covariance matrix is "better" for the investigator, there have been many different proposed methods to place a total ordering on these variance-covariance matrices in order to find the best design. Generally, these methods assume that the investigator is interested relatively equally in the values of all the parameters. Some of these methods are described below.

2.5.4.1 D-efficiency

The D-efficiency of a variance-covariance matrix is one of the first measures of its size that was used in conjoint analysis survey design. The D-efficiency of a variance-covariance matrix V of dimension k is related to the size that a normal distribution with that variance-covariance matrix occupies in the space of the estimated parameters: larger sizes have smaller D-efficiencies. In particular, the size of the occupied space is proportional to the square root of $|det(V)|$, so D-efficiency is usually defined as $D_{\text{eff}}(V) = 1/|det(V)|^{\frac{1}{k}}$. In other words, the D-efficiency is the reciprocal of the geometric mean of the eigenvalues of the matrix.

2.5.4.2 A-efficiency

The A-efficiency of a variance-covariance matrix V of dimension k is related to the average size of the variances for the estimated parameters. In particular, A-efficiency is usually defined as $1/\frac{1}{k}\sum_{i=1}^{k} v_{ii}$. In other words, A-efficiency is the reciprocal of the average size of the variances for the estimated parameters.

2.5.4.3 G-efficiency

The G-efficiency of a variance-covariance matrix V of dimension k is related to the maximum size of the variances for the estimated parameters. In particular, G-efficiency is usually defined as $1/\max_i v_{ii}$. In other words, G-efficiency is the reciprocal of the largest of the variances for the estimated parameters.

2.5.4.4 Comparing the designs

There are myriad similar definitions of statistical efficiency. For simplicity, we will focus on D-efficiency. Suppose that there are two groups of respondents; the first group receives questionnaires with design 1, resulting in a vector of coefficients $\hat{\beta}_1$ with an estimated variance-covariance matrix V_1 and similarly $\hat{\beta}_2$ and V_2 for design 2. Suppose that $D_{\text{eff}}(V_1) > D_{\text{eff}}(V_2)$. Should we conclude that design 1 is better than design 2? Clearly we need a statistical test for the differences in the D-efficiencies. Equivalently, we can use a statistical test for a more natural quantity, the ratio of the D-efficiencies, which is more closely related to equivalent sample sizes between the two designs. The following outlines such a procedure:

1. Resample respondents, with replacement, from group 1. Fit the model to these resampled respondents and obtain the variance-covariance matrix V_1^*.
2. Resample respondents, with replacement, from group 2. Fit the model to these resampled respondents and obtain the variance-covariance matrix V_2^*.
3. Calculate and store the ratio $D_{\text{eff}}(V_1^*)/D_{\text{eff}}(V_2^*)$ in a list d.
4. Repeat steps 1, 2, and 3 to obtain a suitably dense set bootstrapped sample for the ratios of the D-efficiencies; e.g., repeat the steps 100 times so as to calculate and store 100 ratios in d.

5. Form a confidence interval for the ratio of D-efficiencies, e.g., by taking the interval from the 2.5th percentile to the 97.5th percentile of d to obtain a 95% confidence interval.

2.5.5 Discussion

Importantly, this technique assumes nothing about the actual method that respondents use to complete their choice tasks - in particular, the above method does not assume that the respondents complete choice tasks according to an MNL model, but rather analyzes the behavior of the maximum likelihood estimate for an MNL model in a statistically valid way in the presence of the unknown data-generating process provided by the respondents. This idea has been a point of confusion in the literature for some researchers, who have mistakenly assumed that differing designs must be compared in a conceptual framework in which respondents are assumed to complete choice tasks truly according to an MNL model. The only assumption in the above method is that the respondents have in fact been sampled randomly in a representative way from the population of potential respondents.

2.5.6 Usage

In one study, this technique found no statistically significant difference between the inferred MNL parameters or their standard errors between an orthogonal design and a D-efficient design (Kinter et al. 2012). However, in the discussion of that article, it was hypothesized that a larger sample size could lead to discoverable differences in the standard errors.

A larger study compared a design without attribute level overlap to one with attribute level overlap (Bridges 2013). The two designs were found to estimate the same

parameters, and the design with no overlap was found to have statistically significantly higher D-efficiency. The D-efficiency of the design with no overlap was found to be 1.21 times greater than that of the design with overlap, with confidence interval (1.15,1.26). In other words, the statistical test described in Section 2.5.4.4 found that the design without overlap was more efficient. This finding tends to support the usage of designs without attribute level overlap in similar choice situations in the future.

Chapter 3

Novel Statistical Methods in Conjoint Analysis

3.1 Overview

This chapter presents methods in statistical analysis developed to solve real-world problems in conjoint analysis. Specifically, a statistical method is presented for making correct inferences about the distribution of preferences across respondents, especially regarding the inference of the number of natural clusters of preferences. Additionally, this chapter demonstrates how inspection of individual respondents' preferences provides insight into the mixing distribution even when the parametric form that is hypothesized for it during the model fitting process is false. Techniques for visualizing the estimates of individual respondents' preferences are presented which provide greater insight into respondents' preferences than the values of inferred model parameters or goodness-of-fit statistics alone. This chapter also provides an overview of areas statistical challenges in conjoint analysis to give context to the described statistical methods.

Section 3.2 provides an overview of areas of research in the analysis of responses to conjoint analysis surveys. Section 3.3 presents a research problem in the analysis of responses to conjoint analysis surveys that was addressed using statistics, i.e., making inferences on the distribution of preferences in the population, including correctly describing the number of clusters of respondent preferences, when the number of choice tasks per respondent is small. Section 3.4 concludes with a summary of the results.

3.2 Research in statistical analysis

3.2.1 Heterogeneity of preferences across respondents

The formulation of the MNL choice probabilities as arising from an intuitively appealing and relatively simple model has allowed researchers to expand it in various directions to cover increasingly complicated applications by incorporating their research situations into the probabilistic model used. For example, if each respondent makes multiple choices, then the data is panel data, and a model that allows heterogeneity between respondents' preferences is appropriate. This situation is common to stated preference conjoint analysis, where each respondent is typically presented with between 8 and 15 choice tasks.

Conjoint analysis researchers have developed finite mixture models and normal mixture models of multinomial logits for this situation. McFadden (1976) theorized about the practical usefulness of mixture models of multinomial logits; however, computational speed and advances in model estimation techniques (e.g., expectation-maximization algorithms, simulated maximum likelihood, and Gibbs sampling) did not emerge until later to allow the fitting of these kinds of models. These methods are important techniques used currently in the estimation of aggregate preferences and individual respondents' preferences from conjoint analysis surveys.

Conjoint analysis researchers have developed other models, such as the nested logit, which occurs when the error terms in the utility structure are correlated across alternatives, such that certain alternatives are "nested" together in that they tend to be related and so both have higher utility or both have lower utility (i.e., they are similar products). The nested logit has the advantage that a closed form has been derived for it that is similar to the multinomial logit's (except with additional "nesting" coefficients which are estimable). The nested logit model has been generalized to a class of models called generalized extreme value models. Other choice models specialized to various applications are numerous.

3.2.2 Statistical methods for estimating individuals' preferences

It has always been desirable in conjoint analysis to obtain estimates of individual respondents' preferences, but it was previously thought that doing so would require asking too many questions of the respondent (Hauser and Rao 2004). Advances in computational speed and in statistical methods (e.g., finite mixture models, normal mixture models, Gibbs sampling) enabled the estimation of individual-level preferences by borrowing information across conjoint analysis respondents.

Finite mixture models were first applied to conjoint analysis data by Kamakura and Russel (1989). Allenby and Lenk (1994) were the first to use a normal mixture model with conjoint data, which they estimated using a Gibbs sampler. The same model can also be estimated via the method of maximum simulated likelihood, as detailed by Train (2003). This model has increased in popularity in recent years due to increasing speeds at which it can be estimated, its ease of interpretation, and the flexibility of the model in terms of the distributions of aggregate preferences that it can produce. There have also been numerous studies, both empirical and simulation-based, demonstrating

the capability of the model to estimate individuals' preferences even when the number of questions per respondent is low (Lenk et al. 1996; Huber and Train 2001).

3.2.3 Optimization and machine learning

Estimating a normal mixture model of multinomial logits using a Gibbs sampler or maximum simulated likelihood is faster than it once was (on the order of tens of seconds at the time of this writing). However, these methods of estimation are not fast enough for every application. For example, some adaptive conjoint questionnaries require estimation of individual-level preferences in a fraction of a second in order to allow the formulation of the most useful next questions for the respondent while he is taking the questionnaire. New methods continue to advance conjoint analysis both in speed and scope in various application areas, such as in the automatic identification of nonlinear utility structure or attribute interactions.

Among these methods are the analytic-center method designed explicitly to improve adaptive conjoint questionnaires (Evgeniou, Pontil, and Toubia 2007; Toubia, Evgeniou, and Hauser 2007; Dahan et al. 2002), a support vector machine method to estimate aggregate utility structure that is nonlinear (Evgeniou, Boussios, and Zacharia 2005), and an optimization-based method for estimating individual-level preferences that shrinks them toward their mean that is somewhat analogous to estimating the normal mixture model, but can be faster (Evgeniou, Pontil, and Toubia 2007).

3.3 Research problem: Understanding the distribution of preferences among the respondents

A wide variety of methods exist for the statistical modeling of the distribution of preferences among the respondents. In practice, researchers typically specify a parametric distribution, such as multivariate normal, triangular, or lognormal, and estimate its parameters (Train 2008). Super-parametric inference of the mixing distribution has been attempted by Fosgerau et al. (2008), Bajari et al. (2007), and Train (2008) using finite mixtures of point masses or normal distributions to approximate the true preference distribution. Typically, there is not enough information to reliably nonparametrically estimate the distribution of preferences among the respondents. However, there is often enough information to estimate certain aspects of that distribution, such as its mean, its variance-covariance, and its number of natural clusters or modes. Based on research suggesting that there is only a weak dependence of posterior[1] estimates of individual preferences on the choice of mixture model (Huber and Train 2001), I propose in this chapter performing inference on the set of inferred individual preferences from a simple model, such as multivariate normal, as an approach for understanding the true distribution of preferences in practical conjoint analysis applications. Additionally, motivated by Zhu et al. (2009), this chapter suggests regressing the inferred individual preferences on respondent characteristics as the best method to find predictors of respondent preferences.

There is a misconception that the existence of well-fitting super-parametric approaches involving mixtures with several components implies that the true distribution of preferences is comprised of several natural clusters. However, this is not the case. To better understand the extent to which preference distributions exhibit true natural

[1]Whether the model is fit using Bayesian techniques, maximum simulated likelihood, or some other method, there will always be a posterior distribution of a given respondent's MNL model coefficients implied by their responses.

clustering of preferences, this chapter presents a cluster validation technique for responses to choice experiments that is shown to reliably infer the correct number of clusters. This is in contrast to inferring a number of preference clusters using measures of model fit such as Akaike's Information Criterion (AIC) or Bayesian Information Criterion (BIC) to select a number of components for a mixture model and then concluding that the number of clusters is equal to the number of components in the mixture model, a practice that is somewhat widespread in conjoint analysis and which usually overestimates the true number of clusters. Also presented is a method to visualize the distribution of individual-level preferences, so human visual faculties can be used to observe any natural groupings or clusters that exist, instead of relying on mechanical and opaque criteria like AIC and BIC alone.

3.3.1 Choice of mixture distribution in the analysis

Consider the set of modeling techniques that model preference heterogeneity by assigning each respondent his own MNL model, and then assume that the coefficients of those MNL models come from a shared specified mixture distribution with unknown parameters (e.g., normal with unknown mean and variance, triangular with unknown location and width, finite with unknown number and location of points, etc.). Importantly, it has been found that in conjoint analysis applications, the exact parametric form assumed for the mixture distribution has very little effect on the resulting estimates of individuals' coefficients. According to Huber and Train (2001), using a normal mixture distribution achieves very similar results in terms of estimating individuals' preference coefficients to any other, possibly more complicated, mixture distribution, such as mixtures of normals, finite mixtures with many points, etc.[2] These two facts

[2]A normal mixture distribution has distribution $N(\mu, V)$ for unknown mean μ and unknown variance-covariance matrix V. A more general class of candidate distribution is provided by a mixture of normals, which for three components has distribution $\pi_1 N(\mu_1, V_1) + \pi_2 N(\mu_2, V_2) + \pi_3 N(\mu_3, V_3)$ where $\mu_1, \mu_2, \mu_3, V_1, V_2, V_3, \pi_1, \pi_2$, and π_3 are unknown and where $\pi_1, \pi_2, \pi_3 > 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$.

suggest that a) inference on the distribution of heterogeneity is *not* best achieved by making inferences using the estimated values of the parameters of any parametric mixture distribution, b) inference on the distribution of preferences is better achieved by making inferences on the set of estimated individual-level preferences among all the respondents, and c) since most methods give similar results, the analyst should use a neutral approach by assuming a normal mixture distribution when intending to perform inference using the individual-level preferences instead of imposing other parametric assumptions, such as triangularity, lognormality, or other distributions with support smaller than the real line, or overly flexible distributions such as finite mixtures of normals, which do not improve estimation of individual-level preferences in conjoint analysis applications but do impose a parametric assumption of clustered preferences.

3.3.1.1 The finite mixture model is a latent class model

The finite mixture model supposes that there is a specific, finite number, e.g. $K = 3$, of "ideal patients", and that each patient responds exactly as one of those ideal patients does. The choice patterns and proportions in the population of these ideal patients are then estimated. The model is described thoroughly in Section 3.3.5.2. In conjoint analysis, this model is often referred to as a "latent class model" (Vermunt 2014), and its usage to divide respondents into K groups depending on which ideal patient their choice patterns are closest to is referred to as "latent class segmentation" (Bhatnagar and Ghose 2004).

3.3.2 Visualizing the distribution of preferences

Conjoint analysts often perform inference on the distribution of preferences in the population, but they rarely ever visualize it before relying on an inferred parametric

form that strongly biases the result. In the following sections I discuss a technique to visualize in two dimensions the estimated distribution of preferences in the population without relying on parametric inference. In practical applications, when introduced to such visualizations, conjoint analysts have found such images illuminating, providing much greater insight to respondents' preferences than the values of inferred model parameters alone. These visualizations also help guide the modeling process, providing greater insight into features like the adequacy of the mixing distribution, the extent to which there is variation and clustering in preferences, and the general shape of such variation and clustering. For example, McFadden and Train (2000) describe a general test for the adequacy of the mixing distribution, but its power is low, and it offers no suggestion as to what the mixing distribution *should* be in the case that it is deemed inadequate. Visualization techniques help the analyst and investigator discover and describe what the mixing distribution *should* be or *is* in such a case.

3.3.3 Multidimensional scaling

Typically, conjoint questionnaires have at least four attributes, so if individual respondents are modeled as making choices according to an MNL model, their preferences are represented by a vector containing at least four coefficients. Multidimensional scaling is a set of statistical techniques that can be used to plot these preference vectors in a two-dimensional space in a way that attempts to preserve as closely as possible the distances between the preference vectors. In this manner, the major axes of variation in preferences among the respondents and any major clustering behavior can be observed in one graphic, as opposed to making pairwise scatterplots for every possible pair of coefficients and visually inferring the major patterns in the data only from them.

For example, in a conjoint analysis of incentives for participating in a genetic study,

plotting the first two principal components, as shown in Figure 3.1, revealed that the respondents formed three natural clusters, between which the biggest difference was the number of choice tasks for which the respondent chose not to participate in the study (all, some, none). It was found, for example, that people who chose not to opt out also placed greater value on receiving a free examination by a physician. Such a diagram is useful in conjunction with the results of statistical tests to infer clusters of the respondents' preferences.

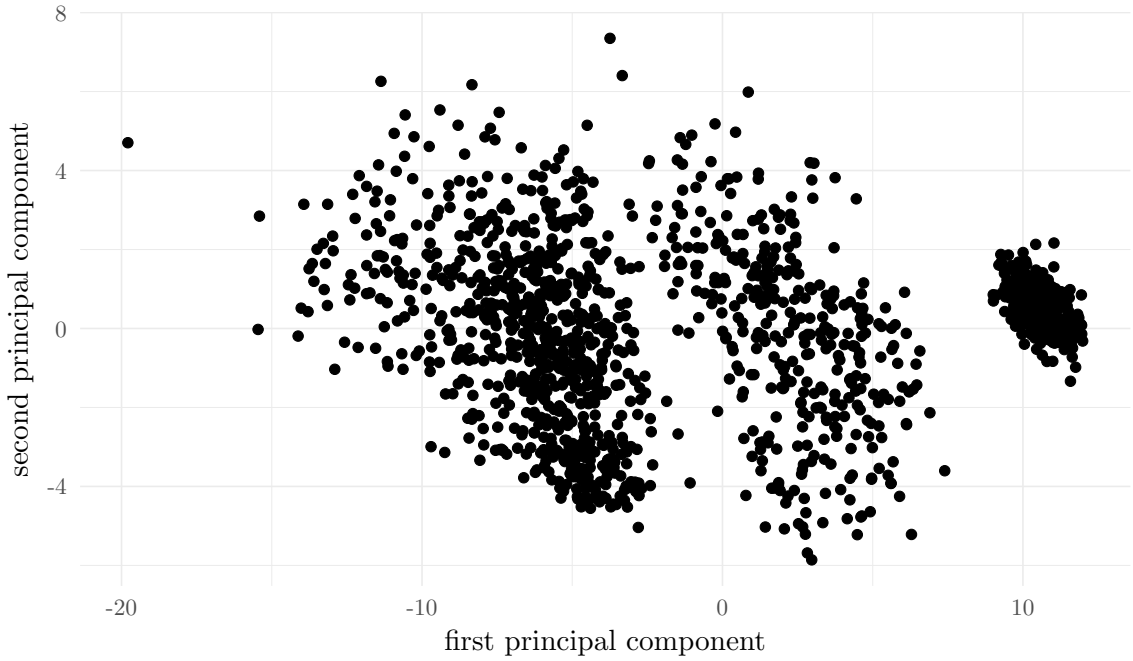

Figure 3.1: Preferences for participation in a genetic study with 12-dimensional preference vectors, scaled

3.3.4 Determining the number of clusters

Conjoint analysts often use AIC and BIC with a finite mixture model of MNLs (referred to as a "latent class" model) to determine the number of clusters when analyzing respondents' preferences. In this method, a sequence of models with different numbers of clusters as a varying hyperparameter are all fit to the data. Then, measures of model fit such as AIC and BIC are compared across the models to select the "best"

one. Unfortunately, choosing models with AIC and choosing models with BIC do not typically result in the same number of clusters. That leaves the investigator to choose heuristically his favorite number of clusters and then proceed to describe the results using that number of clusters. It is not always mentioned when the results are reported that this choice of clustering is one of many different possible clusterings, with varying number of clusters, all of which aptly describe the respondents' preferences. Here we discuss a method to infer the extent to which there is natural clustering inherent in the data. This could be considered inference on the number of clusters, as opposed to what is usually done in conjoint analysis, which is more similar to data description, where an arbitrary and non-unique cloud of points is used to describe the data.

The proposed method is as follows:

1. Fit a normal mixture model of MNLs with flexible variance-covariance matrix to the data.
2. Obtain estimates of the coefficients of each respondents' MNL model from the posterior (e.g., take posterior means).[3]
3. Cluster, e.g., using k-means, the estimates of the respondents' preference vectors.
4. Validate the clustering, e.g., using prediction strength (Tibshirani and Walther 2005).
5. Report the maximum number of clusters n that is statistically validated by the data.

After using this method, the true distribution of preferences can be understood as being composed of at least n clusters. The rest of the variation in preferences can be more accurately described by the analyst as variation within those clusters as opposed to the existence of additional clusters in different locations, which attempting to use

[3]The "posterior" here refers to the distribution of a respondent's MNL model coefficients given his or her responses and given the estimated parameters of the mixing distribution. This posterior is available whether or not a Bayesian method is used to fit the model.

AIC and BIC often results in.

3.3.5 Simulation study A: determining the number of clusters

To demonstrate the utility of this method compared to traditional AIC and BIC, I performed a simulation study where the data-generating process has four clusters of preferences. The advantages of using a normal mixture model are demonstrated. The normal mixture model (see Section 3.3.5.3) correctly identifies the true number of clusters, whereas the finite mixture model (see Section 3.3.5.2) does not. Also, the final determination of the number of clusters using the normal mixture model is greatly aided by a visualization of the preference clusters, whereas the finite mixture model provides no such visual aid, leaving the investigator essentially to guess the true number of clusters from $n \in \{4, 5, 6\}$ based on measures of model fit such as AIC and BIC.

3.3.5.1 Data-generating process

Consider a scenario in which $N = 200$ respondents each answer one conjoint survey, which differs in its attribute levels across choice tasks and respondents, but which has the same $I = 3$ attributes, number of choice tasks $M = 10$, and number of alternatives $J = 2$ for all choice tasks.

Denote the attribute level taken on by the i-th attribute of the j-th alternative for the m-th choice task for the n-th respondent by $x_{(mn)ij}$. Given m, n, i, and j, $x_{(mn)ij}$ is a single scalar real value (often simply 0 or 1 in the common case of categorical attributes).

Denote the responses to the choice tasks as follows: let $y_{(mn)j} = 1$ if the j-th alternative

is chosen by the n-th respondent for his m-th choice task and 0 otherwise.

The true model that generates the data is multinomial logit where the coefficients come from a distribution with four clusters, with equal probability. The clusters have the following centers:

$$\beta_1 = [3, -2, 1] \qquad \beta_2 = [-1, 3, -2] \qquad \beta_3 = [2, 2, 2] \qquad \beta_4 = [-2, -2, -2] \tag{3.1}$$

The rest of the data-generating process can be described as follows, where $\beta_{(n)}$ represents the preference vector assigned to the n-th respondent:

$$\begin{aligned}
\bar{\beta}_{(n)} &\sim i.i.d.\ uniform(\{\beta_1, \beta_2, \beta_3, \beta_4\}) \\
\beta_{(n)} &= \bar{\beta}_{(n)} + \epsilon_{(n)} \\
\epsilon_{(n)} &\sim i.i.d.\ N(0, Id) \\
x_{(mn)ij} &\sim i.i.d.\ N(0, 1) \\
V_{(mn)j} &= \sum_{i=1}^{I} \beta_{(n)i} x_{(mn)ij} \\
p_{(mn)j} &= e^{V_{(mn)j}} / \sum_{k=1}^{J} e^{V_{(mn)k}} \\
\left[y_{(mn)1}, ..., y_{(mn)J}\right] &\sim multinomial\left(p_{(mn)1}, ..., p_{(mn)J}\right)
\end{aligned} \tag{3.2}$$

Only the responses $y_{(mn)j}$ and the choice tasks $x_{(mn)ij}$ are observable to the conjoint analyst. In words, the data-generating process is that there are four preference centers, and each respondent has preferences given by a small, normally distributed perturbation about one of those four preference centers, chosen at random. In particular, the distribution across the respondents is continuous, due to the nature of the perturbation. One data set was generated from this process and used for the

analysis in the following sections.

3.3.5.2 Finite mixture of multinomial logits

The data was first analyzed using a finite mixture of multinomial logits with unknown number of components. The model was fit by maximum likelihood using an expectation-maximization algorithm. In particular, the model is as follows, where D is a distribution on the (three-dimensional) coefficients of MNL models defined as a mixture of K point distributions with probabilities π_k of taking values β_k, respectively:

$$\begin{aligned} p(D = d) &= \sum_{k=1}^{K} \pi_k I(d = \beta_k) \; ; \; \sum_{k=1}^{K} \pi_k = 1 \; ; \; 0 < \pi_k < 1 \\ \beta_{(n)} &\sim i.i.d.\ D \\ V_{(mn)j} &= \sum_{i=1}^{I} \beta_{(n)i} x_{(mn)ij} \\ p_{(mn)j} &= e^{V_{(mn)j}} / \sum_{k=1}^{J} e^{V_{(mn)k}} \\ \left[y_{(mn)1}, ..., y_{(mn)J}\right] &\sim multinomial\left(p_{(mn)1}, ..., p_{(mn)J}\right) \end{aligned} \tag{3.3}$$

The parameters of the model are the probabilities π_k and the locations of the points, β_k, of which there are K each, where K is the number of components. Note that in the first line of (3.3), I represents an "indicator function," not the number of attributes. Note that D is estimated through the estimation of $\{\pi_k\}_k$ and $\{\beta_k\}_k$. In this model, D is the distribution of preferences across respondents, and unlike the distribution of preferences across respondents in the data-generating process, it is discrete, not continuous. Hence, the D estimated from this model cannot equal the true distribution of preferences across respondents of the data-generating process.

The number of clusters was inferred as follows: Individual model fits were performed for $K = 1, 2, 3, 4, 5, 6$ components. Then, the AIC, BIC, and Integrated Likelihood

Criterion (ICL) (Biernacki, Celeux, and Govaert 2000) were obtained for each number of components. The number of components in the model with the smallest model fit statistic was inferred to be the number of clusters. In this example, even though there were four clusters, all three meaures of model fit were minimized for $K = 5$ components. Figure 3.2 plots the three measures of model fit for each number of components used. From the plot, one can see that, no matter which measure of model fit is preferred, the number of clusters inferred is incorrect, and there is little difference in the measures of model fit among $K = 4, 5, 6$ components, so the analyst is left essentially to guess the true number of clusters when using this technique, with little guidance from the data.

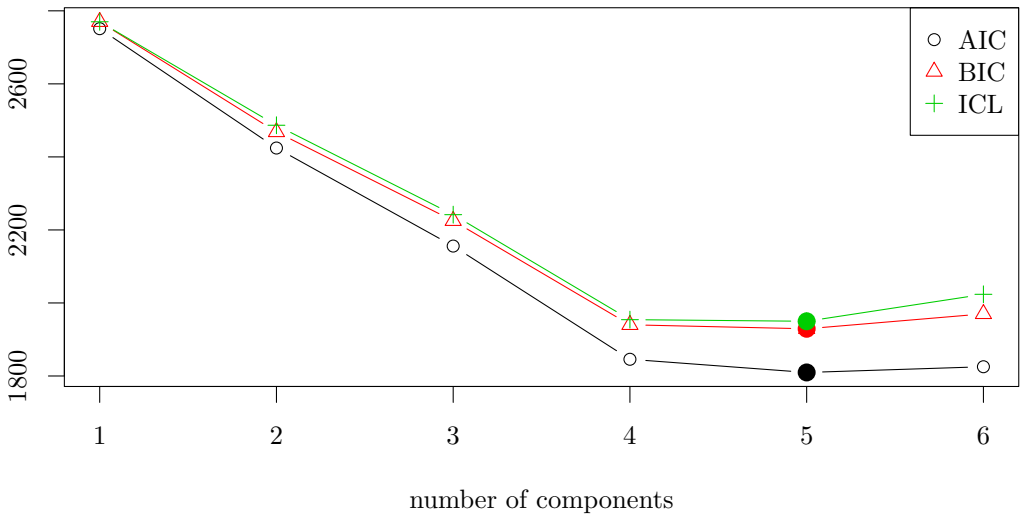

Figure 3.2: AIC, BIC, and ICL by number of components in finite mixtures of multinomial logits

3.3.5.3 Normal mixture of multinomial logits

The data was then analyzed using a normal mixture of multinomial logits. The model was fit using a Gibbs sampler[4] (alternatively, the model can be fit using

[4]The priors used were noninformative. Specifically, the priors were $\mu \sim N(0, 100Id)$ and $V \sim$ inverse Wishart with two degrees of freedom and scale matrix $0.3Id$. In the Gibbs sampling, there were 4,000 iterations, of which the first 2,000 were treated as "burn-in" iterations and discarded.

maximum simulated likelihood). In particular, the model is as follows, where $N(\mu, V)$ is a multivariate normal distribution on the (three-dimensional) coefficients of MNL models with unknown mean and variance-covariance matrix.

$$\begin{aligned} \beta_{(n)} &\sim i.i.d.\ N(\mu, V) \\ V_{(mn)j} &= \sum_{i=1}^{I} \beta_{(n)i} x_{(mn)ij} \\ p_{(mn)j} &= e^{V_{(mn)j}} / \sum_{k=1}^{J} e^{V_{(mn)k}} \\ \left[y_{(mn)1}, ..., y_{(mn)J}\right] &\sim multinomial\left(p_{(mn)1}, ..., p_{(mn)J}\right) \end{aligned} \tag{3.4}$$

The parameters of the model are the mean μ and variance-covariance matrix V of the distribution of preferences (V is conceptually unrelated to $V_{(mn)j}$). In this model, the distribution of preferences across respondents is estimated through the estimation of μ and V. In particular, the model's distribution of preferences across respnodents is unimodal and so cannot equal the true distribution of preferences across respondents of the data-generating process, which has multiple modes.

The number of clusters was inferred as follows: Extracted from the model fit are estimates of each respondents' preferences, $\hat{\beta}_{(n)}$. These estimates come "for free" if Monte Carlo methods are used to fit the model, by taking, for example, the posterior mean of each individual respondent's MNL coefficients. If other methods are used to fit the normal mixture, the individual respondents' MNL coefficients can still be estimated using the posterior of their coefficients, given the responses chosen and the estimates for μ and V.

Because estimates of each respondents' preferences $\hat{\beta}_{(n)}$ are available, it is straightforward to perform more exploratory data analysis with them. For example, the respondents' preferences can be plotted. Since the number of attributes is so small, it is useful to look both at the pairwise scatterplots of the coefficients and at a sum-

mary plot with coordinates from multidimensional scaling. Figure 3.3 shows pairwise scatterplots of the estimated individual respondents' preference coefficients. This plot provides a very helpful visual aid in determining the actual number of clusters. Figure 3.4 provides a similarly useful visual aid, the utility of which increases with the number of attributes involved in the choice tasks.

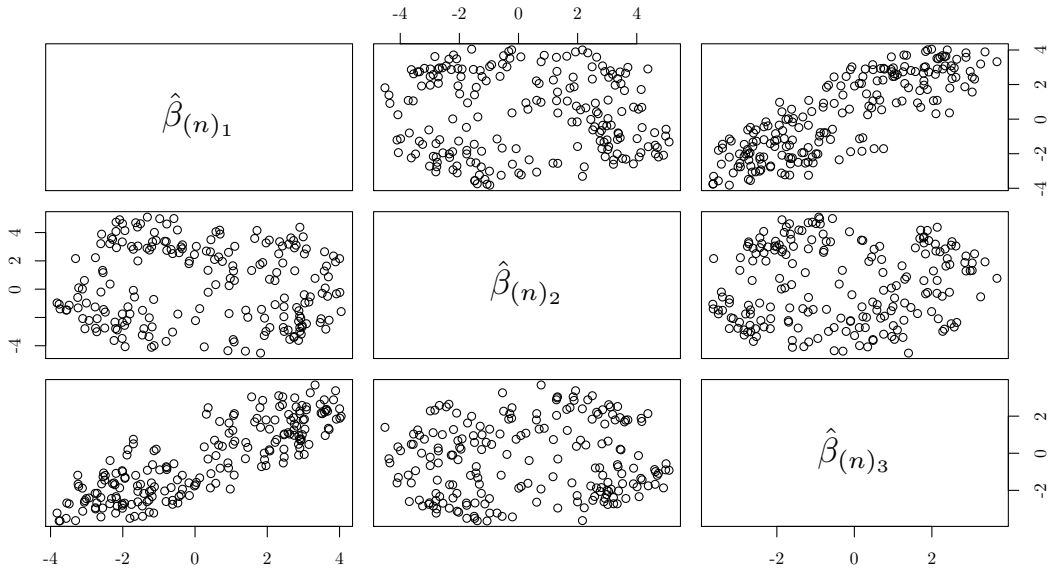

Figure 3.3: Pairwise scatterplots of individual respondents' preference coefficients

The respondents' preference vectors can be clustered using k-means. Using prediction strength to validate the number of clusters shows that the data clearly supports four natural clusters and does not support five. Figure 3.5 shows the prediction strength by number of clusters.

Note that prediction strength is a measure of the reproducibility of the clustering under subsampling. Essentially, prediction strength is computed as follows: Divide the sample equally and randomly into a "test" set and a "training" set. Cluster both sets separately. For each "test" cluster, compute the proportion of observation pairs in that cluster that are also assigned to the same cluster by the "training" set. Take the minimum of this proportion across all "test" clusters. More details can be found in (Tibshirani and Walther 2005). Tibshirani and Walther suggest that prediction

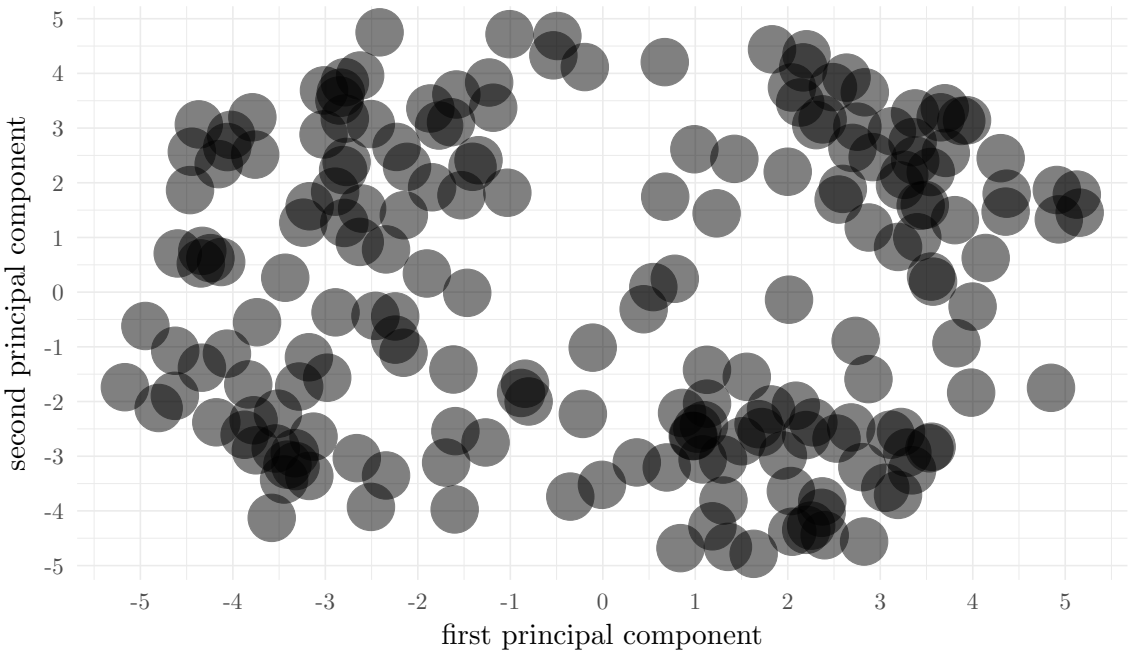

Figure 3.4: Individual respondents' preferences coefficients, after multidimensional scaling

strength greater than 0.8 is indicative of clustering.

3.3.6 Simulation study B: determining the number of clusters when there is only one underlying cluster

This simulation study compares the novel method to infer the number of clusters presented in this Chapter to latent class segmentation using AIC and BIC to infer the number of clusters in the case where there is no real clustering (in other words, there is only one natural cluster) in the data-generating process.

The advantages of using a normal mixture model with posthoc clustering are demonstrated. The normal mixture model (see Section 3.3.5.3) correctly identifies the true number of clusters (which is 1), whereas the finite mixture model (see Section 3.3.5.2) does not. Also, the final determination of the number of clusters using the normal mixture model is greatly aided by a visualization of the preference clusters, whereas

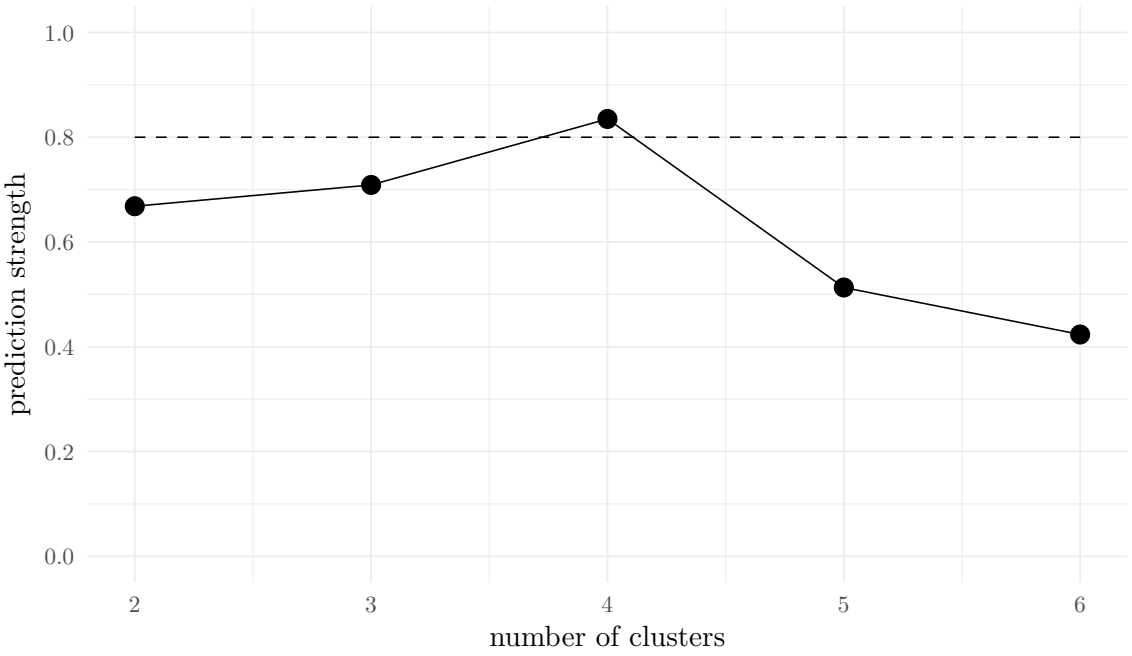

Figure 3.5: Prediction strength for clusters of respondents' preferences

the finite mixture model provides no such visual aid, leaving the investigator essentially to guess the true number of clusters from $n \in \{4, 5, 6\}$ based on measures of model fit such as AIC and BIC.

3.3.6.1 Data-generating process

Consider a scenario in which $N = 200$ respondents each answer one conjoint survey, which differs in its attribute levels across choice tasks and respondents, but which has the same $I = 2$ attributes, number of choice tasks $M = 10$, and number of alternatives $J = 2$ for all choice tasks.

Denote the attribute level taken on by the i-th attribute of the j-th alternative for the m-th choice task for the n-th respondent by $x_{(mn)ij}$. Given m, n, i, and j, $x_{(mn)ij}$ is a single scalar real value (often simply 0 or 1 in the common case of categorical attributes).

Denote the responses to the choice tasks as follows: let $y_{(mn)j} = 1$ if the j-th alternative

is chosen by the n-th respondent for his m-th choice task and 0 otherwise.

The true model that generates the data is multinomial logit where the coefficients come from a distribution with only one natural cluster. Specifically, the distribution used was the uniform distribution on a circle of radius 3. The data-generating process is described in (3.5):

$$\begin{aligned}
\beta_{(n)} &\sim i.i.d.\ uniform(\text{circle of radius 3}) \\
x_{(mn)ij} &\sim i.i.d.\ N(0,1) \\
V_{(mn)j} &= \sum_{i=1}^{I} \beta_{(n)i} x_{(mn)ij} \\
p_{(mn)j} &= e^{V_{(mn)j}} / \sum_{k=1}^{J} e^{V_{(mn)k}} \\
\left[y_{(mn)1}, ..., y_{(mn)J}\right] &\sim multinomial\left(p_{(mn)1}, ..., p_{(mn)J}\right)
\end{aligned} \tag{3.5}$$

3.3.6.2 Finite mixture of multinomial logits

The data was first analyzed using a finite mixture of multinomial logits with unknown number of components. The number of clusters was inferred as follows: Individual model fits were performed for $K = 1, 2, 3, 4, 5, 6$ components. Then, the AIC, BIC, and ICL were obtained for each number of components. The number of components in the model with the smallest model fit statistic was inferred to be the number of clusters. In this example, even though there was only one cluster, all three meaures of model fit were minimized for more than one component. Specifically, minimizing AIC resulted in 6 components, BIC in 5 components, and ICL in 4 components.Figure 3.6 plots the three measures of model fit for each number of components used. From the plot, one can see that, no matter which measure of model fit is preferred, the number of clusters inferred is incorrect, and there is little difference in the measures of model fit among $K = 4, 5, 6$ components, so the analyst is left essentially to guess the true

number of clusters when using this technique, with little guidance from the data.

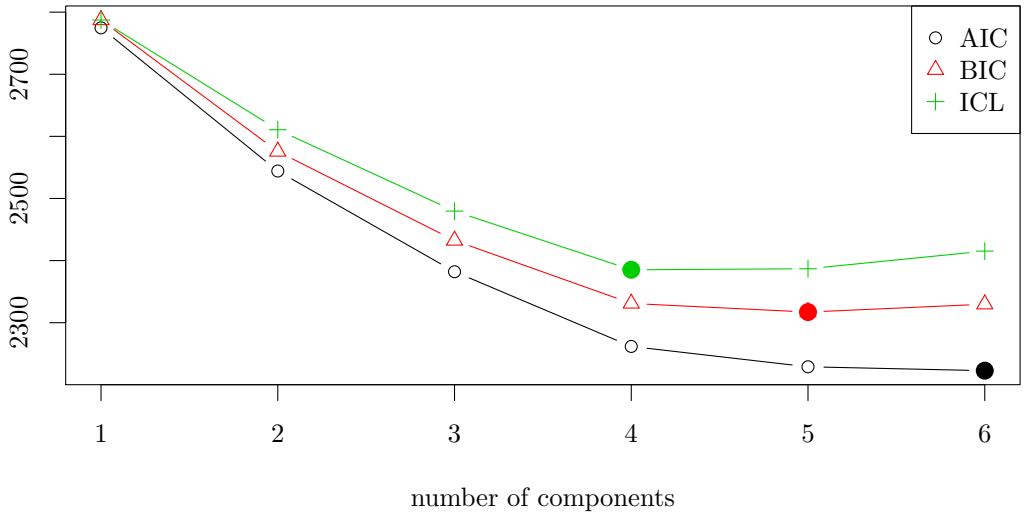

Figure 3.6: AIC, BIC, and ICL by number of components in finite mixtures of multinomial logits

3.3.6.3 Normal mixture of multinomial logits

The data was then analyzed using a normal mixture of multinomial logits. The model was fit using a Gibbs sampler[5] (alternatively, the model can be fit using maximum simulated likelihood). The number of clusters was inferred as follows: Extracted from the model fit are estimates of each respondents' preferences, $\hat{\beta}_{(n)}$. The respondents' estimated preferences were then plotted. Since there are only two attributes, there is no reason to use multidimensional scaling and the attributes can be plotted on a two-dimensional plot. Figure 3.7 shows a plot of the respondents' estimated preference coefficients. This plot provides a very helpful visual aid in determining the actual number of clusters.

The respondents' preference vectors can be clustered using k-means. Using prediction strength to validate the number of clusters shows that the data supports only one

[5]The priors used were noninformative. Specifically, the priors were $\mu \sim N(0, 100Id)$ and $V \sim$ inverse Wishart with two degrees of freedom and scale matrix $0.3Id$. In the Gibbs sampling, there were 4,000 iterations, of which the first 2,000 were treated as "burn-in" iterations and discarded.

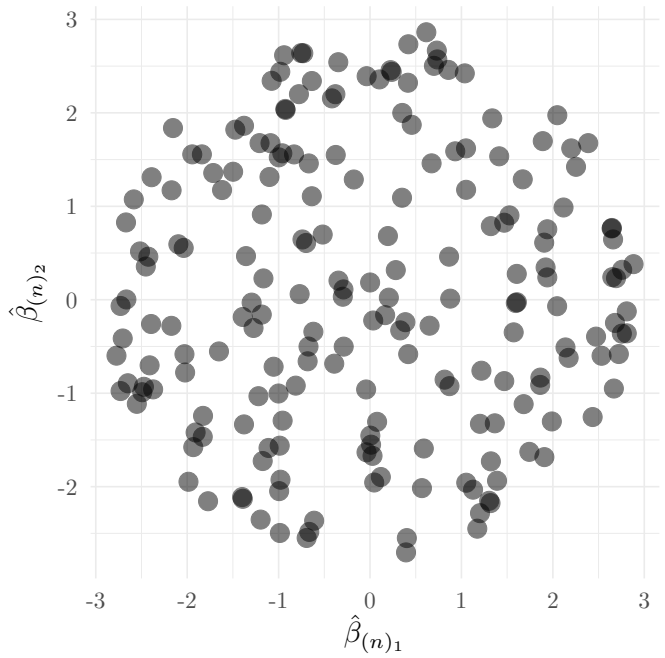

Figure 3.7: Individual respondents' estimated preferences coefficients

natural cluster and does not support four, five, or six. Figure 3.8 shows the prediction strength by number of clusters.

3.3.6.4 Discussion of simulation study B

This example demonstrates that, as long as there is any taste heterogeneity, latent class segmentation will find more than one segment, even if there is no natural clustering. In other words, latent class segmentation conflates the concepts of taste heterogeneity and of clustering of tastes. On the other hand, using prediction strength only finds more than one cluster when there is taste heterogeneity *and* there are distinct clusters of different tastes. The methods presented in this Chapter provide data visualization techniques to aid the investigator in selecting a parametric model to describe any taste variation that exists within each cluster, such as multivariate normal, should this be desired.

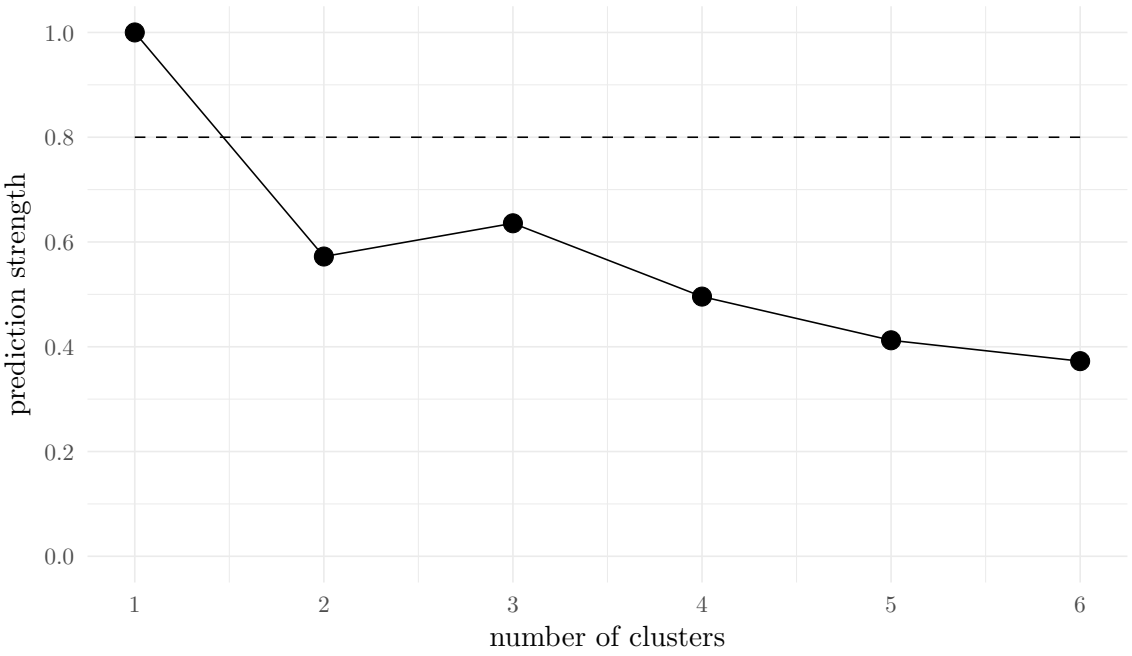

Figure 3.8: Prediction strength for clusters of respondents' preferences

3.3.7 Real-world example: a study of incentives to participate in a hypothetical genetic study

To provide an example of the usage of the methods in this chapter on a real data set, we apply them to the results of a discrete choice experiment that was embedded in a survey of 1,524 respondents. Each respondent was presented with nine choice tasks. The attributes that varied between the alternatives of the choice tasks are presented in the following table. The goal of the study was to assess the impact of different incentives on respondents' desire to participate (or not) in a genetic study.

Table 3.1: A description of the variables involved in the study

Description	Name	0	1	2
Return of individual research results	ind	none	few	all

Description	Name	0	1	2
Researchers with access to study data	res	US only	US and foreign	US and US industry
Length of participation in the study	leng	5 years	10 years	15 years
Compensation for the participant	comp	$0	$50	$100
Method of ascertaining health status	medical records		survey	physical exam
Additional activities in the study	addact	diet journal	fitness test	home visit

The choice tasks each had three alternatives. For each choice task, two of those alternatives were hypothetical genetic studies populated with values for each of the six different incentive attributes. The third alternative was always the option not to participate in either of the two hypothetical genetic studies.

The third alternative was treated in the statistical analysis by adding another variable called "optout" which takes the value 1 for that alternative and 0 for the other two alternatives. The third alternative always took on the value 0 for all of the other six incentive attributes.

Each of the six attributes was treated as categorical and was "dummy coded" with 0 as the "reference level" (Bech and Gyrd-Hansen 2005). Then, a "main effects" multinomial logit model was fit to the data. The results of the model fit are below. In the results table, "comp1" indicates the extent to which $50 is preferred over $0, and "comp2" indicates the extent to which $100 is preferred over $0, and similarly for the other attributes.

Table 3.2: Multinomial logit model fit on 1,524 respondents

Variable	Estimate	Standard error	t-value	p-value
ind1	0.42	0.03	12.9	$< 1 \times 10^{-16}$
ind2	0.55	0.03	17.4	$< 1 \times 10^{-16}$
res1	-0.16	0.03	-5.1	3×10^{-7}
res2	-0.11	0.03	-3.7	2×10^{-4}
leng1	-0.12	0.03	-6.4	1×10^{-10}
leng2	-0.43	0.03	-13.7	$< 1 \times 10^{-16}$
comp1	0.66	0.03	20.0	$< 1 \times 10^{-16}$
comp2	0.99	0.03	30.3	$< 1 \times 10^{-16}$
medical1	0.01	0.03	0.1	0.88
medical2	0.17	0.03	5.3	1×10^{-7}
addact1	0.12	0.03	3.9	9×10^{-5}
addact2	0.04	0.03	1.3	0.21
optout	0.82	0.05	16.3	$< 1 \times 10^{-16}$

The interpretation of each of the coefficients follows.

- (ind): On average, respondents favored studies from which they would receive access to *some* or *all* of the health information that could be gleaned from their genetic data. This information held, on average, significant value to respondents.
- (res): On average, respondents preferred that as few researchers had access to their data as possible. US industry researchers were disliked less than foreign academic researchers. On average, the type of researchers receiving access to the data had relatively little impact on the respondents' decisions compared to

the other incentive attributes.

- (leng): On average, respondents disliked longer studies. Respondents significantly disliked the notion of participating in a study that would take 15 years to complete.
- (comp): On average, respondents favored studies with larger compensation for the respondents. Respondents significantly valued the $50 compensation and even more significantly valued the $100 compensation. The difference in the model between $100 and $0 compensation was the greatest among all the incentive attributes in the study.
- (medical): On average, respondents were indifferent between having their health data accessed through their medical records or through a survey they completed. On average, respondents preferred a free physical exam to those options. On average, this attribute had relatively little impact on the respondents' decisions compared to the other incentive attributes.
- (addact): On average, respondents were largely indifferent between keeping a diet journal, undergoing a fitness test, or having a home visit, while slightly preferring the fitness test to the other options. On average, this attribute had relatively little impact on the respondents' decisions compared to the other incentive attributes.
- (optout): On average, respondents did exercise the option to opt out of participating in either of the presented hypothetical genetic studies. On average, the presence of favorable study conditions, such as a shorter study length, larger compensation and the return of individual research results, could sway a respondent to participate.

3.3.7.1 Finite mixture of multinomial logits

The data was first analyzed using a finite mixture of multinomial logits with unknown number of components. The number of clusters was inferred as follows: Individual model fits were performed for $K = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ components. Then, the AIC, BIC, and ICL were obtained for each number of components. The number of components in the model with the smallest model fit statistic was inferred to be the number of clusters. In this example, AIC, BIC, and ICL were minimized for different numbers of components. Specifically, minimizing AIC resulted in 11 components, BIC in 6 components, and ICL in 4 components. Figure 3.9 plots the three measures of model fit for each number of components used. As AIC, BIC, and ICL disagree, the analyst is left essentially to guess the true number of clusters when using this technique. These opaque measures give little insight to the data, in contrast to the methods proposed in this thesis, whose application is discused in the next section.

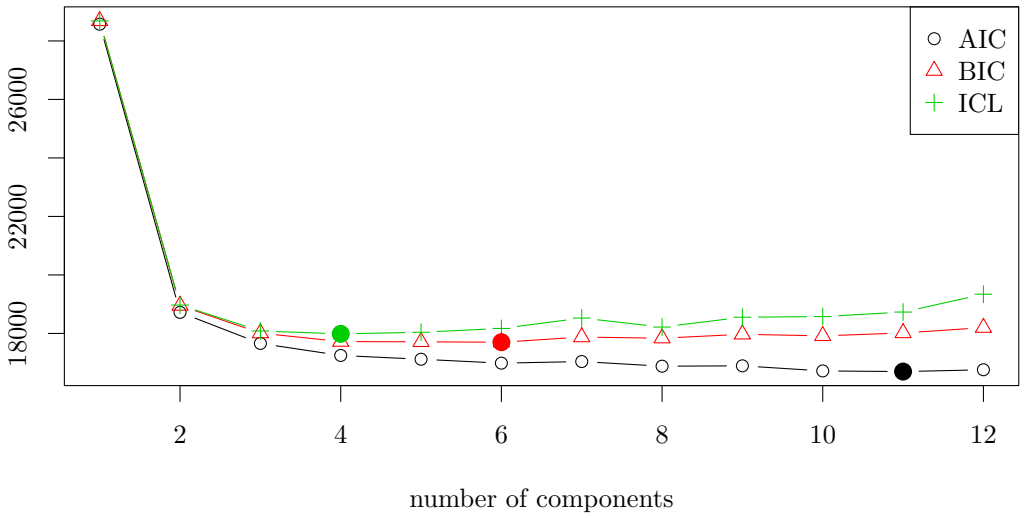

Figure 3.9: AIC, BIC, and ICL by number of components in finite mixtures of multinomial logits

3.3.7.2 Normal mixture of multinomial logits

The data was then analyzed using a normal mixture of multinomial logits. The model was fit using a Gibbs sampler[6] (alternatively, the model can be fit using maximum simulated likelihood). The number of clusters was inferred as follows: Extracted from the model fit are estimates of each respondents' preferences, $\hat{\beta}_{(n)}$. The respondents' estimated preferences were then plotted. Figure 3.10 shows a plot of the respondents' estimated preference coefficients, after multidimensional scaling. This plot provides a very helpful visual aid in determining the actual number of clusters. We immediately see that there are three distinct natural clusters of preferences. This is unlike the approach in Section 3.3.7.1, where no obvious conclusion can be reached using AIC, BIC, and ICL as to the number of clusters. The plot of prediction strength in Figure 3.11 agrees with our visual assessment that the number of natural clusters of preferences is 3. Figure 3.12 plots the same multidimensionally scaled coordinates as Figure 3.10, except with added labels for the clusters.

Table 3.3: Mean multinomial logit coefficients by cluster
on 1,524 respondents

Variable	Cluster 1	Cluster 2	Cluster 3
ind1	0.90	1.15	0.47
ind2	1.32	1.36	1.05
res1	-0.93	-0.26	-1.53
res2	-0.79	-0.10	-1.51
leng1	-0.55	-0.38	-0.69
leng2	-1.36	-1.07	-1.55

[6]The priors used were noninformative. Specifically, the priors were $\mu \sim N(0, 100Id)$ and $V \sim$ inverse Wishart with two degrees of freedom and scale matrix $0.3Id$. In the Gibbs sampling, there were 4,000 iterations, of which the first 2,000 were treated as "burn-in" iterations and discarded.

Variable	Cluster 1	Cluster 2	Cluster 3
comp1	1.73	1.64	1.52
comp2	2.76	2.64	2.33
medical1	-0.34	0.20	-0.90
medical2	-0.06	0.56	-0.67
addact1	0.25	0.44	0.03
addact2	-0.15	0.18	-0.48
optout	1.56	-8.00	10.58
cluster size	361 (24%)	768 (50%)	395 (26%)

Let us examine the kind of conclusions that can be reached by examining the three respective means of the three clusters of preferences. Table 3.3 presents the mean preference coeffcients for each cluster. The most noticeable difference between the respondents in the three clusters is the way they responded to the "optout" attribute, the alternative of not participating in either of the presented genetic studies. Respondents in cluster 3 had a strong preference for not participating in the study; in fact, those respondents chose the alternative of not participating on all nine of their choice tasks. Essentially, the inference here is that about 26% of the population would refuse to participate in a genetic study regardless of the incentives in the range considered in the study. This kind of inference was not obtainable through an aggregate MNL analysis. About 50% of respondents fell into cluster 2. The respondents in this cluster always chose to participate in one of the two hypothetical genetic studies presented to them in their choice tasks. The untempered inference from this is that about 50% of the population would be eager to participate in a genetic study regardless of the incentives offered to them as long as they were in the range considered in the study. About 24% of respondents fell into cluster 1. The respondents in this

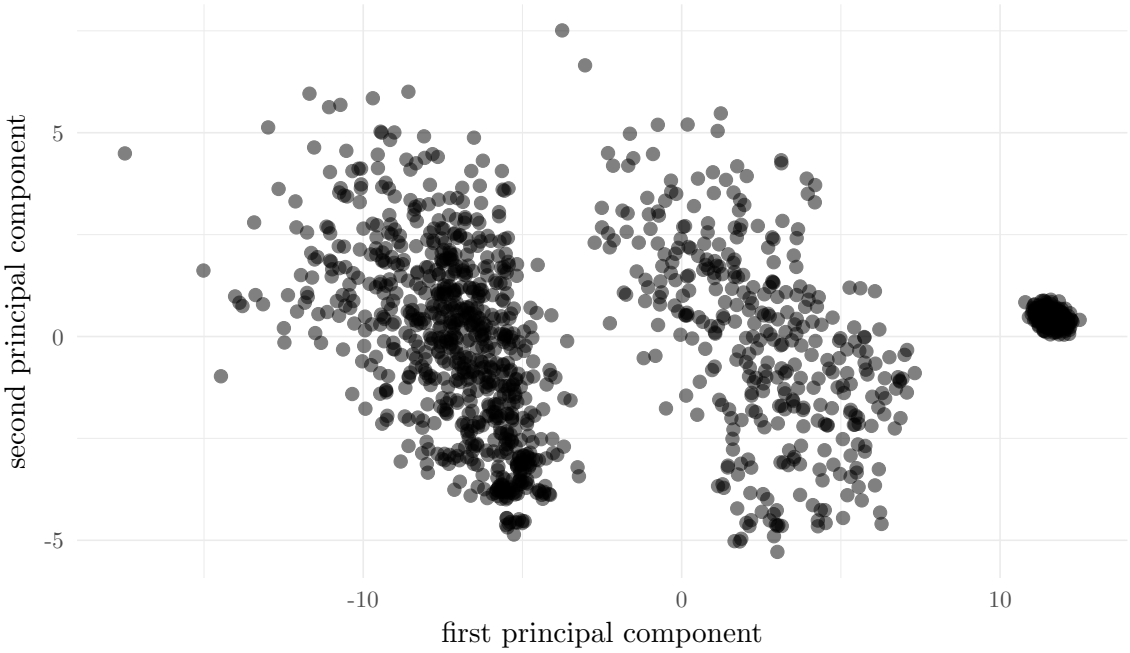

Figure 3.10: Individual respondents' preferences coefficients, after multidimensional scaling

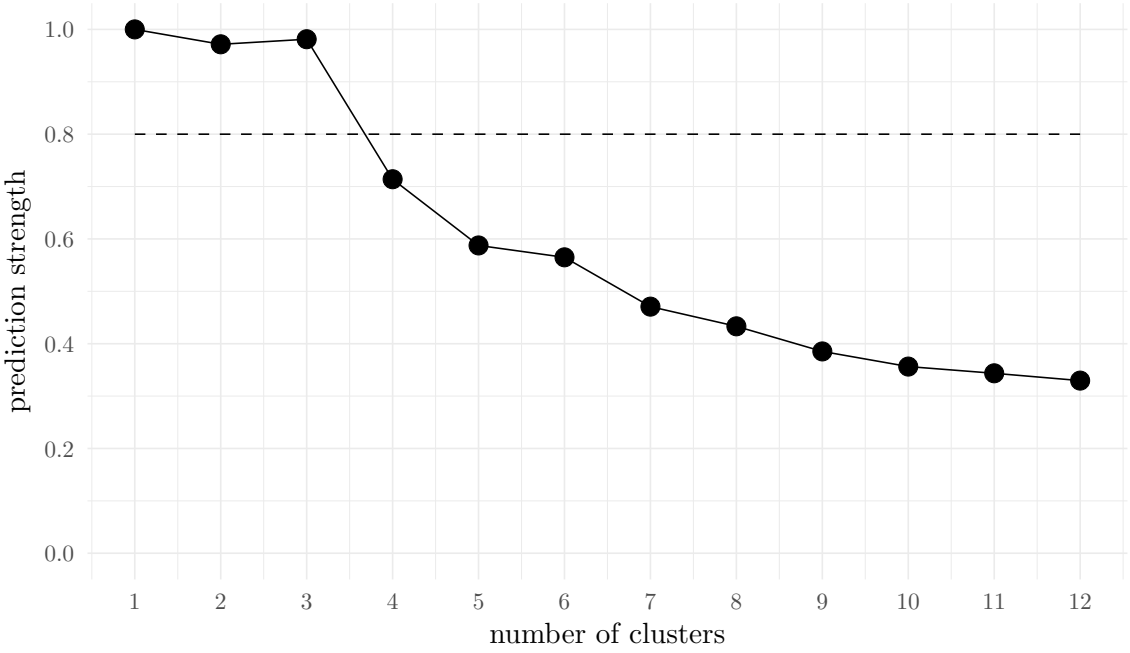

Figure 3.11: Prediction strength for clusters of respondents' preferences

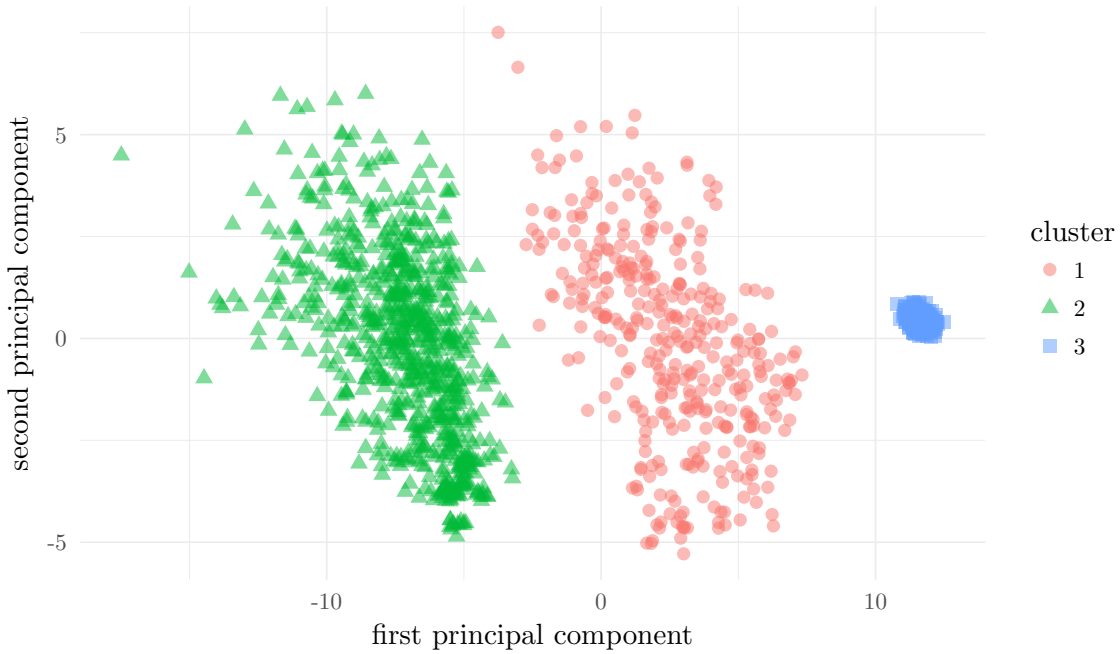

Figure 3.12: Individual respondents' preferences coefficients, after multidimensional scaling

cluster decided not to participate in either genetic study in some of their choice tasks, while choosing to participate in one of the two genetic studies presented in other choice tasks. Analysis of these respondents allows one to infer which incentives might encourage someone to participate in a study who otherwise would not participate. The analyses of clusters 1 and 2 both allow one to infer importances of the incentives as considered relative to one another, but only cluster 1 allows one to, say, place a monetary value on study participation. The coefficients besides "optout" of clusters 1 and 2 are similar, with biggest difference being the "res" attribute. Respondents in cluster 2 were relatively insensitive to changes in this attribute, whereas respondents in cluster 1 reacted negatively to researchers outside US academics' having access to their genetic data. It could be that one of the reasons the respondents in cluster 1 opted out more frequently was because of privacy concerns. Respondents in cluster 2 also appeared to actually gain utility by undergoing a physical exam, whereas a physical exam presented a disutility to respondents in cluster 1. It could be the case

that respondents in cluster 2 are more altruistic and trusting of medical institutions, hence their desire to participate, their eagerness for physical exams, and their lack of privacy concerns. Respondents in cluster 1 also reacted more negatively to longer study periods than respondents in cluster 2. Both respondents in cluster 1 and cluster 2 placed a high value on their compensation.

Further analysis of cluster 1 also shows that respondents who placed higher values on compensation also were more likely to opt out. This kind of inference is obtained from the correlation of the preference attributes for "comp1" and "comp2" with "optout" for the respondents in cluster 1. This kind of inference is not obtainable from an aggregate MNL analysis or a latent class segmentation. This is inference on the taste heterogeneity within cluster 1. Respondents with high "comp1," "comp2," and "optout" who participate in a genetic study are primarily focused on their monetary gain. Since there appears to be a large pool of respondents who are more altruistic and less financially motivated, it may be wise not to target this group by offering large compensation.

3.3.7.3 Discussion of real world example

In addition to providing more insight into the data-generating process, the methods of Section 3.3.7.2 can be used to analyze the data much more quickly than the methods of Section 3.3.7.1. Fitting the finite mixture models for each of $1 \leq K \leq 12$ took over 12 hours on the author's personal computer, whereas fitting the normal mixture model of 3.3.7.2, obtaining individual preference estimates, and validating that clustering with prediction strength took less than 5 minutes.

3.3.8 Discussion

Besides clustering, the estimates of the respondents' preferences can be regressed on demographic variables or other covariates of interest with practically any regression technique. This allows for rapid exploratory analysis and hypothesis generation that fitting simple models for component membership within a finite mixture model does not provide.

3.4 Conclusion

There are myriad techniques for design of experiments for conjoint analysis surveys. However, there is a paucity of comparative studies demonstrating the relative superiority of these different design techniques, despite the practical importance of such studies. This chapter details a statistical test to compare empirically the statistical efficiency of different design techniques. This method has been successfully used to show that a design without attribute overlap outperforms a design with attribute overlap without biasing the estimates of respondent preferences (Bridges 2013). This method uses resampling of model-robust standard errors to create a confidence interval for the relative statistical efficiency of two designs without making assumptions about the data-generating process.

Responses to conjoint analysis surveys are often analyzed using a mixture distribution of multinomial logit models to model preference heterogeneity. This chapter details statistical methods for making correct inferences about the distribution of preferences across respondents, especially regarding the inference of the number of natural clusters of preferences, if any. This method is contrasted with other methods that are commonly used. Importantly, inspection of individual respondents' preferences allows description of the mixing distribution that is more general than any parametric form that is

hypothesized for it during the model fitting process. Additionally, techniques for visualizing the estimates of individual respondents' preferences are discussed, and it is noted that estimates of individual respondents' preferences can be further regressed on other characteristics of the respondents to find predictors of respondent preferences. Overall, this chapter presents a methodology for analyzing data from discrete choice experiments that can replace or be combined with commonly used methods in order to gain a more comprehensive understanding of the data-generating process in a fraction of the time.

Chapter 4

Padé Approximation of the Profile Likelihood

4.1 Overview

In this chapter, a method is presented for obtaining confidence intervals when the sample size is small or there are many nuisance parameters. This method involves the Padé approximation of the profile likelihood function. In particular, this method uses derivatives of the profile likelihood function at its maximum in order to obtain an approximation, and so in that way it is conceptually similar to methods discussed in (Viveros and Sprott 1987) and (DiCiccio and Monti 2001).

4.2 Introduction

The goal of this chapter is to describe a way in which the profile likelihood can be approximated in a computationally inexpensive way for values of the parameter of interest near the maximum likelihood estimate (MLE). The approximation involves no

additional numerical maximizations of the the full likelihood after the MLE is found. Instead, the approximation relies on derivatives of the full likelihood at the MLE. Using this technique, one can compute an approximation to the profile likelihood-based confidence interval that is very close to the actual interval. This approximation by Padé rational functions has the special property that, while it provides a good approximation of the function near the MLE, including allowing for skewness in the distribution of the estimator, the approximate confidence interval resulting from it can still be computed in closed form.

Section 4.3 provides background on the profile likelihood function. Section 4.4 introduces the use of such an approximation for a likelihood with a single scalar parameter of interest and no nuisance parameters. In Section 4.5 we describe the use of such an approximation for a likelihood with a single scalar parameter of interest and a single nuisance parameter. In Section 4.6 we describe the use of such an approximation for a likelihood with a single scalar parameter of interest and a vector nuisance parameter. Section 4.7 describes the use of such an approximation for likelihoods with a vector parameter of interest. In Section 4.8 we discuss possible adjustments to the profile likelihood. Section 4.9 provides examples of the usage of the approximations described in the previous sections. Section 4.10 reviews the results.

4.3 Background

4.3.1 The profile likelihood

The profile likelihood is an important tool for performing statistical inference for a finite-dimensional parameter of interest in the presence of nuisance parameters. The profile likelihood for a parameter of interest ψ is defined as

$$L_p(\psi) = \sup_{\lambda} L(\psi, \lambda) \tag{4.1}$$

where $L(\psi, \lambda)$ is the likelihood of the observed data, which depends on both the finite-dimensional parameter of interest ψ and the nuisance parameter λ. The profile likelihood only depends on the parameter of interest ψ. For example, the parameter of interest could represent the treatment effect of a certain medicine, while the nuisance parameter could represent the effects of other characteristics of a patient, such as age, which are not of direct interest to the investigator. The nuisance parameter could also represent parameters related to the scale of random errors in a statistical model which are in the final analysis not of interest to the investigator, e.g., "σ^2". The nuisance parameter may also be a function, such as a baseline hazard function. For simplicity, in this article, the parameter of interest is taken to be a single scalar value while the nuisance parameter is a single scalar value or a finite-dimensional vector.

4.3.2 Inference with the profile likelihood

If $(\hat{\psi}, \hat{\lambda})$ is a value of the parameters that maximizes the full likelihood function, then $\hat{\psi}$ maximizes the profile likelihood function. So, the profile likelihood achieves its maximum at the ψ-component of the maximum likelihood estimate. Also, under typical regularity conditions (Cramér 1946; Geyer 2013; Patefield 1977), the downward curvature of the profile likelihood at its maximum is an estimate of the variance of $\hat{\psi}$ about the true parameter value. Additionally, the set of values of ψ for which the profile likelihood in Equation (4.1) attains at least 0.15[1] of its maximum value constitutes an approximate 95% confidence interval for the true parameter value.

Denote the natural logarithm of the profile likelihood by $l_p(\psi)$ and denote the natural

[1] $0.15 \approx exp(-\frac{1}{2}\chi^2_{0.95})$ where $\chi^2_{0.95}$ is the 95th percentile of the chi-square distribution with 1 degree of freedom.

logarithm of the full likelihood by $l(\psi, \lambda)$. It is typically the case that the maximum likelihood estimates $(\hat{\psi}, \hat{\lambda})$ satisfy Equation (4.2). In Equation (4.2), subscripts denote partial derivatives, and $l_\lambda(\psi, \lambda)$ is a vector of derivatives if there is more than one nuisance parameter.

$$\begin{aligned} l_\psi(\hat{\psi}, \hat{\lambda}) &= 0 \\ l_\lambda(\hat{\psi}, \hat{\lambda}) &= 0 \end{aligned} \tag{4.2}$$

4.3.3 Drawbacks of using the profile likelihood for applied statistical inference

A drawback of using the profile likelihood for inference in applied statistical settings is that the profile likelihood is often difficult to calculate over a range of values of ψ, even in relatively simple problems. This is because to evaluate the profile likelihood for just a single value of ψ, a maximization needs to be carried out. Even in relatively simple problems, this maximization is often a numerical maximization, which can be time-consuming. For example, evaluating the relative profile likelihood for multiple different values of ψ to find which ones attain a value greater than 0.15 may be very computationally expensive. In many problems, evaluating the profile likelihood for any given *single* value of ψ takes the same amount of time as it takes to compute the full maximum likelihood estimate. The technique described in Section 4.2 and detailed in the following sections obviates those concerns.

4.4 Approximation of a likelihood with a single scalar parameter

Consider a likelihood with a single scalar parameter $L(\theta)$ and its natural logarithm $l(\theta)$. Suppose that there is a value $\hat{\theta}$ that maximizes the likelihood; typically it will also be the case that $l'(\hat{\theta}) = 0$. In some situations it is useful to consider the relative log-likelihood $h(\theta) = l(\theta) - l(\hat{\theta})$. For example, a 95% confidence interval for the parameter can be expressed succinctly in terms of the relative log-likelihood: $\{\theta |\ h(\theta) \geq -1.92\} = \{\theta \mid L(\theta)/L(\hat{\theta}) \geq 0.15\} = \{\theta \mid L(\theta)/L(\hat{\theta}) \geq exp(-\frac{1}{2}\chi^2_{0.95})\}$. The relative log-likelihood only takes negative values, and it takes the value zero at $\hat{\theta}$, the MLE. A typical relative log-likelihood is displayed in Figure 4.1.

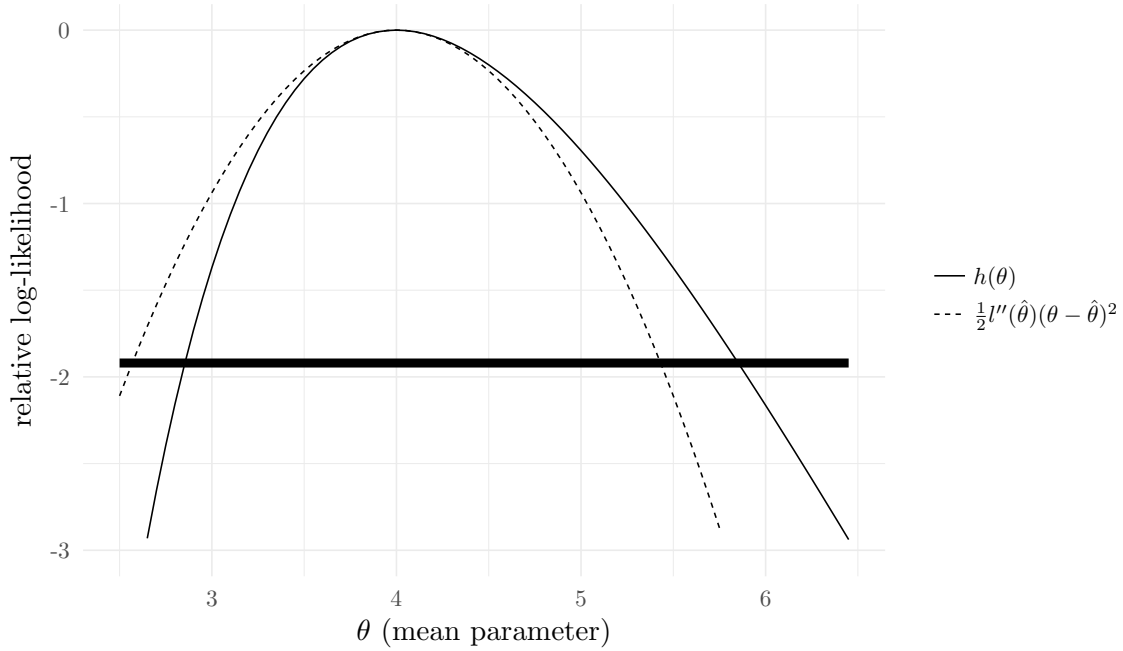

Figure 4.1: log-likelihood for an exponential distribution's mean from a sample with size $n = 30$ and mean $\bar{x} = 4$

The straight line in Figure 4.1 corresponds to the value $-1.92 \cong -\frac{1}{2}\chi^2_{0.95}$, the negate of the 95th percentile of the chi-square distribution. The values of θ for which the relative log-likelihood lies above this line are values of θ with high likelihood. Together they

form the likelihood-based confidence interval described in Section 1. Typically, the values that lie above this line do actually form an interval. In that case, the interval can be characterized by its two endpoints, which solve the equation $h(\theta) = -1.92$. For many likelihoods, this equation can only be solved numerically. In situations where is possible to take derivatives of the log-likelihood at the maximum likelihood estimate, it can be useful to approximate the likelihood by a simpler function such as a parabola and then solve for the points where that approximation intersects the straight line in order to approximate the likelihood-based confidence interval. This is especially useful if it is difficult to solve $h(\theta) = -1.92$ but easy to solve the analogous equation for a simpler approximating function $q(\theta)$ given by $q(\theta) = -1.92$. If a parabola from a second-degree Taylor series approximation is used, then the solutions for the two endpoints follow quickly from the quadratic formula. When the parabola $q_w(\theta) = \frac{1}{2}l''(\hat{\theta})(\theta - \hat{\theta})^2$ is used to approximate the log-likelihood and the likelihood-based confidence interval in this way, the resulting interval is called a Wald confidence interval.

Wald intervals are useful but sometimes have coverage that is too large or too small compared to their nominal level (e.g., of 95%). This often occurs when the likelihood function for values of θ of high likelihood cannot be well-approximated by a parabola, which happens when the sample size is small. By using a better approximation to the likelihood function, coverage closer to the nominal level can often be obtained. Additionally, closeness to the likelihood-based confidence interval can be obtained, which, in addition to often having more correct coverage, has more theoretical and philosophical support as a region of the parameter space that is actually plausible (Berger and Wolpert 1988).

4.4.1 Padé approximation and rationale

An approximation that is similar to the second-degree Taylor approximation that results in the Wald interval, in that it is also based on derivatives of the likelihood at the MLE, is the [2,2] Padé approximant (or simply "Padé approximation" in this dissertation). While the Taylor approximation is a second-degree polynomial, the Padé approximation is the ratio of two second-degree polynomials. The Padé approximation incorporates additional information from the third and fourth derivatives at the MLE and thereby obtains a better approximation. At the same time, the Padé approximation retains the useful property that the equation $q(\theta) = -1.92$ can be solved easily by the quadratic formula. Typically the Padé approximation is much better than a second-degree Taylor approximation at approximating the likelihood near the MLE, and so it results in intervals much closer to the actual likelihood interval. The formula for the Padé approximation for the relative log-likelihood is given by (4.3). The use of the Padé approximation for the relative log-likelihood in this way requires that $l''(\hat{\theta}) < 0$, and its formula is simplified by the fact that $l'(\hat{\theta}) = 0$ and $h(\hat{\theta}) = 0$. Note that the relative log-likelihood $h(\theta) = l(\theta) - l(\hat{\theta})$ differs only by an additive constant from the log-likelihood and so $h(\theta)$ and $l(\theta)$ have equal derivatives of all orders.

$$q_{\text{padé}}(\theta) = \frac{\frac{1}{2}l''(\hat{\theta})(\theta-\hat{\theta})^2}{1 - \frac{l'''(\hat{\theta})}{3l''(\hat{\theta})}(\theta-\hat{\theta}) - \frac{\left(-4l'''(\hat{\theta})^2+3l''(\hat{\theta})l''''(\hat{\theta})\right)}{36l''(\hat{\theta})^2}(\theta-\hat{\theta})^2} \tag{4.3}$$

Perhaps the most succinct way of expressing superiority of the Padé approximation over a second-degree Taylor expansion (parabola) in approximating the likelihood near the MLE is that $h(\theta) = q_w(\theta) + o((\theta-\hat{\theta})^2)$, while $h(\theta) = q_{\text{padé}}(\theta) + o((\theta-\hat{\theta})^4)$. This latter fact about the Padé approximation can be proved directly from (4.3); however, it is also a result that follows from the general theory of Padé approximants,

as described in (Baker and Graves-Morris 1996). The interpretation of this result is that, for values of θ close to the MLE, the Padé approximation is much closer to the actual likelihood than is the second-degree Taylor approximation. Because of this property, it is useful for approximating the likelihood when creating a confidence interval. Note that solving the equation $q_{\text{padé}}(\theta) = -1.92$ is immediate, since the resulting equation is a quadratic polynomial in $(\theta - \hat{\theta})$, and so it has exactly two solutions that can be solved for using the quadratic formula. These two solutions give the endpoints of the approximation to the likelihood-based confidence interval. Unlike for the Wald interval, these two endpoints need not be symmetric about the MLE. As an example, Small (2010) uses this method to obtain an approximate confidence interval for the parameter of an exponential distribution.

4.4.2 About the [2,2] Padé approximant

The [2,2] Padé approximant described in Section 4.4.1 is defined for any function $f(x)$ that is four times differentiable at a given point x_0. In the simplifying case that $f(x_0) = 0$ and $f'(x_0) = 0$, then as long as $f''(x_0) \neq 0$, the [2,2] Padé approximant is given by Equation (4.4), and it satisfies that $f(x) = f_{\text{padé}}(x) + o((x - x_0)^4)$ as $x \to x_0$. In Equation (4.4), $f^{(3)}(x_0)$ and $f^{(4)}(x_0)$ denote the third and fourth derivatives respectively of the function f, evaluated at x_0.

$$f_{\text{padé}}(x) = \frac{\frac{1}{2}f''(x_0)(x - x_0)^2}{1 + \left(-\frac{1}{3}\frac{f^{(3)}(x_0)}{f''(x_0)}\right)(x - x_0) + \left(\frac{4f^{(3)}(x_0)^2 - 3f''(x_0)f^{(4)}(x_0)}{36f''(x_0)^2}\right)(x - x_0)^2} \tag{4.4}$$

The Padé approximation is the only ratio of two second-degree polynomials that satisfies the specified order condition. A corollary of the order condition is that the first four terms of the Taylor series of $f_{\text{padé}}(x)$ about x_0 agree with the first four

terms of the Taylor series of $f(x)$ about x_0. Because, in this chapter, the relative-log likelihood satisfies the simplifying conditions that $h(\hat{\theta}) = 0$, $h'(\hat{\theta}) = 0$, and $h''(\hat{\theta}) \neq 0$, we have described this case first before reviewing Padé approximants more generally.

4.4.3 General [2,2] Padé approximants

Alternatively, if both $f'(x_0) = 0$ and $f''(x_0) = 0$, then there is no ratio of second-degree polynomials that satisfies the order condition. In this latter case, the [2,2] Padé approximant is sometimes defined to be the constant value $f_{\text{padé}}(x) = f(x_0)$, which is deficient since it satisfies $f(x) = f_{\text{padé}}(x) + o((x - x_0)^2)$ but not $f(x) = f_{\text{padé}}(x) + o((x - x_0)^4)$ when $f(x)$ has nonzero higher order derivatives. In general, the order condition will be satisfied when $f'(x_0) f'''(x_0) - f''(x_0)^2 \neq 0$. In particular, at the MLE, it is typically the case that the first derivative is zero and the second derviative is strictly negative, so the order condition is satisfied.

4.4.4 General [*m*,*n*] Padé approximants

Given a function $f(x)$ and a point to expand about, x_0, an $[m, n]$ Padé approximant can be defined in general as $p(x)/q(x)$ where $p(x)$ is an m-th (at most) degree polynomial, $q(x)$ is an n-th (at most) degree polynomial, and

$$p(x) - f(x)q(x) = o((x - x_0)^{m+n}). \tag{4.5}$$

There will exist at least one Padé approximant as long as $f(x)$ has derivatives at x_0 of order up to $m + n$. If the derivatives of $f(x)$ at x_0 satisfy a nondegeneracy condition (see Equation (4.7)) then the Padé approximant is defined uniquely and satisfies the order condition

$$f(x) = p(x)/q(x) + o((x - x_0)^{m+n}). \tag{4.6}$$

In this case where the derivatives are nondegenerate, then the Padé approximant is also definable as the unique ratio of polynomials that satisfies this latter order condition (4.6). The nondegeneracy condition can be expressed as follows: Let c_j denote the j-th derivative of $f(x)$ at x_0 for $j \geq 1$, let c_0 denote $f(x_0)$, and let c_j denote 0 for $j < 0$. Consider the matrix

$$Q_0 = \begin{pmatrix} c_{m-n+1} & c_{m-n+2} & \cdots & c_m \\ c_{m-n+2} & c_{m-n+3} & \cdots & c_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ c_m & c_{m+1} & \cdots & c_{m+n-1} \end{pmatrix}. \tag{4.7}$$

The nondegeneracy condition is that $\det(Q_0) \neq 0$. This corresponds to the existence of a denominator polynomial $q(x)$ with a nonzero constant term in the first definition (4.5) of the Padé approximant. Note that Q_0 is always a square matrix, no matter the relative sizes of m and n. Any combination of $m \geq 0$ and $n \geq 0$ can be used in a Padé approximant, and there is no restriction on their relative size.

The nondegeneracy condition is mild in the sense that it is violated only on a set of Lebesgue measure zero, so if the Padé approximant is of a likelihood from a continuous distribution, the likelihood derivatives will typically satisfy it with probability 1. As noted previously, the $[2, 2]$ Padé approximant about the MLE will typically satisfy the nondegeneracy condition. That is to say, nondegeneracy of Padé approximants does not present difficulties when a Padé approximant is used to approximate a likelihood function.

4.4.5 The coefficients of the [*m*,*n*] Padé approximant

In the case where the derivatives satisfy the nondegeneracy condition, the coefficients of the Padé approximant can be solved for from the linear system of $m+n+1$ equations that results from (4.5) where $f(x)$ is freely replaced by $\sum_{i=0}^{m+n} c_j(x-x_0)^j+o((x-x_0)^{m+n})$. To form the linear system of equations, collect terms by powers of $(x - x_0)$; all the coefficients for $(x-x_0)^j$ with $j \leq m+n$ must equal zero for the asymptotic relationship (4.5) to be satisfied. The polynomial $p(x)$ presents m unknown coefficients, and the polynomial $q(x)$ presents n unknown coefficients. A unique solution will result if the system of equations is supplemented with the additional constraint that the constant term in the polynomial $q(x)$ is set to equal, for example, the value 1; of course, the ratio $p(x)/q(x)$ does not depend on this choice.

4.4.6 Likelihood-based confidence intervals with higher order approximants

There is nothing preventing the use of higher order approximants such as a [3,2] or [4,4] approximant to a log-likelihood function about its maximum. Solving for the equation $h_{\text{padé}}(\theta) = -1.92$ in that case will result in more than two solutions, including possibly complex values, which presents the slight difficulty of deciding which roots are the "right" ones to use for the endpoints of an approximate likelihood interval. When using a Padé approximant of order $[m, n]$, the equation $h_{\text{padé}}(\theta) = -1.92$ will require finding the roots of an order $\max(m, n)$ polynomial. In particular, the roots will typically be found numerically if $\max(m, n) \geq 5$, and the roots are quite simple to find in the case of $m = 2$, $n = 2$ in which case $\max(m, n) = 2$ and the quadratic formula can be used. Generally, the improvement in the approximation of a log-likelihood from using Padé approximants of order greater than $[2, 2]$ is minimal, but the improvement

from using the $[2,2]$ approximant over the $[2,0]$ approximant (the Taylor second degree approximation used in the construction of the Wald confidence interval) can be substantial. For these reasons, this chapter focuses on the use of the [2,2] Padé approximant. Further discussion of Padé approximants can be found in (Baker and Graves-Morris 1996).

4.5 Approximation in the case of a scalar nuisance parameter

Consider a likelihood $L(\psi, \lambda)$ with two scalar parameters, where one is the parameter of interest ψ and the other is a nuisance parameter λ. For example, the parameter of interest could be the log-odds ratio β that represents a treatment effect in the logistic regression $p(x) = \text{logit}^{-1}(\alpha + \beta x)$, where $x = 0$ for the untreated group, $x = 1$ for the treated group, and $p(x)$ is the probability of recovery, while the reference level α might be a nuisance parameter. Denote the natural logarithm of the likelihood by $l(\psi, \lambda)$.

Suppose that there is a value $(\hat{\psi}, \hat{\lambda})$ that maximizes the likelihood; typically Equation (4.2) holds in this case. That is, typically it will be the case that the likelihood will be maximized at a point which lies in the interior of the parameter space in $\mathbb{R}^2$, and so the likelihood will be locally flat there, hence the two partial derivatives are equal to 0 there.

For a general value of ψ, there is often a value λ_ψ that maximizes the constrained likelihood. Suppose that this is indeed the case for general values of ψ; typically it will also be the case that $l_\lambda(\psi, \lambda_\psi) = 0$, by the same reasoning as above. Let $\lambda(\psi)$ denote the function that maps ψ to λ_ψ (for simplicity suppose there is only one such λ_ψ for each ψ). Note that, because of the way $\lambda(\psi)$ is defined, $\lambda(\hat{\psi}) = \hat{\lambda}$.

With this notation, the profile log-likelihood is given by $l_p(\psi) = l(\psi, \lambda(\psi))$. As noted in Section 4.3.2, $\hat{\psi}$ maximizes the profile log-likelihood and the relative profile log-likelihood $h(\psi) = l_p(\psi) - l_p(\hat{\psi})$. Recall that $h(\hat{\psi}) = 0$ and also $h'(\hat{\psi}) = 0$. Because $h(\psi)$ is a function of a single scalar variable that satisfies $h(\hat{\psi}) = 0$ and $h'(\hat{\psi}) = 0$, we can in principle use Equation (4.2) (with l replaced by the profile log-likelihood l_p) to obtain a Padé approximation of $h(\psi)$ about the MLE $\hat{\psi}$. That Padé approximation could then be used to approximate a profile likelihood-based confidence interval.

Note that the relative profile log-likelihood $h(\psi) = l_p(\psi) - l_p(\hat{\psi})$ differs only by an additive constant from the profile log-likelihood and so $h(\psi)$ and $l_p(\psi)$ have equal derivatives of all orders. To calculate the Padé approximation of $h(\psi)$ about $\hat{\psi}$ using Equation (4.2), we need to know the values of the derivatives $h''(\hat{\psi})$, $h^{(3)}(\hat{\psi})$, and $h^{(4)}(\hat{\psi})$, which equal the same derivatives of $l_p(\psi)$ evaluated at $\hat{\psi}$.

It may not be immediately obvious how to calculate the derivatives of $l_p(\psi) = l(\psi, \lambda(\psi))$ at $\hat{\psi}$. Note that the first derivative is typically zero since $\hat{\psi}$ maximizes $l_p(\psi)$. If $\lambda(\psi)$ is obtained by numerical maximization, as is typically the case, then a symbolic derivative of $l(\psi, \lambda(\psi))$ with respect to ψ cannot obviously be obtained through taking derivatives of a closed form for $\lambda(\psi)$. This approach is described in Section 4.5.2; a more straightforward but more computationally expensive approach is described first in Section 4.5.1.

4.5.1 Numerical finite differences

In principle, one way to deal with this difficulty is to use a numerical finite difference approximation for the second derivative of $l_p(\psi)$ such as

$$l_p''(\hat{\psi}) \cong \left(l_p(\hat{\psi} + \Delta) - l_p(\hat{\psi})\right) \Big/ \frac{1}{2}\Delta^2$$

with Δ taken to be small (e.g., $\Delta = 1 \times 10^{-5}$). This involves performing another numerical maximization to obtain $\lambda(\hat{\psi}+\Delta)$. Obtaining the third and fourth derivatives in this way will add one more numerical maximization each, for a total of four numerical maximizations, so obtaining the required derivatives in this way will take about four times as long as finding the MLE. In some settings, this time is neglible and this method works well. Another weakness of this method is that if the numerical maximizer is poor, Δ will have to be taken to be something larger (e.g., $\Delta = 1 \times 10^{-2}$). When larger values of Δ are used, the numerical derivatives suffer in quality and so the curve approximating the likelihood suffers as well.

4.5.2 Symbolic implicit differentiation

Another option for calculating the derivatives of $l_p(\psi) = l(\psi, \lambda(\psi))$ is to use the chain rule to take the total derivative with respect to ψ. Let superscripts denote derivatives so that the equation $l_\lambda(\psi, \lambda_\psi) = 0$ can also be expressed as $l^{(0,1)}(\psi, \lambda(\psi)) = 0$. The chain rule results in Equations (4.8), which are valid for any value of ψ. In the latter two formulas, the arguments of the function h have been suppressed for readability. As discussed earlier, at the MLE we have that $l^{(1,0)}(\hat{\psi}, \lambda(\hat{\psi})) = 0$ and so the first derivative is zero there. Importantly, and as discussed previously, $l^{(0,1)}(\psi, \lambda(\psi)) = 0$ for every value of ψ, which simplifies the formulas below. Note that h can be taken to be either the profile log-likelihood itself or the relative profile log-likelihood $l(\psi, \lambda(\psi)) - l(\hat{\psi}, \lambda(\hat{\psi}))$ in Equations (4.8); they differ only by a constant and so have the same derivatives. It can be seen that derivatives of $\lambda(\psi)$ up to third order are needed to calculate derivatives of $l_p(\psi)$ to fourth order.

$$
\begin{aligned}
\frac{dh(\psi,\lambda(\psi))}{d\psi} &= h^{(1,0)}(\psi,\lambda(\psi)) \\
\frac{d^2h(\psi,\lambda(\psi))}{d\psi^2} &= \lambda'(\psi)^2 h^{(0,2)}(\psi,\lambda(\psi)) + 2\lambda'(\psi)h^{(1,1)}(\psi,\lambda(\psi)) + h^{(2,0)}(\psi,\lambda(\psi)) \\
\frac{d^3h(\psi,\lambda(\psi))}{d\psi^3} &= 3h^{(1,1)}\lambda''(\psi) + h^{(0,3)}\lambda'(\psi)^3 + 3h^{(1,2)}\lambda'(\psi)^2 + 3h^{(2,1)}\lambda'(\psi) \\
&\quad + 3h^{(0,2)}\lambda'(\psi)\lambda''(\psi) + h^{(3,0)} \\
\frac{d^4h(\psi,\lambda(\psi))}{d\psi^4} &= \begin{array}{l} 4h^{(1,1)}\lambda^{(3)}(\psi) + 3h^{(0,2)}\lambda''(\psi)^2 + 6h^{(2,1)}\lambda''(\psi) + h^{(0,4)}\lambda'(\psi)^4 + 4h^{(1,3)}\lambda'(\psi)^3 \\ + 6h^{(2,2)}\lambda'(\psi)^2 + 4h^{(3,1)}\lambda'(\psi) + 4h^{(0,2)}\lambda^{(3)}(\psi)\lambda'(\psi) + 6h^{(0,3)}\lambda'(\psi)^2\lambda''(\psi) \\ + 12h^{(1,2)}\lambda'(\psi)\lambda''(\psi) + h^{(4,0)} \end{array}
\end{aligned}
\tag{4.8}
$$

If no closed form is available for $\lambda(\psi)$, it might appear difficult to obtain its derivatives for use in Equations (4.8). However, the derivatives of $\lambda(\psi)$ can be usefully expressed in terms of the derivatives of the original log-likelihood $l(\psi,\lambda)$. If the original log-likelihood can be differentiated without numerical maximization, thereby so can $\lambda(\psi)$. To see this, recall that $\lambda(\psi)$ solves the equation for local maxima, $l^{(0,1)}(\psi,\lambda(\psi)) = 0$. Taking the total derivative with respect to ψ yields the relationship $\lambda'(\psi)l^{(0,2)}(\psi,\lambda(\psi)) + l^{(1,1)}(\psi,\lambda(\psi)) = 0$. From this we obtain that $\lambda'(\psi) = -l^{(1,1)}(\psi,\lambda(\psi))/l^{(0,2)}(\psi,\lambda(\psi))$. Note that this equation holds for all values of ψ, including at the MLE. Further differentiating $\lambda'(\psi)$ yields the identities (4.9).

$$
\begin{aligned}
\lambda'(\psi) &= -\frac{l^{(1,1)}(\psi, \lambda(\psi))}{l^{(0,2)}(\psi, \lambda(\psi))} \\
\lambda''(\psi) &= -\frac{l^{(2,1)}\left(l^{(0,2)}\right)^2 - 2l^{(1,1)}l^{(1,2)}l^{(0,2)} + l^{(0,3)}\left(l^{(1,1)}\right)^2}{\left(l^{(0,2)}\right)^3} \\
\lambda'''(\psi) &= \frac{1}{\left(l^{(0,2)}\right)^5}\Big[-l^{(3,1)}\left(l^{(0,2)}\right)^4 + 3l^{(1,2)}l^{(2,1)}\left(l^{(0,2)}\right)^3 \\
&\quad + 3l^{(1,1)}l^{(2,2)}\left(l^{(0,2)}\right)^3 - 6l^{(1,1)}\left(l^{(1,2)}\right)^2\left(l^{(0,2)}\right)^2 \\
&\quad - 3\left(l^{(1,1)}\right)^2 l^{(1,3)}\left(l^{(0,2)}\right)^2 - 3l^{(0,3)}l^{(1,1)}l^{(2,1)}\left(l^{(0,2)}\right)^2 \\
&\quad + l^{(0,4)}\left(l^{(1,1)}\right)^3 l^{(0,2)} + 9l^{(0,3)}\left(l^{(1,1)}\right)^2 l^{(1,2)}l^{(0,2)} \\
&\quad -3\left(l^{(0,3)}\right)^2\left(l^{(1,1)}\right)^3\Big]
\end{aligned}
\tag{4.9}
$$

That the derivatives of $\lambda(\psi)$ can be obtained in this way allows the derivatives of $l_p(\psi)$ at a given point ψ to be computed without any additional numerical maximizations besides the ones needed to compute $\lambda(\psi)$. Note that the expression for the first derivative $\lambda'(\hat{\psi})$ and its use in the formula for the second derivative of the relative profile log-likelihood in (4.10) is the main result in (Patefield 1977). Equations (4.8) and (4.9) can be combined to yield Equations (4.10), which express the derivatives of the relative profile log-likelihood at any given ψ in terms of the partial derivatives of the full likelihood at $(\psi, \lambda(\psi))$.

$$
\begin{aligned}
\frac{dh(\psi,\lambda(\psi))}{d\psi} &= h^{(1,0)}(\psi,\lambda(\psi)) \\
\frac{d^2h(\psi,\lambda(\psi))}{d\psi^2} &= h^{(2,0)}(\psi,\lambda(\psi)) - \frac{h^{(1,1)}(\psi,\lambda(\psi))^2}{h^{(0,2)}(\psi,\lambda(\psi))} \\
\frac{d^3h(\psi,\lambda(\psi))}{d\psi^3} &= -\frac{h^{(0,3)}\left(h^{(1,1)}\right)^3}{\left(h^{(0,2)}\right)^3} + \frac{3h^{(1,2)}\left(h^{(1,1)}\right)^2}{\left(h^{(0,2)}\right)^2} - \frac{3h^{(2,1)}h^{(1,1)}}{h^{(0,2)}} + h^{(3,0)} \\
\frac{d^4h(\psi,\lambda(\psi))}{d\psi^4} &= -\frac{3\left(h^{(0,3)}\right)^2\left(h^{(1,1)}\right)^4}{\left(h^{(0,2)}\right)^5} + \frac{h^{(0,4)}\left(h^{(1,1)}\right)^4 + 12h^{(0,3)}h^{(1,2)}\left(h^{(1,1)}\right)^3}{\left(h^{(0,2)}\right)^4} \\
&\quad + \frac{-4h^{(1,3)}\left(h^{(1,1)}\right)^3 - 12\left(h^{(1,2)}\right)^2\left(h^{(1,1)}\right)^2 - 6h^{(0,3)}h^{(2,1)}\left(h^{(1,1)}\right)^2}{\left(h^{(0,2)}\right)^3} \\
&\quad + \frac{6h^{(2,2)}\left(h^{(1,1)}\right)^2 + 12h^{(1,2)}h^{(2,1)}h^{(1,1)}}{\left(h^{(0,2)}\right)^2} + \frac{-3\left(h^{(2,1)}\right)^2 - 4h^{(1,1)}h^{(3,1)}}{h^{(0,2)}} \\
&\quad + h^{(4,0)}
\end{aligned}
\tag{4.10}
$$

4.6 Approximation in the case of a vector nuisance parameter

Conclusions similar to those in Section 4.5 can be reached when the nuisance parameter λ is a vector. Suppose that there are p nuisance parameters, so the log-likelihood can be written as $l\left(\psi,\lambda_1,\lambda_2,...,\lambda_p\right)$. Let λ denote the vector of nuisance parameters, so the log-likelihood is $l(\psi,\lambda)$ and the profile log-likelihood is $l_p(\psi) = l\left(\psi,\lambda(\psi)\right)$. Given a value of ψ, $\lambda(\psi)$ typically solves the equation for the local maxima, $l_\lambda\left(\psi,\lambda(\psi)\right) = 0$. Taking the total derivative with respect to ψ results in a linear system of p equations for the p-dimensional unknown $\lambda'(\psi)$, where the coefficients are derivatives of the full likelihood, just as in Section 4.5. As in (4.9), the solution can be written as $\lambda'(\psi) = -l^{(0,2)}(\psi,\lambda(\psi))^{-1}l^{(1,1)}(\psi,\lambda(\psi))$, where the notation is denoting a matrix

inverse. As in Section 4.5, further differentiation yields $\lambda''(\psi)$ and $\lambda'''(\psi)$, which are both vectors, as well as the required derivatives $h'(\psi)$, $h''(\psi)$, $h'''(\psi)$, and $h^{(4)}(\psi)$.

Specifically, $\lambda'(\psi)$, $\lambda''(\psi)$, and $\lambda'''(\psi)$ are given by the following three formulas, in which a, b, and c are p-dimensional vectors. Note that the formula for $\lambda''(\psi)$ contains references to $\lambda'(\psi)$ and that the formula for $\lambda'''(\psi)$ refers both to $\lambda''(\psi)$ and $\lambda'(\psi)$, which in both cases are readily available from the previous formulas.

$$
\begin{aligned}
\lambda'(\psi) &= -l^{(0,2)}(\psi, \lambda(\psi))^{-1} a \\
\lambda''(\psi) &= -l^{(0,2)}(\psi, \lambda(\psi))^{-1} b \\
\lambda'''(\psi) &= -l^{(0,2)}(\psi, \lambda(\psi))^{-1} c \\
a_m &= \frac{\partial^2 l}{\partial \psi \partial \lambda_m} \\
b_m &= \frac{\partial^3 l}{\partial \psi^2 \partial \lambda_m} + 2 \sum_{i=1}^{p} \frac{\partial^3 l}{\partial \psi \partial \lambda_i \partial \lambda_m} \frac{d\lambda_i}{d\psi} + \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^3 l}{\partial \lambda_j \partial \lambda_i \partial \lambda_m} \frac{d\lambda_j}{d\psi} \frac{d\lambda_i}{d\psi} \\
c_m &= \frac{\partial^4 l}{\partial \psi^3 \partial \lambda_m} + 3 \sum_{i=1}^{p} \frac{\partial^4 l}{\partial \psi^2 \partial \lambda_i \partial \lambda_m} \frac{d\lambda_i}{d\psi} + 3 \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^4 l}{\partial \psi \partial \lambda_j \partial \lambda_i \partial \lambda_m} \frac{d\lambda_j}{d\psi} \frac{d\lambda_i}{d\psi} \\
&\quad + 3 \sum_{i=1}^{p} \frac{\partial^3 l}{\partial \psi \partial \lambda_i \lambda_m} \frac{d^2\lambda_i}{d\psi^2} + \sum_{i=1}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{\partial^4 l}{\partial \lambda_k \partial \lambda_j \partial \lambda_i \partial \lambda_m} \frac{d\lambda_k}{d\psi} \frac{d\lambda_j}{d\psi} \frac{d\lambda_i}{d\psi} \\
&\quad + 3 \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{\partial^3 l}{\partial \lambda_j \partial \lambda_i \lambda_m} \frac{d\lambda_j}{d\psi} \frac{d^2\lambda_i}{d\psi^2}
\end{aligned}
\tag{4.11}
$$

Differentiating the expression for the relative profile log-likelihood $h(\psi) = l_p(\psi) - l_p(\hat{\psi}) = l(\psi, \lambda(\psi)) - l(\hat{\psi}, \lambda(\hat{\psi}))$ with respect to ψ expresses the derivatives required for the Padé approximant, $h'(\psi)$, $h''(\psi)$, $h'''(\psi)$, and $h^{(4)}(\psi)$, in terms of derivatives of the full likelihood and in terms of $\lambda'(\psi)$, $\lambda''(\psi)$, and $\lambda'''(\psi)$. By using the solutions provided by by Equation (4.11) in Equation (4.12), we are provided with the derivatives needed for the Padé approximant of the relative profile log-likelihood $h(\psi)$ in terms of derivatives of the full likelihood.

$$
\begin{aligned}
h'(\psi) &= \frac{\partial l}{\partial \psi} \\
h''(\psi) &= \frac{\partial^2 l}{\partial \psi^2} + \sum_{i=1}^{p} \frac{\partial^2 l}{\partial \psi \partial \lambda_i} \frac{d\lambda_i}{d\psi} \\
h'''(\psi) &= \frac{\partial^3 l}{\partial \psi^3} + 2\sum_{i=1}^{p} \frac{\partial^3 l}{\partial \psi^2 \partial \lambda_i} \frac{d\lambda_i}{d\psi} + \sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^3 l}{\partial \psi \partial \lambda_j \partial \lambda_i} \frac{d\lambda_j}{d\psi}\frac{d\lambda_i}{d\psi} + \sum_{i=1}^{p} \frac{\partial^2 l}{\partial \psi \partial \lambda_i} \frac{d^2\lambda_i}{d\psi^2} \\
h^{(4)}(\psi) &= \frac{\partial^4 l}{\partial \psi^4} + 3\sum_{i=1}^{p} \frac{\partial^4 l}{\partial \psi^3 \partial \lambda_i} \frac{d\lambda_i}{d\psi} + 3\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^4 l}{\partial \psi^2 \partial \lambda_j \partial \lambda_i} \frac{d\lambda_j}{d\psi}\frac{d\lambda_i}{d\psi} + 3\sum_{i=1}^{p} \frac{\partial^3 l}{\partial \psi^2 \partial \lambda_i} \frac{d^2\lambda_i}{d\psi^2} \\
&\quad + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} \frac{\partial^4 l}{\partial \psi \partial \lambda_k \partial \lambda_j \partial \lambda_i} \frac{d\lambda_k}{d\psi}\frac{d\lambda_j}{d\psi}\frac{d\lambda_i}{d\psi} + 3\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\partial^3 l}{\partial \psi \partial \lambda_j \partial \lambda_i} \frac{d\lambda_j}{d\psi}\frac{d^2\lambda_i}{d\psi^2} \\
&\quad + \sum_{i=1}^{p} \frac{\partial^2 l}{\partial \psi \partial \lambda_i} \frac{d^3\lambda_i}{d\psi^3}
\end{aligned}
\tag{4.12}
$$

For an example of the [2,2] Padé approximant of a profile log-likelihood with a vector nuisance parameter, see Example 4.9.4.

4.7 Approximation in the case of a vector parameter of interest

There may be times when there is more than one parameter of interest. In that case, the method of Section 4.6 can be applied separately, once for each parameter of interest. However, there may be occasions where it is desirable to consider two parameters simultaneously. For example, consider the case where the two parameters of interest are the (x, y) coordinates of a 2-D location, and a confidence region is desired for the location. In that case, it makes more sense to consider the two parameters jointly rather than separately. The profile likelihood can be used to create 2-D confidence regions for parameters of this type. In this case, the profile likelihood forms a surface in three-dimensional space, and the two-dimensional region of the parameters for

which it is highest forms a confidence region. The approximate coverage probability for this confidence region is given by the likelihood ratio test: the region of the likelihood for which it attains one twentieth[2] of its maximum value forms an approximate 95% confidence region. In other words, the region of the parameters attaining relative log-likelihood greater than (-3) forms an approximate 95% confidence region.[3] In this section, we will restrict attention to the case where there are exactly two parameters of interest.

Let us first consider the case where there are two parameters of interest and there are no nuisance parameters. It is possible to define a Padé approximant for this case. However, there is no consensus on a single "best" extension of the Padé approximant to the case of more than one variable. The extension in the literature that is probably closest to the univariate Padé approximant is Cuyt's homogeneous Padé approximant (Cuyt 1999); this is what we will use here to approximate the log-likelihood for the two parameters of interest. As an example of the close relationship between the homogeneous and univariate Padé approximant, it can be shown that the $[n,m]$ homogeneous Padé approximant has the property that it reduces to the $[n,m]$ univariate Padé approximant for any coordinate when the other coordinates are fixed (Cuyt 1999). Unfortunately, unlike the [2,2] univariate Padé approximant, the closed form for the [2,2] homogeneous Padé approximant is a very long expression that spans multiple pages. Instead of presenting it directly, we will instead show how to obtain the system of equations that yields its coefficients. An example that demonstrates the usage of the [2,2] Padé approximant on a bivariate log-likelihood is given in Example 4.9.5.

[2] $\frac{1}{20} = exp(-\frac{1}{2}\chi^2_2)$
[3] $-3 \cong -\frac{1}{2}\chi^2_2$

4.7.1 The homogeneous Padé approximant

The [ν,μ] homogeneous Padé approximant can be defined for any function $f(x, y)$ about any expansion point that has partial derivatives up to total degree $\nu + \mu$ at that expansion point, which we will assume without loss of generality to be $(x, y) = (0, 0)$; that is, the function must have derivatives $\frac{\partial^{i+j} f}{\partial^i x \partial^j y}$ for $i+j \leq \nu+\mu$ at its expansion point. The homogeneous Padé approximant is a ratio of two polynomials whose coefficients are a function of those derivatives.

To define the approximant, first define coefficients c_{ij} as follows:

$$c_{ij} = \frac{1}{i!j!} \frac{\partial^{i+j} f}{\partial^i x \partial^j y}$$

Also, define the following polynomials:

$$\begin{aligned}
A_l(x, y) &= \sum_{i+j=\nu\mu+l} a_{ij} x^i y^j, \quad l = 0, ..., \nu \\
B_l(x, y) &= \sum_{i+j=\nu\mu+l} b_{ij} x^i y^j, \quad l = 0, ..., \mu \\
C_l(x, y) &= \sum_{i+j=l} c_{ij} x^i y^j, \quad l = 0, 1, 2, ... \\
p(x, y) &= \sum_{l=0}^{\nu} A_l(x, y) \\
q(x, y) &= \sum_{l=0}^{\mu} B_l(x, y)
\end{aligned} \tag{4.13}$$

The system of equations that yields a_{ij} and b_{ij} is obtained as follows:

Ensure that $(\sum_{l=0}^{\nu+\mu} C_l(x, y))q(x, y) - p(x, y)$ has a coefficient of 0 for all terms of total degree $\nu\mu + \nu + \mu$ or lower.

This condition can also be written as follows:

$$
\begin{aligned}
&C_0(x,y)B_0(x,y) = A_0(x,y) \\
&C_1(x,y)B_0(x,y) + C_0(x,y)B_1(x,y) = A_1(x,y) \\
&\quad\vdots \\
&C_\nu(x,y)B_0(x,y) + \cdots + C_{\nu-\mu}B_\mu(x,y) = A_\nu(x,y) \\
&\\
&C_{\nu+1}(x,y)B_0(x,y) + \cdots + C_{\nu+1-\mu}(x,y)B_\mu(x,y) \equiv 0 \\
&\quad\vdots \\
&C_{\nu+\mu}B_0(x,y) + \cdots + C_\nu(x,y)B_\mu(x,y) \equiv 0
\end{aligned}
\tag{4.14}
$$

where $C_l(x,y) \equiv 0$ for $l < 0$. As in the univariate case, there is one more unknown than the number of equations; one can set $b_{0,\nu\mu} = 1$; the ratio $p(x,y)/q(x,y)$ is unaffected by this choice. Solving the linear system of equations from (4.14) for a_{ij} and b_{ij} and using their values in $p(x,y)/q(x,y)$ yields the [ν,μ] homogeneous Padé approximant.

4.7.2 Vector parameter of interest with vector nuisance parameter

To obtain the derivatives of the profile log-likelihood needed for the multivariate homogeneous Padé approximant when there are one or more nuisance parameters, one can form the equations analagous to (4.11) and (4.12) in order to express the partial derivatives of the profile log-likelihood with respect to the parameters of interest in terms of derivatives of the full log-likelihood. After obtaining the partial derivatives up to total order $\nu + \mu$ of the profile log-likelihood, one can form the [ν,μ] Padé approximant and use it to create a confidence region. Unfortunately, the confidence region does not arise from the quadratic formula or any other exceedingly simple formula in the multivariate case, whether there are nuisance parameters or not. See

Example 4.9.6 for an example with a vector parameter of interest and vector nuisance parameter.

4.8 Other applications and discussion

4.8.1 Adjusted profile likelihood and pseudolikelihoods

When there are many nuisance parameters, the profile likelihood may not perform inferentially as well as adjusted profile likelihoods and other pseudolikelihoods. These pseudolikelihoods often are expressible as small adjustments to the profile likelihood. For example, an adjusted profile log-likelihood from (Severini 2000) is given by $l_a(\psi) = l_p(\psi) - \frac{1}{2}\log|-l_{\lambda\lambda}(\psi, \lambda(\psi))|$. A Padé approximation can be used to in the same way to calculate confidence intervals based on the adjusted profile log-likelihood. The same technique can be used, except the derivatives of the profile log-likelihood will clearly have to be added to the derivatives of the second term (which are also expressible in terms of $\lambda(\psi)$'s derivatives) to obtain the derivatives of the adjusted profile log-likelihood.

4.9 Examples

4.9.1 The mean of an exponential distribution

In this example, we consider a likelihood for a single scalar parameter of interest, the mean parameter θ of an exponential distribution. The probability density function for a single observation $x > 0$ is given by $f(x) = \theta^{-1}e^{-x/\theta}$. Suppose that you have observed a sample of $n = 4$ independently and identically distributed ($i.i.d.$) values with sample mean $\bar{x} = 4$. Then the log-likelihood is $l(\theta) = -16\theta^{-1} - 4\log(\theta)$. This

function achieves its maximum at $\hat{\theta} = 4$, so the relative log-likelihood is given by $h(\theta) = 4 - 16\theta^{-1} + 4\log(4\theta^{-1})$. Calculating the three derivatives of the log-likelihood required for the Padé approximation, we find that $h''(\hat{\theta}) = -1/4$, $h'''(\hat{\theta}) = 1/4$, and $h^{(4)}(\hat{\theta}) = -9/32$. Using Equation (4.4) we find that

$$q_{\text{padé}}(\theta) = \frac{-\frac{1}{8}(\theta - 4)^2}{1 + \frac{1}{3}(\theta - 4) + \frac{5}{288}(\theta - 4)^2}.$$

We can also write the second-degree Taylor expansion as $q_w(\theta) = -\frac{1}{8}(\theta - 4)^2$. Figure 4.2 compares the relative log-likelihood $h(\theta)$, its Padé approximation $q_{\text{padé}}(\theta)$, and its second-degree Taylor expansion $q_w(\theta)$. The bold horizontal line denotes the value -1.92; values of θ for which $h(\theta) > -1.92$ constitute an approximate 95% confidence interval.

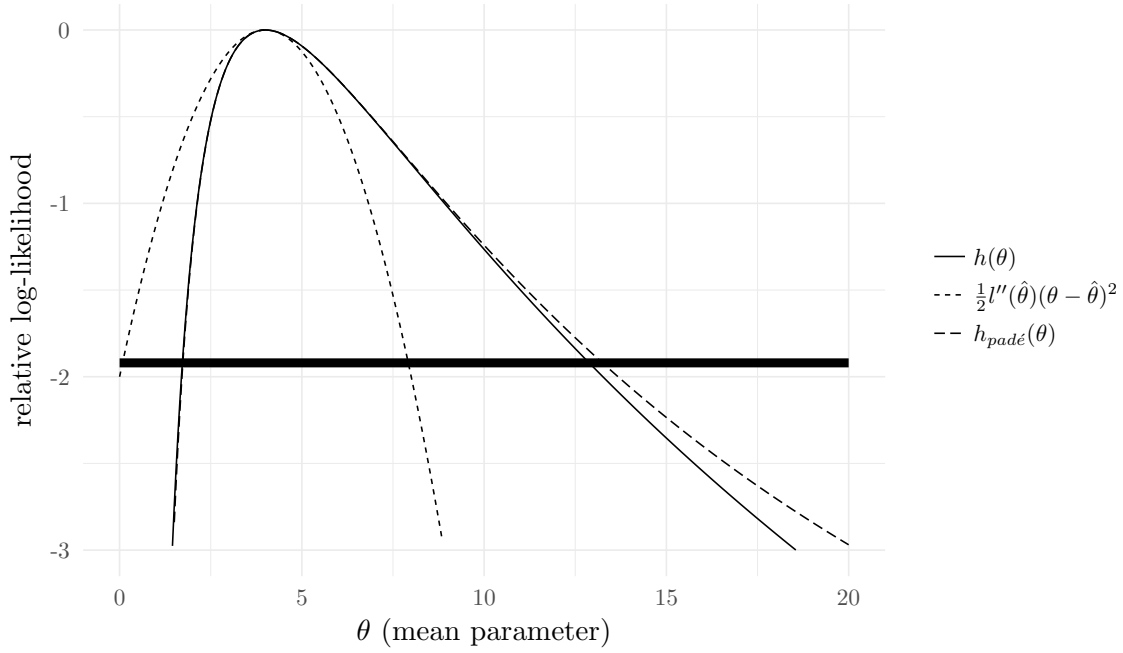

Figure 4.2: log-likelihood for an exponential distribution's mean from a sample with size $n = 4$ and mean $\bar{x} = 4$. The horizontal line is the value $-1.92 \cong -\frac{1}{2}\chi^2_{0.95}$.

4.9.2 The index of a gamma distribution

In this example, we consider the profile likelihood for a scalar parameter of interest in the presence of a scalar nuisance parameter. The parameter of interest is the index α of a gamma distribution and the nuisance parameter is the scale parameter β. The probability density function for a single observation $x > 0$ is given by $f(x) = \Gamma(\alpha)^{-1}\beta^{-\alpha}x^{\alpha-1}e^{-x/\beta}$. Suppose that you have observed a sample of n *i.i.d.* values from a gamma distribution with unknown α and β. Then the log-likelihood is given by $l(\alpha, \beta) = -n\log\Gamma(\alpha) + n(\alpha - 1)\overline{\log}(x) - n\bar{x}/\beta - n\alpha\log(\beta)$ where $\bar{x}$ denotes the observed sample mean and $\overline{\log}(x)$ denotes $\sum_{i=1}^{n}\log(x_i)/n$. In this example, given a value of α, the maximizing value of the nuisance parameter follows quickly from $l_\beta(\alpha, \beta) = 0$, which yields $\beta(\alpha) = \bar{x}/\alpha$. Because $\beta(\alpha)$ has a simple closed form, Equations (4.10) are not needed to obtain the derivatives of the relative profile log-likelihood; the derivatives of the profile likelihood can be obtained from its simple closed form $l_p(\alpha) = l(\alpha, \beta(\alpha)) = -n\log\Gamma(\alpha) + (\alpha - 1)n\overline{\log}(x) - n\alpha - n\alpha\log(\bar{x}/\alpha)$.

The first four derivatives of the profile log-likelihood are given by (4.15). In these formulas, $\psi^{(j)}(\alpha)$ represents the $(j+1)$-st derivative of the function $\log\Gamma(\alpha)$. The maximizer of the profile likelihood $\hat{\alpha}$ is generally found numerically. After $\hat{\alpha}$ has been found, the Padé approximation to the profile log-likelihood about $\hat{\alpha}$ can be computed using (4.4) with the derivatives in (4.15).

$$\begin{aligned}
l_p(\alpha) &= -n\log\Gamma(\alpha) + n(\alpha - 1)\overline{\log}(x) - n\alpha - n\alpha\log(\bar{x}/\alpha) \\
l_p'(\alpha) &= -n\log(\bar{x}/\alpha) + n\overline{\log}(x) - n\psi^{(0)}(\alpha) \\
l_p''(\alpha) &= n/\alpha - n\psi^{(1)}(\alpha) \\
l_p^{(3)}(\alpha) &= -n/\alpha^2 - n\psi^{(2)}(\alpha) \\
l_p^{(4)}(\alpha) &= 2n/\alpha^3 - n\psi^{(3)}(\alpha)
\end{aligned} \tag{4.15}$$

Suppose from a sample size of size $n = 12$ you observe $\bar{x} = 30$ and $\overline{\log}(x) = 3.22537$. Numerically maximizing the likelihood we find that $\hat{\alpha} = 3$ and $\hat{\beta} = 10$. Evaluating the derivatives of the profile log-likelihood and at $\hat{\alpha} = 3$, we find that $l_p''(\hat{\alpha}) = -0.74$, $l_p^{(3)}(\hat{\alpha}) = 0.52$, and $l_p^{(4)}(\hat{\alpha}) = -0.54$. Using Equation (4.4), we can write the Padé approximation to the relative profile log-likelihood as

$$q_{\text{padé}}(\alpha) = \frac{-0.37(\alpha - 3)^2}{1 + 0.23(\alpha - 3) - 0.006(\alpha - 3)^2}.$$

We can also write the second-degree Taylor expansion as $q_w(\alpha) = -0.37(\alpha - 3)^2$. Figure 4.3 compares the relative log-likelihood $h(\alpha)$, its Padé approximation $q_{\text{padé}}(\alpha)$, and its second-degree Taylor expansion $q_w(\alpha)$. The bold horizontal line denotes the value -1.92; values of α for which $h(\alpha) > -1.92$ constitute an approximate 95% confidence interval.

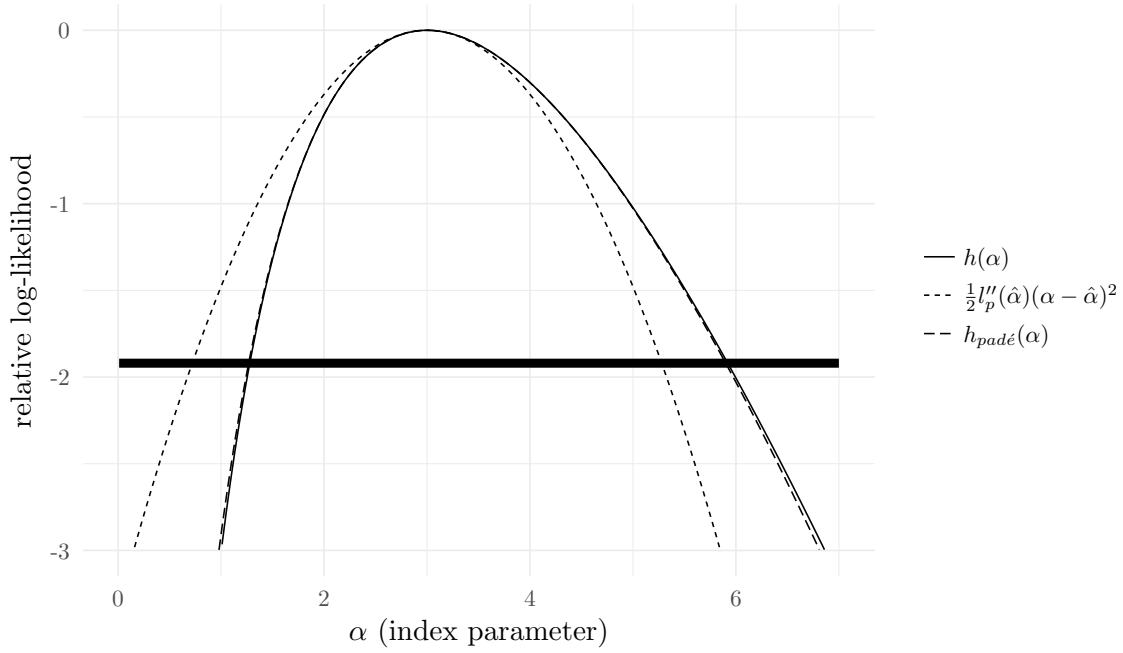

Figure 4.3: profile log-likelihood for a gamma distribution's index from a sample with size $n = 12$, mean $\bar{x} = 30$, and mean log $\overline{\log}(x) = 3.22537$. The horizontal line is the value $-1.92 \cong -\frac{1}{2}\chi^2_{0.95}$.

Because the maximizing value of the nuisance parameter has a simple closed form, it is

particularly easy to understand the effect of using adjustments to the profile likelihood. Consider the adjusted profile likelihood $l_a(\alpha) = l_p(\alpha) - 0.5 \log\left(-l_{\beta\beta}(\alpha, \beta(\alpha))\right)$. Taking derivatives, we see that the additional term adds $-0.5 \ln\left(\alpha^3\right)$ to the profile log-likelihood, which penalizes larger values of α. If $n = 7$ and $\alpha = 3$, then a nominal 95% profile likelihood interval for α has 91.4% coverage, a Wald interval has 97.3% coverage, and the likelihood interval using the adjusted profile likelihood (or a Padé approximation about its maximum) has 94.5% coverage. The form of the adjustment here is from (Severini 2000), which also provides a discussion of why adjustments to the profile likelihood should be considered in cases where the sample size is small or there are many nuisance parameters.

4.9.3 Logistic regression

In this example, we apply Equations (4.10) to the likelihood for a scalar parameter of interest and a scalar nuisance parameter. This example differs from that of Section 4.9.2 in that, here, $\lambda(\psi)$ is difficult to express in closed form and is typically obtained using numerical methods. Consider drawing samples from two different binomial distributions, $x_1 \sim \text{binomial}\,(m_1, p_1)$ and $x_2 \sim \text{binomial}\,(m_2, p_2)$. That is, group 1 is size m_1 and has probability p_1 of success per unit; group 2 is of size m_2 and has probability p_2 of success per unit; and x_1 and x_2 are the numbers of successes observed from those two groups, respectively. Suppose that the parameter of interest is the log odds ratio $\beta = \log\left(p_2/(1-p_2)\right) - \log\left(p_1/(1-p_1)\right)$, which measures the difference in success rates between the two groups. When $p_2 > p_1$, $\beta > 0$. The single scalar nuisance parameter can be chosen with a great deal of freedom, but conventionally it is taken as $\alpha = \log\left(p_1/(1-p_1)\right)$. No other nuisance parameters are needed, since we are considering m_1 and m_2 to be known. The parameters (α, β) are in one-to-one correspondence with the parameters (p_1, p_2) as long as $p_1, p_2 \neq 0, 1$.

In the framework of logistic regression, the problem is $p(x) = \text{logit}^{-1}(\alpha + \beta x)$, where $x = 1$ indicates membership to group 2 and $x = 0$ indicates membership to group 1 so that $p(0) = p_1$ and $p(1) = p_2$. The function $p(x)$ denotes the probability of success given group membership. By the invariance principle of maximum likelihood, the MLE $(\hat{\alpha}, \hat{\beta})$ is easily obtainable by evaluating the definitions of α and β with p_1 and p_2 replaced by their maximum likelihood estimates $\hat{p}_1 = x_1/m_1$ and $\hat{p}_2 = x_2/m_2$. However, constrained estimates of the nuisance parameter α when the parameter of interest β is fixed at any value other than the MLE (as in a profile likelihood) are obtainable only numerically. In this case then, the use of a Padé approximation obtains to great accuracy the profile likelihood interval for the parameter of interest without any numerical maximizations.

The derivatives of the profile log-likelihood can be obtained using the formulas in (4.10), yielding the following:

$$l_p''(\beta) = -\frac{m_1 m_2}{2\left(m_1 \cosh(\alpha(\beta)+\beta) + m_2 \cosh(\alpha(\beta)) + m_1 + m_2\right)}$$

$$l_p^{(3)}(\beta) = \frac{m_1 m_2 \left(m_1^2(2\sinh(\alpha(\beta)+\beta) + \sinh(2(\alpha(\beta)+\beta))) - m_2^2(2\sinh(\alpha(\beta)) + \sinh(2\alpha(\beta)))\right)}{4\left(m_1 \cosh(\alpha(\beta)+\beta) + m_2 \cosh(\alpha(\beta)) + m_1 + m_2\right)^3}$$

The formula for $l_p^{(4)}(\beta)$ is omitted for space considerations.

Suppose from samples of size $m_1 = 10$ and $m_2 = 11$ you observe $x_1 = 1$ and $x_2 = 9$. Then $\hat{p}_1 = 0.10$ and $\hat{p}_2 = 0.82$, and, by the invariance property, $\hat{\alpha} = -2.2$ and $\hat{\beta} = 3.7$. Using the formulas for the derivatives of the profile log-likelihood and evaluating them at the MLE, we find that $l_p''(\beta) = -0.58$, $l_p^{(3)}(\beta) = 0.24$, and $l_p^{(4)}(\beta) = -0.04$. Using formula (4.4), we can write the Padé approximation to the relative profile log-likelihood as

$$q_{\text{padé}}(\alpha) = \frac{-0.29(\beta - 3.7)^2}{1 + 0.14(\beta - 3.7) - 0.013(\beta - 3.7)^2}. \tag{4.16}$$

We can also write the second-degree Taylor expansion, which requires only the second derivative $l_p''(\hat{\beta})$, as $q_w(\alpha) = -0.29(\beta - 3.7)^2$. Figure 4.4 compares the relative log-likelihood $h(\beta)$, its Padé approximation $q_{\text{padé}}(\beta)$, and its second-degree Taylor expansion $q_w(\beta)$. The bold horizontal line denotes the value -1.92; values of β for which $h(\beta) > -1.92$ constitute an approximate 95% confidence interval.

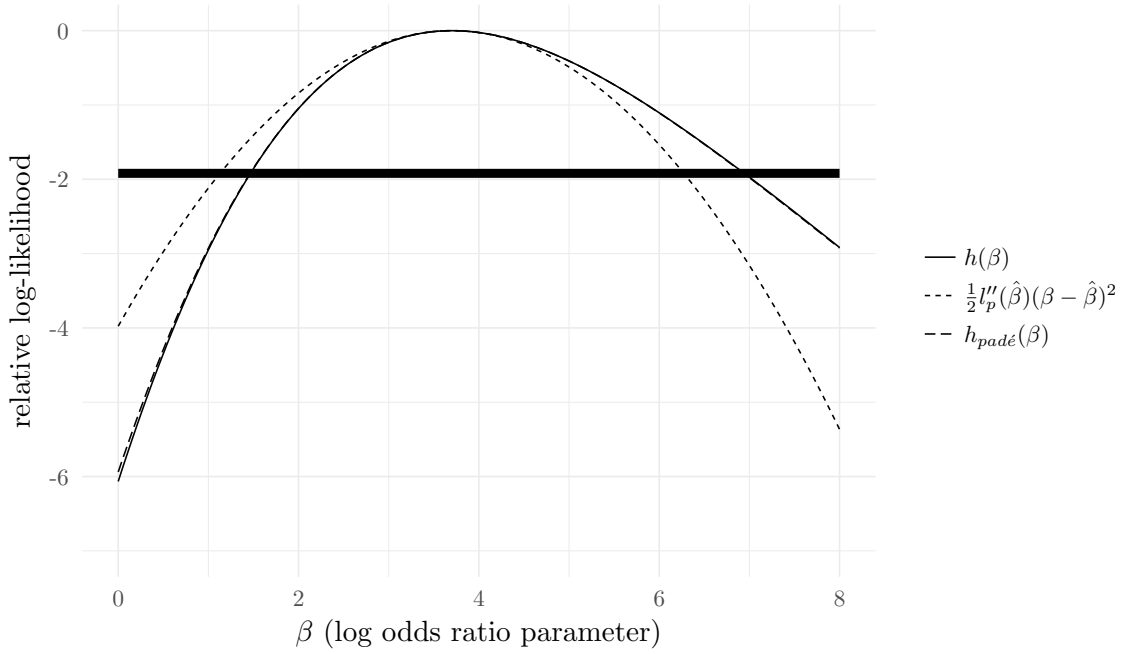

Figure 4.4: profile log-likelihood for the log odds ratio between two binomial distributions from samples with sizes $m_1 = 10$ and $m_2 = 11$ and number of successes $x_1 = 1$ and $x_2 = 9$. The horizontal line is the value $-1.92 \cong -\frac{1}{2}\chi^2_{0.95}$.

Suppose the true values of the parameters are $p_1 = 0.3$ and $p_2 = 0.6$. Then a nominal 95% profile likelihood interval for the log odds ratio β has 95.2% coverage; a Wald interval has 97.0% coverage; and the interval obtained using a Padé approximation also has 95.2% coverage. A likelihood interval based on the likelihood of x_2 given $x_1 + x_2$ (a hypergeometric distribution) has 94.8% coverage. Unlike the profile likelihood or its parabolic approximation, the hypergeometric distribution is a bona fide likelihood,

and it does not depend on the reference level α. A Padé approximation to the hypergeometric likelihood can also be considered, which results in 94.8% coverage as well.

4.9.4 Logistic regression with vector nuisance parameter

In this example, we apply the results of Section 4.6 to the likelihood for a scalar parameter of interest and a vector nuisance parameter. This example differs from that of Section 4.9.3 in that, here, the nuisance parameter is a vector. Consider drawing samples a binomial distribution with probability of success $p(x) = \text{logit}^{-1}(\alpha + \beta_1 x_1 + \beta_2 x_2)$, where $x_1, x_2 = 0, 1$ and where the paramter of interest is α and the vector nuisance parameter is (β_1, β_2). Data is generated from a true model with $\alpha = 0.5$, $\beta_1 = -1.0$, and $\beta_2 = 1.5$. A total of $n = 16$ samples are taken, four where $x_1 = 0, x_2 = 0$, four where $x_1 = 1, x_2 = 0$, four where $x_1 = 0, x_2 = 1$, and four where $x_1 = 1, x_2 = 1$, with observed numbers of successes 2,2,4, and 3 respectively. The full likelihood for the observed data is displayed in Equation (4.17).

$$\begin{aligned} l(\alpha, \beta_1, \beta_2) = {} & 2\log\left(\frac{e^{\alpha+\beta_1}}{e^{\alpha+\beta_1}+1}\right) + 4\log\left(\frac{e^{\alpha+\beta_2}}{e^{\alpha+\beta_2}+1}\right) + 3\log\left(\frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha+\beta_1+\beta_2}+1}\right) \\ & + 2\log\left(1 - \frac{e^{\alpha+\beta_1}}{e^{\alpha+\beta_1}+1}\right) + \log\left(1 - \frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha+\beta_1+\beta_2}+1}\right) + 2\log\left(\frac{e^{\alpha}}{e^{\alpha}+1}\right) \\ & + 2\log\left(1 - \frac{e^{\alpha}}{e^{\alpha}+1}\right) \end{aligned} \tag{4.17}$$

Maximizing the likelihood, we find that the MLE is $\hat{\alpha} = 0.353277, \hat{\beta}_1 = -0.706554, \hat{\beta}_2 = 1.99237$.

Using Equation (4.11), we find that $\beta_1'(\hat{\alpha}) = -0.828126$, $\beta_2'(\hat{\alpha}) = -0.477817$, $\beta_1''(\hat{\alpha}) = -0.0320928$, $\beta_2''(\hat{\alpha}) = 0.143191$, $\beta_1'''(\hat{\alpha}) = -0.025675$, and $\beta_2'''(\hat{\alpha}) = 0.0170632$.

Using Equation (4.12), we then find that $h''(\psi) = -1.13605$, $h'''(\psi) = 0.195582$, and $h^{(4)}(\psi) = 0.47019$.

Figure 4.5 displays the relative profile log-likelihood for α, a quadratic approximation about its maximum, and also the [2,2] Padé approximant about its maximum.

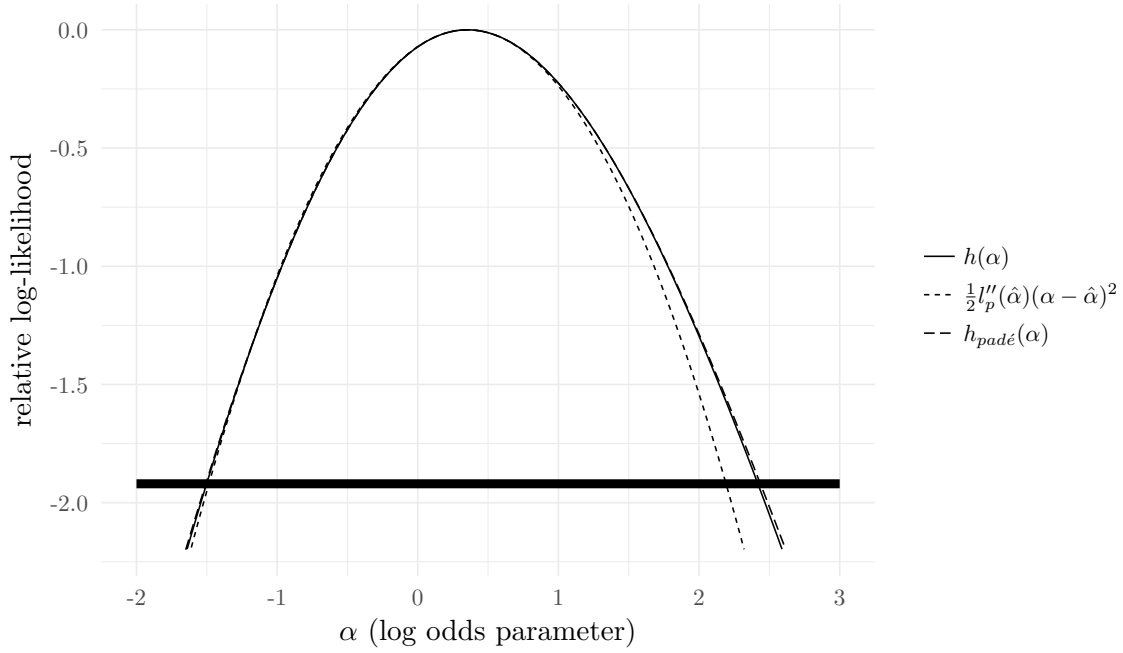

Figure 4.5: profile log-likelihood for α, the log odds of success in the case of $x_1 = 0, x_2 = 0$. The horizontal line is the value $-1.92 \cong -\frac{1}{2}\chi^2_{0.95}$.

4.9.5 Confidence region for parameters of a gamma distribution

In this example, we consider the likelihood for two parameters of interest in the absence of nuisance parameters. The two parameters of interest are the index parameter α and scale parameter β of a gamma distribution. The probability density function for a single observation $x > 0$ is given by $f(x) = \Gamma(\alpha)^{-1}\beta^{-\alpha}x^{\alpha-1}e^{-x/\beta}$. Suppose that you have observed a sample of n *i.i.d.* values from a gamma distribution with unknown α and β. Then the log-likelihood is given by $l(\alpha, \beta) = -n\log\Gamma(\alpha) + n(\alpha - 1)\overline{\log}(x) -$

$n\bar{x}/\beta - n\alpha \log(\beta)$ where $\bar{x}$ denotes the observed sample mean and $\overline{\log}(x)$ denotes $\sum_{i=1}^{n} \log(x_i)/n$.

As noted in Section 4.7, the region of the parameters for which the likelihood attains one twentieth of its maximum value forms an approximate 95% confidence region for the true parameters. Suppose one observes a sample of $n = 150$ independent observations from a gamma distribution with unknown (α, β) with $\bar{x} = 28.2544$ and $\overline{\log}(x) = 3.1791$. Then the MLE is given by $(\hat{\alpha}, \hat{\beta}) = (3.2411, 8.7175)$. Figure 4.6 displays the 95% confidence region arising from the full log-likelihood, its [2,2] homoeogeneous Padé approximant, and the full log-likelihood's quadratic approximation. The derivatives required for the Padé approximant and the quadratic approximation follow straightforwardly from explicitly differentiating the full likelihood. The confidence region from the Padé approximant is closer to the desired region than the confidence region from the quadratic approximation. However, the confidence region from the Padé approximant is not indistinguishable from the desired region.

4.9.6 Confidence region for parameters in a Poisson regression

In this example, we consider the profile likelihood for two parameters of interest in the presence of two nuisance parameters. We consider a Poisson regression where $y \sim Poisson(exp(\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3))$ for $x \in [-2, 2]$. The two parameters of interest are β_1 and β_3; the two others are nuisance parameters.

As noted in Section 4.7, the region of the parameters for which the profile likelihood attains one twentieth of its maximum value forms an approximate 95% confidence region for the true parameters. Suppose one observes a sample of $n = 10$ (x, y) pairs, with $x_i = -2 + \frac{4}{9}(i-1)$ and $y = (0, 0, 0, 0, 1, 0, 3, 1, 31, 377)$. Then the MLE is given

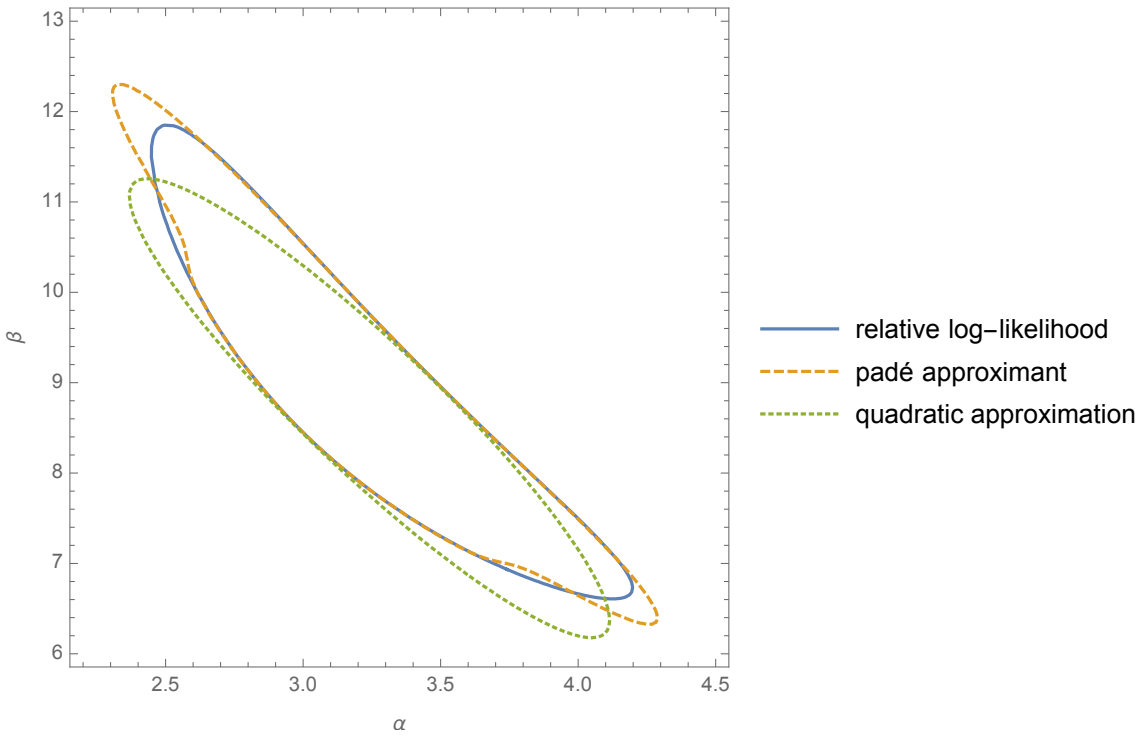

Figure 4.6: 95% Confidence regions for the parameters of a gamma distribution

by $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-0.77, 1.47, 0.06, 0.44)$. Figure 4.7 displays the 95% confidence region for (β_1, β_3) arising from the profile log-likelihood, its [2,2] homoegeneous Padé approximant, and the profile log-likelihood's quadratic approximation. The derivatives required for the Padé approximant and the quadratic approximation follow from the multivariate analogues of Equations (4.11) and (4.12). The confidence region from the Padé approximant is closer to the desired region than the confidence region from the quadratic approximation. However, the confidence region from the Padé approximant is not indistinguishable from the desired region.

4.10 Discussion

The Padé approximation to the profile likelihood discussed in this chapter provides a confidence interval that is numerically similar to a likelihood interval but is computa-

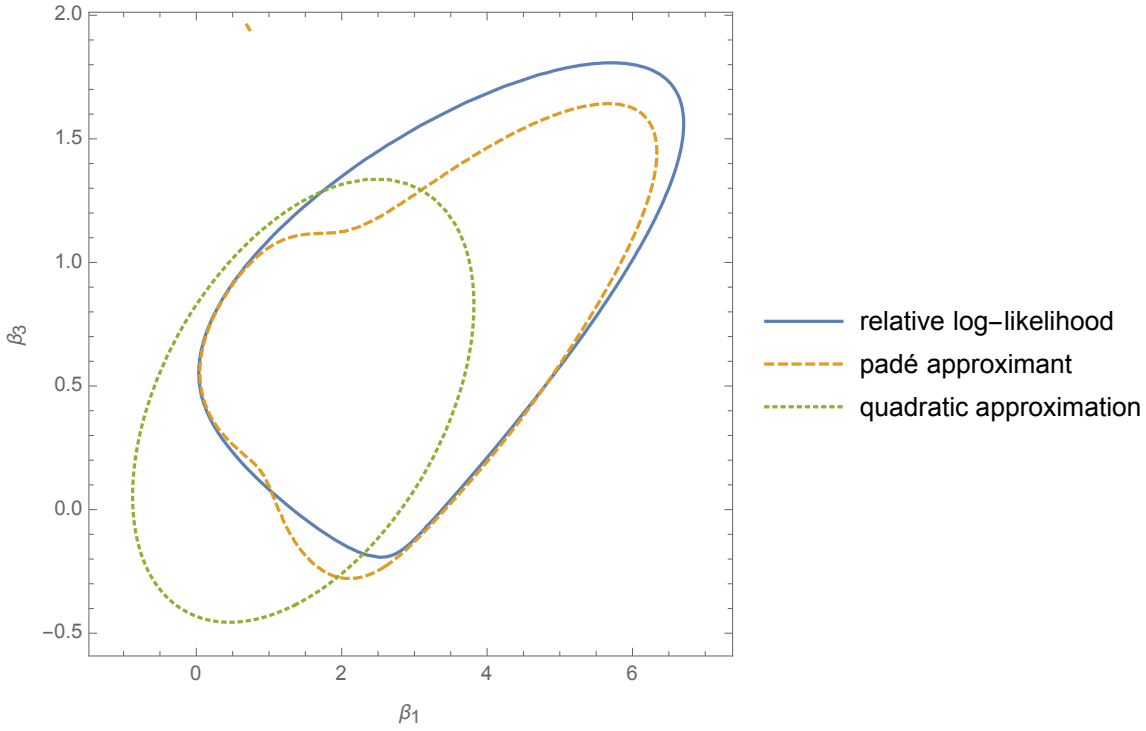

Figure 4.7: 95% Confidence regions for the parameters of a Poisson regression

tionally more comparable to a Wald interval. The method is especially useful when the likelihood is explicit in the parameters, so that it can be explicitly differentiated; in this case, the computation time is essentially the same as that of a Wald interval. The likelihood will generally not be explicit in the parameters when, for example, it is the marginal likelihood of a random effects model with arbitrary mixing distribution. However, as seen in (DiCiccio and Monti 2001), approximations of this kind can be used to usefully decrease computation time even for complicated likelihoods. There, the computation time requirement is made more important since the method is used in construction with the bootstrap. The use of the Padé approximation is not limited to profile likelihoods; in small samples it is seen to be most usefully combined with the technique of adjusting the profile likelihood, such as with the modified profile likelihood or other adjusted likelihoods. When the Padé approximation is used with these techniques, the resulting confidence interval is similar to a Bayesian credible interval with noninformative prior in the way it improves upon the Wald interval.

However, the Padé method requires less computation time and allows the possibility of a confidence interval that is robust to the model, in the sense discussed in (White 1982) and (Royall and Tsou 2003), through the Bartlett correction of the likelihood ratio statistic.

Chapter 5

Conclusions

5.1 Summary

After introducing choice-based conjoint analysis, we introduced a method to empirically assess and compare different conjoint analysis design strategies. A key feature of this method of comparison was that it makes very few assumptions about the data-generating process (essentially just that the respondents answer the surveys independently of one another) while remaining statistically valid; in particular, the respondents are not assumed to use the multinomial logit model.

We then turned to the statistical analysis of conjoint analysis survey results. We introduced a method to plot in two-dimensional space the heterogeneity in preferences across the respondents as inferred from the conjoint analysis survey results that can be used for visualization as well as mixture model assessment. We also introduced a novel method to accurately infer the number of natural clusters in preferences across the respondents. This method was shown in simulation studies to give more accurate results than latent class segmentation with AIC and BIC. We additionally suggested regressing estimated respondent preferences on their demographic variables

as a strategy for hypothesis generation and model building.

We showed that the profile likelihood can be well approximated near the maximum likelihood estimate by the [2,2] Padé approximant. This approximation is shown to be better than that provided by a second-order Taylor series approximation. However, like a second-order Taylor series approximation, it can be used to construct a confidence interval with endpoints found using the quadratic formula. The resulting confidence interval is similar to a profile likelihood interval but with computation time similar to that of a Wald interval.

5.2 Limitations

The methodology for assessing experimental design and for statistical analysis in conjoint analysis studies described in this thesis largely ignores the issue of the scale of respondent's preferences. The issue of respondent scale arises especially when comparing multinomial logit models estimated from different data sources (Swait and Louviere 1993; Hensher, Louviere, and Swait 1998).

There are difficulties extending the Padé approximant to more than one dimension. Because of this, there are difficulties when using the Padé approximant to approximate a profile likelihood with a multidimensional parameter of interest. Additionally, the presence of a multidimensional nuisance parameter complicates the exposition of the Padé approximation.

Bibliography

Allenby, Greg M, and Peter J Lenk. 1994. "Modeling Household Purchase Behavior with Logistic Normal Regression." *Journal of the American Statistical Association* 89 (428): 1218–31.

Anderson, Keaven M, Patricia M Odell, Peter WF Wilson, and William B Kannel. 1991. "Cardiovascular Disease Risk Profiles." *American Heart Journal* 121 (1): 293–98.

Bajari, Patrick, Jeremy T Fox, and Stephen P Ryan. 2007. "Linear Regression Estimation of Discrete Choice Models with Nonparametric Distributions of Random Coefficients." *The American Economic Review* 97 (2): 459–63.

Baker, George A, and Peter Russell Graves-Morris. 1996. *Padé Approximants*. Vol. 59. Encyclopedia of Mathematics and Its Application. Cambridge University Press, Cambridge.

Bech, Mickael, and Dorte Gyrd-Hansen. 2005. "Effects Coding in Discrete Choice Experiments." *Health Economics* 14 (10): 1079–83.

Berger, James O, and Robert L Wolpert. 1988. "The Likelihood Principle." *Lecture Notes-Monograph Series* 6: iii–199.

Bhatnagar, Amit, and Sanjoy Ghose. 2004. "A Latent Class Segmentation Analysis of E-Shoppers." *Journal of Business Research* 57 (7): 758–67.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert. 2000. "Assessing a Mixture

Model for Clustering with the Integrated Completed Likelihood." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (7): 719–25.

Boyd, J Hayden, and Robert E Mellman. 1980. "The Effect of Fuel Economy Standards on the Us Automotive Market: An Hedonic Demand Analysis." *Transportation Research Part A: General* 14 (5-6): 367–78.

Bridges, John FP. 2013. "Overlap and the Efficiency of Discrete-Choice Experiments: Evidence from a Randomized Experiment." In *The 35th Annual Meeting of the Society for Medical Decision Making*. The Society for Medical Decision Making.

Bridges, John FP, A Brett Hauber, Deborah Marshall, Andrew Lloyd, Lisa A Prosser, Dean A Regier, F Reed Johnson, and Josephine Mauskopf. 2011. "Conjoint Analysis Applications in Health—a Checklist: A Report of the Ispor Good Research Practices for Conjoint Analysis Task Force." *Value in Health* 14 (4): 403–13.

Bridges, John FP, Elizabeth T Kinter, Annette Schmeding, Ina Rudolph, and Axel Mühlbacher. 2011. "Can Patients Diagnosed with Schizophrenia Complete Choice-Based Conjoint Analysis Tasks?" *The Patient: Patient-Centered Outcomes Research* 4 (4): 267–75.

Clark, Michael, and Emilia Vynnycky. 2004. "The Use of Maximum Likelihood Methods to Estimate the Risk of Tuberculous Infection and Disease in a Canadian First Nations Population." *International Journal of Epidemiology* 33 (3): 477–84.

Consortium, Breast Cancer Linkage, and others. 1999. "Cancer Risks in Brca2 Mutation Carriers." *Journal of the National Cancer Institute* 91 (15): 1310–6.

Cramér, Harald. 1946. *Mathematical Methods of Statistics (Pms-9)*. Princeton University Press.

Cunningham, Charles E, Beth S Bruce, Anne W Snowdon, Yvonne Chen, Carol Kolga, Caroline Piotrowski, Lynne Warda, Heather Correale, Erica Clark, and Melanie

Barwick. 2011. "Modeling Improvements in Booster Seat Use: A Discrete Choice Conjoint Experiment." *Accident Analysis & Prevention* 43 (6): 1999–2009.

Cunningham, Charles E, Ken Deal, and Yvonne Chen. 2010. "Adaptive Choice-Based Conjoint Analysis." *The Patient: Patient-Centered Outcomes Research* 3 (4): 257–73.

Cuyt, Annie. 1999. "How Well Can the Concept of Padé Approximant Be Generalized to the Multivariate Case?" *Journal of Computational and Applied Mathematics* 105 (1): 25–50.

Dahan, Ely, John Hauser, Duncan Simester, and Olivier Toubia. 2002. "Application and Test of Web-Based Adaptive Polyhedral Conjoint Analysis." *Center for EBusiness @ MIT. Online at Ebusiness.mit.edu*.

DiCiccio, Thomas J, and Anna Clara Monti. 2001. "Approximations to the Profile Empirical Likelihood Function for a Scalar Parameter in the Context of M-Estimation." *Biometrika* 88 (2): 337–51.

Eisen-Hecht, Jon, Randall A Kramer, Joel Huber, and others. 2005. "A Hierarchical Bayes Approach to Modeling Choice Data: A Study of Wetland Restoration Programs." *Unpublished PhD Dissertation, Nicholas School of the Environment, Duke University, Durham, NC*.

Evgeniou, Theodoros, Constantinos Boussios, and Giorgos Zacharia. 2005. "Generalized Robust Conjoint Estimation." *Marketing Science* 24 (3): 415–29.

Evgeniou, Theodoros, Massimiliano Pontil, and Olivier Toubia. 2007. "A Convex Optimization Approach to Modeling Consumer Heterogeneity in Conjoint Estimation." *Marketing Science* 26 (6): 805–18.

Fosgerau, Mogens, and Stephane Hess. 2008. "Competing Methods for Representing Random Taste Heterogeneity in Discrete Choice Models." *MPRA Paper, University*

Library of Munich, Germany.

Geerts, Paul, Guadalupe Martinez, and Andreas Schreiner. 2013. "Attitudes Towards the Administration of Long-Acting Antipsychotics: A Survey of Physicians and Nurses." *BMC Psychiatry* 13: 58.

Geyer, Charles J. 2013. "Asymptotics of Maximum Likelihood Without the Lln or Clt or Sample Size Going to Infinity." In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, 1–24. Institute of Mathematical Statistics.

Green, Paul E, and Vithala R Rao. 1971. "Conjoint Measurement for Quantifying Judgmental Data." *Journal of Marketing Research* 8 (3): 355–63.

Green, Paul E, and Venkatachary Srinivasan. 1978. "Conjoint Analysis in Consumer Research: Issues and Outlook." *Journal of Consumer Research* 5 (2): 103–23.

Hauser, John R, and Vithala R Rao. 2004. "Conjoint Analysis, Related Modeling, and Applications." In *Marketing Research and Modeling: Progress and Prospects*, 141–68. Springer US.

Heiss, Florian, Daniel McFadden, and Joachim Winter. 2010. "Mind the Gap! Consumer Perceptions and Choices of Medicare Part d Prescription Drug Plans." In *Research Findings in the Economics of Aging*, 413–81. University of Chicago Press.

Hensher, David, Jordan Louviere, and Joffre Swait. 1998. "Combining Sources of Preference Data." *Journal of Econometrics* 89 (1): 197–221.

Hiligsmann, Mickael, C van Durme, Piet Geusens, Benedict GC Dellaert, Carmen D Dirksen, TT van der Weijden, Jean-Yves Reginster, and Annelies Boonen. 2013. "Nominal Group Technique to Select Attributes for Discrete Choice Experiments: An Example for Drug Treatment Choice in Osteoporosis." *Patient Preference and*

Adherence 7: 133–39.

Huber, Joel, and Kenneth Train. 2001. "On the Similarity of Classical and Bayesian Estimates of Individual Mean Partworths." *Marketing Letters* 12 (3): 259–69.

Huber, Joel, and Klaus Zwerina. 1996. "The Importance of Utility Balance in Efficient Choice Designs." *Journal of Marketing Research* 33 (3): 307–17.

Johnson, F Reed, Brett Hauber, Corey A Siegel, Steven Hass, and Bruce E Sands. 2010. "Are Gastroenterologists Less Tolerant of Treatment Risks Than Patients? Benefit-Risk Preferences in Crohn's Disease Management." *Journal of Managed Care Pharmacy* 16 (8): 616–28.

Kamakura, Wagner A, and Gary Russell. 1989. "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure." *Journal of Marketing Research* 26: 379–90.

Kinter, Elizabeth T, Thomas J Prior, Christopher I Carswell, and John FP Bridges. 2012. "A Comparison of Two Experimental Design Approaches in Applying Conjoint Analysis in Patient-Centered Outcomes Research." *The Patient: Patient-Centered Outcomes Research* 5 (4): 279–94.

Lenk, Peter J, Wayne S DeSarbo, Paul E Green, and Martin R Young. 1996. "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs." *Marketing Science* 15 (2): 173–91.

Louviere, Jordan J, David Pihlens, and Richard Carson. 2011. "Design of Discrete Choice Experiments: A Discussion of Issues That Matter in Future Applied Research." *Journal of Choice Modelling* 4 (1): 1–8.

Luce, R Duncan, and John W Tukey. 1964. "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement." *Journal of Mathematical Psychology* 1 (1):

1–27.

McFadden, Daniel. 1973. “Conditional Logit Analysis of Qualitative Choice Behavior.” In *Frontiers in Econometrics*, 105–42. Academic Press, New York.

McFadden, Daniel L. 1976. “Quantal Choice Analaysis: A Survey.” In *Annals of Economic and Social Measurement*, 5 (4):363–90.

McFadden, Daniel, and Kenneth Train. 2000. “Mixed Mnl Models for Discrete Response.” *Journal of Applied Econometrics* 15 (5): 447–70.

Morey, Edward, and Kathleen Greer Rossmann. 2003. “Using Stated-Preference Questions to Investigate Variations in Willingness to Pay for Preserving Marble Monuments: Classic Heterogeneity, Random Parameters, and Mixture Models.” *Journal of Cultural Economics* 27 (3-4): 215–29.

Olsen, Søren Bøye, and Jürgen Meyerhoff. 2017. “Will the Alphabet Soup of Design Criteria Affect Discrete Choice Experiment Results?” *European Review of Agricultural Economics* 44 (2): 309–36.

Patefield, WM. 1977. “On the Maximized Likelihood Function.” *Sankhyā: The Indian Journal of Statistics, Series B*, 92–96.

Royall, Richard M. 1986. “Model Robust Confidence Intervals Using Maximum Likelihood Estimators.” *International Statistical Review/Revue Internationale de Statistique* 54 (2): 221–26.

Royall, Richard, and Tsung-Shan Tsou. 2003. “Interpreting Statistical Evidence by Using Imperfect Models: Robust Adjusted Likelihood Functions.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2): 391–404.

Ryan, Mandy, and Diane Skåtun. 2004. “Modelling Non-Demanders in Choice

Experiments." *Health Economics* 13 (4): 397–402.

Savage, Scott J, and Donald M Waldman. 2008. "Learning and Fatigue During Choice Experiments: A Comparison of Online and Mail Survey Modes." *Journal of Applied Econometrics* 23 (3): 351–71.

Scarpa, Riccardo, and John M Rose. 2008. "Design Efficiency for Non-Market Valuation with Choice Modelling: How to Measure It, What to Report and Why." *Australian Journal of Agricultural and Resource Economics* 52 (3): 253–82.

Schellings, Ron, Brigitte AB Essers, Alfons G Kessels, Florian Brunner, Tijmen van de Ven, and Paul BM Robben. 2012. "The Development of Quality Indicators in Mental Healthcare: A Discrete Choice Experiment." *BMC Psychiatry* 12: 103.

Sculpher, Mark, Stirling Bryan, Pat Fry, Patricia de Winter, Heather Payne, and Mark Emberton. 2004. "Patients' Preferences for the Management of Non-Metastatic Prostate Cancer: Discrete Choice Experiment." *BMJ* 328 (7436): 382.

Severini, Thomas A. 2000. *Likelihood Methods in Statistics*. Vol. 22. Oxford Statistical Science Series. Oxford University Press, New York.

Small, Christopher G. 2010. *Expansions and Asymptotics for Statistics*. Vol. 115. Monographs on Statistics and Applied Probability. Chapman Hall/CRC Press, Boca Raton.

Swait, Joffre, and Jordan Louviere. 1993. "The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models." *Journal of Marketing Research* 30 (3): 305–14.

Tibshirani, Robert, and Guenther Walther. 2005. "Cluster Validation by Prediction Strength." *Journal of Computational and Graphical Statistics* 14 (3): 511–28.

Toubia, Olivier, Theodoros Evgeniou, and John Hauser. 2007. "Optimization-Based

and Machine-Learning Methods for Conjoint Analysis: Estimation and Question Design." In *Conjoint Measurement*, 231–58. Springer US.

Train, Kenneth. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

Train, Kenneth E. 2008. "EM Algorithms for Nonparametric Estimation of Mixing Distributions." *Journal of Choice Modelling* 1 (1): 40–69.

Vermunt, Jeroen K. 2014. "Latent Class Model." In *Encyclopedia of Quality of Life and Well-Being Research*, 3509–15. Springer Netherlands.

Viveros, Román, and David A Sprott. 1987. "Allowance for Skewness in Maximum-Likelihood Estimation with Application to the Location-Scale Model." *Canadian Journal of Statistics* 15 (4): 349–61.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica: Journal of the Econometric Society* 50 (1): 1–25.

Whitty, Jennifer A, Paul A Scuffham, and Sharyn R Rundle-Thielee. 2011. "Public and Decision Maker Stated Preferences for Pharmaceutical Subsidy Decisions." *Applied Health Economics and Health Policy* 9 (2): 73–79.

Zhu, Qianqiu, and Zibin Zhang. 2009. "On Using Individual Characteristics in the Mnl Latent Class Conjoint Analysis: An Empirical Comparison of the Nested Approach Versus the Regression Approach." *Marketing Bulletin* 20: 1–12.

Curriculum Vitae

Thomas James Prior

216 Steeplechase Drive, North Wales, PA 19454

(215) 206-1641

tom.james.prior@gmail.com

Born October 2, 1987, Towson, MD

Education

- Doctor of Philosophy in Biostatistics, 2017
  - Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
  - Dissertation title: Novel Statistical Methods in Conjoint Analysis and Padé Approximation of the Profile Likelihood
  - Advisor: Dr. Charles A. Rohde
- Bachelor of Arts in Mathematics, Minor in Computer Science, 2009
  - University of Pennsylvania, Philadelphia, PA
- High School Diploma, 2005
  - North Penn High School, Lansdale, PA

Scholastic Achievements/Scholarships

- Johns Hopkins Bloomberg School of Public Health, Biostatistics Department Tuition Scholarship, 2012 - 2014

- Johns Hopkins Clinical Trials Ophthalmology Pre-Doctoral Traineeship, 2011 - 2012
- Johns Hopkins Bloomberg School of Public Health, Biostatistics Department Teaching Assistantship, 2009 - 2011
- Undergraduate Honors in Mathematics
- Undergraduate GPA: 3.7 overall, 3.9 mathematics
- High school class valedictorian

Teaching and Advising

- Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 2011 - 2013 - led discussions, graded papers and exams, held office hours for one-on-one interactions with students to answer questions, for the following courses:
  - Introduction to Biostatistics
  - Statistical Computing
  - Statistical Methods in Public Health

Employment

- Biostatistician, Johns Hopkins Bloomberg School of Public Health, Department of Health Policy and Mangagement, Dr. John F.P. Bridges. Analysis of Discrete Choice Experiments. 2013 - 2014, part-time.

Publications

- Hauber AB, González JM, Groothuis-Oudshoorn CGM, **Prior T**, Marshall DA, Cunningham C, IJzerman MJ, Bridges JFP. Statistical Methods for the Analysis of Discrete Choice Experiments: A Report of the ISPOR Conjoint Analysis Good Research Practices Task Force. Value in Health 2016;19:300-315.
- Kinter ET, **Prior T**, Carswell, CI, Bridges JFP. A comparison of two experimental design approaches in applying conjoint analysis in patient-centered outcomes research. The Patient: Patient-Centered Outcomes Research 2012;5:279-294.
- **Prior T**, Mele EJ. Comment on "A block slipping on a sphere with friction: Exact and perturbative solutions." Am J Phys 2008;76:92-93.

- **Prior T**, Mele EJ. A block slipping on a sphere with friction: Exact and perturbative solutions. Am J Phys 2007;75:423-426.

Presentations

- **Prior T**. Exploring heterogeneity in preferences for participation in a large genetic cohort study. Healthcare Applications in Conjoint Analysis, Seventeenth Sawtooth Software Conference, Dana Point, California, October 16-18, 2013.

Technical Skills

- Systems: Windows, macOS, Linux
- Statistical Software: R, SAS
- Programming languages, mathematical packages, software development: Matlab, Mathematica, C, C++, Java, Perl, SQL, Angular, Swift/iOS/watchOS
- Other: LaTeX, Word, Excel, PowerPoint, Access