TWO AVENUES OF INference: THE EVIDENTIAL PARADIGM AND RELATIVE BELIEF THEORY

by

Zeynep Baskurt

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Statistical Sciences
University of Toronto

Copyright © 2014 by Zeynep Baskurt
Abstract

Two avenues of inference: The Evidential Paradigm and Relative Belief Theory

Zeynep Baskurt

Doctor of Philosophy

Graduate Department of Statistical Sciences

University of Toronto

2014

This thesis focuses on two inferential theories, the Evidential paradigm (Royall, 1997) and (Bayesian) relative belief inferences (Evans, 1997), both of which are concerned with characterizing statistical evidence. We investigate the success of these theories in quantifying the strength of statistical evidence in a number of statistical problems.

We contribute evidential methodology for the analysis of genetic association in family data using the ratio of composite likelihoods as our evidence function. We show that composite likelihoods, with a robust adjustment, have the two crucial performance properties of the Evidential paradigm: (1) The evidence function will support the true value over the false value by an arbitrarily large factor as the sample size increases, and (2) for large samples, the probability of misleading evidence, is approximated by the bump function which has a small upper bound. Thus, in the Evidential paradigm, composite likelihood ratios can be used as a surrogate for the likelihood ratios constructed from true likelihoods when evaluating the full likelihood is not convenient.

We illustrate an approach to Bayesian point null hypothesis assessment based on relative belief inferences. We develop a methodology to assess possible bias inherent in the relative belief ratio, the evidence function, which measures the change in belief after seeing the data. We then present the application of these inferences to independent binary data in a logistic regression setting. We derive a goodness-of-fit test for logistic regression based on a weakly informative prior. When no lack of fit is detected, we show
how to use these priors to induce priors on model parameters. These priors do not have
the issues that the more commonly used priors have on the model parameters, e.g. a
common approach is to assign diffuse priors on the model coefficients $\beta$s to reflect little
information, which in fact induces an informative prior on the $p$, the mean parameter.
Dedication

This thesis is dedicated to my parents, Aynur and Salim Baskurt, whose love and prayers have been felt from across the Atlantic ocean, and to my husband, Fatih Tandogan, who has been extremely supportive and patient during my PhD study. You all mean the world to me.
Acknowledgements

I would like to express my gratitude to the many people that made the completion of this thesis possible.

My supervisor, Professor Michael Evans, was instrumental in helping me complete my thesis. Not only was he extremely generous with his time, making himself continually available; he constantly offered alternative solutions when I was stuck. His contagious enthusiasm for the subject is what allowed me to persevere and grow as a researcher.

A heartfelt thank you to Dr. Lisa Strug for introducing me to the field of genetic research. Dr. Strug not only allowed me to gain insight into the importance and impact of this area of study, she was also crucial in helping me define a framework and structure for my work and provided me with the necessary data.

Both Dr. Strug and Professor Evans have been remarkable mentors for me, without their guidance this thesis would never have come to be.

I would also like to thank my committee member, Professor Nancy Reid for her supervision in my research and and the external examiner Professor David R. Bickel for his invaluable comments on my thesis. Many thanks to all faculty members in the Department of Statistical Sciences for creating such a positive and productive environment for students. Special thanks to Professor Andrey Feuerverger for his encouragement and guidance that made my admission to the MSc degree possible in 2007, to Dr. Alison Gibbs for her invaluable guidance as a mentor during my TA work with her, to Professors Radford Neal, Jeffrey Rosenthal, Radu Craiu, Lei Sun, Fang Yao for teaching me various courses.

I would like to acknowledge the help of the administrative staff in the Department of Statistical Science, especially Andrea Carter, whose support is always there for us.

I am thankful to my colleagues and friends, especially Elif Acar who helped me to overcome certain challenges during my study, Ximing Xu who meticulously offered his help on certain topics in my work, Barbara Filion, Shivon, Yanling, Avideh, Wei, Tara,
Ramya, Muz, Alex and Liz for being supportive for several years.

Finally, I would like to thank my parents, Salim and Aynur and my brothers Mustafa and Murat for their continuous support. Special thanks to my husband who taught me to remain patient through challenging situations.
Contents

List of tables xi

List of figures xiii

1 Introduction 1

1.1 The Evidential Paradigm ........................................ 2

1.1.1 Error probabilities in EP .................................... 4

1.1.2 Characterization of the evidence function in EP: Two performance properties ........................................ 10

1.1.3 How an Evidential analysis is conducted ....................... 12

1.1.4 Robust adjusted likelihoods for wrong models ................ 14

1.1.5 Developments in Evidential Methodology and its applications . 17

1.2 Relative Belief Inferences ....................................... 19

1.2.1 Evidential properties of relative belief ratio and optimality for relative belief inferences ........................................ 22

1.2.2 The influence of priors on the inference ....................... 26

1.2.3 The overall framework of the inference ....................... 28

1.3 Thesis Outline .................................................. 29

2 A composite likelihood approach for association analysis 31

2.1 Composite likelihood ratio as a valid evidence function ........ 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.1</td>
<td>Some definitions and notations</td>
<td>32</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Composite likelihood inference in the Evidential paradigm</td>
<td>35</td>
</tr>
<tr>
<td>2.2</td>
<td>Modelling correlated binary data</td>
<td>38</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Motivating Example</td>
<td>38</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Common approaches for modelling correlated binary data</td>
<td>40</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Model set-up</td>
<td>40</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Two composite likelihood approaches</td>
<td>43</td>
</tr>
<tr>
<td>2.3</td>
<td>Simulation study</td>
<td>44</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Generating correlated binary data</td>
<td>45</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Simulation design</td>
<td>47</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Simulation results</td>
<td>51</td>
</tr>
<tr>
<td>2.3.4</td>
<td>Summary of the simulation study</td>
<td>60</td>
</tr>
<tr>
<td>2.4</td>
<td>Data example</td>
<td>60</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Our Model</td>
<td>61</td>
</tr>
<tr>
<td>2.4.2</td>
<td>Likelihood Plots</td>
<td>62</td>
</tr>
<tr>
<td>2.5</td>
<td>Summary</td>
<td>66</td>
</tr>
<tr>
<td>3</td>
<td>Hypothesis Assessment using Relative Belief Ratios</td>
<td>68</td>
</tr>
<tr>
<td>3.1</td>
<td>Evidential Interpretation of Relative Belief Ratios for Assessing a Point Null Hypothesis</td>
<td>70</td>
</tr>
<tr>
<td>3.1.1</td>
<td>Calculation of Bayes factors and relative belief ratio for a point null hypothesis</td>
<td>70</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Evidential interpretation of relative belief ratios</td>
<td>76</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Are relative belief ratios incoherent?</td>
<td>84</td>
</tr>
<tr>
<td>3.1.4</td>
<td>A Priori Properties of the Relative Belief Ratio</td>
<td>86</td>
</tr>
<tr>
<td>3.1.5</td>
<td>Assessing the Bias inherent in the Relative Belief Ratios</td>
<td>88</td>
</tr>
<tr>
<td>3.2</td>
<td>Two-way Analysis of Variance</td>
<td>89</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary and Discussion</td>
<td>95</td>
</tr>
</tbody>
</table>
List of Tables

1.1 Probabilities of misleading, weak and strong evidence in Evidential paradigm 5

2.1 A hypothetical family data set for a family study with three siblings ... 39
2.2 Different outcomes with probabilities of occurrence (Le Cessie and Van Houwelingen, 1994). 41
2.3 Simulation design ................................................................. 48
2.4 The relationship between correlations within a binary pair and odds ratio. 49
2.5 The maximum profile composite likelihood estimates of $\beta_1$ for a sibling study with $k = 5$ ............................................... 52
2.6 The maximum profile composite likelihood estimates of $\beta_1$ for a sibling study with $k \in \{2, 3, 4, 5\}$ ......................................... 53
2.7 The maximum profile composite likelihood estimates of $\beta_1$ for a family study. 54
2.8 The maximum profile composite likelihood estimates of $\psi$ for a sibling study with $k = 5$. ............................................... 54
2.9 Distribution of data at SNP rs12393509 ................................. 63

3.1 The Bayes factor and relative belief ratio for $H_0$ under $\Pi_\gamma$. ... 73
3.2 Location Normal example with $n = 50$ and $\tau^2 = 400$ ............ 81
3.3 Population values for Example 11 ............................................ 84
3.5 Values of $RB(0), \Pi_\psi(RB(\psi) \leq RB(\psi_0) \mid T(x)), \psi_{LRSE}(x)$ and $RB(\psi_{LRSE}(x))$ in various two-way analyses. ....... 95
4.1 Values of $RB(\psi_*)$ and the strength of evidence in Example 15. . . . . . . 105
4.2 Values of $RB(\psi_*)$ and the strength of evidence in Example 16 . . . . . . . 106
List of Figures

1.1 Probabilities of weak and misleading evidence for normal model mean
$(\Delta = \sigma)$ ................................. 7
1.2 Plot of the bump function for Normal means. ................................. 9
1.3 Poisson example ................................. 11
1.4 Binomial likelihood function for the probability of heads. ................. 13

2.1 Sibling study with $k = 5$. The probability of misleading evidence under
the independent and pairwise method when $n = 30$. .......................... 55
2.2 Sibling study with $k = 5$. The probability of misleading evidence under
the independent and pairwise method when $n = 500$. ........................ 56
2.3 Sibling study with $k \in \{2, 3, 4, 5\}$. The probability of misleading evidence
under the independent and pairwise method when $n = 30$. ................. 56
2.4 Sibling study with $k \in \{2, 3, 4, 5\}$. The probability of misleading evidence
under the independent and pairwise method when $n = 500$. ............... 57
2.5 Family study with $k = 3$. The probability of misleading evidence under
the independent and pairwise method when $n = 30$. .......................... 58
2.6 Family study with $k = 3$. The probability of misleading evidence under
the independent and pairwise method when $n = 500$. ........................ 58
2.7 Sibling study with $k = 5$. The probability of misleading evidence under
the pairwise method when the parameter of interest is $\delta$. ............... 59
2.8 Standardized $CL_p$ function for OR at three SNPs with 1/8 and 1/32 $CL_p$ intervals. .......................................................... 64

2.9 Evidential analysis of association between SNPs in chromosome X and MI. 66

3.1 Plot of $RB(\mu)$ against $\mu$ when $n = 50$, $\tau^2 = 400$ and $\sqrt{n}\bar{x} = 1.96$ in Example 10. ........................................ 82

4.1 Prior density of $p$ when $\beta_1, \beta_2 \sim N(0, 1)$ (on the left) and $\beta_1, \beta_2 \sim N(0, 20^2)$ (on the right) and $x = 1$. ................................. 102

4.2 Plots of the posterior density of $\psi$ and $RB(\psi)$ in Example 15 when the model is correct, $n_i = 10$, and with the 0.05, 0.10, 0.20 and 0.50 prior quantiles indicated ............................... 105

4.3 Plots of the prior (- -) and posterior densities (-) of $\psi$ and $RB(\psi)$ in Example 16 when the model is wrong and $n_i = 10$. ......................... 106

4.4 Plots of the density of $p = e^\mu/(1 + e^\mu)$ when $\mu$ is $N(0, d^2)$ and $d = 0.5 (-)$, $d = 1.0 (-)$, and $d = 1.702 (...)$. ......................... 108

4.5 Plots of the prior (- -) density, posterior (-) density and relative belief ratio of the squared distance in Example 17. ................................. 111

4.6 Plots of the marginal prior (- -) and posterior (-) densities of the $\beta_i$ coefficients in Example 17. ........................................ 111
Chapter 1

Introduction

A theory of statistical inference is fundamental to statistical science. Such a theory should provide methodology that is efficient, reliable and consistent. A key point is to find a coherent way to quantify the evidence in the observed data, as it is essential in many applications of statistics in science to have the correct evidential interpretation of data (Royall, 2004). Despite the extensive literature on the foundations of statistics, the discussions on the correct way to interpret the data are still under debate. For example, Neyman Pearson theory and the use of $P$-values as measures of strength of evidence have flawed evidential meaning (Royall, 1997, chap.2 and chap. 3) with misinterpretation of $P$ values in statistical practice Blume and Peipert (2003). Blume (2011) discusses how Neyman-Pearson hypothesis testing and significance testing ($P$-values) are wrongly merged in many application of statistics in science. In Royall (1997, chap.6) and Royall (2004), there is a debate over the appropriateness of selecting priors and the use of posterior probabilities as a measure of evidence (we address these issues in Chapter 3 and Chapter 4). According to Blume (2011), the confusions and disagreements are due to the fact that 1) the core of the problem, namely, how to interpret the data as statistical evidence, has a complex philosophical nature and 2) there has not been a commonly recognized framework developed to characterize and evaluate whether paradigms provide
a valid measure of statistical evidence or not.

In this thesis, we discuss how the statistical evidence is studied and quantified under two inferential theories, the Evidential paradigm (Royall, 1997) and (Bayesian) relative belief inferences (Evans, 1997). We believe this thesis contributes to statistical methodology which adequately addresses some of the issues in foundations of statistics. We now provide a summary of the Evidential paradigm, and the relative belief theory. An outline of the thesis is given at the end of the chapter.

1.1 The Evidential Paradigm

The Evidential paradigm (EP), developed by Royall (1997), is based on the Law of Likelihood. EP uses the likelihood ratio (LR) as an objective measure of the relative statistical evidence for a simple versus simple hypothesis. Suppose we observe \( x \) as a realization of a random variable \( X \) with a probability distribution \( \{f(\cdot; \theta), \theta \in \Theta\} \) where \( \theta \) is a fixed dimensional parameter. The Law of Likelihood dictates (Royall, 1997, p. 3):

\[
\text{If hypothesis } H_1 \text{ implies that the probability that a random variable } X \text{ takes the value } x \text{ is } f_1(x), \text{ while hypothesis } H_2 \text{ implies that the probability is } f_2(x), \text{ then the observation } X = x \text{ is evidence supporting } H_1 \text{ over } H_2 \text{ if and only if } f_1(x) > f_2(x), \text{ and the likelihood ratio, } \frac{f_1(x)}{f_2(x)}, \text{ measures the strength of that evidence (Hacking, 1965).}
\]

Then since \( L(\theta) \propto f(x; \theta) \), \( LR = L(\theta_1)/L(\theta_2) \) measures the strength of evidence in favor of \( H_1 : \theta = \theta_1 \) relative to \( H_2 : \theta = \theta_2 \). If \( LR > 1 \) then \( x \) provides evidence that is supporting \( H_1 \) over \( H_2 \), if \( LR < 1 \), \( x \) provides evidence that is supporting \( H_2 \) over \( H_1 \). If \( LR = 1 \), then the evidence is neutral between \( H_1 \) and \( H_2 \). But, when does a likelihood ratio indicate strong evidence? Royall (1997) categorized the scale of likelihood ratios as follows;

- We have strong evidence in favor of \( H_1 \) versus \( H_2 \) if \( L(\theta_1)/L(\theta_2) > k \).
• We have strong evidence in favor of $H_2$ versus $H_1$ if $L(\theta_1)/L(\theta_2) < 1/k$.

• We have weak evidence if $1/k < L(\theta_1)/L(\theta_2) < k$ (the data did not produce sufficiently strong evidence in favor of either hypothesis).

(Royall, 1997) proposed two benchmarks of $k = 8$ and $k = 32$ for fairly strong evidence and strong evidence, respectively. A common criticism is about the choice of $k$, namely, whether the likelihood ratio of $k$ represents statistical evidence of the same strength in different scenarios. Royall (1997, p.11) provided a straightforward way to understand the quantitative meaning of likelihood ratios. Suppose we have two urns, where we know that one contains all white and the other contains half white and half black balls, but we do not know which one is which. We choose one urn randomly and draw balls successively, retuning the ball to the urn after each draw. Suppose we choose three white balls successively. This might be seen as fairly strong evidence that we picked the urn with white balls only. That is, $H_1 : \theta = 1$ and $H_2 : \theta = 1/2$, where $\theta$ is the proportion of white balls in the urn chosen, then $L(\theta_1)/L(\theta_2) = 1/(1/2)^3 = 8$. Similarly, $k = 32$ refers to selecting 5 balls successively, which is strong evidence in favour of $H_1$.

Alternatively, suppose that some prior information is available on the hypotheses, $H_1$ and $H_2$, expressed by $P(\theta_1)$ and $P(\theta_2)$, respectively. Then by Bayes rule we have,

$$
\frac{P(\theta_1 | x)}{P(\theta_2 | x)} = k \frac{P(\theta_1)}{P(\theta_2)},
$$

where $k = f_1(x)/f_2(x)$. We see that a likelihood ratio of $k$ represents evidence that is strong enough to increase the prior probability ratio of $H_1$ to $H_2$ by a factor of $k$, no matter what $H_1$ and $H_2$ represent and what their dimensions and prior probabilities are. The likelihood ratio retains this interpretation in the absence of prior beliefs.

We have argued that the LR measures the strength of evidence the data provide in the EP. However, we also need to show that the EP provides methodology for measuring and controlling the error probabilities, namely, the probability of observing misleading
evidence and weak evidence. As is well known, LR is the essential element in Neyman-Pearson theory. The role of the LRs in this theory is not to represent a quantitative measure of statistical evidence, but to derive the most powerful size $\alpha$ test, where $\alpha$ (type I error) is fixed and type II error is minimized. In the next section, we illustrate the evidential error rates that are analogues to type I and type II error.

### 1.1.1 Error probabilities in EP

Suppose we have a collection of density functions $\{f_\theta : \theta \in \Theta\}$ for an observable $x$ and $L(\theta)$ is the likelihood function for $\theta$ for $n$ such observations. We are interested in quantifying the relative evidence for $\theta_1$ over $\theta_2$, i.e. $H_1: \theta = \theta_1$ and $H_2: \theta = \theta_2$. When we interpret the strength of evidence via the likelihood ratio, $L(\theta_2)/L(\theta_1)$, there are two kinds of undesirable results we could observe;

1. **Misleading evidence**: We can get strong evidence in favor of a wrong hypothesis over the correct hypothesis, i.e. $L(\theta_2)/L(\theta_1) \geq k$ when $H_1$ is true. The probability of getting misleading evidence is denoted by $M_1(n,k) = P_1(L(\theta_2)/L(\theta_1) \geq k)$, which is a function of $k$ and $n$.

2. **Weak evidence**: We can get weak evidence, that is, we fail to get strong evidence in favor of the true hypothesis. The probability of getting weak evidence is denoted by $W_1(n,k) = P_1(1/k \leq L(\theta_2)/L(\theta_1) \leq k)$ or $W_2(n,k) = P_2(1/k \leq L(\theta_2)/L(\theta_1) \leq k)$, which is also a function of $k$ and $n$.

Note that $M_1 + W_1$ is the probability of failing to find strong evidence in favor $H_1$, when $H_1$ is true. Then, $S_1 = 1 - M_1 - W_1$ is the probability of obtaining strong evidence in favor of the true hypothesis. See Table 1.1 for a tabular form of the probabilities of misleading, weak and strong evidence under each hypothesis (reproduced from Table 1 in Strug et al. (2007)).
Table 1.1: Probabilities of misleading, weak and strong evidence in the Evidential paradigm

<table>
<thead>
<tr>
<th>Truth</th>
<th>Misleading evidence</th>
<th>Weak evidence</th>
<th>Strong evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>$M_1 = P_1(L(\theta_2) / L(\theta_1) \geq k)$</td>
<td>$W_1 = P_1(1/k &lt; L(\theta_2) / L(\theta_1) &lt; k)$</td>
<td>$S_1 = 1 - M_1 - W_1$</td>
</tr>
<tr>
<td>$H_2$</td>
<td>$M_2 = P_2(L(\theta_2) / L(\theta_1) \leq 1/k)$</td>
<td>$W_2 = P_2(1/k &lt; L(\theta_2) / L(\theta_1) &lt; k)$</td>
<td>$S_2 = 1 - M_2 - W_2$</td>
</tr>
</tbody>
</table>

I. Type I and Type II error probabilities in the Neyman-Pearson theory

In Neyman-Pearson theory, the best critical region of size $\alpha$ is calculated by solving $P_1(L(\theta_2) / L(\theta_1) \geq c) = \alpha$ for $c$, and it is known that this procedure sometimes leads to a $c$ that is less than 1, which tells us to reject $H_1$ when $L(\theta_1) > L(\theta_2)$ (Blume, 2002).

In the calculation of the probability of misleading evidence $M_1$, $k$ is fixed. Now, $M_1$ is analogous to type I error $\alpha$. And since $M_2 + W_2$ is the probability of failing to find strong evidence in favor of $H_2$ when $H_2$ is true, this can be analogous to a type II error in NP theory. The probability of obtaining strong evidence in favor of $H_1$, $S_1$ is analogous to power (Royall, 1997). Blume (2002) noted that when deciding on the rejection region, if Neyman and Pearson had proposed minimizing a linear combination of the type I and type II error rates, $\min(\alpha + k\beta)$ instead of fixing $\alpha$ and minimizing $\beta$, then the rejection region of $H_1$ would be the same as in EP, that is to reject $H_1$ when $L(\theta_2) / L(\theta_1) > k$.

Next we illustrate an example where these probabilities for a given sample size are calculated.

**Example 1. Weak and misleading probabilities: Normal distribution mean (Royall, 2000)**

Consider a sample of $n$ i.i.d observations from the $N(\theta, \sigma^2)$ distribution with $\sigma^2$ known. Let $H_1 : \theta = \theta_1$ and $H_2 : \theta = \theta_2$. Then the likelihood ratio of $\theta_2$ versus $\theta_1$ is given by;

$$
\frac{L(\theta_2)}{L(\theta_1)} = \exp \left\{ \frac{n(\theta_2 - \theta_1)}{\sigma^2} \left( \bar{x} - \frac{\theta_1 + \theta_2}{2} \right) \right\}
$$
The probability of misleading strong evidence in favor of $\theta_2$ when the true value is $\theta_1$ and the probability of observing weak evidence are given by:

$$M_1(n, k) = P_1\left(\frac{L(\theta_2)}{L(\theta_1)} \geq k\right) = \Phi\left(-\frac{\sqrt{n}\Delta}{2\sigma} - \frac{\sigma \log k}{\sqrt{n}\Delta}\right)$$

$$W_1(n, k) = P_1\left(\frac{1}{k} < \frac{L(\theta_2)}{L(\theta_1)} < k\right)$$

$$= \Phi\left(-\frac{\sqrt{n}\Delta}{2\sigma} + \frac{\sigma \log k}{\sqrt{n}\Delta}\right) - M_1(n, k)$$

where $\Delta = |\theta_2 - \theta_1|$ and $\Phi$ is the standard normal cumulative distribution function.

In Figure 1.1, we see the misleading and weak probabilities as well as Type I and Type II error as sample size increases. It can be shown that $M_1$ is maximized at the sample size, $n^* = (\log k)\sigma^2/\Delta^2$ and its maximum value is $\Phi(-\sqrt{2\log k})$. At this sample size, $M_1(n^*, 8) = 0.021$ (much smaller than 0.05) and $W_1(n^*, 8) = 0.479$, thus $S_1(n^*, 8) = 0.5$. When $n = 11$, we see that both Type I and Type II error rates are 0.05. However, $W_1(11, 8) = 0.14$. If we want $W_1(n', 8) = 0.05$, then we would need $n' = 17$ observations.

It was shown that the sample sizes calculated using the Neyman-Pearson methodology are not big enough to serve the purpose of the Evidential paradigm, that is, the probability of getting strong misleading evidence is not sufficiently high. It can also be shown that $M_i(n, k)$ and $W_i(n, k) \to 0$ as $n \to \infty$. For an in-depth discussion of sample size calculations in EP, see Strug, Rohde, and Corey (2007).
It is important to emphasize that $M_i$ and $W_i$ are only part of the planning stage of the study, they don’t play a role in the interpretation of the evidence the data provide. The control on these probabilities assures that we do not observe misleading evidence or weak evidence often, thus the study design is reliable. Once the data is collected, they become irrelevant (Blume, 2002).

II. Two bounds on the probability of misleading evidence
We showed that the probability of misleading evidence can be controlled as we collect more data. In this section, we investigate whether there is a general upper bound for the probability of misleading evidence for any sample size. Suppose $X$ has density $f_1$ under $H_1$ and $f_2$ under $H_2$, the probability is calculated under the correct model defined in $H_1$.

1. The universal bound: The universal bound applies to all probability density functions (Royall, 2000). From Markov’s inequality, we obtain,

$$P_1 \left( \frac{f_2}{f_1} \geq k \right) \leq \frac{1}{k}$$
This indicates that the probability of misleading evidence is less than or equal to \(1/k\) for \(\forall n\). This is a crude upper bound. See (Royall, 2000) for some implications of this bound on the operational characteristics of the paradigm. The following section presents a smaller upper bound for the probability of misleading evidence.

2. The bound generated from the bump function:

Example 2. Cont. Example 1.

Let \(f_{2n}/f_{1n}\) represent the likelihood ratio for \(n\) iid \(Normal(\theta, \sigma^2)\) random variables. Then from Eq.(1.1) the probability of misleading evidence is

\[
P_1 \left( \frac{f_{2n}}{f_{1n}} \geq k \right) = \Phi \left( -\frac{c}{2} - \frac{\log k}{c} \right)
\]

(1.2)

where \(c = \Delta \sqrt{n}/\sigma\). Note that the distance \(\Delta\) is measured in units of the standard error and the probability of misleading evidence is independent of the sample size at a fixed \(c\). The function in Eq.(1.2) is called the “bump function” in Royall (2000). See Figure 1.2 for the behaviour of the bump function. The \(x\)-axis is the distance from the true mean to alternative in standard error units, e.g. 2 on the \(x\)-axis means \(|\theta_2 - \theta_1| = 2 \times s.e\), where \(s.e. = \sigma/\sqrt{n}\). From the figure, we see that regardless of the sample size, the bump function takes its maximum at \(\Phi(-\sqrt{2 \log k})\) when \(\Delta = (2 \log k)^{1/2}\) and this is the best possible bound (Royall, 2000).
Figure 1.2: Plot of the bump function for Normal means.

Suppose that $f$ is not Normal, but a smooth probability density of $\theta$. Then Royall (2000) proved that when $n$ is large the probability of misleading evidence converges to a bump function,

$$P_1 \left( \frac{f_{2n}}{f_{1n}} \geq k \right) \to \Phi \left( -\frac{c}{2} - \frac{\log k}{c} \right)$$

(1.3)

where $|\theta_1 - \theta_2| = c (nI(\theta_1))^{-1/2}$ and $I(\theta_1)$ is the Fisher information defined as $I(\theta_1) = -E \{[\partial^2 \log f(X; \theta)/\partial \theta^2] \}$.

We prove this property in a more general setting in Chapter 2.

The large sample bound for the probability of misleading evidence is very important as it characterizes an important aspect of the evidence function in EP as will be shown in the next section.
1.1.2 Characterization of the evidence function in EP: Two performance properties

Royall (2000,2003) defined two important performance properties an evidence function should possess in the EP. Let $L(\theta)$ be the likelihood function of $\theta$ where $X_1,\ldots,X_n$ are iid with a smooth probability model $f(\cdot;\theta)$ for $\theta \in \mathcal{R}$.

1. For any false value $\theta \neq \theta_0$, the evidence will eventually support $\theta_0$ over $\theta$ by an arbitrarily large factor:

$$P_0\{ \frac{L(\theta_0)}{L(\theta)} \to \infty \text{ as } n \to \infty\} = 1 \quad (1.4)$$

2. In large samples, the probability of misleading evidence, as a function of $\theta$, is approximated by the bump function,

$$P_0\{ \frac{L(\theta)}{L(\theta_0)} \geq k \} \to \Phi(-\frac{c}{2} - \frac{\log(k)}{c}) \quad (1.5)$$

where $k > 1$, $\Phi$ is the standard normal distribution function and $c$ is proportional to the distance between $\theta$ and $\theta_0$.

The results can be extended to the case where $\theta$ is a fixed dimensional vector parameter. The first property implies that the probability of getting strong evidence in favor of the true value goes to 1. This implies that the probability of weak evidence and strong misleading evidence go to 0 as $n \to \infty$. The second property implies that, when $n$ is sufficiently large, the probability of misleading evidence of strength $k$ is maximized at a fixed constant $\Phi(-\{2\log k\}^{1/2})$, over all $\theta$ (Royall and Tsou, 2003). Those properties ensure that with high probability we will get evidence in favor of the true value and that the probability of strong evidence in favor of a false value is low.

The following example compares the behaviour of the bump function when calculated exactly and when using Eq.(1.5) for a Poisson model.
Example 3. (Royall and Tsou, 2003)

Suppose $X \sim \text{Poisson}(\theta_0)$, where $\theta_0 = 18$ and we choose $k = 8$ (for fairly strong evidence). We calculate the probability of misleading evidence for a set of $\theta_1 = \theta_0 + c\sigma/\sqrt{n}$, where $\sigma = \theta_0$. As $\Delta = |\theta_1 - \theta_0|$ increases, the probability of misleading evidence rises from 0 to its maximum value of $\Phi\left(-\frac{2 \log k}{2}\right) = 0.021$ at $\Delta = (2 \log k)^{1/2}$ and decreases afterwards. From Figure 1.3, we observe that the bump function provides a good approximation for the discrete Poisson distribution when $n = 20$

![Figure 1.3: Poisson example](image)

Suppose now that the parameter of interest is a function of the full parameter. Let $\theta$ be the parameter of interest and $\gamma$ be the nuisance parameter. Then one can use marginal, conditional or partial likelihoods to represent evidence about $\theta$. However, when those are not available for a given model, an alternative approach is to use the profile likelihood, $L_p(\theta) = \max_{\gamma} \{L(\theta, \gamma)\}$. How do we justify that the profile likelihoods are valid evidence functions in EP? For this, Royall and Tsou (2003) proved that the profile likelihood has
the two key performance properties of a likelihood function, namely,

1. For any false value $\theta_2 \neq \theta_1$ the evidence will eventually support $\theta_1$ over $\theta_2$ by an arbitrarily large factor (Royall and Tsou, 2003):

$$P_1 \left\{ \frac{L_p(\theta_1)}{L_p(\theta_2)} \to \infty \text{ as } n \to \infty \right\} = 1.$$  

2. In large samples, the probability of misleading evidence is approximated by the bump function (Royall, 2000):

$$P_1 \left\{ \frac{L_p(\theta_2)}{L_p(\theta_1)} \geq k \right\} \to \Phi \left( -\frac{c}{2} - \frac{\log(k)}{c} \right),$$

where $|\theta_2 - \theta_1| = c(b(\theta_1, \gamma_1)/n)^{1/2}$, $b(\theta, \gamma) = 1/I_{\theta\theta}(1 - \rho_{\theta\gamma}^2)$ and $\rho_{\theta\gamma} = I_{\gamma\gamma}^{-1}I_{\gamma\theta}/I_{\theta\theta}$ and $I_{\theta\theta}$, $I_{\gamma\theta}$ and $I_{\gamma\gamma}$ are the partitions of the information matrix according to $\theta$ and $\gamma$.

The results are shown for a scalar $\theta$ and a fixed dimensional vector $\psi$. They can be extended to the case where $\theta$ is a fixed dimensional vector parameter.

In conclusion, the profile likelihood functions can be used to represent evidence in EP for large samples.

### 1.1.3 How an Evidential analysis is conducted

Suppose now that we have a collection of density functions $\{f_\theta : \theta \in \Theta\}$ for an observed $x$ and $L(\theta)$ is the likelihood function for $\theta$. A likelihood region for the true $\theta$ is a set of the form $\{\theta : LR(\theta) \geq k\}$ for some constant $k$. Royall (1997) refers to a $1/k$ likelihood region as given by $\{\theta : L(\theta)/\sup_\theta L(\theta) \geq 1/k\}$. This is the set of all parameter values that are consistent with the data in the sense that they are not better supported by the maximum likelihood estimate (MLE) by a factor of $k$ or more (Royall, 1997, p.101). We consider an example.
Example 4. Binomial model (Blume, 2002)

Suppose we are interested in the probability of getting heads on a toss of a biased coin. Let $X$ be the number of heads in $n$ tosses. Since each trial is independent, $X \sim \text{Binom}(n, p)$. We flip the coin 50 times and we observe 14 heads so the corresponding likelihood function is $L(\theta) \propto \theta^{14}(1 - \theta)^{36}$. Let $H_1 : \theta = 0.3$ and $H_2 : \theta = 0.5$, then $L(0.3)/L(0.5) = 143$ indicates that $H_1$ is better supported than $H_2$ by a factor of 143. Now, we plot the (standardized) likelihood, $L(\theta)/L(\hat{\theta})$, in Figure 1.4, where $\hat{\theta}$ is the maximum likelihood estimator (MLE).

![Figure 1.4: Binomial likelihood function for the probability of heads (reproduced from Figure 1 in Blume (2002))](image)

The 1/8 likelihood interval for the probability of heads is (0.17,0.42). Any two values within this interval can only be supported over the other in the weak sense (the likelihood ratio is less than 8). These are the parameter values that are consistent with the data because the MLE (the best supported hypothesis) is no better supported than by a factor
of 8.

What about the values outside of this interval? For any value outside, there exists a value inside (e.g. the MLE) that is better supported by a factor of 8 or more.

**Example 5.** $1/k$ likelihood intervals versus $(1 - \alpha)100\%$ confidence intervals (Blume, 2002)

This example aims to show to what values of $\alpha$ do the choices of $k = 8$ and $k = 32$ correspond in a Normal model. Suppose $X \sim N(\mu, \sigma^2)$, where $\sigma^2$ is known. The likelihood interval and $(1 - \alpha)100\%$ confidence interval for the mean $\mu$, is $\bar{x} \pm \sqrt{2 \log k \sigma / \sqrt{n}}$ and $\bar{x} \pm z_{\alpha/2} \sigma / \sqrt{n}$, respectively. Suppose we take $k = 8$, then $\sqrt{2 \log k} = 2.039$. This corresponds to a 95.9% confidence interval. If $k = 32$, then $\sqrt{2 \log k} = 2.633$, and this corresponds to approximately a 99.16% confidence interval. For a normal model, $k = 8$ and $k = 32$ roughly corresponds to $\alpha = 0.05$ and $\alpha = 0.01$. However, when there are multiple hypothesis tests, these limits become increasingly different. This is because, strictly speaking, confidence intervals should also be adjusted for multiple hypothesis tests, and thus is done by dividing $\alpha$ by the number of tests conducted so that the observed Type I error is fixed at the desired level. In the Evidential paradigm, we do not need to do any adjustment on the likelihood interval. It preserves its form in a multiple testing problem, since there is no error rate is fixed.

The next section is particularly important for this thesis as it contains the tools that advocate our proposal in Chapter 2.

### 1.1.4 Robust adjusted likelihoods for wrong models

Until now, we assumed that the working model $f$ is the correct model for $X$. What if the working model is wrong? Would the likelihood function constructed from $f$ still provide reliable evidence? Royall and Tsou (2003) proved that the likelihood ratio constructed from $f$ can continue to be a valid measure for evidential interpretation under certain
conditions. Now, we present these conditions for the LRs to be robust against model misspecification.

Suppose we have \( \{f(.; \theta), \theta \in \Theta\} \) where \( \theta \) is fixed dimensional, as the working model, and there exist a true density \( g \).

The Kullback-Liebler divergence between \( f \) and \( g \) is defined as

\[
K(f : g) = -\{E_g[\log f(.; \theta)] - E_g[\log g(.)]\}
\] (1.6)

The \( g \) subscripts in the expectations mean that the expectations are with respect to the true density \( g \). Let \( \theta_g \) be the value of \( \theta \) that maximizes \( E_g[\log f(.; \theta)] \); that is, \( \theta_g \) minimizes the Kullback-Liebler divergence between \( f \) and \( g \). Then, it can be shown that \( L(\theta_g)/L(\theta) \to \infty \) with probability 1. So the likelihood under the wrong model represents evidence about \( \theta_g \), which is referred to as the object of inference in Royall and Tsou (2003) when the true density is \( g \). Suppose we are interested in \( E_g(X) \) which is referred to as the object of interest. If our working model is wrong, then \( \theta_g \) might not be equal to \( E_g(X) \). If the object of inference \( \neq \) the object of interest then the likelihood ratio will eventually favor the wrong value \( \theta_g \) over the true value \( E_g(X) \) (Royall and Tsou, 2003) since

\[
P_g[\frac{L(\theta_g)}{L(E_g(X))} \to \infty] = 1.
\]

This is the first robustness property that \( f \) should have. We give two examples to illustrate how to assess this condition.

**Example 6.** (Royall and Tsou, 2003) *An example where the object of inference is the object of interest*

Suppose \( X \sim g \) and we choose \( f(X; \theta) \sim \text{Poisson}(\theta) \). We are interested in \( E_g(X) \)(the object of interest). Note that \( E_g(\log\{f(X; \theta)\}) = E_g(X) \log(\theta) - \theta - E_g(\log(X!)) \) is maximized at \( \theta_g = E_g(X) \). In this example, the object of inference is the object of interest, hence the likelihood ratio gives evidence about the true parameter \( E_g(X) = \theta_g \).
Example 7. (Royall and Tsou, 2003) An example where the object of inference is not the object of interest

Suppose we choose \( f(X; \theta) \sim \text{Lognormal}(\theta, \sigma^2) \) with \( \sigma^2 \) known, so \( \log(X) \sim N(\mu, \sigma^2) \).

If we reparametrize in terms of the expected value of \( X \) (the object of interest), \( E_g(X) = \theta = \exp(\mu + \sigma^2/2) \), we get

\[
\begin{align*}
f(x; \theta) &= \left(2\pi\sigma^2x^2\right)^{-1/2} \exp\left\{ \frac{\{\log(x) - \log(\theta) + \sigma^2/2\}^2}{2\sigma^2} \right\} \\
E_g(\log\{f(X; \theta)\}) \text{ is maximized at } \log(\theta) - \sigma^2/2 &= E_g\{\log(X)\}, \text{ so } \theta_g = \exp[E_g\{\log(X)\} + \sigma^2/2].
\end{align*}
\]

Thus, the object of inference is not the object of interest, hence the likelihood ratio does not provide evidence about the mean \( E_g(X) \), but about the quantity \( \theta_g = \exp[E_g\{\log(X)\} + \sigma^2/2] \).

As long as the object of inference remains the object of interest, the first performance property of the likelihood function defined in Section 1.1.2 will hold. We look at the second condition in Section 1.1.2 now, namely, whether the LR from the working model generates the bump function to characterize the large sample probability of misleading evidence. Suppose \( \theta = \theta_g + cn^{-1/2} \). A Taylor expansion of the log likelihood ratio around \( \theta_g \) yields,

\[
l(\theta) - l(\theta_g) = l'(\theta_g)c/n^{1/2} + l''(\theta_g)c^2/2n + O_p(1/n^{1/2}).
\]

Let \( a = E_g\{-\partial^2[\log\{f(X; \theta_g)\}]/\partial\theta^2\} \) and \( b = E_g\{((\partial[\log\{f(X; \theta_g)\}]/\partial\theta))^2\} \). Then it can be shown that \( l(\theta) - l(\theta_g) \to^d N(-c^2/2a, c^2b) \), and \( P_g\{L(\theta)/L(\theta_g) \geq k\} \to \Phi\{-ca/2b^{1/2} - \log(k)/cb^{1/2}\} \), which is not the bump function (Royall and Tsou, 2003).

Royall and Tsou (2003) propose adjusting the likelihood function by raising it to the power \( a/b \), then;

\[
\{l(\theta) - l(\theta_g)\}(a/b) \to^d N(-c^2/2a^2/b, c^2a^2/b)
\]
So

\[ P_g\{(L(\theta)/L(\theta_g))^{a/b} \geq \log k\} \rightarrow \Phi\{-((ca/b^{1/2})/2) - \log(k)/(ca/b^{1/2})\}. \]

To estimate \( a \) and \( b \), we replace them by their consistent estimates, that is,

\[ \hat{a} = \sum (-\partial^2\{\log\{f(X_i; \hat{\theta})\}/\partial \theta^2]\}/n) \text{ and } \hat{b} = \sum (\partial\{\log\{f(X_i; \hat{\theta})\}\}/\partial \theta)^2/n \]

where \( \hat{\theta} \) is the maximum likelihood estimate of \( \theta \).

We summarize the conditions for the likelihood function to possess so that the LR is robust against model misspecification.

1. The object of inference, \( \theta_g \), is the object of interest.

2. In large samples, the probability of misleading evidence converges to the bump function.

A useful result presented in Royall and Tsou (2003) and Stafford (1996) is that as long as the first robustness property (the object of inference is the object of interest) holds, the robust adjusted likelihood function \( L(\theta)^{\hat{a}/\hat{b}} \) has the two performance properties defined in Section 1.1.2 in large samples.

Royall and Tsou (2003) also showed that, with the robust adjustment \( a/b \), profile likelihoods can also be robustified against model misspecification. We provide a detailed proof for more general models in Chapter 2. For details on estimating \( a/b \) in profile likelihoods, we refer the reader to Royall and Tsou (2003).

1.1.5 Developments in Evidential Methodology and its applications

The Evidential Paradigm has many useful applications in statistical practice, especially in genetic studies. Likelihood ratios (LRs), in the form of lod scores (\( \log_{10} LR \)), have already been used as the basis to measure evidence in linkage analysis for several years. Strug and Hodge (2006a) develop EP methodology for linkage analysis by incorporating
operating characteristics, namely, controlling error probabilities (i.e. probability of misleading and weak evidence), as part of the linkage analysis framework. Furthermore, in Strug and Hodge (2006b), they provide a lucid solution to multiple hypothesis problem, a big problem in genetic analysis, by using the concept of decoupling error probabilities and strength of evidence embodied in the EP.

Strug et al. (2010) provide an EP approach for estimating sample size and adjusting for multiple testing in genetic association studies. They also present a useful graphical representation that illustrates association evidence across multiple SNPs in a genomic region, which we also use in Section 2.4.2.

Blume et al. (2007) extended the methods proposed in Royall and Tsou (2003) for adjusting the likelihood function under wrong model specifications to the generalized linear model regression setting. Zhang (2009a) considers EP in nonparameteric and semiparametric situations and proves that empirical likelihood ratios are viable to measure the strength of statistical evidence in the EP framework. Bickel (2008) and Zhang (2009b) (see also Zhang and Zhang (2013)) independently extended the definition of the law of likelihood to the generalized law of likelihood for interpreting evidence for composite hypotheses. Moreover, Bickel (2012) gave the theoretical background for the generalized likelihood ratio to show that it is a valid measure of strength of evidence, where the generalized likelihood ratio is the ratio of the maximized likelihoods over each composite hypothesis.

Now, suppose that we have external information that we want to incorporate into assessing evidence about a hypothesis. This leads us to, an extension of Royall’s approach, namely, relative belief inferences (Evans, 1997).


1.2 Relative Belief Inferences

Bayes factors, developed by Jeffreys (1935, 1961) as a tool for measuring the evidence in favor of a hypothesis $H_0$, play a dominant role in Bayesian analysis. Kass and Raftery (1995) provide an extensive review of Bayes factors, including discussions on the evaluation of evidence about a hypothesis using Bayes factors, on useful computation techniques in the case of intractable integrals, and on the assessment of the uncertainty about model selection. In this chapter, we discuss the relative belief ratio for measuring the evidence in favor of (or against) a hypothesis $H_0$, and its link to Bayes factors.

Suppose we have a sampling model $\{P_\theta : \theta \in \Theta\}$ on $\mathcal{X}$, and a prior $\Pi$ on $\Theta$. Let $T$ denote a minimal sufficient statistic for $\{P_\theta : \theta \in \Theta\}$ and $\Pi(\cdot | T(x))$ denote the posterior distribution of $\theta$ after observing data $x \in \mathcal{X}$. Then for a set $C \subset \Theta$, with $0 < \Pi(C) < 1$, the Bayes factor in favor of $C$ is defined by

$$BF(C) = \frac{\Pi(C | T(x))}{1 - \Pi(C | T(x))} / \frac{\Pi(C)}{1 - \Pi(C)}$$

$BF(C)$ is a measure of how our beliefs in the true value being in $C$ changed from $a$ priori to $a$ posteriori on the odds scale. A more direct measure of this belief can be achieved through the relative belief ratio. A relative belief ratio is a measure of how our beliefs in the true value being in $C$ changed from $a$ priori to $a$ posteriori on the probability scale.

**Definition 1.** The relative belief ratio for $C$ given that we have observed $T(x)$ is given by

$$RB(C) = \frac{\Pi(C | T(x))}{\Pi(C)}$$

The relative belief ratio is closely related to the Bayes factor as we have that

$$BF(C) = \frac{(1 - \Pi(C))RB(C)}{1 - \Pi(C)RB(C)}; \quad RB(C) = \frac{BF(C)}{\Pi(C)BF(C) + 1 - \Pi(C)}, \quad (1.8)$$
and \( BF(C) = RB(C)/RB(C^c) \). If we want to assess a \( H_0 \), then \( BF(H_0) \) or \( RB(H_0) \) can be used to evaluate how the observed data has changed our beliefs in the truth of \( H_0 \).

We now extend the definition of relative belief ratio and Bayes factor to the case where \( \Pi(C) = 0 \). Suppose \( \Pi \) is absolutely continuous on \( \Theta \) with respect to Lebesgue (volume) measure with density \( \pi \). Let \( C(\theta_0) \) be a neighborhood of \( \theta_0 \) with small volume and shrink ‘nicely’ to \( \{\theta_0\} \), then we get

\[
\pi(\theta_0) = \lim_{C(\theta_0) \to \{\theta_0\}} \frac{\Pi(C(\theta_0))}{Vol(C(\theta_0))}
\]

(Rudin (1974), Chapter 8.)

In fact, whenever a version of \( \pi \) exists that is continuous at \( \theta_0 \), then \( \pi(\theta_0) \) is given by this limit. We assume that all our spaces possess sufficient structure and the various mappings we consider are sufficiently smooth, so that the support measures are volume measure on the respective spaces and that any densities used are derived as limits of the ratios of measures of sets converging to points.

Suppose we have a parameter of interest \( \psi = \Psi(\theta) \) where \( \Psi : \Theta \to \Psi \) (we do not distinguish between the function and its range to save notation), and we want to assess the hypothesis \( H_0 = \Psi^{-1}\{\psi_0\} \) with \( \Pi(H_0) = 0 \) (e.g. suppose we are interested in whether there is a difference between the means of two normally distributed populations with known variances then \( \theta = (\mu_1, \mu_2) \) and \( \psi = \Psi(\theta) = \mu_1 - \mu_2 \) with \( H_0 = \{(\mu, \mu) : \mu \in \mathcal{R}\} \)).

We assume that \( P_\theta \) has density \( f_\theta \) with respect to support measure \( \mu \) on \( \mathcal{X} \), \( \Pi \) has density \( \pi \) on \( \Theta \) with respect to support measure \( \nu \) and \( \pi(. \mid T(x)) \) denotes the posterior density on \( \Theta \) with respect to \( \nu \).

Following Tjur (1974), we get simple formulas for marginal and conditional densities. Let \( J_\Psi(\theta = (\det(d\Psi(\theta))d\Psi(\theta))^{1/2}) \) be the volume distortion due to \( \Psi \) at \( \theta \) where \( d\Psi \) is the differential of \( \Psi \). Suppose \( J_\Psi(\theta) \) is finite and positive for all \( \theta \), then the prior probability measure \( \Pi_\Psi \) has density \( \pi_\Psi(\psi) \) with respect to volume measure \( \nu_\Psi \) on \( \Psi \), given by
\[
\pi_{\psi}(\psi) = \int_{\Psi^{-1}(\psi)} \pi(\theta) J_{\Psi}(\theta) \nu_{\Psi^{-1}(\psi)}(d\theta)
\] (1.9)
where \(\nu_{\Psi^{-1}(\psi)}\) is volume measure on \(\Psi^{-1}(\psi)\). Furthermore, the conditional prior density of \(\theta\) given \(\Psi(\theta) = \psi\) is
\[
\pi(\theta | \psi) = \frac{\pi(\theta) J_{\Psi}(\theta)}{\pi_{\psi}(\psi)}
\] (1.10)
where \(\theta \in \Psi^{-1}\{\psi}\).

If we let \(T : \mathcal{X} \to \mathcal{T}\) denote a minimal sufficient statistic for \(\{f_\theta : \theta \in \Theta\}\), then the density of \(T\), with respect to volume measure \(\mu_T\) on \(\mathcal{T}\) is given by
\[
f_{\theta T}(t) = \int_{T^{-1}(t)} f_\theta(x) J_{T}(x) \mu_{T^{-1}(t)}(dx)
\]
where \(\mu_{T^{-1}(t)}\) denotes volume on \(T^{-1}\{t\}\). The prior predictive density, with respect to \(\mu\), of the full data is given by
\[
m(x) = \int_{\Theta} \pi(\theta) f_\theta(x) \nu(d\theta)
\]
The prior predictive density of \(T\), with respect to \(\mu_T\), is given by
\[
m_T(t) = \int_{\Theta} \pi(\theta) f_{\theta T}(t) \nu(d\theta)
= \int_{T^{-1}(t)} m(x) J_{T}(x) \mu_{T^{-1}(t)}(dx)
\]
Since \(\pi_{\psi}(\psi | T(x))/\pi_{\psi}(\psi)\) is the density of \(\Pi_{\Psi}(\cdot | T(x))\) with respect to \(\Pi_{\Psi}\), we have that
\[
\lim_{\epsilon \to 0} \frac{\Pi_{\Psi}(C_\epsilon(\psi) | T(x))}{\Pi_{\Psi}(C_\epsilon(\psi))} = \frac{\pi_{\psi}(\psi | T(x))}{\pi_{\psi}(\psi)}
\] (1.11)
whenever \(C_\epsilon(\psi)\) shrinks ‘nicely’ to \(\{\psi\}\) as \(\epsilon \to 0\) and all densities are continuous at \(\psi\), e.g. \(C_\epsilon(\psi)\) could be a ball of radius epsilon centered at \(\psi\). Thus, \(\pi_{\psi}(\psi | T(x))/\pi_{\psi}(\psi)\) is the limit of the relative belief ratios of sets converging nicely to \(\{\psi\}\). We refer to this ratio as the relative belief ratio of \(\psi\). This extends the relative belief ratio definition in Definition 1.
Definition 2. The relative belief ratio is the limit of relative belief ratios of sets converging nicely to $\psi$ if $C = \Psi^{-1}(\psi)$ with $\Pi(C) = 0$,

$$RB(\psi) = \lim_{\epsilon \to 0} \frac{\Pi_{\Psi}(C_\epsilon(\psi) \mid T(x))}{\Pi_{\Psi}(C_{\epsilon}(\psi))} = \frac{\pi_{\Psi}(\psi \mid T(x))}{\pi_{\Psi}(\psi)} \tag{1.12}$$

From (1.8) and (1.11) we can show that, as $C$ converges nicely to $\{\psi\}$,

$$BF(C) \to \frac{(1 - \Pi(\Psi^{-1}\{\psi\}))RB(\psi)}{1 - \Pi(\Psi^{-1}\{\psi\})RB(\psi)} = RB(\psi) \quad \text{if} \quad \Pi(\Psi^{-1}\{\psi\}) = 0 \tag{1.13}$$

Thus, in the continuous case, $RB(\psi)$ is a limit of Bayes factors with respect to $\Pi$ and so can also be called the Bayes factor in favor of $\psi$ with respect to $\Pi$. We will give an implication of this on assessing point null hypothesis using Bayes factors with mixture priors in Chapter 3. Now, we present relative belief inferences and discuss many optimality properties.

1.2.1 Evidential properties of relative belief ratio and optimality for relative belief inferences

Both the Bayes factor and the relative belief ratio for $H_0 = \Psi^{-1}\{\psi_0\}$ measure the change in belief from a priori to a posteriori. The degree to which our beliefs have changed can be taken as the statistical evidence that $H_0$ is true. For if $RB(\psi_0) > 1$, then the probability of $\psi_0$ has increased by the factor $RB(\psi_0)$ after seeing the data and we have evidence in favor of $H_0$. The larger $RB(\psi_0)$ is, the more evidence we have in favor of $H_0$. Conversely, if $RB(\psi_0) < 1$, then the probability of $\psi_0$ has decreased by the factor $RB(\psi_0)$ after seeing the data and we have evidence against $H_0$. The smaller $RB(\psi_0)$ is, the more evidence we have against of $H_0$. 
This definition of evidence leads to a natural preference ordering in $\Psi$, namely, $\psi_1$ is preferred to $\psi_2$ if the increase in belief for $\psi_1$, from a priori to a posteriori, is greater than $\psi_2$, that is, $RB(\psi_1) > RB(\psi_2)$. The value that has the greatest increase in belief from a priori to a posteriori gives us an estimate of the true value of $\psi$, and is called the least relative surprise estimate, $\psi_{LRSE}(x) = \arg \sup RB(\psi)$. We can construct a $\gamma$-credible region for $\psi$ to assess the accuracy of the least relative surprise estimate, $\psi_{LRSE}(x)$ by choosing $\gamma \in (0,1)$, and looking at the 'size' of the $\gamma$-credible region for $\psi$ (Evans, 1997);

$$C_{\gamma}(x) = \{\psi_0 : \Pi_\Psi(RB(\psi) \leq RB(\psi_0) \mid T(x)) > 1 - \gamma\}$$

The form of the credible region is determined by the ordering for, if $RB(\psi_1) \geq RB(\psi_2)$ and $\psi_2 \in C_{\gamma}(x)$, then we must have $\psi_1 \in C_{\gamma}(x)$. Note that $C_{\gamma_1}(x) \subset C_{\gamma_2}(x)$ when $\gamma_1 \leq \gamma_2$ and $\psi_{LRSE} \in C_{\gamma}(x)$ for each $\gamma$ that leads to a nonempty set.

Suppose that we get $RB(\psi_0) = 20$. We know that this is evidence in favor of $\psi_0$. But how strong is the evidence? So far, we only know that this is more evidence in favor than when $RB(\psi_0) = 17$. Kass and Raftery (1995) discuss Jeffrey’s scale for interpretation of $BF(\psi_0)$, i.e. if $1<BF<3$, we don’t have evidence in favor or against, if $3<BF<10$, we have substantial evidence in favor, if $10<BF<100$, we have strong evidence in favor and if $BF>100$, we have decisive evidence in favor of $\psi_0$. However, it is difficult to see how such a universal scale is to be determined, and in any case, this does not tell us how well the data support alternatives to $H_0$. When we assess $H_0 : \Psi^{-1}\{\psi_0\}$, we can consider the relative belief ratios for other values of $\psi$, e.g. there might be a $\psi \neq \psi_0$ which produces a relative belief ratio that is much larger than that for $\psi_0$. Then how can we measure the strength of evidence $RB(\psi_0)$ provides? We propose to compare $RB(\psi_0)$ to each of the possible values of $RB(\psi)$ as part of assessing $H_0$, as opposed to just considering the hypothesis testing problem $H_0$ versus $H_0^c$. To measure the strength of evidence provided by $RB(\psi_0)$, we calculate a posterior tail probability,

$$\Pi_\Psi(RB(\psi) \leq RB(\psi_0) \mid T(x)) \quad (1.14)$$
This is the posterior probability that the true value of $\psi$ has a relative belief ratio no greater than $RB(\psi_0)$. Note that the $\gamma$-credible region for $\psi$, $C_\gamma(x) = \{\psi_0 : \Pi_\Psi(RB(\psi) < RB(\psi_0 | T(x)) \geq 1 - \gamma)\}$, and $\Pi_\Psi(RB(\psi) \leq RB(\psi_0) | T(x)) = 1 - \inf\{\gamma : \psi_0 \in C \gamma(x)\}$.

So our measure of accuracy for estimation and our measure of strength of evidence for hypothesis assessment are intimately related. Also note that the interpretation of Eq.(1.14) depends on whether we get a $RB(\psi_0)$ bigger than 1 or less than 1. If $RB(\psi_0) > 1$, so that we get evidence in favor of $H_0$, then big values of Eq(1.14) suggests that the evidence in favor of $H_0$ is strong. If $RB(\psi_0) < 1$, so that we get evidence against $H_0$, then small values of Eq(1.14) suggests that the evidence against $H_0$ is strong. In Evans (1997), Eq. (1.14) was introduced as a measure of surprise and was used as a p-value to assess a hypothesis. It was subsequently discovered, however, see Baskurt and Evans (2013) that this approach was misleading as it is not based on a valid measure of evidence. This was corrected in Baskurt and Evans (2013). Also note that Eq. (1.14) is formally equivalent to the quantity considered in Aitkin (1997). However in Aitkin (1997), the quantity does not serve as a measure of evidence, i.e. larger values are not evidence in favor of a hypothesis.

We elaborate on the interpretation of Eq (1.14), provide additional properties of $RB(\psi_0)$ as a measure of the evidence in favor of $H_0$ and derive some relevant inequalities in Chapter 3. We now give a simple example to illustrate how we assess the evidence using $RB(\psi)$ and Eq. (1.14).

**Example 8.** *Discrete prior with $\Psi$ specified (Baskurt and Evans, 2012)*

Suppose $X$ and $Y$ are independent and $X \sim \text{Binomial}(m, \theta_1)$ and $Y \sim \text{Binomial}(n, \theta_2)$ where $\Theta = \{1/2, 1/4\}^2$ and $\Pi$ is uniform on $\Theta$. Let $\psi = \Psi(\theta) = \theta_1 - \theta_2$. We wish to assess the hypothesis $\Psi(\theta) = \psi_0 = 0$. Then $\Theta_0 = \Psi^{-1}\{\psi_0\} = \{(1/4, 1/4), (1/2, 1/2)\} = \{(\omega, \omega) : \omega \in \{1/4, 1/2\}\}$ and $\Pi(\Theta_0) = 1/2$. Suppose we observe $x = 2, y = 2, m = 4$ and
Chapter 1. Introduction

$n = 5$. Then,

\[
\begin{align*}
\Pi(H_0|x = 2, y = 2) &= \Pi((1/4, 1/4) | x, y) + \Pi((1/2, 1/2) | x, y) \\
\Pi((1/4, 1/4) | x, y) &= \frac{P(x = 2, y = 2|\theta_1 = 1/4, \theta_2 = 1/4)\Pi((1/4, 1/4))}{P(x = 2, y = 2)} 
\end{align*}
\]

where

\[
\begin{align*}
P(x = 2, y = 2) &= P(x = 2, y = 2|\theta_1 = 1/4, \theta_2 = 1/4)\Pi((1/4, 1/4)) \\
&\quad + P(x = 2, y = 2|\theta_1 = 1/4, \theta_2 = 1/2)\Pi((1/4, 1/2)) \\
&\quad + P(x = 2, y = 2|\theta_1 = 1/2, \theta_2 = 1/4)\Pi((1/2, 1/4)) \\
&\quad + P(x = 2, y = 2|\theta_1 = 1/2, \theta_2 = 1/2)\Pi((1/2, 1/2))
\end{align*}
\]

Then \(\Pi((1/4, 1/4) | x = 2, y = 2) = 0.165\). Similarly \(\Pi((1/2, 1/4) | x = 2, y = 2) = 0.347\).

It follows that \(\Pi(H_0|x = 2, y = 2) = 0.512, RB(H_0) = 0.512/0.5 = 1.024\) and \(BF(H_0) = 1.049\). Both the Bayes factor and the relative belief ratio suggest that the beliefs in the truth of \(H_0\) have slightly increased after having seen the data. So we have evidence in favor of \(H_0\). But to evaluate the strength of this evidence, we calculate Eq (1.14). Possible values of \(\psi\) are \{-1/4, 0, 1/4\}. We already calculated \(RB(\{0\}) = 1.024\).

\[
RB(\{-1/4\}) = \frac{\Pi((1/4, 1/2) | x = 2, y = 2)}{\Pi((1/4, 1/2))} = \frac{0.195}{0.25} = 0.781
\]

and

\[
RB(\{1/4\}) = \frac{\Pi((1/2, 1/4) | x = 2, y = 2)}{\Pi((1/2, 1/4))} = \frac{0.293}{0.25} = 1.172
\]

Thus, \(\Pi(RB(\Psi(\theta)) \leq RB(\{0\})) | x = 2, y = 2) = \Pi((1/4, 1/4), (1/4, 1/2), (1/2, 1/2) | x = 2, y = 2) = 0.165 + 0.195 + 0.347 = 0.707\), that is, the posterior probability of obtaining a larger value of the relative belief ratio is 0.293. We conclude that we have moderate evidence in favor of \(H_0\).

One attractive property of relative belief inferences is that they are invariant under smooth reparameterizations. This is due to the fact that, if \(\omega = \Omega(\psi)\) for some 1-1, smooth function \(\Omega\), then \(RB(\omega) = RB(\psi)\) as Jacobians cancel in the numerator and
denominator. Furthermore, various optimality properties have been established for $\psi_{LRSE}$ and $C_\gamma(x)$: Evans, Guttman, and Swartz (2006) proved that the $\gamma$--relative surprise region, $C_\gamma(x)$, has the smallest prior measure among all measurable sets $C \in \Psi$ satisfying $\Pi_\psi(C|x) \geq \Pi_\psi(C_\gamma(x)|x)$. Note that in the calculation of $\gamma$-hpd regions, one usually uses densities taken with respect to the volume measure, i.e. $\pi_\Psi(\psi|x_0) = [d\Pi_\Psi(.|x)/dv](\psi)$ and the $\gamma$--hpd region is, $B_\gamma(x) = \{\psi : \pi_\Psi(\psi|x) \geq k\}$ where $k$ is chosen such that $\Pi_\Psi(B_\gamma(x)|x) \geq \gamma$. It is known that the $\gamma$--hpd regions have the smallest volume among all regions having the same posterior content. However, $\gamma$--hpd regions are not invariant under 1-1 smooth parameter transformations, whereas the $\gamma$--relative surprise regions are (see Jang (2010) for proof). Evans and Shakhatreh (2008) showed that the $\gamma$--relative surprise region maximizes both the relative belief ratio and the Bayes factor among all posterior regions having posterior probability at least $\gamma$, and these maximized values are always bounded below by 1, i.e. $RB(C_\gamma) \geq 1$ and $BF(C_\gamma) > 1$. This is a property that is not possessed by other rules for forming credible regions. So $C_\gamma(x)$ reveals the region where the increase in belief from a priori to a posteriori is maximum and as such, $C_\gamma(x)$ is letting the data speak the loudest among all such credible regions. Evans and Shakhatreh (2008) also proved that among all credible regions, $C_\gamma$ minimizes the prior probability of covering a false value and that $C_\gamma$ is unbiased because the probability of covering a false value is always bounded above by the prior probability of covering the true value. Evans and Jang (2011c) showed that the estimate $\psi_{LRSE}(x)$ is unbiased with respect to a general family of loss functions and, is either a Bayes rule or a limit of Bayes rules with respect to a simple loss function based on the prior.

1.2.2 The influence of priors on the inference

A common criticism of Bayesian analysis is the influence of the prior distribution on the inference. Since two statisticians can choose different priors for the same problem, their conclusions might be different. A prior can be chosen to reflect the analyst’s prior
knowledge and belief about the unknown parameter. In this case, it is a subjective prior.
It is known that to choose a prior or elicit a prior is still rather difficult. A more common
approach is to choose a prior which provides little information, e.g. a noninformative
prior. These priors are typically improper. In this paper we restrict attention to proper
priors although limiting results can often be obtained when considering a sequence of
increasingly diffuse priors. Moreover, there are some issues with noninformative priors;
see Example 10 in Chapter 3 for the effect of diffuse prior on the inference (Lindley’s
paradox) and Chapter 4 for how noninformative priors for the regression parameters
induce more informative priors on log-odds in a logistic regression analysis.

To address some of the issues due to priors, Evans and Moshonov (2006) provide
methodology to assess whether or not a prior-data conflict exists. A prior-data conflict
refers to the situations where the prior puts most of its mass on the distributions in
the sampling model for which the observed data is surprising. They discuss that, after
the model checking is done and the model is found to be appropriate, the statistician
should check if the prior chosen is appropriate. To assess this, they provide a P-value
which is calculated from comparing the observed value of minimal sufficient statistics
to its prior predictive distribution. They also argue that if there is a prior which never
leads to a prior-data conflict, then such a prior can be called noninformative, that is,
this methodology gives us a necessary characteristic of a noninformative prior. Jang
(2010) improved this methodology by providing a method to make this P-value invariant
under 1-1, smooth transformations of the data, that is, any choice of a form of the
minimal sufficient statistics will lead to the same P-value. Evans and Jang (2011a)
develop a method for measuring the amount of information a prior incorporates into an
analysis with respect to the base prior, where a base prior is the one that best reflects
the information available. They characterize the concept of weakly informative priors.
Thus, if we want our prior to put less information, when compared to what base prior
puts, into the analysis, we can adopt their method. This approach can be used to modify
a prior when prior-data conflict is detected.

Furthermore, Kass and Raftery (1995) provide some discussions on the sensitivity analysis of the Bayes factors with respect to the choice of priors. Sinharay and Stern (2002) present an approach for understanding the sensitivity of the Bayes factor to the prior distribution in a model selection context. They propose calculating the Bayes factor over a grid of values of the parameter that are present in the full model and absent in the nested model, e.g., the variance component in a random effects model. They then plot these Bayes factors with respect to the grid of values, and note the effect of the prior for that extra parameter on the Bayes factor.

1.2.3 The overall framework of the inference

The only objective part of a statistical analysis is the data itself, if collected under the assumptions of the appropriate analysis, e.g., if it represents a random sample of the population of interest. In a Bayesian analysis, we have two more ingredients, the prior and the statistical model, both of which our inferences depend on. For this reason, we need to check the ingredients we are introducing to the analysis against data. Below, we summarize a general framework of statistical inference, where items 3 and 4 are covered in relative belief inferences. For item 2, we have recently proposed an application of the relative belief inference on the goodness-of-fit for a logistic regression model, which will be covered in Chapter 4.

1. Collect data under the assumptions of the appropriate analysis.

2. Choose a model \( \{f_\theta : \theta \in \Theta\} \) and check for goodness-of-fit.

3. Choose a prior and check the prior (checking for prior-data conflict).

4. Calculate the relative belief ratio to assess the evidence for a hypothesis. Check whether the prior chosen induces any bias in favour or against the hypothesis.
Proceed with the hypothesis assessment approach using the relative belief inferences (these procedures are presented in Chapter 3).

The sample size plays an important role in a statistical analysis. Suppose there are two statisticians following this framework on the same data set. If they get different conclusions, one way to clear up the situation is to collect more data. After collecting more data, the analyses in item 2 and 3 become more accurate and at least one of the statistician will fail at item 2 and/or 3, that is, at least one of them will have to modify their model and/or prior.

1.3 Thesis Outline

In Chapter 1, we presented two different inferential methodologies, the Evidential paradigm and the (Bayesian) relative belief theory. A common theme of these methodologies is to define a measure of statistical evidence provided by the data.

In Chapter 2, we develop new methodology for the analysis of genetic association using family data in the EP framework. Specifically, we are interested in assessing genetic associations between a binary trait Y and gene X using observations from related individuals. Since evaluating the full likelihood is complicated due to the high dimensional binary outcome and complex dependencies within family members, we propose using the ratio of composite likelihoods as our evidence function. We prove that composite likelihoods, with a robust adjustment, have the two crucial performance properties of the evidential paradigm (defined in Section 1.1.2). We also present various simulation studies and a real data application.

In Chapter 3, we present a novel approach to Bayesian hypothesis assessment based on the relative belief ratio. We focus on assessing a point null hypothesis. A common approach to calculate a Bayes factor in this case is to assign a prior probability for $H_0$ with a prior on $H_0$ and use the mixture prior. In our approach, we show that, this causes
an inconsistency in prior assignments and we can avoid it by choosing the conditional prior on $H_0$ that is induced by the original prior assigned. In this case, there is no need to assign a prior probability for $H_0$ and the calculation of the Bayes factor is straightforward.

Our hypothesis assessment approach is based on 1) the computation of a Bayes factor/relative belief ratio, as our evidence function, 2) the computation of a measure of the strength of this evidence via a posterior tail probability, and 3) the point where the Bayes factor/relative belief ratio is maximized. We also propose a method to consider a priori properties of a Bayes factor/relative belief ratio. This method evaluates possible bias inherent in the Bayes factor/relative belief ratio when assessing the evidence for a hypothesis. We present an application to a two-way analysis.

In Chapter 4, we consider the application of the relative belief inference to univariate binary data based on the logistic regression model. We develop goodness-of-fit tests for these models based on weakly informative priors and then use these priors conditionally to make inference about aspects of the models when no lack of fit is detected. These priors are seen to resolve several issues associated with more common approaches.
Chapter 2

A composite likelihood approach to the analysis of correlated binary data in genetic association studies

In this chapter, we present an Evidential approach to the analysis of genetic association using family data when the response is a binary trait. Since evaluating the full likelihood is complicated due to the binary outcome and complex dependencies within family members, we propose using the ratio of composite likelihoods as our evidence function. Composite likelihoods are constructed by multiplying together lower dimensional likelihood objects (Lindsay, 1988). They have proved to be useful tools for inference when the full likelihood is intractable or impractical to construct.

In Section 2.1, we provide the definition and the required notation for the composite likelihood approach. Therein, we show that for composite likelihood functions, the two key performance properties of an evidence function hold (Eq. (1.4) and (1.5)). Thus, the composite likelihood ratio can be used as a surrogate of the canonical likelihood ratio to measure the strength of statistical evidence provided by data.

In Section 2.2, we discuss the motivating example and present a model for bivariate
binary data. The model set up and notations for constructing the composite likelihoods that we propose are also presented. We adopt two different composite likelihood approaches, using independent and pairwise likelihoods, for modelling multivariate binary data.

In Section 2.3, we demonstrate the theoretical results obtained in Section 2.1 via simulation under various settings, with different family structures. In Section 2.4, we illustrate the evidential analysis of genetic association between a trait and a single-nucleotide polymorphism (SNP) in a real data example.

We conclude this chapter with a summary in Section 2.5.

### 2.1 Composite likelihood ratio as a valid evidence function

#### 2.1.1 Some definitions and notations

Suppose \( Y = (Y_1, Y_2, \ldots, Y_m) \) is an \( m \)-dimensional random variable with a specified joint density function, \( f(y; \theta) \), where \( \theta \in \Theta \subset \mathcal{R}^d \) is some unknown parameter. Considering this parametric model and a set of measurable events \( \{A_k; i = 1, \ldots, K\} \), a composite likelihood is defined as

\[
CL(\theta; y) = \prod_{k=1}^{K} f(y \in A_k; \theta)^{w_k},
\]

where \( w_k, k = 1, \ldots, K \) are positive weights. The associated composite log-likelihood is denoted by \( cl(\theta; y) = \log CL(\theta; y) \) following the notation in Varin (2008). When we consider a random sample of size \( n \), the composite likelihood becomes

\[
CL(\theta; y) \sim = CL(\theta) = \prod_{i=1}^{n} CL(\theta; y_i) = \prod_{i=1}^{n} \prod_{k=1}^{K} f(y_i \in A_k; \theta)^{w_k};
\]

(2.1)

with the composite score function \( u(\theta; y) = u(\theta) = \Delta_{\theta} cl(\theta) \), where \( \Delta_{\theta} \) is the differentiation operation with respect to the parameter \( \theta \). In the following, we drop the argument
Chapter 2. A composite likelihood approach for association analysis

\( y \) for notational simplicity.

Note that the parametric statistical model \( \{ f(y; \theta), \ y \in Y \subset \mathcal{R}^m, \ \theta \in \Theta \subset \mathcal{R}^d \} \) may or may not contain the true density \( g(y) \) of \( Y \). Varin and Vidoni (2005) defined the Composite Kullback-Leibler divergence between the assumed model \( f \) and the true model \( g \) as

\[
K(g : f; \theta) = \sum_{k=1}^{K} E_g [\log g(Y \in A_k) - \log f(Y \in A_k; \theta)] w_k. \tag{2.2}
\]

This is a linear combination of the Kullback-Leibler divergence associated with individual components of the composite likelihood. In the case where \( f(y \in A_k; \theta) \neq g_k(y) \) for some \( k \), the estimating equation \( u(\theta) = 0 \) is not unbiased, i.e. \( E_g[u(\theta)] \neq 0 \ \forall \theta \). However, for the parameter value \( \theta_g \), which uniquely minimizes the composite Kullback-Leibler divergence in (2.2), \( E_g[u(\theta_g)] = 0 \) holds. Then, under some regularity conditions, the maximum composite likelihood estimator (MCLE), \( \hat{\theta}_{CL} = \arg \max_{\Theta} CL(\theta) \), converges to this pseudo-true value \( \theta_g \). Note that \( \theta_g \) depends on the choice of \( A_k \). Xu (2012) provided a rigorous proof of the \( \theta_g \)-consistency of \( \hat{\theta}_{CL} \) under model misspecification. Furthermore, when \( f(y \in A_k; \theta_0) = g_k(y) \) for all \( k \), \( \hat{\theta}_{CL} \) is a consistent estimator of the true parameter value \( \theta_0 \) (Xu, 2012).

In many practical settings, the parameter of interest is only a subset of the parameter space. In such cases, we partition \( \theta \) into \( \theta = (\psi, \lambda) \in \Theta \subset \mathcal{R}^d \), where \( \psi \in \mathcal{R}^p \) is the parameter of interest and \( \lambda \in \mathcal{R}^q \) is the nuisance parameter, with \( p + q = d \). Then, \( \hat{\theta}_\psi = (\psi, \hat{\lambda}(\psi)) \) denotes the constrained composite maximum likelihood estimator of \( \theta \) for fixed \( \psi \), and \( CL_p(\psi) \) is the profile composite likelihood function, \( CL_p(\psi) = CL(\hat{\theta}_\psi) = \max_\lambda CL(\psi, \lambda) \). The composite score function is partitioned as

\[
u(\theta) = \begin{bmatrix} u_\psi(\theta) \\ u_\lambda(\theta) \end{bmatrix} = \begin{bmatrix} \partial cl(\theta)/\partial \psi \\ \partial cl(\theta)/\partial \lambda \end{bmatrix}.
\]
Taking the expectation of its second moments, we obtain the variability matrix,

\[ J = \begin{bmatrix} J_{\psi\psi} & J_{\psi\lambda} \\ J_{\lambda\psi} & J_{\lambda\lambda} \end{bmatrix}, \]

with \( J_{\psi\psi} = E_g\{ (\partial cl(\theta; Y)/\partial \psi)^2 \} \) and \( J_{\psi\lambda} = E_g\{ (\partial cl(\theta; Y)/\partial \psi)(\partial cl(\theta)/\partial \lambda) \} \). Moreover, we have the following partitions of the sensitivity matrix \( H \), and the Godambe information \( G \), as well as their inverse matrices.

\[ H = \begin{bmatrix} H_{\psi\psi} & H_{\psi\lambda} \\ H_{\lambda\psi} & H_{\lambda\lambda} \end{bmatrix}, \quad H^{-1} = \begin{bmatrix} H_{\psi\psi} & H_{\psi\lambda} \\ H_{\lambda\psi} & H_{\lambda\lambda} \end{bmatrix}, \]

with \( H_{\psi\psi} = E_g\{ -\partial^2 cl(\theta; Y)/\partial \psi \partial \psi \} \) and \( H_{\psi\lambda} = E_g\{ -\partial^2 cl(\theta; Y)/\partial \psi \partial \lambda \} \) and,

\[ G = \begin{bmatrix} G_{\psi\psi} & G_{\psi\lambda} \\ G_{\lambda\psi} & G_{\lambda\lambda} \end{bmatrix}, \quad G^{-1} = \begin{bmatrix} G_{\psi\psi} & G_{\psi\lambda} \\ G_{\lambda\psi} & G_{\lambda\lambda} \end{bmatrix}. \]

The information in the composite score function is given by the Godambe information \( G \), which is defined as

\[ G(\theta) = H(\theta)J(\theta)^{-1}H(\theta). \tag{2.3} \]

An important issue in composite likelihood is that the specified low dimensional margins may not correspond to a joint distribution. In general, compatibility of the low dimensional margins is required in order to have a valid inference (Varin, 2008). On the other hand, when the parameter of interest is included in low dimensional margins, e.g. regression model settings, and if the pairwise likelihoods are used, it is not necessary that a joint model corresponding to low dimensional margins exists. The philosophy is similar to the one in generalized estimating equations (GEE), where a joint distribution of the data is not needed to be defined and to exist (Varin (2008), Varin, Reid, and Firth (2011)). Xu and Reid (2011) presented an example where the MCLE, \( \hat{\theta}_{CL} \), converged to the true value, even though there was no compatible joint density. A general overview of composite likelihood methods can be found in Varin, Reid, and Firth (2011).
2.1.2 Composite likelihood inference in the Evidential paradigm

The likelihood paradigm uses likelihood ratios to measure the strength of statistical evidence. That is, \( L(\theta_1)/L(\theta_2) \) measures the strength of evidence in favour of \( H_1 : \theta = \theta_1 \) relative to \( H_2 : \theta = \theta_2 \) (Hacking (1965), Royall (1997)). We propose that the composite likelihood ratios can be used as a surrogate for the real likelihood ratios to measure strength of statistical evidence provided by data. For this, we need to prove that the composite likelihood functions have the two crucial performance properties of the evidential paradigm. Note that since composite likelihoods can be seen as misspecified likelihoods, we need to derive the robust adjustment factor defined in Section 1.1.4, so that the inference becomes robust against model misspecification. Eventually, we want to show that the composite likelihood ratio with the robust adjustment can be used as the evidence function for evidential interpretation of data. As a first condition, we need to determine whether the object of inference is equal to the object of interest, which can only be checked after the working model \( f \) is chosen. Suppose the object of interest is \( E_g(Y) \) and the object of inference is \( \theta_g \). We essentially need to understand what \( \theta_g \) represents in our working model \( f \). Specifically, we need to check if \( \theta_g \) corresponds to \( E_g(Y) \). This is very important because the (composite) likelihood function constructed from \( f \) will provide evidence only about the quantity \( \theta_g \). We refer the reader to Examples 6 and 7 in Section 1.1.4 where we check analytically whether \( \theta_g \) is equal to \( E_g(Y) \). When analytical derivations are difficult, one can also conduct simulations to check this condition as illustrated in Section 2.4.

In Theorem 1, we show that the composite likelihood functions, with the robust adjustment factor, have the two important performance properties of the Evidential paradigm.

**Theorem 1.** Assume \( Y = (Y_1, Y_2, ..., Y_m) \) is a random vector from an unknown distribution \( g(y) \). The parametric model \( f(\theta; y) \) is chosen as the working model, with \( \theta \in \Theta \subset \mathcal{R} \). Let \( \theta_g \) be the (unique) minimizer of the composite Kullback-Leibler divergence between \( f \)
and $g$. Assume $Y_1, \ldots, Y_n$ is $n$ independent and identically distributed observations from the model $g(.)$. Under regularity conditions on the component log densities in Appendix A, the following properties hold.

(a) For any value $\theta \neq \theta_g$, the evidence will eventually support $\theta_g$ over $\theta$ by an arbitrarily large factor;

$$
P_g \left\{ \frac{CL(\theta_g)}{CL(\theta)} \rightarrow \infty \text{ as } n \rightarrow \infty \right\} = 1. \tag{2.4}$$

(b) In large samples, the probability of misleading evidence, as a function of $\theta$, is approximated by the bump function,

$$
P_g \left\{ \left[ \frac{CL(\theta)}{CL(\theta_g)} \right]^{a/b} \geq k \right\} \rightarrow \Phi(-c/2 - \log(k)/c), \tag{2.5}$$

where $k > 1$, $\Phi$ is the standard normal distribution function, $c$ is proportional to the distance between $\theta$ and $\theta_g$, $a = E_g(\Delta_g u(\theta_g; Y))$ and $b = Var_g(u(\theta_g; Y))$.

The results can be extended to the case where $\theta$ is a fixed dimensional vector parameter.

**Proof.** (a) The composite likelihood function for $n$ observations is $CL(\theta) = \prod_{i=1}^n CL(\theta; y_i)$.

Let $R_n = \prod_{i=1}^n CL(\theta_g; y_i)/\prod_{i=1}^n CL(\theta; y_i)$. We want to show $R_n \rightarrow \infty$.

Let $cl(\theta; y) = \log CL(\theta; y)$.

$$
\log \left( \frac{\prod_{i=1}^n CL(\theta_g; y_i)}{\prod_{i=1}^n CL(\theta; y_i)} \right)^{1/n} = \frac{1}{n} \left( \sum_{i=1}^n cl(\theta_g; y_i) - \sum_{i=1}^n cl(\theta; y_i) \right)
= \frac{1}{n} \left( \sum_{k=1}^K w_k \left\{ \sum_{i=1}^n \log(f(y_i \in A_k; \theta_g)) - \sum_{i=1}^n \log(f(y_i \in A_k; \theta)) \right\} \right)
\rightarrow a.s. \sum_{k=1}^K w_k \{ E_g[\log(f(Y \in A_k; \theta_g)) - \log(f(Y \in A_k; \theta))] \} > 0 \tag{2.6}
$$

...
since \( w_k \)'s are positive and \( \theta_g \) minimizes the \( K(g : f; \theta) \). (2.6) is true by the Strong Law of Large Numbers. Since \( (1/n) \log R_n \to c > 0 \), where \( c \) is a finite positive number, then \( R_n = \prod_{i=1}^{n} CL(\theta_g, y_i)/\Pi_{i=1}^{n} CL(\theta, y_i) \to \infty \).

(b) See Appendix B.

Note that we can substitute the \( a \) and \( b \) terms with the consistent estimates \( \hat{a} = n^{-1} \sum_{i=1}^{n} u(\hat{\theta}_{CL}; Y_i) \) and \( \hat{b} = n^{-1} \sum_{i=1}^{n} (u(\hat{\theta}_{CL}; Y_i))^2 \).

The first property implies that the probability of getting strong evidence in favor of the true value goes to 1 as the sample size, \( n \) increases. The second property implies that, when \( n \) is large, the probability of misleading evidence of strength \( k \) is maximized at a fixed constant \( \Phi(-\{2 \log k\}^{1/2}) \), over all \( \theta \) (Royall and Tsou, 2003). See Example 3 in Chapter 1 for the behaviour of the bump function. Those properties ensure that with high probability we will get evidence in favor of the true value and that the probability of strong evidence in favor of a false value is low.

Suppose \( \theta \in \Theta \subset \mathcal{R}^d \) is partitioned as \( \theta = (\psi, \lambda) \) and \( \psi \) is a parameter of interest. In some situations, it is possible to obtain a marginal or a conditional likelihood to be used as an evidence function. When they are not available, an alternative approach is to use the profile likelihood. It was shown in Royall(2000) that the large-sample bound for the probability of misleading evidence, \( \Phi(-\{2 \log k\}^{1/2}) \) still holds for profile likelihoods. In Theorem 2, we show that the profile composite likelihood also has these two properties.

**Theorem 2.** Assume \( Y = (Y_1, Y_2, ..., Y_m) \) is a random variable from an unknown distribution \( g(y) \), the model \( f(\theta; y) \) is the assumed model, with \( \theta \in \theta \subset \mathcal{R}^2 \) partitioned as \( \theta = (\psi, \lambda) \) and \( \psi \) is a parameter of interest. Let \( Y_1, ..., Y_n \) be \( n \) independent and identically distributed observations from the model \( g(.) \). Under regularity conditions on the component log densities in Appendix A, the following properties hold.

(a) For any false value \( \psi \neq \psi_g \), the evidence will eventually support \( \psi_g \) over \( \psi \) by an
arbitrarily large factor;

\[
\frac{CL_p(\psi; y)}{CL_p(\psi_g; y)} \rightarrow^p \infty \quad n \to \infty \quad (2.7)
\]

(b) In large samples, the probability of misleading evidence, as a function of \( \psi \), is approximated by the bump function,

\[
P_g \left\{ \left[ \frac{CL_p(\psi; y)}{CL_p(\psi_g; y)} \right]^{a/b} \geq k \right\} \to \Phi(-c^*/2 - \log(k)/c^*) \quad (2.8)
\]

where \( k > 1 \), \( \Phi \) is the standard normal distribution function, \( c^* = ca/b^{1/2} \), \( c \) is proportional to the distance between \( \psi \) and \( \psi_g \), \( a = H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \) and \( b = H^{\psi\psi}(\psi_g, \lambda_g)^{-1}G^{\psi\psi}(\psi_g, \lambda_g)H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \).

**Proof.** See Appendix B. \( \square \)

The results can be extended to the case where \( \psi \) and \( \lambda \) are fixed dimensional vector parameters. Again, we substitute the \( a \) and \( b \) terms with consistent estimates. Note that the adjustment factor \( a/b \) simplifies to \( H^{\psi\psi}/G^{\psi\psi} \) since we assume \( \psi \) is a scalar. This ratio is equal to the adjustment factor proposed by Pace et al. (2011) in order to get a composite likelihood ratio test converging to a \( \chi^2 \) distribution instead of converging to \( \sum \nu_i \chi^2_{(1)i} \), where \( \nu_i \)'s are the eigenvalues of the matrix \( (H^{\psi\psi})^{-1}G^{\psi\psi} \).

### 2.2 Modelling correlated binary data

#### 2.2.1 Motivating Example

Our main motivation is genetic association studies for family data. When the observations are independent, a common approach for genetic association studies is to use generalized linear models (GLM). Sometimes genetic data from families are collected for linkage analysis. This type of data can be used for association analysis and conducting population based association analysis rather than family based analysis can increase the statistical
Chapter 2. A composite likelihood approach for association analysis

power when using the whole perigree (Browning et al., 2005). When related individuals are involved in the analysis, GEE or generalized linear mixed effects models are commonly used in a frequentist framework. Here we propose applying the Evidential paradigm with composite likelihood ratios to genetic association studies when we have binary data with correlated outcomes.

A hypothetical data set for a family study with three siblings is illustrated in Table 2.1. The phenotype $Y$ is a binary response, taking $Y = 1$ if the individual has the disease, $Y = 0$ otherwise. The SNP genotypes, $X \in \{0, 1, 2\}$, represent the number of minor allele of the selected SNP.

Table 2.1: A hypothetical family data set for a family study with three siblings. Individual ID 1 and 2 represent "mother" and father", respectively and 3, 4 and 5 represent offspring.

| Family ID | Individual ID | Phenotype(Y) | SNP1 | SNP2 | ...
|-----------|---------------|--------------|------|------|-----
| 1         | 1             | 1            | 0    | 1    | ...
| 1         | 2             | 0            | 2    | 0    | ...
| 1         | 3             | 1            | 1    | 1    | ...
| 1         | 4             | 1            | 1    | 2    | ...
| 1         | 5             | 0            | 0    | 2    | ...
| 2         | 1             | 0            | 1    | 0    | ...
| 2         | 2             | 0            | 1    | 0    | ...
|           | ...           | ...          | ...  | ...  | ...

The main difficulty to model the data in Table 2.1 is to define a joint probability model.
2.2.2 Common approaches for modelling correlated binary data

In general, constructing a fully specified probabilistic model for correlated binary data is challenging. A joint probability mass function (pmf) for correlated binary variables was first proposed by Bahadur (1961). Using the Bahadur representation, the joint probabilities are written as functions of marginal probabilities and second and higher order correlations. Although the Bahadur representation provides a tractable expression of a pmf, it has some drawbacks. A practical approach when using the Bahadur’s model is to assign 0 to higher order correlations and consider only two-way correlations in the model. However, when the higher order correlations are excluded from the model, the correlation between two responses becomes highly constrained so that the pmf remains valid. Bahadur (1961) discussed the restrictions on the parameter space when the third and higher order correlations are set to 0 (Molenberghs and Verbeke, 2005, chap. 7).

In Zi (2010), a full likelihood function based on the Bahadur representation and a pairwise likelihood function were used to investigate the properties of composite likelihood inference for correlated binary data under a frequentist framework. It was shown that the pairwise likelihood approach provides more accurate estimates, has higher efficiency and is also more robust to model misspecification. Other approaches for modelling the joint pmf for correlated binary data include constructing multivariate probit models or Dale models. (Molenberghs and Verbeke, 2005). However, these are computationally intensive hence intractable in high dimensional data. To overcome this challenge, we propose using composite likelihoods constructed from lower dimensional margins using 1) the working independence assumption and 2) the pairwise approach.

2.2.3 Model set-up

Suppose there are N independent clusters (e.g. families), with \( n_i \) observations in the \( i^{th} \) cluster. Let \( \mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i}) \) be a binary response for the \( i^{th} \) cluster, where \( Y_{ij} \)
indicates whether the individual \( j \) in the \( i^{th} \) family has the trait or not. Similarly the genotype data vector at a particular SNP is defined as \( X_i = (X_{i1},...,X_{in_i}) \). It is common to assume an underlying logistic regression model with an additive effect of the genotype on a binary response as in Eq.(2.9),

\[
\log \frac{p_{ij}}{1-p_{ij}} = \beta_0 + \beta_1 x_{ij}, \quad (2.9)
\]

where \( p_{ij} = P(Y_{ij} = 1|x_{ij}) = E(Y_{ij}|x_{ij}) \) is the marginal probability that the individual \( ij \) has the disease trait given \( x_{ij} = 0, 1 \) or 2.

Suppose \( Y_1 = (Y_{i1}, Y_{i2}) \) with \( i = 1, ..., N \) where observations within pairs are correlated but observations from different pairs are independent. Let \( p_{i1} \) and \( p_{i2} \) be the marginal outcome probabilities (see Table 2.2) each of which has an underlying logistic regression model in Eq. (2.9).

Table 2.2: Different outcomes with probabilities of occurrence (Le Cessie and Van Houwelingen, 1994).

<table>
<thead>
<tr>
<th>( Y_{i2} )</th>
<th>( Y_{i2} = 1 )</th>
<th>( Y_{i2} = 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{i1} = 1 )</td>
<td>( p_{i11} )</td>
<td>( p_{i10} )</td>
</tr>
<tr>
<td>( Y_{i1} = 0 )</td>
<td>( p_{i01} )</td>
<td>( p_{i00} )</td>
</tr>
<tr>
<td>( p_{i2} )</td>
<td>( 1 - p_{i2} )</td>
<td>( 1 )</td>
</tr>
</tbody>
</table>

The log-likelihood function for \( N \) pairs is,

\[
l(\beta_0, \beta_1, \psi; y) = \sum_{i=1}^{N} \left( Y_{i11} \log p_{i11} + Y_{i10} \log p_{i10} + Y_{i01} \log p_{i01} + Y_{i00} \log p_{i00} \right) \quad (2.10)
\]

where \( \psi \) is the parameter for quantifying the dependence between related pairs (Le Cessie and Van Houwelingen, 1994) and,

\[
Y_{ijk} = \begin{cases} 
1 & Y_{i1} = j \text{ and } Y_{i2} = k, \\
0 & otherwise, 
\end{cases} \quad (2.11)
\]
Chapter 2. A composite likelihood approach for association analysis

e.g. \(Y_{i11} = 1\) if \(Y_{i1} = 1\) and \(Y_{i2} = 1\) or \(Y_{i10} = 1\) if \(Y_{i1} = 1\) and \(Y_{i2} = 0\), etc.

Moreover \(p_{ijk} = Pr(Y_{ijk} = 1 \mid x_i)\), e.g. \(p_{i11} = P(Y_{i1} = 1, Y_{i2} = 1 \mid x_i)\), where \(x_i\) is the covariate vector, \(x_i = (x_{i1}, x_{i2})\). Note that \(p_{i01}, p_{i10}\) and \(p_{i00}\) can be written in terms of \(p_{i11}\),

\[
p_{i10} = p_{i1} - p_{i11},
\]
\[
p_{i01} = p_{i2} - p_{i11},
\]
\[
p_{i00} = 1 - p_{i1} - p_{i2} + p_{i11}.
\]

There are different approaches to quantify the dependence between a pair of binary observations. One approach is to quantify the dependence using the correlation between \(Y_{i1}\) and \(Y_{i2}\), however, the correlation gets constrained depending on the marginal probabilities, \(p_{i1}\) and \(p_{i2}\) (Prentice, 1988). Here, we describe the association among a 2 \(\times\) 2 table (Table 2.2) via the odds ratio (Dale, 1986).

\[
\psi = \frac{p_{i11}p_{i00}}{p_{i10}p_{i01}}. \tag{2.12}
\]

Since we assume a common odds ratio across any pairs, we drop the subscripts for \(\psi\). It can be seen that Eq.(2.12) is the ratio of the odds \(Y_{i1} = 1\) given that \(Y_{i2} = 1\) and the odds of \(Y_{i1} = 1\) given that \(Y_{i2} = 0\), which is interpreted as the odds of concordant pairs to discordant pairs.

\[
\psi = \frac{Pr(Y_{i1} = 1 \mid Y_{i2} = 1, x_i)/Pr(Y_{i1} = 0 \mid Y_{i2} = 1x_i)}{Pr(Y_{i1} = 1 \mid Y_{i2} = 0, x_i)/Pr(Y_{i1} = 0 \mid Y_{i2} = 0 x_i)}
\]
\[
= \frac{Pr(Y_{i1} = 1, Y_{i2} = 1, x_i)/Pr(Y_{i1} = 0, Y_{i2} = 1, x_i)}{Pr(Y_{i1} = 1, Y_{i2} = 0, x_i)/Pr(Y_{i1} = 0, Y_{i2} = 0, x_i)}
\]
\[
= \frac{p_{i11}p_{i00}}{p_{i10}p_{i01}} \quad \text{(Using Table 2.2 notation)}
\]
\[
= \frac{p_{i11}(1 - p_{i1} - p_{i2} + p_{i11})}{(p_{i1} - p_{i11})(p_{i2} - p_{i11})}.
\]

The joint probability of \(p_{i11}\) is written in terms of the marginal probabilities, \(p_{i1}\) and \(p_{i2}\) and the odds ratio, \(\psi\). (Plackett (1965).
Chapter 2. A composite likelihood approach for association analysis

\[ p_{i1} = \begin{cases} 
1 + (p_{i1} + p_{i2})(\psi - 1) - S(p_{i1}, p_{i2}, \psi) \\
2(\psi - 1) & \text{if } \psi \neq 1, \\
p_{i1}p_{i2} & \text{if } \psi = 1,
\end{cases} \tag{2.13} \]

where

\[ S(p_{i1}, p_{i2}, \psi) = \sqrt{1 + (p_1 + p_2)(\psi - 1)^2 + 4\psi(1-\psi)p_1p_2}, \]

for \( p_{i1}, p_{i2} \in (0, 1), \psi \geq 0 \). We use \( \delta = \log \psi \in \mathcal{R} \) in the computations to remove the restriction on the parameter space of \( \psi \). If \( Y_{i1} \) and \( Y_{i2} \) are independent then \( \psi = 1 \).

Thus, the pairwise likelihood in (2.10) is a function of marginal probabilities, \( p_{i1} \) and \( p_{i2} \) (Eq.(2.9)) and the parameter \( \psi \) which characterizes the dependence between pairs in odds scale.

### 2.2.4 Two composite likelihood approaches

We propose modelling high dimensional binary \( Y_i \) using composite marginal likelihoods in two different ways:

1. The composite likelihood constructed from independent marginals:

   This is the simplest composite marginal likelihood to construct. It is useful if we are interested only in marginal parameters (Varin et al., 2011). The composite likelihood constructed under the working independence assumption is,

   \[ CL_{\text{ind}}(\beta_0, \beta_1) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} P(Y_{ij} = y_{ij} | x_{ij}) \]

   \[ = \prod_{i=1}^{N} \left[ \prod_{j=1}^{n_i} (p_{ij})^{y_{ij}} (1 - p_{ij})^{1-y_{ij}} \right], \tag{2.14} \]

   where \( p_{ij} = \exp(\beta_0 + \beta_1 x_{ij})/(1 + \exp(\beta_0 + \beta_1 x_{ij}) \). The parameter of interest is \( \beta_1 \).

   We can determine the profile composite likelihood \( CL_p(\beta_1) = \max_{\beta_0} \{ L(\beta_0, \beta_1) \} \) and compute the profile composite likelihood estimate, \( \hat{\beta}_{1CL_{\text{ind}}} = \max_{\beta_1} \log CL_{\text{ind}}(\beta_0, \beta_1) \).

2. Composite likelihood constructed from pairwise marginals: When the dependence parameter is also of interest, we need to construct the composite likelihood from
pairwise (or higher order) likelihood components (Varin et al., 2011). The composite likelihood constructed from pairwise likelihood components is,

\[ CL_{\text{pair}}(\beta_0, \beta_1, \psi) = \prod_{i=1}^{N} \left[ \prod_{j=1}^{n_i-1} \prod_{k=j+1}^{n_i} P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | x_i) \right]^{1/(n_i - 1)}, \quad (2.15) \]

where \( \prod_{j=1}^{n_i-1} \prod_{k=j+1}^{n_i} P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | x_i) \) denotes the pairwise likelihood for the \( i \)th family (Eq. (2.10)). The weight \( 1/(n_i - 1) \) is used to weigh the contribution of each family according to its size when the parameter of interest is the marginal parameter, since each observation in a family of size \( k \) presents in \( k - 1 \) pairs (Zhao and Joe, 2005).

When the parameter of interest is \( \beta_1 \), we can determine the profile composite likelihood \( CL_p(\beta_1) = \max_{\beta_0, \beta_1} \{ CL_{\text{pair}}(\beta_0, \beta_1, \psi) \} \) and compute the profile composite likelihood estimate, \( \hat{\beta}_1^{CL_{\text{pair}}} = \max_{\beta_1} \log CL_p(\beta_0, \beta_1, \psi) \). However, when we want to get inference about \( \psi \), we use the composite likelihood in Eq. (2.15) without the weights for families. We do not need to use the weights since the dependence parameter \( \psi \) appears the right amount of times in the pairwise likelihood function for a family of size \( k \), where there are \( k - 1 \) pairs.

\[ CL_{\psi_{\text{pair}}}^{\psi}(\beta_0, \beta_1, \psi) = \prod_{i=1}^{N} \left[ \prod_{j=1}^{n_i-1} \prod_{k=j+1}^{n_i} P(Y_{ij} = y_{ij}, Y_{ik} = y_{ik} | x_i) \right] \quad (2.16) \]

and calculate \( CL_p(\psi) = \max_{\beta_0, \beta_1} \{ CL_{\psi_{\text{pair}}}^{\psi}(\beta_0, \beta_1, \psi) \} \).

### 2.3 Simulation study

In this section, we investigate the performance of the composite likelihood ratio as an evidence function. We consider three different family structures:

1. **Sibling study with \( k = 5 \)**: Data consist of only siblings, where the number of siblings is 5 in each family.

2. **Sibling study with \( k \in \{2, 3, 4, 5\} \)**: Data consist of only siblings, where the number of siblings is 2, 3, 4 or 5 in each family.
3. **Family study with \( k = 3 \):** Data consist of nuclear families with 3 siblings, i.e., 2 parents and 3 offspring.

### 2.3.1 Generating correlated binary data

We use a simulation program called SIMLA (Schmidt et al., 2005) to generate genotype data \((X)\) for families under three different family structures. To generate the binary phenotype \(Y\) given \(X\), we use the method of Emrich and Piedmonte (1991) in order to have control over the dependence parameter, \(\psi\). This method uses a discretised normal approach to generate correlated binary variates with specified marginal probabilities and pairwise correlations. Suppose we want to generate a \(k\)-dimensional correlated binary vector, \(Y = (Y_1, \ldots, Y_k)\) given \(X = (X_1, \ldots, X_k)\), such that \(E[Y_i \mid x_i] = p_i\) for \(i = 1, \ldots, k\), and the pairwise correlation, \(\text{corr}(Y_i, Y_j \mid x_{ij}) = \delta_{ij}\) where \(x_{ij} = (x_i, x_j)\) is

\[
\delta_{ij} = \frac{p_{ij} - p_ipj}{(p_i(1 - p_i)p_j(1 - p_j))^{1/2}}, \tag{2.17}
\]

where \(p_{ij} = E[Y_iY_j \mid x_{ij}] = Pr(Y_i = 1, Y_j = 1 \mid x_{ij})\) for \(i = 1, \ldots, k - 1\) and \(j = 2, \ldots k\).

Now, let \(Z = (Z_1, \ldots, Z_k)\) be a standard multivariate random variable with mean 0 and correlation matrix \(\Sigma = (\rho_{ij})\) with \(i = 1, \ldots, k - 1\) and \(j = 2, \ldots k\). Then set \(Y_i = 1\) if \(Z_i \leq z(p_i)\) and set \(Y_i = 0\) otherwise for \(i = 1, \ldots, k\), where \(z(p_i)\) is the \(p_i^{th}\) quantile of the standard normal distribution. This leads to

\[
E[Y_i \mid x_i] = Pr(Y_i = 1 \mid x_i) = Pr(Z_i \leq z(p_i)) = p_i, \tag{2.18}
\]

and

\[
E[Y_i Y_j \mid x_{ij}] = Pr(Y_i = 1, Y_j = 1 \mid x_{ij}) = Pr(Z_i \leq z(p_i), Z_j \leq z(p_j)) = \Phi(z(p_1), z(p_2), \rho_{ij}),
\]

where

\[
\Phi(z(p_1), z(p_2), \rho_{ij}) = \int_{-\infty}^{z(p_1)} \int_{-\infty}^{z(p_2)} f(z_1, z_2, \rho) dz_1 dz_2,
\]
and \( f(z_1, z_2, \rho) \) is the probability density function of a standard bivariate normal random variable with mean \( \mathbf{0} \) and correlation coefficient \( \rho \).

Note from Eq.(2.17) that \( E[Y_iY_j \mid x_{ij}] = p_ip_j + (p_i(1-p_i)p_j(1-p_j))^{1/2} \). Then,

\[
\Phi(z(p_1), z(p_2), \rho_{ij}) = p_ip_j + (p_i(1-p_i)p_j(1-p_j))^{1/2}.
\] (2.19)

To solve Eq.(2.19) for \( \rho_{ij} \), Emrich and Piedmonte (1991) suggested using a bisection technique. This method does not ensure that the pairwise probabilities \( p_{ij} \), or the correlation matrix composed of binary correlations \( \delta_{ij} \) are valid. Below are the compatibility conditions that are needed to be checked (Leisch et al., 1998);

1. \( 0 \leq p_i \leq 1 \) for \( i = 1, \ldots, k \).

2. \( \max(0, p_i + p_j - 1) \leq p_{ij} \leq \min(p_i, p_j) \) for \( i \neq j \).

3. \( p_i + p_j + p_l - p_{ij} - p_{il} - p_{jl} \leq 1 \) for \( i \neq j, j \neq l, l \neq i \).

These conditions are necessary in order to get a nonnegative joint mass function for \( \mathbf{Y} \) (Emrich and Piedmonte, 1991).

In Section 2.2.3, we define the bivariate joint probability, \( p_{ij} \), in terms of the marginal probabilities \( p_i \) and \( p_j \), and a common odds ratio \( \psi \). However, for this simulation method, we need to use pairwise correlations \( \delta_{ij} \) instead of \( \psi \). The relationship between \( \delta_{ij} \), and \( \psi \) can be easily established when we plug Eq.(2.13) in Eq.(2.17). A numerical evaluation of this relationship is given in Table 2.4.

Below, we provide a step-by-step summary of the simulation algorithm for generating a \( k \)-dimensional binary vector:

1. Set \( \beta_0, \beta_1 \) and \( \psi \) values. Determine the marginal probabilities and second order probabilities given in Eq.(2.9) and Eq.(2.13). Check the compatibility conditions. Calculate pairwise correlations \( \delta_{ij} \) between each binary pair.

2. Calculate \( z(p_i) \) from Eq.(2.18).
3. Solve Eq. (2.19) to obtain the elements of the correlation matrix $\Sigma = (\rho_{ij})$ for the multivariate normal distribution.

4. Generate a $k$-dimensional multivariate normal vector $\mathbf{Z}$ with mean $\mu_i$ and correlation matrix $\Sigma = (\rho_{ij})$.

5. Set $y_i = 1$ if $z_i \leq \mu_i$, and $y_i = 0$ otherwise, for $i = 1, ..., k$.

We check the compatibility conditions in Step 1 using the R package \textit{bindata} (Leisch et al. (2012)), and for Steps 2–5, we used the R package \textit{mvtnormEP} (By and Qaqish (2011)) in the R Statistical Software. Other methods for generating correlated data can be found in Jin (2010).

Note that in the sibling study with $k = 5$ and $k \in \{2, 3, 4, 5\}$, there is one common $\psi$ that quantifies the dependence between siblings. In the family study with $k = 3$, there are two values assumed for the dependence parameter; $\psi_1$ which quantifies the dependence between siblings and $\psi_2$ which quantifies the dependence between a parent and an offspring. This should be taken into account in Step 1.

Another important point is that this method generates correlated binary data that satisfy the first and second order marginal distributions. There might be more than one joint distribution that generate the same lower dimensional marginal distributions, in which case the inference from a composite likelihood approach will be the same for that family of distributions (Varin et al., 2011). This property of composite likelihood inference is viewed as being robust by many authors (Xu (2012), Varin et al. (2011), Jin (2010)).

### 2.3.2 Simulation design

We follow a simulation setting similar to the one in Jin (2010). We keep the regression parameters constant at $\beta_0 = -1$ and $\beta_1 = 2$ throughout our simulations. Given these values, we get the following marginal probabilities using Eq. (2.9), given genotype $X$,
Chapter 2. A composite likelihood approach for association analysis

$P(Y_i = 1|x_i = 0) = 0.27$, $P(Y_i = 1|x_i = 1) = 0.73$ and $P(Y_i = 1|x_i = 2) = 0.95$. The genotype data $X$ are generated in SIMLA with a minor allele frequency of 0.25. Table 2.3 presents our simulation settings under each family structure. The model parameters used to generate the correlated binary data are given in the second column. The third column gives the parameter of interest, and the fourth column specifies the composite likelihood approach undertaken.

Table 2.3: Simulation design

<table>
<thead>
<tr>
<th>Family structure</th>
<th>Model Parameters</th>
<th>Parameter of Interest</th>
<th>CL approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sibling study</td>
<td>$\beta_0 = -1$</td>
<td>$\beta_1$</td>
<td>Independent and pairwise</td>
</tr>
<tr>
<td></td>
<td>$\beta_1 = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>with $k = 5$</td>
<td>$\psi = 1, 2, 3, 6$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sibling study</td>
<td>$\beta_0 = -1$</td>
<td>$\psi$</td>
<td>Pairwise</td>
</tr>
<tr>
<td>with $k \in {2, 3, 4, 5}$</td>
<td>$\beta_1 = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\psi = 3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family study</td>
<td>$\beta_0 = -1$</td>
<td>$\beta_1$</td>
<td>Independent and pairwise</td>
</tr>
<tr>
<td>with $k = 3$</td>
<td>$\beta_1 = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\psi_1 = 2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\psi_2 = 3$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Our main purpose is to conduct inference for $\beta_1$, however, in one simulation study, we choose our parameter of interest to be the dependence parameter $\psi$ in order to see the performance of the pairwise composite likelihood approach.

In the Sibling study with $k = 5$, we choose three different values for the dependence
parameter $\psi$ to indicate weak dependence ($\psi = 1.2$), moderate dependence ($\psi = 3$) and strong dependence ($\psi = 6$) within family members. Here we are interested in whether the inference about $\beta_1$ is affected by different strengths of dependence. For the other family structures, we only take into account $\psi = 3$. To get an idea of how much correlation these $\psi$ values induce within a binary pair, we regenerate the table presented in Jin (2010, p.115) according to our values in Table 2.4. We only assume positive dependence within pairs, as the odds of having the trait in a family will increase if another member of the family is also affected.

Table 2.4: The relationship between correlations within a binary pair and odds ratio.

| odds ratio | $\delta_{ij|0,0}$ | $\delta_{ij|0,1}$ | $\delta_{ij|1,1}$ | $\delta_{ij|0,2}$ | $\delta_{ij|1,2}$ | $\delta_{ij|2,2}$ |
|------------|------------------|------------------|------------------|------------------|------------------|------------------|
| $\psi$     |                  |                  |                  |                  |                  |                  |
| 1.2        | 0.037            | 0.035            | 0.037            | 0.017            | 0.018            | 0.009            |
| 3          | 0.232            | 0.184            | 0.232            | 0.078            | 0.121            | 0.077            |
| 6          | 0.379            | 0.260            | 0.379            | 0.105            | 0.200            | 0.160            |

where $\delta_{ij|k,l} = \text{corr}(Y_i, Y_j \mid x_i = k, x_j = l)$ in Eq.(2.17) and $(Y_i,Y_j)$ is a sibling pair.

For each setting in Table 2.3, we generate family data where the number of families vary with $N = 30, 100, 300, 500$ and 1000. There are two main results from the simulation that we focus on: (1) we first estimate the profile MCLE of the parameter of interest and check whether it converges to the true value as sample size increases, as this will indicate that the object of inference is the same as the object of interest. That is, the composite likelihoods provide evidence about the true parameter; and (2) we then calculate the behaviour of the probability of observing misleading evidence defined in Eq.(2.8) of Theorem 2.

For the sibling studies where there is one dependence parameter $\psi$, we have $\theta = (\beta_0, \beta_1, \psi)$, where $\beta_1$ is the parameter of interest and $(\beta_0, \psi)$ are the nuisance parameters. In another simulation setting, $\psi$ is the parameter of interest and $(\beta_0, \beta_1)$ are the nuisance
parameters. For the family study, we have \( \theta = (\beta_0, \beta_1, \psi_1, \psi_2) \), where \( \beta_1 \) is the parameter of interest and \((\beta_0, \psi_1, \psi_2)\) are the nuisance parameters. We follow the steps below to find the profile MCLEs of the parameter of interest, for example, when the parameter of interest is \( \beta_1 \) and \( \theta = (\beta_0, \beta_1, \psi) \).

1. Set a grid for \( \beta_1 \), i.e. \( \{\beta_{11}, \beta_{12}, \ldots, \beta_{1g}\} \).

2. For each \( \beta_{1i}, i = 1, \ldots, g \), maximize the composite likelihood chosen with respect to the nuisance parameters as a function of \( \beta_{1i} \). Obtain the composite likelihood value for each \((\beta_{1i}, \hat{\beta}_0(\beta_{1i}), \hat{\psi}(\beta_{1i}))\). Note that \((\hat{\beta}_0(\beta_{1i}), \hat{\psi}(\beta_{1i}))\) are the MCLEs when \( \beta_1 \) is taken as fixed. For this, we use the Newton Raphson algorithm implemented in the \( R \) Statistical Software.

3. Observe the profile MCLE, \( \hat{\beta}_{1CL} \), which maximizes the composite likelihoods calculated in Step 2.

We generate 10,000 simulated data sets and get our parameter estimations by averaging over 10,000 profile MCLEs:

\[
\hat{\beta}_{1CL} = \frac{1}{10000} \sum_{j=1}^{10000} \hat{\beta}_{1CL}^{(j)} \quad \hat{\psi}_{CL} = \frac{1}{10000} \sum_{j=1}^{10000} \hat{\psi}_{CL}^{(j)} \quad (2.20)
\]

The next step is to estimate the probability of misleading evidence in Eq.(2.8) of Theorem 2. For this, we first estimate the robust adjustment factor \( a/b \). Remember that \( a = H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \) and \( b = H^{\psi\psi}(\psi_g, \lambda_g)^{-1}G^{\psi\psi}(\psi_g, \lambda_g)H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \) where \( \psi \) is the parameter of interest, \( \lambda \) is the nuisance vector. Note that for \( \theta = (\psi, \lambda) \), \( G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) \) with \( H(\theta) = E_g\{-\partial^2 cl(\theta; Y)/\partial \theta \partial \theta^T\} \) and \( J = E_g\{(\partial cl(\theta; Y)/\partial \theta)(\partial cl(\theta; Y)/\partial \theta)^T\} \). We estimate \( J(\theta) \), using

\[
\hat{J}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} u(\hat{\theta}_{CL}; y_i)u(\hat{\theta}_{CL}; y_i)^T \quad (2.21)
\]

where \( u(\theta; y_i) \) is the elements of the composite score function, \( y_i \) is the observations vector, \( \hat{\theta} \) is the global MCLEs of \( \theta = (\beta_0, \beta_1, \psi) \), and \( N \) is the sample size (number of families).
Then we estimate \( H(\theta) \), using
\[
\hat{H}(\hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \text{cl}(\theta; y_i)}{\partial \theta \partial \theta^T}
\] (2.22)

For the parameter of interest \( \beta_1 \), \( \hat{H}^{\beta_1,\beta_1}(\hat{\theta}) \) and \( \hat{G}^{\beta_1,\beta_1}(\hat{\theta}) \) are the entities of the matrices \( \hat{H}^{-1}(\hat{\theta}) \) and \( \hat{G}^{-1} \) that belong to the parameter \( \beta_1 \). Then,
\[
\frac{\hat{a}}{\hat{b}} = \frac{\hat{H}^{\beta_1,\beta_1}(\hat{\theta})^{-1}}{\hat{H}^{\beta_1,\beta_1}(\hat{\theta})^{-1} \hat{G}^{\beta_1,\beta_1}(\hat{\theta}) \hat{H}^{\beta_1,\beta_1}(\hat{\theta})^{-1}} = \frac{\hat{H}^{\beta_1,\beta_1}(\hat{\theta})}{\hat{G}^{\beta_1,\beta_1}(\hat{\theta})}
\] since \( \beta_1 \) is a scalar. (2.23)

We are now ready to estimate the probability of misleading evidence in Eq.(2.8) of Theorem 2. For each simulated dataset, we check whether the composite likelihood ratio with the robust adjustment factor is greater than the pre-specified threshold. Then we estimate the probability of misleading evidence using
\[
\frac{1}{10000} \sum_{j=1}^{10000} \{ \frac{CL_p(\beta_1; y^{(j)})}{CL_p(\beta_1; \tilde{y}^{(j)})} \} \frac{\hat{a}}{\hat{b}} \geq k \}
\] (2.24)

where \( y^{(j)} \) is the \( j \)th simulated dataset under the model parameter \( \beta_{1g} \), \( \beta_1 \) is a parameter value that is different than \( \beta_{1g} \) and \( I \) is the indicator function.

### 2.3.3 Simulation results

Here, we first give the results for checking whether the profile MCLEs in Eq (2.20) are consistent estimators of the true parameters, that is, whether the object of inference is equal to the object of interest. We then determine the behaviour of the probability of observing misleading evidence under different simulation settings. Our findings are summarized below.

**Checking whether the object of inference is equal to the object of interest**

The simulation results for checking whether the profile MCLEs of \( \beta_1 \) converge to the true parameter value for different family structures and sample sizes based on different
composite likelihoods are given in Tables 2.5, 2.6, and 2.7. The simulation results for checking whether the profile MCLE of $\psi$ converge to the true parameter value for the Sibling study with $k = 5$ based on a pairwise likelihood is given in Table 2.8.

In Table 2.5, for the Sibling study with $k = 5$ siblings, we see that as sample size increases, both composite likelihood approaches provide consistent estimates for the true parameter value $\beta_1$. This ensures that the object of inference is equal to the object of interest; that is, both composite likelihood approach provides evidence about the true parameter value. Hence, both composite likelihood approaches can be used to model correlated binary data when families have the same number of siblings. The results do not change whether the dependence between sibling pairs are weak or strong.

Table 2.5: Sibling study with $k = 5$. The maximum profile composite likelihood estimates of $\beta_1$, using independent and pairwise likelihood methods, under a weak, moderate and strong dependence parameter ($\psi = 1.2, 3$ and $6$) and with different numbers of families ($n$).

<table>
<thead>
<tr>
<th>$\psi$</th>
<th>$n$</th>
<th>independent</th>
<th>pairwise</th>
<th>$\psi$</th>
<th>$n$</th>
<th>independent</th>
<th>pairwise</th>
<th>$\psi$</th>
<th>$n$</th>
<th>independent</th>
<th>pairwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>30</td>
<td>2.048</td>
<td>2.048</td>
<td>2.064</td>
<td>2.062</td>
<td>2.096</td>
<td>2.090</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>2.016</td>
<td>2.016</td>
<td>2.023</td>
<td>2.023</td>
<td>2.024</td>
<td>2.024</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>2.002</td>
<td>2.002</td>
<td>2.006</td>
<td>2.006</td>
<td>2.006</td>
<td>2.006</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>2.002</td>
<td>2.002</td>
<td>2.003</td>
<td>2.003</td>
<td>2.003</td>
<td>2.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>2.000</td>
<td>2.000</td>
<td>2.002</td>
<td>2.002</td>
<td>2.002</td>
<td>2.002</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 2.6, for the Sibling study with $k \in \{2, 3, 4, 5\}$ siblings, we also see that as
sample size increases, both composite likelihood approaches provide consistent estimates for the true parameter value $\beta_1$. Thus, both composite likelihood approaches provide evidence about the true parameter, thus can be used to model correlated binary data when families have the different number of siblings.

Table 2.6: Sibling study with $k \in \{2, 3, 4, 5\}$. The maximum profile composite likelihood estimates of $\beta_1$, using independent and pairwise likelihood methods, under the moderate dependence parameter $\psi = 3$ and with different number of families $n$.

<table>
<thead>
<tr>
<th>Model parameter, $\beta_1 = 2$</th>
<th>$\hat{\beta}_{1CL_p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>independent</td>
</tr>
<tr>
<td>30</td>
<td>2.106</td>
</tr>
<tr>
<td>100</td>
<td>2.029</td>
</tr>
<tr>
<td>300</td>
<td>2.008</td>
</tr>
<tr>
<td>500</td>
<td>2.006</td>
</tr>
<tr>
<td>1000</td>
<td>2.002</td>
</tr>
</tbody>
</table>

In Table 2.7, for the Family study with $k = 3$ siblings and parents, we see that as sample size increases, the independent likelihood provides consistent estimates for the true parameter value $\beta_1$. However, the pairwise likelihood does not. We conclude that when there is more than one type of relationship in a family (e.g. within sibling pairs and between a parent and the offspring), the pairwise likelihood approach fails to provide reliable evidence for the mean parameter, $\beta_1$. This is due to that fact that the two different parameter for dependence, $\psi_1$ and $\psi_2$, induce some constraints on the true mean parameters ($\beta_0, \beta_1$). This changes the meaning of the mean parameters that are represented by the pairwise likelihood. Other composite likelihood approaches, e.g. a composite conditional likelihood may be suitable for this type of problem, however, evaluation of this is beyond the scope of this thesis.
Table 2.7: Family study with $k = 3$ siblings. The maximum profile composite likelihood estimates of $\beta_1$, using independent and pairwise likelihood methods, under a moderate dependence parameter $\psi = 3$ and with different number of families ($n$).

<table>
<thead>
<tr>
<th>Model parameter, $\beta_1 = 2$</th>
<th>$\hat{\beta}_{1CL_p}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>independent</td>
</tr>
<tr>
<td>30</td>
<td>2.062</td>
</tr>
<tr>
<td>100</td>
<td>2.022</td>
</tr>
<tr>
<td>300</td>
<td>2.003</td>
</tr>
<tr>
<td>500</td>
<td>2.003</td>
</tr>
<tr>
<td>1000</td>
<td>2.002</td>
</tr>
</tbody>
</table>

In Table 2.8, we see that as sample size increases, the pairwise likelihood provides consistent estimates for the true parameter value $\psi$. Thus, when the parameter is the dependence parameter, the pairwise likelihood can be used to get inference for $\psi$.

Table 2.8: Sibling study with $k = 5$. The maximum profile composite likelihood estimates of $\delta = \log(\psi)$, using the pairwise likelihood method and with different number of families ($n$), $\beta_0 = -1$, $\beta_1 = 2$, $\delta = \log(3) = 1.099$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>30</th>
<th>100</th>
<th>300</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_{CL_p}$</td>
<td>1.031</td>
<td>1.079</td>
<td>1.092</td>
<td>1.094</td>
<td>1.096</td>
</tr>
</tbody>
</table>

**Checking the behaviour of the probability of observing misleading evidence**

We now investigate whether the probability of observing misleading evidence can be described by the bump function (Figure 1.2 in Chapter 1) under the two composite likelihood approaches as the sample size increases. The bump function has the maximum value of $\Phi(-\sqrt{2\log 8}) = 0.021$ when the strong evidence criterion for the likelihood ratio is 8. The horizontal line at 0.021 in all figures below indicates this maximum.
Figure 2.1: Sibling study with $k = 5$. Plots for the probability of misleading evidence under the independent (---) and pairwise (−) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1, \beta_1 = 2, \psi = 3$ and $n = 30$.

From Figures 2.1 and 2.2 we see the behaviour of the probability of observing misleading evidence for $\beta_1$ for the Sibling study with $k = 5$ siblings for sample sizes 30 and 500, respectively. These figures illustrate, first, before the robust adjustment is applied, the probability of observing misleading evidence is not approximated by the bump function, as expected. The pairwise likelihood approach performs slightly better than the independent likelihood approach, i.e. it produces smaller probabilities of observing misleading evidence, however, the results are still not satisfactory (Figures 2.1(a) and 2.2(a)). Second, after the robust adjustment is applied, the probability of observing misleading evidence can be approximated by the bump function for large samples. This result holds for both independent and pairwise likelihood approaches because both composite likelihood approaches provide MCLEs that are consistent estimates of the true parameter. We see from Figures 2.3 and 2.4 that the same conclusions hold for the Sibling study with $k \in \{2, 3, 4, 5\}$ siblings.
Figure 2.2: Sibling study with $k = 5$. Plots for the probability of misleading evidence under the independent (- -) and pairwise (–) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi = 3$ and $n = 500$.

Figure 2.3: Sibling study with $k \in \{2, 3, 4, 5\}$. Plots for the probability of misleading evidence under the independent (- -) and pairwise (–) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi = 3$ and $n = 30$. 
Chapter 2. A composite likelihood approach for association analysis

Figure 2.4: Sibling study with $k \in \{2, 3, 4, 5\}$. Plots for the probability of misleading evidence under the independent (---) and pairwise (–) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi = 3$ and $n = 500$.

For the family study with 3 siblings, we have observed that the object of inference is not the object of interest when the pairwise likelihood approach is used (Table 2.7). Thus, the probability of misleading evidence will not converge to a bump function as the sample size increases even after the robust adjustment is applied. We see this occurs in Figures 2.5 and 2.6 (solid line). When we use the independent likelihood approach with the robust adjustment, we get the bump function as expected.

From Figure 2.7, we see that the probability of observing misleading evidence can be described by the bump function when the pairwise likelihood approach with the robust adjustment is used for making inference about $\psi$. However, empirical evidence suggests that we need big sample sizes to attain a better symmetric shape, which is predicted asymptotically.
Figure 2.5: Family study with $k = 3$. Plots for the probability of misleading evidence under the independent (−−) and pairwise (−) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi_1 = 2$, $\psi_2 = 3$ and $n = 30$.

Figure 2.6: Family study with $k = 3$. Plots for the probability of misleading evidence under the independent (−−) and pairwise (−) methods (a) before robust adjustment (b) after robust adjustment. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi_1 = 2$, $\psi_2 = 3$ and $n = 500$. 
Figure 2.7: Sibling study with $k = 5$. Plots for the probability of misleading evidence under pairwise methods (−) before robust adjustment (−−) after robust adjustment (a)$n=30$, (b)$n=100$, (c)$n=300$, (d)$n=500$, and (e)$n=1000$. Data were generated under $\beta_0 = -1$, $\beta_1 = 2$, $\psi = 3$ ($\delta = \log(3) = 1.099$ on the x-axis).
2.3.4 Summary of the simulation study

We end this section with a summary of some important aspects.

1. As long as the object of inference is the same as the object of interest, both composite likelihood approaches after the robust adjustment achieve correct inference when the sample size is large. One can argue that the independent likelihood approach is preferable as it is easier to implement. However, for smaller samples, we see a slight gain in estimation performance, and in obtaining smaller probabilities of misleading evidence when using the pairwise likelihood approach.

2. If there is more than one type of dependence structure within a family, the object of inference is not equal to the object of interest when the pairwise likelihood is used. Hence, the pairwise likelihood approach would be misleading. This is not a problem in the independent likelihood approach.

3. The pairwise likelihood approach enables inference of second order parameters, such as the dependence.

2.4 Data example

In this section, we demonstrate the independent composite likelihood approach proposed in Section 2.2.4 to an earlier study of the analysis of genetic association with meconium ileus (MI) in individuals with cystic fibrosis (CF) (Sun et al., 2012).

It is known that carrying two disease causing mutations in the CFTR gene is responsible for CF and that about 15% of CF patients have intestinal obstruction at birth, a complication known as MI (Sun et al., 2012). There is evidence that other genes, called modifier genes, contribute to MI risk. Sun et al. (2012) identify three other gene regions that contribute to MI. The original study consists of siblings and unrelated CF patients.
GEE was used to conduct a genome-wide association study that identified eight SNPs from the three regions as associated at the genome-wide significance level of $p < 5 \times 10^{-8}$.

Using the subset of that data set, which was collected by the Canadian Consortium for Cystic Fibrosis Gene Modifiers, we analyzed 1438 families with singletons, 107 families with two siblings and three families with three siblings for a total of 1661 individuals with CF. Among those there are 252 individuals with MI. Here we present our findings from an analysis of SNPs on the X chromosome, which contained the SNPs with the smallest $p$-values at the previously identified regions in Sun et al. (2012).

### 2.4.1 Our Model

The majority of the data set consists of independent individuals (1438 out of 1661). There are 107 families with two siblings and only three families with three siblings. We constructed a composite marginal likelihood under a working independence assumption with a robust adjustment factor, to correct for the misspecified model for correlated individuals and we assume an underlying logistic regression model with an additive genetic effect of the genotype:

$$
\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i
$$

where $E(Y_i|x_i) = p_i$ is the marginal probability that subject $i$ has the disease given the number of minor alleles at the SNP, $Y_i = 1$ if subject $i$ has the disease and zero otherwise. The number of minor alleles, $X_i$, is coded as 0 and 2 for males and 0, 1 and 2 for females.

The composite likelihood function with the robust adjustment factor, $a/b$, is:

$$
CL(\beta_0, \beta_1) = \left[ \prod_{i=1}^{n} \left( \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right)^{y_i} \left( 1 - \left( \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right) \right)^{1-y_i} \right]^{a/b}
$$

where $a/b = H_{\beta_0, \beta_1}(\beta_0, \beta_1)/G_{\beta_0, \beta_1}(\beta_0, \beta_1)$. See Appendix B for the derivation of $a/b$ and Eq. (2.23) for estimating $a/b$. 
In genetic association studies, it is common to choose the odds ratio (OR), $e^{\beta_1}$ as the interest parameter, and plot the likelihood function for $e^{\beta_1}$ (Strug et al., 2010). Under the null hypothesis of no association, the OR is equal to 1, and the OR is some value greater than 1 under the alternative. Note that since $\beta_0$ is a nuisance parameter, we profile out the baseline odds, $e^{\beta_0}$, and use the profile composite likelihood ($CL_p$) ratio as our evidence function.

### 2.4.2 Likelihood Plots

**Likelihood plot for a single SNP**

In Sun et al. (2012), three SNPs from Chromosome X were found to be associated with MI with rs3788766 providing the smallest p-value. In Figure 2.8, we plot the $CL_p$ function of the OR at rs3788766, and for comparison, two other SNPs that are not associated with MI to illustrate an evidential analysis of genetic association. By plotting the $CL_p$ function, we can assess all the evidence about association the data set provides.

In Figure 2.8(a), the $1/8$ $CL_p$ interval for the OR is 1.3 to 1.8. The OR values between this interval are consistent with the data at the level $k=8$, i.e. there are no other values outside this interval that are better supported than the values within the interval, by a factor of as large as 8 (Royall, 1997). We see that OR=1 is outside of the $1/8$ $CL_p$ interval. That tells us that there are some parameter values of the OR, namely the MLE, $\hat{OR}_{mle} = 1.5$, that are better supported than an OR=1 by a factor of greater than 8.

The $1/32$ $CL_p$ interval shows that an OR=1 is also not supported by the data at level $k=32$. The adjustment factor $\hat{a}/\hat{b}$ is 0.98, which is very close to 1, suggesting that the composite likelihood is not too discrepant from the true likelihood. This is due to the fact that most observations in our data are independent.

In Figure 2.8(b), we see that both $1/8$ and $1/32$ $CL_p$ intervals include OR=1 as a plausible value. This tells us that there is no value that the data supports over OR=1.
by a factor of 8 or more. In Figure 2.8(c), we noticed that the $CL_p$ is skewed, suggesting that there is sparsity in the data. (see Table 2.9).

Table 2.9: Distribution of data at SNP rs12393509

<table>
<thead>
<tr>
<th>SNP</th>
<th># of minor allele</th>
<th>Disease status</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>Yes</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>1306</td>
<td>238</td>
<td>1544</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1308</td>
<td>239</td>
<td>1547</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.8: Standardized $CL_p$ function for OR at three SNPs. $1/8$ and $1/32$ $CL_p$ intervals ($CL_p I$) as well as the estimated robust adjustment factor $\hat{a}/\hat{b}$ are provided.
Chapter 2. A composite likelihood approach for association analysis

Likelihood plot for a region of typed SNP

In genetic association analysis, one needs to evaluate many SNPs in a region. Plotting hundreds of individual likelihood functions corresponding to each SNP may not be practical. Instead, Strug, et al (2010) provide a single plot that represents the association information in a region of interest. It highlights the region of SNPs by base pair position where the evidence strength for association is the highest. Figure 2.9 presents an example with 16 SNPs on Chromosome X from 1661 individuals. We have the SNPs by base pair position on the x-axis and OR on the y-axis. The vertical lines for each SNP on the x-axis represent the $1/k \text{CL}_p$ intervals where $k = 32, 100$ and 1000. $k=32$ is chosen as the criterion to define whether the SNP is associated with MI or not. We see that some of the $\text{CL}_p$ intervals are grey. These $\text{CL}_p$ intervals include OR=1 (horizontal line at OR=1) as a plausible value so the SNPs that produced these intervals are not concluded to be associated with MI. If the $1/k \text{CL}_p$ interval does not include OR=1, then it will be coloured in green, red or navy blue for $k=32,100$ or 1000 respectively. For example, for the SNP rs12839137, the green portion of the $\text{CL}_p$ interval is above the OR=1 line, the red and navy blue portions are either touching or below the OR=1 line. Thus, the $1/32 \text{CL}_p$ interval for this SNP does not include the OR=1 value as a plausible value, however, $1/100$ and $1/1000 \text{CL}_p$ intervals include OR=1. This tells us that the evidence strength provided for this SNP is greater than 32 but less than 100. For SNP rs3788766, the $1/k \text{CL}_p$ interval for $k=32, 100$ and 1000 do not include OR=1 as a plausible value, indicating an association between this SNP and MI at the level $>1000$. Note that the $1/32 \text{CL}_p$ interval for rs3788766 (the green portion) is the same as provided in Figure 2.8(a). Also notice the long grey $\text{CL}_p$ interval at the middle-right of the plot. This corresponds to the SNP in Figure 2.8(c). Therefore, Figure 2.9 also highlights if there is something unusual with any of the SNPs in the data.

The small horizontal tick on each $\text{CL}_p$ interval represents the MLE for the ORs at the SNPs that were found to be associated with MI at some $k$. The max LRs (=
$CL_p(\hat{OR}_{mle})/CL_p(1)$ for the three SNPS where the strength of evidence for association is the largest are also provided on the plot. These findings are consistent with the GEE results in the original study (Sun et al., 2012).

![Evidential analysis of association between SNPs in chromosome X and MI using $CL_p$ ratio with a robust adjustment.](image)

Figure 2.9: Evidential analysis of association between SNPs in chromosome X and MI using $CL_p$ ratio with a robust adjustment.

### 2.5 Summary

We showed that the composite likelihood ratio, with a robust adjustment, is a valid statistical evidence function in EP as it has the two performance properties of an evidence function. We need to use the robust adjustment on the composite likelihood even though the likelihood objects in the composite likelihood are correctly specified, since most of the time multiplying them to construct the composite likelihood does not lead to a probability density function.

We presented an extensive simulation study based on a genetic association analysis of family data where a logistic regression model with an additive effect of the genotype on
the marginal binary response is assumed. To generate the correlated binary data, we used a discretised normal approach with specified marginal probabilities and pairwise correlations. We proposed constructing the composite likelihood 1) from independent marginals, and 2) from pairwise marginals on three different family structures. We concluded that when there is more than one type of relationship in a family, the pairwise likelihood approach can not be used to obtain evidence about the mean parameters (regression coefficients) of the model. We also observed that we get reliable evidence about the mean parameter when the composite likelihoods constructed from independent marginals are used. We can extend the use of the independent composite likelihood approach to the cases where we have more complicated and bigger families. Moreover, using independent likelihoods reduces the computations considerably.

Lastly, we illustrate the evidential analysis of genetic association between a trait and a SNP in a real data example, where we conducted the evidential analysis using the composite likelihood constructed from independent marginals.

In conclusion, we developed an EP approach for the analysis of genetic association in family data by providing a reliable evidence function when the real likelihoods are intractable or impractical to construct.
Chapter 3

Hypothesis Assessment and
Inequalities for Relative Belief Ratios

In this chapter, we present in detail our approach to hypothesis assessment that we introduced in Chapter 1. We are particularly interested in the problem of assessing a point null hypothesis $H_0$ based on the choice of a single prior $\Pi$ on the parameter space, where $\Pi(H_0) = 0$. In the Bayesian approach, the Bayes factor (BF) is commonly used for assessing the evidence for or against a given hypothesis $H_0$. We proposed using the relative belief ratio for assessing the evidence for or against a given hypothesis $H_0$ and illustrate the close relationship between Bayes factors and relative belief ratios in Chapter 1. When the prior probability of $H_0$ is 0, the calculations of $BF(H_0)$ and $RB(0)$ are ambiguous. The most common practice in such situations for Bayes factors is to replace $\Pi$ by a mixture prior $\Pi_\gamma = \gamma \Pi_0 + (1 - \gamma)\Pi$, where $\gamma$ is a positive probability assigned to $H_0$ and $\Pi_0$ is a probability measure on $H_0$ and compute $BF(H_0)$ with respect to $\Pi_\gamma$ (Jeffreys, 1935, 1961).

In Section 3.1, we provide the definition of the $BF(H_0)$ arising from the single prior $\Pi$, which is basically the relative belief ratio at $H_0$ when $\Pi(H_0) = 0$ and show its connection to the Bayes factor arising from a mixture prior, $\Pi_\gamma$. We point out a general inconsistency
in the assignment of prior beliefs when $\Pi_0$ is not given by the conditional prior of the parameter of interest given $H_0$. We also consider the problem when there is no obvious parameter of interest that generates $H_0$, e.g. in the Behrens-Fisher problem.

We then provide additional evidential properties of the relative belief ratio as a measure of evidence for $H_0$ and present our approach to hypothesis assessment with examples. Our approach is based on 1) the computation of the relative belief ratio to assess the evidence given by data, 2) the computation of a measure of the strength of the evidence given by the RB ratio via a posterior probability, which is effectively the calibration of relative belief ratios, and Bayes factors and 3) the point where the relative belief ratio is maximized, which is useful in determining whether or not we have detected a deviation from $H_0$ of practical significance.

In Section 3.1.4, we establish some close parallels between the use of relative belief ratios to assess statistical evidence and the approach to assessing statistical evidence via likelihood ratios (Royall, 1997) discussed in Chapter 1. Royall (2000) derived bounds for probability of observing strong misleading evidence (see Section 1.1.2). We derive similar inequalities for relative belief inferences and this procedure allows us to evaluate the bias inherent in the relative belief ratio due to prior assigned. This is useful for eliciting priors in a Bayesian analysis.

In Section 3.2, we present an application of relative belief inferences to a two-way analysis. We conclude with a short summary and discussion in Section 3.3.

The contents of this Chapter can be found in Baskurt and Evans (2012) and Baskurt and Evans (2013).
3.1 Evidential Interpretation of Relative Belief Ratios for Assessing a Point Null Hypothesis

3.1.1 Calculation of Bayes factors and relative belief ratio for a point null hypothesis

Suppose we wish to assess a hypothesis \( H_0 = \Psi^{-1}\{\psi_0\} \) for some parameter of interest \( \psi = \Psi(\theta) \) and with \( \Pi(H_0) = 0 \). We showed in Chapter 1.2 that the relative belief ratio at \( \psi \), \( RB(\psi) = \pi(\psi|x)/\pi(\psi) \), is a limit of Bayes factors with respect to \( \Pi \), and so can also be called the Bayes factor in favor of \( \psi \) with respect to \( \Pi \). However, if \( \Pi(H_0) > 0 \), then \( RB(\psi) \) is not a Bayes factor with respect to \( \Pi \) but is related to the Bayes factor through Eq (1.8). We also defined the density of \( T \), where \( T : \mathcal{X} \rightarrow \mathcal{T} \) denotes a minimal sufficient statistic for \( \{f_\theta : \theta \in \Theta\} \) in Chapter 1.2 using the definitions in Tjur (1974). Then \( f_\theta T(t) = \int_{T^{-1}\{t\}} f_\theta(x)J_T(x)\mu_{T^{-1}\{t\}}(dx) \), where \( \mu_{T^{-1}\{t\}} \) denotes volume on \( T^{-1}\{t\} \). The prior predictive density, with respect to \( \mu \), of the full data is given by, \( m(x) = \int_\Theta \pi(\theta)f_\theta(x)\nu(d\theta) \). And the prior predictive density of \( T \), with respect to \( \mu_T \), is given by \( m_T(t) = \int_\Theta \pi(\theta)f_\theta T(x)\nu(d\theta) = \int_{T^{-1}\{t\}} m(x)J_T(x)\mu_{T^{-1}\{t\}}(dx) \). These derivations lead us to a generalization of the Savage-Dickey ratio result (Dickey, 1971), as we do not require coordinates for nuisance parameters.

**Theorem 3. (Savage-Dickey)**

\[
\frac{\pi_\psi(\psi \mid T(x))}{\pi_\psi(\psi)} = \frac{m_T(T(x) \mid \psi)}{m_T(T(x))}. \tag{3.1}
\]

**Proof.** The posterior density of \( \theta \) is \( \pi(\theta|T(x)) = \pi(\theta)f_{\theta T}(T(x))/m_T(T(x)) \). The posterior
density of $\psi = \Psi(\theta)$, with respect to $\nu_\Psi$, is;

$$
\pi_\Psi(\psi \mid T(x)) = \int_{\Psi^{-1}\{\psi\}} \pi(\theta \mid T(x)) J_\Psi(\theta) \nu_{\Psi^{-1}\{\psi\}}(d\theta) \quad \text{(from Eq. (1.9))}
$$

$$
= \int_{\Psi^{-1}\{\psi\}} \frac{\pi(\theta) f_{\theta T}(T(x))}{m_T(T(x))} J_\Psi(\theta) \nu_{\Psi^{-1}\{\psi\}}(d\theta)
$$

$$
= \int_{\Psi^{-1}\{\psi\}} \frac{\pi(\theta \mid \psi) \pi_\Psi(\psi) f_{\theta T}(T(x))}{m_T(T(x))} \nu_{\Psi^{-1}\{\psi\}}(d\theta) \quad \text{(from Eq. (1.10))}
$$

$$
= \pi_\Psi(\psi) \int_{\Psi^{-1}\{\psi\}} \frac{\pi(\theta \mid \psi) f_{\theta T}(T(x))}{m_T(T(x))} \nu_{\Psi^{-1}\{\psi\}}(d\theta)
$$

and so

$$
\frac{\pi_\Psi(\psi \mid T(x))}{\pi_\Psi(\psi)} = \frac{m_T(T(x) \mid \psi)}{m_T(T(x))}
$$

where $m_T(T(x) \mid \psi)$ is the conditional prior predictive density of $T$ given $\Psi(\theta) = \psi$ and $m_T(T(x)) = \int_\Psi m_T(T(x) \mid \psi) \pi_\Psi(\psi) \nu_\Psi(d\psi)$.}

As we noted in Definition 2 in Chapter 1, Eq (3.1) is the relative belief ratio of $\psi$ when $\Pi(\Psi^{-1}\{\psi_0\}) = 0$. This theorem provides computational convenience in calculating the relative belief ratio.

Below, we illustrate three cases. In the first two cases, we construct the connection between relative belief ratios and Bayes factors when the prior distribution on $\Psi$ is 1) discrete and 2) continuous and when the parameter of interest $\Psi(\theta)$ is specified in the problem. In case 3) we look into the situation where there is no parameter of interest clearly specified in the problem (see Example 9 below).

1. Suppose that the prior distribution on $\Psi$ is discrete with $0 < \pi_\Psi(\psi) < 1$. The prior predictive density of $T$ is $m_T(T(x)) = \pi_\Psi(\psi_0) m_T(T(x) \mid \psi_0) + (1 - \pi_\Psi(\psi_0)) m_T(T(x) \mid \{\psi_0\}^c)$ where

$$
m_T(T(x) \mid \psi_0) = \int_{\Psi^{-1}\{\psi_0\}} \pi(\theta \mid \psi_0) f_{\theta T}(T(x)) \nu_{\Psi^{-1}\{\psi_0\}}(d\theta)
$$

$$
m_T(T(x) \mid \{\psi_0\}^c) = \int_{\Psi \backslash \{\psi_0\}} m_T(T(x) \mid \psi) \pi_\Psi(\psi) \nu_\Psi(d\psi)
$$
are the prior predictive densities given $\Psi(\theta) = \psi_0$ and $\Psi(\theta) \neq \psi_0$, respectively.

We have that,

$$
\frac{m_T(T(x) \mid \psi_0)}{m_T(T(x))} = \frac{m_T(T(x) \mid \psi_0)}{m_T(T(x))} = \frac{(1 - \pi_\psi(\psi_0)) m_T(T(x) \mid \{\psi_0\}^c)}{(1 - \pi_\psi(\psi_0)) m_T(T(x) \mid \{\psi_0\}^c) + (1 - \pi_\psi(\psi_0)) (3.2)
$$

$m_T(T(x) \mid \psi_0) / m_T(T(x))$ is the relative belief ratio for $\psi_0$ due to Eq.(3.1). We see that

$$
\frac{m_T(T(x) \mid \psi_0)}{m_T(T(x) \mid \{\psi_0\}^c)} = \frac{(1 - \pi_\psi(\psi_0)) m_T(T(x) \mid \psi_0) / m_T(T(x))}{1 - \pi_\psi(\psi_0) m_T(T(x) \mid \psi_0) / m_T(T(x))}
$$

The Bayes factor for $H_0$ is given by

$$
BF(H_0) = \frac{\pi_\psi(\psi_0 \mid T(x))}{(1 - \pi_\psi(\psi_0 \mid T(x)))} \div \frac{\pi_\psi(\psi_0)}{(1 - \pi_\psi(\psi_0))}
$$

$$
= \frac{m_T(T(x) \mid \psi_0) / m_T(T(x) \mid \{\psi_0\}^c)}{m_T(T(x) \mid \{\psi_0\}^c)} \quad \text{(from Eq. (3.2))}
$$

Thus, $BF(H_0)$ is a 1-1 increasing function of the relative belief ratio of $\psi_0$ as also established in Eq.(1.8).

2. Now consider the situation where $\Pi(H_0) = 0$. Let $H_0^\epsilon \subset \Psi$ be such that $H_0^\epsilon \subset \Psi$ converges nicely to $\{\psi_0\}$. Then we have that

$$
BF(H_0^\epsilon) = \frac{\Pi_\psi(H_0^\epsilon \mid T(x))}{(1 - \Pi_\psi(H_0^\epsilon \mid T(x)))} \div \frac{\Pi_\psi(H_0^\epsilon)}{(1 - \Pi_\psi(H_0^\epsilon))}
$$

$$
\to \frac{\pi_\psi(\psi \mid T(x))}{\pi_\psi(\psi)} = \frac{m_T(T(x) \mid \psi_0)}{m_T(T(x))} \quad \text{as } \epsilon \to 0.
$$

So the relative belief ratio at $\psi_0$ is a limit of Bayes factors.

Suppose we assign a prior probability $\gamma \in (0, 1)$ to $H_0$ and replace $\Pi$ by $\Pi_\gamma = \gamma \Pi_0 + (1 - \gamma) \Pi$ where $\Pi_0$ is a probability measure on $H_0$ so $\Pi_\gamma(H_0) = \gamma$. Then, let $m_{0T}$ denote the prior predictive density of $T$ under $\Pi_0$, which is given by $m_{0T} = \,...
\[
\int_{\Psi^{-1}\{\psi_0\}} \pi_0(\theta) f_{\theta T}(T(x)) \nu_{\Psi^{-1}\{\psi_0\}}(d\theta)
\]
where \(\pi_0\) is the prior density of \(\theta\) under \(\Pi_0\).

We obtain the Bayes factor and relative belief ratio for \(H_0\) with respect to \(\Pi_\gamma\) as in Table 3.1. These are generally not equal.

**Table 3.1: The Bayes factor and relative belief ratio for \(H_0\) under \(\Pi_\gamma\).**

<table>
<thead>
<tr>
<th>(BF(H_0))</th>
<th>(RB(H_0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\frac{m_{0T}(T(x))}{m_T(T(x))})</td>
<td>(\frac{m_{0T}(T(x))}{m_T(T(x))} / \frac{m_T(T(x))}{m_T(T(x))})</td>
</tr>
</tbody>
</table>

We now show that in certain circumstances \(BF_{\Pi_\gamma}(\psi_0) = RB(\psi_0)\) where \(RB(\psi_0)\) is the relative belief ratio with respect to \(\Pi\). The Savage-Dickey ratio result in Eq.(3.1) holds when using the conditional prior \(\Pi(\cdot | \psi_0)\) on \(H_0\), i.e. \(\Pi_0 = \Pi(\cdot | \psi_0)\). Verdinelli and Wasserman (1995) extends the Savage-Dickey ratio result when \(\Pi_0 \neq \Pi(\cdot | \psi_0)\). In the following theorem, we generalize their result as we do not require coordinates for nuisance parameters.

**Theorem 4.** (Verdinelli-Wasserman) When \(H_0 = \Psi^{-1}\{\psi_0\}\) for some \(\Psi\) and \(\psi_0\) and \(\Pi(H_0) = 0\), then the Bayes factor in favor of \(H_0\) with respect to \(\Pi_\gamma\) is

\[
\frac{m_{0T}(T(x))}{m_T(T(x))} = RB(\psi_0) \frac{E_{\Pi_0} \left( \frac{\pi(\theta | \psi_0, T(x))}{\pi(\theta | \psi_0)} \right)}{\pi(\theta | \psi_0)}
\]

(3.3)

where \(E_{\Pi_0}\) refers to expectation with respect to \(\Pi_0\).
Chapter 3. Hypothesis Assessment using Relative Belief Ratios

\[ m_{\Theta T}(T(x)) \]  
\[ = \frac{\int_{\Psi_0} f_{\Theta T}(T(x)) \pi_0(\theta) \nu_{\theta^{-1} \{\psi_0\}}(d\theta)}{m_T(T(x))} \]
\[ = \pi_{\Psi}(\psi_0 | T(x)) \int_{\Psi_0} f_{\Theta T}(T(x)) \pi(\theta | \psi_0, T(x)) \pi_0(\theta) \nu_{\theta^{-1} \{\psi_0\}}(d\theta) \]
\[ = \pi_{\Psi}(\psi_0 | T(x)) \int_{\Psi_0} \pi(\theta | \psi_0, T(x)) \pi_0(\theta) \nu_{\theta^{-1} \{\psi_0\}}(d\theta) \]
\[ = \pi_{\Psi}(\psi_0 | T(x)) \int_{\Psi_0} \pi(\theta | \psi_0, T(x)) \pi_0(\theta) \nu_{\theta^{-1} \{\psi_0\}}(d\theta) \]

**Proof.**

\[ \text{Corollary 1. If } \Pi_0 = \Pi(., | \psi_0), \text{ then } BF_{\Pi_0}(\psi_0) = RB(\psi_0). \]

**Proof.** Since \( \pi_0(\theta) = \pi(\theta | \psi_0) \), we have \( E_{\Pi_0}(\pi(\theta | \psi_0, T(x)) / \pi(\theta | \psi)) = 1 \) which establishes the result.

In general, Eq.(3.3) establishes the relationship between the Bayes factor when using the conditional prior \( \Pi(., | \psi_0) \) on \( H_0 \) and the Bayes factor when using an arbitrary prior \( \Pi_0 \) on \( H_0 \). The adjustment is the expected value, with respect to \( \Pi_0 \), of the conditional relative belief ratio \( \pi(\theta | \psi_0, T(x)) / \pi(\theta | \psi_0) \) for \( \theta \in H_0 \), given \( H_0 \). This can also be written as \( E_{\Pi(., | \psi_0, T(x))}(\pi_0(\theta) / \pi(\theta | \psi_0)) \) and so measures the discrepancy between the conditional priors given \( H_0 \) under \( \Pi \) and \( \Pi_\gamma \). So when \( \pi_0 \) is substantially different than \( \pi(., | \psi_0) \), we can expect a significant difference in the Bayes factors. To maintain consistency in the prior assignments, we require here that \( \Pi_0 \) equal \( \Pi(., | \psi_0) \) for some smooth \( \Psi \) and \( \psi_0 \). Note that in the discrete case, it is clear that we should choose \( \Pi_0 \) equal to \( \Pi(., | \psi_0) \).
Chapter 3. Hypothesis Assessment using Relative Belief Ratios

Also, in the continuous case, Jeffreys’ approach requires completely different modifications on $\Pi$ to obtain Bayes factors for different values of $\psi_0$. By contrast, $RB(\psi_0)$ is defined for every value of $\psi_0$ without any modification of $\Psi$. Clearly, there is no need to introduce the mixture prior if $\pi_0(\theta) = \pi(\theta | \psi_0)$. See Example 2 in Baskurt and Evans (2012) for a simple application of this.

3. So far, we have assumed there is a parameter of interest specified in the problem. Consider the Behrens-Fisher example, where we are testing whether the means of two normal populations are equal or not. In this case there are many $\Psi$ that generate $H_0$, e.g. $\Psi(\theta) = \theta_1 - \theta_2$, $\Psi(\theta) = \theta_1/\theta_2$, etc. So, the question is, when $\Pi(H_0) = 0$, and there is no particular parameter of interest, which of the many $\Psi$ functions that generate $H_0$ do we use to obtain $RB(\psi_0)$? Suppose that $\Psi$ is smooth, $H_0 = \Psi^{-1}\{\psi_0\}$ and $J_\Psi(\theta)$ is constant and positive for $\theta \in H_0$. We will refer to such a $\Psi$ as a constant volume distortion generator of $H_0$ and denote the class of such transformations by $T_{H_0}$. We have the following result.

**Theorem 5.** For every $\Psi \in T_{H_0}$, the conditional relative prior belief of $\theta_1, \theta_2 \in H_0$ given $\psi$ is $\pi(\theta_1)/\pi(\theta_2)$. Furthermore, the relative belief ratio of $H_0$ is independent of $\Psi \in T_{H_0}$.

**Proof.** The conditional prior beliefs about $\theta_1, \theta_2 \in H_0$ given $\Psi(\theta) = \psi$ is

$$
\frac{\pi(\theta_1 \mid \psi_0)}{\pi(\theta_2 \mid \psi_0)} = \frac{\pi(\theta_1)J_\Psi(\theta_1)/\pi_\Psi(\psi_0)}{\pi(\theta_2)J_\Psi(\theta_2)/\pi_\Psi(\psi_0)} \quad \text{from Eq. (1.10)}
$$

$$
= \frac{\pi(\theta_1)}{\pi(\theta_2)} \quad \text{since } \Psi \in T_{H_0}.
$$

This completes the first part of the theorem.

$$
RB(\psi_0) = \frac{\pi_\Psi(\psi_0 \mid x)}{\pi_\Psi(\psi_0)} = \frac{\int_{\Psi^{-1}\{\psi_0\}} f_{\theta T}(T(x)) \pi(\theta) J_\Psi(\theta) \nu_{\Psi^{-1}}(d\theta)/m_T(T(x))}{\int_{\Psi^{-1}\{\psi_0\}} \pi(\theta) J_\Psi(\theta) \nu_{\Psi^{-1}}(d\theta)}
$$

$$
= \frac{\int_{\Theta_0} f_{\theta T}(T(x)) \pi(\theta) \nu_{\Theta_0}(d\theta)/m_T(T(x))}{\int_{\Theta_0} \pi(\theta) \nu_{\Theta_0}(d\theta)}
$$
where the last equality follows from $\Theta_0 = \Psi^{-1}\{\psi_0\}$ and $\nu_{\Psi^{-1}\{\psi_0\}} = \nu_{\Theta_0}$ since this measure is determined by the geometry of $\Theta_0$ alone. It is seen from the last equality that $RB(\psi_0)$ is independent of $\Psi \in T_{H_0}$. This complete the second part of the theorem.

This says that when $\Psi \in T_{H_0}$, the volume distortions induced by $\Psi$, as measured by $J_{\Psi}(\theta)$, do not affect our conditional prior beliefs about $\theta \in H_0$ and these beliefs are essentially given by the values of $\pi(\theta)$ for $\theta \in H_0$. These results suggest that we choose a $\Psi$ that belongs to $T_{H_0}$, so that we do not have volume distortions affecting our inference. In other words, the relative belief ratio and the Bayes factor will be invariant to the choice of $\Psi$, if $\Psi \in T_{H_0}$.

**Example 9. Behrens-Fisher problem**

Suppose that $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ and $x, y$ are independent random samples form the corresponding distributions where $(\mu_i, \sigma_i^2) \in R \times R^+$ are unknown for $i = 1, 2$. We want to assess the hypothesis $H_0 : \mu_1 = \mu_2$. There are a number of different possible $\Psi$ that generate $H_0$. One common choice is $\psi = \Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \mu_1 - \mu_2$ and $H_0 = \Psi^{-1}\{0\}$. There are many other $\Psi \in T_{H_0}$ that lead to the same value of the relative belief ratio and the Bayes factor. Suppose $\psi = \Psi(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \mu_1/\mu_2$ for $H_0 = \Psi^{-1}\{1\}$. Then $J_{\Psi}(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = (1/\mu_1^2 + \mu_2^2/\mu_2^2)^{-1/2}$ which equals to $\mu/\sqrt{2}$ and this is not constant on $H_0$ when $\mu_1 = \mu_2 = \mu$. Therefore, $\psi = \mu_1 - \mu_2$ would a better choice since the relative belief ratio will be invariant to this choice of $\psi$.

See Example 3 in Baskurt and Evans (2012) for another example.

### 3.1.2 Evidential interpretation of relative belief ratios

This section is an extension of Section 1.2.1. In Bayesian inference, belief is measured by probability. Evidence is measured by how data change our beliefs. The Bayes factor or
relative belief ratio for \( H_0 = \Psi^{-1}\{\psi_0\} \) measures how our beliefs in \( H_0 \) have changed after seeing the data. If \( RB(\psi_0) > 1 \), the data increased our belief in \( H_0 \) and we have evidence in favor of \( H_0 \). Conversely, if \( RB(\psi_0) < 1 \), the data decreased our belief in \( H_0 \) and we have evidence against \( H_0 \). And we say that we have more evidence in \( \psi_1 \) compared to \( \psi_2 \), if \( RB(\psi_1) > RB(\psi_2) \). We will compare \( RB(\psi_0) \) to each of the possible values of \( RB(\psi) \) as part of assessing \( H_0 \).

The following theorem presents another property of \( RB(\psi_0) \) as a measure of the evidence in favor of \( H_0 \).

**Theorem 6.** \( RB(\psi_0) = E_{\Pi(\psi_0)}(RB(\theta)) \)

**Proof.** First note that \( RB(\theta) = f_{\theta T}(T(x)) \). Then, we have that

\[
RB(\psi_0) = \frac{m_T(T(x) \mid T(x))}{m_T(T(x))} = \frac{\int_{\Psi^{-1}\{\psi_0\}} f_{\theta T}(T(x)) \pi(\theta \mid \psi_0) \nu_{\Psi^{-1}\{\psi\}}(d\theta)}{m_T(T(x))} = \frac{\int_{\Psi^{-1}\{\psi_0\}} RB(\theta) \pi(\theta \mid \psi_0) \nu_{\Psi^{-1}\{\psi\}}(d\theta)}{E_{\Pi(\psi_0)}(RB(\theta))}.
\]

Theorem 6 says that evidence in favor of \( H_0 \) is obtained by averaging the evidence in favor of each value of the full parameter that makes \( H_0 \) true, using the conditional prior given that \( H_0 \) is true. Furthermore, based on the asymptotic distribution of the posterior density, under quite general conditions, we will have that \( RB(\psi_0) \to 0 \) when \( H_0 \) is false and, when \( H_0 \) is true, \( RB(\psi_0) \to \infty \) in the continuous case and \( RB(\psi_0) \to 1/\pi_{\Psi}(\psi_0) \) in the discrete case, as we increase the amount of data.

To measure the strength of evidence that \( RB(\psi_0) \) provides, we propose a posterior
tail probability, namely,

\[ \Pi_\psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) \]  

(3.4)

This is the posterior probability that the true value of \( \psi \) has a relative belief ratio no greater than \( RB(\psi_0) \). We mentioned in Section 1.2.1 that if we get a \( RB(\psi_0) > 1 \) and a big value for Eq.(3.4), we conclude the evidence in favor of \( H_0 \) is strong. If we get a \( RB(\psi_0) < 1 \) and a small value for Eq.(3.4), then we conclude the evidence against \( H_0 \) is strong. What if we get a \( RB(\psi_0) > 1 \) and a small value for Eq.(3.4), or a \( RB(\psi_0) < 1 \) and a big value for Eq.(3.4)? Before considering this, we present the following theorem.

**Theorem 7.** When \( RB(\psi_0) < 1 \), then

\[ \Pi_\psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) \leq RB(\psi_0) \]  

(3.5)

**Proof.** We have that

\[
\Pi_\psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) = \int_{\{\psi : RB(\psi) \leq RB(\psi_0)\}} \pi_\psi (\psi \mid T(x)) \nu_\psi (d\psi)
\]

\[
= \int_{\{\psi : RB(\psi) \leq RB(\psi_0)\}} RB(\psi) \pi_\psi (\psi) \nu_\psi (d\psi)
\]

\[
\leq \int_{\{\psi : RB(\psi) \leq RB(\psi_0)\}} RB(\psi_0) \pi_\psi (\psi) \nu_\psi (d\psi)
\]

\[
\leq RB(\psi_0)
\]

We see from Eq.(3.5) that when we have a small value of \( RB(\psi_0) \), then we have strong evidence against \( RB(\psi_0) \) and in fact, there is no need to compute Eq.(3.4). Note that while Eq.(3.5) always holds, it does not provide a useful bound when \( RB(\psi_0) > 1 \).

As previously discussed, when \( \Pi(\Psi^{-1}\{\psi\}) = 0 \), we can also interpret \( RB(\psi_0) \) as the Bayes factor with respect to \( \Pi \) in favor of \( H_0 \). Thus, Eq.(3.5) is also an a posteriori measure of the strength of the evidence provided by the Bayes factor. When \( \psi \) has a discrete distribution, we have the following result where we interpret \( BF(\psi) \) in the obvious way.
Corollary 2. If $\Pi_\Psi$ is discrete, then

$$
\Pi_\Psi (BF(\psi) \leq BF(\psi_0) \mid T(x)) \leq BF(\psi) \times E_{\Pi} \left( \{1 + \pi_\Psi(\Psi(\theta))(BF(\psi_0) - 1)\}^{-1} \right)
$$

Proof. Using (1.8) we have that $BF(\psi) \leq BF(\psi_0)$ if and only if $RB(\psi) \leq BF(\psi_0)/\{1 + \pi(\psi)(BF(\psi_0) - 1)\}$ and as in the proof of Theorem 7 this implies the inequality. Also $1 + \pi(\psi)(BF(\psi_0) - 1) \geq 1 + \max_\psi \pi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) \leq 1$ and $1 + \pi(\psi)(BF(\psi_0) - 1) \geq 1 + \min_\psi \pi(\psi)(BF(\psi_0) - 1)$ when $BF(\psi_0) > 1$ which completes the proof.

We see from Corollary 2 that a small value of $BF(\psi_0)$ is, in both the discrete and continuous case, strong evidence against $H_0$.

Suppose $RB(\psi_0) > 1$, so that we have evidence in favor of $H_0$, and that Eq.(3.4) is small. This tells us that there is a large posterior probability that the true value of $\psi$ has an even larger relative belief ratio and so this evidence in favor of $H_0$ does not seem to be strong. Alternatively, large values of Eq.(3.4) with $RB(\psi_0) > 1$ suggest that the evidence in favor of $H_0$ is strong, since there is a high posterior probability that the true value is in $\{\psi : RB(\psi) \leq RB(\psi_0)\}$ and based on the evidential preference ordering, $\psi_0$ is the best estimate in this set.

The interpretation of evidence in favor of $H_0$ is somewhat more involved than evidence against $H_0$ and the following example illustrates this.

Example 10. (Location normal)
Suppose we have a sample $x = (x_1, ..., x_n)$ from a $N(\mu, 1)$ distribution, where $\mu \in \mathbb{R}^1$ is unknown, so $T(x) = \bar{x}$, we take $\mu \sim N(0, \tau^2)$, $\Psi(\mu) = \mu$, and we want to assess $H_0 : \mu = 0$. Then $\bar{x}$ has $N(\mu, 1/n)$ and $\mu \mid \bar{x} \sim N(n\tau^2\bar{x}/(1 + n\tau^2), \tau^2/(1 + n\tau^2))$. The relative belief ratio at $H_0$ and the strength of evidence, $\Pi(RB(\mu) \leq RB(0) \mid x)$, are as
follows.

\[
RB(0) = \frac{\pi(0 \mid \bar{x})}{\pi(0)} = (1 + n\tau^2)^{1/2} \exp\{-n \left(1 + 1/n\tau^2\right)^{-1} \bar{x}^2/2\}
\]  

(3.6)

and

\[
\Pi(RB(\mu) \leq RB(0) \mid x) = 1 - \Phi\left((1 + 1/n\tau^2)^{1/2} (|\sqrt{n}\bar{x}| + (1 + n\tau^2)^{-1} \sqrt{n}\bar{x})\right)
\]

\[
\times \Phi\left((1 + 1/n\tau^2)^{1/2} (-|\sqrt{n}\bar{x}| + (1 + n\tau^2)^{-1} \sqrt{n}\bar{x})\right)
\]  

(3.7)

First, we look at what happens when we make the prior more diffuse via \(\tau^2 \to \infty\). It is straightforward to see that \(RB(0)\) in Eq. (3.6) goes to \(\infty\) as \(\tau^2 \to \infty\). That says the evidence in favor of \(H_0\) becomes arbitrarily large as we make the prior more diffuse. So we can bias the evidence \textit{a priori} in favor of \(H_0\) by choosing \(\tau^2\) very large. The reason \(RB(0)\) gives arbitrarily large evidence in favor of \(\mu = 0\) as we make the prior diffuse is that we are placing the bulk of the prior mass further and further away from \(\bar{x}\) as \(\tau^2\) gets larger. Then, \(\mu = 0\) looks more and more like a plausible value when compared to the values where the prior mass is being allocated. On the other hand, the strength of evidence Eq.(3.7) converges to \(1 - 2 \left(1 - \Phi(|\sqrt{n}\bar{x}|)\right)\) as \(\tau^2 \to \infty\) (see Appendix C for proof) and it may be a small value depending on the value of \(1 - 2 \left(1 - \Phi(|\sqrt{n}\bar{x}|)\right)\). Since this bias is induced by the value of \(\tau^2\), we need to address this issue \textit{a priori} and we present an approach to deal with this in Section 3.1.4.

We note that \(1 - 2 \left(1 - \Phi(|\sqrt{n}\bar{x}|)\right)\) is the frequentist p-value for this problem. \textit{Lindley’s paradox} (Lindley, 1957) refers to the situation in which classical frequentists methods and the Bayesian approach to a point-hypothesis testing do not agree. As large values of \(\tau^2\) are associated with noninformativity, these two approaches are expected to agree, however, diffuse priors lead to large values of \(RB(0)\) or \(BF(0)\) whereas the frequentist p-value can be very small. We argue that if we use the measure of strength of evidence in Eq.(3.7), then the paradox disappears as we can have evidence in favour of \(H_0\) while, at the same time, this evidence is not strong.
We also show that, for a fixed value of $RB(0)$, Eq.(3.7) converges to 0 as $n$ or $\tau^2$ get larger (see Appendix C). This says that a higher standard is set for establishing that a fixed value of $RB(0)$ is strong evidence in favour of $H_0$, as we increase the amount of data or make the prior more diffuse.

Second, we consider the behaviour of $RB(0)$ and the measure of strength of evidence as $n \to \infty$ (see Appendix C). For this we have that

$$RB(0) \to \begin{cases} \infty & H_0 \text{ true} \\ 0 & H_0 \text{ false,} \end{cases}$$

$$\Pi_\Psi (RB(\mu) \leq RB(0) | T(x)) \to \begin{cases} \mathcal{U}(0,1) & H_0 \text{ true} \\ 0 & H_0 \text{ false} \end{cases}$$

where $\mathcal{U}(0,1)$ is a uniform distribution on (0,1). We see that as the sample size increases, $RB(0)$ correctly identifies whether we have evidence against or in favor of $H_0$. And the strength of evidence goes to zero when $H_0$ is false, which tells us that we have strong evidence against $H_0$. However, when $H_0$ is true, although it is true that we will eventually have evidence in favor of $H_0$, the strength of this evidence is not always going to be strong at least measured by Eq.(3.4). That is, we could get a small value $\Pi_\Psi (RB(\mu) \leq RB(0) | T(x))$ which tells us the evidence in favor of $H_0$ is not strong, as other values of $\mu$ have larger relative belief ratios. If we have evidence against $H_0$, we should look at $\psi_{LRSE}(x)$ to see if we have detected a deviation from $H_0$ that is of practical importance.

We now give a numerical example. Suppose that $n = 50$, $\tau^2 = 400$ and we observe $\sqrt{n\bar{x}} = 1.96$. We plot $RB(\mu)$ versus $\mu$ in Figure 3.1 and present the results in Table 3.2.

<table>
<thead>
<tr>
<th>$RB(0)$</th>
<th>The strength of evidence (Eq. (3.7))</th>
<th>$\psi_{LRSE}(x)$</th>
<th>$RB(\mu_{LRSE}(x))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.72</td>
<td>0.05</td>
<td>0.28</td>
<td>141.40</td>
</tr>
</tbody>
</table>
Figure 3.1: Plot of $RB(\mu)$ against $\mu$ when $n = 50$, $\tau^2 = 400$ and $\sqrt{n\bar{x}} = 1.96$ in Example 10.

We see that we obtain $RB(0) = 20.72$, which represents strong evidence in favor of $H_0$ according to Jeffreys’ scale. However, we see that the strength of evidence is only 0.05, which states that there is a larger posterior probability that the true value has a larger relative belief ratio. So we cannot say that the evidence in favor of $H_0$ is strong. Then, we see that $\mu_{LRSE}(x) = 0.28$ and $RB(\mu_{LRSE}(x)) = 141.40$. Whether the difference between $\mu_{LRSE}(x) = 0.28$ and 0 is of practical importance or not depends on the application context. If the application suggests that a value of 0.28 is practically speaking close enough to 0 to be treated as 0, then it is reasonable to proceed as of $H_0$ is correct and this is supported by the value of the Bayes factor (20.72).

It is important that we have a clear idea of the size of a significant difference in any practical approach to hypothesis testing. We must always take into account practical significance when we have evidence against $H_0$. When we have evidence in favor of $H_0$,
the value of the strength of evidence in (3.4) tells us when it is necessary to do this.

When $\psi_0$ is false, we have $RB(\psi_0)$ and the strength of evidence,

$$\Pi_\Psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) \to 0$$
as the sample size increases. However, it is not always the case that $\Pi_\Psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) \to 1$ when $\psi_0$ is true. The following result shows that this is simply an artifact of continuity.

**Theorem 8.** Suppose that $\Theta = \{\theta_0, ..., \theta_k\}$, $\pi(\theta) > 0$ for each $\theta$, $H_0 = \Psi^{-1}\{\psi_0\}$ and $x = (x_1, ..., x_n)$ is a sample from $f_\theta$. Then we have that $\Pi_\Psi (RB(\psi) \leq RB(\psi_0) \mid T(x)) \to 1$ as $n \to \infty$ whenever $H_0$ is true.

**Proof.** See Appendix C

So if we think of continuous models as approximations to situations that are in reality finite, then we see that Eq.(3.4) may not be providing a good approximation. One possible solution is to use a metric $d$ on $\Psi$ and a distance $\delta$ such that $d(\psi, \psi') \leq \delta$ means that $\psi$ and $\psi'$ are practically indistinguishable. We can use this to discretize $\Psi$ and compute both the relative belief ratio for $H_0 = \{\psi : d(\psi, \psi_0) \leq \delta\}$ and its strength in this discretized version of the problem. It can be easily implemented computationally and is implicit in our computations. From a practical point-of-view, computing Eq.(3.4) and if it is small (as in Example 10), looking at $d(\psi_{LRSE}, \psi_0)$ to see if a deviation of any practical importance has been detected is a useful approach.

We emphasize that we should always quote $\psi_{LRSE}$ when we get evidence against $H_0$ as it gives an indication of whether we have detected a deviation from $H_0$ of practical significance or we have detected a very small difference because of a large sample size. Quoting $\psi_{LRSE}$ resolves this. On the other hand, when we have evidence in favor of $H_0$ and a small Eq.(3.4), we can still quote $\psi_{LRSE}$; however, the problem could be caused by the artifact of continuity. In this case the solution is to discretize according to a specified delta.
### 3.1.3 Are relative belief ratios incoherent?

In this section, we emphasize the fact that $RB(\psi)$ is a measure of evidence not belief, where belief is measured by probability and evidence is the change in beliefs after we see the data. Lavine and Schervish (1999) defined a coherence condition for any measure of evidence for a hypothesis, following the definition of coherence tests for nested hypothesis by Gabriel (1969): If $H_1$ implies $H_2$, then the support for $H_1$ can not be greater than the support for $H_2$. They argue via an example that the Bayes factor violates this condition, that is, $BF(H_1) > BF(H_2)$ when $H_1 \subset H_2$, and that it is not coherent measure of evidence.

We now illustrate a similar behaviour of the relative belief ratios with an example and explain the phenomena. We are particularly interested in the situation where $RB(H_1) > 1$ and $RB(H_2) < 1$ when $H_1 \subset H_2$.

**Example 11. (Evidence of a crime)**

Suppose that a murder is committed in a town and it is known that an adult member of a town with $m$ adult citizens committed the crime. The evidence from the crime scene indicates that a person of ethnic origin $E$ committed this crime. There are $m_1 < m$ adult members of this ethnic group residing in the town. Suppose further that the town has a university with $n$ students of which $n_1$ are of ethnic origin $E$. The values are summarized in Table 3.3. We assume uniform beliefs before the evidence is collected.

<table>
<thead>
<tr>
<th>University Student</th>
<th>Ethnic Origin</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E$</td>
<td>$E^c$</td>
</tr>
<tr>
<td>$U$</td>
<td>$n_1$</td>
<td>$n - n_1$</td>
</tr>
<tr>
<td>$U^c$</td>
<td>$m_1 - n_1$</td>
<td>$m - n - m_1 + n_1$</td>
</tr>
<tr>
<td></td>
<td>$m_1$</td>
<td>$m - m_1$</td>
</tr>
</tbody>
</table>
where $U$ denotes the set of university students in the city and $E$ denotes the set of adult members of the ethnic group. Then,

$H_1$: A university student of ethnic origin $E$ committed the crime and,

$H_2$: A university student committed the crime.

After the evidence is obtained, we get the following results. The probability that a university student committed the crime is, $P(U \mid E) = P(U \cap E)/P(E) = n_1/m_1$. And the probability that a university student of ethnic origin $E$ committed the crime is $P(U \cap E \mid E) = n_1/m_1$. Then $RB(H_1) = P(U \cap E \mid E)/P(U \cap E) = (n_1/m_1)/(n_1/m) = m_1/m_1$ and $RB(H_2) = P(U \mid E)/P(U) = (n_1/m_1)/(n/m) = n_1m/m_1n$. Note that $m_1/m_1 > 1$ but $n_1m/m_1n < 1$ when $n_1/n$ is small enough, namely, whenever the students of ethnic origin $E$ comprise a small enough fraction of the student population.

Here we see that there can be evidence in favor of the statement that a student of ethnic origin $E$ committed the crime while at the same time there is evidence against the statement that a student committed the crime. We do not view this result as an incoherent property of the relative belief ratio as, a very low proportion of students of ethnic origin $E$ to all students, weakens the evidence (correctly) that a student committed the crime. It would be unfair to indicate evidence in favor of a student having committed the crime unless students of ethnic origin $E$ comprise a big fraction of the student population.

We have the following lemma which explains further the phenomena.

**Lemma 1.** (Additivity property of the relative belief ratio)

If $A \subseteq B$ and $P(B \setminus A) = 0$, then $RB(A) = RB(B)$ and when $P(B \setminus A) > 0$

$$RB(B) = RB(A)P(A \mid B) + RB(B \setminus A)P(B \setminus A \mid B)$$
Proof.

\[ RB(B) = \frac{P(B \mid X)}{P(B)} = \frac{P(A \cap B \mid C) + P(A^c \cap B \mid C)}{P(B)} \]
\[ = \frac{P(A \cap B \mid X)}{P(A \cap B)} \frac{P(A \cap B)}{P(B)} + \frac{P(A^c \cap B \mid X)}{P(A^c \cap B)} \frac{P(A^c \cap B)}{P(B)} \]
\[ = RB(A)P(A \mid B) + RB(B \setminus A)P(B \setminus A \mid B) \quad (3.8) \]

It is seen from Eq.(3.8) that when \( H_1 \subset H_2 \), it is possible to get \( RB(H_2) < 1 \) and \( RB(H_1) > 1 \). We know that \( P(H_1) \leq P(H_2) \), which indicates our beliefs in \( H_1 \) and \( H_2 \), however, the relative belief ratio measures the change in beliefs, not beliefs.

### 3.1.4 A Priori Properties of the Relative Belief Ratio

In this Section, we consider the a priori behaviour of the relative belief ratio. First, we follow Royall (2000) and present the prior probability of getting a small value of \( RB(\psi_0) \) when \( H_0 \) is true, as we know this would be misleading evidence. We have the following result, where \( M_T \) denotes the prior predictive measure of the minimal sufficient statistic \( T \).

**Theorem 9.** The prior probability that \( RB(\psi_0) \leq q \), given that \( H_0 \) is true, is bounded above by \( q \), namely,

\[ M_T \left( \frac{m_T(t \mid \psi_0)}{m_T(t)} \leq q \mid \psi_0 \right) \leq q \quad (3.9) \]

**Proof.** Using Theorem 3, the prior probability that \( RB(\psi_0) \leq q \) is given by

\[
\Pi \times P_\theta \left( \frac{\pi_\Psi(\psi_0 \mid T(X))}{\pi_\Psi(\psi_0)} \leq q \mid \psi_0 \right) = \Pi \times P_\theta \left( \frac{m_T(T(X) \mid \psi_0)}{m_T(T(X))} \leq q \mid \psi_0 \right) 
\]
\[
= \int_{\{ \frac{m_T(t \mid \psi_0)}{m_T(t)} \leq q \}} m_T(t \mid \psi_0) \mu_T(dt) 
\]
\[
\leq \int_{\{ \frac{m_T(t \mid \psi_0)}{m_T(t)} \leq q \}} qm_T(t) \mu_T(dt) 
\]
\[
\leq q.
\]
Theorem 9 tells us that, \textit{a priori}, the relative belief ratio for $H_0$ is unlikely to be small when $H_0$ is true.

Theorem 9 is concerned with $RB(\psi_0)$ providing misleading evidence when $H_0$ is true. Again following Royall (2000), we also consider the prior probability that $RB(\psi_0)$ is large when $H_0$ is false, that is, when $\psi_0 \neq \psi_{\text{true}}$. We can calculate the prior probability that $RB(\psi_0) \geq q$ when $\theta \sim \Pi(\cdot \mid \psi_{\text{true}})$, $x \sim P_\theta$ and $\psi_0 \sim \Pi_\Psi$ independently of $(\theta, x)$ as discussed in Evans and Shakhatreh (2008). Here, $\psi_0$ is a false value in the generalized sense that it has no connection with the true value of the parameter and the data. We have the following result.

**Theorem 10.** The prior probability that $RB(\psi_0) \geq q$, when $\theta \sim \Pi(\cdot \mid \psi_0)$, $x \sim P_\theta$ and $\psi_0 \sim \Pi_\Psi$ independently of $(\theta, x)$, is bounded above by $1/q$.

\textit{Proof.}

\[
\begin{align*}
\Pi(\cdot \mid \psi_{\text{true}}) \times P_\theta \times \Pi_\Psi \left( \frac{\pi_\Psi(\psi_0 \mid T(x))}{\pi_\Psi(\psi_0)} \geq q \right) &= M_T(\cdot \mid \psi_{\text{true}}) \times \Pi_\Psi \left( \frac{\pi_\Psi(\psi_0 \mid t)}{\pi_\Psi(\psi_0)} \geq q \right) \\
&= \int_T \int_{\{\pi_\Psi(\psi_0 \mid t) / \pi_\Psi(\psi_0) \geq q\}} \pi_\Psi(\psi_0) m_T(t \mid \psi_{\text{true}}) \nu_{\Psi}(d\psi_0) \mu_T(dt) \\
&\leq \int_T \int_{\{\pi_\Psi(\psi_0 \mid t) / \pi_\Psi(\psi_0) \geq q\}} \pi_\Psi(\psi_0 \mid t) m_T(t \mid \psi_{\text{true}}) \nu_{\Psi}(d\psi_0) \mu_T(dt) \\
&\leq \frac{1}{q}.
\end{align*}
\]

Theorem 10 tells us that it is \textit{a priori} very unlikely that $RB(\psi_0)$ will be large when $\psi_0$ is a false value. This reinforces the interpretation that large values of $RB(\psi_0)$ are evidence in favor of $H_0$. 

\[\square\]
3.1.5 Assessing the Bias inherent in the Relative Belief Ratios

In Example 10, we looked at the behaviour of \( RB(\mu) \) as \( \tau^2 \to \infty \). We see that \( RB(\mu) \to \infty \) for every \( \mu \) (Eq.(C.1)). This suggests that the choice of prior possibly induces bias into the analysis in the sense that it makes it more likely to find evidence in favor of \( H_0 \) or against \( H_0 \). While the calibration of \( RB(\psi_0) \) in Eq.(3.4) takes into account the actual size of \( RB(\psi_0) \), it does not tell us whether the prior induces \textit{a priori} bias either for or against \( H_0 \). We propose assessing the bias against \( H_0 \) in the prior by

\[
MT \left( \frac{m_t(t | \psi_0)}{m_T(t)} \leq 1 | \psi_0 \right). \tag{3.10}
\]

If Eq.(3.10) is large, then we have a \textit{a priori} small chance of detecting evidence in favor of \( H_0 \) when \( H_0 \) is true. Obviously, we would want Eq.(3.10) to be small. We can use Eq.(3.10) as a design tool by choosing the sample size to make Eq.(3.10) small.

Similarly, we propose assessing the bias in favor of \( H_0 \) in the prior by

\[
MT \left( \frac{m_t(t | \psi_0)}{m_T(t)} \leq 1 | \psi_* \right). \tag{3.11}
\]

for various values of \( \psi_* \neq \psi_0 \) that represent practically significant deviations from \( \psi_0 \). If Eq.(3.11) is small, then this indicates that the prior is biasing the evidence in favor of \( \psi_0 \). Again, we can use this as a design tool by choosing the sample size so that Eq.(3.11) is large.

We illustrated how to assess the bias inherent in the relative belief ratio using the \textit{location normal} example in Baskurt and Evans (2013). In the next section, we present an application of these concepts to a two-way analysis.

Overall we believe that priors should be based on beliefs and elicited, and the assessment of \textit{a priori} bias as well as assessment for priori-data conflict, discussed in Section 1.2.2 are necessary steps to be taken in a Bayesian inference. Ideally, \textit{a priori} bias assessment should be done at the design stage, but even if it is done \textit{post hoc}, this seems preferable to just ignoring the possibility that such biasing can occur.


3.2 Two-way Analysis of Variance

To illustrate the results of this chapter, we consider testing for no interaction in a two way ANOVA. Suppose we have two categorical factors $A$ and $B$, and observe $x_{ijk} \sim N(\mu_{ij}, \nu^{-1})$ for $1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq n_{ij}$. A minimal sufficient statistic is given by $T(x) = (\bar{x}, s^2)$ where

\[
\bar{x} \sim N_{ab}(\mu, \nu^{-1}D^{-1}(n))
\]
\[
(n_. - ab)s^2 \sim \text{Gamma}_{\text{rate}}((n_. - ab)/2, (2\nu)^{-1})
\]

where $\bar{x}$ and $s^2$ are independent, $(n_. - ab)s^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} (x_{ijk} - \bar{x}_{ij})^2$, $D(n) = \text{diag}(n_{11}, n_{12}, \ldots, n_{ab})$ and $n_. = \sum_{i,j} n_{ij}$.

Suppose we use the conjugate priors $\mu | \nu \sim N_{ab}(\mu_0, \nu^{-1}\Sigma_0)$ with $\Sigma_0 = \tau_0^2 I$, and $\nu \sim \text{Gamma}_{\text{rate}}(\alpha_0, \beta_0)$. Then, we have the posterior,

\[
\nu | x \sim \text{Gamma}_{\text{rate}}(\alpha_0(x), \beta_0(x))
\]
\[
\mu | \nu, x \sim N_{ab}(\mu_0(x), \nu^{-1}\Sigma_0(x))
\]

where

\[
\mu_0(x) = \Sigma_0(x)(D(n)\bar{x} + \tau_0^{-2}\mu_0),
\]
\[
\Sigma_0(x) = (D(n) + \tau_0^{-2}I)^{-1},
\]
\[
\alpha_0(x) = \alpha_0 + (n_. - ab)/2,
\]
\[
\beta_0(x) = \beta_0 + (\bar{x} - \mu_0)'(D^{-1}(n) + \tau_0^2 I)^{-1}(\bar{x} - \mu_0)/2 + (n_. - ab)s^2/2.
\]

As is common in this situation, we test first for interactions between $A$ and $B$ and, if no interactions are found then we test for any main effects. For this we let $C_A = (c_{A1} c_{A2} \ldots c_{Aa}) \in R^{a \times a}, C_B = (c_{B1} c_{B2} \ldots c_{Bb}) \in R^{b \times b}$ denote contrast matrices (orthogonal and first column constant) for $A$ and $B$, respectively, and put $C = C_A \otimes C_B = (c_{11} c_{12} \ldots c_{ab})$ where $c_{ij} = c_{Ai} \otimes c_{Bj}$ and $\otimes$ denotes Kronecker product. Here is a numerical example showing how contrasts work.
Example 12. (Two-way ANOVA) Suppose \( a = 3 \) and \( b = 2 \), then

\[
C_A = \begin{pmatrix}
1/\sqrt{3} & -1/\sqrt{2} & -1/\sqrt{6} \\
1/\sqrt{3} & 1/\sqrt{2} & -1/\sqrt{6} \\
1/\sqrt{3} & 0 & 2/\sqrt{6}
\end{pmatrix}, \quad C_B = \begin{pmatrix}
1/\sqrt{2} & -1/\sqrt{2} \\
1/\sqrt{2} & 1/\sqrt{2}
\end{pmatrix}
\]

and

\[
C = (C_A \otimes C_B) = \begin{pmatrix}
(c_{A1} \otimes c_{B1}) & (c_{A1} \otimes c_{B2}) & (c_{A2} \otimes c_{B1}) & (c_{A2} \otimes c_{B2}) & (c_{A3} \otimes c_{B1}) & (c_{A3} \otimes c_{B2})
\end{pmatrix}
\]

The contrasts are denoted by \( \alpha = C' \mu \) where \( \alpha_{ij} = c'_{ij} \mu \), \( \alpha = [\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}, \alpha_{31}, \alpha_{32}]' \) and \( \mu = [\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \mu_{31}, \mu_{32}]' \). Suppose we are interested in the main effect of \( B \), that is, we want to see if there is a difference between the two levels of \( B \). It can be seen that \( c'_{12} = c_{A1} \otimes c_{B2} = [-1/\sqrt{6}, 1/\sqrt{6}, -1/\sqrt{6}, 1/\sqrt{6}, -1/\sqrt{6}, 1/\sqrt{6}] \). Then \( \alpha_{12} = c'_{12} \mu \) yields;

\[
\alpha_{12} = -\mu_{11} + \mu_{12} - \mu_{21} + \mu_{22} - \mu_{31} + \mu_{32}.
\]

Then \( H_0: \alpha_{12} = 0 \) is equivalent to \( H_0: \mu_{11} + \mu_{21} + \mu_{31} = \mu_{12} + \mu_{22} + \mu_{32} \).

The contrasts \( \alpha = C' \mu \) have the following joint prior and posterior distributions,

\[
\alpha | \nu \sim N_{ab}(C' \mu_0, \nu^{-1} C' \Sigma_0 C) = N_{ab}(C' \mu_0, \nu^{-1} \Sigma_0), \text{since } C \text{ is orthogonal}
\]

\[
\alpha | \nu, y \sim N_{ab}(C' \mu_0(y), \nu^{-1} C' \Sigma_0(x) C).
\]

From this we obtain the marginal prior and posterior distributions of the contrasts,

\[
\alpha \sim \text{Student}_a(2 \alpha_0, C' \mu_0, (\beta_0/\alpha_0)C' \Sigma_0 C), \quad \text{(3.12)}
\]

\[
\alpha | x \sim \text{Student}_a(2 \alpha_0(x), C' \mu_0(x), (\beta_0(x)/\alpha_0(x))C' \Sigma_0(x) C), \quad \text{(3.13)}
\]

where we say \( w \sim \text{Student}_k(\lambda, m, M) \) with \( m \in \mathbb{R}^k \) and \( M \in \mathbb{R}^{k \times k} \) positive definite, when \( w \) has density

\[
\frac{\Gamma((\lambda + k)/2)}{\Gamma(\lambda/2)\Gamma(k/2)} (\det(M))^{-1/2} (1 + (w - m)' M^{-1}(w - m)/\lambda)^{-(\lambda+k)/2} \lambda^{-k/2}
\]
on $R^k$. Recall that, if $w \sim \text{Student}_k(\lambda, m, M)$ then, for distinct $i_j$ with $1 \leq j \leq l \leq k$, we have that $(w_{i_1}, \ldots, w_{i_l}) \sim \text{Student}_l(\lambda, m(i_1, \ldots, i_l), M(i_1, \ldots, i_l))$ where $m(i_1, \ldots, i_l)$ and $M(i_1, \ldots, i_l)$ are formed by taking the elements of $m$ and $M$ as specified by $(i_1, \ldots, i_l)$.

We will now continue with Example 12 to show the rest of the calculations.

**Example 13. (Two-way ANOVA - cont. Example 12)**

We want to assess whether an interaction between $A$ and $B$ exists. This corresponds to $H_0 : \psi = (\alpha_{22}, \alpha_{32}) = (0, 0)$ where $\psi = \Psi(\mu, \nu^{-1})$. We can easily get the marginal prior and posterior density of $\psi = (\alpha_{22}, \alpha_{32})$ from Eq.(3.12) and Eq.(3.13), respectively and obtain the exact expression for $RB(0)$ (3.14) and compute $\Pi_\Psi(RB(\psi) \leq RB(0) \mid T(x))$ by simulations as shown below.

\[
RB(0) = \frac{\text{Student}_2(2\alpha_0(x), \mu_{22,32}(x), \Sigma_{22,32}(x))}{\text{Student}_2(2\alpha_0, \mu_{22,32}, \Sigma_{22,32})}
\]  
(3.14)

where $\mu_{22,32}$ and $\Sigma_{22,32}$ are the sub-vectors of $C'\mu_0$ and $(\beta_0/\alpha_0)C'\Sigma_0C$ in Eq.(3.12) respectively, and $\mu_{22,32}(x)$ and $\Sigma_{22,32}(x)$ are the sub-vectors of $C'\mu_0(x)$ and $(\beta_0(x)/\alpha_0(x))C'\Sigma_0(x)C$ in Eq.(3.13) respectively, for the corresponding indices.

We follow the steps below to complete the analysis.

1. Choose a prior and determine $\alpha_0, \beta_0, \mu_0$ and $\tau_0$.

2. Calculate $T(x) = (\bar{x}, s^2)$ from the data at hand. Calculate $RB(\psi_0)$ from Eq.(3.14) to assess the evidence for the hypothesis.

3. Generate $\psi^{(i)} = (\alpha_{22}^{(i)}, \alpha_{32}^{(i)})$ from its prior and posterior densities obtained from Eq.(3.12) and Eq.(3.13) for $i = 1, \ldots 10^4$

4. Evaluate $RB(\psi^{(i)}) = \Pi_\Psi(\psi^{(i)} \mid T(x))/\Pi_\Psi(\psi^{(i)})$ and estimate the strength of evidence by

\[
\frac{1}{10000} \sum_{j=1}^{10000} \mathcal{I}(RB(\psi^{(i)}) \leq RB(0)).
\]  
(3.15)
5. Check whether the prior chosen induces any bias in favour or against the hypothesis.

To evaluate the *a priori* bias against $H_0$ based on a given prior, we need to compute $M_T(RB(0) \leq 1 \mid (\alpha_{22}, \alpha_{32}))$. For this we need to be able to generate $T(x) = (\bar{x}, s^2)$ from the conditional prior predictive $M_T(\cdot \mid (\alpha_{22}, \alpha_{32}))$. This is easily accomplished by generating $(\mu, \nu)$ from the conditional prior given $(\alpha_{22}, \alpha_{32})$ and then generating $\bar{x} \sim N_6(\mu, \nu^{-1}D^{-1}(n))$ independent of $(n_\cdot - ab)s^2 \sim \text{Gamma}_{\text{rate}}((n_\cdot - ab)/2, (2\nu)^{-1})$.

For this we need the conditional prior distribution of $\mu$ given $\nu$ and $(\alpha_{22}, \alpha_{32})$. We have that $\alpha = C'\mu$ and $\mu = C\alpha$. As noted above, $\alpha \mid \nu \sim N_6(C'\mu_0, \nu^{-1}C'S_0C)$ and so we can generate $\mu$ from this conditional distribution by generating $\alpha$ from the conditional distribution obtained from the $N_6(C'\mu_0, \nu^{-1}S_0)$ distribution by conditioning on $(\alpha_{22}, \alpha_{32})$ and putting $\mu = C\alpha$.

Here are the summary of the steps to compute $M_T(RB(0) \leq 1 \mid (\alpha_{22}, \alpha_{32}))$ where $i = 1, ..., 10000$.

1. Generate $\nu^{(i)} \sim \text{Gamma}_{\text{rate}}(\alpha_0, \beta_0)$ and $\alpha^{(i)} \mid \nu^{(i)} \sim N_6(C'\mu_0, \nu^{-1}C'S_0C)$.

2. Note that we need to generate $\mu \mid \nu$ conditional on $\alpha$ where $(\alpha_{22}, \alpha_{32})$ are fixed. For this we take the $\alpha^{(i)}$ generated in Step 1, replace the values $(\alpha_{22}, \alpha_{32})$ with the desired values then set $\mu^{(i)} = C\alpha^{(i)}$ since the contrasts are *a priori* independent given $\nu$.

3. Generate data by $\bar{x}^{(i)} \mid \mu^{(i)}, \nu^{(i)} \sim N_6(\mu^{(i)}, (\nu^{(i)})^{-1}D^{-1}(n))$ independent of $(n_\cdot - ab)(s^2)^{(i)} \mid \nu^{(i)} \sim \text{Gamma}_{\text{rate}}((n_\cdot - ab)/2, (2\nu^{(i)})^{-1})$.

4. Calculate the posterior distribution of $\alpha^{(i)}$ using Eq.(3.13) then obtain the marginal posterior distribution of $(\alpha_{22}^{(i)}, \alpha_{32}^{(i)})$.

5. Calculate $RB(0)^{(i)}$ and estimate $M_T(RB(0) \leq 1 \mid (\alpha_{22}, \alpha_{32}))$ by

$$
\frac{1}{10000} \sum_{j=1}^{10000} I(RB(0)^{(i)} \leq 1).
$$

(3.16)
As a specific numerical example, we take \((n_{11}, n_{12}, n_{21}, n_{22}, n_{31}, n_{32}) = (55, 50, 45, 43, 56, 48)\), \(\mu_0 = 0, \alpha_0 = 3, \beta_0 = 3\).

The prior for \(\nu^{-1}\) has mean 1.5 and variance 2.25 and we now consider the choice of \(\tau_0^2\) as this has the primary effect on the \textit{a priori} bias for \(H_0\). In the first row of Table 3.4 we present the values of the \textit{a priori} bias against \(H_0\) for several values of \(\tau_0^2\) and see that the bias against \(H_0\) is large when \(\tau_0^2\) is small. In the subsequent rows of Table 2 we present the bias in favor of \(H_0\) when \(H_0\) is false. For this we record \(M_T(RB(0) \leq 1 \mid \alpha_{22}^2 + \alpha_{32}^2 = \delta)\) for various \(\delta\) so we are averaging over all \((\alpha_{22}, \alpha_{32})\) that are the same distance from \(H_0\). To generate \(T(x) = (\bar{x}, s^2)\) from \(M_T(\cdot \mid \alpha_{22}^2 + \alpha_{32}^2 = \delta) = \int_{\{\alpha_{22}^2 + \alpha_{32}^2 = \delta\}} M_T(\cdot \mid \alpha_{22}, \alpha_{32})\pi(\alpha_{22}, \alpha_{32} \mid \alpha_{22}^2 + \alpha_{32}^2 = \delta) \, d\alpha_{22} d\alpha_{32}\), we generate \((\alpha_{22}, \alpha_{32})\) from the conditional prior given \(\alpha_{22}^2 + \alpha_{32}^2 = \delta\), and this is a uniform on the circle of radius \(\delta^{1/2}\), and then generate from \(M_T(\cdot \mid \alpha_{22}, \alpha_{32})\). As expected, we see that there is bias in favor of \(H_0\) only when \(\tau_0^2\) is large and we are concerned with detecting values of \((\alpha_{22}, \alpha_{32})\) that are close to \(H_0\).
Table 3.4: Values of $M_T(RB(0) \leq 1 \mid \alpha_{22}^2 + \alpha_{32}^2 = \delta)$ for various $\delta$ and $\tau_0^2$ in two-way analysis.

<table>
<thead>
<tr>
<th>$\tau_0^2$</th>
<th>0.01</th>
<th>0.05</th>
<th>0.08</th>
<th>0.10</th>
<th>0.50</th>
<th>5.00</th>
<th>10.00</th>
<th>100.00</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.00$</td>
<td>0.53</td>
<td>0.28</td>
<td>0.26</td>
<td>0.24</td>
<td>0.16</td>
<td>0.10</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.05$</td>
<td>0.74</td>
<td>0.58</td>
<td>0.51</td>
<td>0.47</td>
<td>0.30</td>
<td>0.19</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.10$</td>
<td>0.85</td>
<td>0.77</td>
<td>0.71</td>
<td>0.67</td>
<td>0.46</td>
<td>0.27</td>
<td>0.24</td>
<td>0.17</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.20$</td>
<td>0.95</td>
<td>0.93</td>
<td>0.91</td>
<td>0.90</td>
<td>0.74</td>
<td>0.50</td>
<td>0.43</td>
<td>0.31</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.30$</td>
<td>0.98</td>
<td>0.98</td>
<td>0.97</td>
<td>0.90</td>
<td>0.69</td>
<td>0.62</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.40$</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.97</td>
<td>0.84</td>
<td>0.78</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.50$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.93</td>
<td>0.89</td>
<td>0.73</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.60$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.97</td>
<td>0.95</td>
<td>0.84</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 0.80$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>0.95</td>
</tr>
<tr>
<td>$\alpha_{22}^2 + \alpha_{32}^2 = 1.00$</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Suppose now that our prior beliefs lead us to choose $\tau_0^2 = 0.10$. In Table 3.5, we present some selected cases of assessing $H_0$ based on simulated data sets where the data is generated in such a way that we know there is no prior-data conflict. Recall that $\psi = (\alpha_{22}, \alpha_{32})$ and Eq.(3.4) is measuring the strength of the evidence that $\psi = 0$. For the first 4 cases $H_0$ is true and we always get evidence in favor $H_0$. Notice that in case 4, where we only have marginal evidence in favor, the strength of this evidence is also quite low (recall that strong means Eq.(3.4) is small when we have evidence against and Eq.(3.4) is big when we have evidence in favor). In cases 5 and 6, the hypothesis $H_0$ is marginally false and we get strong evidence against in one of these cases. The other cases indicate that we can still get misleading evidence (evidence in favor when $H_0$ is false) but the strength of the evidence is not large in these cases. Also, as we increase the effect size, the evidence becomes more definitive against $H_0$ and also stronger. Overall we see that, based on the sample sizes and the prior, we never get evidence in favor of $H_0$ that
can be considered extremely strong when $H_0$ is false. In case 3, we get the most evidence in favor of $H_0$ but Eq.(3.4) says that the posterior probability of the true value having a larger relative belief value is 0.45. The best estimate of the true value in this case is $\psi_{LRSE}(x) = (-0.02, -0.12)$ with $RB(\psi_{LRSE}(x)) = 8.62$. Depending on the application, these values can add further support to evidence in favor of $H_0$.

Table 3.5: Values of $RB(0), \Pi_\psi(RB(\psi) \leq RB(\psi_0) \mid T(x)), \psi_{LRSE}(x)$ and $RB(\psi_{LRSE}(x))$ in various two-way analyses.

<table>
<thead>
<tr>
<th>Case</th>
<th>$\psi_{true}$</th>
<th>$RB(0)$</th>
<th>Eq.(3.4)</th>
<th>$\psi_{LRSE}(x)$</th>
<th>$RB(\psi_{LRSE}(x))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(0.00, 0.00)</td>
<td>3.50</td>
<td>0.62</td>
<td>(0.10, 0.11)</td>
<td>5.10</td>
</tr>
<tr>
<td>2</td>
<td>(0.00, 0.00)</td>
<td>3.16</td>
<td>0.22</td>
<td>(−0.10, −0.13)</td>
<td>12.76</td>
</tr>
<tr>
<td>3</td>
<td>(0.00, 0.00)</td>
<td>5.11</td>
<td>0.55</td>
<td>(−0.02, −0.12)</td>
<td>8.62</td>
</tr>
<tr>
<td>4</td>
<td>(0.00, 0.00)</td>
<td>1.22</td>
<td>0.17</td>
<td>(−0.14, −0.32)</td>
<td>5.59</td>
</tr>
<tr>
<td>5</td>
<td>(0.01, 0.01)</td>
<td>3.07</td>
<td>0.55</td>
<td>(−0.09, −0.16)</td>
<td>4.94</td>
</tr>
<tr>
<td>6</td>
<td>(0.01, 0.01)</td>
<td>0.09</td>
<td>0.00</td>
<td>(−0.22, 0.18)</td>
<td>25.60</td>
</tr>
<tr>
<td>7</td>
<td>(0.10, 0.10)</td>
<td>0.02</td>
<td>0.00</td>
<td>(0.36, 0.05)</td>
<td>24.75</td>
</tr>
<tr>
<td>8</td>
<td>(0.10, 0.10)</td>
<td>1.96</td>
<td>0.35</td>
<td>(0.24, −0.17)</td>
<td>4.42</td>
</tr>
<tr>
<td>9</td>
<td>(0.20, 0.20)</td>
<td>0.04</td>
<td>0.00</td>
<td>(0.19, 0.35)</td>
<td>19.28</td>
</tr>
<tr>
<td>10</td>
<td>(0.20, 0.20)</td>
<td>1.84</td>
<td>0.11</td>
<td>(0.13, 0.15)</td>
<td>14.74</td>
</tr>
<tr>
<td>11</td>
<td>(0.30, 0.30)</td>
<td>0.27</td>
<td>0.02</td>
<td>(0.22, 0.23)</td>
<td>14.55</td>
</tr>
<tr>
<td>12</td>
<td>(0.30, 0.30)</td>
<td>0.00</td>
<td>0.00</td>
<td>(0.23, 0.31)</td>
<td>32.12</td>
</tr>
</tbody>
</table>

### 3.3 Summary and Discussion

We have shown that, when a hypothesis $H_0$ has 0 prior probability with respect to a prior on $\Theta$, a Bayes factor and a relative belief ratio of $H_0$ can be defined via limits without the need to introduce a discrete probability on $H_0$. We have argued that the evidence in favor of or against $H_0$ is assessed by computing the relative belief ratio
$RB(\psi_0)$ and the measure of the strength of this evidence via Eq.(3.4). We also advocate looking at $(\psi_{LRSE}(x), RB(\psi_{LRSE}(x)))$ as part of hypothesis assessment, where the value of $RB(\psi_{LRSE}(x))$ tells us the maximum increase in belief for any value of $\psi$. Furthermore, we have presented an approach to assessing the a priori bias induced by a particular prior, either in favor or against a hypothesis, and have shown that this can be controlled via experimental design.

$RB(\psi_0)$ could be considered as a standardized integrated likelihood, however, we cannot multiply $RB(\psi_0)$ by a positive constant as we can do with a likelihood, since it destroys the role of the relative belief ratio as a measure of evidence and we also lose many inequalities we have derived. We also would like to emphasize that the measure of strength of evidence in Eq.(3.4) is not to be interpreted as a P-value.
Chapter 4

Logistic Regression Analysis via Relative Belief

In this chapter, we present the application of relative belief inferences to univariate binary data based on the logistic regression model. We suppose that we have a binary response $Y = 0, 1$ related to $k$ quantitative predictors $(X_1, X_2, ..., X_k)$ via

$$\log \frac{p}{1-p} = \sum_{i=1}^{k} \beta_i x_i,$$  \hspace{1cm} (4.1)

where $p = P(Y|X_1 = x_1, ..., X_k = x_k)$ and $(\beta_1, ..., \beta_k) \in \mathbb{R}^k$.

There are two main problems we consider here. First, we are concerned with the overall fit of the model. It is necessary in general to have methods for checking this model after observing data $(x_{1i}, ..., x_{ki}, y_i)$ for $i = 1, ..., n$. Second, given the nonlinear relationship in Eq. (4.1), it is not obvious what an appropriate prior is for $(\beta_1, ..., \beta_k)$. It is common to put diffuse normal priors on the $\beta_i$, however, Evans and Jang (2011a) demonstrate that this is not accomplishing the goal of being noninformative. In fact, as the diffuseness of a prior grows, the prior can become more informative and not less.

We discussed in Chapter 3 that the relative belief ratio for a hypothesis $H_0$, $RB(H_0)$, measures the evidence that $H_0$ is true. We also presented a posterior measure of the strength of this evidence. In Section 3.1, when $\Pi(H_0) = 0$, we explained how to compute
Chapter 4. Logistic Regression Analysis via Relative Belief

When there is a parameter of interest, $\Psi$, specified in the problem and when there is no particular parameter of interest (see Example 9). Basically, when the parameter of interest that generates $H_0$ is not clearly specified in the problem, as occurs in most goodness-of-fit approaches, we want to choose a $\Psi \in \mathcal{T}_{H_0}$ so that the relative belief ratio and the Bayes factor are invariant to the choice of $\Psi$ (see Theorem 5).

In Section 4.1, we discuss a general method, called the method of concentration (Evans, Gilula, and Guttman (1993) and Evans, Gilula, Guttman, and Swartz (1997)), for computing $RB(H_0)$ in such circumstances. The method basically compares the concentration of the prior around the subset of the parameter space specified by $H_0 \subset \Theta$ with the concentration of the posterior around this subset by choosing a measure of distance $\Psi = d_{H_0}$ of $\theta$ from $H_0$, such that $d_{H_0}(\theta) = 0$ if and only if $\theta \in H_0$. We then discuss a computational difficulty in calculating $RB(H_0)$ when both $\pi_{\psi}(0) = 0$ and $\pi_{\psi}(0 | x) = 0$.

In Section 4.2, we use this approach to obtain a Bayesian goodness-of-fit test for logistic regression based on a weakly informative prior and present examples where we apply the goodness-of-fit test when the model is correct and when the model is wrong. When no lack of fit detected, we show that this prior induces a prior on $(\beta_1, \ldots, \beta_k)$, which avoids some problems that the commonly used priors on $\beta$ possess. The priors presented here are basically from the family of multivariate normal priors, but one has to be careful with the scaling of such prior to avoid problems.

### 4.1 Method of Concentration

There is no guarantee that we can find a $\Psi \in \mathcal{T}_{H_0}$, when the parameter of interest is not specified in the problem. In this section, we illustrate how the method of concentration can be use to choose a $\Psi$ that generates $H_0$. If $H_0$ is true, then we expect the observed data to lead to the posterior distribution of $\theta$ being more concentrated about $H_0$ than the prior distribution of $\theta$. To measure the concentration of a distribution about $H_0$ we
choose a measure \( d_{H_0}(\theta) \) of the distance of \( \theta \) from \( H_0 \), such that \( d_{H_0}(\theta) = 0 \) if and only if \( \theta \in H_0 \), and then see how closely the distribution of \( d_{H_0}(\theta) \) concentrates about 0. Note that \( \Psi(\theta) = d_{H_0}(\theta) \) generates \( H_0 \) via \( H_0 = \Psi^{-1}\{0\} \). We present an example where a natural choice of \( d_{H_0}(\theta) \) arises from \( H_0 \).

**Example 14. Squared Euclidean distance - Linear regression**

Suppose that \( \Theta \) is an open subset of \( \mathbb{R}^k \) and let \( d_{H_0}(\theta) = ||\theta - \theta_0(\theta)||^2 \) where \( \theta_0(\theta) \) is a point in \( H_0 \) that is closest to \( \theta \). For example, suppose \( Y \in \mathbb{R}^n \), \( \mu = E(Y \mid x) \), \( \text{Var}(Y) = \sigma^2 I \) and we want to assess the linearity assumption, namely, \( H_0 : \mu = X\beta \in \mathcal{L}(\mathcal{X}) \) for some \( X \in \mathbb{R}^{n \times k} \) of rank \( k \) and the columns of \( X \) form a basis for \( H_0 \). Then \( P = X(X^tX)^{-1}X^t \) is an orthogonal projection matrix on \( \mu \in H_0 \). Using squared Euclidean distance \( d_{H_0}(\mu) = \Psi(\mu) = ||\mu - P\mu||^2 = \mu^t(I - X(X^tX)^{-1}X^t)\mu \) where \( P\mu \) is the closest point in \( \mathcal{L}(\mathcal{X}) \) to \( \mu \in \mathbb{R}^n \). So, \( H_0 = \Psi^{-1}\{0\} \) and \( \Psi'(\mu) = 2\mu^t(I - X(X^tX)^{-1}X^t) \) and \( J_{\Psi}(\theta) = ||\mu - P\mu||^{-1}/\sqrt{2} = \Psi^{-1/2}(\mu)/\sqrt{2} \) so \( \Psi(\mu) \) has constant volume distortion. This generalizes to affine subspaces of \( \mathbb{R}^k \).

While there is nothing that forces us to choose \( d_{H_0} \) to be squared Euclidean distance, this often has some computational convenience and can be considered as a generalization of the use of variance in statistics as a measure of spread. In Evans, Gilula, and Guttman (1993) and Evans, Gilula, Guttman, and Swartz (1997) \( d_{H_0} \) was taken to be squared Euclidean distance and the prior and posterior distributions of \( \Psi \) were compared graphically. Here we propose a more formal approach via computing \( RB(H_0) \) and the strength of evidence.

### 4.1.1 Computations

When using the method of concentration, namely, \( \Psi(\theta) = d_{H_0}(\theta) \), a computational difficulty arises as it can happen that both \( \pi_\Psi(0) = 0 \) and \( \pi_\Psi(0 \mid x) = 0 \) and we don’t have closed form expressions for \( \pi_\Psi(\psi) \) and \( \pi_\Psi(\psi \mid x) \) to assist in the computation of
the limiting ratio $RB(0)$. In this situation it makes sense to approximate $RB(0)$ by $RB(d_{H_0}(\theta) \leq \psi_*)$ for small $\psi_*$. Since we will commonly use simulation methods, the fact that the prior and posterior densities vanish at 0 indicates there will not be many sampled values of $\psi$ near 0 from either the prior or posterior. So our approach is to choose $\psi_*$ to be a left tail quantile of $\Pi_\Psi$ that can be reliably estimated, e.g., the 0.05 quantile. Furthermore, we will approximate the measure of strength of evidence in Eq.(3.4) via a discrete approximation to $\Pi_\Psi(RB(\psi) \leq RB(\psi_*) \mid x)$ that we subsequently describe as part of our computational algorithm. Actually, rather than relying on a single choice of $\psi_*$, it makes sense to look at $\Pi_\Psi(RB(\psi) \leq RB(\psi_*) \mid x)$ for several choices of $\psi_*$ in the left tail of the prior, to determine how the posterior distribution of $\Psi$ has concentrated about 0 relative to its prior distribution.

We use the following algorithm for the approximations when using $\Psi = d_{H_0}$. This requires that we have available exact or approximate samplers for both $\Pi_\Psi$ and $\Pi_\Psi(\cdot \mid x)$. Let $N$ be a positive integer and $\psi_{i/N}$ denote the $i/N$-th prior quantile of $\Psi$ for $i = 0, \ldots, N$ where $\psi_0 = 0$ and $\psi_1 = \infty$. Also, let $\hat{F}$ denote an empirical prior cdf, $\hat{F}(\cdot \mid x)$ denote an empirical posterior cdf, $\psi_* = \psi_{i_0/N}$ and $i_0 \in \{1, \ldots, N\}$.

(1) Select $M_1$ and generate a sample $\psi_1, \ldots, \psi_{M_1}$ from $\Pi_\Psi$.

(2) Compute the estimates $\hat{\psi}_0, \hat{\psi}_{1/N}, \hat{\psi}_{2/N}, \ldots, \hat{\psi}_1$, using interpolation between sample quantiles, where $\hat{\psi}_0 = 0$ and $\hat{\psi}_1$ is the largest sample value.

(3) Select $M_2$ and generate a sample $\psi_{1,x}, \ldots, \psi_{M_2,x}$ from $\Pi_\Psi(\cdot \mid T(x))$.

(4) For $\psi \in [\hat{\psi}_{i/N}, \hat{\psi}_{(i+1)/N})$ estimate $RB(\psi)$ by

$$\hat{RB}(\psi) = \frac{\hat{F}(\hat{\psi}_{(i+1)/N} \mid x) - \hat{F}(\hat{\psi}_{i/N} \mid x)}{\hat{F}(\hat{\psi}_{(i+1)/N}) - \hat{F}(\hat{\psi}_{i/N})} = N(\hat{F}(\hat{\psi}_{(i+1)/N} \mid x) - \hat{F}(\hat{\psi}_{i/N} \mid x)).$$

(5) Estimate $\Pi_\Psi(RB(\psi) \leq RB(\psi_{i_0/N}) \mid x)$ by the finite sum

$$\sum_{\{i : \hat{RB}(\hat{\psi}_{i/N}) \leq \hat{RB}(\psi_{i_0/N})\}} (\hat{F}(\hat{\psi}_{(i+1)/N} \mid x) - \hat{F}(\hat{\psi}_{i/N} \mid x)).$$

The following result establishes the convergence of Eq.(4.2) to the measure of strength of evidence in (3.4) as $N, M_1,$ and $M_2$ grow.
Theorem 11. Suppose that $RB(\psi)$ is continuous in $\psi$ and $RB(\psi)$ has a continuous posterior distribution. Then Eq.(4.2) converges almost surely to to $\Pi_\Psi(RB(\psi) \leq RB(0) \mid x)$ as $N \to \infty, M_1 \to \infty$ and $M_2 \to \infty$.

Proof. See Appendix C.

As in any problem we must be aware of whether or not we have detected a difference from a hypothesized value $\psi_0$ that is of practical significance. This entails specifying a $\delta > 0$ such that, if the true value of $\psi$ differs from $\psi_0$ by no more than $\delta$, then we consider $H_0$ as being true. The use of $\Psi = d_{H_0}$ allows us to deal with this issue in a very natural way.

4.2 Logistic Regression

In this section, we first discuss an issue about assigning priors on model parameters. We then develop a test for goodness of fit based on the method of concentration and show that this leads to a method for assigning a prior on model parameters that avoids problems with more typical choices. For simplicity we will present the case of a single predictor but note that adding more predictors is completely feasible.

4.2.1 Assigning Priors for $p_i$

Let $x_i \in R^1$ denote the value of a predictor $X, p_i = P(Y = 1 \mid X = x_i)$ and $\mu_i = \log (p_i/(1 - p_i))$. Suppose first that $X$ is a categorical predictor taking $k$ values and we observe $n_i$ observations when $X = x_i$. If the observations are independent, then the appropriate model is a $k$-fold product of binomial models. Depending on what we know a priori, we could select independent beta priors for the $p_i$, although other choices are certainly possible. For example, if we felt that we know virtually nothing about $p_i$, then we could take $p_i$ to be uniformly distributed. Since $\mu_i = \log (p_i/(1 - p_i))$, it follows that
the log-odds, \( \mu_i \), has a standard logistic distribution given by

\[
\pi(\mu_i) = \frac{\exp(\mu_i)}{(1 + \exp(\mu_i))^2} \quad \mu \in \mathbb{R}
\]  

(4.3)

In this case, it is relatively simple to assign a prior to the \( p_i \) or \( \mu_i \). In particular, assign one that is weakly informative with respect to the \( p_i \).

Suppose now that \( X \) is quantitative. In this situation it is common to relate the \( p_i \) to \( X \) via \( \mu = \beta_1 + \beta_2 x \), when \( X = x \), or even use a higher degree polynomial. In this case, we want to assess the hypothesis \( H_0 : \beta_2 = 0 \) as this corresponds to no effect due to the predictor. It is common to put \( N(0, \sigma_i^2) \) priors on the \( \beta_i \) where the \( \sigma_i^2 \) are chosen large to reflect little information \textit{a priori}. As is well-known, see, for example, Evans and Jang (2011a), there are problems with such a choice. We illustrate this in Figure 4.1 where we have plotted the prior density of \( p \) when \( \beta_1, \beta_2 \sim N(0, 1) \), and \( \beta_1, \beta_2 \sim N(0, \sigma^2) \), \( x = 1 \). As \( \sigma \) grows all the prior probability for \( p \) piles up at 0 and 1 and so this is clearly a poor choice. For a general \( (p_1, \ldots, p_k) \) it is not clear how we should choose a normal prior on \( (\beta_1, \beta_2) \) to reflect the information about the \( p_i \). Thus, choosing a diffuse prior on \( (\beta_1, \beta_2) \) to reflect little information in fact induces an informative prior on the \( p_i \).

![Figure 4.1: Prior density of of \( p \) when \( \beta_1, \beta_2 \sim N(0, 1) \) (on the left) and \( \beta_1, \beta_2 \sim N(0, 20^2) \) (on the right) and \( x = 1 \).](image-url)
This behavior of diffuse normal priors has been noted by other authors. Bedrick, Christensen and Johnson (1996, 1997), based on earlier work in Tsutukawa and Lin (1986), make the recommendation that priors should instead be placed on the \( p_i \), as these are the parameters for which we typically have prior information. Their recommendation is that, when \( \mu_i = \beta_1 + \beta_2 x_i \), two of the \( x_i \) values be selected and then Beta\((\alpha_i, \gamma_i)\) priors be placed on the corresponding \( p_i \). While this results in more sensible priors, it is not clear how to select the two \( x_i \) values, as the induced prior on the \((\beta_1, \beta_2)\) will depend on this choice.

4.2.2 Checking the logistic regression model

Suppose that we have no information about the relationship between \( Y \) and \( X \) other than that the \( y_i \) are conditionally independent given the \( x_i \). It is reasonable to assume that the \( p_i \) are i.i.d. \( U(0, 1) \). This implies that \textit{a posteriori} \( p_1 \ldots, p_k \) are independent with \( p_i \mid \bar{y} \sim \text{Beta}(1 + n_i \bar{y}_i, 1 + n_i(1 - \bar{y}_i)) \), where \( n_i \) is the number of observations collected for \( x_i \) and \( \bar{y}_i \) denotes the proportion of successes observed when \( X = x_i \). If we have prior information about some or all of the \( p_i \), then other beta distributions can be used for the priors.

We now assess the logistic regression hypothesis using the method of concentration via generating from the prior and posterior distributions of the \( p_i \) and transforming to \( \mu_i = \log p_i/(1 - p_i) \). For if \( \mu \in R^k \) is the vector of log-odds, then the logistic regression model holds if and only if \( \mu \in H_0 = L(V) \) where \( V = (1_k, x), 1_k = (1, \ldots, 1)^t \in R^k \) and \( x = (x_1, \ldots, x_k)^t \). Then we choose \( \Psi(\mu) = d_{H_0}(\mu) \) as in Example 14.

We follow the steps below for assessing \( H_0 \).

1. Generate a sample \( p_i \) from \( \pi(p_i) \) where \( p_i \sim U(0, 1) \) and transform to \( \mu_i = \log p_i/(1 - p_i) \) for \( i = 1, \ldots, k \).

2. Calculate the prior distance \( d_{\text{prior}} = ||\mu_{\text{prior}} - P\mu_{\text{prior}}||^2 \), where \( P = V(V^t V)^{-1}V^t \).
3. Generate a sample $p_i$ from $\pi(p_i \mid y)$, where $p_i \mid \bar{y}_i \sim \text{Beta}(1 + n_i\bar{y}_i, 1 + n_i(1 - \bar{y}_i))$. and transform to $\mu_{\text{prior}} = \log p_i/(1 - p_i)$ for $i = 1, ..., k$.

4. Calculate the posterior distance $d_{\text{post}} = ||\mu_{\text{post}} - P\mu_{\text{post}}||^2$.

5. Repeat Steps 1-4 for $10^4$ times. Plot the posterior and prior densities of $d_{\text{prior}}$ and $d_{\text{post}}$ to get a visual opinion how their distributions concentrate around 0. Using the algorithm explained in Section 4.1.1, estimate $RB(0)$ and the strength of evidence to evaluate the hypothesis.

We show the application of this method in the next two examples.

**Example 15. When the model is correct**

First we apply the goodness-of-fit test when the model is correct and we place uniform priors on the $p_i$. For this, we take $k = 3$, $x = (0, 1, 2)$ and $\beta_1 = 0.5, \beta_2 = -1$ so $p = (0.62, 0.38, 0.18)$. When we generate data from the model, we consider three different $n_i$, where $n_i$ is the number of observations for each $x = (0, 1, 2)$. In Table 4.1, we present the generated values of the $n\bar{y}_i$, and the values of $RB(\psi_*)$ and $\Pi_{\Psi} (RB(\psi) \leq RB(\psi_*) \mid x)$ for various left-tail prior quantiles $\psi_*$ of the prior.

We see that the evidence, when using $RB(\psi_*)$ to approximate $RB(0)$, is in each case in favor of the hypothesis that the logistic regression model holds and that this evidence is either strong or fairly strong. In Figure 4.2 we have plotted the posterior density of $\psi$ and $RB(\psi)$ when $n_i = 10$, with the 0.05, 0.10, 0.20 and 0.50 prior quantiles marked on the $x$-axis. We see that the posterior distribution of $\Psi$ has concentrated in the left tail of its prior distribution.
Table 4.1: Values of $RB(\psi_*)$ and the strength of evidence in Example 15 when $\psi_*$ equals to 0.01, 0.05 and 0.10 prior quantiles and the model is correct.

| Quantile | $RB(\psi_*)$ (\(\Pi_\psi(RB(\psi) \leq RB(\psi_*)|x))\) |
|----------|--------------------------------------------------|
|          | $n_i = 1$                                      | $n_i = 5$                                      | $n_i = 10$                                      |
| $(n_i \bar{y})$ | (1,0,0)                                      | (3,2,1)                                      | (7,4,1)                                      |
| 0.01     | 1.059(0.828)                                  | 2.073(1.000)                                | 2.714(1.000)                                |
| 0.05     | 1.048(0.702)                                  | 2.067(0.979)                                | 2.611(0.841)                                |
| 0.10     | 1.046(0.670)                                  | 2.011(0.856)                                | 2.567(0.763)                                |

Figure 4.2: Plots of the posterior density of $\psi$ and $RB(\psi)$ in Example 15 when the model is correct, $n_i = 10$, and with the 0.05, 0.10, 0.20 and 0.50 prior quantiles indicated

Example 16. When the model is wrong

We also consider an example where the simple logistic regression model is wrong. Let $k = 5$ with $x = (1, 3, 5, 7, 9)$. To achieve a non-linear relationship between $\mu = \log p/(1-p)$ and $x$, we generate $p_i \sim U(0, 1)$, e.g. $p = (0.875, 0.327, 0.107, 0.198, 0.908)$ so the relationship $\mu_i = \beta_1 + \beta_2 x_i$ does not hold for any choice of $(\beta_1, \beta_2)$. 
In Table 4.2, we present the results for generated data from the model with these probabilities and we see that the evidence, when using $RB(\psi^*)$ to approximate $RB(0)$, is in each case against the hypothesis that the logistic regression model holds and that this evidence is either strong or fairly strong. In Figure 4.3 we plot the prior and posterior densities of $\psi$ and $RB(\psi)$ and we see clear evidence that the model does not hold.

Table 4.2: Values of $RB(\psi^*)$ and the strength of evidence in Example 16 when $\psi^*$ equals to 0.01, 0.05 and 0.10 prior quantiles and the model is wrong.

| Quantile | $RB(\psi^*)$ ($\Pi_\psi(RB(\psi) \leq RB(\psi^*|x))$) |
|----------|-----------------------------------------------|
|          | $n_i = 1$                                      |
|          | $n_i = 5$                                      |
|          | $n_i = 10$                                     |
| $(n_i \bar{y})$ | (1,1,0,0,1)                                    |
|          | (5,3,0,2,1)                                    |
|          | (9,2,1,2,10)                                   |
| 0.01     | 0.761(0.066)                                   |
|          | 0.017(0.000)                                   |
|          | 0.000(0.000)                                   |
| 0.05     | 0.732(0.028)                                   |
|          | 0.046(0.002)                                   |
|          | 0.001(0.000)                                   |
| 0.10     | 0.793(0.082)                                   |
|          | 0.060(0.003)                                   |
|          | 0.003(0.000)                                   |

Figure 4.3: Plots of the prior (---) and posterior densities (−) of $\psi$ and $RB(\psi)$ in Example 16 when the model is wrong and $n_i = 10$. 
This test of fit has low power when many of \( n_i \) are small. In design contexts we can select the values \((x_i, n_i)\) to ensure sensitivity. In some situations, however, the data have many \( n_i = 1 \). In these circumstances we can typically group the \( x_i \) values and also fit higher order quadratic or cubic terms, and test for the higher order terms for the model checking. We illustrate this in Example 17 in the next section.

### 4.2.3 Assigning priors on \( \beta \)s

Suppose that we obtain no evidence against the logistic regression model holding. Then appropriate inferences for the \( \beta_i \) are obtained using the conditional prior and posterior of \( \mu \) given that \( \mu \in H_0 = \mathcal{L}(X) \). Suppose the prior is given by a product of uniforms on the \( p_i \). Then, since \( d_{H_0} \) has constant volume distortion on \( H_0 \), the conditional prior density of \( \mu \) given that \( d_{H_0}(\mu) = 0 \), is proportional to the prior density of \( \mu \) which is given by

\[
\prod_{i=1}^{k} \exp\{\mu_i\}/(1 + \exp\{\mu_i\})^2 \text{ by Eq. (4.3)}.
\]

We then coordinatize \( H_0 \) by \((\beta_1, \beta_2)\), and obtain the conditional prior of \((\beta_1, \beta_2)\) which is proportional to

\[
\pi(\beta_1, \beta_1) \propto g(\beta_1, \beta_2) = \prod_{i=1}^{k} \frac{\exp\{\beta_1 + \beta_2 x_i\}}{(1 + \exp\{\beta_1 + \beta_2 x_i\})^2}
\]

A similar result holds for the conditional posterior. It is difficult to see exactly what these priors are saying about the \( \beta_i \). Camilli (1994) discussed a \( N(0, d^2) \) approximation to the standard logistic distribution. Using this approximation, we can get conditional priors that are much easier to work with. The optimal choice of \( d \), in the sense that it minimizes \( \max_{x \in \mathbb{R}} |\Phi(x/d) - e^x/(1 + e^x)| \) is given by \( d = 1.702 \) and this leads to a maximum difference less than 0.01. Thus, when \( \mu \) is distributed \( N(0, 1.702^2) \) we have that \( e^\mu/(1 + e^\mu) \) is approximately distributed \( U(0, 1) \) with the same maximum error. In Figure 4.4 we have plotted the density of \( p = e^\mu/(1 + e^\mu) \) when \( \mu \) is \( N(0, d^2) \) for various choices of \( d \) and we see that it is indeed approximately uniform when \( d = 1.702 \).
The prior $\prod_{i=1}^{k} \exp\{\mu_i\}/(1 + \exp\{\mu_i\})^2$ induced on the logits by the uniform prior on $(p_1, \ldots, p_k)$ is then approximated by the $N_k(0, d^2 I)$ prior with $d = 1.702$. Now suppose that we condition on $\mu = X\beta$ where $X \in \mathbb{R}^{k \times 2}$ and $X'X = I$, namely, $X$ is column orthonormal. The conditional prior on $\beta$ is proportional to $\exp\{-\beta'X'X\beta/2d^2\} = \exp\{-\beta'\beta/2d^2\}$, namely, the conditional prior on $\beta$ is $N_2(0, d^2 I)$. Note that this says that the $\beta_i$ are independent and approximately distributed standard logistic. Then, applying the standard logistic cdf $F$ component wise, we have that $p = F(u)$ where $u = X\beta \sim N_k(0, d^2 XX')$. Note that the $N_k(0, d^2 XX')$ distribution is singular but we can generate a value from it easily via $u = dXz$ where $z \sim N_2(0, I)$. From this we can simulate to obtain the conditional joint prior on $p$.

The prior $\prod_{i=1}^{k} \exp\{\mu_i\}/(1 + \exp\{\mu_i\})^2$ on the logits makes sense when we know absolutely nothing about the $p_i$. In situations where there is some information about $p_i$, e.g. we know that $p_i$ is very small, we try to place a prior on $p_i$ that reflects this. One could then choose location and scaling parameters for the logistic distribution to reflect what our knowledge about $p_i$ says about $\mu_i$. Again we can approximate such a
distribution by a normal distribution as it is easier to work with.

Another issue is dealing with the fact that when two \( x_i \) values are close, then the corresponding \( p_i \) values are close too. We would want our prior to reflect this. This is accomplished by imposing a correlation between \( \mu_i \) and \( \mu_j \) that depends upon \(|x_i - x_j|\). In the end we choose a \( N_k(\alpha, \Sigma) \) prior on \( \mu \). Then the conditional prior on \( \beta \) is proportional to \( \exp\left\{-(X\beta - \alpha)'\Sigma^{-1}(X\beta - \alpha)/2d^2\right\} \) which is in turn proportional to \( \exp\left\{-(\beta - (X'\Sigma^{-1}X)^{-1}V\alpha)'X'\Sigma^{-1}X(\beta - (X'\Sigma^{-1}X)^{-1}X'\alpha)/2d^2\right\} \). Therefore, the conditional prior distribution of \( \beta \) is

\[
N_2((X'\Sigma^{-1}X)^{-1}X'\alpha, d^2(X'\Sigma^{-1}X)^{-1}) \tag{4.4}
\]

and \( p = F(u) \) with

\[
u = X\beta \sim N_k(X(X'\Sigma^{-1}X)^{-1}X'\alpha, d^2X(X'\Sigma^{-1}X)^{-1}X'). \tag{4.5}\]

So we choose \((\alpha_i, \sigma_{ii})\) to reflect what we know about \( p_i \) through \( p_i = F((z - \alpha_i)/d\sqrt{\sigma_{ii}}) \) where \( z \sim N(0,1) \). For example, if we think that \( p_i \) is in the interval \((0.1, 0.2)\) with prior probability 0.9 then \( 0.9 = \Phi(\alpha_i + d\sqrt{\sigma_{ii}}F^{-1}(0.2)) - \Phi(\alpha_i + d\sqrt{\sigma_{ii}}F^{-1}(0.1)) \) and placing another prior probability restriction on \( p_i \) will allow us to solve for \( \alpha_i \) and \( \sigma_{ii} \).

Choosing \((\alpha_i, \sigma_{ii}) = (0, d^2)\) corresponds to being noninformative about \( p_i \). The issue of correlations is more difficult but a plausible approach here is to take \( \text{Corr}(\mu_i, \mu_j) = \exp\left\{-(x_i - x_j)^2/2l^2\right\} \) where \( l \geq 0 \) is a hyperparameter chosen to reflect how quickly we believe \( p_i \) and \( p_j \) will become alike as \(|x_i - x_j| \to 0 \). Note that if we take \( l = 0 \), then the \( p_i \) are \textit{a priori} uncorrelated and when \( l = \infty \), then the \( p_i \) are completely dependent.

The choice of \( l \) reflects what we know about the dependence of the probabilities on the predictor. An approach to choosing \( l \) is to select \((\mu_i, \mu_j)\) and then choose \( l \) so that \( \text{Corr}(\mu_i, \mu_j) \) equals some specific value. Alternatively, we could consider placing a prior on \( l \). Note that, when \( x_i \to x_j \), then the \( N_k(\alpha, \Sigma) \) prior converges to the \( N_{k-1}(\alpha_{-i}, \Sigma_{-i}) \) distribution where \((\alpha_{-i}, \Sigma_{-i})\) is obtained from \((\alpha, \Sigma)\) by deleting all entries associated with the \( i \)-th coordinate. So there is a coherency among priors between the situations
where we consider \( \mu_i \) and \( \mu_j \) as possibly very different, because \( |x_i - x_j| \) is large, and when we think of them as virtually identical because \( |x_i - x_j| \) is quite small. Of course ‘large’ and ‘small’ are application dependent as this depends on the meaning of the predictor.

We present a real data application of this to a quadratic logistic regression model in Example 17. When performing the goodness of fit approach, we use Metropolis-within-Gibbs algorithm with systematic scan to generate \( \mu \) from its posterior. The prior and posterior densities of \( \beta \) are calculated using a multivariate Student importance sampling method.

**Example 17. Gender-Height data**

We consider predicting gender from height (\( Ht \)) measurements in cm from a data set on 102 male and 100 female athletes collected by the Australian Institute of Sport. There are 147 distinct values of \( Ht \) and we grouped these to form 21 cells each of length 3 cm ranging from 147 cm to 210 cm. The value of the predictor variable was taken to be the midpoint of each interval. This was done to reduce the size of the correlation matrix. The raw data can be found in Sheather (2008).

We considered the model \( \logit(p) = \beta_1 + \beta_2 Ht + \beta_3 (Ht)^2 \). For the prior on the 21 distinct \( \mu_i \) we chose a \( N_{21}(0, d^2 R) \) distribution with \( l = 2 \) so that the prior correlation between the \( \mu_i \) in adjacent cells equals 0.32. In Figure 4.5 we have plotted the prior density, posterior density and the relative belief ratio of the squared distance for the goodness of fit test. We can see from this that there is evidence in favour of the logistic model. In fact the goodness-of-fit test for the quadratic model gave the values for \( (\Pi_\psi (RB(\psi) \leq RB(\psi_*) | x), RB(\psi_*)) \) as (1.000, 8.708), (0.773, 3.473) and (0.6015, 2.220) when \( \psi_* \) is the 0.01, 0.05 and 0.10 prior quantile, respectively. In Figure 4.6 we have plotted the marginal priors and posteriors of the \( \beta_i \).

For the hypothesis \( H_0 : \beta_3 = 0 \) we obtained \( RB(0) = 1.42 \) with the strength of evidence equal to 0.40 and \( (\hat{\beta}_3, RB(\hat{\beta}_3)) = (1.12, 1.81) \) so this is evidence in favour of \( \beta_3 = 0 \). For the hypothesis \( H_0 : \beta_2 = 0 \) we obtained \( RB(0) = 3.2 \times 10^{-13} \) which we know
immediately provides strong evidence against $H_0$ as the strength of evidence is bounded above by $RB(\psi_0)$ by Theorem 7 in Chapter 3.

Figure 4.5: Plots of the prior (---) density, posterior (–) density and relative belief ratio of the squared distance in Example 17.

Figure 4.6: Plots of the marginal prior (---) and posterior (–) densities of the $\beta_i$ coefficients in Example 17.
4.3 Conclusion and current work

We illustrated the application of the relative belief inferences to the analysis of logistic regression model when we have univariate binary data. We developed a goodness of fit test and showed how this method led to sensible priors on model parameters when no lack of fit is detected. This approach is seen to resolve several issues associated with the analysis of these models.

We have expanded the models presented here to the case when the binary responses are correlated. Our model for correlated binary data allows the values of predictor variables to be different for each element of a response vector. The contents of this chapter and our approach to analyzing correlated binary data are currently being written up in a manuscript.
Appendix A

Regularity condition

The regularity conditions (A1-A6) are provided in (Knight, 2000, p.245). Let \( l(\theta; y) = \log f(y; \theta) \) and let \( l_{\theta}(\theta; y), l_{\theta\theta}(\theta; y) \) and \( l_{\theta\theta\theta}(\theta; y) \) be the first three partial derivatives of \( l(\theta; y) \) with respect to \( \theta \). These conditions apply on the component log densities of a composite likelihood.

A1. The parameter space \( \Theta \) is an open subset of the real line.

A2. The set \( A = \{ y : f(x; \theta) > 0 \} \) does not depend on \( \theta \).

A3. \( f(y; \theta) \) is three times continuously differentiable with respect to \( \theta \) for all \( y \) in \( A \).

A4. \( E[l_{\theta}(\theta; y)] = 0 \) for all \( \theta \) and \( \text{Var}[l_{\theta}(\theta; y)] = I(\theta) \) where \( 0 < I(\theta) < \infty \) for all \( \theta \).

A5. \( E[l_{\theta\theta}(\theta; y)] = -J(\theta) \) where \( 0 < J(\theta) < \infty \) for all \( \theta \).

A6. For each \( \theta \) and \( \delta > 0 \), \( |l_{\theta\theta\theta}(t; y)| \leq M(x)| \) for \( |\theta - t| \leq \delta \) where \( E_{\theta}[M(Y_i)] < \infty \)

A7. There exists a unique point \( \theta_g \in \Theta \) which minimizes the composite Kullback-Lebler divergence 2.2 (Xu, 2012).

Condition A4 changes when there is model misspecification (e.g. setting up wrong marginal or conditional densities in composite likelihoods), e.g. \( E_{\theta}[l_{\theta}(\theta; y)] = 0 \) only for \( \theta = \theta_g \), where the expectation is taken under the correct (unknown) model \( g \).
It can be deduced that the mean and variance of the log likelihood derivatives, \(l_\theta(\theta; y), l_{\theta\theta}(\theta; y)\) and \(l_{\theta\theta\theta}(\theta; y)\) are of order \(O(n)\). The higher order derivatives are, in general, of order \(O_p(n)\) (Severini, 2000, p.88).

According to Severini (2000, p.106), sufficient conditions for the consistency of the MLE for regular models are:

1. \(\Theta\) is a compact subset of \(\mathcal{R}^d\).

2. \(\sup_{\theta \in \Theta} |n^{-1}l(\theta) - n^{-1}E\{l(\theta)\}| \to^p 0\) as \(n \to \infty\).

We need the second condition to hold on the component log densities. Let \(l_i\) be the \(i^{th}\) component log density in the composite likelihood with \(i = 1, ..., d\), such that \(cl(\theta) = \sum_{i=1}^{d} n^{-1}l_i(\theta)\) then

\[
\sup_{\theta \in \Theta} \left| \sum_{i=1}^{d} n^{-1}l_i(\theta) - \sum_{i=1}^{d} n^{-1}E\{l_i(\theta)\} \right| \leq \sum_{i=1}^{d} \left| n^{-1}l_i(\theta) - n^{-1}E\{l_i(\theta)\} \right|
\]

\[
\sum_{i=1}^{d} \sup_{\theta \in \Theta} \left| n^{-1}l_i(\theta) - n^{-1}E\{l_i(\theta)\} \right| \leq \sum_{i=1}^{d} \sup_{\theta \in \Theta} \left| n^{-1}l_i(\theta) - n^{-1}E\{l_i(\theta)\} \right| = \sum_{i=1}^{d} \sup_{\theta \in \Theta} \left| n^{-1}l_i(\theta) - n^{-1}E\{l_i(\theta)\} \right| \quad (A.1)
\]

If each component in Eq.(A.1) goes to 0 in probability, then the term on the left side goes to 0 in probability.
Appendix B

Proofs for Chapter 2

Proof of Theorem 1b): Note that $CL(\theta; y) = \prod_{k=1}^{K} f(y \in A_k; \theta)^{\omega_k}$ and $cl(\theta; y) = \log CL(\theta; y) = \sum_{k=1}^{K} \omega_k \log f(y \in A_k; \theta)$.

Composite score function: $u(\theta; y) = \nabla_{\theta} cl(\theta; y) = \sum_{k=1}^{K} \omega_j \nabla \log f(y \in A_k; \theta)$

The sensitivity matrix under the correct model:

$$H(\theta) = E_g (\nabla_{\theta} u(\theta; Y)) = \int -\nabla_{\theta} u(\theta; y) g(y) dy$$

The variability matrix under the correct model:

$$J(\theta) = \text{var}_g (u(\theta; Y))$$

The Godambe information matrix (Godambe, 1960) under the correct model is

$$G(\theta) = H(\theta) J(\theta)^{-1} H(\theta).$$

Then for $n$ independent and identically distributed observations $Y_1, ..., Y_n$ from the model $g(\cdot)$, as $n \to \infty$, under the regularity conditions, we have:

$$\frac{\sum_{i=1}^{n} u(\theta_g; y_i)}{\sqrt{n}} \xrightarrow{d} N(0, J(\theta_g)) \quad \text{since} \quad E_g(u(\theta_g; Y)) = 0 \quad (B.1)$$

$$\frac{\sum_{i=1}^{n} \nabla_{\theta} u(\theta_g; y_i)}{n} \to -H(\theta_g) \quad (B.2)$$
where \( \theta_g \) is the (unique) minimizer of the composite Kullback-Leibler divergence between \( f \) and \( g \). Let \( \theta = \theta_g + c/\sqrt{n} \), then the Taylor expansion of the log composite likelihood around \( \theta_g \) is:

\[
cl(\theta) - cl(\theta_g) = u(\theta_g)(\theta - \theta_g) + \nabla_\theta u(\theta_g)\frac{(\theta - \theta_g)^2}{2} + O_p(n^{-1/2})
\]

\[
= \sum_{i=1}^{n} u(\theta_g; y_i) \frac{c}{\sqrt{n}} + \sum_{i=1}^{n} \nabla u(\theta_g; y_i) \frac{c^2}{2n} + O_p(n^{-1/2})
\]

\[
\to d \mathcal{N} \left( \begin{array}{c} 
\frac{c^2}{2} \text{E}(\nabla u(\theta_g; Y)), \\
\frac{c^2}{2} \text{var}(u(\theta_g; Y)) 
\end{array} \right) \quad \text{(from Eq. (B.1) and (B.2))}
\]

\[
\to N \left( -\frac{c^2}{2}a, c^2b \right) \quad \text{(B.3)}
\]

\( O_p(n^{-1/2}) \) can be justified since the log composite likelihood is a finite sum of genuine log likelihoods, which are of the same order. Note that Eq.(B.3) does not generate the bump function since the mean is not the negative half of the variance in the asymptotic normal distribution. In order to obtain the bump function, we can adjust the ratio of composite likelihoods by raising it to the power \((a/b)\):

\[
\frac{a}{b} \log \frac{CL(\theta; y)}{CL(\theta_g; y)} \to N \left( -\frac{a^2c^2}{2b}, -\frac{a^2c^2}{b} \right) \quad \text{(B.4)}
\]

\[
\lim_{n \to \infty} P \left( \left( \frac{CL(\theta; y)}{CL(\theta_g; y)} \right)^{a/b} \geq k \right) = \Phi \left( -\frac{c^*}{2} - \frac{\log k}{c^*} \right)
\]

where \( c^* = \frac{ac}{\sqrt{b}} \)

We can estimate \( a/b \) through consistent estimates of \( J(\theta) \) and \( H(\theta) \). Let \( \hat{\theta}_{CL} \) be the maximum likelihood estimator of \( \theta \), which is a consistent estimator of \( \theta_g \) (Xu, 2012) then

\[
\hat{a} = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta u(\hat{\theta}_{CL}; y_i)
\]
\[ \hat{b} = \frac{1}{n} \sum_{i=1}^{n} u^{2}(\hat{\theta}_{CL}; y_{i}) \]

**Proof of Theorem 2:**

(a)  
\[
\frac{CL_{p}(\psi_{g}, y)}{CL_{p}(\psi, y)} \xrightarrow{p} \infty \quad n \to \infty \quad (B.5)
\]

\[ CL_{p}(\psi_{g}; y) = CL(\psi_{g}, \hat{\lambda}(\psi_{g}); y) = \sup_{\lambda} CL(\psi_{g}, \lambda; y) \geq CL(\psi_{g}, \lambda; y) \quad \forall \lambda \] is thus true for \( \lambda_{g} \).

Then it would be enough to show \[ \frac{CL(\psi_{g}, \lambda_{g}; y)}{CL_{p}(\psi, y)} \xrightarrow{p} \infty \] since \[ CL(\psi_{g}, \lambda_{g}; y) \geq CL(\psi_{g}, \lambda_{g}; y) \]. This will imply that Eq. (B.5) holds.

In Severini (2000) on page 127, it was shown that the difference between a profile log-likelihood function from a genuine log likelihood function is of order \( O_{p}(1) \), i.e. \( l_{p}(\psi; y) = l(\psi, \lambda(\psi); y) + O_{p}(1) \), here \( l(\psi, \lambda; y) \) refers to a genuine log likelihood function as it can be obtained from an actual model for the data using a Taylor expansion \( l_{p}(\psi; y) = l(\psi, \hat{\lambda}(\psi); y) \) about \( l(\psi, \lambda; y) \). Following a similar Taylor expansion for the composite likelihood, we get;

\[
cl(\psi, \hat{\lambda}(\psi)) = cl(\psi, \lambda(\psi)) + cl_{\lambda}(\psi, \lambda(\psi))^{T}(\hat{\lambda}(\psi) - \lambda(\psi))
\]

\[
+ \frac{1}{2} (\hat{\lambda}(\psi) - \lambda(\psi))^{T} cl_{\lambda\lambda}(\psi, \lambda(\psi))(\hat{\lambda}(\psi) - \lambda(\psi)) + \ldots
\]

Then if \( \hat{\lambda}(\psi) = \lambda(\psi) + O_{p}(n^{1/2}) \) is true then we conclude that \( cl_{p}(\psi; y) = cl(\psi, \lambda(\psi)) + O_{p}(1) \). Since \( cl(\psi, \lambda(\psi)) \) is a finite sum of genuine log-likelihood functions, under regularity condition on genuine likelihood functions, \( cl_{\lambda}(\psi, \lambda(\psi)) = O_{p}(\sqrt{n}) \) and \( cl_{\lambda\lambda}(\psi, \lambda(\psi)) = O_{p}(n) \).

Why is \( \hat{\lambda}(\psi) = \lambda(\psi) + O_{p}(n^{1/2}) \) true?

Remember \( \theta_{g} = (\psi_{g}, \lambda_{g}) \) is the value of the parameter that minimizes the K-L divergence between the assumed model \( f \) and the true model \( g \). In the profile composite likelihood \( CL_{p}(\psi) = CL(\psi, \hat{\lambda}(\psi)) = \sup_{\lambda} CL(\psi, \lambda), \hat{\lambda}_{\psi} \) is the maximum likelihood estimate of \( \lambda \) for a fixed \( \psi \). In general \( \hat{\lambda}(\psi) \) is not a consistent estimator of \( \lambda_{g} \) unless \( \psi \) is fixed at the ‘true’ value, \( \psi_{g} \). Note that:
\[
\frac{1}{n} \text{cl}(\psi, \lambda) - \frac{1}{n} E_g[\text{cl}(\psi, \lambda)] \xrightarrow{\mathbb{P}} 0.
\]

Following the arguments in Severini (2000), section 4.2.1, the maximizer of \(\text{cl}(\psi, \lambda)/n\) should converge in probability to the maximizer of \(E_g[\text{cl}(\psi, \lambda)]/n\), which is \((\psi_g, \lambda_g)\). It was shown in Xu (2012) that the maximum composite likelihood estimator, \(\hat{\theta}_{CL}\), converges almost surely to \(\theta_g\) where \(\theta_g\) is the parameter that minimizes the Kullback-Leibler divergence between the working model \(f\) and the true model \(g\) (Eq. (2.2)). Here, we treat \(\psi\) fixed, then \(\lambda(\psi)\) becomes the only parameter and \(\hat{\lambda}(\psi)\) is the MLE of \(\lambda(\psi)\) for a fixed \(\psi\). Note that when \(\psi = \psi_g\), \(\lambda(\psi_g) = \lambda_g\). By following the regular arguments about the composite MLEs in Xu (2012), it can be shown that \(\hat{\lambda}(\psi) \to \lambda(\psi)\) as \(n \to 0\), where \(\lambda(\psi)\) is the value of \(\lambda\) that maximizes \(n^{-1}E_g[l(\psi, \lambda)]\) when \(\psi\) is fixed. The asymptotic distribution of \(\sqrt{n} \left(\hat{\lambda}(\psi) - \lambda(\psi)\right)\) is derived in Eq.(B.12) and Eq.(B.15). Thus \(\hat{\lambda}(\psi) = \lambda(\psi) + O_p(n^{-1/2})\) and \(cl_p(\psi) = cl(\psi, \lambda(\psi)) + O_p(1)\). □

It follows that:

\[
(1/n) \log \frac{CL(\psi_g, \lambda_g)}{CL_p(\psi)} = (1/n)(cl(\psi_g, \lambda_g) - cl(\psi, \lambda(\psi))) + (1/n)O_p(1) \tag{B.6}
\]

where \(cl(\theta) = \log CL(\theta)\).

\[
\log \left(\frac{\prod_{i=1}^{n} CL(\psi_g, \lambda_g; y_i)}{\prod_{i=1}^{n} CL(\psi, \lambda(\psi); y_i)}\right)^{1/n} = \frac{1}{n} \left(\sum_{i=1}^{n} cl(\psi_g, \lambda_g; y_i) - \sum_{i=1}^{n} cl(\psi, \lambda(\psi); y_i)\right) + (1/n)O_p(1)
\]

\[
= \frac{1}{n} \left(\sum_{k=1}^{K} w_k \left\{\sum_{i=1}^{n} \log(f(y_i \in A_k; \psi_g, \lambda_g)) - \sum_{i=1}^{n} \log(f(y_i \in A_k; \psi, \lambda(\psi)))\right\}\right) + (1/n)O_p(1)
\]

\[
\xrightarrow{\mathbb{P}} \sum_{k=1}^{K} w_k \left\{E_g \left[\log(f(Y \in A_k; \psi_g, \lambda_g)) - \log(f(Y \in A_k; \psi, \lambda(\psi)))\right]\right\} > 0
\]

since \(w_k\)'s are positive and \(\theta_g = (\psi_g, \lambda_g)\) minimizes the Kullback-Leibler divergence in
Let $R_n = \prod_{i=1}^{n} CL(g \psi, \lambda; y_i) / \prod_{i=1}^{n} CL(g \psi, \lambda; y_i)$. We get $1/n \log R_n \to c > 0$, where $c$ is a finite positive number, then $R_n = \prod_{i=1}^{n} CL(g \psi, \lambda; y_i) / \prod_{i=1}^{n} CL(g \psi, \lambda; y_i) \to \infty$. \hfill \Box

(b) $\lim_{n \to \infty} P_g \left( \frac{CL_p(g \psi)}{CL_p(g \psi)} \geq k \right) = \Phi \left( -\frac{c}{2} - \frac{\log k}{c} \right)$ where $c$ is proportional to the distance between $\psi$ and $\psi_g$.

Following the proof of Royall (2000) for profile likelihoods: Let $cl_p(g \psi) = \log CL_p(g \psi)$.

For $\psi = \psi_g + c/\sqrt{n}$,

$$cl_p(g \psi) - cl_p(g \psi_g) = cA_n + (c^2/2)B_n + R_n$$

where $A_n = \frac{1}{\sqrt{n}} \left. \frac{dcl_p(g \psi)}{d\psi} \right|_{(\psi_g)}$, $B_n = \frac{1}{n} \left. \frac{d^2cl_p(g \psi)}{d\psi^2} \right|_{(\psi_g)}$ and $R_n = O_p(n^{-1/2})$.

We make a Taylor expansion of $\left. \frac{dcl(g \psi, \hat{\lambda}(g \psi))}{d\psi} \right|_{(\psi_g)}$ about $\left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda(g \psi))}$.

$$\left. \frac{dcl(g \psi, \hat{\lambda}(g \psi))}{d\psi} \right|_{(\psi_g)} = \left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda(g \psi))} + \left. \frac{dcl(g \psi, \lambda)}{d\lambda} \right|_{(\psi_g, \lambda(g \psi))} \left. \frac{d\hat{\lambda}(g \psi)}{d\psi} \right|_{(\psi_g)}$$

$$= \left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda(g \psi))}$$

We make a Taylor expansion of $\left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda(g \psi))}$ about $\left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda_g)}$.

$$\left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda(g \psi))} = \left. \frac{dcl(g \psi, \lambda)}{d\psi} \right|_{(\psi_g, \lambda_g)} + \left. \frac{d^2cl(g \psi, \lambda)}{d\psi d\lambda} \right|_{(\psi_g, \lambda_g)} (\hat{\lambda}(g \psi) - \lambda_g) + R^*_n$$

We observe that $\lambda(g \psi) = \lambda_g$ and that $cl(g \psi, \lambda(g \psi))$ is a finite sum of genuine log-likelihood functions. Then $R^*_n$ in Eq.(B.10) is of $O_p(1)$ since the higher order derivatives of log-likelihood function are of order $O_p(n)$ and $(\hat{\lambda}(g \psi) - \lambda_g) = O_p(n^{-1/2})$ due to Eq.(B.15).
Then (B.8) becomes:

\[
\frac{1}{\sqrt{n}} \left. \frac{dcl_p(\psi)}{d\psi} \right|_{(\psi_g, \lambda_g)} = \frac{1}{\sqrt{n}} \left. \frac{\partial cl(\psi, \lambda)}{\partial \psi} \right|_{(\psi_g, \lambda_g)} + \frac{1}{n} \left. \frac{\partial^2 cl(\psi, \lambda)}{\partial \psi \partial \lambda} \right|_{(\psi_g, \lambda_g)} \sqrt{n}(\hat{\lambda}(\psi_g) - \lambda_g) \quad (\text{B.11})
\]

(I) and (II) in Eq.(B.11) become:

\[
\frac{1}{\sqrt{n}} \left. \frac{\partial cl(\psi, \lambda)}{\partial \psi} \right|_{(\psi_g, \lambda_g)} \rightarrow N(0, J_{\psi \psi}(\psi_g, \lambda_g)) \quad \text{(by CLT)}
\]

\[
\frac{1}{n} \left. \frac{\partial^2 cl(\psi, \lambda)}{\partial \psi \partial \lambda} \right|_{(\psi_g, \lambda_g)} \rightarrow E_g \left( \frac{\partial^2 cl(\psi, \lambda)}{\partial \psi \partial \lambda} \right|_{(\psi_g, \lambda_g)}) \quad \text{(by LLN)}
\]

\[
= -H_{\psi \lambda}(\psi_g, \lambda_g)
\]

What about (III) in Eq.(B.11)?

Note that \( \hat{\lambda}(\psi_g) \) is the solution to

\[
\left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\hat{\lambda}(\psi_g)} = 0.
\]

Then a Taylor expansion of

\[
\left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\hat{\lambda}(\psi_g)} \text{ about } \left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\lambda_g}
\]

gives,

\[
\left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\hat{\lambda}(\psi_g)} = 0 = \left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\lambda_g} + \left. \frac{d^2 cl(\psi_g, \lambda)}{d\lambda^2} \right|_{\lambda_g} (\hat{\lambda}(\psi_g) - \lambda_g) + R_n^{**}.
\]

(B.12)

Dividing both sides by \( \sqrt{n} \), we get

\[
\frac{1}{\sqrt{n}} \left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\hat{\lambda}(\psi_g)} = \frac{1}{\sqrt{n}} \left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\lambda_g} + \frac{1}{n} \left. \frac{d^2 cl(\psi_g, \lambda)}{d\lambda^2} \right|_{\lambda_g} \sqrt{n}(\hat{\lambda}(\psi_g) - \lambda_g) + \frac{R_n^{**}}{\sqrt{n}}.
\]

(B.13)

In (B.13), \( \frac{R_n^{**}}{\sqrt{n}} = \frac{1}{n} \left. \frac{d^3 cl(\psi_g, \lambda)}{d\lambda^3} \right| \sqrt{n}(\hat{\lambda}(\psi_g) - \lambda_g)^2 \) where \( |\lambda_g - t| \leq \delta \). Then by A6 and \( (\hat{\lambda}(\psi_g) - \lambda_g) \rightarrow p 0 \), the following argument holds (Knight, 2000, chap.5),

\[
(\hat{\lambda}(\psi_g) - \lambda_g) \frac{1}{n} \left. \frac{d^3 cl(\psi_g, \lambda)}{d\lambda^3} \right|_t \rightarrow p 0.
\]

(B.14)

Then (III) in Eq.(B.11) becomes

\[
\sqrt{n}(\hat{\lambda}(\psi_g) - \lambda_g) \doteq \frac{1}{\sqrt{n}} \left. \frac{dcl(\psi_g, \lambda)}{d\lambda} \right|_{\lambda_g} - \frac{1}{n} \left. \frac{d^2 cl(\psi_g, \lambda)}{d\lambda^2} \right|_{\lambda_g}
\]

\[
\doteq \frac{1}{H_{\lambda \lambda}(\psi_g, \lambda_g)} N(0, J_{\lambda \lambda}(\psi_g, \lambda_g))
\]

(B.15)
Substituting (B.15) in (B.11)

\[ A_n = \frac{1}{\sqrt{n}} \frac{dcl_p(\psi)}{d\psi} \Bigg|_{(\psi_g)} \xrightarrow{d} z_1 - H_{\psi\lambda}(\psi_g, \lambda_g) \frac{z_2}{H_{\lambda\lambda}(\psi_g, \lambda_g)} \]

where

\[
\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \xrightarrow{d} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} J_{\psi\psi}(\psi_g, \lambda_g) & J_{\psi\lambda}(\psi_g, \lambda_g) \\ J_{\lambda\psi}(\psi_g, \lambda_g) & J_{\lambda\lambda}(\psi_g, \lambda_g) \end{bmatrix} \right) \]

Then

\[ A_n \xrightarrow{d} N \left( 0, [J_{\psi\psi}(\psi_g, \lambda_g) + \frac{H_{\psi\lambda}(\psi_g, \lambda_g)}{H_{\lambda\lambda}(\psi_g, \lambda_g)}]^2 J_{\lambda\lambda}(\psi_g, \lambda_g) - 2 \frac{H_{\psi\lambda}(\psi_g, \lambda_g)}{H_{\lambda\lambda}(\psi_g, \lambda_g)} J_{\psi\lambda}(\psi_g, \lambda_g) \right) \]

(B.16)

What about \( B_n \) in (B.7)?

\[ B_n = \frac{1}{n} \frac{d^2cl_p(\psi)}{d\psi^2} \Bigg|_{(\psi_g)} \]

(B.17)
where

\[
\frac{d^2 c_{p}(\psi)}{d\psi^2} \bigg|_{(\psi_g)} = \frac{d}{d\psi} \left[ \frac{d c_{p}(\psi)}{d\psi} \right] \bigg|_{(\psi_g)} \\
= \frac{\partial}{\partial \psi} \left[ \frac{\partial c_{p}(\psi, \lambda)}{\partial \psi} + \frac{\partial c_{p}(\psi, \lambda)}{\partial \lambda} \frac{d\lambda}{d\psi} \right] \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \\
+ \left\{ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi \partial \lambda} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} + \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \lambda^2} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \right\} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{(\psi_g)} \\
+ \frac{d^2 \hat{\lambda}(\psi)}{d\psi^2} \bigg|_{(\psi_g)} \frac{\partial c_{p}(\psi, \lambda)}{\partial \lambda} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} = 0 \\
= \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi^2} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} + 2 \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi \partial \lambda} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{(\psi_g)} \\
+ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \lambda^2} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \left( \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{(\psi_g)} \right)^2
\]

Why is \( R_n \) of \( O_p(n^{-1/2}) \) in Eq.(B.7)? \( R_n = \frac{1}{\delta^{3/2}} \frac{d^3 c_{p}(\psi)}{d\psi^3} \bigg|_{t} \) where \( |\psi_g - t| \leq \delta \).

\[
\frac{d^3 c_{p}(\psi)}{d\psi^3} \bigg|_{t} = \frac{d}{d\psi} \left[ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi^2} \bigg|_{(t, \hat{\lambda}(t))} + 2 \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi \partial \lambda} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{t} \right] \\
+ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \lambda^2} \bigg|_{(t, \hat{\lambda}(t))} \left( \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{t} \right)^2
\]

\[
= \frac{d}{d\psi} \left[ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi^2} \bigg|_{(t, \hat{\lambda}(t))} \right] + \frac{d}{d\psi} \left[ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \psi \partial \lambda} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{t} \right] \\
+ \frac{d}{d\psi} \left[ \frac{\partial^2 c_{p}(\psi, \lambda)}{\partial \lambda^2} \bigg|_{(t, \hat{\lambda}(t))} \left( \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{t} \right)^2 \right]
\]
Appendix B. Proofs for Chapter 2

$A_1 = \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^3} \bigg|_{(t, \hat{\lambda}(t))} + \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^2 \partial \lambda} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t$

$A_2 = 2 \left\{ \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^2 \partial \lambda} \bigg|_{(t, \hat{\lambda}(t))} + \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^3} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t \right\} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t$

$A_3 = \left[ \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi \partial \lambda^2} \bigg|_{(t, \hat{\lambda}(t))} + \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^3} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t \frac{d^2\hat{\lambda}(\psi)}{d\psi^2} \bigg|_t \right]^2$

$$+ 2 \frac{\partial^2 cl(\psi, \lambda)}{\partial \psi \partial \lambda} \bigg|_{(t, \hat{\lambda}(t))} \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t \left[ \frac{d^2\hat{\lambda}(\psi)}{d\psi^2} \bigg|_t \right]$$

From Lemma 3 below, $\frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t = O_p(1)$. By taking the second derivative of $\frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_t = O_p(1)$. Also, following the arguments presented in Lemma 2, we observe that the second and higher order derivatives of composite log-likelihood functions in $A_1$, $A_2$ and $A_3$ are of $O_p(n)$, since the mean of the log likelihood derivatives are of order $O(n)$ (they are $O(1)$ for one observation) and the log composite likelihood is a finite sum of log likelihoods, e.g. the first term in $A_1$ is $O_p(n)$ since

$$1 \frac{\partial^3 cl(\psi, \lambda)}{n \partial \psi^3} \bigg|_{(t, \hat{\lambda}(t))} \to E_g \left( \left. \frac{\partial^3 cl(\psi, \lambda)}{\partial \psi^3} \right|_{(t, \lambda(t))} \right).$$

$\therefore R_n = O_p(n^{-1/2})$. \hfill $\Box$

**Lemma 2. (Royall, 2000)**

$$1 \frac{\partial^2 cl(\psi, \lambda)}{n \partial \psi^2} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \to -H_{\psi\psi}(\psi_g, \lambda_g)$$

$$1 \frac{\partial^2 cl(\psi, \lambda)}{n \partial \psi \partial \lambda} \bigg|_{(\psi_g, \hat{\lambda}(\psi_g))} \to -H_{\psi\lambda}(\psi_g, \lambda_g)$$

This follows from the Law of Large Numbers, since $\hat{\lambda}(\psi)$ is the MLE in the one dimensional model with a fixed $\psi$. 
Lemma 3. (Royall, 2000)

\[ \frac{d\hat{\lambda}(\psi)}{d\psi} \bigg|_{(\psi_g)} \rightarrow -\frac{H_{\psi\lambda}(\psi_g, \lambda_g)}{H_{\lambda\lambda}(\psi_g, \lambda_g)} \quad (B.18) \]

Proof. Let \( g(\psi, \hat{\lambda}(\psi)) = \frac{\partial cl(\psi, \lambda)}{\partial \lambda} \bigg|_{\hat{\lambda}(\psi)} \). Since \( g(\psi, \hat{\lambda}(\psi)) = 0 \) \( \forall \psi \), \( g(\psi, \hat{\lambda}(\psi)) \) is a constant. Thus \( dg/d\psi = 0 \).

\[ \frac{dg}{d\psi} = \frac{\partial g(\psi, \lambda)}{\partial \psi} \bigg|_{\hat{\lambda}(\psi)} + \frac{\partial g(\psi, \lambda)}{\partial \lambda} \bigg|_{\hat{\lambda}(\psi)} \frac{d\hat{\lambda}(\psi)}{d\psi} \]

Thus \( \frac{d\hat{\lambda}(\psi)}{d\psi} = -\frac{\frac{\partial^2 cl(\psi, \lambda)}{\partial \psi \partial \lambda}}{\frac{\partial^2 cl(\psi, \lambda)}{\partial \psi^2}} \bigg|_{\hat{\lambda}(\psi)} \) The conclusion follows from Lemma 2.

Then

\[ B_n \rightarrow -H_{\psi\psi} + H_{\psi\lambda} \frac{H_{\psi\lambda}}{H_{\psi\psi}} \]

Then (B.7) becomes;

\[ cl_p(\theta_1) - cl_p(\theta_g) \overset{d}{\rightarrow} N(-\frac{c^2}{2}a, c^2b) \quad (B.19) \]

where

\[ a = H_{\psi\psi} - H_{\psi\lambda} \frac{H_{\psi\lambda}}{H_{\psi\psi}} = H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \]

\[ b = [J_{\psi\psi}(\psi_g, \lambda_g) + \left( \frac{H_{\psi\lambda}(\psi_g, \lambda_g)}{H_{\lambda\lambda}(\psi_g, \lambda_g)} \right)^2 J_{\lambda\lambda}(\psi_g, \lambda_g) - 2 \frac{H_{\psi\lambda}(\psi_g, \lambda_g)}{H_{\lambda\lambda}(\psi_g, \lambda_g)} J_{\psi\lambda}(\psi_g, \lambda_g)] \quad \text{(from (B.16))} \]

\[ = H^{\psi\psi}(\psi_g, \lambda_g)^{-1}G^{\psi\psi}(\psi_g, \lambda_g)H^{\psi\psi}(\psi_g, \lambda_g)^{-1} \]

We see that (B.19) does not produce a bump function (the mean is not the negative half of the variance). If we take \( \left( \frac{cl_p(\theta_{g_1})}{cl_p(\theta_g)} \right)^{a/b} \) then

\[ \left( \frac{cl_p(\psi_1)}{cl_p(\psi_g)} \right)^{a/b} \rightarrow N(-\frac{c^2}{2b}, \frac{c^2a^2}{b}) \]
which results in a bump function. Then the probability of misleading evidence will be

\[
P_g \left\{ \left( \frac{cl_p(\psi_1)^{a/b}}{cl_p(\psi_g)} \right) \geq k \right\} \to \Phi \left\{ -\frac{(ca/b^{1/2})}{2} - \frac{\log(k)}{(ca/b^{1/2})} \right\} \quad (\text{the bump function}).
\]

In Theorem 2, \( c* = ca/b^{1/2} \).
Appendix C

Proofs for Chapter 3 and Chapter 4

Chapter 3

Lemma 4. (Location Normal) Suppose that \( x = (x_1, x_2, ..., x_n) \) is a sample from a \( N(\mu, 1) \) distribution where \( \mu \in \mathbb{R}^1 \) is unknown. Then a minimal sufficient statistic is \( T(x) = \bar{x} \), which follows a \( N(\mu, 1/n) \) distribution. If \( \mu \sim N(0, \tau^2) \) and \( \Psi(\mu) = \mu \), then the strength of evidence at \( \mu = 0 \) is given by,

\[
\Pi\left( \frac{\pi(\mu|\bar{x})}{\pi(\mu)} \leq \frac{\pi(0|\bar{x})}{\pi(0)} \right) = 1 - \Phi\left( \left( 1 + \frac{1}{n\tau^2} \right)^{1/2} \left( |\sqrt{n}\bar{x}_n| + \frac{\sqrt{n}\bar{x}}{1 + n\tau^2} \right) \right) \\
+ \Phi\left( \left( 1 + \frac{1}{n\tau^2} \right)^{1/2} \left( -|\sqrt{n}\bar{x}_n| + \frac{\sqrt{n}\bar{x}}{1 + n\tau^2} \right) \right).
\]

As \( \tau^2 \to \infty \), the strength of evidence converges to \( 2(1 - \Phi(|\sqrt{n}\bar{x}|)) \). As \( n \to \infty \), the strength of evidence converges in distribution to a \( Unif(0, 1) \) distribution, when the true value of \( \mu \) is 0, and it converges to 0 almost surely when \( \mu \neq 0 \).

Proof. The relative belief ratio at \( \mu \) is given by

\[
RB(\mu) = \frac{\pi(\mu|\bar{x})}{\pi(\mu)} = (n\tau^2 + 1)^{1/2} \exp\left\{ -\frac{(\mu - \bar{x})^2}{2/n} + \frac{\bar{x}^2}{(n\tau^2 + 1)2/n} \right\}. \tag{C.1}
\]
We have that
\[
\Pi\left(\frac{\pi(\mu|x)}{\pi(\mu)} \leq \frac{\pi(0|x)}{\pi(0)} | \bar{x} \right) = \Pi\left(\exp\left\{ -\frac{(\mu - \bar{x})^2}{2/n} \right\} \leq \exp\left\{ -\frac{(\bar{x})^2}{2/n} \right\} \right)
\]
\[
= \Pi(\mu \geq |\bar{x} + \bar{x} - \frac{n\tau^2}{n\tau^2 + 1}\bar{x}) + \Pi(\mu \leq -|\bar{x} + \bar{x} - \frac{n\tau^2}{n\tau^2 + 1}\bar{x})
\]
\[
= 1 - \Phi\left(\frac{|\bar{x} + \bar{x} - \frac{n\tau^2}{n\tau^2 + 1}\bar{x}}{\sqrt{\frac{\tau^2}{n\tau^2 + 1}}}\right) + \Phi\left(\frac{-|\bar{x} + \bar{x} - \frac{n\tau^2}{n\tau^2 + 1}\bar{x}}{\sqrt{\frac{\tau^2}{n\tau^2 + 1}}}\right)
\]
\[
= 1 - \Phi\left(\frac{\sqrt{n\bar{x}_n}}{1 + \frac{1}{n\tau^2}}\right)^{1/2} \left(|\sqrt{n\bar{x}_n}| + \frac{\sqrt{n\bar{x}}}{1 + n\tau^2}\right)
\]
\[
+ \Phi\left(\frac{\sqrt{n\bar{x}_n}}{1 + \frac{1}{n\tau^2}}\right)^{1/2} \left(-|\sqrt{n\bar{x}_n}| + \frac{\sqrt{n\bar{x}}}{1 + n\tau^2}\right). \quad (C.2)
\]

Then Eq. (C.2) \(\rightarrow 2(1 - \Phi(|\sqrt{n\bar{x}}|))\) as \(\tau^2 \rightarrow \infty\).

As \(n \rightarrow \infty\), and when \(\mu = 0\), \(\sqrt{n\bar{x}} \rightarrow^d z\), where \(z \sim N(0, 1)\). Then by the Continuous Mapping theorem, Eq. (C.2) converges in distribution to \(2(1 - \Phi(|z|))\), which has a uniform distribution on \((0,1)\). When \(\mu \neq 0\), since \(\bar{x} \rightarrow \mu\) a.s., where \(\mu \neq 0\), Eq. (C.2) converges to 0 a.s. \(\square\)

**Theorem 12. (Location Normal)** If \(T(x) = \bar{x}\) where \(\bar{x} \sim N(\mu, 1/n)\) and \(\mu \sim N(0, \tau^2)\), then for a fixed value of \(RB(0)\), the strength of evidence goes to 0, as either \(n\) or \(\tau^2 \rightarrow \infty\).

**Proof.**
\[
RB(0) = \frac{\pi(0 | \bar{x})}{\pi(0)} = (1 + n\tau^2)^{1/2} \exp\{-n \left(1 + 1/n\tau^2\right)^{-1} \bar{x}^2/2\}
\]

Then \(\bar{x} = \pm \sqrt{-2(1 + n\tau^2)/n^2\tau^2}log\{RB_0(n\tau^2 + 1)^{-0.5}\}\) for a fixed \(RB(0)\). Note that as either \(n\) or \(\tau^2 \rightarrow \infty\), \(|\bar{x}_n| \rightarrow \infty\). By Lemma 4, the measure of strength of evidence is
\[
\Pi\left(\frac{\pi(\mu|\bar{x}_n)}{\pi(\mu)} \leq \frac{\pi(0|\bar{x})}{\pi(0)} | \bar{x}_n \right) = 1 - \Phi\left(\frac{\sqrt{n\bar{x}_n}}{1 + \frac{1}{n\tau^2}}\right)^{1/2} \left(|\sqrt{n\bar{x}_n}| + \frac{\sqrt{n\bar{x}}}{1 + n\tau^2}\right)
\]
\[
+ \Phi\left(\frac{\sqrt{n\bar{x}_n}}{1 + \frac{1}{n\tau^2}}\right)^{1/2} \left(-|\sqrt{n\bar{x}_n}| + \frac{\sqrt{n\bar{x}}}{1 + n\tau^2}\right).
\]
Then $|\sqrt{n}x_n| \to \infty$ and
\[
\frac{\sqrt{n}x}{1 + n\tau^2} = \pm \sqrt{\frac{-2}{(1 + n\tau^2)(n\tau^2)}\log\{RB_0(n\tau^2 + 1)^{-0.5}\}} \to 0
\]
as either $n$ or $\tau^2 \to \infty$. Thus, the measure of strength of evidence converges to $1 - \Phi(\infty) + \Phi(-\infty) = 0$ as either $n$ or $\tau^2 \to \infty$. 

\[\square\]

**Proof of Theorem 8.** We have that
\[
\frac{RB(\psi)}{\psi_0} = \frac{\pi_{\psi}(\psi)}{\pi_{\psi_0}(\psi_0)} \sum_{\theta : \Psi(\theta) = \psi} \Pi(\theta | \psi) f_{\theta,n}(x)
\]
and, for $\theta_0$ such that $\Psi(\theta) = \psi$, let $A_n(\theta_0) = \{\theta : n^{-1} \log (RB(\Psi(\theta))/RB(\psi_0) \leq 0)\}$.

Note that $\theta_0 \in A_n(\theta_0)$. Now,
\[
\frac{1}{n} \log \left( \frac{RB(\psi)}{RB(\psi_0)} \right) = \frac{1}{n} \log \left( \frac{\pi_{\psi}(\psi)}{\pi_{\psi_0}(\psi_0)} \right) + \frac{1}{n} \log \left( \frac{f_{\theta(n)}(x)}{f_{\theta_0}(x)} \right) + \frac{1}{n} \log \left( \frac{\sum_{\theta : \Psi(\theta) = \psi} \pi(\theta | \psi) f_{\theta,n}(x)/f_{\theta_0}(x)}{\sum_{\theta : \Psi(\theta) = \psi} \pi(\theta | \psi) f_{\theta,n}(x)/f_{\theta_0}(x)} \right)
\]
(C.3)
where $f_{\theta(n)}(x) = \sum_{\theta : \Psi(\theta) = \psi} f_{\theta,n}(x)$. Observe that, as $n \to \infty$ the first term on the right hand side of (C.3) converges to 0 as does the third term since $0 < \min \{\pi(\theta | \psi) : \Psi(\theta) = \psi\} < 1$. Now putting $f_{\theta(n)}(x) = \max \{f_{\theta,n}(x) : \Psi(\theta) = \psi\}$, the second term on the right-hand side of (C.3) equals
\[
\frac{1}{n} \log \left( \frac{f_{\theta(n)}(x)}{f_{\theta_0}(x)} \right) - \frac{1}{n} \log \left( \frac{f_{\theta(n)}(x)}{f_{\theta_0}(x)} \right) + \frac{1}{n} \log \left( \frac{f_{\theta(n)}(x)}{f_{\theta_0}(x)} \right).
\]
(C.4)
Note that the third term in Eq. (C.4) is bounded above by $n^{-1} \log (\#\{\theta : \Psi(\theta) = \psi\})$ which converges to 0. Now by the strong law, when $\theta_0$ is true, as $n \to \infty$, then
\[
n^{-1} \log (f_{\theta,n}(x)/f_{\theta_0,n}(x)) \to E_{\theta_0}(\log (f_{\theta}(X)/f_{\theta_0}(X))).
\]
By Jensen’s inequality $E_{\theta_0}(\log (f_{\theta}(X)/f_{\theta_0}(X))) \leq \log E_{\theta_0}(f_{\theta}(X)/f_{\theta_0}(X)) = 0$ and the inequality is strict when $\theta \neq \theta_0$ while $E_{\theta_0}(\log (f_{\theta_0}(X)/f_{\theta_0}(X))) = 0$. Therefore, using $\#\{\theta : \Psi(\theta) = \psi\} < \infty$, the first term in Eq. (C.4) converges to $\max \{E_{\theta_0}(\log (f_{\theta}(X)/f_{\theta_0}(X))) : \Psi(\theta) = \psi\}$ while the second term converges to $\max \{E_{\theta_0}(\log (f_{\theta}(X)/f_{\theta_0}(X))) : \Psi(\theta) = \psi\} = 0$. Therefore, we have that there exists $n_0$ such that $A_n(\theta_0) = \Theta$ for all $n \geq n_0$ and so $\Pi_{\psi} (RB(\psi) \leq RB(\psi_0) | T(x)) = \Pi_{\psi} (A_n(\theta_0) | T(x)) = 1$. 

\[\square\]
Chapter 4

Proof of Theorem 11. If \( C = \{ \psi : RB(\psi) = RB(0) \} \), then \( \Pi_\psi(C \mid x) = 0 \) and so \( \Pi_\psi(RB(\psi) \leq RB(0) \mid x) = \Pi_\psi(RB(\psi) < RB(0) \mid x) \). We use the following result.

Lemma (i) \( \Pi_\psi(\{ \psi : RB(\psi) < RB(0) \} \Delta \lim(A_N \setminus C) \mid x) = 0 \) where \( A_N = \bigcup \{(\psi_{i/N}, \psi_{(i+1)/N}] : N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) < NF(\psi_{1/N} \mid x)\} \),
(ii) \( \Pi_\psi(\lim sup B_N \mid x) = 0 \) where \( B_N = \bigcup \{(\psi_{i/N}, \psi_{(i+1)/N}] : N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) = NF(\psi_{1/N} \mid x)\} \).

Proof: (i) Suppose \( RB(\psi) < RB(0) \) and let \( \delta = RB(0) - RB(\psi) \). Since \( RB(0) = \lim_{N \to \infty} F(\psi_{1/N} \mid x)/F(\psi_{1/N}) = \lim_{N \to \infty} NF(\psi_{1/N} \mid x) \) there exists \( N_\psi \) such that for all \( N > N_\psi \) we have \( NF(\psi_{1/N} \mid x) > RB(0) - \delta/2 \). For each \( N \) there exist unique prior quantiles such that \( \psi_{i/N}(\psi) < \psi \leq \psi_{(i+1)/N}(\psi) \). By the convergence of these quantiles to \( \psi \) as \( N \) increases, which implies the convergence of \( N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) \) to \( RB(\psi) \), there exists \( N'_\psi > N_\psi \) such that for all \( N > N'_\psi \) we have that \( N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) < RB(0) - \delta/2 \). This proves that \( \{ \psi : RB(\psi) < RB(0) \} \subset \liminf A_N \).

Suppose now that \( \psi \in \lim sup A_N \). Then there are infinitely many values of \( N \) such that \( N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) < NF(\psi_{1/N} \mid x) \). This implies that \( RB(\psi) \leq RB(0) \) and so \( \lim sup(A_N \setminus C) = (\lim sup A_N) \setminus C \subset \{ \psi : RB(\psi) < RB(0) \} \subset (\liminf A_N) \setminus C = \liminf(A_N \setminus C) \) and we are done.

(ii) Suppose that \( \psi \in \lim sup B_N \). Then \( (\psi_{i/N}(\psi), \psi_{(i+1)/N}(\psi)] \subset B_N \) for infinitely many \( N \) and this implies that \( \psi \in C \) and the result follows.

Now let \( \epsilon > 0 \) and suppose \( i_0 = 1 \). From part (i) of the Lemma we have that \( \Pi_\psi(RB(\psi) \leq RB(0) \mid x) = \lim \Pi_\psi(A_N \setminus C \mid x) = \lim \Pi_\psi(A_N \mid x) \) so there exists \( N_0 \) such that \( |\Pi_\psi(RB(\psi) \leq RB(0) \mid x) - \Pi_\psi(A_N \mid x)| < \epsilon/3 \) for all \( N > N_0 \). By part (ii) of the Lemma there exits \( N_{00} \) such that for all \( N > N_{00} \) we have that \( \Pi_\psi(B_N \mid x) < \epsilon/6 \). Hereafter we suppose that \( N > \max\{N_0, N_{00}\} \).

We now prove that for all \( M_1 \) and \( M_2 \) large enough, (1) is within \( \epsilon \) of \( \Pi_\psi(RB(\psi) \leq \)
$RB(0) \mid x)$. Let $S_N = \{i : N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) < NF(\psi_{1/N} \mid x)\}$ and note that $\Pi_\Psi(A_N \mid x) = \sum_{i \in S_N} (F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x))$. By Theorem 2.3.1 of Serfling (1980) we have that $\hat{\psi}_{i/N} \rightarrow \psi_{i/N}$ almost surely as $M_1 \rightarrow \infty$ and $M_2 \rightarrow \infty$. Also, by the Glivenko-Cantelli theorem, we have that $\hat{F}(\psi)$ converges almost surely and uniformly to $F(\psi)$ and similarly $\hat{F}(\psi \mid x)$ converges to $F(\psi \mid x)$ as $M_1 \rightarrow \infty$ and $M_2 \rightarrow \infty$. In particular, this implies that $\hat{R}B(0)$ converges almost surely to $NF(\psi_{1/N} \mid x)$.

Suppose $i \in S_N$. Let $\delta = NF(\psi_{1/N} \mid x) - N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x))$. Then, for all $M_1, M_2$ large enough, $\hat{R}B(0) > NF(\psi_{1/N} \mid x) - \delta/2$ and

$$N(\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x)) < N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) + \delta/2.$$ 

This implies that $i \in \hat{S}_N = \{i : \hat{R}B(\psi_{i/N}) \leq \hat{R}B(0)\}$ for all $M_1, M_2$ large enough and so $S_N \subset \hat{S}_N$ for all $M_1, M_2$ large enough. Furthermore, the contribution $(\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x))$ that this index makes to the sum (1) converges to $(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x))$ as $M_1 \rightarrow \infty$ and $M_2 \rightarrow \infty$.

If $i \notin S_N$ and $N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) > NF(\psi_{1/N} \mid x)$, then the same argument shows that $i \notin \hat{S}_N$ for all $M_1, M_2$ large enough. Since $(\hat{S}_N) \leq N$, then for all $M_1, M_2$ large enough, we have that $\sum_{i \in \hat{S}_N} (\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x)) = \sum_{i \in S_N} (\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x)) + \sum_{i \in S_N \setminus \hat{S}_N} (\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x))$ where the second sum contains at most those terms corresponding to $i$ where $N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) = NF(\psi_{1/N} \mid x)$. The first sum converges almost surely to $\Pi_\Psi(A_N \mid x)$ and the limit supremum of the second term is bounded above by $\sum_{D(N,x)} (F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) = \Pi_\Psi(B_N \mid x)$ where $D(N,x) = \{i : N(F(\psi_{(i+1)/N} \mid x) - F(\psi_{i/N} \mid x)) = NF(\psi_{1/N} \mid x)\}$. So for all $M_1, M_2$ large enough we have $|\sum_{i \in S_N} (\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x)) - \Pi_\Psi(A_N \mid x)| < \epsilon/3$ and

$$\left| \sum_{i \in \hat{S}_N \setminus S_N} (\hat{F}(\psi_{(i+1)/N} \mid x) - \hat{F}(\psi_{i/N} \mid x)) \right| < \epsilon/3$$

and this finishes the proof. The proof of the more general case, where $i_0$ is not constrained to be 1, follows easily by noting that $\psi_{i_0/N} \rightarrow 0$ as $N \rightarrow \infty$. 

Appendix C. Proofs for Chapter 3 and Chapter 4
Bibliography


BIBLIOGRAPHY


