**OBJECTIVE BAYESIAN ANALYSIS OF
THE 2×2 CONTINGENCY TABLE AND
THE NEGATIVE BINOMIAL DISTRIBUTION**

A Thesis presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by
John Christian Snyder
Dr. Dongchu Sun, Thesis Supervisor
MAY 2018

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

**OBJECTIVE BAYESIAN ANALYSIS OF
THE 2×2 CONTINGENCY TABLE AND
THE NEGATIVE BINOMIAL DISTRIBUTION**

presented by John Snyder,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dongchu Sun

Dr. Zhuoqiong He

Dr. Athanasios Micheas

Dr. Suhwon Lee

Dr. X. H. Wang

To Qing, whose enduring love and support fuels the
passion for all that I do.

ACKNOWLEDGMENTS

I would like to extend my sincerest gratitude to my advisor, Dr. Dongchu Sun. Dr. Sun has, through sheer force of his academic personality, pushed me to approach the peak of my mathematical ability. By being unsatisfied with anything other than rigorous proofs and exact results, he has made me appreciate the excitement of creating knowledge through research, and his advising has allowed me to gaze into a world of technical detail which would have been inaccessible otherwise. Thanks to him, I have an appreciation for how beautiful and elegant the fusing of statistics and mathematics can be, and I will be forever grateful for that. I would also like to thank Dr. Chong He, whom has taught me and mentored me in various projects and selected me to work for her at the Missouri Department of Conservation. Additionally, her comments over the years in our research group meetings have been invaluable for advancing my projects. Third, I thank Dr. Sakis Micheas, who has taught me the most challenging mathematical statistics courses I have ever seen, and has always been willing to talk to me about new research applications I have in mind. His passionate teaching style and approachable personality have left a strong impression on me, and I will forever instill these values deep inside me. Next, I am thankful to Dr. Suhwon Lee, who has facilitated my involvement with students at many levels both in and out of our department through Datafest and other activities. These student interactions have kept me grounded and have constantly reminded me about why the work we do as Statisticians is so important. Finally, I would like to thank Dr. X.H. Wang. Dr. Wang's experience with objective Bayesian analysis is very valuable to the completion of this work.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	ii
LIST OF TABLES	vii
LIST OF FIGURES	ix
ABSTRACT	xi
CHAPTER	
1 Introduction	1
1.1 Overview of Bayesian Analysis	2
1.2 Objective Bayesian Analysis	3
1.2.1 Non-informative Priors	4
1.2.2 The Jeffreys Prior	6
1.3 Reference Priors	8
1.3.1 Quantifying Information	9
1.3.2 Reference Priors for Nuisance Parameters	11
1.3.3 Reference Priors for many Nuisance Parameters	12
1.3.4 Reparameterizations of the Parameter Space	17
1.4 Objective Priors for Discrete Parameter Spaces	17
1.4.1 Reference Priors for Discrete Parameters	18
1.4.2 A Loss Based Approach	20
1.5 Outline	22

2	Objective Bayesian Analysis of the 2x2 Contingency Table	24
2.1	Introduction	24
2.1.1	Sampling Schemes	25
2.2	Objective Bayesian Analysis for Unrestricted Poisson Sampling	31
2.2.1	Relative Risk	32
3	A Reference Prior for the Odds Ratio of the 2x2 Contingency Table	34
3.1	Parameterization	34
3.2	Proof of Theorem 3.1.1	38
3.2.1	Derivation of $\pi_{3k}(\eta_3 \eta_2, \eta_1, \theta)$ and $\pi_{2k}(\eta_2 \eta_1, \theta)$	38
3.2.2	Derivation of $\pi_1(\eta_1 \theta)$	39
3.2.3	Derivation of $\pi(\theta)$	41
3.3	Behavior of the Normalizing Constant $H(\theta)$ and $G(\theta)$	43
3.4	Behavior of the Marginal Reference Prior $\pi^R(\theta)$	51
3.4.1	Reference Prior for the Log odds Ratio	53
3.5	Behavior of the Conditional Reference Prior $\pi^R(\eta_1 \theta)$	55
3.6	Figures and Tables	58
4	Reference Posterior Analysis and Computation for the Odds Ratio of the 2x2 Contingency Table	61
4.1	Posterior Distributions	62
4.2	Posterior Sampling and Computation	69
4.3	Numerical Examples	70
4.3.1	Simulation Study	70
4.3.2	Data Analysis	72

4.4	A Note on the Requirement $n_1 \geq 1$	77
4.4.1	One large value in only one cell	79
4.5	Software Implementations	83
4.5.1	R Package: OBayes2by2Table	83
4.5.2	GPU Computation	88
4.6	Discussion and Future Work	91
4.7	Figures and Tables	93
5	Objective Bayesian Analysis of the Negative Binomial Distribution	102
5.1	Continuous Dispersion Parameterization	103
5.1.1	Hierarchical Objective Prior	106
5.1.2	Hierarchical Priors	107
5.1.3	Integrated Likelihood for α	108
5.1.4	Posterior Impropropriety	109
5.1.5	Direct Computation of $\pi(\mu, \alpha)$	114
5.2	Discrete Parameterization	127
5.2.1	Objective Bayesian Analysis	128
5.3	Discussion and Future Work	136
5.4	Figures and Tables	137
6	Summary and Concluding Remarks	148
APPENDIX		
A	Miscellaneous Results	151
A.1	Closed Form for $G(\theta)$	152

A.2	The Jeffreys Rule Prior for the Odds Ratio θ	158
A.3	Numerical Computation of $G(\theta)$	160
BIBLIOGRAPHY		161
VITA		167

LIST OF TABLES

Table		Page
2.1	a 2×2 table	24
4.1	Fisher’s Tea Tasting Experiment	72
4.2	Comparison to Fisher’s Exact Test	73
4.3	Data on Cell phone Computer types	75
4.4	A “Table” with one value	77
4.5	Comparison of Techniques	78
4.6	Comparison of Techniques on log-scale	78
4.7	A “Table” with one value	79
4.8	Log-scale Posterior results for θ for $\pi^R(\theta)$ as well as $\pi^R(\theta n)$ for various n	80
4.9	Mean Absolute Deviation	96
4.10	95% Frequentist Coverage	96
4.11	Mean Absolute Deviation: $\eta_2 = 5$	97
4.12	95% Frequentist Coverage: $\eta_2 = 5$	97
4.13	Mean Absolute Deviation: $\eta_2 = 10$	98
4.14	95% Frequentist Coverage: $\eta_2 = 10$	98
4.15	Mean Absolute Deviation: $\eta_2 = 20$	99

4.16	95% Frequentist Coverage: $\eta_2 = 20$	99
4.17	Mean Absolute Deviation: $\eta_2 = 50$	100
4.18	95% Frequentist Coverage: $\eta_2 = 50$	100
4.19	Simulation Results: Fully Expanded Mean Absolute Deviations . . .	101
4.20	Simulation Results: Fully Expanded 95% Frequentist Coverage	102
5.1	Polynomial coefficients of $\sqrt{\alpha}\pi(\alpha)$ and taylor series coefficients of $\psi(\alpha+1)$	137
5.2	Ratio of MSEs for $E[\mu y]$ and $\hat{\mu}$	137
5.3	Ratio of MSEs for $median[\mu y]$ and $\hat{\mu}$	137
5.4	Ratio of MSEs for $E[\alpha y]$ and $\hat{\alpha}$	138
5.5	Ratio of MSEs for $median[\alpha y]$ and $\hat{\alpha}$	138

LIST OF FIGURES

Figure		Page
3.1	Plot of $\pi^R(\theta)$ compared to $\pi^*(\theta) \propto 1/\theta$	58
3.2	Plot of $\pi^R(\theta)$ compared to $\pi^*(\theta) \propto 1/\theta$ as well as a transformed Laplace Prior.	59
3.3	Plot of $\pi(\eta_1 \theta)$ for some θ	60
3.4	Plot of $\pi(\theta, \eta_1)$	60
4.1	Illustration of a CUDA Block	90
4.2	Posterior samples of parameters for the Cell phone data	93
4.3	Histogram of 1,000,000 samples from $\pi(\theta \mathbf{n})$	94
4.4	Reference Posteriors of $\xi = \frac{1}{\theta}$ with only one cell populated by a large n value	95
5.1	Plot of estimated prior $\pi(\alpha)$ for $n = 90$, $c = 1000$, $m = 750$, and $k = 1000$ 139	
5.2	Plot of estimated prior $\pi(\alpha)$ for $n = 90$, $c = 1000$, $m = 750$, and $k = 1000$ 140	
5.3	Contour plot of $\pi(\lambda, \alpha \mathbf{y})$ for various levels of overdispersion with $\lambda = 5$ 141	
5.4	Plots of the Marginal Posterior of α for $n = 50$ generated from a Poisson distribution with various λ	142

5.5	Plots of the Marginal Posterior of α for $n = 200$ generated from a Poisson distribution with various λ	143
5.6	Plots of the Marginal Posterior of α for $n = 500$ generated from a Poisson distribution with various λ	144
5.7	Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 5$	145
5.8	Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 25$	146
5.9	Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 100$	147
5.10	Plot of $\pi(r, p x)$ using the objective prior for $\pi(r p)$ and a uniform prior for p	148

ABSTRACT

In Bayesian analysis, the “objective” Bayesian approach seeks to select a prior distribution not by using (often subjective) scientific belief or by mathematical convenience, but rather by deriving it under a pre-specified criteria. This approach takes the decision of prior selection out of the hands of the researcher. Ideally, for a given data model, we would like to have a prior which represents a “neutral” prior belief in the phenomenon we are studying. In categorical data analysis, the odds ratio is one of several approaches to quantify how strongly the presence or absence of one property is associated with the presence or absence of another property. In this project, we present a Reference prior for the odds ratio of an unrestricted 2×2 table. Posterior simulation can be conducted without MCMC and is implemented on a GPU via the CUDA extensions for C. Simulation results indicate that the proposed approach to this problem is far superior to the widely used Frequentist approaches that dominate this area. Real data examples also typically yield much more sensible results, especially for small sample sizes or for tables that contain zeros. An R package is also presented to allow for easy implementation of this methodology. Next, we develop an approximate reference prior for the negative binomial distribution, applying this methodology to a continuous parameterization often used for modeling over-dispersed count data as well as the typical discrete case. Results indicate that the developed prior equals the performance of the MLE in estimating the mean of the distribution but is far superior when estimating the dispersion parameter.

Chapter 1

Introduction

The structure of all statistical approaches generally begins by considering a random process where we observe data (called x) which can be described by a model $f(x|\theta)$ where θ is an unknown quantity of interest. The behavior of θ is typically the driving factor in most analysis. For instance, if we are studying the public's feeling about a new governmental policy, we may consider x to be a “yes” or “no” response to a favorable feeling when asked in a survey, and θ to be the proportion of voters which favor (and thus answer “yes”) this policy. In manufacturing, x could be how long it takes a component to fail, and θ could be the average time until failure of all such components. Typically, it is impractical to consider every unit in a population to determine θ exactly, so we must use properties of the probability model $f(x|\theta)$ along with our understanding about randomness to *infer* about the nature of θ . There are many approaches to conduct this inference, and we focus our work on the Bayesian approach.

1.1 Overview of Bayesian Analysis

The general framework for Bayesian analysis involves starting from a *prior* $\pi(\theta)$ describing the *a priori* knowledge about the parameters(θ) of a model to a *posterior* distribution $\pi(\theta|\mathbf{x})$ which summarizes all the information known about θ after observation of the data $\mathbf{x}$. In a sense, the prior $\pi(\theta)$ describes our initial understanding (or guess, if no such knowledge is available) about the behavior of the quantity of interest and the posterior $\pi(\theta|\mathbf{x})$ describes the behavior of θ after observation of data from the process $f(x|\theta)$. Formally, the posterior distribution is obtained by an application of Bayes' rule as

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta), \quad (1.1)$$

where $f(\mathbf{x}|\theta)$ represents the likelihood after observation of an independent and identically distributed(i.i.d.) random sample from $f(x|\theta)$. Furthermore, $\mathbf{x}$ represents the vector of data observations $(x_1, \dots, x_n)$ and the parameter θ can be a scalar quantity or a vector of d parameters $\theta = (\theta_1, \dots, \theta_d)$.

Because the prior $\pi(\theta)$ represents our initial uncertainty of θ 's true value, it must be specified. There are two schools of thought for defining a prior, the *objective* approach and the *subjective* approach. Subjective priors often assume some informative structure on θ based on knowledge of the process and/or historical data. For example, *Robust Bayesian Analysis* is a general technique which allows the user to provide some a-priori information for the parameter θ and consider a wide variety of prior functions $\pi_i(\theta)$ each of which conform to this specification, which is typically derived from scientific knowledge (Berger, 1985). The answer is then reported as

an interval obtained from a combination of the posteriors $\pi_i(\theta|\mathbf{x})$, each of which was constructed from $\pi_i(\theta)$ through (1.1). Goldstein (2006) provides convincing arguments for the use of subjective Bayesian Procedures, especially when extremely little data is available or in situations which have exceptionally complicated model structures. Still, subjective priors are not the focus of this work, but the reader is referred to Ramsey (1964), French (1982), Lindley (1987), and Berger (2004) for both philosophical as well as practical discussions and examples.

In summary, subjective Bayesian approaches rely on historical knowledge or scientific belief, but many times historical data does not exist. Even if it does, many modern modeling applications are exceedingly complicated, making it unreasonable to claim enough knowledge to place informative priors on a large number of parameters. Still, if a subjective prior is to be used, it can easily spawn a debate about the choice since it could potentially have a large impact on the corresponding posterior distribution. The objective Bayesian approach seeks to *derive* a prior which is determined once the *data* model $f(x|\theta)$ is chosen. The general idea is to create a prior from the data model such that it has certain mathematical advantages; for example, the reference prior technique can be applied to maximize the difference between the prior and posterior distributions in an information-theoretic sense so that the data can have the maximum possible impact on the resulting posterior.

1.2 Objective Bayesian Analysis

Performing Bayesian analysis according to the objective paradigm is advantageous in a number of ways. Principally, it takes prior specification out of the hands of the

researcher, allowing them to focus on building the data model. Furthermore, it is well known that statistical software packages are becoming more accessible to researchers outside the field of statistics. Objective Bayesian inference has the potential to be a major part of the statistical “machine,” so to speak. That is, once a researcher decides on a data model, the subsequent analysis has already been laid out by the framework of the objective prior and corresponding posterior distribution. Most importantly, since the entire framework only needs to be derived once, the author can (and should) develop robust and easy-to-use software so that researchers and scientists can apply their methodology to their data. A few general methods to obtain $\pi(\theta)$ after model specification are the Jeffreys prior (Jeffreys, 1961), Reference Priors (Bernardo, 1979) and Probability Matching Priors. The “hands-off” approach implied by the Objective Bayesian paradigm has a rich history dating back to Reverend Thomas Bayes original 1763 work.

1.2.1 Non-informative Priors

In his famous essay, Bayes (1763) did not mention the various aspects of Bayesian analysis that we know and use today. In fact, there was no mention of a “prior” $\pi(\theta)$ in this exposition at all; however it was clear that in his proposed methodology to estimate the unknown probability of some physical phenomenon, we should use a uniform distribution to represent our neutral prior belief about the nature of the unknown parameter describing it. Bayes described a series of independent experiments roughly as follows:

Say we randomly place a white ball on a pool table of “length” 1, with every position having equal probability. Our task is to guess the position of the ball on the table(i.e. the distance to one end of the table). Our method to accomplish this is to throw another ball on the table, and if it ends up closer to the edge, we record this as a “success.” We repeat this process, and from the number of successes we observe we can deduce the position of the white ball.

Bayes considered that the position of the white ball as both random as well as having an equal chance of originating at any position on the table. In modern terms, this experiment is a series of Bernoulli trials with probability of success $p \in (0, 1)$. Furthermore, this structure also supposes p to be a uniformly distributed random variable on the interval $(0, 1)$. Thus, if we toss the second ball n times and observe x “successes,” this experiment has $f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$ and $\pi(p) = 1$. The posterior distribution $\pi(p|x)$ would be computed using (1.1)

$$\pi(p|x) = \frac{p^x (1-p)^{n-x} \cdot 1}{\int_0^1 p^x (1-p)^{n-x} \cdot 1 dp} = \frac{p^x (1-p)^{n-x}}{B(x+1, n-x+1)}$$

which is a $Beta(x+1, n-x+1)$ distribution. If we happen to observe many successes relative to our trials, p is likely to be large in that it is more likely that we would observe a success on our $n+1$ th trial. Bayes essentially considered the conditional distribution of p , given we have observed our data(successes and failures), and this is where the modern notion of a posterior distribution comes from. However, the formalization of this process would come several years later.

As presented in Stigler (1974), Laplace independently formalized this process in 1774. Laplace formed his “Principle” as

$$P(\theta_i|F) = \frac{P(F|\theta_i)}{\sum_{j=1}^n P(F|\theta_j)}$$

where $\theta_1, \theta_2, \dots, \theta_n$ represent the n *causes* (or parameter support in today’s terminology) and F is the “event.” This is a formalization of Bayes’s earlier work, and was the first time the Bayesian philosophy was presented in this way. We can see here that Laplace’s interpretation puts a uniform prior on θ_i , and this is reasonable from an axiomatic perspective. This philosophy, later known as the *principle of indifference*, states that if we know nothing about the “causes” $\theta_1, \theta_2, \dots, \theta_n$, they should be assigned a probability equal to $\frac{1}{n}$. In his later work, Laplace formally declared that the posterior distribution for θ should be proportional to the likelihood for the data (Feinberg, 2006). In current terminology we represent this as

$$\pi(\theta|x) \propto f(x|\theta).$$

Expressing the posterior in this way is what led to the concept of *inverse probability* being the primary approach to statistical analysis up until the invention of the frequentist machinery (Feinberg, 2006). It is “inverse” in the sense that we use information from the data observations x to directly infer backwards to the parameter value θ .

1.2.2 The Jeffreys Prior

The main problem with the inverse probability approach is that assuming a uniform prior does not actually represent ignorance. In fact, Bernardo and Smith (2000) point out that “...there is no prior that represents ignorance.” They discuss that every prior possible carries *some* information, and this concept will be discussed further in the section on *reference priors*. In fact, the uniform prior actually carries a substantial

amount of information, as the assumption that θ can take any possible value with equal probability is fairly strong (Bernardo and Smith, 2000). To illustrate this, consider a model $f(x|\theta)$ and choose the uniform prior for θ . In this case, the prior for $\psi = g(\theta)$ would *not* be uniform. The philosophical justifications for a constant prior break down almost immediately here, as it is very common to transform a parameter space even in the simple rescaling of variables and there is no reason our ignorance on one scale should transform to anything other than ignorance on another scale. In contrast to the uniform approach, the Jeffreys rule prior (Jeffreys, 1961) is a way of obtaining the prior for θ based on the Fisher's Information Matrix I_θ of the model $f(\mathbf{x}|\theta)$, which is defined as

$$I_\theta = -E_\theta \left[\frac{d^2}{d\theta^2} \log f(\mathbf{x}|\theta) \right].$$

This quantity describes the amount of information about θ that the model carries because the expectation is with respect to the model $f(x|\theta)$ and thus the effect of the random variable x has been averaged out. Once we have I_θ , the Jeffreys prior is then defined as

$$\pi_J(\theta) \propto \sqrt{|I_\theta|}.$$

The main advantage of Jeffreys prior is that it is invariant under re-parameterization. If we have a prior $\pi(\theta)$ and the re-parameterization $\psi = g(\theta)$, this gives us

$$\pi(\psi) \propto \pi(g(\theta))|J|, \tag{1.2}$$

where J is the Jacobian of the transformation. This property makes Jeffreys prior especially appealing for scale parameters. Essentially, this property tells us that once we derive a prior distribution for θ , the prior for ψ is *exactly* the same after adjusting the density with the Jacobian to account for the stretching or shrinking of the parameter space induced by the transformation. This invariance also makes for an attractive justification for applying this structure because it is not dependent on the scale of the set of parameters that are chosen to describe parameter space.

1.3 Reference Priors

Recall that the posterior distribution arises as the combination of prior knowledge about a parameter θ and the information gained by observing the data x through the likelihood. Therefore, if we had a procedure for quantifying the amount of information gained about the process through the observation of the likelihood (i.e. the change in knowledge from prior to posterior), we know that this change must be describing the change in our knowledge about the parameter θ which is obtained from observation of the data itself. Recalling that a key philosophy behind the objective Bayesian paradigm is to obtain a posterior distribution which reflects the data as much as possible, a concept reinforced by Berger (2004), the key behind the Reference Prior approach is to select a prior which maximizes the increase in (expected) information from prior to posterior after observation of the data, so that we can make the “bridge” representing our gain in knowledge about θ after the observation of the experiment as large as possible.

1.3.1 Quantifying Information

Lindley (1992), building on the work of Shannon (1948), discussed the concept of how much information is gained about θ after conducting an experiment. Bernardo (1979) used this idea to formulate the idea of “missing information,” which he then used to first introduce the concept of a reference prior.

First, consider independent and identically distributed data $x = (x_1, x_2, \dots, x_n)$ from a statistical model $f(x|\theta)$, where θ is an unknown parameter which describes the nature of the data. The amount of information gained about θ after observation of the data x could be described as the amount of information gained when we consider the posterior for θ given the data, $\pi(\theta|x)$ instead of the prior distribution for θ , $\pi(\theta)$. The amount of information gained is given by the *Kullback-Leibler* divergence from prior to posterior, $D_{KL}(\pi(\theta|x)||\pi(\theta))$, is defined as the expectation under the posterior of the difference in the logarithms between the posterior and prior, or

$$D_{KL}(\pi(\theta|x)||\pi(\theta)) = \int_{\Theta} \pi(\theta|x) \log \frac{\pi(\theta|x)}{\pi(\theta)} d\theta.$$

We note $D_{KL}(\pi(\theta|x)||\pi(\theta))$ is the divergence between $\pi(\theta|x)$ and $\pi(\theta)$ for *all* θ . It is one number which quantifies how different two distributions are. If we then take the expectation of this quantity with respect to the marginal of $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$, $m(\mathbf{x}) = \int_{\Theta} f(\mathbf{x}|\theta)\pi(\theta)$, we get the *expected gain in information* under the prior π , I_k^π , expressed as

$$I_k^\pi = E_{(x)} [D_{KL}(\pi(\theta|x)||\pi(\theta))]. \quad (1.3)$$

Note here that each $\mathbf{x}_i$ is a sample of size n from $f(x|\theta)$, so we would essentially

interpret the expectation in (1.3) as being how much information we gain on average across the k data sets. At this point, we seek the prior π which induces the maximum possible expected gain in information after observation of many data sets, i.e. for some large k ,

$$\sup_{\pi} I_k^{\pi}. \tag{1.4}$$

In particular, Bernardo (1979) considers the reference prior $\pi(\theta)$ to be a function which maximizes this missing information as $k \rightarrow \infty$. Because I_k^{π} is computed as the expected KL divergence over all possible data sets, this may explain why Reference priors tend to perform so well in small sample settings. Now, when we have a one-dimensional parameter vector θ , Berger and Bernardo (1992a) showed that, as $k \rightarrow \infty$, the prior which maximizes the expected gain in information in (1.4) is

$$\pi(\theta) \propto \sqrt{|I_{\theta}|},$$

which is exactly the same as the Jeffreys prior! In fact, under some regularity conditions (Berger et al., 2009), in the one-dimensional case, reference priors are exactly the same as the Jeffreys prior. However, when there are multiple parameters, the reference prior and Jeffreys prior will typically differ, and, unfortunately, the Jeffreys prior typically performs very poorly in multidimensional problems (Berger et al., 2009). These multidimensional problems provided the motivation to formalize a reference prior structure for dealing with “nuisance” parameters. This paved the way for Berger and Bernardo (1992b), which showed that the multiparameter problem can be conducted using a series of one parameter steps.

1.3.2 Reference Priors for Nuisance Parameters

Consider $f(x|\theta_1, \theta_2)$ where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$. Under this parameterization, we may say θ_1 is “of interest” and θ_2 is a “nuisance” parameter. Furthermore, often times the parameter of interest may be a reparameterization of another parameter space, and the nuisance parameters are constructed to facilitate the study of θ_1 . We desire a reference prior $\pi^R(\theta_1, \theta_2)$ for (θ_1, θ_2) , and the reference prior approach to this problem is as follows.

1. Noting that $f(x|\theta_1, \theta_2)$ only depends on θ_2 if we condition on θ_1 , we begin by finding a conditional prior $\pi(\theta_2|\theta_1)$ by choosing *any* marginal prior $\pi^*(\theta_1)$ for θ_1 and finding the $\pi(\theta_2|\theta_1)$ which satisfies

$$\begin{aligned} & \sup_{\pi(\theta_2|\theta_1)} E_{(x, \theta_1)} \left[\int_{\Theta_2} \pi(\theta_2|\theta_1; x) \log \frac{\pi(\theta_2|\theta_1; x)}{\pi(\theta_2|\theta_1)} d\theta_2 \right] \\ &= \sup_{\pi(\theta_2|\theta_1)} \int_{\Theta_1} \int_{\mathcal{X}} f(x|\theta_1) \pi^*(\theta_1) \int_{\Theta_2} \pi(\theta_2|\theta_1; x) \log \frac{\pi(\theta_2|\theta_1; x)}{\pi(\theta_2|\theta_1)} d\theta_2 dx d\theta_1, \end{aligned} \quad (1.5)$$

where

$$f(x|\theta_1) = \int_{\Theta_2} f(x|\theta_1, \theta_2) \pi(\theta_2|\theta_1) d\theta_2$$

is the integrated marginal likelihood for θ_1 . Bernardo (1979) points out that one *could* choose $\pi(\theta_2|\theta_1)$ in a subjective way, contingent on some domain knowledge of the problem at hand, but as shown by Berger and Bernardo (1992a), Sun and Berger (1998), and others, we typically obtain it by finding the conditional reference prior $\pi^R(\theta_2|\theta_1)$ which satisfies 1.5.

2. The prior $\pi^R(\theta_2|\theta_1)$ is the prior for θ_2 given we know θ_1 , and is called the

conditional reference prior, and is used to form the marginal model for the parameter of interest θ_1 as

$$f(x|\theta_1) = \int_{\Theta_2} f(x|\theta_1, \theta_2) \pi^R(\theta_2|\theta_1) d\theta_2.$$

We then find the prior $\pi(\theta)$ which satisfies

$$\sup_{\pi(\theta_1)} \int_{\mathcal{X}} m(x) \int_{\Theta_1} \pi(\theta_1|x) \log \frac{\pi(\theta_1|x)}{\pi(\theta_1)} d\theta_1 dx, \quad (1.6)$$

where

$$m(x) = \int_{\Theta_1} \int_{\Theta_2} f(x|\theta_1, \theta_2) \pi^R(\theta_2|\theta_1) \pi(\theta_1) d\theta_2 d\theta_1,$$

$$\pi(\theta_1|x) = \frac{f(x|\theta_1) \pi(\theta_1)}{m(x)}.$$

We call the prior satisfying (1.6) the *marginal reference prior* and denote it by $\pi^R(\theta)$. It is important to note that the conditional reference prior for $\pi(\theta_2|\theta_1)$ will typically not be proper (Berger and Bernardo, 1992a). Because this function is used to find the marginal likelihood for θ , Berger and Bernardo (1992a) discusses how the conditional reference priors must be truncated on compact supports in order to induce propriety.

1.3.3 Reference Priors for many Nuisance Parameters

The aforementioned approach for conducting reference analysis with a parameter of interest and a nuisance parameter can be extended to models where the number of parameters is greater than two. When we have many nuisance parameters, we

can find reference priors for multidimensional parameter spaces through a stepwise procedure. Generally, as previously mentioned, in this setup a subset of parameters are considered to be “of interest,” and the remaining are considered as “nuisance” parameters. Berger and Bernardo (1992b) showed that if we have a regular model (i.e. replicated data $f(\mathbf{x}|\theta)$ is asymptotically normal) with parameter vector $\theta = (\theta_1, \theta_2, \dots, \theta_m)$ which is ordered with respect to inferential interest, we can derive the joint reference prior for θ using successive corners of the inverse Fisher information matrix I_θ^{-1} . Note that in the overview that follows, it is assumed that the parameters are not grouped, which is called a “one at a time” reference prior. Even though this approach is mathematically more tedious and generally more difficult, it simplifies the exposition in a way relevant to the problems discussed in this work. Despite this, the notation for this methodology is still cumbersome, and the reader is referred to a brief example at the end of this section for a concrete demonstration. For notational convenience, Berger and Bernardo (1992b) first define I_θ^{-1} as

$$I_\theta^{-1} = S = \begin{pmatrix} s_{11} & s_{21} & \dots & s_{m1} \\ s_{21} & s_{22} & \dots & s_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{m1} & s_{m2} & \dots & s_{mm} \end{pmatrix},$$

so that the upper left corners of S can be easily represented as

$$S_q = \begin{pmatrix} s_{11} & s_{21} & \dots & s_{q1} \\ s_{21} & s_{22} & \dots & s_{q2} \\ \vdots & \vdots & \ddots & \vdots \\ s_{q1} & s_{q2} & \dots & s_{qq} \end{pmatrix}.$$

Here, S_q is the $q \times q$ upper left submatrix of S , so that $S_m \equiv S$. Furthermore, they define $H_q \equiv S_q^{-1}$ as the inverse of the upper $j \times j$ corner of S , and finally, the most important quantity

$$h_j \equiv \text{lower right most element of } H_j, \quad j = 1 \dots m$$

will be used to successively obtain the conditional prior distributions. First, noting that $\theta_{[j]} = (\theta_1, \theta_2, \dots, \theta_j)$, we can begin by defining the prior for the last (least interesting w.r.t. inferential interest) parameter θ_m given all other parameters $\theta_{[m-1]}$ as

$$\pi_m^l(\theta_m | \theta_{[m-1]}) = \frac{|h_m(\theta)|^{1/2}}{\int_{\Theta_m^l} |h_m(\theta)|^{1/2} d\theta_m},$$

where Θ_m^l is the (typically compact) support of θ_m . For the remaining parameters $q = m-1, m-2, \dots, 1$ we would iteratively obtain $\pi_q^l(\theta_q | \theta_{[q-1]})$ by

$$\pi_q^l(\theta_q | \theta_{[q-1]}) =$$

$$\frac{\exp \left[\frac{1}{2} \int_{\Theta_{q+1}^l} \cdots \int_{\Theta_m^l} \log |h_q(\theta)| \cdot \pi_m^l(\theta_m | \theta_{[m-1]}) \cdots \pi_{q+1}^l(\theta_{q+1} | \theta_{[q]}) d\theta_m \cdots d\theta_{q+1} \right]}{\int_{\Theta_q^l} \exp \left[\frac{1}{2} \int_{\Theta_{q+1}^l} \cdots \int_{\Theta_m^l} \log |h_q(\theta)| \cdot \pi_m^l(\theta_m | \theta_{[m-1]}) \cdots \pi_{q+1}^l(\theta_{q+1} | \theta_{[q]}) d\theta_m \cdots d\theta_{q+1} \right] d\theta_q} . \quad (1.7)$$

A practical simplification is noted as a result of Lemma 2.1 of Datta and Ghosh (1996). There, they showed that

$$|h_q(\theta)| = \frac{|I_{\theta[\sim q-1, q-1]}|}{|I_{\theta[\sim q, q]}|} \quad (1.8)$$

where $I_{\theta[\sim j, j]}$ is defined as the submatrix with elements

$$((I_{\theta_{ik}}))_{i=j+1, \dots, m, k=j+1 \dots m} .$$

In other words, $I_{\theta[\sim j, j]}$ represents the lower right corner of the Fisher information matrix I_θ *after* the j^{th} row and column. Furthermore, $|I_{\theta[\sim m, m]}|$ is understood to be 1. Obtaining the $|h_q(\theta)|$ through (1.8) typically makes the successive computations of (1.7) substantially easier. After all conditional distributions are obtained, the joint prior of all nuisance parameters can be obtained by

$$\pi(\theta) = \lim_{l \rightarrow \infty} \frac{\pi^l(\theta)}{\pi^l(\theta^*)},$$

where θ^* is any interior point of Θ .

Nuisance parameter Example

Say we have $f(\mathbf{x}|\theta, \eta)$, where θ is the parameter of interest and η is the nuisance parameter. The Fisher information matrix is

$$I_\theta = \begin{pmatrix} I_{\theta,\theta} & I_{\eta,\theta} \\ I_{\eta,\theta} & I_{\eta,\eta} \end{pmatrix}.$$

We would first define $\pi_2^l(\eta|\theta) \propto \sqrt{|I_{\eta,\eta}|}$, but note that this function will usually be improper. To obtain the marginal model and eventually find the reference prior for θ , we normalize on the compact support Θ_2^l to get

$$\pi_2^l(\eta|\theta) = \frac{|I_{\eta,\eta}|^{1/2}}{\int_{\Theta_2^l} |I_{\eta,\eta}|^{1/2} d\eta},$$

which we note is *proper* Θ_2^l . Then we find (1.8) as

$$|h_q(\theta)| = \frac{\left| \begin{pmatrix} I_{\theta,\theta} & I_{\eta,\theta} \\ I_{\eta,\theta} & I_{\eta,\eta} \end{pmatrix} \right|}{|I_{\eta,\eta}|}.$$

This quantity is then applied to (1.7) to get

$$\pi_1^l(\theta) \propto \int_{\Theta_2^l} \log |h_q(\theta)| \pi_2^l(\eta|\theta_1) d\eta.$$

At this point, one would make a limiting argument on the compact supports and find the joint prior $\pi(\theta, \eta)$.

1.3.4 Reparameterizations of the Parameter Space

Additionally, the main theorem of Datta and Ghosh (1996) states that the reference prior for $(\theta_1, \theta_2, \dots, \theta_m)$ is invariant to one-to-one transformations of the form

$$\begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_m \end{pmatrix} \mapsto \begin{pmatrix} g_1(\theta_1) \\ g_2(\theta_1, \theta_2) \\ g_3(\theta_1, \theta_2, \theta_3) \\ \vdots \\ g_m(\theta_1, \theta_2, \theta_3, \dots, \theta_m) \end{pmatrix}. \quad (1.9)$$

This result is similar to that of the Jeffreys prior shown in (1.2), with the additional restriction that the transformation needs to satisfy the triangular structure shown in (1.9). The usefulness of this result should not be understated. Typically, the derivation of reference priors requires considerable effort and can even prove to be mathematically intractable in many cases. However, once we have obtained the reference prior for one parameterization, we immediately have the reference prior for entire classes of reparameterizations of this model.

1.4 Objective Priors for Discrete Parameter Spaces

Sometimes we wish to obtain a prior for the parameter θ when it is discrete. For example, the number of trials n in a Binomial model, the number of populations units R that have a certain property from a Hypergeometric model, or in the case of this work, the number of failures r before stopping an experiment from a negative

binomial model. The previously discussed methods for obtaining objective priors do not apply because of the lack of continuity of θ . Berger et al. (2012) discussed approaching this problem using four different approaches, and Villa and Walker (2014) approached this problem from an entirely new loss-based perspective.

1.4.1 Reference Priors for Discrete Parameters

Developing an objective prior for discrete parameters is a difficult and sometimes inconsistent process. Many priors are developed and studied for different models in an ad hoc and problem-specific manner. Berger et al. (2012) discussed this lack of a general methodology to develop prior distributions for discrete parameter spaces, and presented four approaches which can be used to develop an objective prior for a discrete parameter θ .

Approach 1: Treat the Discrete Parameter as Continuous

The most straightforward approach to the discrete parameter problem is to simply treat θ as a continuous parameter and standard apply objective methods. Berger et al. (2012) mentions that this approach, while convenient, sometimes yields a model with a new normalizing constant as a function of θ , thus changing the underlying structure of the problem from the discrete case. For this reason, this is not recommended approach unless there is no change in the structure of the problem. In Chapter 5, we will see that the Negative Binomial distribution has a natural representation where the discrete parameter is treated as continuous.

Approach 2: Introduce a Continuous Hyperparameter

The second approach outlined in Berger et al. (2012) is hierarchical in nature. The idea is to embed the discrete problem into a continuous one. For example, Jeffreys (1961) discussed that for a *Hypergeometric*(N, R, n) distribution, an unknown number of “success” elements R arises as the binomial distribution $B(R|N, p)$ with an unknown p . The problem can therefore, according to Berger et al. (2012), be reduced to finding an objective prior $\pi(p)$ and computing

$$\pi^*(R) = \int B(R|N, p)\pi(p)dp.$$

This procedure is appealing; however it is mentioned that it is both exceptionally uncommon, and different hierarchical structures can be constructed, leading to different priors for the parameter of interest. In the case of the Negative Binomial distribution, it is unclear if an underlying structure for the number of failures k exists, so this method will not be employed.

Approach 3: Use of a Consistent Estimator

This approach relies on the fact that the development of reference priors is based on considering a sequence of k “imaginary” sets of replications of observations from the data model as $k \rightarrow \infty$. Approach 3 analogously proceeds by choosing a consistent estimator $\hat{\theta}_k$ based on each imaginary random sample $\mathbf{x}^{(k)} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$ which is continuous as $k \rightarrow \infty$. The asymptotic distribution of $\hat{\theta}_k$ is found and the reference prior is computed. The primary issue with this method, as discussed in Berger et al. (2012), is that it requires the use of inefficient estimators because efficient estimators

have no possible continuous embedding. Because the estimators used are inefficient and also not unique, this method suffers from a philosophical perspective. Still, it is recommended that it can be used to suggest an objective prior whose behavior under simulation could be studied.

Approach 4: Using Parameter Based Asymptotics

This approach proceeds by first applying the limiting operation $\theta \rightarrow \infty$ to the discrete parameter θ to obtain the limiting asymptotic distribution of the density under consideration. Then consider θ as continuous and apply reference prior methods to the asymptotic distribution. The main drawback of this approach is that it yields a prior for large values of θ , and the performance of the prior for small values of θ may not be satisfactory.

1.4.2 A Loss Based Approach

Villa and Walker (2014) proposed a novel approach for computing a prior function $\pi(\theta)$ based on the *worth* of each element θ in the discrete parameter space. This idea assigns value to each element in the parameter space through the self-information loss $-\log \pi(\theta)$ (Merhav and Feder, 1998) rather than through probability directly. Essentially, they examined the loss in information if the true value θ_0 is removed from the parameter space.

The concept of self-information quantifies the relationship between uncertainty and information for an event e as

$$I(e) = \log \left(\frac{1}{P(e)} \right) = -\log P(e).$$

This means that the more likely an event, the less information it carries, should it occur. This means that events with probability one contain no information and that independent events have joint information equal to the sum of their individual information values. Following Merhav and Feder (1998), loss is then assigned to each element of the parameter space by

$$l(\theta) = -\log \pi(\theta)$$

for the prior $\pi = \{\pi(\theta), \theta \in \Theta\}$. The loss $l(\theta)$ for each $\theta \in \Theta$ is assigned by linking this loss function to the “worth” of the elements. Note that according to (Berk 1966), if the true value θ_0 is removed from the parameter space, the posterior distribution asymptotically accumulates to the value θ' for which $f(\cdot|\theta')$ is closest to $f(\cdot|\theta_0)$ with respect to Kullback-Leibler divergence. Because of this, the worth was defined as a numerical measure of what is lost when removing a particular θ from Θ should it be the true value; i.e. how “far away” would our best model be if that model were misspecified? This means that the *loss* of each θ would be

$$l(\theta) = -D_{KL}(f(\cdot|\theta)||f(\cdot|\theta')).$$

$l(\theta)$ is then linked to the self-information loss so that the prior mass for each θ is an exercise in finding the θ' such that the Kullback-Leibler divergence between the model *with* θ to the model *without* θ is minimized. The Villa-Walker prior is given by

$$\pi(\theta) \propto \exp \left\{ \min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta)||f(\cdot|\theta')) \right\} - 1.$$

In this formulation, 1 is subtracted because it has the desirable property of assigning null mass to elements with null worth. In summary, to obtain the Villa-Walker prior, one must first find the θ' that minimizes the aforementioned Kullback-Leibler divergence and use the above formula to compute the objective prior.

1.5 Outline

In Chapter 2, we introduce the background of the 2×2 contingency table, sampling schemes, and a review of current techniques for conducting data analysis. In Chapter 3, we present a reference prior for the odds ratio of an unrestricted 2×2 table along with several constructed but interesting nuisance parameters. As we will see, the derived reference priors have many useful, intuitive properties and many sub-steps infused into the various proofs have a very elegant construction. In Chapter 4, we present the posterior distributions and an efficient sampling algorithm. A robust simulation study is also presented, the results of which indicate that the proposed approach to this problem is far superior to the straightforward and widely used Frequentist approaches that dominate this area, as well as other proposed Bayesian approaches. Additionally, real data examples also typically yield much more sensible results, especially for small sample sizes or for tables that contain zeros. The accessible nature of the 2×2 tables give us the ability to evaluate our methodology on a wide variety of interesting datasets, such as the Fisher's tea tasting example. What's more, the practical computation of the methodology in this chapter is implemented in a robust software package which is free to download and use by anyone. The power of using Graphics Processing Units(GPUs) towards solving statistical problems is also

demonstrated, with this approach increasing the speed of computation by 200x when compared to C and over 10,000x when compared to R.

In Chapter 5 we focus on two cases of the Negative Binomial distribution. First, a parameterization useful for capturing overdispersion is considered. For this model, because the Fisher information matrix is intractable to work with analytically, an approximation to the reference prior is developed. This approximation yields a functional form, which is then studied under simulation, yields much better performance than the frequentist MLE in estimating the dispersion parameter, and comparable performance estimating the mean parameter. Even though the methodology developed here is an approximation to the reference prior, the properties of this prior are good. This spurs a discussion about the merits of approximate methods in practice and the potential for this approach to be used in the “black box” of statistics. Next, we use the discrete approach for obtaining an objective Bayesian prior outlined in 1.4.2 to develop a joint prior for the negative binomial distribution using the standard discrete parameterization. Here, the prior and posterior distribution are successfully obtained and studied. Specifically, they are compared to the results of the MLE when applied to both real data as well as under simulation.

Finally, in Chapter 6, we conclude with a discussion on the objective Bayesian philosophy and its implications for the future of both practical and theoretical statistics, as well as a discussion on the importance of computational efficiency for the modern day statistician.

Chapter 2

Objective Bayesian Analysis of the 2x2 Contingency Table

2.1 Introduction

In clinical trials and other studies, it is common to evaluate whether a treatment yields a positive response relative to a control. Each individual in the study will have either a positive or negative response, and they can further be cross-classified into their respective treatment group. This cross classified data can be arranged into a 2×2 contingency table of counts such as in Table 2.1. This table is a matrix that

Table 2.1: a 2×2 table

		Outcome		
		-	+	
Group	Control	n_1	n_2	$n_1 + n_2$
	Treatment	n_3	n_4	$n_3 + n_4$
		$n_1 + n_3$	$n_2 + n_4$	n

essentially displays the frequency distribution of the variables, and is commonly used when conducting surveys and clinical trials. In Table 2.1, n_i represents the individual cell counts, and $n = n_1 + n_2 + n_3 + n_4$ the overall total for the table. The sampling structure present in the data collection process will determine the type of model used, with different sampling decisions having profound implications for the resulting model structure. There are four common ways of describing realizations of the cell counts n_i , each of them nested within the former by increasingly restrictive conditioning.

2.1.1 Sampling Schemes

The first type of sampling is called *unrestricted* sampling. The resulting model structure for this design requires *no* conditioning on either the marginal totals or overall table total. Because of this lack of conditioning, this is the most general possible design for a 2×2 table and therefore does not need the researcher to specify sampling totals prior to the study. This is a liberating structure, as often times researchers use methodologies developed for restricted sampling even if their sampling was essentially conducted independent of this condition (Agresti, 2013).

Unrestricted Sampling

When we view the counts n_i as independent variables, this is equivalent to sampling randomly with no stopping criteria for the total n . In this case, each n_i is modeled according to a Poisson distribution with mean λ_i as

$$f(n_i|\lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{n_i}}{n_i!}, \quad (2.1)$$

which leads to the joint distribution

$$\ell(\mathbf{n}|\lambda) = \prod_{i=1}^4 \frac{e^{-\lambda_i} \lambda_i^{n_i}}{n_i!}. \quad (2.2)$$

To date, there do not seem to be any Bayesian treatments of this data structure, and this will be the main focus of our work.

Fixed Sample Size n

Consider conditioning the previous structure on the total number of observations n but allowing the marginal totals to remain random. This is equivalent to considering each cell count as a Poisson random variable but conditioning on the total sample size n , which is also a Poisson with mean $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$.

$$\begin{aligned} f(n_1, n_2, n_3|n) &= \frac{P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N = n)}{P(N = n)} \\ &= \frac{P(N_1 = n_1, N_2 = n_2, N_3 = n_3, N_4 = n_4)}{P(N = n)} \\ &= \frac{\frac{e^{-\lambda_1} \lambda_1^{n_1}}{n_1!} \frac{e^{-\lambda_2} \lambda_2^{n_2}}{n_2!} \frac{e^{-\lambda_3} \lambda_3^{n_3}}{n_3!} \frac{e^{-\lambda_4} \lambda_4^{n_4}}{n_4!}}{\frac{e^{-\lambda} \lambda^n}{n!}} \\ &= \frac{n!}{n_1! n_2! n_3! n_4!} \left(\frac{\lambda_1}{\lambda} \right)^{n_1} \left(\frac{\lambda_2}{\lambda} \right)^{n_2} \left(\frac{\lambda_3}{\lambda} \right)^{n_3} \left(\frac{\lambda_4}{\lambda} \right)^{n_4}. \end{aligned} \quad (2.3)$$

So that

$$N_1, N_2, N_3, N_4 | N = n \sim \text{Multinomial} \left(n, \frac{\lambda_1}{\lambda}, \frac{\lambda_2}{\lambda}, \frac{\lambda_3}{\lambda}, \frac{\lambda_4}{\lambda} \right).$$

Because of this, data collection may be restricted to examine a predetermined number of subjects due to budgetary considerations or researchers may take a random

subsample of data. In this fixed total sample size setting, the multinomial distribution is used. Conducting Bayesian analysis on functions of the parameters under the multinomial model is also a topic that has not been explored in the literature, and will be a topic of future work for us.

Fixed Marginal Totals

Another way of conditioning is to take the Poisson counts n_i and condition on knowing both totals of one margin. Each of the marginal totals $n_1 + n_2$ and $n_3 + n_4$ are independently Poisson with means $\lambda_1 + \lambda_2$ and $\lambda_3 + \lambda_4$. This gives us

$$\begin{aligned}
f(n_2, n_4 | n_1 + n_2, n_3 + n_4) &= \frac{P(N_2 = n_2, N_4 = n_4, N_{0.} = n_1 + n_2, N_{1.} = n_3 + n_4)}{P(N_{0.} = n_1 + n_2, N_{1.} = n_3 + n_4)} \\
&= \frac{P(N_2 = n_2, N_4 = n_4, N_1 = n_1, N_3 = n_3)}{P(N_{0.} = n_1 + n_2, N_{1.} = n_3 + n_4)} \\
&= \frac{\frac{e^{-\lambda_2} \lambda_2^{n_2}}{n_2!} \frac{e^{-\lambda_4} \lambda_4^{n_4}}{n_4!} \frac{e^{-\lambda_1} \lambda_1^{n_1}}{n_1!} \frac{e^{-\lambda_3} \lambda_3^{n_3}}{n_3!}}{\frac{e^{-\lambda_1 + \lambda_2} (\lambda_1 + \lambda_2)^{n_1 + n_2}}{(n_1 + n_2)!} \frac{e^{-\lambda_3 + \lambda_4} (\lambda_3 + \lambda_4)^{n_3 + n_4}}{(n_3 + n_4)!}} \\
&= \frac{n_1 + n_2!}{n_2! (n_1 + n_2 - n_2)!} \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n_2} \left(1 - \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n_1 + n_2 - n_2} \\
&\times \frac{n_3 + n_4!}{n_4! (n_3 + n_4 - n_4)!} \left(\frac{\lambda_4}{\lambda_3 + \lambda_4} \right)^{n_4} \left(1 - \frac{\lambda_4}{\lambda_3 + \lambda_4} \right)^{n_3 + n_4 - n_4} \\
&\equiv \text{Bin} \left(n_1 + n_2, \frac{\lambda_2}{\lambda_1 + \lambda_2} \right) \times \text{Bin} \left(n_3 + n_4, \frac{\lambda_4}{\lambda_3 + \lambda_4} \right).
\end{aligned} \tag{2.4}$$

The practical implication of this method of conditioning is that data may be collected on a predetermined number of individuals from each category of one of the variables and classified them according to the other variable (stratified sampling). This is extremely useful if one category is much more rare than the other and we would like

to ensure that one category is equally represented. Under this structure, a reference prior for the log odds ratio was obtained and studied by Sun and Berger (1998).

Both Marginal Totals are Fixed

The fourth rather uncommon sampling setup arises when the marginal totals of both rows and columns are fixed. In a 2x2 contingency table, when both marginals are fixed, knowing the value of one element of the table determines the values of all other elements. Therefore, the distribution of any cell n_i in the table can be expressed solely in terms of any one element according to the hypergeometric distribution (Agresti, 2013). This sampling setup has been immortalized in R.A. Fisher's "Lady Tasting Tea" experiment (Fisher, 1971). The experiment provided a lady (Dr. Muriel Bristol) with 8 randomly ordered cups of tea. 4 of the cups prepared by adding milk before the tea and the other 4 were prepared by adding the tea before the milk. She was to select the 4 cups prepared by one method. This offered her the advantage of judging cups by comparison as she was informed of the setup of the experiment. Under the independence assumption, we have

$$f(n_1) = P(N_1 = n_1) = \frac{\binom{n_1+n_2}{n_1} \binom{n_3+n_4}{n_1+n_3-n_1}}{\binom{n}{n_1+n_3}}, \quad (2.5)$$

where the support of n_1 is $n_1 \in [\max(0, n_1 + n_2 + n_1 + n_3 - n), \min(n_1 + n_2, n_1 + n_3)]$. This formulation was then generalized to Fisher's noncentral hypergeometric distribution, which is expressed as

$$f(n_4|\omega) = \frac{\binom{n_3+n_4}{n_4} \binom{n_2}{n_2+n_4-n_4} \omega^x}{\sum_{y=x_{min}}^{x_{max}} \binom{n_3+n_4}{y} \binom{n_2}{n_2+n_4-y} \omega^y}. \quad (2.6)$$

Here, ω represents the noncentrality parameter as well as the odds ratio in the contingency table. Furthermore, it is clear to see that if $\omega = 1$, (2.6) reduces to (2.5). The frequentist properties of this distribution were discussed extensively in Kou and Ying (1996).

Once data are collected under one of these schemes, it is generally of interest to conduct analysis on some measure of association between treatment and outcome such as the increased risk, relative risk, or odds or log-odds ratio. The odds ratio is one of the most common parameters used to assess the relationship in a 2x2 contingency table, it quantifies how strongly the presence or absence of one property is associated with the presence or absence of another property. In the unrestricted 2×2 table, it is represented by the cross product of the elements $\theta = \frac{\lambda_1 \lambda_4}{\lambda_2 \lambda_3}$. We note that without loss of generality, if a methodology is available to estimate θ in the unrestricted case, we can apply it to the other cases as well because those other sampling schemes are special cases of unrestricted sampling.

Lindley (1964) provides a Bayesian framework for estimating the odds ratio and log odds ratio with Beta and Dirichlet priors for both the Binomial and Multinomial sampling schemes. Sun and Berger (1998) constructed a reference prior for the log-odds ratio under binomial sampling assuming independence of this quantity and the log odds of one group. Erdogan and Dickey (1974) presented a unified analysis to obtain Bayes factors for association using all 4 models and applying a conjugate structure. Agresti and Hitchcock (2005) provides an excellent overview of Bayesian techniques for categorical data analysis. Bayesian approaches to contingency tables generally take advantage of the mathematical convenience of a conjugate prior

A fundamental issue with the analysis of 2×2 data is the possibility of having

very low or even 0 values for cell counts. As pointed out by Ware (1989), sometimes cutting edge treatments for rare and serious diseases may offer profoundly positive benefits to patients. In these cases, we may have a very small amount of data. In particular, if a zero count for any cell value exists, the frequentist estimator of the odds ratio, $\hat{\theta} = \frac{n_1 n_4}{n_2 n_3}$, is either 0 or ∞ . The earliest approaches to treating this issue were treated in Haldane (1955) and Gart and Zweifel (1967), who suggested that in the presence of a zero cell count, one may simply add a constant c to each cell. They also demonstrated that $c = 0.5$ yielded the smallest first-order finite sample bias for any constant c , achieving a bias of $\mathcal{O}(n^{-2})$ for large samples when estimating $\log \theta$. This may be why this is a common and easily implemented “fix” one can use to solve this problem. Agresti and Yang (1987) criticized this approach from a philosophical standpoint, arguing that adding a constant was essentially adding “fake data” to the table. However, Agresti (1999) showed through simulation that $c = .5$ yields coverage close to nominal in small sample settings when the true odds ratio is less than 4. More recently, Parzen et al. (2002) presented a nonparametric technique to obtain a median unbiased estimate of θ under an independent binomial setup by fully enumerating the entire bootstrap distribution of the marginal probability of success in the two groups. Agresti and Hitchcock (2005) and references therein presented a summary of many Bayesian techniques for this problem, and most of them consider a multinomial or binomial sampling structure and use subjective beta priors on the proportion of success in each group. Each of the discussed methods place some restrictions on the table in the form of fixing the overall total or the totals of one margin.

2.2 Objective Bayesian Analysis for Unrestricted Poisson Sampling

In this work, we focus on deriving the reference prior for the odds ratio θ under the unrestricted Poisson representation of the 2×2 table. When we view the counts n_i as independent variables, this is equivalent to sampling randomly with no stopping criteria for the total n . In this case, the joint distribution for $\mathbf{n} = (n_1, n_2, n_3, n_4)$ is

$$\ell(\mathbf{n}|\lambda) = \prod_{j=1}^4 \frac{e^{-\lambda_j} \lambda_j^{n_j}}{n_j!}, \quad (2.7)$$

and the fisher information matrix is given by $I_\lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$.

The remainder of this chapter is organized as follows. In section 3.1, we present the core of our work, including the reference prior, reference posterior, and sampling strategies, including a description of the computational efficiency of the developed estimation technology. Section 4.3 presents a robust simulation study comparing the proposed reference structure to established frequentist approaches as well as other Bayesian methods. Finally, section 4.3.2 applies the methodology to several datasets and compares the results with frequentist and Bayesian procedures.

The Fisher information matrix of the joint Poisson distribution in (2.7) is

$$I_\lambda = \text{diag} \left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \frac{1}{\lambda_3}, \frac{1}{\lambda_4} \right),$$

Assuming independence of λ_i , and by theorem 4 of Sun and Berger (1998) the

independent reference prior, and also the Jeffreys rule prior is given by

$$\pi^r(\lambda) \propto \prod_j \lambda_j^{-1/2}, \quad (2.8)$$

which leads to the joint posterior

$$\pi^r(\lambda|\mathbf{n}) \propto \prod_j \text{Gamma}(n_j + 1/2, 1). \quad (2.9)$$

It is interesting to note that with $\lambda_j|n_j \sim \text{Gamma}(n_j + 1/2, 1)$, if $n_1 = n_2 = n_3 = n_4 = 0$, each marginal posterior is proper. This means that we could apply Jeffreys prior to obtain posterior for $\lambda_j|n_j$, as well as any 1-to-1 function of the λ_j , even with no data! This is unreasonable, and as we will see, the reference prior approach will yield much more sensible conditions.

Recall that with the 2×2 table, we will typically desire to make inference on *functions* of the parameters rather than the λ_j parameters themselves. As somewhat of a toy example, the first quantity we will consider is the *relative risk* of the table.

2.2.1 Relative Risk

The relative risk (also called “risk ratio”) of a table is defined as the probability of success in the treatment group divided by the probability of success in the control group. This quantity gives us an idea about how the risks compare between the two groups. Since we consider sampling as unrestricted, we can consider the proportions

of "success" in the row variables as

$$p_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}.$$

The relative risk is therefore defined as

$$\omega = \frac{p_4}{p_2} = \frac{\lambda_4}{\lambda_2}.$$

Say we have $\theta = (\omega, \psi) = \left(\frac{\lambda_4}{\lambda_2}, \lambda_4 \lambda_2\right)$ where ω is the relative risk and ψ is the nuisance parameter. We proceed as with the example in 1.3.3 by first noting that the fisher information of this parameterization is

$$I_\theta = \frac{1}{4} \begin{bmatrix} \psi^{1/2} \frac{\omega+1}{\omega^{5/2}} & \psi^{-1/2} \frac{\omega-1}{\omega^{3/2}} \\ \psi^{-1/2} \frac{\omega-1}{\omega^{3/2}} & \psi^{-3/2} \frac{\omega+1}{\omega^{1/2}} \end{bmatrix}.$$

So $\pi(\psi|\omega) \propto \psi^{-3/4}$, $\psi \in [a_k, b_k]$. Then we note that

$$\frac{|I_\theta|}{I_{\theta[2,2]}} = \frac{\omega^{-2} \psi^{-1}}{\psi^{-3/2} \frac{\omega+1}{\omega^{1/2}}} = \frac{\psi^{1/2}}{\omega^{3/2} (1 + \omega)},$$

to obtain

$$\pi(\omega) \propto \exp \left\{ \frac{1}{2} \int_{a_k}^{b_k} \log \left[\frac{\psi^{1/2}}{\omega^{3/2} (1 + \omega)} \right] \pi(\psi|\omega) d\psi \right\} = \omega^{-3/4} (1 + \omega)^{-1/2}.$$

which is a *beta-prime* distribution with $\alpha = \beta = \frac{1}{4}$.

Chapter 3

A Reference Prior for the Odds Ratio of the 2x2 Contingency Table

3.1 Parameterization

The odds ratio is the most common parameter used to assess the relationship in a 2×2 contingency table, it quantifies how strongly the presence or absence of one property is associated with the presence or absence of another property. In the unrestricted 2×2 table, it is represented by the cross product of the elements $\theta = \frac{\lambda_1 \lambda_4}{\lambda_2 \lambda_3}$. The procedure in section 2.2.1 was rather straightforward because the relative risk is a function of only two parameters, hence one nuisance parameter needed to be constructed. Because the odds ratio uses all 4 λ_i , we must construct a set of 3 nuisance parameters to ensure that we have a four dimensional 1-to-1 transformation which θ belongs to. Consider

the joint transformation

$$\theta = \frac{\lambda_1 \lambda_4}{\lambda_2 \lambda_3}, \quad \eta_1 = \frac{\lambda_1}{\lambda_1 + \lambda_3}, \quad \eta_2 = \lambda_1 + \lambda_3, \quad \eta_3 = \frac{\lambda_2}{\lambda_1}, \quad (3.1)$$

where $\eta_1 \in (0, 1)$, $\eta_2 > 0$, and $\eta_3 > 0$ are nuisance parameters. Note that (3.1) is equivalent to

$$\lambda_1 = \eta_1 \eta_2, \quad \lambda_2 = \eta_1 \eta_2 \eta_3, \quad \lambda_3 = \eta_2 (1 - \eta_1), \quad \lambda_4 = \theta \eta_2 \eta_3 (1 - \eta_1). \quad (3.2)$$

Generally the choice of nuisance parameters in these problems is a key concern, as the iterative nature of the integrations presented in (1.7) make the process difficult in practice. The three η_i in (3.1) were initially chosen by trial and error, and, using the computer algebra system MAPLE 2017, the Fisher matrix was obtained in order to subjectively evaluate the mathematical feasibility of the inputted parameterization. If a parameterization seemed tractable, considerable effort was spent attempting to derive the reference prior. Much time was spent attempting to find a parameterization that “worked” mathematically, and finally we obtained one with remarkable interpretations and properties. We first note that despite being “nuisance” parameters, the three η_i have highly informative interpretations. η_1 represents the proportion of negative/positive outcomes in the control/treatment group. η_2 represents the total average number of negative/positive outcomes. We can further note from (3.2) that all four λ_i tend to grow with η_2 when θ , η_1 , and η_3 remain fixed. Finally, η_3 represents the relative risk of a positive outcome in the control group. As the first step in obtaining the fisher information matrix for the

parameterization in (3.1). The Jacobian of (3.1) is

$$J = \begin{bmatrix} 0 & \eta_2 & \eta_1 & 0 \\ 0 & -\eta_2 & -\eta_1 + 1 & 0 \\ 0 & \eta_3 \eta_2 & \eta_3 \eta_1 & \eta_1 \eta_2 \\ -\eta_3 \eta_1 \eta_2 + \eta_3 \eta_2 & -\theta \eta_2 \eta_3 & -\theta \eta_1 \eta_3 + \theta \eta_3 & -\theta \eta_1 \eta_2 + \theta \eta_2 \end{bmatrix}. \quad (3.3)$$

Using (2.7) leads to the transformed likelihood

$$\begin{aligned} f(n|\theta, \eta_1, \eta_2, \eta_3) &= \frac{[\eta_1 \eta_2]^{n_1} e^{-\eta_1 \eta_2}}{n_1!} \frac{[\eta_1 \eta_2 \eta_3]^{n_2} e^{-\eta_1 \eta_2 \eta_3}}{n_2!} \frac{[\eta_2(1 - \eta_1)]^{n_3} e^{-\eta_2(1 - \eta_1)}}{n_3!} \times \\ &\times \frac{[\theta \eta_2 \eta_3(1 - \eta_1)]^{n_4} e^{-\theta \eta_2 \eta_3(1 - \eta_1)}}{n_4!} \eta_1 (1 - \eta_1) \eta_2^3 \eta_3 \\ &= \frac{1}{n_1! n_2! n_3! n_4!} \eta_1^{n_1 + n_2} (1 - \eta_1)^{n_3 + n_4} \eta_2^{n_1 + n_2 + n_3 + n_4} \eta_3^{n_2 + n_4} \theta^{n_4} \times \\ &\times \exp \{ -\eta_2 [\eta_1(1 + \eta_3) + (1 - \eta_1)(1 + \theta \eta_3)] \}. \end{aligned} \quad (3.4)$$

To obtain the fisher information matrix of this model we note that

$$\begin{aligned} I_{(\theta, \eta_1, \eta_2, \eta_3)} &= J' I_{(\lambda_1, \lambda_2, \lambda_3, \lambda_4)} J = J' \text{diag} \left(\frac{1}{\eta_1 \eta_2}, \frac{1}{\eta_2(1 - \eta_1)}, \frac{1}{\eta_1 \eta_2 \eta_3}, \frac{1}{\theta \eta_2 \eta_3(1 - \eta_1)} \right) J \\ &= \begin{bmatrix} \frac{\eta_3 \eta_2(1 - \eta_1)}{\theta} & -\eta_3 \eta_2 & \eta_3(1 - \eta_1) & \eta_2(1 - \eta_1) \\ -\eta_3 \eta_2 & \frac{\eta_2(1 + \eta_3(1 - \eta_1) + \theta \eta_1 \eta_3)}{\eta_1(1 - \eta_1)} & (1 - \theta) \eta_3 & (1 - \theta) \eta_2 \\ \eta_3(1 - \eta_1) & (1 - \theta) \eta_3 & \frac{1 + \eta_1 \eta_3 + \theta \eta_3(1 - \eta_1)}{\eta_2} & \theta(1 - \eta_1) + \eta_1 \\ \eta_2(1 - \eta_1) & (1 - \theta) \eta_2 & \theta(1 - \eta_1) + \eta_1 & \frac{\eta_2(\theta(1 - \eta_1) + \eta_1)}{\eta_3} \end{bmatrix}. \end{aligned}$$

Examining the determinants of the submatrices of $I_{(\theta, \eta_1, \eta_2, \eta_3)}$ we have

$$\begin{aligned} |I_4^*| &= |I_{(\theta, \eta_1, \eta_2, \eta_3)}| = \frac{\eta_2^2}{\theta}, \\ |I_3^*| &= |I_{(\eta_1, \eta_2, \eta_3)}| = \frac{\eta_2 (\theta \eta_1 - \theta \eta_3 - \theta - \eta_1)}{\eta_1 (\eta_1 - 1) \eta_3}, \\ |I_2^*| &= |I_{(\eta_2, \eta_3)}| = \frac{\theta(1 - \eta_1) + \eta_1}{\eta_3}, \text{ and} \\ |I_1^*| &= |I_{\eta_3}| = \frac{(\theta(1 - \eta_1) + \eta_1) \eta_2}{\eta_3}. \end{aligned}$$

These quantities will be used to successively derive the conditional reference priors for the nuisance parameters η_3 , η_2 , and η_1 following the procedure in Datta and Ghosh (1996). Now, we present the core of our work, the joint reference prior for the parameterization in (3.1).

Theorem 3.1.1. Consider the functions

$$H(\theta) = \int_0^1 \frac{ds}{\sqrt{s(1-s)(\theta(1-s)+s)}} \quad \text{and} \quad G(\theta) = \int_0^1 \frac{\log \sqrt{t(1-t)} dt}{\sqrt{t(1-t)(\theta(1-t)+t)}}. \quad (3.5)$$

The one at a time reference prior for $\{\theta, \eta_1, \eta_2, \eta_3\}$ is

$$\pi^R(\theta, \eta_1, \eta_2, \eta_3) \propto \pi^R(\theta, \eta_1) \pi^R(\eta_2) \pi^R(\eta_3), \quad (3.6)$$

where

$$\pi^R(\theta, \eta_1) \propto \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} \frac{[H(\theta)]^{-1}}{\sqrt{\eta_1(1-\eta_1)} \sqrt{\theta(1-\eta_1) + \eta_1}}, \quad \theta > 0, \eta_1 \in (0, 1), \quad (3.7)$$

$$\pi^R(\eta_2) \propto \frac{1}{\sqrt{\eta_2}}, \quad \eta_2 > 0, \quad (3.8)$$

$$\pi^R(\eta_3) \propto \frac{1}{\sqrt{\eta_3}}, \quad \eta_3 > 0. \quad (3.9)$$

3.2 Proof of Theorem 3.1.1

3.2.1 Derivation of $\pi_{3k}(\eta_3|\eta_2, \eta_1, \theta)$ and $\pi_{2k}(\eta_2|\eta_1, \theta)$

First, note that $|I_1^*| = \frac{g(\theta, \eta_1, \eta_3)}{\eta_3}$ so that

$$\pi(\eta_3|\eta_2, \eta_1, \theta) \propto \sqrt{|I_1^*|} \propto \frac{1}{\sqrt{\eta_3}}.$$

With the compact support argument, the conditional prior density of η_3 given (θ, η_1, η_2) is

$$\pi_{3k}(\eta_3|\eta_2, \eta_1, \theta) = \frac{1}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \frac{1}{\sqrt{\eta_3}}, \quad \eta_3 \in [a_{3k}, b_{3k}]. \quad (3.10)$$

Next, note that $\frac{|I_2^*|}{|I_1^*|} = \frac{1}{\eta_2}$ to obtain

$$\begin{aligned} \pi_{2k}(\eta_2|\eta_1, \theta) &\propto \exp \left\{ \frac{1}{2} \int_{a_{3k}}^{b_{3k}} \log \left[\frac{|I_2^*|}{|I_1^*|} \right] \pi_{3k}(\eta_3|\eta_2, \eta_1, \theta) d\eta_3 \right\} \\ &\propto \exp \left\{ \frac{1}{2} \int_{a_{3k}}^{b_{3k}} \log \left[\frac{1}{\eta_2} \right] \frac{1}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \frac{1}{\sqrt{\eta_3}} d\eta_3 \right\} \\ &\propto \frac{1}{\sqrt{\eta_2}}. \end{aligned}$$

We similarly define η_2 to be defined on its compact support $[a_{2k}, b_{2k}]$ to obtain the conditional prior density of η_2 given (θ, η_1) as

$$\pi_{2k}(\eta_2|\eta_1, \theta) = \frac{1}{2(\sqrt{b_{2k}} - \sqrt{a_{2k}})} \frac{1}{\sqrt{\eta_2}}, \quad \eta_2 \in [a_{2k}, b_{2k}]. \quad (3.11)$$

3.2.2 Derivation of $\pi_1(\eta_1|\theta)$

Next, note that

$$\frac{|I_3^*|}{|I_2^*|} = \frac{\eta_2 ([\theta(1 - \eta_1) + \eta_1] + \theta\eta_3)}{\eta_1 (1 - \eta_1) (\theta(1 - \eta_1) + \eta_1)}$$

to obtain

$$\begin{aligned} \pi_1(\eta_1|\theta) &\propto \exp \left\{ \frac{1}{2} \int_{a_{2k}}^{b_{2k}} \int_{a_{3k}}^{b_{3k}} \log \left(\frac{|I_3^*|}{|I_2^*|} \right) \pi_{3k}(\eta_3|\eta_2, \eta_1, \theta) \pi_{2k}(\eta_2|\eta_1, \theta) d\eta_3 d\eta_2 \right\} \\ &= \exp \left\{ \frac{1}{2} \int_{a_{2k}}^{b_{2k}} \int_{a_{3k}}^{b_{3k}} \log \left[\frac{\eta_2 \{[\theta(1 - \eta_1) + \eta_1] + \theta\eta_3\}}{\eta_1 (1 - \eta_1) (\theta(1 - \eta_1) + \eta_1)} \right] \right. \\ &\quad \times \left. \frac{1}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \frac{1}{\sqrt{\eta_3}} \frac{1}{2(\sqrt{b_{2k}} - \sqrt{a_{2k}})} \frac{1}{\sqrt{\eta_2}} d\eta_3 d\eta_2 \right\} \\ &\propto \exp \left\{ \frac{1}{2} \int_{a_{3k}}^{b_{3k}} \log \left[\frac{[\theta(1 - \eta_1) + \eta_1] + \theta\eta_3}{\eta_1 (1 - \eta_1) (\theta(1 - \eta_1) + \eta_1)} \right] \frac{1}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \frac{1}{\sqrt{\eta_3}} d\eta_3 \right\} \\ &= \frac{1}{\sqrt{\eta_1(1 - \eta_1)}} \exp \left\{ \frac{1}{4} \frac{1}{\sqrt{b_{3k}} - \sqrt{a_{3k}}} \int_{a_{3k}}^{b_{3k}} \log \left[1 + \frac{\theta}{\theta(1 - \eta_1) + \eta_1} \eta_3 \right] \frac{1}{\sqrt{\eta_3}} d\eta_3 \right\}. \end{aligned} \quad (3.12)$$

If we let $\beta = \frac{\theta}{\theta(1 - \eta_1) + \eta_1}$, the integrand in (3.12) can be represented as

$$\int_{a_{3k}}^{b_{3k}} \log(1 + \beta\eta_3) \frac{1}{\sqrt{\eta_3}} d\eta_3 = 2\eta_3^{\frac{1}{2}} \log(1 + \beta\eta_3) \Big|_{a_{3k}}^{b_{3k}} - 2\beta \int_{a_{3k}}^{b_{3k}} \frac{\eta_3^{\frac{1}{2}}}{1 + \beta\eta_3} d\eta_3. \quad (3.13)$$

For the second term of (3.13), consider the substitution $s = \sqrt{\beta\eta_3}$ to see

$$2\beta \int_{a_{3k}}^{b_{3k}} \frac{\eta_3^{\frac{1}{2}}}{1 + \beta\eta_3} d\eta_3 = \frac{4}{\sqrt{\beta}} \int_{a_{3k}}^{b_{3k}} \frac{s^2}{1 + s^2} ds = \frac{4}{\sqrt{\beta}} \left[\sqrt{\beta\eta_3} - \arctan(\sqrt{\beta\eta_3}) \right] \Big|_{a_{3k}}^{b_{3k}}.$$

so that

$$\begin{aligned} \pi_{1k}(\eta_1|\theta) &\propto \frac{1}{\sqrt{\eta_1(1-\eta_1)}} \exp \left\{ \frac{\eta_3^{\frac{1}{2}} \log[1 + \beta\eta_3] \Big|_{a_{3k}}^{b_{3k}} - 2 \left[\sqrt{\eta_3} - \frac{\arctan(\sqrt{\beta\eta_3})}{\sqrt{\beta}} \right] \Big|_{a_{3k}}^{b_{3k}}}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \right\} \\ &\propto \frac{1}{\sqrt{\eta_1(1-\eta_1)}} \exp \left\{ \frac{\sqrt{b_{3k}} \log(1 + \beta \cdot b_{3k}) - \sqrt{a_{3k}} \log(1 + \beta \cdot a_{3k})}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} + \right. \\ &\quad \left. + \frac{[\arctan(\sqrt{\beta \cdot b_{3k}}) - \arctan(\sqrt{\beta \cdot a_{3k}})]}{\sqrt{\beta}(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \right\}. \end{aligned} \quad (3.14)$$

Note that $\arctan(x)$ is bounded as $x \rightarrow 0$ and $x \rightarrow \infty$. So, the terms of (3.14) behave the following way as $a_{3k} \rightarrow 0$ and $b_{3k} \rightarrow \infty$:

$$\begin{aligned} \frac{\sqrt{b_{3k}} \log(1 + \beta \cdot b_{3k})}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} &\approx \frac{1}{2} \log(\beta \cdot b_{3k}), \\ \frac{\sqrt{a_{3k}} \log(1 + \beta \cdot a_{3k})}{2(\sqrt{b_{3k}} - \sqrt{a_{3k}})} &\approx \frac{1}{2} \log(1) \text{ and} \\ \frac{2[\arctan(\sqrt{\beta \cdot b_{3k}}) - \arctan(\sqrt{\beta \cdot a_{3k}})]}{\sqrt{\beta}(\sqrt{b_{3k}} - \sqrt{a_{3k}})} &\approx \frac{c}{\sqrt{a}\sqrt{b_{3k}}} \end{aligned}$$

for some constant $c > 0$. Therefore the leading term inside the exponential of (3.14) is $\frac{1}{2} \log[\beta \cdot b_{3k}]$, so we finally arrive at

$$\begin{aligned} \pi_1(\eta_1|\theta) &= \lim_{k \rightarrow \infty} \pi_{1k}(\eta_1|\theta) \\ &\propto \frac{1}{\sqrt{\eta_1(1-\eta_1)}} \frac{1}{\sqrt{\theta(1-\eta_1) + \eta_1}}. \end{aligned}$$

Now, in this case, we can define

$$H(\theta) = \int_0^1 \frac{1}{\sqrt{\eta_1(1-\eta_1)}} \frac{1}{\sqrt{\theta(1-\eta_1) + \eta_1}} d\eta_1 < \infty$$

so that

$$\pi_1(\eta_1|\theta) = \frac{1}{H(\theta)\sqrt{\eta_1(1-\eta_1)}} \frac{1}{\sqrt{\theta(1-\eta_1) + \eta_1}}, \quad \eta_1 \in (0, 1) \quad (3.15)$$

is a proper density.

3.2.3 Derivation of $\pi(\theta)$

To shift our attention back to the final parameter of interest θ and derive the prior of interest $\pi(\theta)$, we note

$$\frac{|I_4^*|}{|I_3^*|} = \frac{\eta_1(1-\eta_1)\eta_2\eta_3}{\theta(\theta(1-\eta_1) + \eta_1 + \theta\eta_3)},$$

giving us

$$\begin{aligned} \pi(\theta) &\propto \exp \left\{ \frac{1}{2} \int_0^1 \int_{a_{2k}}^{b_{2k}} \int_{a_{3k}}^{b_{3k}} \log \left[\frac{\eta_1(1-\eta_1)\eta_2\eta_3}{\theta(\theta(1-\eta_1) + \eta_1 + \theta\eta_3)} \right] \right. \\ &\quad \left. \times \pi_{3k}(\eta_3|\eta_2, \eta_1, \theta) \pi_{2k}(\eta_2|\eta_1, \theta) \pi_1(\eta_1|\theta) d\eta_3 d\eta_2 d\eta_1 \right\} \\ &\propto \frac{1}{\sqrt{\theta}} \exp \left\{ \frac{1}{2} \int_0^1 \int_{a_{3k}}^{b_{3k}} \log \left[\frac{\eta_1(1-\eta_1)}{\theta(1-\eta_1) + \eta_1 + \theta\eta_3} \right] \pi_{3k}(\eta_3|\eta_2, \eta_1, \theta) \pi_{1k}(\eta_1|\theta) d\eta_3 d\eta_1 \right\} \\ &= \frac{1}{\sqrt{\theta}} \exp \left\{ \frac{1}{2} \int_0^1 \int_{a_{3k}}^{b_{3k}} \log \left[\frac{\eta_1(1-\eta_1)}{\theta(1-\eta_1) + \eta_1 + \theta\eta_3} \right] \frac{\pi_{1k}(\eta_1|\theta) d\eta_3 d\eta_1}{2\sqrt{\eta_3}(\sqrt{b_{3k}} - \sqrt{a_{3k}})} \right\}. \end{aligned}$$

Again letting $\beta = \frac{\theta}{\theta(1-\eta_1)+\eta_1}$, note that

$$\log \left[\frac{\eta_1(1-\eta_1)}{\theta(1-\eta_1)+\eta_1+\theta\eta_3} \right] = \log [\eta_1(1-\eta_1)] - \log [\theta(1-\eta_1)+\eta_1] - \log [1+\beta\eta_3],$$

which yields

$$\begin{aligned} \pi_k(\theta) &\propto \frac{1}{\sqrt{\theta}} \exp \left\{ \frac{1}{2} \int_0^1 \int_{a_{3k}}^{b_{3k}} \left[\log \left\{ \frac{\eta_1(1-\eta_1)}{\theta(1-\eta_1)+\eta_1} \right\} - \log (1+a\eta_3) \right] \right. \\ &\quad \left. \times \frac{\pi_1(\eta_1|\theta) d\eta_3 d\eta_1}{2\sqrt{\eta_3}(\sqrt{b_{3k}}-\sqrt{a_{3k}})} \right\}. \end{aligned}$$

Using 3.13, we have

$$\begin{aligned} \pi_k(\theta) &\propto \frac{1}{\sqrt{\theta}} \exp \left\{ \frac{1}{2} \int_0^1 \log \left[\frac{\eta_1(1-\eta_1)}{\theta(1-\eta_1)+\eta_1} \right] \pi_1(\eta_1|\theta) d\eta_1 \right. \\ &\quad \left. - \frac{1}{2} \int_0^1 \log \left[\frac{\theta}{\theta(1-\eta_1)+\eta_1} \right] \pi_1(\eta_1|\theta) d\eta_1 \right\} \\ &= \frac{1}{\sqrt{\theta}} \exp \left\{ \frac{1}{2} \int_0^1 \log \left[\frac{\eta_1(1-\eta_1)}{\theta} \right] \pi_1(\eta_1|\theta) d\eta_1 \right\} \\ &= \frac{1}{\theta} \exp \left\{ \frac{1}{2} \int_0^1 \log [\eta_1(1-\eta_1)] \pi_1(\eta_1|\theta) d\eta_1 \right\} \\ &= \frac{1}{\theta} \exp \left\{ \frac{1}{H(\theta)} \int_0^1 \frac{\log [\sqrt{\eta_1(1-\eta_1)}]}{\sqrt{\eta_1(1-\eta_1)(\theta(1-\eta_1)+\eta_1)}} d\eta_1 \right\} \\ &= \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\}. \end{aligned} \tag{3.16}$$

Here, $G(\theta)$ is an integration with the dummy variable t

$$G(\theta) = \int_0^1 \frac{\log \sqrt{t(1-t)}}{\sqrt{t(1-t)(\theta(1-t)+t)}} dt.$$

Note that a closed form is available for $G(\theta)$, though its functional form is not helpful in this case as it depends on a more complex numerical integration. More work can be done in the future to try to manipulate this into a more manageable form. Still, the reader is referred to Appendix A.1 for a study of its derivation.

3.3 Behavior of the Normalizing Constant $H(\theta)$ and $G(\theta)$

The functions $H(\theta)$ and $G(\theta)$ have interesting behaviors that lead us to even more interesting behavior about the marginal priors $\pi^R(\theta)$ and $\pi^R(\eta_1|\theta)$. These results are extremely useful in demonstrating various properties of the reference prior $\pi^R(\theta)$ and $\pi^R(\eta_1|\theta)$.

Lemma 3.3.1. The function $H(\theta)$ as defined in (3.5) satisfies the following properties

- (a) $H(1) = \pi$.
- (b) $H\left(\frac{1}{\theta}\right) = \sqrt{\theta}H(\theta), \quad \forall \theta > 0$.
- (c) $H(\theta)$ is a strictly monotone *decreasing* function in θ .
- (d) $H(\theta)$ can be represented by the following equivalent expressions using the *Gauss Hypergeometric function* ${}_2F_1(a, b; c; z)$:

$$H(\theta) = \frac{\pi}{\sqrt{\theta}} {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; \frac{\theta-1}{\theta}\right) = \pi {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; 1-\theta\right).$$

(e) We have the following limits for $H(\theta)$:

$$\lim_{\theta \rightarrow 0} \frac{H(\theta)}{-\log(\theta)} = 1, \quad \lim_{\theta \rightarrow \infty} \frac{H(\theta)}{\frac{\log \theta}{\sqrt{\theta}}} = 1.$$

Proof. For (a), simply substitute $\theta = 1$ and note that $H(1)$ is the beta integral. For (b), we have

$$\begin{aligned} H\left(\frac{1}{\theta}\right) &= \int_0^1 \frac{1}{\sqrt{s(1-s) \left[\frac{1}{\theta}(1-s) + s\right]}} ds \\ &= \sqrt{\theta} \int_0^1 \frac{1}{\sqrt{s(1-s) [(1-s) + \theta s]}} ds, \end{aligned}$$

which is equal to $\sqrt{\theta}H(\theta)$ after the substitution $t = 1 - s$.

For (c), simply note that for any $0 < \theta_1 < \theta_2$ and $s \in (0, 1)$,

$$\frac{1}{\sqrt{\theta_1(1-s) + s}} > \frac{1}{\sqrt{\theta_2(1-s) + s}}.$$

For (d), we note that the *Gauss hypergeometric* function ${}_2F_1(a, b; c; z)$ is represented by the series

$${}_2F_1(a, b; c; z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}, \quad (3.17)$$

where $(a)_n$ is called the *rising Pochhammer symbol* and when $|z| < 1$ and by analytic continuation (Bateman, 1953) elsewhere. As presented in Bateman (1953), Euler(1748) showed that

$$B(b, c-b) {}_2F_1(a, b; c; z) = \int_0^1 x^{b-1} (1-x)^{c-b-1} (1-zx)^{-a} dx, \quad (3.18)$$

provided that $z \in \mathbb{R} < 1$ and $c > b$. Noting that $\theta - 1 < \theta \Rightarrow \frac{\theta-1}{\theta} < 1 \quad \forall \quad \theta > 0$, and choosing $a = \frac{1}{2}$, $b = \frac{1}{2}$, and $c = 1$, this gives us

$$\begin{aligned} H(\theta) &= \int_0^1 \frac{1}{\sqrt{\eta_1(1-\eta_1)}} \frac{1}{\sqrt{\theta(1-\eta_1) + \eta_1}} d\eta_1 \\ &= \frac{1}{\sqrt{\theta}} \int_0^1 \eta_1^{\frac{1}{2}-1} (1-\eta_1)^{1-\frac{1}{2}-1} \left[1 - \frac{\theta-1}{\theta} \eta_1 \right]^{-\frac{1}{2}} d\eta_1 \\ &= \frac{\pi}{\sqrt{\theta}} {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; \frac{\theta-1}{\theta} \right). \end{aligned} \quad (3.19)$$

Furthermore, the Pfaff transformation for the hypergeometric function is defined by

$${}_2F_1(a, b; c; z) = (1-z)^{-a} {}_2F_1 \left(a, c-b; c; \frac{z}{z-1} \right), \quad (3.20)$$

which yields the alternative form

$$H(\theta) = \pi \cdot {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; 1-\theta \right). \quad (3.21)$$

Beyond aiding in finding the asymptotic behavior of $H(\theta)$, this alternative parameterization is extremely useful for computation, as practical numerical evaluation of ${}_2F_1(a, b; c; z)$ can be unstable for $|z| > 1$, even though mathematically there is no problem analytically continuing the function into this region (Bateman, 1953). Now, to show (e), note that equation (15.4.21) in Olver et al. (2010) states that, if $c = a + b$,

$$\lim_{z \rightarrow 1^-} \frac{{}_2F_1(a, b; a+b; z)}{-\log(1-z)} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}. \quad (3.22)$$

Applying this to (3.21) and noting that $\theta \rightarrow 0^+$ is equivalent to $z = (1-\theta) \rightarrow 1^-$,

we have

$$\lim_{(1-\theta) \rightarrow 1^-} \frac{{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; 1-\theta\right)}{-\log \theta} = \frac{1}{\pi} \iff \lim_{\theta \rightarrow 0} \frac{H(\theta)}{-\log \theta} = 1, \quad (3.23)$$

which is the first part of (e). Next, noting that $(1 - \frac{1}{\theta}) = \frac{\theta-1}{\theta} \rightarrow 1^-$ is the same as $\theta \rightarrow \infty$, we apply Lemma 3.3.1 to (3.23) and see that

$$\lim_{\theta \rightarrow \infty} \frac{H(\theta)}{\frac{\log \theta}{\sqrt{\theta}}} = 1, \quad (3.24)$$

thus proving the second part of (e). Because of the strictly decreasing nature of $H(\theta)$ as well as the stable and positive tail behavior, we can conclude that $H(\theta)$ is a positive and strictly decreasing function of θ . $\square$

Now we look at $G(\theta)$, which also has some very interesting and helpful properties.

Lemma 3.3.2. The function $G(\theta)$ as defined in (3.5) satisfies the following properties

- (a) $G\left(\frac{1}{\theta}\right) = \sqrt{\theta}G(\theta), \quad \forall \theta > 0.$
- (b) $G(\theta)$ is a strictly monotone *increasing* function of θ .
- (c) We have the following limits for $G(\theta)$:

$$\lim_{\theta \rightarrow 0} \frac{G(\theta)}{-\log^2 \theta} = 1, \quad \lim_{\theta \rightarrow \infty} \frac{G(\theta)}{-\frac{\log^2 \theta}{\sqrt{\theta}}} = 1.$$

Proof. The proof of (a) is similar to lemma 3.3.1 part (a). The proof of (b) is similar to lemma 3.3.1 part (b) after noting that for $t \in (0, 1)$, $\log [t(1-t)] < 0$. Recall that

if $z = \frac{\theta-1}{\theta}$

$$\begin{aligned}
G(\theta) &= \int_0^1 \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t)+t)}} dt \\
&= \frac{1}{\sqrt{\theta}} \int_0^1 \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(1-zt)}} dt \\
&= \frac{1}{2\sqrt{\theta}} \left(\int_0^1 \frac{\log t}{\sqrt{t(1-t)(1-zt)}} dt + \int_0^1 \frac{\log(1-t)}{\sqrt{t(1-t)(1-zt)}} dt \right)
\end{aligned}$$

Now, $\theta \rightarrow \infty \equiv z \rightarrow 1$. From an application of equation 4.254(6) in Gradshteyn and Ryzhik (2007) with $q = 1/2$, we can see that as $\theta \rightarrow \infty$, we have

$$\int_0^1 \frac{\log t}{\sqrt{t(1-t)(1-zt)}} dt \rightarrow \int_0^1 \frac{\log t}{\sqrt{t(1-t)}} dt = -\frac{\pi^2}{2},$$

and therefore,

$$G(\theta) \approx \frac{1}{2\sqrt{\theta}} \left(\int_0^1 \frac{\log(1-t)}{\sqrt{t(1-t)(1-zt)}} dt - \frac{\pi^2}{2} \right). \quad (3.25)$$

Using (3.18), we can choose $a = \frac{1}{2}$, $b = \frac{1}{2}$, and $c = T + 1$, we have

$$B\left(\frac{1}{2}, T + \frac{1}{2}\right) {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; T + 1; z\right) = \int_0^1 \frac{(1-\eta_1)^T}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1. \quad (3.26)$$

If we differentiate the RHS of (A.5) with respect to T then set $T = 0$ we see that

$$\begin{aligned}
\frac{d}{dT} \int_0^1 \frac{(1-\eta_1)^T}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 &= \int_0^1 \frac{(1-\eta_1)^T \log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \Big|_{T=0} \\
&= \int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1. \quad (3.27)
\end{aligned}$$

Using differentiation rules for the Gamma function, we have

$$\begin{aligned} \frac{d}{dT} \left[B \left(\frac{1}{2}, T + \frac{1}{2} \right) {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; T + 1; z \right) \right] \Big|_{T=0} = \\ {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) \left\{ B \left(\frac{1}{2}, \frac{1}{2} \right) \left[\psi \left(\frac{1}{2} \right) - \psi(1) \right] \right\} + B \left(\frac{1}{2}, \frac{1}{2} \right) \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; T + 1; z \right) \Big|_{T=0}. \end{aligned}$$

Using the definition of the hypergeometric series in (3.17), we have

$$\begin{aligned} \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; T + 1; z \right) &= \frac{d}{dT} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(T+1)_n} \frac{z^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right)}{\Gamma^2 \left(\frac{1}{2} \right)} \frac{z^n}{n!} \frac{d}{dT} \left[\frac{\Gamma(T+1)}{\Gamma(T+n+1)} \right] \Big|_{T=0} \\ &= \sum_{n=0}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right)}{\Gamma^2 \left(\frac{1}{2} \right)} \frac{z^n}{n!} \frac{\psi(1) - \psi(n+1)}{\Gamma(n+1)}. \end{aligned}$$

Then,

$$\begin{aligned} \int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 = \\ -2\pi \log 2 \cdot {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) + \sum_{n=0}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right)}{\Gamma(n+1)} \frac{z^n}{n!} (\psi(1) - \psi(n+1)). \end{aligned} \quad (3.28)$$

To deal with the second term, note that

$$\psi(1) - \psi(n+1) = -H_n \approx -[\gamma + \log(n)],$$

where $H_n = \sum_{k=1}^n \frac{1}{k}$ is the n^{th} Harmonic Number and $\gamma = -\psi(1)$ is the Euler-Mascheroni constant. This tells us that (A.7) is

$$\int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \approx -(2\pi \log 2 + \gamma) {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) - \sum_{n=1}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right) \cdot \log n}{\Gamma^2(n+1)} z^n \quad (3.29)$$

From here, we can apply Stirling's formula to the second term of (3.29) and reduce to see that

$$\sum_{n=1}^{\infty} \frac{\Gamma^2(n + \frac{1}{2}) \cdot \log n}{\Gamma^2(n + 1)} z^n \approx \sum_{n=1}^{\infty} \frac{\log n}{n} z^n \quad (3.30)$$

To evaluate the sum on the RHS of (3.30), we apply lemma 3 of Vargas (2014), which states that for a given function $\psi : [N, \infty) \rightarrow \mathbb{R}^+$ suppose that the map $t \mapsto \psi(t)x^t$ is unimodal with maximum at $t = t_x \geq N$, $0 < x < 1$. Then

$$\sum_{n \geq N} \psi(n)x^n = \int_N^{\infty} \psi(t)x^t dt + O(\psi(t_x)x^{t_x}) + O(1)$$

as $x \rightarrow 1^-$. Now, the map $t \mapsto \frac{\log t}{t} z^t$ is indeed unimodal, and $\frac{\log t}{t} z^t \leq 1/e$ for all $0 \leq z \leq 1$.

Thus

$$\sum_{n=1}^{\infty} \frac{\log n}{n} z^n = \int_1^{\infty} \frac{\log t}{t} z^t dt + O(1).$$

as $z \rightarrow 1^-$. If we set $\lambda = -1/\log z$ (so that $\lambda \rightarrow \infty$ as $z \rightarrow 1^-$) and make the substitution $s = t/\lambda$, the integral becomes

$$\int_1^{\infty} \frac{\log t}{t} z^t dt = \int_1^{\infty} \frac{\log t}{t} e^{-t/\lambda} dt = \log \lambda \int_{1/\lambda}^{\infty} \frac{e^{-s}}{s} ds + \int_{1/\lambda}^{\infty} \frac{e^{-s} \log s}{s} ds. \quad (3.31)$$

Evaluating the terms individually, we have,

$$\int_{1/\lambda}^{\infty} \frac{e^{-s}}{s} ds \sim \int_{1/\lambda}^1 \frac{e^{-s}}{s} ds \sim \int_{1/\lambda}^1 \frac{1}{s} ds = \log \lambda$$

and

$$\int_{1/\lambda}^{\infty} \frac{e^{-s} \log s}{s} ds \sim \int_{1/\lambda}^1 \frac{e^{-s} \log s}{s} ds \sim \int_{1/\lambda}^1 \frac{\log s}{s} ds = -\frac{(\log \lambda)^2}{2}, \text{ as } \lambda \rightarrow \infty.$$

By substituting these into (3.31) we get

$$\int_1^\infty \frac{\log t}{t} z^t dt \sim \frac{(\log \lambda)^2}{2} = \frac{[-\log(-\log z)]^2}{2} \sim \frac{[\log(1-z)]^2}{2},$$

and hence by (3.31) we have

$$\sum_{n=1}^\infty \frac{\log n}{n} z^n \sim \frac{[\log(1-z)]^2}{2}, \text{ as } z \rightarrow 1^-.$$

Recalling the limiting behavior of ${}_2F_1(\cdot)$ in (3.22), we now have that

$$\int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \approx (2\pi \log 2 + \gamma) \log(1-z) - \frac{[\log(1-z)]^2}{2}. \quad (3.32)$$

Noting the leading term of $\frac{[\log(1-z)]^2}{2}$, substituting $z = \frac{\theta-1}{\theta}$ back into (3.32) and substituting this into (3.25), gives the result

$$\lim_{\theta \rightarrow \infty} \frac{G(\theta)}{-\frac{\log^2 \theta}{\sqrt{\theta}}} = 1.$$

Further, as a result of the symmetry of $G(\theta)$, we also have that

$$\lim_{\theta \rightarrow 0} \frac{G(\theta)}{\log^2 \theta} = 1. \quad (3.33)$$

□

Both H and G are non-autonomous Schröder functional equations, monotonic, and behave as relatively simple functions near 0 and ∞ . We further note that the properties in lemma 3.3.1(b) and 3.3.2(a) should hold for *any* function of $t(t-1)$ in the numerator of $H(\theta)$.

3.4 Behavior of the Marginal Reference Prior $\pi^R(\theta)$

Now we will study some properties of $\pi^R(\theta)$ and $\pi^R(\eta_1|\theta)$. These properties are remarkable.

Theorem 3.4.1. The marginal reference prior $\pi^R(\theta)$ in (3.7) for the odds ratio θ is

$$\pi^R(\theta) = \frac{c}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\}, \quad \frac{1}{c} = \int_0^\infty \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} d\theta, \quad (3.34)$$

and has the following properties:

(a) The following limits hold for $\pi^R(\theta)$

$$\lim_{\theta \rightarrow 0} \frac{\pi^R(\theta)}{\theta^{-3/4}} = 1, \quad \lim_{\theta \rightarrow \infty} \frac{\pi^R(\theta)}{\theta^{-5/4}} = 1.$$

(b) It is a proper distribution with normalizing constant $c \approx 3.585$

(c) $\pi^R\left(\frac{1}{\theta}\right) = \theta^2 \pi^R(\theta)$, $\theta > 0$.

(d) The prior density of $\frac{1}{\theta}$ and θ are identical.

(e) $\int_0^1 \pi^R(\theta) d\theta = 0.5$. That is, the median of $\pi^R(\theta)$ is 1.

(f) $\int_0^\infty \theta \pi^R(\theta) d\theta = \infty$. That is, the mean of $\pi^R(\theta)$ does not exist.

Proof. For (a), we first note that as a result of part lemma 3.3.1 (e) and 3.3.2 (c), we can see

$$\lim_{\theta \rightarrow 0} \frac{\frac{G(\theta)}{H(\theta)}}{\frac{1}{4} \log \theta} = 1 \quad \text{and} \quad \lim_{\theta \rightarrow \infty} \frac{\frac{G(\theta)}{H(\theta)}}{-\frac{1}{4} \log \theta} = 1. \quad (3.35)$$

Combining this with $\frac{1}{\theta}$ completes the proof. (b) is a direct result of (a) demonstrating integrability of $\pi(\theta)$ at 0 and ∞ , with the approximation of the normalizing constant being given by numerical integration. (c) can be shown in the same way as in the proof of lemma 3.3.1, without the substitution.

To show (e), Consider the integration

$$\int_0^1 \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} d\theta,$$

and perform the substitution $\gamma = \frac{1}{\theta} \Rightarrow \theta = \frac{1}{\gamma} \Rightarrow \frac{d\theta}{d\gamma} = -\frac{1}{\gamma^2}$, so that

$$\int_0^1 \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} d\theta = \int_{\infty}^1 \gamma \exp \left\{ \frac{G(1/\gamma)}{H(1/\gamma)} \right\} \left(-\frac{1}{\gamma^2} \right) d\gamma = \int_1^{\infty} \frac{1}{\gamma} \exp \left\{ \frac{G(\gamma)}{H(\gamma)} \right\} d\gamma.$$

This gives us $\int_0^1 \pi(\theta) d\theta = \int_1^{\infty} \pi(\theta) d\theta$, proving the result.

For (d), we apply property (c) of this Theorem as well as the corresponding Jacobian

$\omega = \frac{1}{\theta} \Rightarrow \frac{d\theta}{d\omega} = \omega^{-2}$ to see

$$\pi^*(\omega) = \pi \left(\frac{1}{\omega} \right) \omega^{-2} = \omega^2 \pi(\omega) \omega^{-2} = \pi(\omega).$$

For (f), we apply (3.36) and note that

$$\lim_{\theta \rightarrow 0} \frac{\theta \pi(\theta)}{\theta^{1/4}} = 1 \quad \text{and} \quad \lim_{\theta \rightarrow \infty} \frac{\theta \pi(\theta)}{\theta^{-1/4}} = 1. \quad (3.36)$$

Thus,

$$\int_0^{\infty} \theta \pi(\theta) d\theta > \int_1^{\infty} \theta \pi(\theta) d\theta = \infty.$$

□

We note that $\pi^R(\theta) \approx \theta^{-3/4}$ and $\pi^R(\theta) \approx \theta^{-5/4}$ for small and large θ , respectively. Since $\int_0^\varepsilon \theta^{-3/4} d\theta < \infty$ and $\int_M^\infty \theta^{-5/4} d\theta < \infty$ for large M , this shows that $\int_0^\infty \pi^R(\theta) d\theta < \infty$. It is curious to note that an initial approach to find $\pi^R(\theta)$ was to apply the numerical reference prior algorithm in Berger et al. (2009). The result of the algorithm suggested that $\pi(\theta) \propto \frac{1}{\theta}$, which is improper. Indeed, the analytical solution is very close to $\frac{1}{\theta}$, however it is remarkable that the factor $\exp\left\{\frac{G(\theta)}{H(\theta)}\right\}$ pushes this density below $\frac{1}{\theta}$ at both boundaries of the parameter space thus making it a proper density. This is illustrated in Figure 3.1 where we can see a plot of the reference prior for θ $\pi^R(\theta)$ and overlaid on $\pi^*(\theta) \propto 1/\theta$ over the range $\theta \in [0.1, 4]$ in the left panel. They are almost indistinguishable, but in the right side of Figure 3.1 we can see that $\pi^R(\theta)/\pi^*(\theta) > 1$ in a neighborhood around $\theta = 1$ and less than one at the tails. Also from (a) we note that this prior is exceptionally heavy tailed. As $\theta \rightarrow \infty$, it's behavior of $\theta^{-5/4}$ is much heavier than even a Cauchy prior for a location parameter, which behaves as θ^{-2} for large θ . Noting that $\frac{1}{\theta} = \frac{\lambda_2 \lambda_3}{\lambda_1 \lambda_4}$, property (d) tells us that $\pi^R(\theta)$ is invariant to what we consider “positive” and “negative” outcomes. Finally, (e) and (f) demonstrate that the median of this prior density corresponds to $\theta = 1$, which is the quantity of θ which indicates no relationship between the row and column variables, and despite having this reasonable value for the median, the mean still does not exist.

3.4.1 Reference Prior for the Log odds Ratio

Often it is of interest to study the log odds ratio of a 2×2 table. Because we have the marginal reference prior for the odds ratio, we can apply a simple transformation

obtain the reference prior for the log odds ratio as well. As we will now see, this prior has remarkable properties, as well.

Theorem 3.4.2. The marginal reference prior $\pi_\xi^R(\xi)$ for the log odds ratio $\xi = \log(\theta)$ can be obtained from the reference prior for the odds ratio $\pi^R(\theta)$ in (3.34) as

$$\pi_\xi^R(\xi) = \pi^R(e^\xi) \cdot e^\xi, \quad (3.37)$$

and has the following properties:

- (a) For any $\xi \in \mathcal{R}$, $\pi_\xi^R(\xi) = \pi_\xi^R(-\xi)$, that is, the prior is symmetric.
- (b) The following limits hold for $\pi_\xi^R(\xi)$

$$\lim_{\xi \rightarrow -\infty} \frac{\pi_\xi^R(\xi)}{e^{-\xi/4}} = 1, \quad \lim_{\xi \rightarrow \infty} \frac{\pi_\xi^R(\xi)}{e^{\xi/4}} = 1.$$

Proof. For (a), we have $\pi_\xi^R(-\xi) = \pi^R(e^{-\xi})e^{-\xi}$, but $\pi^R(e^{-\xi}) = e^{2\xi}\pi^R(e^\xi)$ by Theorem 3.4.1(c). Therefore $\pi_\xi^R(-\xi) = e^{2\xi}\pi^R(e^\xi)e^{-\xi} = \pi^R(e^\xi) \cdot e^\xi = \pi_\xi^R(\xi)$. To show (b), note the limiting behavior of $\pi(\theta)$ in Theorem 3.4.1(a) and apply the transformation $\xi = \log \theta$ as in (3.37). $\square$

Theorem 3.4.2(a) is a desirable property, giving equal probability for the log odds to be above or below 0. This symmetry of the prior for the log odds ratio was also observed in Sun and Berger (1998), where they derived the reference prior for the log odds ratio under joint binomial(fixed marginal) sampling. Combining this with (b) shows that the unbalanced tail behavior of the prior for the odds ratio $\pi(\theta)$ becomes balanced on log scale. Further, we can see from (b) that the tail behavior of the reference prior on log scale is exactly the same as a Laplace(0,4) distribution. Figure

3.2 shows the priors on the odds ratio scale as well as on the log odds scale. On log scale, the Laplace prior has much more weight at 0 than the reference prior, and $\pi^*(\theta) \propto 1/\theta$ is not shown on log scale because it is constant. The behavior of the Laplace prior on $\xi = \log \theta$ is apparent after transformation to $\theta = e^\xi$ space as well. The Laplace prior puts more prior mass on $\theta = 1$, which indicates no relationship. This behavior is often used in problems where we would like to induce sparsity into the solutions of high dimensional problems like the “large p ” regression setup.

3.5 Behavior of the Conditional Reference Prior $\pi^R(\eta_1|\theta)$

Theorem 3.5.1. The conditional reference prior $\pi^R(\eta_1|\theta)$ for the proportion of successes in the treatment group η_1 takes its form from (3.7) as

$$\pi^R(\eta_1|\theta) = \frac{[H(\theta)]^{-1}}{\sqrt{\eta_1(1-\eta_1)}\sqrt{\theta(1-\eta_1) + \eta_1}}, \quad (3.38)$$

and has the following properties:

- (a) For any $\theta > 0$, it is a proper distribution.
- (b) $\pi^R(1 - \eta_1|\theta) = \pi^R(\eta_1|\frac{1}{\theta})$ and $\pi^R(1 - \eta_1|\frac{1}{\theta}) = \pi^R(\eta_1|\theta)$, $\forall \eta_1 \in (0, 1)$ and $\theta > 0$.
- (c) $\pi^R(\eta_1|\theta = 1)$ is the $Beta(1/2, 1/2)$ distribution.

Proof. (a) is shown in the proof for Theorem 3.1.1. (b) can be shown in the same way as in the proof of lemma 3.3.1, without the substitution. (c) is trivially obtained by substituting $\theta = 1$ into (3.38). □

Points 2 and 3 illustrate that $\pi^R(\eta_1|\theta)$ is a weighted beta distribution, of sorts. As θ increases, the diagonal elements λ_1 and λ_4 tend to grow. With $\eta_1 = \frac{\lambda_1}{\lambda_1 + \lambda_3}$, this pushes the mass of $\pi^R(\eta_1|\theta)$ towards 1, this is illustrated in Figure 3.3. Finally, when this conditional prior is combined with the marginal prior for the parameter of interest, we have one more interesting prior property.

Lemma 3.5.2. The joint prior $\pi^R(\theta, \eta_1)$ in (3.7) has the property

$$\pi^R\left(\frac{1}{\theta}, 1 - \eta_1\right) = \theta \pi^R(\theta, \eta_1). \quad (3.39)$$

Proof. This immediately follows as a result of Theorem 3.4.1 (b) and Theorem 3.5.1 (b). □

As a result of this, when we examine a plot of (3.7), we can see that in a sense, the relationship in Lemma 3.5.1(b) reverses (see Figure 3.4). When we factor in the parameter of interest θ , we can see that for $\theta > 1$, there is slightly more mass on $\eta_1 < .5$, which is not what we would “expect” as large values of θ would tend to increase the values of λ_1 and λ_4 . Further, note that in the middle region for η_1 , the marginal prior for θ approaches 0 considerably slower. Recalling that $\eta_1 = \frac{\lambda_1}{\lambda_1 + \lambda_3}$, this may tell us that when λ_1 and λ_3 are well balanced, we are more unsure of θ , but when either λ_1 or λ_3 is much bigger than the other (i.e. η_1 is close to 0 or 1), the prior for θ is strong. Furthermore, note that for values of θ less than 1, this joint prior places more mass on values of η_1 *greater* than 0.5, and vice versa for values of $\theta > 1$. This is curious, because since $\theta = \frac{\lambda_1 \lambda_4}{\lambda_3 \lambda_2}$, large values of θ would mean that $\eta_1 = \frac{\lambda_1}{\lambda_1 + \lambda_3}$ would *tend* to be closer to 1 because λ_1 would *tend* to be greater than λ_3 . Of course this is not true in general, however the fact that this joint prior is

reflecting the opposite of this indicates that in some sense, it is a prior which is “least favorable” to the data.

3.6 Figures and Tables

Figure 3.1: Plot of $\pi^R(\theta)$ compared to $\pi^*(\theta) \propto 1/\theta$.

Plot of the Reference prior and Laplace(0,4) on Log scale

Figure 3.2: Plot of $\pi^R(\theta)$ compared to $\pi^*(\theta) \propto 1/\theta$ as well as a transformed Laplace Prior.

Figure 3.3: Plot of $\pi(\eta_1|\theta)$ for some θ .

Figure 3.4: Plot of $\pi(\theta, \eta_1)$.

Chapter 4

Reference Posterior Analysis and Computation for the Odds Ratio of the 2x2 Contingency Table

Let $(n_{i1}, n_{i2}, n_{i3}, n_{i4})$, $i = 1, \dots, N$ be a random sample of tables each having the joint Poisson model given by (2.7). Since $n_{ij} \sim \text{Poisson}(\lambda_j)$, $\sum_{i=1}^N n_{ij} \sim \text{Poisson}(N\lambda_j)$ is a sufficient statistic for the means of each of the $j = 1, \dots, 4$ cells of the sum of the N tables. Using the same joint transformation in (3.1) on the means of the parameters of this summary table yields

$$\theta' = \frac{N^2 \lambda_1 \lambda_4}{N^2 \lambda_2 \lambda_3} = \theta, \quad \eta'_1 = \frac{N \lambda_1}{N \lambda_1 + N \lambda_3} = \eta_1, \quad \eta'_2 = N(\lambda_1 + \lambda_3), \quad \eta'_3 = \frac{N \lambda_2}{N \lambda_1} = \eta_3. \quad (4.1)$$

This means that with respect to θ , η_1 , and η_3 , one table is sufficiently able to represent the information in these parameters. On the other hand, η_2 is related to N , so it, in a sense, is a scale parameter for this model and can be used as a proxy for

studying “sample size” in the simulation study later. Note that, despite all three nuisance parameters having informative interpretations, it likely is the case that only η_1 , the proportion of negatives in the control group, would be of interest. It is indeed convenient that the joint posterior $\pi^R(\theta, \eta_1, \eta_2, \eta_3 | \mathbf{n})$ which arises when combining the results of theorem 3.1.1 with the transformed likelihood can be marginalized marginalized η_2 and η_3 in the following result.

4.1 Posterior Distributions

Theorem 4.1.1. Consider the likelihood in (3.4) under the reference prior (3.6).

(a) Recalling that $n = n_1 + n_2 + n_3 + n_4$,

$$\pi^R(\eta_2 | \theta, \eta_1, \eta_3; \mathbf{n}) \sim \text{Gamma} \left(n + \frac{1}{2}, \eta_1(1 + \eta_3) + (1 - \eta_1)(1 + \theta\eta_3) \right). \quad (4.2)$$

(b) If $n_1 + n_3 \geq 1$,

$$\pi^R(\eta_3 | \theta, \eta_1; \mathbf{n}) \stackrel{d}{=} \frac{1}{\eta_1 + \theta(1 - \eta_1)} \frac{\gamma}{1 - \gamma}, \quad \gamma \sim \text{Beta}(n_2 + n_4 + 1/2, n_1 + n_3). \quad (4.3)$$

(c) If $n_1 + n_3 \geq 1$, the marginal likelihood for (θ, η_1) is

$$L_1(\theta, \eta_1) = \int_0^\infty \int_0^\infty L(\theta, \eta_1, \eta_2, \eta_3) \pi^R(\eta_2, \eta_3) d\eta_2 d\eta_3 \quad (4.4)$$

$$= \frac{\Gamma(n_2 + n_4 + 1/2) \Gamma(n_1 + n_3)}{n_1! n_2! n_3! n_4!} \frac{\theta^{n_4} \eta_1^{n_1 + n_2} (1 - \eta_1)^{n_3 + n_4}}{(\eta_1 + \theta(1 - \eta_1))^{n_2 + n_4 + 1/2}}. \quad (4.5)$$

(d) If $n_1 \geq 1$, $L_1(\theta, \eta_1)$ is bounded by $\frac{\Gamma(n_2 + n_4 + 1/2) \Gamma(n_1 + n_3)}{n_1! n_2! n_3! n_4!}$ so that the joint posterior $\pi^R(\theta, \eta_1, \eta_2, \eta_3 | \mathbf{n})$ is proper.

(e) If $n_1 \geq 1$, we have

$$\pi^R(\theta, \eta_1 | \mathbf{n}) \propto \frac{\theta^{n_4-1}}{H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} \frac{\eta_1^{n_1+n_2-1/2} (1-\eta_1)^{n_3+n_4-1/2}}{(\eta_1 + \theta(1-\eta_1))^{n_2+n_4+1}}. \quad (4.6)$$

(f) As a result of (e), if $n_1 \geq 1$, we have

$$\pi^R(\eta_1 | \theta; \mathbf{n}) \propto \frac{\eta_1^{n_1+n_2-1/2} (1-\eta_1)^{n_3+n_4-1/2}}{(\eta_1 + \theta(1-\eta_1))^{n_2+n_4+1}}. \quad (4.7)$$

(g) If $n_1 \geq 1$,

$$\pi^R(\theta | \mathbf{n}) \propto \frac{{}_2F_1 \left(n_2 + n_4 + 1, n_1 + n_2 + \frac{1}{2}; \sum_{j=1}^4 n_j + 1; \frac{\theta-1}{\theta} \right)}{\theta^{n_2+2} H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\}. \quad (4.8)$$

(h) If n_2 and n_3 are both at least $m \geq 1$, the m^{th} moment of θ exists, i.e.,

$$\mathbb{E}(\theta^m | \mathbf{n}) = \int_0^\infty \theta^m \pi^R(\theta | \mathbf{n}) d\theta < \infty, \quad (4.9)$$

Proof. From the likelihood (3.4) and results of Theorem 3.1.1, we can immediately conclude

$$\pi(\eta_2 | \theta, \eta_1, \eta_3, n) \sim \text{Gamma} \left(n_1 + n_2 + n_3 + n_4 + \frac{1}{2}, \eta_1(1 + \eta_3) + (1 - \eta_1)(1 + \theta\eta_3) \right)$$

First, note that the joint conditional prior for $(\eta_2, \eta_3 | \theta, \eta_1)$ is

$$\pi(\eta_2, \eta_3 | \theta, \eta_1) \propto \frac{1}{\sqrt{\eta_2 \eta_3}}. \quad (4.10)$$

Next, combining (3.4) with (4.10) and integrating η_3 and η_2 , we have

$$L(\theta, \eta_1) = \frac{\eta_1^{n_1+n_2}(1-\eta_1)^{n_3+n_4}\theta^{n_4}}{n_1!n_2!n_3!n_4!} \int_0^\infty \eta_3^{n_2+n_4-1/2} \int_0^\infty \eta_2^{n_1+n_2+n_3+n_4-1/2} \times \\ \times \exp\{-\eta_2[\eta_1(1+\eta_3) + (1-\eta_1)(1+\theta\eta_3)]\} d\eta_2 d\eta_3. \quad (4.11)$$

Using the properties of the Gamma density, we can rewrite the inner integral to get,

$$\int_0^\infty \eta_2^{n_1+n_2+n_3+n_4-1/2} \exp\{-\eta_2[\eta_1(1+\eta_3) + (1-\eta_1)(1+\theta\eta_3)]\} d\eta_2 = \\ = \frac{\Gamma(n_1+n_2+n_3+n_4+1/2)}{[\eta_1(1+\eta_3) + (1-\eta_1)(1+\theta\eta_3)]^{n_1+n_2+n_3+n_4+1/2}}.$$

This gives us

$$L(\theta, \eta_1) = \frac{\Gamma(h_2)}{n_1!n_2!n_3!n_4!} \theta^{n_4} \eta_1^{n_1+n_2} (1-\eta_1)^{n_3+n_4} \int_0^\infty \eta_3^{h_1-1} (1+\eta_3\xi)^{-h_2} d\eta_3, \quad (4.12)$$

where $h_1 = n_2 + n_4 + 1/2$, $h_2 = n_1 + n_2 + n_3 + n_4 + 1/2$, and $\xi = \eta_1 + \theta(1-\eta_1)$. Then, consider the change of variables $t = \frac{\xi\eta_3}{1+\xi\eta_3} \in (0, 1)$, we therefore have $\eta_3 = \frac{t}{\xi(1-t)}$ and $d\eta_3 = \frac{1}{\xi(1-t)^2} dt$, which gives

$$\int_0^\infty \eta_3^{h_1-1} (1+\eta_3\xi)^{-h_2} d\eta_3 = \frac{1}{\xi^{h_1}} \int_0^1 \left(\frac{t}{1-t}\right)^{h_1-1} \left(1 + \frac{t}{1-t}\right)^{-h_2} \frac{1}{(1-t)^2} dt \\ = \frac{1}{\xi^{h_1}} \int_0^1 t^{h_1-1} (1-t)^{h_2-h_1-1} dt. \quad (4.13)$$

In (4.13) note that $h_1 > 0$, so if $h_2 - h_1 = n_1 + n_3 \geq 1$ the integrand in (4.12) finally

becomes

$$\int_0^\infty \eta_3^{h_1-1} (1 + \eta_3 \xi)^{-h_2} d\eta_3 = \frac{1}{(\eta_1 + \theta(1 - \eta_1))^{n_2+n_4+1/2}} \frac{\Gamma(n_2 + n_4 + 1/2) \Gamma(n_1 + n_3)}{\Gamma(n_1 + n_2 + n_3 + n_4 + 1/2)}. \quad (4.14)$$

Also, note that as a result of (4.13), we can say define $t = \frac{\xi \eta_3}{1 + \xi \eta_3}$ so that

$$t \sim \text{Beta}(n_2 + n_4 + 1/2, n_1 + n_3).$$

Therefore, we can show (c) by noting $\eta_3 = \frac{t}{\xi(1-t)} \sim \pi(\eta_3 | \theta, \eta_1, n)$. Continuing from (4.14), we have

$$L(\theta, \eta_1) = \frac{\Gamma(n_2 + n_4 + 1/2) \Gamma(n_1 + n_3)}{n_1! n_2! n_3! n_4!} \frac{\eta_1^{n_1+n_2} (1 - \eta_1)^{n_3+n_4} \theta^{n_4}}{(\eta_1 + \theta(1 - \eta_1))^{n_2+n_4+1/2}} \text{ if } n_1 + n_3 \geq 1,$$

which completes the proof for part (a). Now, note that

$$\frac{\eta_1^{n_1+n_2} (1 - \eta_1)^{n_3+n_4} \theta^{n_4}}{(\eta_1 + \theta(1 - \eta_1))^{n_2+n_4+1/2}} = \frac{\eta_1^{n_1} (1 - \eta_1)^{n_3}}{\sqrt{\eta_1 + \theta(1 - \eta_1)}} \left(\frac{\eta_1}{\eta_1 + \theta(1 - \eta_1)} \right)^{n_2} \left(\frac{\theta(1 - \eta_1)}{\eta_1 + \theta(1 - \eta_1)} \right)^{n_4}. \quad (4.15)$$

The second two factors are clearly less than or equal to 1. for the first factor, we have

$$\frac{\eta_1^{n_1} (1 - \eta_1)^{n_3}}{\sqrt{\eta_1 + \theta(1 - \eta_1)}} \leq \eta_1^{n_1 - \frac{1}{2}} (1 - \eta_1)^{n_3} \leq 1 \quad \text{if } n_1 \geq 1,$$

which proves part (b). Recalling (3.7), the posterior $\pi(\theta, \eta_1 | n)$ is therefore

$$\pi(\theta, \eta_1 | n) \propto m(n | \theta, \eta_1) \pi(\theta, \eta_1) \quad (4.16)$$

$$\propto \frac{\eta_1^{n_1+n_2-1/2}(1-\eta_1)^{n_3+n_4-1/2}\theta^{n_4-1}}{(\eta_1+\theta(1-\eta_1))^{n_2+n_4+1}} \frac{1}{H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\}$$

Finally, we integrate η_1 from (4.16) to obtain $\pi(\theta|n)$

$$\begin{aligned} \pi(\theta|n) &\propto \frac{\theta^{n_4-1}}{H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} \int_0^1 \frac{\eta_1^{n_1+n_2-1/2}(1-\eta_1)^{n_3+n_4-1/2}}{(\eta_1+\theta(1-\eta_1))^{n_2+n_4+1}} d\eta_1 \\ &\propto \frac{\theta^{-n_2-2}}{H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} \\ &\quad \times \int_0^1 \eta_1^{n_1+n_2-1/2}(1-\eta_1)^{n_3+n_4-1/2} \left(1 - \frac{\theta-1}{\theta}\eta_1 \right)^{-(n_2+n_4+1)} d\eta_1 \end{aligned}$$

Choosing $a = n_2 + n_4 + 1$, $b = n_1 + n_2 + \frac{1}{2}$, and $c = n_1 + n_2 + n_3 + n_4 + 1$, applying (3.18) yields the result. To prove (h), we can demonstrate that (4.8) is integrable at 0 and ∞ , this is sufficient to show that $\int_0^\infty \theta \pi(\theta|n) d\theta < \infty$.

As $\theta \rightarrow 0$, recall that $H(\theta) \approx -\log(\theta)$. We can again use the Pfaff transformation in (3.20) to rewrite the marginal posterior (4.8) in the form

$$\pi(\theta|n) \propto \frac{\theta^{n_4-1} {}_2F_1 \left(n_2 + n_4 + 1, n_3 + n_4 + \frac{1}{2}; \sum_{j=1}^4 n_j + 1; 1 - \theta \right)}{H(\theta)} \exp \left\{ \frac{G(\theta)}{H(\theta)} \right\} \quad (4.17)$$

Next, Note that equations 15.4.20 and 15.4.23 in Olver et al. (2010) tell us

$$\lim_{z \rightarrow 1^-} {}_2F_1(a, b; c; z) = \frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} \quad \text{if } c-a-b > 0 \text{ and} \quad (4.18)$$

$$\lim_{z \rightarrow 1^-} \frac{{}_2F_1(a, b; c; z)}{(1-z)^{c-a-b}} = \frac{\Gamma(c)\Gamma(a+b-c)}{\Gamma(a)\Gamma(b)} \quad \text{if } c-a-b < 0. \quad (4.19)$$

Considering the numerator of (4.17), $c-a-b = n_1 - n_4 - \frac{1}{2}$, and this quantity being greater or less than 0 is equivalent to considering $n_1 > n_4$ and $n_1 \leq n_4$, respectively.

Recalling that $z = \frac{\theta-1}{\theta}$ and that $z \rightarrow 1^- \equiv \theta \rightarrow \infty$, we can see that as $\theta \rightarrow \infty$ we have

$${}_2F_1\left(n_2 + n_4 + 1, n_3 + n_4 + \frac{1}{2}; n_1 + n_2 + n_3 + n_4 + 1; 1 - \theta\right) \rightarrow C \text{ if } n_1 > n_4$$

and

$${}_2F_1\left(n_2 + n_4 + 1, n_3 + n_4 + \frac{1}{2}; n_1 + n_2 + n_3 + n_4 + 1; 1 - \theta\right) \approx \theta^{n_1 - n_4 - \frac{1}{2}} \text{ if } n_1 \leq n_4.$$

Recalling the behavior of $\frac{G(\theta)}{H(\theta)}$ noted in (3.36). This tells us that for $n_1 > n_4$, we have

$$\lim_{\theta \rightarrow 0} \frac{\theta^m \pi(\theta|n)}{-\frac{\theta^{n_4 - \frac{3}{4} + m}}{\log \theta}} = 1$$

and when $n_1 \leq n_4$

$$\lim_{\theta \rightarrow 0} \frac{\theta^m \pi(\theta|n)}{-\frac{\theta^{n_1 - \frac{5}{4} + m}}{\log \theta}} = 1.$$

Since $\int_0^\varepsilon -\frac{\theta^{n_4 - \frac{3}{4} + m}}{\log \theta} d\theta < \infty$ for any $n_4 \geq 0$, $m \geq 1$ and $\int_0^\varepsilon -\frac{\theta^{n_1 - \frac{5}{4} + m}}{\log \theta} d\theta < \infty$ for any $n_1 \geq 0$, $m \geq 1$, we conclude that $\int_0^\varepsilon \theta^m \pi(\theta|n) d\theta < \infty$ with no restrictions on the data.

As $\theta \rightarrow \infty$, recall that $H(\theta) \approx \frac{\log \theta}{\sqrt{\theta}}$. Further we also know the behavior of $\frac{G(\theta)}{H(\theta)}$ from (3.36). We can now proceed similarly to the case when $\theta \rightarrow 0$ by considering the numerator of (4.8) directly and apply equations (4.18) and (4.19). In this case, $c - a - b = n_3 - n_2 - \frac{1}{2}$, and this quantity being greater or less than 0 is equivalent to considering $n_3 > n_2$ and $n_3 \leq n_2$, respectively. Recalling that $z = \frac{\theta-1}{\theta}$ and that

$z \rightarrow 1^- \equiv \theta \rightarrow \infty$, we can see that as $\theta \rightarrow \infty$ we have

$${}_2F_1\left(n_2 + n_4 + 1, n_1 + n_2 + \frac{1}{2}; n_1 + n_2 + n_3 + n_4 + 1; \frac{\theta - 1}{\theta}\right) \rightarrow C \text{ if } n_3 > n_2$$

and

$${}_2F_1\left(n_2 + n_4 + 1, n_1 + n_2 + \frac{1}{2}; n_1 + n_2 + n_3 + n_4 + 1; \frac{\theta - 1}{\theta}\right) \approx \theta^{n_2 - n_3 + \frac{1}{2}} \text{ if } n_3 \leq n_2$$

So noting the behavior of the other factors in (4.8), for $n_3 > n_2$,

$$\lim_{\theta \rightarrow 0} \frac{\theta^m \pi(\theta|n)}{\frac{1}{\theta^{n_2 + \frac{7}{4} - m} \log(\theta)}} = 1$$

Similarly, for $n_3 \leq n_2$, we have

$$\lim_{\theta \rightarrow 0} \frac{\theta^m \pi(\theta|n)}{\frac{1}{\theta^{n_3 + \frac{5}{4} - m} \log(\theta)}} = 1.$$

Now, $\int_M^\infty \frac{1}{\theta^a \log(\theta)} d\theta < \infty$ if $a > 1$, this means that $\int_M^\infty \frac{1}{\theta^{n_2 + \frac{7}{4} - m} \log(\theta)} d\theta < \infty$ for any $n_2 \geq m$ and $\int_M^\infty \frac{1}{\theta^{n_3 + \frac{5}{4} - m} \log(\theta)} d\theta < \infty$ for any $n_3 \geq m$. So, n_2 must be at least m when it is less than n_3 , and n_3 must be at least m when it is less than or equal to n_2 , so we can equivalently conclude that $\int_M^\infty \theta^m \pi(\theta|n) d\theta < \infty$ when *both* n_2 and n_3 are at least m . $\square$

It is important to note that the posteriors in Theorem 4.1.1 are expressed in a hierarchical manner. That is, we know the posterior of θ conditioned on *only* the data in the table, we know η_1 conditioned on *only* θ and the data, and so on. This rare structure allows us to study the posterior of θ using the data and none of the

nuisance parameters and will be exploited when posterior sampling is conducted. It is remarkable that even with the four parameter problem here, the only condition for all the posteriors in Theorem 4.1.1 to exist and be proper is that $n_1 \geq 1$. Generally speaking, when a statistical problem has p parameters, we need at least p observations to estimate those parameters, but in this problem we only need one value in the table! This is especially interesting when we consider that the priors for both η_3 and η_2 are improper.

4.2 Posterior Sampling and Computation

Because of the hierarchical structure of the posteriors presented in Theorem 4.1.1 (a), (b), (f), and (g), we do not need to conduct Markov Chain Monte Carlo(MCMC) to sample from the parameters. Instead, we provide an efficient Monte Carlo(MC) algorithm for obtaining posterior samples from the parameters $(\theta, \eta_1, \eta_2, \eta_3)$ in (3.1). We can conduct Monte Carlo sampling by iterating through the following steps.

1. Sample $\theta^* \sim \pi^R(\theta|n)$;
2. Sample $\eta_1^*|\theta^*; \mathbf{n} \sim \pi^R(\eta_1|\theta^*; \mathbf{n})$;
3. Sample $\eta_3^*|\theta^*, \eta_1^*; \mathbf{n} \sim \pi^R(\eta_3|\theta^*, \eta_1^*; \mathbf{n})$ by first generating a value γ^* from the distribution $\gamma^* \sim \text{Beta}(n_2 + n_4 + 1/2, n_1 + n_3)$ and then compute

$$\eta_3^* = \frac{1}{\eta_1^* + \theta^*(1 - \eta_1^*)} \cdot \frac{\gamma^*}{(1 - \gamma^*)}; \quad (4.20)$$

4. Sample

$$\eta_2^*|\theta^*, \eta_1^*, \eta_3^*; \mathbf{n} \sim \text{Gamma} \left(n_1 + n_2 + n_3 + n_4 + \frac{1}{2}, \eta_1^*(1 + \eta_3^*) + (1 - \eta_1^*)(1 + \theta^* \eta_3^*) \right).$$

For each sample, θ^* was sampled with the ratio of uniforms method (Kinderman and Monahan, 1977). Functional evaluations of the hypergeometric function were conducted by numerically evaluating the integral representation of ${}_2F_1(a, b; c; z)$ (Bateman, 1953). A sample from $\pi^R(\eta_1^*|\theta^*, n)$ can then be obtained by using the simulated value from θ^* from Step 1 and conducting ratio of uniforms again to sample from (4.7). η_3^* can then be sampled using θ^* and η_1^* , and finally η_2^* can easily be sampled by using θ^* , η_1^* , and η_3^* . This procedure can be conducted at whatever depth a practitioner desires. Of course, if one only desires samples from $\pi^R(\theta|n)$ (usually this would be the case), no other steps need to be conducted. If one desires samples of θ and η_1 , they may stop at Step 2, etc, stopping whenever they have obtained the desired number of samples. Because we are able to sample in this way, we can avoid the often computationally inefficient MCMC algorithm, which must sample sequentially from the full conditional posteriors, and must converge before analysis can be done.

4.3 Numerical Examples

4.3.1 Simulation Study

The values of the parameters in (3.1) were fixed to be $\theta^* = (.5, 1, 2)$, $\eta_1^* = (.25, .5, .75)$, $\eta_2^* = (5, 10, 20, 50)$, and $\eta_3^* = (1/2, 1, 2/1)$. For all 108 combinations of parameter values,

1000 random 2×2 tables were generated by sampling $n_i \sim Poi(\lambda_i)$ for $i = 1, 2, 3, 4$ where the λ_i were fixed by solving (3.1) for each λ_i . Quantiles of the posterior under the reference prior can easily be found using numerical integration. For comparison, these results are compared to those found using the Jeffreys prior, The results under the Laplace prior transformed to θ scale, as well as Fisher’s exact test. The details for posterior analysis using Jeffreys prior is presented in Appendix A.2. Quantiles from the transformed Laplace Prior were found using numerical integration. For Fisher’s exact test, the R function `fisher.test()` was used for the generated table. The posterior median for $\theta^*|\mathbf{n}$ under the Reference prior was compared to that of the transformed Laplace prior, Jeffreys Prior, as well as the odds ratio as computed by Fisher’s exact test by evaluating the mean absolute deviation(MAD) of the estimate compared to the true value for the various combinations of parameter values. 95% one sided frequentist coverage was also evaluated by computing

$$P(\theta^* \leq \theta_{.95}(\mathbf{n})|\theta^*, \eta_1^*, \eta_2, \eta_3^*).$$

Aggregate results across all nuisance parameters are shown in Tables 4.9 and 4.10. Deviation and coverage results aggregated across η_3 are shown in tables 4.11 - Table 4.18. These tables are interesting after recalling that from 3.2, we can see that as η_2 grows, all of the λ_j parameters tend to grow together. So we can study the effect of “sample size” on the analysis of θ . Finally, fully expanded deviation and coverage results are shown in Tables, 4.19 and 4.20, respectively.

With respect to the deviation from θ , the reference prior is clearly superior to all other candidates. Coverage is slightly below nominal in some cases, which is likely due to the posterior under the reference prior yielding medians that are smaller than

all the others, however we note that it is much closer to nominal most of the time.

4.3.2 Data Analysis

Fisher’s Tea Tasting

A classic problem in the study of experimental design is the “lady tasting tea” experiment presented in Fisher (1937). As the story goes, psychologist Muriel Bristol claimed to have the ability to tell whether tea or milk was added to a cup first. Fisher gave her eight cups, four of each variety, in random order. Additionally, she was made aware of the design, so this is an example of *both* margins of the 2×2 table being fixed. In other words, this table, shown in Table 4.1 essentially has one degree of freedom, in that if we know one value(say n_1) we know the rest of the table. This

Table 4.1: Fisher’s Tea Tasting Experiment

		Guess		
		Milk	Tea	
Poured First	Milk	n_1	$4 - n_1$	4
	Tea	$4 - n_1$	n_1	4
		4	4	8

was the keynote example for Fisher’s exact test, and to attempt to verify her ability, we would conduct the one sided hypothesis test on the odds ratio θ as $H_a : \theta > 1$. To compare this to our proposed Reference Prior, we could compute the Bayes Factor BF using the *entire* table as

$$BF = \frac{P(\theta > 1|\mathbf{n})}{P(\theta < 1|\mathbf{n})} \frac{P(\theta > 1)}{P(\theta < 1)} = \frac{P(\theta > 1|\mathbf{n})}{P(\theta < 1|\mathbf{n})}, \quad (4.21)$$

because $\int_0^1 \pi(\theta) d\theta = 0.5$. Now, noting that n_1 can take values of 0, 1, 2, 3, 4, we have the results in Table 4.2. We can see that Fisher's Exact Test is overly

Table 4.2: Comparison to Fisher's Exact Test

n_1	0	1	2	3	4
p-value for testing $H_a : \theta > 1$	1	.986	.757	.243	.014
Bayes Factor for $\theta > 1$	.0005	.031	.555	6.523	364.978

conservative here, with the only way to demonstrate her ability being to obtain a perfect result (which would still not be significant with $\alpha = .01$). On the other hand, using the Reference prior, a perfect result provides decisive evidence for her ability, and correctly identifying 6 out of 8 ($n_1 = 3$) still allows us to say that she is 6 and a half times as likely to have the ability than to not. Perhaps acknowledging the overly conservative nature of the p-value, Fisher followed up with a discussion on the virtues of repeated trials in experiments.

A Clinical Trial

Parzen et al. (2002) considered data from a randomized phase II clinical trial evaluating whether two new chemotherapy treatments differed with respect to patient responses,

	Response		
	Normal	Life-Threatening	
Treatment 1	10	1	
Treatment 2	12	2	

We compare the modified median unbiased estimator(MMUE) of Parzen et al. (2002), the odds ratio as according to Fishers exact test, as well as the posterior for θ under both Jeffreys prior and the Reference prior.

	Estimate	2.5% Quantile	97.5% Quantile
MMUE	1.53	0.1042	12.4456
Fisher's Exact Test	1.63	0.0745	107.9371
$\pi(\theta n)$ under $\pi^J(\theta)$	1.54	0.1528	23.2051
$\pi(\theta n)$ under $\pi^R(\theta)$	1.06	0.1043	12.9026

We can see that the MMUE and posterior under the proposed reference prior have the tightest and similar credible intervals.

ECMO

Over the last 20 years, Extra Corporeal Membrane Oxygenation (ECMO) therapy has become an incredibly successful treatment for individuals experiencing significant cardiac and/or pulmonary dysfunction (Makdisi and Wang, 2015). As pointed out by Ware (1989), because of the clear and significant practical benefit of ECMO therapy, ethical concerns were raised over the implementation of a clinical trial where infants would receive a clearly less effective control therapy. Because of this, even though the practice was around since 1977, Ware (1989) provided the first clinical trial to demonstrate its effectiveness. In their study, 10 patients received a control therapy, with 6 surviving, and 9 received ECMO therapy, with all surviving. Kass and Greenhouse (1989), later provided a Bayesian analysis of this problem using 84 different proper priors, some informative and some “diffuse.” Under a Binomial sampling scheme, they examined the probability that the log odds ratio δ was positive,

indicating that the ECMO therapy was an effective treatment. Ultimately, they provided justification for an informative prior which resulted in $P(\delta|data > 0) = 0.95$. In this case, their examining of the probability that δ is greater than 0 is equivalent to examining the probability that the odds ratio θ is greater than 1. Using the techniques outlined in section 4.2, we can quickly arrive at the conclusion that $P(\pi(\theta|\mathbf{n}) > 1) = 0.972$. It is remarkable to note that, using the posterior obtained from the reference prior, this result can be obtained *without* conditioning on the marginal totals as is required by the binomial sampling scheme. Additionally, based on 10,000 posterior samples for $\pi(\eta_1|\theta)$, we find that the posterior median is 0.8674 and the credible interval formed the 2.5% and 97.5% quantiles is (0.5611, .9923). Because in this case η_1 represents the proportion of survivals which are in the ECMO group, by borrowing strength from the odds ratio θ , we have further evidence that ECMO treatment improves survival.

Tables with 0 counts

As a final example, in a junior level statistics class, data were collected on the type of phone students use (Iphone or Android) as well as the operating system (Mac or Windows) of their primary computing device. The data were cross classified into the table below:

Table 4.3: Data on Cell phone Computer types

	Windows	Mac
Android	5	0
Iphone	8	15

First, note that there were no subjects who use an android phone and a mac

computer, so the frequentist odds ratio

$$\hat{\theta} = \frac{n_1 n_4}{n_2 n_3},$$

is undefined. Using the `fisher.test()` function in R gives a 95% CI for θ as $(1.29, \infty)$. 1,000,000 samples from $\pi(\theta, \eta_1 | n)$ were obtained using the procedure outlined in section 4.2. Despite the 0 count for n_2 , the marginal posterior for $\theta | n$ under the reference prior is still able to be sampled from, and yields a reasonable median of 16.97, indicating that if someone owns an Iphone, the odds of owning a Mac are about 17 times larger than if they own an Android phone. The 95% credible interval for this quantity is $(1.59, 928.39)$, which is obviously much more precise than the Fisher’s Exact Test. We can also easily compute the Bayes factor for testing $\theta > 1$ via numerical integration of the posterior, yielding 120.2424, which is very strong evidence in favor of a relationship between phone type and computer type. It is also interesting to note the very large posterior mean for θ of 248.62, this is due to the condition for a finite first moment in Theorem 4.1.1 part (h) not being satisfied ($n_2 = 0$). If n_2 is artificially adjusted to be 1 in this problem, the posterior mean immediately stabilizes to a much more reasonable value of 12.53 (median 5.77). Furthermore, the posterior distributions for η_1 , η_2 , and η_3 are shown in Figure 4.2. Based on Table 4.3, the reported estimates in Figure 4.2 for η_1 and η_2 seem reasonable. From the posterior of η_3 we can see that, even though we did not observe any Android users with a Mac computer, we can infer that they are 11% as likely to own one. It is also interesting to note that a histogram of log transformed posterior samples of θ is remarkably symmetric (Figure 4.3). This will most likely be an area of future study. Finally, the CUDA C implementation of the posterior sampling for this dataset yielded

phenomenal speedups. Computation time in R was 40644.67 seconds(over 11 hours!), a C program completed in 481.68 seconds, and the CUDA implementation finished in .0901 seconds. This is a concrete example of a statistical application showing the tremendous potential that GPU computing could have on the field of Statistics.

4.4 A Note on the Requirement $n_1 \geq 1$

Recall from the properties of Theorem 4.1.1 that the only condition needed to guarantee the existence *and propriety* of all posteriors is that $n_1 \geq 1$. This result is especially interesting because the conditional priors for η_3 and η_2 are both improper, but we only need to observe one count to have a proper posterior. Table 4.7 is an example of a table where $n_1 = 1$ and all other table entries $n_2 = n_3 = n_4 = 0$. We note that the results for the reference prior in Table 4.5 are very reasonable.

Table 4.4: A “Table” with one value

1	0
0	0

Because this table has virtually no information, the posterior median of the odds ratio θ is very close to 1, but slightly greater than 1 due to the entry being on the diagonal. We similarly examine adding 0.1 to all cells as well. Furthermore, the 95% credible interval is *very* wide, which reinforces the fact that we have almost no information. On the other hand, the “standard” of adding 0.5 to all cells yields a very unreasonable result, with .5 introducing far too much information and artificially pushing the confidence interval towards 1. When we add 0.1 to all cells instead, the estimate becomes even more unreasonable and the upper bound of the confidence

Table 4.5: Comparison of Techniques

	Estimate	95% CI
Fishers Exact Test	∞	$(0, \infty)$
Add 0.5 to all cells	3	$(0.019, 473.058)$
Add 0.1 to all cells	11	$(0.0002, 593, 920.3)$
Reference Bayesian Analysis	1.064	$(0.000105, 151, 297)$

interval is much larger. For ease of interpretation, Table 4.6 shows the results of Table

Table 4.6: Comparison of Techniques on log-scale

	Estimate	95% CI
Fishers Exact Test	∞	$(-\infty, \infty)$
Add 0.5 to all cells	1.099	$(-3.962, 6.159)$
Add 0.1 to all cells	2.398	$(-8.499, 13.295)$
Reference Bayesian Analysis	.062	$(-9.161, 11.927)$

4.5 on a log scale. This reinforces the reasonable values obtained by the Reference Prior, as the posterior credible interval is roughly symmetric but shifted slightly to the right of 0, again because the only non-zero entry is on the diagonal. The condition $n_1 > 1$ is in fact the minimal possible restriction on the dataset. However, if $n_1 = 0$ but some other $n_j \geq 0, j \neq 1$, we can simply permute the table and adjust θ as needed. These simple permutations show that the methodology can be applied to *any* table with at least one count.

- Case 1: $n_4 = 1, \theta = \frac{\lambda_1 \lambda_4}{\lambda_2 \lambda_3} = \frac{\lambda_4 \lambda_1}{\lambda_3 \lambda_2}$ is the same.

$$\begin{array}{|c|c|} \hline 0 & 0 \\ \hline 0 & 1 \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 0 & 0 \\ \hline \end{array}$$

- Case 2: $n_3 = 1, \xi = \frac{1}{\theta} = \frac{\lambda_2 \lambda_3}{\lambda_1 \lambda_4}$.

0	0
1	0

 $\Rightarrow$

1	0
0	0

For this case, we can compute the posterior of ξ as

$$\pi^R(\xi|\mathbf{n}) \propto \frac{{}_2F_1\left(n_2 + n_4 + 1, n_3 + n_4 + \frac{1}{2}; \sum_{j=1}^4 n_j + 1; \frac{\xi-1}{\xi}\right)}{\xi^{n_4+1} H(\xi)} \exp\left\{\frac{G(\xi)}{H(\xi)}\right\}.$$

- Case 3: $n_2 = 1$, conduct the permutations for case 1 and 2.

0	1
0	0

 $\Rightarrow$

0	0
0	1

 $\Rightarrow$

1	0
0	0

Of course, when an open source software package implementing this structure is developed, this procedure can be conducted automatically.

4.4.1 One large value in only one cell

In a conversation with Scott Holan, a faculty member of the MU stat department, it was discussed that a table with only $n_1 = 1$ contains basically no information, and that the results shown in Table 4.6 may be the result of the posterior “falling back” on the prior. To study this, consider Table 4.7, which an example of a table where $n_1 = n$ and all other table entries $n_2 = n_3 = n_4 = 0$. The results, shown in

Table 4.7: A “Table” with one value

n	0
0	0

Table 4.8, show a very curious pattern indeed. First, note that the width of a 95%

confidence interval on the prior $\pi^R(\theta)$ is around 22.6 on log scale. When we increase n to 1, as was the case in Table 4.6, we can see the width of the posterior confidence interval is 21.089 on log scale. So, observing *only* $n_1 = 1$ makes our knowledge about θ to be $e^{22.602-21.089} \approx 4.5$ times more precise. However, note that as c is increased, the width of the credible intervals stay relatively similar on log scale! The intuition behind this curious result presents itself if we recall the elementary interpretation of an odds ratio in a 2×2 table. The odds ratio θ allows us to quantify the relationship between the row and column variables of this table. If we have only observed counts for one combination of table variables, conceptually it doesn't matter how many we have observed, because we have no information across any of the other categories to assess this potential relationship. Plots of these posteriors are shown in Figure 4.4. The following theorem provides a general result as the n in Table 4.7.

Table 4.8: Log-scale Posterior results for θ for $\pi^R(\theta)$ as well as $\pi^R(\theta|n)$ for various n

	Median	95% CI	Width
$\pi^R(\theta)$	0	(−11.301, 11.301)	22.602
$n = 1$	.062	(−9.161, 11.927)	21.089
$n = 10$	2.711	(−8.119, 13.338)	21.456
$n = 100$	4.809	(−6.247, 15.682)	21.928
$n = 1000$	6.945	(−4.564, 17.964)	22.527
$n = 10000$	8.857	(−3.113, 19.277)	22.389
$n \rightarrow \infty$	9.287	(−1.912, 20.686)	22.597

Theorem 4.4.1. Consider Table 4.7. We have the following asymptotic behavior for the marginal posterior $\pi(\theta|n)$ in (4.8).

(a)

$$\lim_{n \rightarrow \infty} \pi^R(\theta|\mathbf{n}) = \frac{1}{0.284} \frac{\pi^R(\theta)}{H(\theta)}, \theta \in (0, 1) \quad (4.22)$$

(b)

$$\lim_{n \rightarrow \infty} \pi^R(\theta|\mathbf{n}) = \frac{1}{0.284} \frac{\pi^R(\theta)}{\sqrt{\theta} H(\theta)}, \theta \in [1, \infty) \quad (4.23)$$

Proof. For (a), substitution of $n_1 = n$ and $n_2 = n_3 = n_4 = 0$ into (4.8) gives

$$\pi^R(\theta|\mathbf{n}) \propto \frac{{}_2F_1\left(1, n + \frac{1}{2}; n + 1; \frac{\theta-1}{\theta}\right)}{\theta^2 H(\theta)} \exp\left\{\frac{G(\theta)}{H(\theta)}\right\}.$$

Now, the behavior of the hypergeometric function ${}_2F_1(a, b; c; z)$ for large parameters b and c has been studied at length in the mathematical literature since Watson (1918) first explored them 100 years ago. The most complete and approachable modern explanation appears in Temme (2003), who demonstrates that

$${}_2F_1(a, b + n; c + n; z) = (1 - z)^{-a} \left[1 + \frac{a(c - b)}{c + n} \zeta + \mathcal{O}(n^{-2}) \right],$$

as $\lambda \rightarrow \infty$ with $\zeta = \frac{z}{z-1}$ fixed and $|\zeta| < 1$ being required for series convergence.

Noting $a = 1$, $b = 1/2$, $c = 1$, $1 - z = 1/\theta$, and $\zeta = 1 - \theta$, we have

$${}_2F_1\left(1, n + \frac{1}{2}; n + 1; \frac{\theta - 1}{\theta}\right) = \theta \left[1 + \frac{1 - \theta}{2(1 + n)} + \mathcal{O}(n^{-2}) \right]. \quad (4.24)$$

Now, since $|1 - \theta| < 1$, the leading term of (4.24) is simply θ , so we therefore have

$$\begin{aligned} \pi^R(\theta|\mathbf{n}) &\propto \frac{{}_2F_1\left(1, n + \frac{1}{2}; n + 1; \frac{\theta-1}{\theta}\right)}{\theta^2 H(\theta)} \exp\left\{\frac{G(\theta)}{H(\theta)}\right\} \\ &\approx \frac{\theta}{\theta^2 H(\theta)} \exp\left\{\frac{G(\theta)}{H(\theta)}\right\} \\ &= \frac{1}{\theta H(\theta)} \exp\left\{\frac{G(\theta)}{H(\theta)}\right\} = \frac{\pi^R(\theta)}{H(\theta)} \end{aligned}$$

For (b), we can conduct the transformation $\xi = \frac{1}{\theta}$ with Jacobian $J = \xi^{-2}$ to see

$$\pi^R(\xi|\mathbf{n}) \propto \frac{{}_2F_1\left(1, n + \frac{1}{2}; n + 1; 1 - \xi\right)}{H(\xi)} \exp\left\{\frac{G(\xi)}{H(\xi)}\right\}. \quad (4.25)$$

From the definition of the hypergeometric function in (3.17), we have

$${}_2F_1\left(1, n + \frac{1}{2}; n + 1; 1 - \xi\right) = \sum_{s=0}^{\infty} \frac{(1)_s (n + 1/2)_s}{(n + 1)_s s!} (1 - \xi)^s = \sum_{s=0}^{\infty} \frac{(n + 1/2)_s}{(n + 1)_s} (1 - \xi)^s.$$

Expanding this sum, we can see

$$\sum_{s=0}^{\infty} \frac{(n + 1/2)_s}{(n + 1)_s} (1 - \xi)^s = 1 + \frac{n + \frac{1}{2}}{n + 1} (1 - \xi) + \frac{(n + \frac{1}{2})(n + \frac{3}{2})}{(n + 1)(n + 2)} (1 - \xi)^2 + \dots$$

Therefore, as $n \rightarrow \infty$, by the Dominated Convergence Theorem, we can see that

$$\begin{aligned} {}_2F_1\left(1, n + \frac{1}{2}; n + 1; 1 - \xi\right) &= 1 + (1 - \xi) + (1 - \xi)^2 + \dots \\ &= \frac{1}{1 - (1 - \xi)} = \frac{1}{\xi} \end{aligned}$$

Substituting this into (4.25) we finally get

$$\pi^R(\xi|\mathbf{n}) \propto \frac{1}{\xi H(\xi)} \exp\left\{\frac{G(\xi)}{H(\xi)}\right\} = \frac{\pi^R(\xi)}{H(\xi)}. \quad (4.26)$$

which is the same as (4.23) after transforming back to θ space. $\square$

A particularly interesting interpretation about this result is that (4.22) and (4.26) together indicate that the posterior densities of $\theta|n$ and $\frac{1}{\theta}|n$ are the *identical* if we have only observed one very large value in the table.

4.5 Software Implementations

As discussed in section 1.5, it is the author’s firm belief that robust and efficient software should be developed and provided with any developed objective Bayesian methodology. This is in line with the “default” nature of this approach to statistics and allows applied researchers to easily use the results of research that take, in many cases, years to develop. Here, we present the two key developments with respect to the software implementations of the work in this chapter, an easy to use software package downloadable from within R using the “devtools” package as well as a cutting edge approach to posterior sampling using a Graphics Processing Unit(GPU).

4.5.1 R Package: OBayes2by2Table

The program “R” (<https://www.r-project.org/>) is a programming language used for data based computing which is widely used by statisticians, data scientists and applied researchers from many disciplines. Its high and growing popularity likely depends on its ease of use, high level of community support, and ability produce publication-quality graphics. Furthermore, a large quantity of downloadable libraries are available to extend its functionality in a seemly unlimited number of ways. R is freely available under the GNU General Public License, and easily installable software is provided for most operating systems. At its core, R is a command line interface, however several graphical user interfaces are available. The most popular of them being “RStudio”(<https://www.rstudio.com/>), which adds, among many other things, syntax highlighting and other “quality of life” upgrades to help make coding much easier. Here, we detail the install process of the developed R package for

the methodology of this chapter, and walk through several examples on usage. Code will be provided here.

Package Installation and Use

The initial release of this software is on “GitHub,” which is a hosting service for computer code that uses the protocol “git” for version control. Installing the software can be accomplished easily using the R package “devtools.”

```
> install.packages("devtools")
```

Once the devtools library finishes installing, we can load it and connect to the author’s github repository to download and install the package.

```
> library(devtools)
> devtools::install_github("John-Snyder/OBayes2by2Table")
```

This will install the package and all its dependencies. Now, say we would like to conduct analysis on the data in Table 4.3. We can load the software, load the data into R, and use the “OR_Ref” function to conduct analysis.

```
> library(OBayes2by2Table)
> MyTable <- rbind(c(5,0),
                  c(8,15))
> dimnames(MyTable) <- list(Phone = c("Android", "Iphone"),
Computer = c("Windows", "Mac"))
```

```

> MyTable
Computer
Phone      Windows Mac
Android      5      0
Iphone      8      15

> Ref_Results <- OR_Ref(MyTable, conf.int = TRUE,
post.sample = FALSE)
Computing CI via numerical integration...
2.5% quantile finished
97.5% quantile finished

>
> Ref_Results$CI
[1] 1.590054 1049.888803

```

Several options are currently available, with more on the horizon. In the preceding example, we can see that an exact equal tailed $100(1 - \alpha)\%$ credible interval can be obtained via numerical integration. If the “post.sample” argument is set to TRUE, we can also conduct posterior sampling. Because of the hierarchical nature of the sampling setup of this problem, we can control the depth of the sampling. This depth is adjusted by the user via the “sampling.depth” argument. This argument controls the sampling in a way that is identical to the 4 step process outlined in section 4.2. If we set this argument to 1, we will only sample θ , if we set it to 2, we will sample θ and η_1 , etc. When posterior sampling is conducted, messages and a progress bar informs us on the status of the Monte Carlo samplers. For example:

```

> Ref_Results <- OR_Ref(MyTable, conf.int = FALSE,
  post.sample = TRUE,
  num.samples = 1000,
  sampling.depth = 4)
Generating 1000 posterior samples from theta | data
=====| 100%
Generating 1000 posterior samples from eta1 | theta, data
=====| 100%
Generating 1000 posterior samples from eta3 | theta, eta1, data
Generating 1000 posterior samples from eta2 |.
> quantile(Ref_Results$Theta.Samples, c(.025, .5, .975))
2.5%          50%          97.5%
1.776729    15.968372   808.415745
> quantile(Ref_Results$Eta1.Samples, c(.025, .5, .975))
2.5%          50%          97.5%
0.1439363   0.3607117   0.6394058
> quantile(Ref_Results$Eta2.Samples, c(.025, .5, .975))
2.5%          50%          97.5%
6.871325    12.665737   20.440763
> quantile(Ref_Results$Eta3.Samples, c(.025, .5, .975))
2.5%          50%          97.5%
0.002471525 0.117761900 0.934028943

```

The posterior samples in the “Ref_Results\$Theta.Samples” object can be used in many ways. We could compute quantiles as in the code example, posterior odds of

$\theta > 1$, make plots, etc.

Currently, this software implementation is usable by researchers, but many significant improvements are planned. Options will be enhanced, with the user being able to specify arbitrary posterior quantiles, conduct Bayesian model comparison with Bayes factors, and automatically display plots of the posterior. Performance will be improved as well. Currently, the source of the package is coded completely in R. C modules for core functions have been developed, and, in the very near future will be integrated into the production version of the software. This will take computation time from a minute or so to becoming nearly instant. Finally, accessibility will improve as well. After the integration of the C modules, the package will be obtainable *without* the devtools package but rather from R's CRAN repository. When this happens, a user could simply type the following into the R console.

```
install.packages("OBayes2by2Table")
```

For the current list of options available, interested parties may install the package and examine the help menu by typing `?OR_Ref` into the R console window. The current version of this software is released freely under the GNU General Public License(<https://www.gnu.org/licenses/gpl.html>), making the source and all potential derivatives of it available to the public.

4.5.2 GPU Computation

Since many θ^* can be sampled independently, this structure presented the potential to be computed in a parallelized way. The posterior sampling algorithm was implemented in CUDA, a set of extensions for C used for programming directly on

Nvidia GPUs. The structure of the program breaks the problem into a number of parallel “blocks,” and within each block, we can run a loop to generate a small number of θ s, and if we desire, use this value to progress through the other parameters as shown above. This structure took posterior computation time for a test dataset from around 5 minutes in R to under 0.1 seconds on the GPU.

Computation on the GPU Device

Driven by the tremendous demand for high-definition 3D graphics by the video game industry, the Graphics Processing Unit(GPU) has become a highly parallel processor with many computing units that has a mind-boggling amount of computational power. While a standard computer CPU may have 2, 4, or even 8 cores, modern GPUs often have several thousand cores capable of computing many simple mathematical computations in parallel.

CUDA is a set of extensions for the C language which allows programmers to use General-Purpose computing on Graphics Processing Units(GPGPU) to conduct scientific computation on Nvidia graphics cards. To program in CUDA, we write *kernels* which execute on the GPU. When a CUDA program calls a kernel, a collection of “blocks” each of which has a programmer specified number of threads is launched. The kernel is then launched on these blocks in parallel and loop through the threads however the programmer specifies. For our problem, each of the parallel blocks is built to sample up to the four parameters in sequence, depending on the sampling depth the user would like.

Figure 4.1 illustrates the structure of the parallel blocks. The number of blocks that can run at once is hardware dependent, and there is significant room for

Figure 4.1: Illustration of a CUDA Block

optimization.

Structure of the CUDA program GPUs provide programmers a mind boggling amount of computational power. But what we gain in power, we lose in flexibility due to the architecture of the GPU being completely different than CPUs. Because of this, the many years of libraries available for C are unavailable for GPU computing. This became an issue when we desired to evaluate the marginal posterior in (4.8) as well as $H(\theta)$ and $G(\theta)$. For evaluations of the Gauss hypergeometric function in $H(\theta)$ and (4.8), the source code for the function “gsl_sf_hyp2F1” from the GNU scientific library(<https://www.gnu.org/software/gsl/>) was ported to CUDA. Computation of $G(\theta)$ was less straightforward, as porting a numerical integration routine to handle the singularities in the integrand was not feasible. Fortunately, as shown in Appendix A.3, the functional form of the integrand in $G(\theta)$ can be approximated very well for $t \approx 0$ and $t \approx 1$. After the development of the CUDA program, R’s .Call() function is evoked to call the external function which is compiled as a shared-object(*.so). This structure is similar to how most functions in base R are developed. Generally, the R user will have access a high level function

which takes in data, performs various checks and processing steps, and finally submits it to an external precompiled function using `.Call()`. In the future, this software will be integrated into the “`OBayes2by2Table`” R package via a GPU option if needed.

4.6 Discussion and Future Work

In this work, we provided an Objective Bayesian structure to conduct analysis on the odds ratio of a 2×2 table using reference priors. The priors have a large number of very intuitive and elegant properties, lending credibility to this methodology being a reasonable answer to the 2×2 contingency table problem. We also demonstrated that under a robust simulation, its performance exceeded those of Frequentist methods as well as the Jeffreys prior and Laplace prior. Furthermore, the provided real data examples demonstrate that the results of this structure provide reasonable and more precise answers to the proposed questions. What’s more, all of the real data examples were traditionally analyzed using methodologies which assumed some conditioning on the 2×2 table, so we have demonstrated that we can use the reference prior to obtain more precise analysis even when we do not take advantage of our knowledge of the totals. This has profound implications for sampling, where in the real world often times issues arise that suggest we can never truly know the totals. An easy R-package was also provided and will be updated continuously and supported long term.

In the future, using the main results of Datta and Ghosh (1996), this work can be extended to consider transformations of the joint reference prior. Because the odds ratio θ is the first (most interesting) parameter in our ordering, the reference prior for any 1 to 1 function $g(\theta)$ can be trivially obtained. We also feel that it may

be interesting to try to transform to some other parameterizations of the nuisance parameters. For example, we may consider transforming η_1 as

$$\eta_1^* = g_2(\theta, \eta_1) = \theta \frac{1 - \eta_1}{\eta_1} = \frac{\lambda_4}{\lambda_2}, \quad \text{or} \quad g_2(\theta, \eta_1) = 1 - \eta_1 = \frac{\lambda_3}{\lambda_1 + \lambda_3}.$$

η_3 could also be transformed to have the space $(0, 1)$ by

$$\eta_3^* = g_4(\theta, \eta_1, \eta_2, \eta_3) = \frac{1}{1 + \eta_3} = \frac{1}{1 + \frac{\lambda_2}{\lambda_1}} = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Other parameterizations may be related to the sum of the entire table, such as changing η_2 to $\eta_2^* = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$. A next step in the development of Bayesian theory for 2×2 contingency tables would be to derive the reference prior for the odds ratio of a 2×2 table under multinomial sampling. This would be a major step in completing the story of this simply stated data structure.

4.7 Figures and Tables

Figure 4.2: Posterior samples of parameters for the Cell phone data

Figure 4.3: Histogram of 1,000,000 samples from $\pi(\theta|\mathbf{n})$

Figure 4.4: Reference Posteriors of $\xi = \frac{1}{\theta}$ with only one cell populated by a large n value

θ	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.50	0.4170	0.4921	0.5165	0.5256
1.00	0.7100	0.8924	0.8237	0.8357
2.00	1.3162	1.5826	1.4840	1.4944

Table 4.9: Mean Absolute Deviation

θ	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.50	0.9589	0.9719	0.9898	0.9875
1.00	0.9513	0.9701	0.9876	0.9860
2.00	0.9470	0.9685	0.9835	0.9842

Table 4.10: 95% Frequentist Coverage

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.4438	0.5204	0.6459	0.6394
0.5	0.5	0.5264	0.6734	0.7382	0.7292
0.5	0.75	0.6602	0.8798	0.9253	0.895
1	0.25	0.8484	0.9832	1.1461	1.0996
1	0.5	0.8747	1.0932	1.0701	1.0327
1	0.75	0.8627	1.1408	1.0583	1.0129
2	0.25	1.6471	1.7957	2.1171	1.9894
2	0.5	1.7756	2.1182	2.1063	1.9874
2	0.75	1.4379	1.5959	1.5599	1.479

Table 4.11: Mean Absolute Deviation: $\eta_2 = 5$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.988	0.9973	0.992	0.998
0.5	0.5	0.9813	0.9967	0.989	0.9987
0.5	0.75	0.988	0.9993	0.9917	0.9997
1	0.25	0.9757	0.994	0.9897	1
1	0.5	0.9703	0.991	0.9903	0.9993
1	0.75	0.9757	0.9927	0.9883	0.999
2	0.25	0.97	0.9897	0.9893	0.998
2	0.5	0.95	0.975	0.9813	0.9983
2	0.75	0.9613	0.9833	0.9863	0.9977

Table 4.12: 95% Frequentist Coverage: $\eta_2 = 5$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.3925	0.4086	0.4894	0.5125
0.5	0.5	0.5058	0.5468	0.6224	0.6357
0.5	0.75	0.6681	0.7274	0.8017	0.8097
1	0.25	0.7474	0.803	0.91	0.9395
1	0.5	0.881	1.0263	1.0332	1.0511
1	0.75	1.0201	1.1854	1.1299	1.1504
2	0.25	1.5109	1.6067	1.8383	1.8836
2	0.5	1.6367	1.8861	1.8602	1.8652
2	0.75	1.7583	1.9867	1.8702	1.8942

Table 4.13: Mean Absolute Deviation: $\eta_2 = 10$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.9587	0.986	0.9827	0.9967
0.5	0.5	0.9493	0.982	0.966	0.9887
0.5	0.75	0.973	0.9957	0.9767	0.9967
1	0.25	0.937	0.9797	0.9777	0.9963
1	0.5	0.939	0.9813	0.9623	0.9883
1	0.75	0.9657	0.989	0.977	0.9937
2	0.25	0.937	0.9797	0.973	0.9927
2	0.5	0.942	0.9743	0.9633	0.9863
2	0.75	0.9617	0.9867	0.976	0.9913

Table 4.14: 95% Frequentist Coverage: $\eta_2 = 10$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.317	0.3549	0.3713	0.4003
0.5	0.5	0.333	0.3878	0.3733	0.3909
0.5	0.75	0.4976	0.5685	0.5381	0.5608
1	0.25	0.6243	0.759	0.7318	0.7857
1	0.5	0.6247	0.821	0.695	0.7265
1	0.75	0.8181	1.0325	0.8379	0.8811
2	0.25	1.2167	1.4538	1.4333	1.5408
2	0.5	1.1156	1.5002	1.2271	1.2723
2	0.75	1.4825	1.9076	1.4771	1.5644

Table 4.15: Mean Absolute Deviation: $\eta_2 = 20$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.9257	0.9777	0.964	0.989
0.5	0.5	0.9517	0.985	0.963	0.9797
0.5	0.75	0.96	0.9867	0.9633	0.9847
1	0.25	0.9347	0.977	0.9627	0.9837
1	0.5	0.9427	0.9853	0.9597	0.979
1	0.75	0.9443	0.986	0.9563	0.9827
2	0.25	0.9367	0.9753	0.9633	0.9857
2	0.5	0.9437	0.9833	0.962	0.9757
2	0.75	0.9353	0.9793	0.949	0.9757

Table 4.16: 95% Frequentist Coverage: $\eta_2 = 20$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.2141	0.2656	0.2325	0.2489
0.5	0.5	0.1975	0.2556	0.2054	0.2163
0.5	0.75	0.248	0.3167	0.2541	0.2685
1	0.25	0.413	0.6046	0.4495	0.4813
1	0.5	0.3555	0.566	0.3679	0.3855
1	0.75	0.4506	0.6935	0.4549	0.4816
2	0.25	0.801	1.067	0.8758	0.9381
2	0.5	0.6441	1.0439	0.6693	0.7048
2	0.75	0.7677	1.0293	0.7738	0.8137

Table 4.17: Mean Absolute Deviation: $\eta_2 = 50$

θ	η_1	$\pi^R(\theta \mathbf{n})$	$\pi^L(\theta \mathbf{n})$	$\pi^J(\theta \mathbf{n})$	Fisher
0.5	0.25	0.943	0.9883	0.9577	0.9723
0.5	0.5	0.9417	0.9927	0.9587	0.9707
0.5	0.75	0.946	0.9897	0.958	0.9757
1	0.25	0.943	0.9877	0.957	0.969
1	0.5	0.9403	0.995	0.96	0.9687
1	0.75	0.947	0.9923	0.96	0.9727
2	0.25	0.9367	0.988	0.9557	0.967
2	0.5	0.945	0.9937	0.9627	0.9697
2	0.75	0.9447	0.9933	0.9603	0.9727

Table 4.18: 95% Frequentist Coverage: $\eta_2 = 50$

			η_3											
θ	η_1	η_2	$\pi^R(\theta)$			$\pi^J(\theta)$			$\pi^L(\theta)$			Fishers Exact Test		
			.5	1	2	.5	1	2	.5	1	2	.5	1	2
0.5	0.25	5	0.382	0.399	0.550	0.493	0.581	0.864	0.412	0.472	0.677	0.492	0.584	0.842
0.5	0.25	10	0.356	0.394	0.427	0.420	0.514	0.534	0.352	0.412	0.462	0.439	0.545	0.553
0.5	0.25	20	0.336	0.333	0.282	0.411	0.397	0.307	0.358	0.378	0.329	0.453	0.427	0.321
0.5	0.25	50	0.267	0.209	0.166	0.305	0.222	0.171	0.323	0.259	0.215	0.333	0.234	0.180
0.5	0.50	5	0.540	0.540	0.499	0.858	0.756	0.600	0.715	0.674	0.631	0.833	0.739	0.616
0.5	0.50	10	0.539	0.561	0.417	0.748	0.670	0.450	0.575	0.599	0.466	0.753	0.680	0.474
0.5	0.50	20	0.400	0.335	0.265	0.483	0.363	0.274	0.450	0.391	0.322	0.508	0.377	0.288
0.5	0.50	50	0.234	0.201	0.158	0.248	0.207	0.161	0.291	0.258	0.217	0.261	0.217	0.171
0.5	0.75	5	0.871	0.678	0.432	1.376	0.932	0.468	1.169	0.902	0.568	1.290	0.903	0.492
0.5	0.75	10	0.968	0.599	0.437	1.301	0.674	0.430	1.036	0.665	0.481	1.249	0.705	0.475
0.5	0.75	20	0.688	0.454	0.351	0.807	0.464	0.343	0.775	0.515	0.416	0.812	0.495	0.376
0.5	0.75	50	0.304	0.242	0.198	0.319	0.246	0.198	0.380	0.310	0.260	0.335	0.260	0.211
1.0	0.25	5	0.767	0.775	1.003	0.822	1.030	1.586	0.750	0.878	1.321	0.777	0.995	1.526
1.0	0.25	10	0.707	0.747	0.789	0.787	0.945	0.998	0.689	0.798	0.921	0.811	0.985	1.023
1.0	0.25	20	0.638	0.649	0.586	0.772	0.769	0.654	0.704	0.795	0.778	0.851	0.822	0.684
1.0	0.25	50	0.482	0.409	0.348	0.552	0.437	0.360	0.661	0.603	0.550	0.603	0.462	0.379
1.0	0.50	5	0.825	0.887	0.912	1.024	1.128	1.059	0.970	1.109	1.201	0.961	1.075	1.063
1.0	0.50	10	0.939	0.872	0.833	1.210	1.009	0.881	1.045	1.024	1.010	1.207	1.024	0.922
1.0	0.50	20	0.785	0.604	0.485	0.931	0.653	0.501	0.982	0.797	0.684	0.977	0.677	0.526
1.0	0.50	50	0.422	0.348	0.296	0.444	0.360	0.299	0.617	0.571	0.510	0.464	0.379	0.314
1.0	0.75	5	0.924	0.883	0.782	1.337	1.049	0.788	1.316	1.159	0.948	1.225	1.007	0.807
1.0	0.75	10	1.290	1.004	0.766	1.603	1.048	0.738	1.477	1.173	0.906	1.537	1.099	0.815
1.0	0.75	20	1.006	0.766	0.683	1.102	0.759	0.653	1.243	0.953	0.902	1.118	0.805	0.720
1.0	0.75	50	0.535	0.437	0.380	0.548	0.438	0.379	0.760	0.670	0.650	0.573	0.464	0.408
2.0	0.25	5	1.423	1.550	1.968	1.328	2.061	2.962	1.218	1.696	2.473	1.240	1.934	2.794
2.0	0.25	10	1.326	1.429	1.777	1.461	1.825	2.229	1.251	1.513	2.057	1.495	1.890	2.266
2.0	0.25	20	1.178	1.381	1.091	1.463	1.676	1.161	1.307	1.701	1.354	1.632	1.788	1.203
2.0	0.25	50	0.955	0.808	0.640	1.096	0.878	0.654	1.096	1.054	1.052	1.199	0.932	0.683
2.0	0.50	5	1.554	1.781	1.992	1.751	2.173	2.395	1.618	2.106	2.631	1.605	2.022	2.335
2.0	0.50	10	1.655	1.718	1.537	1.976	1.966	1.638	1.773	1.990	1.895	1.940	1.959	1.697
2.0	0.50	20	1.268	1.125	0.954	1.492	1.202	0.987	1.617	1.506	1.378	1.553	1.229	1.035
2.0	0.50	50	0.756	0.614	0.563	0.804	0.632	0.572	1.059	1.033	1.040	0.844	0.664	0.606
2.0	0.75	5	1.466	1.530	1.317	1.691	1.708	1.280	1.593	1.731	1.464	1.531	1.610	1.297
2.0	0.75	10	2.110	1.741	1.424	2.490	1.775	1.346	2.359	1.963	1.638	2.367	1.832	1.484
2.0	0.75	20	1.668	1.414	1.365	1.750	1.392	1.289	2.052	1.823	1.847	1.781	1.481	1.431
2.0	0.75	50	0.924	0.724	0.655	0.944	0.727	0.650	1.067	1.036	0.985	0.979	0.764	0.697

Table 4.19: Simulation Results: Fully Expanded Mean Absolute Deviations

			η_3											
θ	η_1	η_2	$\pi^R(\theta)$			$\pi^J(\theta)$			$\pi^L(\theta)$			Fishers Exact Test		
			.5	1	2	.5	1	2	.5	1	2	.5	1	2
0.5	0.25	5	0.993	0.991	0.980	0.997	0.997	0.982	0.999	0.999	0.994	0.998	1.000	0.996
0.5	0.25	10	0.969	0.958	0.949	0.994	0.984	0.970	0.994	0.986	0.978	1.000	0.999	0.991
0.5	0.25	20	0.945	0.915	0.917	0.970	0.962	0.960	0.978	0.976	0.979	0.994	0.984	0.989
0.5	0.25	50	0.942	0.942	0.945	0.953	0.959	0.961	0.986	0.987	0.992	0.979	0.969	0.969
0.5	0.50	5	0.998	0.977	0.969	0.998	0.990	0.979	1.000	0.998	0.992	1.000	0.997	0.999
0.5	0.50	10	0.976	0.944	0.928	0.983	0.959	0.956	0.992	0.980	0.974	0.995	0.985	0.986
0.5	0.50	20	0.965	0.952	0.938	0.968	0.964	0.957	0.987	0.988	0.980	0.982	0.980	0.977
0.5	0.50	50	0.950	0.929	0.946	0.964	0.951	0.961	0.992	0.991	0.995	0.982	0.962	0.968
0.5	0.75	5	0.996	0.990	0.978	0.993	0.988	0.994	1.000	1.000	0.998	1.000	1.000	0.999
0.5	0.75	10	0.976	0.975	0.968	0.971	0.980	0.979	0.996	0.998	0.993	0.995	0.998	0.997
0.5	0.75	20	0.968	0.966	0.946	0.961	0.968	0.961	0.988	0.984	0.988	0.986	0.983	0.985
0.5	0.75	50	0.959	0.950	0.929	0.964	0.966	0.944	0.992	0.993	0.984	0.985	0.980	0.962
1.0	0.25	5	0.974	0.974	0.979	0.998	0.992	0.979	0.996	0.991	0.995	1.000	1.000	1.000
1.0	0.25	10	0.929	0.928	0.954	0.981	0.975	0.977	0.976	0.975	0.988	0.998	0.997	0.994
1.0	0.25	20	0.951	0.931	0.922	0.977	0.956	0.955	0.983	0.972	0.976	0.992	0.984	0.975
1.0	0.25	50	0.951	0.948	0.930	0.958	0.957	0.956	0.985	0.988	0.990	0.974	0.969	0.964
1.0	0.50	5	0.985	0.960	0.966	0.999	0.984	0.988	0.998	0.983	0.992	1.000	0.999	0.999
1.0	0.50	10	0.958	0.930	0.929	0.975	0.956	0.956	0.989	0.974	0.981	0.995	0.983	0.987
1.0	0.50	20	0.947	0.938	0.943	0.956	0.960	0.963	0.991	0.988	0.977	0.986	0.981	0.970
1.0	0.50	50	0.944	0.936	0.941	0.959	0.960	0.961	0.992	0.996	0.997	0.973	0.968	0.965
1.0	0.75	5	0.995	0.981	0.951	0.995	0.994	0.976	0.998	0.998	0.982	1.000	1.000	0.997
1.0	0.75	10	0.982	0.959	0.956	0.983	0.975	0.973	0.997	0.986	0.984	0.998	0.992	0.991
1.0	0.75	20	0.960	0.941	0.932	0.971	0.954	0.944	0.991	0.985	0.982	0.988	0.984	0.976
1.0	0.75	50	0.947	0.949	0.945	0.962	0.958	0.960	0.995	0.991	0.991	0.980	0.968	0.970
2.0	0.25	5	0.956	0.979	0.975	0.998	0.993	0.977	0.983	0.995	0.991	1.000	0.999	0.995
2.0	0.25	10	0.928	0.941	0.942	0.982	0.977	0.960	0.973	0.980	0.986	0.998	0.997	0.983
2.0	0.25	20	0.949	0.926	0.935	0.975	0.949	0.966	0.975	0.973	0.978	0.990	0.980	0.987
2.0	0.25	50	0.948	0.931	0.931	0.968	0.943	0.956	0.992	0.982	0.990	0.984	0.955	0.962
2.0	0.50	5	0.956	0.956	0.938	0.992	0.982	0.970	0.978	0.978	0.969	1.000	0.998	0.997
2.0	0.50	10	0.947	0.943	0.936	0.972	0.965	0.953	0.973	0.984	0.966	0.991	0.993	0.975
2.0	0.50	20	0.954	0.942	0.935	0.964	0.963	0.959	0.984	0.981	0.985	0.984	0.973	0.970
2.0	0.50	50	0.949	0.947	0.939	0.965	0.961	0.962	0.994	0.994	0.993	0.973	0.968	0.968
2.0	0.75	5	0.980	0.949	0.955	0.996	0.981	0.982	0.996	0.975	0.979	0.999	0.995	0.999
2.0	0.75	10	0.966	0.967	0.952	0.980	0.980	0.968	0.990	0.988	0.982	0.994	0.992	0.988
2.0	0.75	20	0.944	0.934	0.928	0.953	0.948	0.946	0.981	0.972	0.985	0.982	0.968	0.977
2.0	0.75	50	0.930	0.956	0.948	0.947	0.969	0.965	0.989	0.996	0.995	0.968	0.978	0.972

Table 4.20: Simulation Results: Fully Expanded 95% Frequentist Coverage

Chapter 5

Objective Bayesian Analysis of the Negative Binomial Distribution

5.1 Continuous Dispersion Parameterization

In the statistical analysis of count processes, overdispersion often leads practitioners to consider the Negative Binomial distribution as a robust alternative to the Poisson distribution because it has the advantage of having variance that is not equal to its mean in general. In this chapter, we study the negative binomial distribution in the context of a continuous dispersion parameter α as well as the standard discrete case. Currently, there is no reference prior for the negative binomial distribution when both parameters are unknown, making objective Bayesian analysis difficult. We first present the specific form of the parameterization used for this distribution. In section 5.1.1, we demonstrate that the common approach of modeling overdispersed count data with a Poisson-Gamma mixture yields an improper posterior when objective

hyperpriors are applied to the Gamma distribution in the lower level of the hierarchy. In section 5.1.5, we use the methodology of Berger et al. (2009) to numerically approximate the marginal prior for the dispersion parameter. The approximate prior which is obtained is likely not exactly the reference prior, however its properties are studied anyway, and its performance is superior to the frequentist MLE. In section 5.2, we use the method of Villa and Walker (2014) to obtain an objective prior for the negative binomial distribution under the standard discrete parameterization. A proper posterior for the discrete parameter is obtained for known p , but a lack of literature about this topic prevents us from considering a robust comparative simulation study as we did in Chapter 4.

The negative binomial distribution is a relatively common two parameter discrete distribution. In elementary considerations, it is interpreted as the distribution of the number of “successes” which we will observe in a sequence of independent bernoulli trials before we observe a prespecified number of “failures.” This leads to the pdf

$$f(k; r, p) = \binom{k+r-1}{k} (1-p)^r p^k \quad (5.1)$$

where k is the number of successes, r is the number of failures and p is the probability of a success on each trial. In addition to this, there are many other alternative variations in the interpretation and form of (5.1). The definition in (5.1) can be extended to consider the parameter r as a positive real value. Although we can not conceptualize a non-integer number of “failures”, we can still define the pdf (5.1) by extending the binomial coefficient $\binom{k+r-1}{k}$ to the equivalent form written with the gamma function. To represent the continuous nature of this space, we use α in place of r and further

consider the transformation $p = \frac{\mu}{\mu + \alpha}$ to get the formulation

$$p(y; \alpha, \mu) = \frac{\Gamma(y + \alpha)}{\Gamma(y + 1)\Gamma(\alpha)} \left(\frac{\alpha}{\mu + \alpha} \right)^\alpha \left(\frac{\mu}{\mu + \alpha} \right)^y, y = 0, 1, \dots \quad (5.2)$$

The negative binomial distribution as parameterized by (5.2) can be used as an alternative to the Poisson distribution. This is especially useful for discrete data whose support is unbounded and positive and the sample variance s^2 greatly exceeds the sample mean $\bar{x}$. In this case, the observations are said to be overdispersed with respect to the Poisson distribution. Because the Poisson distribution only has one parameter λ , all moments are restricted to be functions of this one parameter. In fact, the first two theoretical moments of the Poisson distribution are equal, and therefore it should not be used to model data where the variance is much higher than the mean. The parameterization in (5.2) is of particular interest because it yields the centered moments

$$\begin{aligned} E(Y) &= \mu \\ Var(Y) &= \mu + \frac{\mu^2}{\alpha}. \end{aligned} \quad (5.3)$$

We can see that here, μ controls the mean and α controls how large the variance is in relation to that mean. This gives a very nice framework for modeling overdispersed count data, as smaller values of α represent data that are more overdispersed. We can also note that

$$\lim_{\alpha \rightarrow \infty} p(y; \alpha, \mu) = \frac{\mu^y}{y!} e^{-\mu} \equiv Poi(\mu).$$

So, for very large values of α , the distribution approaches a Poisson distribution.

This parameterization is useful, then, for models for which we wish to account for overdispersion. These models have been studied reasonably extensively in the frequentist framework (see Lawless (1987) and references therein).

Even though this elementary distribution only has two parameters, analysis is not straightforward when both parameters are unknown. Even the frequentist MLE requires numerical methods like the EM algorithm or Newton's method to find a solution for both μ and α , and asymptotics or bootstrapping will need to be used to construct confidence intervals or do hypothesis testing. Currently, there is no known objective prior for the negative binomial distribution when both parameters are unknown. We desire an objective prior for (μ, α) so that we are able to assess both mean as well as the amount of over-dispersion present in the data.

5.1.1 Hierarchical Objective Prior

We first note that the negative binomial distribution in (5.2) may be obtained by considering a poisson-gamma mixture. We have

$$\begin{aligned} y_i | \lambda_i &\sim \text{Poi}(\lambda_i), \\ \lambda_i | \alpha, \mu &\sim \text{Gamma}\left(\alpha, \frac{\alpha}{\mu}\right). \end{aligned} \tag{5.4}$$

The marginal distribution of y_i is therefore

$$\begin{aligned} p(y_i | \alpha, \mu) &= \int_0^\infty \frac{\lambda_i^{y_i} e^{-\lambda_i}}{\Gamma(y_i + 1)} \frac{\alpha^\alpha \lambda_i^{\alpha-1} e^{-\frac{\alpha}{\mu} \lambda_i}}{\mu^\alpha \Gamma(\alpha)} d\lambda_i \\ &= \frac{(\alpha/\mu)^\alpha}{\Gamma(y_i + 1) \Gamma(\alpha)} \left(\frac{\mu}{\mu + \alpha} \right)^{y_i + \alpha} \int_0^\infty \lambda_i^{y_i + \alpha - 1} e^{-\lambda_i} d\lambda_i \end{aligned}$$

$$= \frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)\Gamma(\alpha)} \left(\frac{\alpha}{\mu}\right)^\alpha \left(\frac{\mu}{\mu + \alpha}\right)^{y_i + \alpha}, y = 0, 1, \dots$$

which is the same as (5.2). Because the negative binomial distribution can be obtained from a Poisson-Gamma mixture, an initial approach to obtain an objective prior for the negative binomial distribution may be to use established methods for the gamma distribution on (α, μ) . This hierarchical structure was explored at length in Berger et al. (2005), which demonstrated that for a normal hierarchical model with $y_j|\theta_j \sim N(\theta_j, 1)$ and $\theta_j \sim N(\mu, \tau^2)$, using the reference prior for (μ, τ) yields an improper posterior. As we will see, the same thing happens in this case if we use established methods to construct an objective prior for (μ, α) in (5.2) through the gamma distribution in (5.4).

5.1.2 Hierarchical Priors

Objective priors for the gamma distribution were presented in Sun and Ye (1996). The Jeffery's, Reference, and the H prior of Liseo (1993) are

$$\begin{aligned}\pi_J(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu}, \\ \pi_R(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu\sqrt{\alpha}}, \\ \pi_H(\alpha, \mu) &\propto \frac{\alpha\Psi'(\alpha) - 1}{\mu\sqrt{\alpha}},\end{aligned}$$

where $\Psi'(x) = \frac{d^2}{dx^2} \log \Gamma(x)$ is the *trigamma* function. We note that the prior for μ for each of the above priors is proportional to $\frac{1}{\mu}$, a fact which will make the derivations that follow much easier.

5.1.3 Integrated Likelihood for α

First, the likelihood for equation 5.2 is

$$p(\mathbf{y}|\mu, \alpha) = \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{\Gamma(\alpha)^n} \left(\frac{\alpha}{\mu + \alpha} \right)^{n\alpha} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i}. \quad (5.5)$$

Because the prior for μ in each of the above is proportional to $\frac{1}{\mu}$, the integrated likelihood for α is as follows for each of the priors defined in section 5.1.2.

$$\begin{aligned} L^*(\mathbf{y}|\alpha) &= \int_0^\infty \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{\Gamma(\alpha)^n} \left(\frac{\alpha}{\mu + \alpha} \right)^{n\alpha} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} \frac{1}{\mu} d\mu \\ &= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{\Gamma(\alpha)^n} \int_0^\infty \frac{\mu^{\sum y_i - 1}}{(\mu + \alpha)^{\sum y_i + n\alpha}} d\mu. \end{aligned}$$

Now consider the transformation $p = \frac{\mu}{\mu + \alpha}$, we have

$$\begin{aligned} L^*(\mathbf{y}|\alpha) &= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{[\Gamma(\alpha)]^n} \int_0^1 p^{\sum_{i=1}^n y_i - 1} \sqrt{\frac{1-p}{\alpha p}} \left(\frac{1-p}{\alpha} \right)^{n\alpha + \frac{1}{2}} \frac{\alpha}{(1-p)^2} d\alpha \\ &= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{[\Gamma(\alpha)]^n} \int_0^1 p^{\sum_{i=1}^n y_i - 1} (1-p)^{n\alpha - 1} dp \\ &= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\Gamma(n\alpha) \cdot \Gamma(\sum y_i)}{\Gamma(\alpha)^n \cdot \Gamma(\sum y_i + n\alpha)} \\ &\propto \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)}. \end{aligned} \quad (5.6)$$

Unfortunately, the gamma functions make the form of 5.6 much too difficult to work with analytically. Because of this, we study the behavior of the function as $\alpha \rightarrow 0$ and $\alpha \rightarrow \infty$, and proceed similarly for other formulations with many functions later.

5.1.4 Posterior Impropriety

Because the integrated likelihood in (5.6) is difficult to work with, the posterior distribution which would be obtained by mixing the priors in 5.1.2 with the integrated likelihood (5.6) must be studied as $\alpha \rightarrow 0$ and $\alpha \rightarrow \infty$ as it cannot be obtained analytically. By studying the behavior of the priors as well as the integrated likelihood we can determine the behavior of the posterior at each boundary. As it turns out, the posterior would be improper for all the priors in Sun and Ye (1996).

Posterior Behavior as $\alpha \rightarrow 0$

Lemma 5.1.1. The integrated likelihood $L^*(\mathbf{y}|\alpha)$ is constant as $\alpha \rightarrow 0$

Proof. First note that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \Rightarrow \Gamma(\alpha) \approx \frac{1}{\alpha}$ as $\alpha \rightarrow \infty$.

We consider 3 cases:

1. If $y_i = 0 \forall i$,

$$L^*(\mathbf{y}|\alpha) \approx \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)} = 1.$$

2. If $y_i \neq 0$ for a single i ,

$$L^*(\mathbf{y}|\alpha) \approx \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)} = \frac{\Gamma(y_1 + \alpha)}{\Gamma(\alpha)} \frac{\Gamma(n\alpha)}{\Gamma(y_1 + n\alpha)}.$$

As $\alpha \rightarrow 0$,

$$\frac{\Gamma(y_1 + \alpha)}{\Gamma(\alpha)} \frac{\Gamma(n\alpha)}{\Gamma(y_1 + n\alpha)} \approx \frac{\Gamma(y_i)\alpha}{n\alpha\Gamma(y_i)} = \frac{1}{n}.$$

3. If $y_i, y_j \neq 0$ for a pair $i \neq j$,

$$\begin{aligned} L^*(\mathbf{y}|\alpha) &\approx \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)} \\ &= \frac{\Gamma(y_i + \alpha)\Gamma(y_j + \alpha)}{\Gamma(\alpha)} \frac{\Gamma(n\alpha)}{\Gamma(y_i + y_j + n\alpha)} \approx \frac{\Gamma(y_i)\Gamma(y_j)\alpha}{n\alpha\Gamma(y_i + y_j)} = \frac{\Gamma(y_i)\Gamma(y_j)}{n\Gamma(y_i + y_j)}. \end{aligned}$$

So the likelihood is constant as $\alpha \rightarrow 0$.

Finally, say $y_i \neq 0 \forall i$, we have

$$L^*(\mathbf{y}|\alpha) \approx \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)} \rightarrow \frac{\prod_{i=1}^n \Gamma(y_i)}{1/\alpha^n} \frac{1/n\alpha}{\Gamma(\sum y_i)} = \frac{\prod_{i=1}^n \Gamma(y_i)}{\Gamma(\sum y_i)} \frac{\alpha^{n-1}}{n},$$

which is 0 as $\alpha \rightarrow 0$. □

Lemma 5.1.2. As $\alpha \rightarrow 0$, the order of the priors for (μ, α) from Sun and Ye (1996) are

$$\begin{aligned} \pi_J(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu} \approx \frac{1}{\sqrt{\alpha}}, \\ \pi_R(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu\sqrt{\alpha}} \approx \frac{1}{\alpha}, \\ \pi_H(\alpha, \mu) &\propto \frac{\alpha\Psi'(\alpha) - 1}{\mu\sqrt{\alpha}} \approx \frac{1}{\alpha^{3/2}}. \end{aligned}$$

Proof. We note that $\Psi'(x)$ has the expansion.

$$\Psi'(x) = \sum_{j=0}^{\infty} \frac{1}{(\alpha + j)^2} = \frac{1}{\alpha^2} + \sum_{j=1}^{\infty} \frac{1}{(\alpha + j)^2},$$

so that $\Psi'(x)$ has order $\frac{1}{\alpha^2}$ as $\alpha \rightarrow 0$. Combining this with the forms of π_J , π_R , and π_H yield the aforementioned orders. □

Theorem 5.1.3. The posterior obtained using the integrated likelihood (2) and π_J from Sun and Ye (1996) is integrable as $\alpha \rightarrow 0$.

Proof. Because Lemma 5.1.4 showed that the likelihood is constant, we have that the posteriors under π_J , π_R , or π_H have the same order as shown in Lemma 5.1.5 as $\alpha \rightarrow 0$, hence only the posterior obtained under π_J is integrable. $\square$

Posterior Behavior as $\alpha \rightarrow \infty$

Lemma 5.1.4. The integrated likelihood $L^*(\mathbf{y}|\alpha)$ is constant as $\alpha \rightarrow \infty$

Proof. First we recall Stirling's approximation, which gives

$$\Gamma(x+1) \approx \frac{\sqrt{2\pi x} x^x}{e^x} \Rightarrow \Gamma(x) \approx \frac{\sqrt{2\pi(x-1)}(x-1)^{x-1}}{e^{x-1}} = \sqrt{2\pi} \frac{(x-1)^{x-1/2}}{e^{x-1}}.$$

We then note

$$\sqrt{2\pi} \frac{(x-1)^{x-1/2}}{e^{x-1}} \rightarrow \frac{x^{x-1/2}}{e^x} \text{ as } x \rightarrow \infty.$$

So we use $\Gamma(x) \approx \frac{x^{x-1/2}}{e^x}$ as $x \rightarrow \infty$.

$$\begin{aligned} L^*(\mathbf{y}|\alpha) &\propto \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \cdot \frac{\Gamma(n\alpha)}{\Gamma(\sum y_i + n\alpha)} \\ &\approx \prod_{i=1}^n \left(\frac{(y_i + \alpha)^{y_i + \alpha - 1/2}}{e^{y_i + \alpha}} \right) \frac{e^{n\alpha}}{\alpha^{n\alpha - n/2}} \frac{(n\alpha)^{n\alpha - 1/2}}{e^{n\alpha}} \frac{e^{\sum y_i + n\alpha}}{(\sum y_i + n\alpha)^{\sum y_i + n\alpha - 1/2}} \\ &= {}_c \frac{\prod_{i=1}^n (y_i + \alpha)^{y_i + \alpha - 1/2}}{(\sum y_i + n\alpha)^{\sum y_i + n\alpha - 1/2}} \cdot n^{n\alpha} \alpha^{\frac{n-1}{2}}. \end{aligned}$$

Examining the first term of the exponents in the first fraction, we have

$$\left(\frac{y_i + \alpha}{\sum y_i + n\alpha} \right)^{y_i} \rightarrow \left(\frac{1}{n} \right)^{y_i} \approx c \text{ as } \alpha \rightarrow \infty,$$

giving us

$$L^*(\mathbf{y}|\alpha) \propto c \frac{\prod_{i=1}^n (y_i + \alpha)^{\alpha-1/2}}{(\sum y_i + n\alpha)^{n\alpha-1/2}} \cdot n^{n\alpha} \alpha^{\frac{n-1}{2}} \text{ as } \alpha \rightarrow \infty.$$

Next note that

$$\prod_{i=1}^n (y_i + \alpha)^\alpha = \alpha^{n\alpha} \prod_{i=1}^n \frac{(y_i + \alpha)^\alpha}{\alpha^\alpha} = \alpha^{n\alpha} \prod_{i=1}^n \left(1 + \frac{y_i}{\alpha} \right)^\alpha \rightarrow \alpha^{n\alpha} e^{\sum y_i} \text{ as } \alpha \rightarrow \infty,$$

which tells us

$$L^*(\mathbf{y}|\alpha) \propto c \frac{\prod_{i=1}^n (y_i + \alpha)^{-1/2}}{(\sum y_i + n\alpha)^{n\alpha-1/2}} \cdot (n\alpha)^{n\alpha} \alpha^{\frac{n-1}{2}} \text{ as } \alpha \rightarrow \infty.$$

Now,

$$\left(\frac{\sum y_i + n\alpha}{n\alpha} \right)^{n\alpha} = \left(1 + \frac{\sum y_i}{n\alpha} \right)^{n\alpha} \rightarrow e^{\sum y_i} \text{ as } \alpha \rightarrow \infty,$$

which further reduces the likelihood to

$$L^*(\mathbf{y}|\alpha) \propto c \frac{\prod_{i=1}^n (y_i + \alpha)^{-1/2}}{(\sum y_i + n\alpha)^{-1/2}} \cdot \alpha^{\frac{n-1}{2}} \text{ as } \alpha \rightarrow \infty,$$

which is constant as $\alpha \rightarrow \infty$. □

Lemma 5.1.5. As $\alpha \rightarrow \infty$, the order of the priors for (μ, α) from Sun and Ye (1996)

are

$$\begin{aligned}\pi_J(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu} = O(1), \\ \pi_R(\alpha, \mu) &\propto \frac{\sqrt{\alpha\Psi'(\alpha) - 1}}{\mu\sqrt{\alpha}} = O(\alpha^{-1/2}), \\ \pi_H(\alpha, \mu) &\propto \frac{\alpha\Psi'(\alpha) - 1}{\mu\sqrt{\alpha}} = O(\alpha^{-1/2}).\end{aligned}$$

Proof. First, note that $\Psi'(x)$ is a special case of the polygamma function

$$\psi^{(m)}(x) = \frac{d^{m+1}}{dx^{m+1}} \ln \Gamma(x),$$

with $m = 1$. For large values of x , $\psi^{(m)}(x)$ has the series expansion

$$\psi^{(m)}(x) \sim (-1)^{m+1} \sum_{k=0}^{\infty} \frac{(k+m-1)!}{k!} \frac{B_k}{x^{k+m}},$$

where B_k are the Bernoulli numbers with $B_1 = \frac{1}{2}$ and $m \geq 1$. Therefore

$$\Psi'(x) \sim \sum_{k=0}^{\infty} \frac{B_k}{x^{k+1}} \Rightarrow \Psi'(x) \approx \frac{1}{x} + \frac{1}{x^2} + \frac{1}{x^3} + \dots,$$

so that $\Psi'(x) = O(x^{-1})$. When $\Psi'(\alpha) \approx \alpha^{-1}$ is substituted into the priors, the result follows. $\square$

Theorem 5.1.6. The posterior obtained using the integrated likelihood (2) and any prior from Sun and Ye (1996) is not integrable as $\alpha \rightarrow \infty$

Proof. Because Lemma 5.1.4 showed that the likelihood is constant, we have that the posteriors under π_J , π_R , or π_H have the same order as shown in Lemma 5.1.5 and are

thus not integrable at ∞ . □

Because the posterior is not integrable at ∞ under any prior, the posterior is improper under any of the priors from Sun and Ye (1996).

5.1.5 Direct Computation of $\pi(\mu, \alpha)$

Exact Marginal Reference Prior Method Berger et al. (2001) detailed a procedure to find a marginal reference prior based on conditioning. To obtain the joint reference prior for μ and α we begin by factoring it as $\pi^R(\mu, \alpha) \propto \pi^R(\mu|\alpha)\pi^R(\alpha)$. We then take $\pi^R(\mu|\alpha)$ to be the Jeffery's rule prior, which is the standard reference prior for the univariate case. We then obtain the integrated likelihood

$$L(\alpha|\mathbf{y}) = \int_{\Theta} L(\mu, \alpha|\mathbf{y})\pi^R(\mu|\alpha)d\mu,$$

and take the marginal reference prior $\pi^R(\alpha)$ to be the Jefferey's rule prior for this marginal likelihood. The method with respect to this problem proceeds as follows.

The log of the p.m.f in (5.2) is

$$\log p(y; \alpha, \mu) = c + \log \Gamma(y + \alpha) - \log \Gamma(\alpha) + \alpha \log(\alpha) + y \log \mu - (y + \alpha) \log(\mu + \alpha),$$

which yields the first and second derivatives

$$\begin{aligned} \frac{d}{d\mu} \log p(y; \alpha, \mu) &= \frac{y}{\mu} - \frac{y+\alpha}{\mu+\alpha}, \\ \frac{d}{d\alpha} \log p(y; \alpha, \mu) &= \Psi(y + \alpha) - \Psi(\alpha) + \log(\alpha) + 1 - \left[\frac{y+\alpha}{\mu+\alpha} + \log(\mu + \alpha) \right], \\ \frac{d^2}{d\mu^2} \log p(y; \alpha, \mu) &= -\frac{y}{\mu^2} + \frac{y+\alpha}{(\mu+\alpha)^2}, \\ \frac{d^2}{d\alpha^2} \log p(y; \alpha, \mu) &= \Psi'(y + \alpha) - \Psi'(\alpha) + \frac{1}{\alpha} - \left[\frac{(\mu+\alpha)-(y+\alpha)}{(\mu+\alpha)^2} + \frac{1}{\mu+\alpha} \right], \text{ and} \end{aligned}$$

$$\frac{d^2}{d\mu d\alpha} \log p(y; \alpha, \mu) = -\frac{(\mu+\alpha)-(y+\alpha)}{(\mu+\alpha)^2}.$$

Where $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ and $\Psi'(x) = \frac{d^2}{dx^2} \log \Gamma(x)$ are the *digamma* and *trigamma* functions.

Recall that for (5.2) we have $E[Y] = \mu$ and thus obtain the fisher information matrix as

$$I(\mu, \alpha) = \begin{pmatrix} \frac{\alpha}{\mu(\mu+\alpha)} & 0 \\ 0 & \Psi'(\alpha) - E[\Psi'(y+\alpha)] - \frac{\mu}{\alpha(\mu+\alpha)} \end{pmatrix}$$

. We can see that if α is known, the reference prior μ is just $\pi(\mu|\alpha) \propto \frac{1}{\sqrt{\mu(\mu+\alpha)}}$. However if we desire a joint prior $\pi(\mu, \alpha)$ using the asymptotic method of Berger and Bernardo (1992a), the expectation required to evaluate $I_{22}(\mu, \alpha)$ for $|I(\mu, \alpha)|$ is analytically difficult. An approximation to the marginal reference prior will be needed. The likelihood for equation 5.2 is

$$p(\mathbf{y}|\mu, \alpha) = \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{\Gamma(\alpha)^n} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} (\mu + \alpha)^{-n\alpha}.$$

Recall that $\pi(\mu|\alpha) \propto \frac{1}{\sqrt{\mu(\mu+\alpha)}}$ is the Jeffery's rule prior when α is known and obtain the marginal likelihood for α by integrating over μ as follows

$$\begin{aligned} p(\mathbf{y}|\alpha) &= \int_0^\infty \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{\Gamma(\alpha)^n} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} (\mu + \alpha)^{-n\alpha} \frac{1}{\sqrt{\mu(\mu + \alpha)}} d\mu \\ &= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{[\Gamma(\alpha)]^n} \int_0^\infty \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} \frac{1}{\sqrt{\mu}(\mu + \alpha)^{n\alpha - \frac{1}{2}}} d\mu. \end{aligned}$$

Consider the transformation $p = \frac{\mu}{\mu+\alpha}$ and the corresponding Jacobian, we have

$$p(\mathbf{y}|\alpha) = \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{[\Gamma(\alpha)]^n} \int_0^1 p^{\sum_{i=1}^n y_i} \sqrt{\frac{1-p}{\alpha p}} \left(\frac{1-p}{\alpha} \right)^{n\alpha + \frac{1}{2}} \frac{\alpha}{(1-p)^2} d\alpha$$

$$\begin{aligned}
&= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{[\Gamma(\alpha)]^n} \int_0^1 p^{\sum_{i=1}^n y_i + \frac{1}{2} - 1} (1-p)^{n\alpha + \frac{1}{2} - 1} dp \\
&= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\Gamma(n\alpha) \cdot \Gamma(\sum y_i + 1)}{\Gamma(\alpha)^n \cdot \Gamma(\sum y_i + n\alpha + 1)}.
\end{aligned}$$

This likelihood, in addition to being difficult to work with analytically as before, is improper due to the impropriety of $\pi^R(\mu|\alpha)$, so calculating the Jeffery's prior for this is not an option. A solution is to truncate the prior $\pi^R(\mu|\alpha)$ at c , so we may induce propriety in the marginal likelihood for α , and then approximate the reference prior using the method of Berger et al. (2009) since we are now in the univariate case. Truncating $\pi^R(\mu|\alpha)$ at c yields the following density and normalizing constant.

$$\pi^{R*}(\mu|\alpha) \propto \frac{1}{\sqrt{\mu(\mu + \alpha)}}, \mu \in (0, c). \quad (5.7)$$

The normalizing constant is

$$\int_0^c \frac{1}{\sqrt{\mu(\mu + \alpha)}} d\mu = 2 \log(\sqrt{c + \alpha} + \sqrt{c}) - \log \alpha.$$

So we then use

$$\pi^{R*}(\mu|\alpha) = \frac{1}{2 \log(\sqrt{c + \alpha} + \sqrt{c}) - \log \alpha} \frac{1}{\sqrt{\mu(\mu + \alpha)}}, \mu \in (0, c)$$

to compute $p^*(\alpha|\mathbf{y})$, the truncated marginal likelihood. This proceeds by first integrating μ over the truncated space. We have

$$p^*(\alpha|\mathbf{y}) = \int_0^c L(\mu, \alpha|\mathbf{y}) \pi^{R*}(\mu|\alpha) d\mu$$

$$\begin{aligned}
&= \int_0^c \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{\alpha^{n\alpha}}{\Gamma(\alpha)^n} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} (\mu + \alpha)^{-(n\alpha)} \times \\
&\quad \times \frac{1}{2 \log(\sqrt{c + \alpha} + \sqrt{c}) - \log \alpha} \frac{1}{\sqrt{\mu(\mu + \alpha)}} d\mu \\
&= \frac{B\left(\frac{c}{c + \alpha}; \sum y_i + 1, n\alpha\right)}{2 \log(\sqrt{c + \alpha} + \sqrt{c}) - \log \alpha} \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{\Gamma(\alpha)^n},
\end{aligned}$$

where $B(x; a, b) = \int_0^x p^{a-1}(1-p)^{b-1}dp$ is the *incomplete beta function*, with $x = 1$ giving us the beta function $B(a, b)$. Note that the calculations are similar to the above and are not detailed. The truncated marginal likelihood, while proper, is still very difficult to work with analytically. Instead, since the problem is now univariate in nature, the reference prior can be approximated with the numerical algorithm described in Berger et al. (2009, pg. 920), which is a numerical procedure designed to mimic theorem 7 in the article.

The algorithm with respect to this problem, proceeds as follows: For each α over a grid for with the approximation is desired, repeat the following steps from $j = 1, \dots, m$.

1. Simulate a random sample $\{\mathbf{y}_{1j}, \dots, \mathbf{y}_{kj}\}$ of size k from $p(\mathbf{y}|\mu, \alpha)$.
 - (a) Generate μ from $\pi^{R^*}(\mu|\alpha)$ given in (5.7).
 - (b) Use the generated μ and the fixed α to generate k random samples of size n from the negative binomial distribution in (5.2).
2. Numerically evaluate

$$c_j = \int_0^\infty \prod_{i=1}^k p(\mathbf{y}_{ij}|\alpha) \pi^*(\alpha) d\alpha.$$

3. Compute

$$r_j(\alpha) = \log \left[\frac{\prod_{i=1}^k p(\mathbf{y}_{ij}|\alpha)\pi^*(\alpha)}{c_j} \right].$$

4. Compute and store $\pi(\alpha) = \exp \left[\frac{\sum_{j=1}^m r_j(\alpha)}{m} \right]$.

This is done over a grid of α values for which the estimate of the prior $\pi(\alpha)$ is desired. Note that $\pi^{R*}(\alpha)$ is an arbitrary choice, so it can be chosen to aid computation. To facilitate computation of steps 2, 3 and 4, the "Rmpfr" library in R was used because the products in these steps often induce overflow to ∞ or underflow to 0. The Rmpfr library is an R package which interfaces to the LGPL'ed MPFR (Multiple Precision Floating-Point Reliable) Library which itself is based on the GMP (GNU Multiple Precision) Library. Specifically, MPFR is a C library which provides the ability to conduct accurate software based multiple-precision floating-point computations (Lefevres, 2011). This R package allows the use of an arbitrary level of precision to avoid both overflow and underflow in the computations at the expense of computation time (Machler, 2012). Here, octuple (256-bit) precision was used in all intermittent calculations in steps 2, 3, and 4, and the resulting estimations of $\pi(\alpha)$ were converted to doubles (64-bit) for analysis. After the approximated $\pi(\alpha)$ values are computed, standard interpolation techniques can be used to provide a functional approximation to $\pi(\alpha)$.

Results

First, it was desired to examine if the shape of the distribution for $\pi(\alpha)$ changed for different values of the truncation constant c used for $\pi(\mu|\alpha)$ as well as the sample size n used in the algorithm. For this portion of the study, $k = 100$ and $m = 250$

were used due to the complexity of the algorithm and the desire for a broad range of results. Furthermore, $c = 100, 500$, and 1000 and $n = 20, 30, 40$, and 50 were used to estimate $\pi(\alpha)$ for α between $.1$ and $.65$, in steps of $.05$. The normalized results can be seen in Figure 5.1 and indicate that the shape of the density does not change for different values of c or n . We focus on smaller values of α because in the parameterization shown in (5.2), smaller values indicate more over-dispersion, which is the case we would like to focus on with this problem.

After this, the algorithm was finally run with $m = 750$, $k = 1000$, $n = 90$, and $c = 1000$. $\pi(\alpha)$ was computed over 11 points for α between $.1$ and $.70$ which were equally spaced on the log scale. Figure 5.2 shows the plot of these $(\alpha, \pi(\alpha))$ estimates. It appears that there is infinite mass at 0, so the functional form of the prior should have this quality as well. To arrive at a functional form for $\pi(\alpha)$, a polynomial regression was fit to $g(\alpha)\pi(\alpha)$. The signs of these regression coefficients always seemed to alternate, and after some experimentation, the Taylor series for $\psi'(\alpha + 1)$,

$$\psi'(\alpha + 1) = \sum_{k=0}^{\infty} (-1)^{k+2} \frac{(k+1)!}{k!} \zeta(k+2) z^k,$$

where $\zeta()$ is the Riemann zeta function, seemed to match the coefficients of this polynomial fit very well. In Table 5.1, we can see the polynomial coefficients for the first three terms of a polynomial fit to $\sqrt{\alpha}\pi(\alpha)$ are very close, therefore it appears that $\pi(\alpha)$ has the form $\pi(\alpha) \propto \frac{\psi'(1+\alpha)}{\sqrt{\alpha}}$. Note that Maple 17 gives the result $\int_0^{\infty} \frac{\psi'(1+\alpha)}{\sqrt{\alpha}} d\alpha = \infty$, therefore we define and normalize on the compact support $\alpha \in [b_0, c_0]$, which yields

$$\pi_{b_0, c_0}(\alpha) = \frac{1}{m(\pi_{b_0, c_0})} \frac{\psi'(1+\alpha)}{\sqrt{\alpha}}, \quad \alpha \in [b_0, c_0], \quad (5.8)$$

where a closed form for the normalizing constant $m(\pi_{b_0, c_0})$ is not needed.

Using (5.8) and proceeding as in Berger et al. (2009) we take the limits of the compact supports defined for $\pi_{c_1}(\mu|\alpha)$ and $\pi_{c_0}(\alpha)$

$$\pi(\alpha, \mu) = \lim_{c_i \rightarrow \infty} \frac{\pi_{c_1}(\mu|\alpha)\pi_{b_0, c_0}(\alpha)}{\pi_{c_1}(\mu^*|\alpha^*)\pi_{b_0, c_0}(\alpha^*)},$$

where μ^* and α^* are any interior points of the supports of μ and α . Using $\mu^* = 1$, $\alpha^* = 1$, and noting that $\psi'(2) = \frac{\pi^2}{6} - 1$ we have

$$\begin{aligned} \pi(\alpha, \mu) &= \lim_{c \rightarrow \infty} \frac{\pi_c(\mu|\alpha)\pi_c(\alpha)}{\pi_c(\mu^*|\alpha^*)\pi_c(\alpha^*)} = \lim_{c \rightarrow \infty} \frac{2 \log(\sqrt{c+1} + \sqrt{c}) - \log 1}{2 \log(\sqrt{c+\alpha} + \sqrt{c}) - \log \alpha} \\ &\quad \times \frac{\sqrt{1(1+1)}}{\sqrt{\mu(\mu+\alpha)}} \frac{\psi'(1+\alpha)}{(\frac{\pi^2}{6} - 1) \sqrt{\alpha}} \\ &= \frac{\sqrt{2}}{\sqrt{\mu(\mu+\alpha)}} \frac{\psi'(1+\alpha)}{(\frac{\pi^2}{6} - 1) \sqrt{\alpha}} \\ &\quad \times \lim_{c \rightarrow \infty} \frac{2 \log(\sqrt{c+1} + \sqrt{c})}{2 \log(\sqrt{c+\alpha} + \sqrt{c}) - \log \alpha}. \end{aligned} \tag{5.9}$$

Now, note that

$$\begin{aligned} \lim_{c \rightarrow \infty} \frac{2 \log(\sqrt{c+1} + \sqrt{c})}{2 \log(\sqrt{c+\alpha} + \sqrt{c}) - \log \alpha} &= \lim_{c \rightarrow \infty} \frac{\log(\sqrt{c+1} + \sqrt{c})}{\log(\sqrt{c+\alpha} + \sqrt{c})} \\ &= \lim_{c \rightarrow \infty} \frac{\log\left(\frac{\sqrt{c+1}}{\sqrt{c}} + 1\right) + \log(\sqrt{c})}{\log\left(\frac{\sqrt{c+\alpha}}{\sqrt{c}} + 1\right) + \log(\sqrt{c})} = 1, \end{aligned}$$

which gives us

$$\pi(\alpha, \mu) = \frac{\sqrt{2}\psi'(1+\alpha)}{(\frac{\pi^2}{6} - 1) \sqrt{\alpha\mu(\mu+\alpha)}}. \tag{5.10}$$

This prior, while not necessarily the reference prior, still may be an approximation to it, and may yield good properties under simulation. Therefore, we next obtain the posterior and proceed with simulation.

Posterior

Combining the joint prior in (5.10) with the likelihood in (5.5) we obtain

$$\begin{aligned}
p(\mu, \alpha|y) &\propto \pi(y|\mu, \alpha)\pi(\mu, \alpha) \\
&= \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\Gamma(y_i + 1)} \right] \frac{1}{\Gamma(\alpha)^n} \left(\frac{\alpha}{\mu + \alpha} \right)^{n\alpha} \left(\frac{\mu}{\mu + \alpha} \right)^{\sum y_i} \frac{\sqrt{2}\psi'(1 + \alpha)}{\left(\frac{\pi^2}{6} - 1\right) \sqrt{\alpha\mu(\mu + \alpha)}} \\
&\propto \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \psi'(1 + \alpha) \alpha^{n\alpha-1/2} \mu^{\sum y_i - \frac{1}{2}} (\mu + \alpha)^{-(\sum y_i + n\alpha + \frac{1}{2})}. \quad (5.11)
\end{aligned}$$

We have $\int_0^\infty \pi(\alpha) d\alpha = \infty$, so we must verify $p(y) < \infty$ so that the posterior in (5.11) is a proper distribution.

Theorem 5.1.7. The posterior distribution in (5.11) is a proper distribution.

Proof. First, obtain the marginal likelihood for y as

$$\begin{aligned}
p(y) &= \int_0^\infty \int_0^\infty \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \psi'(1 + \alpha) \alpha^{n\alpha-1/2} \mu^{\sum y_i - \frac{1}{2}} (\mu + \alpha)^{-(\sum y_i + n\alpha + \frac{1}{2})} d\mu d\alpha \\
&= \int_0^\infty \frac{\psi'(1 + \alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \int_0^\infty \alpha^{n\alpha} \mu^{\sum y_i - \frac{1}{2}} (\mu + \alpha)^{-(\sum y_i + n\alpha + \frac{1}{2})} d\mu d\alpha \\
&= \int_0^\infty \frac{\psi'(1 + \alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \int_0^\infty \left(\frac{\alpha}{\mu + \alpha} \right)^{n\alpha} \mu^{\sum y_i - \frac{1}{2}} (\mu + \alpha)^{-\sum y_i - \frac{1}{2}} d\mu d\alpha.
\end{aligned}$$

Considering the transformation $p = \frac{\alpha}{\mu + \alpha} \in (0, 1) \Rightarrow d\mu = \left(\frac{\alpha}{\mu + \alpha} \right)^{-1} (\mu + \alpha) dp$ we have

$$p(y) = \int_0^\infty \frac{\psi'(1 + \alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \int_0^1 p^{n\alpha-1} (1-p)^{\sum y_i - \frac{1}{2}} dp d\alpha$$

$$\begin{aligned}
&= \int_0^\infty \frac{\psi'(1+\alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} B\left(n\alpha, \sum y_i + \frac{1}{2}\right) d\alpha \\
&\propto \int_0^\infty \frac{\psi'(1+\alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} d\alpha \\
&= \int_0^\infty p(\alpha|y) d\alpha.
\end{aligned} \tag{5.12}$$

There is likely no analytical solution to (5.12). However $p(y) < \infty$ if $p(\alpha|y)$ is integrable as $\alpha \rightarrow 0$ and as $\alpha \rightarrow \infty$

Behavior of $p(\alpha|y)$ as $\alpha \rightarrow 0$ Examining each factor of $p(\alpha|y)$ we have

$$\lim_{\alpha \rightarrow 0} \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} = \frac{\prod_{i=1}^n \Gamma(y_i)}{\Gamma(\sum y_i + \frac{1}{2})} \text{ as } \alpha \rightarrow 0,$$

and

$$\lim_{\alpha \rightarrow 0} \frac{\psi'(1+\alpha) \Gamma(n\alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \approx \lim_{\alpha \rightarrow 0} \psi'(1) \frac{1/n\alpha}{\alpha^{1/2} \cdot 1/\alpha^n} = \psi'(1) \lim_{\alpha \rightarrow 0} \frac{\alpha^{n-3/2}}{n} = 0 \text{ if } n > 1,$$

so $p(\alpha|y)$ is integrable as $\alpha \rightarrow 0$ provided we have more than 1 observation.

Behavior of $p(\alpha|y)$ as $\alpha \rightarrow \infty$ First we have the following Lemma.

Lemma 5.1.8. As $\alpha \rightarrow \infty$

$$\frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} = O\left(\alpha^{-\frac{1}{2}}\right).$$

Proof. It is sufficient to show

$$\alpha^{\frac{1}{2}} \cdot \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \rightarrow c \text{ as } \alpha \rightarrow \infty.$$

We have

$$\begin{aligned} & \lim_{\alpha \rightarrow \infty} \alpha^{\frac{1}{2}} \cdot \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \\ &= \lim_{\alpha \rightarrow \infty} \alpha^{\frac{1}{2} + \sum y_i} \cdot \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\alpha^{y_i} \Gamma(\alpha)} \right] \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \\ &= \frac{1}{n^{\sum y_i + 1/2}} \lim_{\alpha \rightarrow \infty} \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\alpha^{y_i} \Gamma(\alpha)} \right] \frac{(n\alpha)^{\sum y_i + 1/2} \Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \\ &= \frac{1}{n^{\sum y_i + 1/2}} \lim_{\alpha \rightarrow \infty} \left\{ \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\alpha^{y_i} \Gamma(\alpha)} \right] \right\} \lim_{\alpha \rightarrow \infty} \left\{ \frac{(n\alpha)^{\sum y_i + 1/2} \Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \right\}. \end{aligned}$$

We note that because y_i is an integer each factor in the first limit has the form

$$\frac{\Gamma(y_i + \alpha)}{\alpha^{y_i} \Gamma(\alpha)} = \frac{f(\alpha) \Gamma(\alpha)}{\alpha^{y_i} \Gamma(\alpha)},$$

where $f(\alpha) = o(\alpha^{y_i})$. Hence,

$$\lim_{\alpha \rightarrow \infty} \left\{ \prod_{i=1}^n \left[\frac{\Gamma(y_i + \alpha)}{\alpha^{y_i} \Gamma(\alpha)} \right] \right\} = 1,$$

with the second limit proceeding similarly. With this, we have

$$\alpha^{\frac{1}{2}} \cdot \frac{\prod_{i=1}^n \Gamma(y_i + \alpha)}{\Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})} \rightarrow \frac{1}{n^{\sum y_i + 1/2}} \text{ as } \alpha \rightarrow \infty.$$

□

When combining this with $\frac{\psi'(1+\alpha)}{\sqrt{\alpha}}$ and noting that $\psi'(1+\alpha) = o(\alpha^{-1})$, we have that $p(\alpha|y) = o(\alpha^{-2})$, therefore $p(\alpha|y)$ is integrable as $\alpha \rightarrow \infty$. Combining results, we have that $p(\alpha|y)$ is integrable at 0 and ∞ . $\square$

Posterior Visualization

To visualize the posterior distribution for a dataset y , a random sample of $n = 500$ from a negative binomial distribution with known μ and α was generated. In Figure 5.3, μ was fixed to be 5, and α was varied between 2, 5, 10, and ∞ to represent high, moderate, low, and no over-dispersion, respectively. To simulate y for the case of $\alpha \rightarrow \infty$, a sample of $n = 250$ was taken from the Poisson distribution. Once y was obtained the joint posterior was evaluated and plotted on a grid with a resolution of .01. The posterior behaves in an expected way. When there is data which is highly over-dispersed (i.e. $\alpha \rightarrow 0$), the marginal posterior for α becomes highly concentrated on the true value. As the data becomes less and less overdispersed, we notice that the posterior of α becomes much less concentrated, and much more mass is located at large values of α . To study this further, note that equation (5.12) tells us that the marginal posterior of $\alpha|\mathbf{y}$ is

$$\pi(\alpha|\mathbf{y}) \propto \frac{\psi'(1+\alpha) \prod_{i=1}^n \Gamma(y_i + \alpha)}{\alpha^{1/2} \Gamma(\alpha)^n} \frac{\Gamma(n\alpha)}{\Gamma(n\alpha + \sum y_i + \frac{1}{2})}. \quad (5.13)$$

This marginal posterior was plotted with random Poisson data which were generated for various n and μ . Results are shown in Figures 5.4, 5.5, and 5.6. Note that for a given n , the shape of the marginal posterior is roughly the same for different μ when considering the larger plot range. This may be an important fact, as we may later be

able to derive properties of this marginal posterior

Posterior Simulation

Adaptive Rejection Metropolis Sampling (ARMS) was used to sample from the posterior distribution. This method, introduced by Gilks et al. (1995), introduces rejection sampling into the Metropolis algorithm by adapting an envelope function to the density when and where proposals are rejected, with successful proposals being used as proposals in the Metropolis-Hastings algorithm. This technique, while inefficient at sampling from higher dimensional densities, works well in this case and does not require log concavity of the density. The "arms" function in the R package "HI" was used to facilitate simulation. In this methodology, strictly speaking, the support of the target density must be a bounded convex set, however it can be restricted to a bounded set having probability of practically one to achieve satisfactory results. In the following plot, simulated posterior values are generated for the same setup as in the previous section. For the three negative binomial cases, 1000 samples was adequate to see that ARMS can effectively be used to sample from the posterior (5.11).

Simulation Study

To study the performance of the Bayesian structure, we compare it to the MLE in a simulation study. The following tables examine the relative performance in MSE between the posterior mean and median and the MLE. Vectors of count data were created using sample sizes of $n = 10, 30, 200$ to represent small, moderate, and large datasets. The mean μ was varied between 10, 50, and 200. To represent a range of

over-dispersion situations, α was varied not according to fixed values, but to represent prespecified levels of over-dispersion. Recall from (5.3) that the variance of the data Y is inflated by μ^2/α , so μ and α work together to determine the over-dispersion. For example, say $\mu = 10$, $\alpha = 1$ yields $V(Y) = 10 + 100 = 110$, so the variance is $V(Y)/E(Y) = 11$ times larger than the mean. However for $\mu = 200$, $\alpha = 1$ gives $V(Y) = 200 + 40,000 = 40,200$, making the variance $40,200/200 = 201$ times larger than the mean. So instead of specifying α directly, an over-dispersion factor F is specified and α is computed according to the formula

$$\alpha = \frac{\mu}{F - 1},$$

which results in a value for α which can be used to generate data Y such that $E(Y) = \mu$ and $V(Y) = F\mu$. Factors of over-dispersion were set at $F = 10, 5, 2, 1.1$ so as to represent very high, high, moderate, and low levels of overdispersion.

For each (n, μ, α) combination, $K = 500$ datasets were generated. For each dataset, 5000 posterior samples were obtained from (5.11). The MSE of the posterior means and medians for μ were obtained by

$$\frac{1}{K} \sum_{j=1}^K (E[\mu|y_j] - \mu^*)^2,$$

$$\frac{1}{K} \sum_{j=1}^K (\text{median}[\mu_i|y_j] - \mu^*)^2,$$

and similarly for α . These were compared to the MSE of the MLE as computed by the “fitdistr” function in R’s MASS library with a ratio. The results, which are in Tables 5.2, 5.3, 5.4, and 5.5, indicate that the proposed prior in (5.10) performs no

worse at estimating μ compared to the frequentist method, but *substantially* better than the frequentist method when estimating α , particularly when sample sizes are small and/or overdispersion is minimal.

Frequentist coverage was also computed by numerically evaluating

$$\frac{1}{K} \sum_{j=1}^K 1 \left(\int_0^{\mu^*} \pi(\mu|y_j) < \alpha^* \right),$$

for μ and a similar expression for α . This was done for a range α^* between 0 and 1. The results are shown for $n = 5$ in Figure 5.7, for $n = 25$ in 5.8, and for $n = 100$ in Figure 5.9. We can see that, despite superior performance in estimation, frequentist coverage for α can be quite bad, especially for smaller sample sizes. Coverage for μ is great for all sample sizes, and all coverage is good as the sample size becomes large.

5.2 Discrete Parameterization

The negative binomial distribution is a sequence of Bernoulli trials often parameterized as

$$f(x; r, p) = \binom{x+r-1}{x} (1-p)^r p^x. \quad x = 0, 1, \dots; \quad r = 1, 2, \dots; \quad p \in (0, 1), \quad (5.14)$$

where r is the number of failures until the experiment is stopped, p is the probability of a success at each trial, and x is the number of successes observed. This distribution can be used in sequential experiments, and an objective prior distribution has been developed in the case where r is a known constant (Sun and Berger, 2008).

As we proceed with the methodology outlined in 1.4.2, let f_r represent $f(x; r, p)$.

5.2.1 Objective Bayesian Analysis

The Kullback-Leibler divergence between f_r and $f_{r'}$ is

$$\begin{aligned}
D_{KL}(f_r||f_{r'}) &= E_{f_r} \left\{ \log \left(\frac{f_r}{f_{r'}} \right) \right\} \\
&= E_{f_r} \left\{ \log \left[\binom{x+r-1}{x} \right] - \log \left[\binom{x+r'-1}{x} \right] \right\} + \\
&\quad + (r-r') \log(1-p) \\
&= E_{f_r} \left\{ \log \left[\frac{\frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)}}{\frac{\Gamma(x+r')}{\Gamma(x+1)\Gamma(r')}} \right] \right\} + (r-r') \log(1-p) \\
&= E_{f_r} \left\{ \log \left[\frac{\Gamma(x+r)}{\Gamma(x+r')} \right] \right\} + \log \left[\frac{\Gamma(r')}{\Gamma(r)} \right] + (r-r') \log(1-p). \quad (5.15)
\end{aligned}$$

The next step is to find the value of r' that minimizes $D_{KL}(f_r||f_{r'})$. As we will see, this value is $r' = r + 1$. To demonstrate this, it must be shown that $D_{KL}(f_r||f_{r+1}) < D_{KL}(f_r||f_{r+2})$ so that $r' = r + 1$ is the closest for $r' > r$, $D_{KL}(f_r||f_{r-1}) < D_{KL}(f_r||f_{r-2})$ so that $r' = r - 1$ is the closest for $r' < r$, and finally $D_{KL}(f_r||f_{r+1}) < D_{KL}(f_r||f_{r-1})$. We demonstrate the first two with two lemmas, and the last one numerically.

Lemma 5.2.1. For any $r \geq 1$, $D_{KL}(f_r||f_{r+1}) < D_{KL}(f_r||f_{r+2})$.

Proof.

$$\begin{aligned}
D_{KL}(f_r||f_{r+1}) &= E_{f_r} \left\{ \log \left[\frac{\Gamma(x+r)}{\Gamma(x+r+1)} \right] \right\} + \log \left[\frac{\Gamma(r+1)}{\Gamma(r)} \right] + (r-r-1) \log(1-p) \\
&= -E_{f_r} \{ \log(x+r) \} + \log \left[\frac{r}{1-p} \right].
\end{aligned}$$

$$\begin{aligned}
D_{KL}(f_r||f_{r+2}) &= E_{f_r} \left\{ \log \left[\frac{\Gamma(x+r)}{\Gamma(x+r+2)} \right] \right\} + \log \left[\frac{\Gamma(r+2)}{\Gamma(r)} \right] + (r-r-2) \log(1-p) \\
&= -E_{f_r} \{ \log(x+r)(x+r+1) \} + \log[r(r+1)] - 2 \log(1-p).
\end{aligned}$$

So we wish to show $D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r+2}) < 0$,

$$\begin{aligned}
D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r+2}) &= -E_{f_r} \{ \log(x+r) \} + \log(r) - \log(1-p) + \\
&+ E_{f_r} \{ \log(x+r)(x+r+1) \} - \log[r(r+1)] + 2 \log(1-p) \\
&= E_{f_r} \left\{ \log \left[\frac{(x+r-1)}{(x+r-1)(x+r-2)} \right] \right\} + \log(r-2) - \log(1-p) \\
&= -E_{f_r} \{ \log[(x+r-2)] \} + \log(r-2) - \log(1-p).
\end{aligned}$$

□

Lemma 5.2.2. For any $r \geq 3$, $D_{KL}(f_r||f_{r-1}) < D_{KL}(f_r||f_{r-2})$.

Proof. First noting

$$\begin{aligned}
D_{KL}(f_r||f_{r-1}) &= E_{f_r} \left\{ \log \left[\frac{\Gamma(x+r)}{\Gamma(x+r-1)} \right] \right\} + \log \left[\frac{\Gamma(r-1)}{\Gamma(r)} \right] + (r-r+1) \log(1-p) \\
&= E_{f_r} \{ \log(x+r-1) \} - \log(r-1) + \log(1-p),
\end{aligned}$$

and

$$\begin{aligned}
D_{KL}(f_r||f_{r-2}) &= E_{f_r} \left\{ \log \left[\frac{\Gamma(x+r)}{\Gamma(x+r-2)} \right] \right\} + \log \left[\frac{\Gamma(r-2)}{\Gamma(r)} \right] + (r-r+2) \log(1-p) \\
&= E_{f_r} \{ \log[(x+r-1)(x+r-2)] \} \\
&\quad - \log(r-1) - \log(r-2) + 2 \log(1-p).
\end{aligned}$$

So we wish to show $D_{KL}(f_r||f_{r-1}).D_{KL}(f_r||f_{r-2}) < 0$

$$\begin{aligned}
D_{KL}(f_r||f_{r-1}) - D_{KL}(f_r||f_{r-2}) &= E_{f_r} \{ \log(x+r-1) \} - \log(r-1) + \log(1-p) - \\
&\quad - E_{f_r} \{ \log[(x+r-1)(x+r-2)] \} + \log(r-1) + \log(r-2) - 2\log(1-p) \\
&= E_{f_r} \left\{ \log \left[\frac{(x+r-1)}{(x+r-1)(x+r-2)} \right] \right\} + \log(r-2) - \log(1-p) \\
&= - E_{f_r} \{ \log [(x+r-2)] \} + \log(r-2) - \log(1-p).
\end{aligned}$$

Applying Jenson's inequality, we have

$$\begin{aligned}
E_{f_r} \{ \log [(x+r-2)] \} &\geq \log \{ E_{f_r} [x+r-2] \} = \log \left\{ \frac{pr}{1-p} + r-2 \right\} \\
&= \log \left(\frac{r-2+2p}{1-p} \right).
\end{aligned}$$

so that

$$-E_{f_r} \{ \log [(x+r-2)] \} \leq \log(1-p) - \log(r-2+2p).$$

This, along with noting that $p \in (0, 1) \Rightarrow r-2+2p > r-2$, gives us

$$D_{KL}(f_r||f_{r-1}) - D_{KL}(f_r||f_{r-2}) < \log(r-2) - \log(r-2+2p) < 0.$$

□

Lemma 5.2.3. For any $r \geq 2$, $D_{KL}(f_r||f_{r+1}) < D_{KL}(f_r||f_{r-1})$.

Proof. This is the same as showing that $D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r-1}) < 0$.

$$\begin{aligned}
D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r-1}) &= -E_{f_r} \{ \log(x+r) \} + \log(r) - \log(1-p) - \\
&\quad - E_{f_r} \{ \log[(x+r-1)] \} - \log(r-1) + \log(1-p) \\
&= -E_{f_r} \{ \log[(x+r)(x+r-1)] \} + \log[r(r-1)] - 2\log(1-p) \\
&= -E_{f_r} \{ \log[x^2 + x(2r-1) + r(r-1)] \} + \log[r(r-1)] - 2\log(1-p). \quad (5.16)
\end{aligned}$$

Jenson's inequality says that

$$E_{f_r} \{ \log [x^2 + x(2r-1) + r(r-1)] \} \geq \log \{ E_{f_r}(X^2) + E_{f_r}(X)(2r-1) + r(r-1) \}.$$

Noting that $E[X^2] = Var(X) + E[X]^2 = \frac{pr}{(1-p)^2} + \frac{(pr)^2}{(1-p)^2} = \frac{pr(1+pr)}{(1-p)^2}$, we have

$$\begin{aligned}
\log \{ E_{f_r}(X^2) + E_{f_r}(X)(2r-1) + r(r-1) \} &= \log \left\{ \frac{pr(1+pr) + (1-p)pr(2r-1)}{(1-p)^2} \right\} \\
&\quad + \log(r(r-1)),
\end{aligned}$$

so that

$$\begin{aligned}
-E_{f_r} \{ \log [x^2 + x(2r-1) + r(r-1)] \} &\leq -\log \left\{ \frac{pr(1+pr) + (1-p)pr(2r-1)}{(1-p)^2} \right\} \\
&\quad - \log(r(r-1)),
\end{aligned}$$

and therefore

$$D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r-1}) \leq -\log \{ pr(1+pr) + (1-p)pr(2r-1) \}.$$

Unfortunately, $pr(1 + pr) + (1 - p)pr(2r - 1)$ is not always greater than 1 for $r \geq 2$, but it is rather close. For example, if $r = 2$, p must be larger than $\frac{2-\sqrt{14}}{2} \approx .129$ for the inequality to be true. If $r = 3$, $p > \frac{9-5\sqrt{3}}{6} \approx .057$ satisfies it, and for $r = 4$, it drops to $p > \frac{8-\sqrt{61}}{6} \approx .032$. Note, however, that the conjecture $D_{KL}(f_r||f_{r+1}) - D_{KL}(f_r||f_{r-1}) < 0$ may still be true as Jensen's inequality only provides an upper bound for this quantity. To explore this, it was evaluated numerically, with the expectation in (5.16) being evaluated over the support $x \in [0, 10^6]$. First, all combinations of $r \in [2, 500]$ and $p \in [.01, .99]$ in steps of .01 were run and for all combinations the numerical results indicated that the difference represented by (5.16) was never non-negative. Furthermore, an infinite loop was run, where r was sampled from a discrete uniform distribution from 2 to 500, and p was sampled from a $Unif(0, 1)$ distribution. The loop was set to "break" if a non-negative value for (5.16) was ever computed. After 24 hours and approximately 16×10^6 total computations, no non-negative values were found. From this numerical evaluation, we believe that the conjecture $D_{KL}(f_r||f_{r+1}) < D_{KL}(f_r||f_{r-1})$ is true. $\square$

Now, recalling that the objective prior for the discrete parameter θ will have the form

$$\pi(\theta) \propto \exp \left\{ \min_{\theta' \neq \theta \in \Theta} D_{KL}(f(\cdot|\theta)||f(\cdot|\theta')) \right\} - 1,$$

we find the prior for $r|p$ as

$$\begin{aligned} \pi(r|p) &\propto \exp \{ D_{KL}(f_r||f_{r+1}) \} - 1 \\ &= \frac{r}{1-p} \exp \left\{ E_{f_r} \left[\log \left(\frac{1}{x+r} \right) \right] \right\} - 1 \end{aligned}$$

$$= \frac{r}{1-p} \exp \left\{ - \sum_{x=0}^{\infty} \log(x+r) \frac{\Gamma(x+r)}{x! \Gamma(r)} (1-p)^r p^x \right\} - 1. \quad (5.17)$$

We note that the sum in (5.17) may not have a closed form, however, this is a $E[\log(g(x))]$ and instead, a second order Taylor expansion can be used to approximate the expectation in the exponent of the prior. This approach was used in Teh et al. (2006) with great success. A second order Taylor expansion of $E[f(x)]$ around the point $x_0 = E[x]$ is

$$E[f(x)] \approx f(E[X]) + \frac{f''(E[X])}{2} \text{Var}(X). \quad (5.18)$$

Here, $f(x) = \log(x+r)$, so that $\log[E(x+r)] = \log\left(\frac{pr}{1-p} + r\right) = \log\left(\frac{r}{1-p}\right)$. We also have $f''(x) = -\frac{1}{(x+r)^2}$ which gives us $f''(E[x]) = -\frac{(1-p)^2}{r^2}$. Noting that $\text{Var}(X) = \frac{pr}{(1-p)^2}$ and substituting into (5.18) gives us

$$\begin{aligned} \pi(r|p) &\propto \frac{r}{1-p} \exp \left\{ \frac{p}{2r} - \log \left(r + \frac{pr}{1-p} \right) \right\} - 1 \\ &= \exp \left\{ \frac{p}{2r} \right\} - 1. \end{aligned} \quad (5.19)$$

Lemma 5.2.4. The approximation of $\pi(r|p)$ given by (5.19) is improper.

Proof. Expanding the sum and noting the series definition of the exponential function, we have

$$\begin{aligned} \sum_{r=1}^{\infty} \left[\exp \left\{ \frac{p}{2r} \right\} - 1 \right] &= \frac{p}{2} \left(1 + \frac{1}{2} + \frac{1}{3} + \dots \right) + \\ &\quad + \frac{p^2}{2!} \left(\left(\frac{1}{2} \right)^2 + \left(\frac{1}{4} \right)^2 + \left(\frac{1}{6} \right)^2 + \dots \right) + \dots \end{aligned}$$

Note that $(1 + \frac{1}{2} + \frac{1}{3} + \dots) = \infty$, and all other terms in this expression are positive, therefore the entire sum diverges. $\square$

This gives us the posterior

$$\pi(r|x_1, p) = \left(\exp \left\{ \frac{p}{2r} \right\} - 1 \right) \prod_{i=1}^n \binom{x_i + r - 1}{x_i} (1-p)^r p^{x_i}. \quad (5.20)$$

It turns out that this posterior is proper!

Lemma 5.2.5. The posterior distribution in (5.20) is proper after only one observation.

Proof. We have

$$\sum_{r=1}^{\infty} \pi(r|x_1, p) = \sum_{r=1}^{\infty} \left(\exp \left\{ \frac{p}{2r} \right\} - 1 \right) \binom{x_1 + r - 1}{x_1} (1-p)^r p^{x_1}.$$

Each term in the distributed sum can be shown to be finite by noting that

$$\frac{(x_1 + r - 1)!}{(r - 1)!} = (x_1 + r - 1) \cdot (x_1 + r - 2) \cdot \dots \cdot (r) \leq (x_1 + r - 1)^{x_1},$$

which gives us,

$$\sum_{r=1}^{\infty} \exp \left\{ \frac{p}{2r} \right\} \binom{x_1 + r - 1}{x_1} (1-p)^r p^{x_1} \leq \frac{p^{x_1}}{x_1!} \sum_{r=1}^{\infty} \exp \left\{ \frac{p}{2r} \right\} (x_1 + r - 1)^{x_1} (1-p)^r. \quad (5.21)$$

Applying the ratio test to the terms in the sum of the RHS, we have

$$\begin{aligned} \left| \frac{a_{n+1}}{a_n} \right| &= \frac{\exp \left\{ \frac{p}{2(r+1)} \right\} (x_1 + (r+1) - 1)^{x_1} (1-p)^{r+1}}{\exp \left\{ \frac{p}{2r} \right\} (x_1 + r - 1)^{x_1} (1-p)^r} \\ &= \exp \left\{ \frac{-p}{2r(r+1)} \right\} (1-p) \left(\frac{x_1 + r}{x_1 + r - 1} \right)^{x_1}. \end{aligned}$$

Taking the limit, we have

$$\begin{aligned}
\lim_{r \rightarrow \infty} \left| \frac{a_{n+1}}{a_n} \right| &= \lim_{r \rightarrow \infty} \left\{ \exp \left[\frac{-p}{2r(r+1)} \right] (1-p) \left(\frac{x_1 + r}{x_1 + r - 1} \right)^{x_1} \right\} \\
&= (1-p) \cdot \lim_{r \rightarrow \infty} \exp \left[\frac{-p}{2r(r+1)} \right] \cdot \lim_{r \rightarrow \infty} \left(\frac{x_1 + r}{x_1 + r - 1} \right)^{x_1} \\
&= (1-p) \cdot 1 \cdot 1.
\end{aligned}$$

Because $p \in (0, 1) \Rightarrow 1-p < 1$, the sum found in (5.21) for the upper bound of (5.20) converges. □

The posterior distribution in 5.20 was combined with a uniform prior for p and plotted for a random dataset where $r = 5$ and $p = .3$. The result is in Figure 5.10. We can see that there is considerable correlation between r and p . This makes the use of this prior potentially questionable.

5.3 Discussion and Future Work

In this chapter, we explored an application of the numerical method in Berger et al. (2009) to obtain an approximation to the reference prior $\pi(\alpha)$. We note that this prior is likely not equal to the true reference prior for α , but that it ended up having good properties anyway. A similar result was found in an application of the numerical method in Berger et al. (2009) on the model $U(\theta, \theta^2)$. The numerical algorithm resulted in a ‘guess’ for the reference prior for θ as

$$\pi^R(\theta) \propto \frac{1}{\theta - 1}$$

But the analytical solution turned out to be

$$\pi^R(\theta) \propto \frac{2\theta - 1}{\theta(\theta - 1)} \exp \left\{ \psi \left(\frac{2\theta}{2\theta - 1} \right) \right\}.$$

Over their computed range, the two functions were almost identical to one another, and may not yield significantly different results in real data or simulated applications. Finally, with the popularity of Integrated Nested Laplace Approximations (INLA) in recent years, the numerical procedure to obtain estimates for $\pi(\alpha)$ may be able to be used to input into the popular R-INLA package directly.

5.4 Figures and Tables

Table 5.1: Polynomial coefficients of $\sqrt{\alpha}\pi(\alpha)$ and taylor series coefficients of $\psi(\alpha+1)$

	Intercept	α	α^2
Polynomial Regression	1.485342	-2.4952	4.204111
Taylor Series	1.644934	-2.40411	3.24697

Table 5.2: Ratio of MSEs for $E[\mu|y]$ and $\hat{\mu}$

		(smaller is better)								
		$n = 5$			$n = 25$			$n = 100$		
		$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$
F	20	48.52	99.71	99.99	99.92	100.97	100.80	100.38	99.88	99.53
	5	100.59	99.27	99.88	100.67	100.16	100.75	100.06	100.65	100.09
	2	100.01	100.17	99.57	99.90	100.73	100.31	100.74	100.08	100.09
	1.5	99.89	99.37	99.69	99.17	100.22	100.10	100.30	99.20	100.01
	1.25	99.04	99.21	99.07	100.46	99.86	99.90	100.37	99.45	99.31

Table 5.3: Ratio of MSEs for $median[\mu|y]$ and $\hat{\mu}$

		(smaller is better)								
		$n = 5$			$n = 25$			$n = 100$		
		$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$
F	20	53.01	100.26	100.86	100.07	99.41	100.84	99.09	100.67	100.58
	5	95.16	100.43	99.91	100.58	100.72	100.27	100.85	100.21	100.10
	2	100.87	99.83	100.01	100.80	100.17	100.17	100.19	100.01	100.11
	1.5	99.96	100.95	100.23	100.40	100.02	100.00	100.09	99.10	99.95
	1.25	100.97	100.96	100.42	100.16	99.70	99.91	100.15	99.38	99.38

Table 5.4: Ratio of MSEs for $E[\alpha|y]$ and $\hat{\alpha}$

		(smaller is better)								
		$n = 5$			$n = 25$			$n = 100$		
		$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$
F	20	92.37	79.80	77.43	98.71	95.55	94.16	99.92	99.06	98.31
	5	77.63	73.50	70.80	98.97	99.82	99.59	100.08	99.56	100.04
	2	25.06	24.94	25.74	100.11	100.57	100.20	99.88	99.32	100.28
	1.5	6.04	12.17	27.78	86.89	87.87	98.30	100.67	99.67	99.11
	1.25	10.95	24.21	34.87	57.41	66.09	66.41	99.21	98.29	91.80

Table 5.5: Ratio of MSEs for $median[\alpha|y]$ and $\hat{\alpha}$

		(smaller is better)								
		$n = 5$			$n = 25$			$n = 100$		
		$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$	$\mu = 5$	$\mu = 25$	$\mu = 100$
F	20	68.35	70.40	76.05	97.22	94.49	94.22	100.46	98.14	97.82
	5	63.34	67.26	66.29	93.87	97.91	98.03	100.16	98.21	99.57
	2	23.36	28.05	31.65	84.99	89.66	90.72	95.83	98.11	99.89
	1.5	10.73	22.82	47.85	66.32	70.58	87.25	96.19	95.43	100.30
	1.25	19.14	36.36	47.59	59.17	68.22	69.58	79.62	86.69	85.15

Figure 5.1: Plot of estimated prior $\pi(\alpha)$ for $n = 90$, $c = 1000$, $m = 750$, and $k = 1000$

Figure 5.2: Plot of estimated prior $\pi(\alpha)$ for $n = 90$, $c = 1000$, $m = 750$, and $k = 1000$

(a) $\lambda = 5, \alpha = 1$

(b) $\lambda = 5, \alpha = 5$

(c) $\lambda = 5, \alpha = 10$

(d) $\lambda = 5, \alpha = \infty$

Figure 5.3: Contour plot of $\pi(\lambda, \alpha|y)$ for various levels of overdispersion with $\lambda = 5$

Figure 5.4: Plots of the Marginal Posterior of α for $n = 50$ generated from a Poisson distribution with various λ

Figure 5.5: Plots of the Marginal Posterior of α for $n = 250$ generated from a Poisson distribution with various λ

Figure 5.6: Plots of the Marginal Posterior of α for $n = 500$ generated from a Poisson distribution with various λ

(a) $\mu = 5, n = 5$

(b) $\mu = 25, n = 5$

(c) $\mu = 100, n = 5$

Figure 5.7: Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 5$

(a) $\mu = 5, n = 25$

(b) $\mu = 25, n = 25$

(c) $\mu = 100, n = 25$

Figure 5.8: Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 25$

(a) $\mu = 5, n = 100$

(b) $\mu = 25, n = 100$

(c) $\mu = 100, n = 100$

Figure 5.9: Frequentist Coverage for μ and α with various levels of overdispersion factor F for $n = 100$

Figure 5.10: Plot of $\pi(r, p|x)$ using the objective prior for $\pi(r|p)$ and a uniform prior for p

Chapter 6

Summary and Concluding Remarks

In this dissertation, we studied Objective Bayesian approaches to several commonly used statistical procedures which have been untouched by the Objective Bayesian Community. As was seen, Objective Bayesian Analysis often outperforms frequentist procedures when evaluated by typically frequentist metrics, even for very simple statistical structures which have been well established for a very long time. In Chapter 3, we developed an elegant Bayesian structure for reference analysis of various measures of association for a 2×2 and discovered a wide variety of interesting properties for the various pieces of the puzzle. In Chapter 4 we presented a hierarchical posterior structure which led to the development of a highly efficient posterior sampler. We also demonstrated our priors superiority to the frequentist procedure both under simulation and also real examples, with a chief advantage being the powerful ability to not let 0 cell counts in our table stop us from conducting analysis. Posterior sampling on a GPU via CUDA demonstrated that, despite the lack of flexibility available to the programmer, GPUs have a mind boggling amount

of computational power, and could represent a new chapter for all of scientific computing. The simple structure of the 2×2 table also allowed us to consider a wide array of interesting datasets. From a motivating (and silly) example the author took from his undergraduate students, to a table with only one value, to Fisher’s tea tasting experiment. Furthermore, an easy to use R package has been developed to apply the reference prior theory to any 2×2 table. The author believes that the almost “automatic” nature of Objective Bayesian Analysis is a cornerstone of statistics and is excited about the potential for this elegant and powerful structure. The future of this project is likely to be rich mathematically and philosophically, with new discoveries about the structure of the problem being made often. We note that special functions such as the hypergeometric function appeared a lot in this chapter, and it is the authors belief that the statistical focus of this work is only scratching the surface of the mathematical structure of this problem. In particular, recall that $G(\theta)$ and $H(\theta)$ share the property $f\left(\frac{1}{x}\right) = \sqrt{x}f(x)$. This exact type of symmetry is present in the Jacobi Θ function, and this is used to prove the canonical reflection formula for the Reimann ζ function. In the future, we may be able to find some deeper mathematical meanings from these results. Also in this Chapter, an effort was made to conduct sampling in an extremely efficient way. In Chapter 5, we overcame the analytical difficulties with the function form of a mean and dispersion parameterization of the Negative Binomial distribution to derive an approximation to the reference prior which allowed us to more efficiently estimate the dispersion parameter α , and also used a new method to obtain an objective prior for discrete parameter spaces to work with the typical form of the Negative Binomial Distribution. The main challenge of the Objective Bayesian approach is that objective priors are often very difficult to compute analytically,

requiring considerable effort from a specialized group of researchers. As we saw in this dissertation, even very simply stated statistical structures can lead to a mountains of mathematical challenges. The advantage to overcoming these challenges, however, is that the Bayesian structure is now defined for anyone who wishes to apply a Bayesian approach to these methodologies. In the future, we believe that the numerical method used to approximate the reference prior in this chapter could be somehow programmed into the “black box” of statistics and used to conduct objective Bayesian Analysis.

Appendix A

Miscellaneous Results

A.1 Closed Form for $G(\theta)$

Note that (3.16) can be written as

$$= \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{2} \frac{1}{\sqrt{\theta}} \int_0^1 \frac{\log [\eta_1(1 - \eta_1)]}{\sqrt{\eta_1(1 - \eta_1)(1 - z\eta_1)}} d\eta_1 \right\}$$

where $z = \frac{\theta-1}{\theta}$. Using (3.18) and choosing $a = \frac{1}{2}$, $b = T + \frac{1}{2}$, and $c = T + 1$, we have

$$B\left(T + \frac{1}{2}, \frac{1}{2}\right) {}_2F_1\left(\frac{1}{2}, T + \frac{1}{2}; T + 1; z\right) = \int_0^1 \frac{\eta_1^T}{\sqrt{\eta_1(1 - \eta_1)(1 - z\eta_1)}} d\eta_1 \quad (\text{A.1})$$

If we differentiate the RHS of (A.1) with respect to T then set $T = 0$ we see that

$$\frac{d}{dT} \int_0^1 \frac{\eta_1^T}{\sqrt{\eta_1(1 - \eta_1)(1 - z\eta_1)}} d\eta_1 = \int_0^1 \frac{\eta_1^T \log \eta_1}{\sqrt{\eta_1(1 - \eta_1)(1 - z\eta_1)}} d\eta_1 \Big|_{T=0}$$

$$= \int_0^1 \frac{\log \eta_1}{\sqrt{\eta_1(1-\eta_1)}(1-z\eta_1)} d\eta_1 \quad (\text{A.2})$$

To differentiate the LHS of (A.1), note that

$$\frac{d}{dx}\Gamma(x) = \Gamma(x)\psi(x)$$

where $\psi(x)$ is the *digamma* function, and by extension of this we have

$$\frac{d}{dx}B(x, y) = B(x, y)[\psi(x) - \psi(x + y)].$$

Using these two facts, we have

$$\begin{aligned} & \frac{d}{dT} \left[B \left(T + \frac{1}{2}, \frac{1}{2} \right) {}_2F_1 \left(\frac{1}{2}, T + \frac{1}{2}; T + 1; z \right) \right] \Big|_{T=0} = \\ & {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) \left\{ B \left(\frac{1}{2}, \frac{1}{2} \right) \left[\psi \left(\frac{1}{2} \right) - \psi(1) \right] \right\} \\ & + B \left(\frac{1}{2}, \frac{1}{2} \right) \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, T + \frac{1}{2}; T + 1; z \right) \Big|_{T=0} \end{aligned}$$

Using (3.17), we have

$$\begin{aligned} \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, T + \frac{1}{2}; T + 1; z \right) &= \frac{d}{dT} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(T + \frac{1}{2}\right)_n}{(T + 1)_n} \frac{z^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{\Gamma \left(n + \frac{1}{2} \right)}{\Gamma \left(\frac{1}{2} \right)} \frac{z^n}{n!} \cdot \frac{d}{dT} \left[\frac{\Gamma \left(T + n + \frac{1}{2} \right)}{\Gamma \left(T + \frac{1}{2} \right)} \frac{\Gamma \left(T + 1 \right)}{\Gamma \left(T + n + 1 \right)} \right]. \end{aligned}$$

Noting that

$$\frac{d}{dT} \left[\frac{g_1(T)g_2(T)}{g_3(T)g_4(T)} \right] \Big|_{T=0} = \frac{g_1(T)g_2(T)}{g_3(T)g_4(T)} \Big|_{T=0} \cdot \frac{d}{dT} \log \left[\frac{g_1(T)g_2(T)}{g_3(T)g_4(T)} \right] \Big|_{T=0}$$

we have

$$\begin{aligned}
& \left. \frac{d}{dT} \left[\frac{\Gamma(T+n+\frac{1}{2})}{\Gamma(T+\frac{1}{2})} \frac{\Gamma(T+1)}{\Gamma(T+n+1)} \right] \right|_{T=0} = \\
& = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)} \left[\left. \frac{d}{dT} \log \Gamma\left(T+n+\frac{1}{2}\right) \right|_{T=0} - \left. \frac{d}{dT} \log \Gamma\left(T+\frac{1}{2}\right) \right|_{T=0} + \right. \\
& \quad \left. + \left. \frac{d}{dT} \log \Gamma(T+1) \right|_{T=0} - \left. \frac{d}{dT} \log \Gamma(T+n+1) \right|_{T=0} \right] \\
& = \frac{1}{\sqrt{\pi}} \frac{\Gamma(n+\frac{1}{2})}{\Gamma(n+1)} \left[\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n+\frac{1}{2}\right) - \psi(n+1) \right]
\end{aligned}$$

Noting that $\psi\left(\frac{1}{2}\right) - \psi(1) = -2\log 2$, the differentiation of the LHS of (A.1) with respect to T and then setting $T = 0$ yields

$$\begin{aligned}
& \int_0^1 \frac{\log \eta_1}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 = \\
& - 2\pi \log 2 \cdot {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; z\right)
\end{aligned} \tag{A.3}$$

$$+ \sum_{n=0}^{\infty} \frac{\Gamma^2(n+\frac{1}{2})}{\Gamma(n+1)} \frac{z^n}{n!} \left(\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n+\frac{1}{2}\right) - \psi(n+1) \right) \tag{A.4}$$

Proceeding similarly we can choose $a = \frac{1}{2}$, $b = \frac{1}{2}$, and $c = T + 1$, we have

$$B\left(\frac{1}{2}, T + \frac{1}{2}\right) {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; T + 1; z\right) = \int_0^1 \frac{(1-\eta_1)^T}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \tag{A.5}$$

If we differentiate the RHS of (A.5) with respect to T then set $T = 0$ we see that

$$\begin{aligned}
\frac{d}{dT} \int_0^1 \frac{(1-\eta_1)^T}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 & = \int_0^1 \frac{(1-\eta_1)^T \log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \Big|_{T=0} \\
& = \int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1
\end{aligned} \tag{A.6}$$

Using the derivative equations from earlier, we have

$$\begin{aligned} & \frac{d}{dT} \left[B \left(T + \frac{1}{2}, \frac{1}{2} \right) {}_2F_1 \left(\frac{1}{2}, T + \frac{1}{2}; T + 1; z \right) \right] \Big|_{T=0} = \\ & {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) \left\{ B \left(\frac{1}{2}, \frac{1}{2} \right) \left[\psi \left(\frac{1}{2} \right) - \psi(1) \right] \right\} \\ & + B \left(\frac{1}{2}, \frac{1}{2} \right) \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; T + 1; z \right) \Big|_{T=0} \end{aligned}$$

Again using (3.17), we have

$$\begin{aligned} \frac{d}{dT} {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; T + 1; z \right) &= \frac{d}{dT} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(T+1)_n} \frac{z^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right)}{\Gamma^2 \left(\frac{1}{2} \right)} \frac{z^n}{n!} \cdot \frac{d}{dT} \left[\frac{\Gamma(T+1)}{\Gamma(T+n+1)} \right] \end{aligned}$$

Using the same approach as was used to compute (A.5), we have

$$\frac{d}{dT} \left[\frac{\Gamma(T+1)}{\Gamma(T+n+1)} \right] \Big|_{T=0} = \frac{\psi(1) - \psi(n+1)}{\Gamma(n+1)}$$

which yields so that

$$\begin{aligned} & \int_0^1 \frac{\log(1-\eta_1)}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 = \\ & -2\pi \log 2 \cdot {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) + \sum_{n=0}^{\infty} \frac{\Gamma^2 \left(n + \frac{1}{2} \right)}{\Gamma(n+1)} \frac{z^n}{n!} (\psi(1) - \psi(n+1)) \end{aligned} \quad (\text{A.7})$$

combining (A.3) and (A.7) we have

$$\begin{aligned} & \int_0^1 \frac{\log[\eta_1(1-\eta_1)]}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 = \\ & -4\pi \log 2 \cdot {}_2F_1 \left(\frac{1}{2}, \frac{1}{2}; 1; z \right) \end{aligned} \quad (\text{A.8})$$

$$+ \sum_{n=0}^{\infty} \frac{\Gamma^2\left(n + \frac{1}{2}\right)}{\Gamma(n+1)} \frac{z^n}{n!} \left(2\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n + \frac{1}{2}\right) - 2\psi(n+1) \right) \quad (\text{A.9})$$

Continuing from (3.16) and recalling the form of $G(\theta)$ presented in (3.19), and inserting the value of the expectation in (A.8), we have

$$\begin{aligned} \pi(\theta) &\propto \frac{1}{\theta} \exp \left\{ \frac{G(\theta)}{2} \frac{1}{\sqrt{\theta}} \int_0^1 \frac{\log [\eta_1(1-\eta_1)]}{\sqrt{\eta_1(1-\eta_1)(1-z\eta_1)}} d\eta_1 \right\} \\ &= \frac{1}{\theta} \exp \left\{ -2 \log 2 + \frac{1}{2\pi {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; z\right)} \sum_{n=0}^{\infty} \frac{\Gamma^2\left(n + \frac{1}{2}\right)}{\Gamma(n+1)} \frac{z^n}{n!} \times \right. \\ &\quad \left. \times \left(2\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n + \frac{1}{2}\right) - 2\psi(n+1) \right) \right\} \\ &= \frac{1}{\theta} \exp \left\{ -2 \log 2 + \frac{1}{2 {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; z\right)} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} \times \right. \\ &\quad \left. \times \left(2\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n + \frac{1}{2}\right) - 2\psi(n+1) \right) \right\} \quad (\text{A.10}) \end{aligned}$$

Furthermore, note that the digamma function has the following useful properties, among others

$$\begin{aligned} \psi(n+1) &= \psi(1) + \sum_{k=1}^n \frac{1}{k}, \\ \psi\left(n + \frac{1}{2}\right) &= \psi\left(\frac{1}{2}\right) + \sum_{k=1}^n \frac{2}{2k-1} \end{aligned}$$

so that

$$2\psi(1) - \psi\left(\frac{1}{2}\right) + \psi\left(n + \frac{1}{2}\right) - 2\psi(n+1) = \sum_{k=1}^n \frac{2}{2k-1} - 2 \sum_{k=1}^n \frac{1}{k}$$

So that

$$\pi(\theta) \propto \frac{1}{\theta} \exp \left\{ \frac{1}{{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; 1; z\right)} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} \left[\sum_{k=1}^n \frac{2}{2k-1} - 2 \sum_{k=1}^n \frac{1}{k} \right] \right\} \quad (\text{A.11})$$

Now, note that we have

$$\sum_{k=1}^n \frac{2}{2k-1} - \sum_{k=1}^n \frac{1}{k} = 2 \log 2 + H_{n-1/2} - H_n = 2 \log 2 - 2 \int_0^1 \frac{t^{2n}}{1+t} dt$$

where $H_n = \sum_{k=1}^n \frac{1}{k}$ represents the n -th *harmonic number*. This tells us that the summation in (A.11) can be written as

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} \left[\sum_{k=1}^n \frac{2}{2k-1} - 2 \sum_{k=1}^n \frac{1}{k} \right] &= 2 \log 2 \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} + \\ &+ \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} \left(-2 \int_0^1 \frac{t^{2n}}{1+t} dt \right) - \\ &- \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} H_n \end{aligned} \quad (\text{A.12})$$

The middle term can be rearranged as

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} \left(-2 \int_0^1 \frac{t^{2n}}{1+t} dt \right) &= -2 \int_0^1 \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{(zt^2)^n}{n!} \cdot \frac{1}{1+t} dt \\ &= -2 \int_0^1 \frac{{}_2F_1\left(\frac{1}{2}, \frac{1}{2}, 1, zt^2\right)}{1+t} dt \end{aligned}$$

Recalling the definition of the hypergeometric function in (3.17), we can see that the quantity in (A.12) exactly equals

$$2 \log 2 {}_2F_1\left(\frac{1}{2}, \frac{1}{2}, 1, z\right) - 2 \int_0^1 \frac{{}_2F_1\left(\frac{1}{2}, \frac{1}{2}, 1, zt^2\right)}{1+t} dt - \sum_{n=0}^{\infty} \frac{\left(\frac{1}{2}\right)_n \left(\frac{1}{2}\right)_n}{(1)_n} \frac{z^n}{n!} H_n$$

A.2 The Jeffreys Rule Prior for the Odds Ratio θ

Recall from (2.9) that the posteriors for $\lambda_i|n_i$ are $Gamma(n_i + \frac{1}{2}, 1)$ for the Reference prior on $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$. Because for the joint independent Poisson model the Fisher information matrix is $diag(\lambda_1^{-1}, \lambda_2^{-1}, \lambda_3^{-1}, \lambda_4^{-1})$, the Jeffreys prior is identical. Furthermore, recall that Jeffreys prior is invariant to 1-1 transformations. Because of this, we have the following theorem.

Theorem A.2.1. The marginal posterior for $\theta|\mathbf{n}$ under Jeffreys prior is

$$\begin{aligned} \pi(\theta|\mathbf{n}) = & \frac{B(n_2 + n_4 + 1, n_1 + n_3 + 1)}{B(n_1 + 1/2, n_2 + 1/2)B(n_3 + 1/2, n_4 + 1/2)} \\ & \times \theta^{n_4-1/2} {}_2F_1(n_3 + n_4 + 1, n_2 + n_4 + 1; n. + 2; 1 - \theta) \end{aligned} \quad (\text{A.13})$$

Proof. First let $X_j \sim Gamma(n_j + 1/2, 1)$ so that we can express θ by

$$\theta = R_1 R_2$$

where $R_1 = \frac{X_1}{X_2}$ and $R_2 = \frac{X_4}{X_3}$. By the properties of the gamma distribution, R_1 and R_2 each have a standard beta prime distribution with pdfs

$$\begin{aligned} f_{R_1}(r_1) &= \frac{r_1^{n_1-1/2} (1 + r_1)^{-(n_1+n_2+1)}}{B(n_1 + 1/2, n_2 + 1/2)}, r_1 > 0 \\ f_{R_2}(r_2) &= \frac{r_2^{n_4-1/2} (1 + r_2)^{-(n_3+n_4+1)}}{B(n_3 + 1/2, n_4 + 1/2)}, r_2 > 0 \end{aligned}$$

Now, to find the distribution of the odds ratio $\theta = R_1 R_2$, we use the standard way to

find the formula of a product distribution

$$\begin{aligned}
\pi^J(\theta) &= \int_{-\infty}^{\infty} f_{R_1}(r_1) f_{R_2}(\theta/r_1) \frac{1}{|r_1|} dr_1 \\
&= c \int_0^{\infty} r_1^{n_1-1/2} (1+r_1)^{-(n_1+n_2+1)} (\theta/r_1)^{n_4-1/2} (1+\theta/r_1)^{-(n_3+n_4+1)} dr_1 \\
&= c \cdot \theta^{n_4-1/2} \int_0^{\infty} r_1^{n_1-n_4-1} (1+r_1)^{-(n_1+n_2+1)} (1+\theta/r_1)^{-(n_3+n_4+1)} dr_1 \\
&= c \cdot \theta^{n_4-1/2} \int_0^{\infty} r_1^{n_1+n_3} (1+r_1)^{-(n_1+n_2+1)} (\theta+r_1)^{-(n_3+n_4+1)} dr_1, \quad (\text{A.14})
\end{aligned}$$

where $c = [B(n_1 + 1/2, n_2 + 1/2)B(n_3 + 1/2, n_4 + 1/2)]^{-1}$. The result now follows from an application of (3.18) to the integral in (A.14). $\square$

It is remarkable to note that this marginal posterior is always proper for all $n_j \geq 0$, even when $n_1 = n_2 = n_3 = n_4 = 0$! In particular, note the following form.

Lemma A.2.2. If $n_1 = n_2 = n_3 = n_4 = 0$,

$$\pi(\theta|\mathbf{n}) = \frac{1}{\pi^2} \frac{\log(\theta)}{\sqrt{\theta}(\theta-1)} \quad (\text{A.15})$$

Proof. Substituting $n_1 = n_2 = n_3 = n_4 = 0$, we can quickly see

$$\pi(\theta|\mathbf{n}) = \frac{B(1,1)}{B(1/2, 1/2)B(1/2, 1/2)} \theta^{-1/2} {}_2F_1(1, 1; 2; 1-\theta) \quad (\text{A.16})$$

Next, from the properties of the hypergeometric function, we have that

$$\log(1+z) = z {}_2F_1(1, 1; 2; -z) \quad (\text{A.17})$$

Using (A.17) with $z = \theta - 1$ in (A.16) proves the result. $\square$

If we desire to obtain a posterior sample for the odds ratio θ under the Jeffreys prior, we can conduct the following two steps:

1. For a given 2×2 table, sample $\lambda_i \sim \text{Gamma}(n_i + \frac{1}{2})$ for $i = 1, 2, 3, 4$, and
2. Compute $\theta = \frac{\lambda_1 \lambda_4}{\lambda_2 \lambda_3}$.

This straightforward procedure provides a good baseline for comparison against other Bayesian procedures.

A.3 Numerical Computation of $G(\theta)$

Recall that $G(\theta)$ is defined as

$$G(\theta) = \int_0^1 \frac{\log \left[\sqrt{t(1-t)} \right] dt}{\sqrt{t(1-t)(\theta(1-t) + t)}}.$$

Because there is no numerical integration procedure is available for the GPU architecture, computation of $G(\theta)$ was an issue due to the singularities of the integrands.

$$\frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t) + t)}} \approx \begin{cases} \frac{\log \sqrt{t}}{\sqrt{\theta t}} & t \in (0, \varepsilon) \\ \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t) + t)}} & t \in (\varepsilon, 1 - \varepsilon) \\ \frac{\log \sqrt{1-t}}{\sqrt{1-t}} & t \in (1 - \varepsilon, 1) \end{cases}$$

when then gives us

$$G(\theta) \approx \int_0^\varepsilon \frac{\log \sqrt{t}}{\sqrt{\theta t}} dt + \int_\varepsilon^{1-\varepsilon} \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t) + t)}} dt + \int_{1-\varepsilon}^1 \frac{\log \sqrt{1-t}}{\sqrt{1-t}} dt$$

$$\begin{aligned}
&= \frac{\sqrt{\varepsilon}(\log \varepsilon - 2)}{\sqrt{\theta}} + \sqrt{\varepsilon}(\log \varepsilon - 2) + \int_{\varepsilon}^{1-\varepsilon} \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t) + t)}} dt \\
&= \sqrt{\varepsilon}(\log \varepsilon - 2) \frac{1 + \sqrt{\theta}}{\sqrt{\theta}} + \int_{\varepsilon}^{1-\varepsilon} \frac{\log \left[\sqrt{t(1-t)} \right]}{\sqrt{t(1-t)(\theta(1-t) + t)}} dt \tag{A.18}
\end{aligned}$$

The second term in (A.18) has no singularities, so it can be integrated numerically with relatively simple techniques.

Bibliography

A. Agresti. On logit confidence intervals for the odds ratio with small samples. *Biometrics*, 55:597602, 1999.

A. Agresti. *Categorical Data Analysis*. John Wiley and Sons, 3 edition, 2013. ISBN 9780470463635.

A. Agresti and D.B. Hitchcock. Bayesian inference for categorical data analysis. *Stat Methods Appt*, 14:297330, 2005.

A. Agresti and M. Yang. An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, 5:921, 1987.

H. Bateman. *Higher Transcendental Functions*. McGraw-Hill Book Company, Inc., 1953.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

James O. Berger and Jose M. Bernardo. Ordered group reference priors with application to the multinomial problem. *Biometrika*, 79(1):25–37, 1992a.

- James O. Berger, Victor De Oliveira, and Bruno Sanso. Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.
- James O. Berger, Jose M. Bernardo, and Dongchu Sun. The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938, 2009.
- J.O. Berger. *Statistical decision theory and Bayesian analysis*. Number 1. Springer-Verlag Inc., 3 edition, 1985. ISBN 9780387960982.
- J.O. Berger. The case for objective bayesian analysis. *Bayesian Statistics*, 1:1–17, 2004.
- J.O. Berger and J.M. Bernardo. On the development of reference priors. *Bayesian Statistics*, 4(1):35–60, 1992b.
- J.O. Berger, J. Strawderman, and D. Tang. Posterior propriety and admissibility of hyperpriors in normal hierarchical models. *Annals of Statistics*, 33:606–646, 2005.
- J.O. Berger, J. Bernardo, and D. Sun. Reference priors for discrete parameter spaces. *Journal of the American Statistical Association*, 107(3):636–648, 2012.
- J.M. Bernardo. Reference posterior distributions for bayesian reference. *Journal of the Royal Statistical Society*, 41:113–147, 1979.
- J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Wiley, 2000.
- G.S. Datta and M. Ghosh. On the invariance of noninformative priors. *The Annals of Statistics*, 24(1):141–159, 1996.

- G. Erdogan and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, 1974.
- S.E. Feinberg. When did bayesian inference become "bayesian"? *Bayesian Analysis*, 1(1):1–40, 2006.
- R.A. Fisher. *The Design of Experiments*. Oliver And Boyd; Edinburgh; London, 1937.
- R.A. Fisher. *The Design of Experiments*. Macmillan, 9 edition, 1971. ISBN 0-02-844690-9.
- S. French. On the axiomatization of subjective probabilities. *Theory and Decision*, 14:1933, 1982.
- J.J. Gart and J.R. Zweifel. On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika*, 14:181187, 1967.
- W.R. Gilks, N.G. Best, and K.C. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society*, 44(4):455–472, 1995.
- M. Goldstein. Subjective bayesian analysis: Principles and practice. *Bayesian Analysis*, 3:403420, 2006.
- I.S. Gradshteyn and I.M. Ryzhik. *Table of Integrals, Series, and Products*. Elsevier, 7 edition, 2007.
- J.B.S. Haldane. The estimation and significance of the logarithm of a ratio frequencies. *Annals of Human Genetics*, 20:309311, 1955.
- H. Jeffreys. *Theory of Probability*. London: Oxford University Press, 1961.

- R.E. Kass and J. Greenhouse. Investigating therapies of potentially great benefit: Ecmo. comment on paper by j.h ware. *Statistical Science*, 4(4):310–314, 1989.
- A. J. Kinderman and J. F. Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):257–260, 1977.
- S.G. Kou and Z. Ying. Asymptotics for a 2x2 table with fixed margins. *Statistica Sinica*, 44(6):809–829, 1996.
- Jerald F. Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15(3):209–225, 1987.
- V. Lefevres. Introduction to the gnu mpfr library. GNU Hackers Meeting, Paris, 2011.
URL <https://www.gnu.org/ghm/2011/paris/slides/vincent-lefevre-mpfr.pdf>.
- D.V. Lindley. The bayesian analysis of contingency tables. *Annals of Mathematical Statistics*, 35:1622–1643, 1964.
- D.V. Lindley. *Bayesian Statistics: A Review*. Society for Industrial and Applied Mathematics, 1987.
- D.V. Lindley. On the measure of information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1992.
- B. Liseo. Elimination of nuisance parameters with reference priors. *Biometrika*, 80(2):295–304, 1993.

- M. Machler. Arbitrarily accurate computation with r: The rmpfr package. 2012. URL <https://cran.r-project.org/web/packages/Rmpfr/vignettes/Rmpfr-pkg.pdf>.
- G. Makdisi and I. Wang. Extra corporeal membrane oxygenation (ecmo) review of a lifesaving technology. *Journal of Thoracic Disease*, 7(7):166–176, 2015.
- Neri Merhav and Mike Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.
- F.W. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.
- M. Parzen, S. Lipsitz, J. Ibrahim, and N. Klar. An estimate of the odds ratio that always exists. *Journal of Computational and Graphical Statistics*, 11(2):420–436, 2002.
- F.P. Ramsey. *Studies in Subjective Probability*. Wiley, New York, 1964. ISBN 97811131329918.
- C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 1948.
- S. M. Stigler. Memoir on the probability of the causes of events. *Statistical Science*, 1(3):364–378, 1974.
- D. Sun and J.O. Berger. Reference priors with partial information. *Biometrika*, 85(1):55–71, 1998.

- D. Sun and J.O. Berger. Objective bayesian analysis under sequential experimentation. *IMS Collections, Pushing The Limits of Contemporary Statistics: Contributions in Honour of Jayanta K. Ghosh*, 3:16–32, 2008.
- D. Sun and K. Ye. Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika*, 83(1):55–65, 1996.
- Y.W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. *NIPS 2006*, page 13531360, 2006.
- N.M. Temme. Large parameter cases of the gauss hypergeometric function. *Journal of Computational and Applied Mathematics*, 153:441–462, 2003.
- A.R. Vargas. Boundary asymptotics for power series, September 2014.
- C. Villa and S.G. Walker. An objective approach to prior mass functions for discrete parameter spaces. *To appear in Journal of the American Statistical Association*, 2014.
- J.H. Ware. Investigating therapies of potentially great benefit: Ecmo. *Statistical Science*, 4(4):298–306, 1989.
- G.N. Watson. Asymptotic expansions of hypergeometric functions. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22:277308, 1918.

VITA

John Snyder was born on April 4th 1985 in Latham, New York. He earned a Bachelors degree in Mathematics from The College of St. Rose, in Albany, New York, in 2009. Through most of his undergraduate career, he was an Education major, but discovered a budding passion for statistics through an introduction to probability course taken in the Fall of 2007. From there, he obtained a Masters degree in Statistics from Washington State University in 2011, and in that same year began PhD study at the University of Missouri-Columbia. In the fall of 2012, he met his advisor, Professor Dongchu Sun, and, after many years of intense PhD study, will graduate with a PhD in Statistics in the spring of 2018.

Post graduation, he will begin his career at Monsanto Company as a data scientist, and will marry the love of his life in the fall of 2018.