COMPLEX BAYESIAN MODELS: CONSTRUCTION, 
AND SAMPLING STRATEGIES

by

Carolyn M. Huston
B.Sc., University of Alberta, 2002
M.Sc., University of Alberta, 2005

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE
DEPARTMENT OF STATISTICS AND ACTUARIAL SCIENCE
FACULTY OF SCIENCE

© Carolyn M. Huston 2011
SIMON FRASER UNIVERSITY
Fall 2011

All rights reserved. However, in accordance with the Copyright Act of Canada, this work may be reproduced, without authorization, under the conditions for Fair Dealing. Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review, and news reporting is likely to be in accordance with the law, particularly if cited appropriately.
APPROVAL

Name: Carolyn M. Huston

Degree: Doctor of Philosophy

Title of Thesis: Complex Bayesian Models: Construction, and Sampling Strategies

Examining Committee: Dr. Derek Bingham (Chair)

________________________________________
Dr. Steven Thompson, Senior Supervisor

________________________________________
Dr. Carl Schwarz, Supervisor

________________________________________
Dr. Richard Lockhart, Supervisor

________________________________________
Dr. Joan Hu, Internal Examiner

________________________________________
Dr. Jeff Holt, External Examiner
University of Virginia

Date Approved: October 17, 2011
Declaration of Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the “Institutional Repository” link of the SFU Library website <www.lib.sfu.ca> at: <http://ir.lib.sfu.ca/handle/1892/112>) and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, BC, Canada
Abstract

Bayesian models are useful tools for realistically modeling processes occurring in the real world. In particular, we consider models for spatio-temporal data where the response vector is compositional, i.e. has components that sum-to-one.

A unique multivariate conditional hierarchical model (MVCAR) is proposed. Statistical methods for MVCAR models are well developed and we extend these tools for use with a discrete compositional response. We harness the advantages of an MVCAR model when the response variables of interest are relational, rather than absolute measures. Drawbacks that exist in current modeling approaches for such data are addressed.

Following this, we consider the role of sample selection as a way to support, and to improve the robustness, of Bayesian hierarchical models. We develop guidelines for creating ignorable sampling approaches for complex Bayesian models. This is demonstrated through development of approaches appropriate for our MVCAR model. In particular, a response dependent adaptive approach based on exact sample size requirements for multinomial data is offered.

We initiate a context for considering ‘optimality’ of different sampling methods when the criteria being optimized is a surface, not a scalar. Our optimality evaluation approach is unified with literature about Bayesianly justifiable simulation approaches, including posterior predictive checks.

An example from Fraser River Sockeye salmon fisheries where compositional data provides information about stock run-timings during spawning migration and motivates this work. Such monitoring data with spatial or temporal components occur in a wide variety of applications. Technologies for both measurement and data storage have improved; data are better and there is more of it. Concurrently, society has become more aware of its important relationship to understanding and managing complex natural systems.
Keywords: Compositional data; Multivariate Conditional Autoregression (MVCAR); Bayesian hierarchical model; adaptive sampling; multinomial; exact confidence interval
To my family - thanks.
Acknowledgments

There is a saying that it takes a community to raise a child...well, it also took a community for me to have a successful PhD experience.

Many thanks to Dr. Steven Thompson my supervisor. You were a source of thoughtful advice, and I always appreciated your support in allowing me to explore new ideas while still providing me guidance.

To the other members of my academic committee, I thank them also for their guidance, and for providing me with the opportunity to solicit alternate viewpoints about my studies. Most especially I thank them for being interested in my success.

Other members of the department also deserve my thanks. In particular I want to notice Sadika Jungic who has walked me through many of the intricacies involved with academic administration. Without her help, I would never have been able to complete the forms needed for my scholarships, RAs, TAs, reimbursements, etc.

To the whole administrative group in statistics, I thank them for their unfailing helpfulness, and their many kind words when I came by for assistance.

Outside of academia, I want to thank the friends who supported me throughout, and made sure that I was dragged away from my desk at least once in a while. To Jessica Tilley, Laura Nielson, Fareeza Khurshed, Sandi Ford, and Brenda Black - you rock!

In my extended academic community, I want to thank the team at IRMACS. Special thanks to Pam Borghardt and Kelly Gardiner. It was the attention to detail of the group at IRMACS that often made my life easier so I was able to focus on research instead of distractions. Many thanks also go to the computing technical staff whom I have worked with over the years. Your patience with crashes, software hiccups, and just teaching me new computing skills was life-saving.

Most importantly, a special thank you to my graduate student peers who shared the
ups, downs, uncertainties, and elations of the graduate student experience. Simon Bonner, Elizabeth Juarez, Kelly Burkett, Darby Thompson, Lihui Zhao, Jean Shin, Jeong-Eun Min, Cindy Feng, Crystal Linkletter, Kyle Vincent, Harlan Campbell, Joslin Goh, Ryan Lekivetz - I will miss you.

To all those who should be included in these acknowledgements, if you think that your name belongs in here then no doubt it should. Please consider it so!
# Contents

Approval ii  
Abstract iii  
Dedication v  
Acknowledgments vi  
Contents viii  
List of Tables xi  
List of Figures xiii

## 1 Introduction 1

1.1 Motivations and Example Data ................................................. 2  
1.2 Background On Compositional Data .......................................... 5  
1.3 Common Statistical Methodology .............................................. 8  
  1.3.1 Bayesian Methods .......................................................... 8  
  1.3.2 Sampling ................................................................. 12

## 2 Hierarchical Bayesian Model For Compositions 17

2.1 Background ............................................................... 17  
2.2 Model Development ....................................................... 20  
  2.2.1 Sampling Protocols ...................................................... 20  
  2.2.2 A Hierarchical Model For Estimating Stock Proportions ............ 20
## CONTENTS

2.2.3 Fitting the Model .......................................... 23
2.2.4 Goodness-Of-Fit Assessment .............................. 23

2.3 Example .......................................................... 24
  2.3.1 Motivating Data ............................................ 24
  2.3.2 Parameter Estimation ..................................... 25
  2.3.3 Sensitivity Results ........................................ 29
  2.3.4 Goodness-Of-Fit Results ................................. 30
  2.3.5 Convergence ............................................... 30

2.4 Discussion ........................................................ 33

3 Exact Approach To Multinomial Sample Size .................. 36
  3.1 Background ..................................................... 36
  3.2 Methods And Notation ....................................... 39
    3.2.1 Binomial Definitions .................................. 39
    3.2.2 Binomial Exact Criteria ................................. 40
    3.2.3 Multinomial Definitions ................................. 41
    3.2.4 Multinomial Exact Criteria ............................. 41
    3.2.5 Computational Approach To Worst Case Parameter Vectors ................................. 42
  3.3 Results .......................................................... 46
    3.3.1 Binomial Results ........................................ 46
    3.3.2 Multinomial Results ..................................... 49
  3.4 Discussion ........................................................ 54

4 Bayesian Ignorable Sampling .................................... 57
  4.1 Background ..................................................... 57
    4.1.1 Superpopulation Simulations To Examine Operating Characteristics Of Bayesian Models ............................. 58
    4.1.2 Ignorable Sampling ....................................... 59
  4.2 Notation and Simulation Justification ........................ 60
  4.3 Simulation Model .............................................. 61
  4.4 Ignorable Sampling Approaches .............................. 65
    4.4.1 Conventional and Non-Adaptive Sampling Approaches ................................. 67
    4.4.2 Response Dependent Adaptive Approaches ............................................... 69
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5 Discussion</td>
<td>72</td>
</tr>
<tr>
<td><strong>5 Optimality For Spatial And Other Complex Models</strong></td>
<td>73</td>
</tr>
<tr>
<td>5.1 Justification</td>
<td>75</td>
</tr>
<tr>
<td>5.2 Example</td>
<td>78</td>
</tr>
<tr>
<td>5.2.1 Data Simulation</td>
<td>78</td>
</tr>
<tr>
<td>5.2.2 Discrepancy Results</td>
<td>79</td>
</tr>
<tr>
<td>5.3 Discussion</td>
<td>89</td>
</tr>
<tr>
<td><strong>6 Conclusions</strong></td>
<td>91</td>
</tr>
<tr>
<td>Bibliography</td>
<td>96</td>
</tr>
<tr>
<td><strong>Appendix A</strong> R/WinBUGS Code For Running CH MVCAR Model</td>
<td>103</td>
</tr>
<tr>
<td><strong>Appendix B</strong> Bayesian P-Values For Focused Tests</td>
<td>110</td>
</tr>
<tr>
<td><strong>Appendix C</strong> Binomial and Multinomial Sample Sizes</td>
<td>113</td>
</tr>
<tr>
<td><strong>Appendix D</strong> Discrepancy Graphics</td>
<td>126</td>
</tr>
</tbody>
</table>
List of Tables

2.1 Sample of observed data on stock proportions. Days when sampling does not occur are indicated by rows of NAs. Day refers chronologically to the day during sampling that the observation was taken. The number of fish in the sample is denoted $n_t$, and there are a total of 49 days over which sampling occurs. The remaining columns represent observed stock proportions. 25

2.2 Posterior estimates for standard deviations and correlations of $\varepsilon_i$ terms; standard deviations for the point estimates given in brackets . . . . . . . . 29

3.1 Exact multinomial sample size results when the half interval width (d) is set to 0.05 across all simplexes evaluated. We identify $m$ near where the worst case sample size results are occurring. . . . . . . . . . 54

3.2 Normal approximation multinomial sample size results when the half interval width (d) is set to 0.05 across all simplexes. We identify $m$ where the worst case sample size results are occurring. Based on interval definitions from Thompson (1987). . . . . . . . . . . . . . . . . . . . . . . 54

5.1 Settings used to generate proportion data for simulation study . . . . . . . 78

C.1 Exact multinomial sample size results when the half interval width (d) is set to 0.2 across all simplexes evaluated. We identify $m$ near where the worst case sample size results are occurring. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124

C.2 Exact multinomial sample size results when the half interval width (d) is set to 0.1 across all simplexes evaluated. We identify $m$ near where the worst case sample size results are occurring. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 124
C.3 Exact multinomial sample size results when the half interval width (d) is set to 0.01 across all simplexes evaluated. We identify \( m \) near where the worst case sample size results are occurring. . . . . . . . . . . . . . . . . . 125
# List of Figures

1.1 Fraser River catchment area map ............................................. 2  
1.2 Theoretical run-timing and abundance curves for Fraser River salmon stocks 4  
1.3 A 3-simplex ............................................................................ 6  
1.4 A 3-simplex represented as a ternary plot ................................. 7  

2.1 Graphs of smoothed and frequentist proportions. Grey dots represent observed proportions based on fish counts. The black lines represent estimates based on the mean of the posterior. Hollow black dots represent days where no sampling was taken, and the point estimated has been taken from the MVCAR model. Graphs are ordered in terms of early to late run stocks, with the final graph showing information from all stocks at once. . . 26  
2.2 Comparison of interval estimates based on CH MVCAR highest posterior densities (HPDs) and frequentist intervals based on a multinomial assumption. Black lines with dots represent the point-wise frequentist 95% confidence intervals based on a multinomial assumption. Gaps between bars indicate days where no sampling occurred. The grey shadow represents the corresponding to 95% highest posterior density estimates. Relative interval widths are calculated as (HPD interval width)/(frequentist interval width). . 28
2.3 Traceplots showing generally good convergence results based on observing three independent MCMC chains. The first graph shows the traces for the overall model deviance. The second plot represents traces from MCMC draws on a proportion estimate on a day when fish were sampled. The third plot represents MCMC chains for proportions estimates on a day when fish were not sampled. The final plot represents chains for correlation estimates between stocks 1 and 5; very few observations were taken from stock 1. 32

3.1 Exact highest probability sample sizes required for binomial distributions when $d=0.1$, values of $\theta$ are between 0 and 0.5, and $\alpha$ values are 0.4, 0.1, 0.05, 0.01, 0.001. 48

3.2 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.2$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively small sample sizes). Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues. 50

3.3 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively large sample sizes). Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues. 51

3.4 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Figure shows only the largest 10% of sample size requirements. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively large sample sizes). Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues. 52

4.1 Theoretical run timing curves in lower graph, added together to conceptually show individual stock contributions to overall abundance in the upper graph. 62

4.2 Theoretical relative abundance curves based on converting normal abundances into proportions assuming one sampling event daily. 64
4.3 Theoretical samples based on normal underlying abundance curves converted to proportions. Proportions are then given Dirichlet process error, and multinomial sampling error to reflect real world causes of variability. Missing days occur with the sampling method illustrated, such days appear as gaps in the lines. .......................... 66

5.1 Boxplots of discrepancies between \( P \) and \( \hat{P} \) for 3 stocks and 6 sampling methods. ................................. 83
5.2 Lineplots of discrepancies between \( P \) and \( \hat{P} \) for 6 sampling methods demonstrated using Stock 2. .................. 84
5.3 Boxplots of discrepancies between \( \Pi \) and \( \hat{P} \) for 3 stocks and 6 sampling methods. ................................. 87
5.4 Lineplots of discrepancies between \( \Pi i \) and \( \hat{P} \) for 6 sampling methods demonstrated using Stock 2. .................. 88

C.1 Exact highest probability sample sizes required for binomial distributions when \( d=0.2 \), values of \( \theta \) are between 0 and 0.5, and \( \alpha \) values are 0.4, 0.1, 0.05, 0.01, 0.001. ................................. 114
C.2 Exact highest probability sample sizes required for binomial distributions when \( d=0.05 \), values of \( \theta \) are between 0 and 0.5, and \( \alpha \) values are 0.4, 0.1, 0.05, 0.01, 0.001. ................................. 116
C.3 Exact highest probability sample sizes required for binomial distributions when \( d=0.01 \), values of \( \theta \) are between 0 and 0.5, and \( \alpha \) values are 0.4, 0.1, 0.05, 0.01, 0.001. ................................. 117
C.4 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when \( d = 0.1 \) at \( \alpha = 0.4, 0.1, 0.05, \) and 0.01. Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues. ................................. 119
C.5 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when \( d = 0.05 \) at \( \alpha = 0.4, 0.1, 0.05, \) and 0.01. Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues. ................................. 120
C.6 Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Only largest 10% of sample size requirements illustrated. Graphs shown when \( d = 0.05 \) at \( \alpha = 0.4, 0.1, 0.05, \) and 0.01. Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.

D.1 Lineplots of discrepancies between \( P \) and \( \tilde{P} \) for Stock 1 (simulated data) . . 128
D.2 Lineplots of discrepancies between \( P \) and \( \tilde{P} \) for Stock 3 (simulated data) . . 130
D.3 Lineplots of discrepancies between \( \Pi \) and \( \tilde{P} \) for Stock 1 (simulated data) . . 132
D.4 Lineplots of discrepancies between \( \Pi \) and \( \tilde{P} \) for Stock 3 (simulated data) . . 134
Chapter 1

Introduction

There is a growing awareness in societies of the importance of preserving, protecting, and understanding our natural world due to both tangible and intangible contributions that a healthy environment makes to human quality of life. Tangible measures can include factors such as clean air, water, and access to resources allowing us to produce goods or food. Intangible contributions can include social, cultural, and spiritual roles that the natural environment can shape by supporting traditional lifestyles, or by providing symbols that are culturally or spiritually iconic.

Despite their importance to preserving the natural world, monitoring, modeling, and understanding complex ecological systems are challenging tasks. Ecological data tend to be multivariate, ‘noisy,’ and are often expensive or difficult to sample. Our desire for knowledge is often counterbalanced by financial and practical limitations. In addition to these challenges, animal populations and ecological systems are often evolving in response to anthropogenic influences such as climate change, competition between humans and animals for scarce resources, or due to human consumption of naturally occurring species.

From a statistical perspective, a challenge presented by the ‘noisiness’ of data collected on ecological systems is identifying ways to meaningfully account for, interpret, and manage variation that is observed. Our interest is in estimating parameters of interest, or ‘signal’ in the data; this is done by separating it from error or ‘process’ variation. The research developed in this thesis represents several approaches to managing variation by taking a close look at modeling and sampling approaches that balance model complexity with real world pragmatism. The research presented is motivated by data collected in a single ecological
monitoring program. The methods developed are intended to be more general, and have the potential to be useful across a much broader range of applications.

1.1 Motivations and Example Data

Sockeye salmon are a socially and economically important keystone species on the Pacific West Coast. Understanding the migration patterns of spawning salmon is critical to maintaining healthy commercial, aboriginal, and sport fisheries. Data collection on this complex ecological system is constrained by limited time, financial resources, and often competing social interests.

Once considered the greatest salmon river in the world, historic run sizes in the Fraser River were (and sometimes are still) among the largest in North America. As illustrated in Figure 1.1, the Fraser River catchment area consists of numerous major and minor tributaries, some beginning as far away as the Rocky Mountains. The Fraser River finally empties into the Strait of Georgia, near Vancouver, British Columbia.
Salmon spawn in all of these tributary rivers in the catchment area. It is populations returning to these unique tributaries that make up the commercial and related sockeye salmon fisheries that occur in the Lower Fraser River, Strait of Georgia, and Strait of Juan de Fuca. Sockeye members migrate to their source tributaries, and source tributaries are used to help define the stocks being managed in the fishery. Salmon from different tributary stocks are physically and genetically differentiated due to the reproductive isolation and environmental differences in these different breeding grounds. The fisheries consisting of such distinct stock groups are known as ‘mixed-stock’ because they consist of the many smaller stocks migrating to their tributary points of origin. When the commercial fisheries are opened, there is an almost complete removal of migrating salmon in the river during the length of the opening.

For management and understanding purposes, salmon stocks are often divided into 5 major tributary/stock groups:

1. Early Stuart
2. Early Summer
3. Summer
4. Birkenhead
5. Late Summer.

The ordering chosen here reflects the chronological run-timings of the different stocks. In addition to genetic differences, the characteristic that most dominates the inclusion of a tributary into a stock group is the run-timing of when salmon from the tributary enter the Fraser River. Sockeye salmon spawn in the fall, regardless of their source tributary. The time when fish first enter the mouth of the Fraser River relates to the distance/difficulty of their migration route to spawning grounds. Fish that have further or more difficult travel generally need to enter the river earlier in order to reach their spawning grounds at the appropriate time.

Key tasks for fisheries managers are to ensure both that the overall sockeye population size is maintained in order to support commercial interests on the river; managers also
have to ensure that each of the source stocks maintains an adequate population size. Maintaining the genetic diversity of salmon stocks enhances the species’ ability to respond to environmental change by preserving as many potentially helpful genes as possible.

Figure 1.2 shows a conceptual picture demonstrating theoretical run-timings for the different salmon stocks. Unfortunately, this figure is an inadequate representation of reality. There are two major problems with this representation of run-timings. One is that curves are too idealized and smoothed, which does not reflect the actual day-to-day variability in stock makeup that occurs in actual salmon runs. The second problem is that the picture represents an historic understanding of salmon runs. Environmental changes in terms of climate and landscape are occurring and salmon run timings today behave differently than the curves pictured. Real stock runs exhibit greater variability in stock abundances; runs start earlier or later relative to one another; and the relative abundance of stocks has changed.

Figure 1.2: Theoretical run-timing and abundance curves for Fraser River salmon stocks

Some sockeye salmon stocks that migrate up the Fraser River have been listed as endangered. Understanding salmon migration patterns as they are currently occurring is a key part of managing the fishery so that fish from vulnerable stocks can be protected.

To help manage the stocks, the Pacific Salmon Commission began a sampling program whereby fish in the river are sampled on an approximately daily basis. Fish in the sam-
samples are assigned to one of the five stock groups based on their genetic information. The proportion of fish that come from each stock group on a given day are observed. Over time, this acts as a gauge of run-timings and relative stock sizes. This sampling does not provide information about sockeye salmon abundances. Sample sizes are targeted to be about 50 fish sampled daily. As with much observational ecological data, these data show high amounts of variability that requires careful statistical modeling in order to gain useful inferences about sockeye salmon migration patterns.

1.2 Background On Compositional Data

Multivariate data in the form of a continuous vector of proportions summing to one are often referred to as compositional data. Let \( \Pi \) be a vector of proportions containing \( D \) elements where \( 0 \leq p_i \leq 1 \) \( \forall i = 1, 2, \ldots, D \), and \( \sum_{i=1}^{D} p_i = 1 \). Such data are compositional, and drawn from a sample space which is a \( D - 1 \) dimensional simplex denoted \( \mathcal{S}^{D-1} = \{ \Pi \in \mathbb{R}^D : \Pi' 1 = 1 \} \). This means that although there are \( D \) different proportions in \( \Pi \), due to the positivity \( (0 < p_i < 1 \ \forall \ i = 1, 2, \ldots, D) \) and unit sum constraints, \( \sum_{i=1}^{D} p_i = 1 \), the sample space is a triangular plane (or hyperplane) in \( \mathbb{R}^D \).

Consider a composition consisting of three proportions, \( \Pi = (p_1, p_2, p_3) \), \( \sum_{i=1}^{3} p_i = 1 \). The sample space of such a composition occurs on a planar (two-dimensional) triangle formed by the points \( l_1 = (1, 0, 0) \), \( l_2 = (0, 1, 0) \), and \( l_3 = (0, 0, 1) \) as shown in Figure 1.3. This figure makes it clear that the sample space is a two-dimensional object occurring in three-dimensional space, \( \mathcal{S}^2 = \{ \Pi \in \mathbb{R}^3 : \Pi' 1 = 1 \} \).

The simplex shown in Figure 1.3 is often used without showing its orientation in three dimensional space. Such a plot is shown without data in Figure 1.4. Plots such as this are often referred to as a ternary diagram. Plotting observed points on such a ternary diagram is a common way to visualize three component data. Compositional data with more than three components are more difficult to visualize.

Statistical analyses on such a simplex are challenging. Due to the unit sum constraint, as one element of \( \Pi \) increases, other elements relatively decrease. This can be seen in the correlation/covariance structure of distributions occurring on the simplex such as the multinomial or the Dirichlet. Such constraints do not reflect data where some variables may co-vary positively (Aitchison, 1982; Mosimann, 1962; El-Shaarawi and Piegorsch, 2002).
Problems in estimating correlations when data are relational, such as compositional data, have been recognized as early as 1896 in an article by Karl Pearson (1896-1897). Aitchison (1986); Pawlowsky-Glahn and Olea (2004) have focused on such problems when analyzing compositional data.

A variety of estimation approaches have been proposed for the analysis of compositional data. These methods start with transformations from the simplex $S^{D-1}$ to the real numbers, $\mathbb{R}^{D-1}$. One example is the additive log ratio (ALR) transform $z_i = \log\left(\frac{y_i}{y_1}\right)$ (Aitchison, 1982). Another alternative is to use a latent variable model such as $\log\left(\frac{p_i}{p_1}\right)$. These transformations share common drawbacks. Observed proportions of zero are handled poorly. In the case of the ALR transform, it is not possible to calculate $z_i = \log\left(\frac{0}{y_1}\right) = \log(0)$. In latent variable models, the likelihood can become infinite. Other transformations have similar issues with observed zeros. Additionally, when projecting from the simplex to all the real numbers, the shape of the relationships between variables can be distorted. Finally, it is not always clear how parameters like covariance on the transformed scale with $D - 1$ values in the observation vector should be interpreted on the original scale, which has $D$ components whose covariance is of interest (Billheimer et al, 2001).

A related approach is to create Dirichlet or multinomial-like distributions that have less restrictive variance-covariance structures. This involves creating a compound or conjugate
Dirichlet or multinomial distribution (Grunwald et al, 1993; Mosimann, 1962; Ord et al, 1993; Harvey and Fernandes, 1989). The principles are similar to those when a beta distribution is used with a binomial observation model, allowing for over or underdispersion. Our model follows the spirit of these approaches.

Compositional data similar to that collected in the Fraser River sockeye fishery can occur in a variety of applications. One of these applications is palynology, where counts of different types of pollen are tracked in ice cores and bogs to help to understand vegetation and climate changes through space and time (Niklasson and Lindbladh, 2002). Such counts are often termed a species assemblage. An example in statistical genetics occurs where patterns of genetic variation can be tracked through time and geography, providing insight into processes contributing to genetic drift (Vounatsou et al, 2000). Compositional data are commonly recorded in geologic applications when records consist of counts of different rock types. (Reyment and Savazzi, 1999). In geology, compositional data are synonymous with the sum-to-one problem. When consumers need to choose between a discrete number of different purchasing options, the data recorded are multinomial, and the application is consumer choice (Guimarães and Lindrooth, 2007). Finally, responses in sample surveys are often recorded as categorical response variables (Moura and Migon, 2002).
CHAPTER 1. INTRODUCTION

1.3 Common Statistical Methodology

1.3.1 Bayesian Methods

Bayesian inference is a philosophical approach to inference which allows probabilities and probability statements to specify degrees of belief. This contrasts a frequentist approach where probabilities are associated with the relative frequency that an event occurs in repetitions of an experiment (Silvey, 1975). This occurs because in Bayesian inference we assign prior belief about parameters included in inference. Such prior information summarizes our belief in possible parameter values for the model before observations are made/analyzed.

In its simplest form, we assign some prior distribution \( p(\theta) \) on possible values that the parameters of interest \( \theta \) can take. We then define a sampling distribution (or likelihood when more than one observation is present) as \( p(y|\theta) \) for the data. Symbolically,

\[
p(\theta, y) = p(\theta)p(y|\theta).
\]

By conditioning on the data, \( y \) and applying Bayes’ rule a posterior density can be obtained as follows:

\[
p(\theta|y) = \frac{p(\theta, y)}{p(y)} = p(\theta) \frac{p(y|\theta)}{p(y)}.
\]

Given the observed (fixed) data and the prior, \( p(y) = \int_{\Theta} p(\theta)p(y|\theta) d\theta \) can be considered constant. Because of this, posterior densities are often expressed in an unnormalized form

\[
p(\theta|y) \propto p(\theta)p(y|\theta). \tag{1.1}
\]

Above, Equation 1.1 represents a very simple, standard expression for a Bayesian model; it consists of only the prior and the likelihood. Based on the same principles and application of Bayes rule, we can have a model with more than one level of prior distribution. For example,

\[
p(\theta, \alpha|x) \propto p(y|\theta)p(\theta|\alpha)p(\alpha).
\]
CHAPTER 1. INTRODUCTION

Higher level priors are often referred to as hyperpriors, and models with multiple levels of priors are often called hierarchical Bayesian models. Such hierarchical models can be useful because hyperpriors allow us to incorporate extra belief and structure about the data that is not possible with the likelihood and prior structure alone. Being able to build this kind of hierarchical model allows us a flexible mathematical framework for including important information about the data. For more theoretical background on Bayesian models, or Bayesian hierarchical models, see Carlin and Louis (2009); Gelman et al (2003). For examples of fitted hierarchical models, refer to Gelman and Hill (2007) or Chapter 2.

Historically, one of the challenges for conducting Bayesian inference rested in how to calculate estimates of parameters of interest. In simple problems, such as those involving conjugate pairings of priors and likelihoods, such parameters can be found analytically. As models become more complex due to non-conjugate prior posterior pairings, or through a larger hierarchy, or both, analytic solutions are rarely possible. With increases in computational power, and the introduction of appropriate algorithms, it becomes possible to approximate the posterior and to sample from it. Summary statistics for the distribution can then be inferred based on the samples of the posterior distribution provided by the sampling algorithm. For example,

\[ \bar{g} = \frac{1}{m} \sum_{t=1}^{m} g(\theta^t) \]

where \( \theta^t \) represents a draw from the posterior distribution of \( \theta \) from the \( t^{th} \) time step of the algorithm sampling from the posterior distribution.

The most prevalently used of these approximation algorithms are Markov Chain Monte Carlo (MCMC) methods. Such MCMC chains are based on a random walk that after a sufficient number of steps will be representative of the target distribution.

One commonly used MCMC algorithm is known as the Metropolis-Hastings algorithm (Metropolis et al, 1953; Hastings, 1970). In this method, a Markov chain is initialized by providing values to each of the parameters in the model. New potential parameter values are created from a proposal distribution having density \( R(\theta) \). Whether this parameter draw will be accepted as the next value of the parameter draw is determined with probability
\[ A = \min \left( 1, \frac{p(\theta'|y)R(\theta)}{p(\theta|y)R(\theta')} \right) . \]  

In Equation 1.2, \( \theta \) denotes the current value of the parameter, and \( \theta' \). If the proposal is accepted, the parameter value(s) are updated to \( \theta' \); otherwise they remain at \( \theta \). This process continues for many successive iterations in order to generate a sample adequate to represent the distribution of interest.

A special case of the Metropolis-Hastings method is the Gibbs sampler (Casella and George, 1992; Gelman et al, 2003, pp. 287-289), also known as alternating conditional sampling. In this Markov chain sampling method, \( \theta \) is divided into a vector of sub-parameters \( (\theta_1, \theta_2, ..., \theta_a) \). The Gibbs sampler then cycles through the different \( \theta_j' \) and updates them based on their conditional distribution \( p(\theta_j'|\theta_1', ..., \theta_{j-1}', \theta_{j+1}', ..., \theta_a') \). Because the choice of \( \theta_j' \) is conditional on \( (\theta_1', ..., \theta_{j-1}', \theta_{j+1}', ..., \theta_a') \), it is independent of \( \theta_{j-1}' \) and thus always accepted. This conditional form is also convenient. If our model can be expressed in terms of conditional distributions, or partially conditional distributions, Gibbs methods can be used (Besag et al, 1995).

With all MCMC methods for Bayesian estimation, we only want to consider a subset of samples when the Markov chain is exploring the correct target distribution. Depending on the starting points, it can take many iterations for the sampler to converge on the correct posterior, and even more iterations for the samples from the chain to provide a good approximation of the posterior distribution. Two common ways of assessing convergence of a chain (or multiple chains) is to assess them using traceplots and the Gelman-Rubin-Brooks convergence diagnostic (Gelman and Rubin, 1992; Brooks and Gelman, 1998). The general concept behind these methods is that if multiple chains are run from different starting points, and if the chains ultimately sample from relatively the same region of the parameter space, then they are thought to be converged.

Calibrating and testing Bayesian models are important tasks. Such evaluations of models are often achieved through the use of Bayesianly relevant frequency approaches (Rubin, 1984). Such frequency (often simulation) based evaluations are consistent with a Bayesian modeling philosophy if the calculations are both Bayesianly justifiable and Bayesianly relevant. Bayesianly justifiable implies that known values are considered observations of random variables, and unknown values are unobserved random variables. Conditional dis-
tributions given the knowns and a model are calculated using Bayes’ theorem. Bayesian relevance is achieved if the frequency calculation aids in improving, communicating, or validating solutions based on Bayesian models.

One important application of such frequency calculations is to assess the goodness-of-fit of how well estimates from the specified model fit the observed data. In a Bayesian context, assessing goodness of fit generally relies on the use of posterior predictive checks (Gelman et al, 1996, 2003, pp.157-192). The general idea behind posterior predictive checks is that if we create replicate data based on our model, then it should look similar to the observed data. To do this type of check, we need to generate posterior predictive observation values, $y^{rep}$, creating new ‘data’ based on draws of parameter values from the posterior distribution. A test quantity (or discrepancy measure) is then specified based on these $y^{rep}$ values. Such a test quantity can be used to calculate a Bayesian p-value using the following definition,

$$p_B = Pr[T(y^{rep}, \theta) \geq T(y, \theta) | y].$$ (1.3)

There is a relationship between Bayesian p-values and classical p-values in that both have a goal of evaluating the fit of a model. Both are attempting to quantify differences between the assumed model and the data. At the same time, Bayesian tests are different from classical ones.

Classical theory tends to focus on whether a model fits or not with a goal of evaluating it as correct or incorrect. Such an evaluation is related to Neyman-Pearson style testing where models are tested based on their Type 1 and Type 2 errors. Type 1 and Type 2 errors are not such a focus in Bayesian statistics where analysis is focused more on what degree of belief we have in our models and subsequent inference. Instead, model checking in Bayesian inference can focus on learning in what ways a specified model departs from the data, giving us insight into how such a departure might influence our belief in the model (Gelman, 2003).

With the different outlook that goodness of fit testing in a Bayesian context has, our test quantity $T(y, \theta)$ can be defined much more flexibly than in classical statistics. While summary goodness of fit tests such as a $\chi^2$ can be calculated by defining $T(y, \theta)$ in Equation 1.3 appropriately, it is also possible to base an evaluation of model fit on more focused discrep-
ancy measures. Such measures can even be at a level as detailed as the data itself. Focused discrepancy measures (which will be further elaborated in Chapter 2.2.4) are more difficult to define in the classical goodness of fit context due to what their asymptotic requirements would be.

Further elaborating on the choice of $T(y, \theta)$ in a Bayesian paradigm, $T(y, \theta)$ can be a graphical display (Gelman, 2004). While such a definition does not yield a p-value, graphing can be an extremely valuable way to compare the relationship between model and data. Also, discrepancies discovered in a graphical review of fit might then suggest a discrepancy measure that could be summarized as a Bayesian p-value.

Posterior predictive testing as an approach to evaluating goodness of fit in a Bayesian context has many useful qualities. In statistics, it is common to take the view that models are incorrect, but that they can still be useful in describing the data and quantifying uncertainty. By choosing appropriate posterior predictive quantities, we are able to gain insight past the question of whether the model is correct; it will not be. Instead we can evaluate where the model fits well, and where it fits poorly. We can assess these discrepancies both in terms of their statistical magnitude, and also whether that magnitude represents an important lack of fit in real world application. Posterior predictive measures can also be used to discuss regions of the data where a model makes realistic predictions, and also conditions where it fails to adequately describe the data.

### 1.3.2 Sampling

How data are collected is fundamental to how they are later modeled and interpreted. It is at or prior to the sampling stage that key decisions are made about the research being conducted, and consequently what inference methods will be appropriate. These decisions include defining a population of interest; determining a measurable attribute relevant to our questions of interest based on this population; deciding on an appropriate sampling method for the data (such as simple random sampling); determining a sample size; and actually implementing the sampling and data collection.

By the time data reaches the analysis stage, and statisticians make assertions about it such as ‘observations are *independently and identically distributed* (iid),’ a host of decisions must have been made during sampling for this assertion to hold true. For example, iid
generally implies some sort of randomization mechanism during sampling.

If appropriate decisions about sampling were not made, the iid assumption will not be true, and any inference based on this assumption will be invalid. Similarly, the way randomization is incorporated into sampling also influences how and what analyses should occur. For example, a stratified random sample should be analyzed differently than a randomized cluster sample. Data analysis needs to account for information about data collection.

Much of the literature about sampling methods has been developed in the context of survey sampling (Cochran, 1977; Valliant et al, 2000; Skinner and Holt, 1989). For a resource that discusses sampling in the scientific context, Thompson (2002) provides an excellent overview.

Although simple random sampling (SRS) will generally yield data that can be analyzed using standard statistical methods, it is often not the first choice of sampling method in practice. We often want to take the ‘best’ sample possible with the resources that are available. Here, ‘best’ is evaluated according to level of uncertainty that is associated with the chosen estimator. If we are using an unbiased estimation method and comparing between two sampling strategies, the strategy that has the lowest variance \( \sigma^2(\hat{\theta}) \) is considered the optimal one. In the case of biased estimators, interest focuses on minimizing the mean square error \( (mse) \), \( mse(\hat{\theta}) = \sigma^2(\hat{\theta}) + bias^2(\hat{\theta}, \theta) \). In all cases, the most efficient estimator for the data is desirable.

Where different conventional samplers should be used for optimal results is well described in the literature cited above. More current discussion on optimal sampling strategies also includes so-called adaptive sampling methods (Thompson, 2002; Thompson et al, 1996; Thompson, 2006). In conventional sampling, the selection of units to be included in the sample is determined before sampling begins; in adaptive sampling, the selection of the sample may depend on values of interest that have already been observed. Based on these observed values, the sampling plan may change (‘adapt’). This means that selection of units for the sample occurs in a sequential manner. It involves recording both the data, and the order in which they were selected.

With a known population model, it has been demonstrated that the optimal sampling strategy will generally be adaptive (Thompson et al, 1996). In life statisticians often exist somewhere between having a known model, and having no model at all for data. A consequence of imperfect knowledge is that it might not be possible to identify one optimal
sampling strategy. At the same time, partial knowledge is often enough to suggest useful and easy to implement adaptive sampling approaches that might offer improvement on conventional designs.

As with other forms of estimation, Bayesian models should account for how data were collected in the model (although this is often ignored/forgotten in practice). In Equation 1.1 in Chapter 1.3.1, we expressed an unnormalized posterior density as

\[ p(\theta|y) \propto p(\theta)p(y|\theta). \]

This expression relies only on our priors for \( \theta \) and our observed data likelihood.

In considering sampling methods in a Bayesian context, it is common to conceptualize the analysis in terms of both the data we did observe, and the data that we could have observed (Rubin, 1976; Gelman et al, 2003, pp.197-245). Complete data therefore consists of observed data and missing or potential data. Inference on such complete data involves both the observed data, and the pattern of the missing data. In notation, if we consider \( y = (y_1, \ldots, y_N) \) as the matrix of potential data, then we must also have a vector of indicator variables \( I = (I_1, \ldots, I_N) \). Indicator \( I_i \) takes the value 1 if datum \( y_i \) is observed, and takes the value 0 if \( y_i \) is not observed. This means that our complete data likelihood is expanded from \( p(y|\theta) \) and to

\[ p(y, I|\theta, \phi) = p(y|\theta)p(I|y, \phi). \]  

(1.4)

\( \theta \) represents parameters of interest from the data, and \( \phi \) represents parameters related to inclusion or exclusion in the model. Generally speaking, the \( \phi \) parameter(s) are not generally of much scientific interest. We should also consider the case where there is a fully observed covariate, \( x \), included in the model, in which case Equation 1.4 becomes

\[ p(y, I|x, \theta, \phi) = p(y|x, \theta)p(I|x, y, \phi) \]

because we need to condition on \( x \). One of the benefits of this formulation in the sampling context has been the introduction of multiple imputation techniques (Rubin and Schenker, 1986) to simulate distributions of missing values and provide better inference when such
missing data are occurring. In such a situation, analysis is divided into estimation of 
\( p(\theta, \phi|x, y_{obs}, I) \), and draws from the posterior of this can be used to generate 
\( p(y_{mis}|x, y_{obs}, I, \theta, \phi) \).

Practically speaking, it is often desirable to ignore the missing data process, in which 
case it is valid to evaluate the posterior distribution of \( \theta \) as shown in Equation 1.1. This 
is only a valid strategy if \( I \) provides no information about \( \theta \), and posterior predictions of 
\( y_{mis} \) can be specified by only the data model and \( y_{obs} \). For this to be true, two conditions 
need to be met (Gelman et al, 2003, p.204). First, missing data must be missing at random. 
This means that the probability of missing data depends only on \( x \), \( y_{obs} \), and any observed 
covariates \( y_{mis} \). It cannot depend on the values of \( y_{mis} \). Put symbolically,

\[
p(I|x, y, \phi) = p(I|x, y_{obs}, \phi) . \tag{1.5}
\]

Secondly is the condition of distinct parameters. This means that \( \phi \) must be distinct 
from \( \theta \) so that the process that is generating the missing data is independent of the param-
eters of interest. Put symbolically,

\[
p(\phi|x, \theta) = p(\phi|x) . \tag{1.6}
\]

The consequence of these two conditions is that we can write our posterior as

\[
p(\theta|x, y_{obs} = p(\theta|x, y_{obs}, I).
\]

At first glance, it may appear that the only way to meet the criteria demonstrated in 
Equations 1.5 and 1.6 is through some sort of randomization mechanism when sampling. 
With care though, the idea of ignorability can be elaborated to other situations.

Consider a case in accounting with an observed covariate, \( x \), which represents declared 
income on a tax return (and is fully observed on all tax returns). Now imagine an inclusion 
rule, based on \( \phi \) such that every person whose declared income is greater than $1000000 
will be audited. The outcome of the audits correspond \( y_{obs} \), and our parameters of interest 
\( \theta \) are based on these outcomes. At first glance, it may not appear that data collection in this 
case could possibly be ignorable, but with some assumptions they can be considered so.

First, the missing data are missing at random. Whether a unit is included in the sample
depends on the fully observed covariates, $x$ (declared income). Further, if the outcome being searched for in the audit is thought to be unrelated to income, then the second criteria, that $\phi$ and $\theta$ be independent is also met and the sampling could be considered ignorable. However, many reasons I could think of to audit individuals with a high income implied that $\theta$ and $\phi$ would be related (e.g., high income people have more opportunity for error on their tax returns).

A few important summary points about sampling in the Bayesian context should be mentioned. Conditioning on enough covariates makes many sampling methods become ignorable (exceptions occur when the sampling mechanism does not depend on the response except through covariates). This does not make collecting many covariates practical. Second, when we engage in non-random sampling designs such as the income example above, it is possible to conceive of non-random ignorable designs as long as enough assumptions are met. Unfortunately, if we are unaware of ways in which ignorability assumptions fail, an analysis could be wrong whether we are cognizant of it or not. Lack of awareness does not translate into a correct analysis, so including randomization mechanisms where possible is always advisable.
Chapter 2

Hierarchical Bayesian Model For Compositions

2.1 Background

In this chapter, a Bayesian hierarchy is proposed which uses a multivariate conditional autoregressive (MVCAR) model applied to a compositional response vector i.e. the response has components that sum to one. The data are sampled through time. This model develops ideas from both compositional data analysis and MVCAR models to enhance our ability to analyze and interpret compositional data in a spatial/temporal context. We focus on successfully modeling components when measurements take the form of discrete counts, and observed zero counts occur in component categories. We also allow for components that co-vary positively through time. Some of the challenges posed by this type of data are outlined in Chapter 1.2.

This work is motivated by a problem in managing the socially and economically important wild sockeye salmon fishery in the Fraser River of British Columbia. Understanding salmon migration behaviour is an important part of successfully managing the sockeye fisheries for the future. Our components are genetically identified stock groups from this fishery. Small samples of fish are taken daily, yielding information about the relative magnitudes in the form of proportions of fish from different stocks during the migration season.

Along with the compositional nature of our data, we also wish to account for the corre-
lated relationships between observations at adjacent time points. Besag (1974) introduced a markov random field (MRF) approach to analyze correlated spatial data. In such models, the estimated value of the response at the current site is conditional on the responses at neighbouring sites. Such a Markov random field approach can be constructed in a Bayesian context using pairwise difference priors (Besag et al, 1995, 1991).

Assume that \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \) is a random vector of responses where 1, \ldots, \( n \) index the temporal/spatial locations where the variable of interest is being measured, and \( \sigma \) is a scale parameter. A difference prior can be written

\[
\pi(\varepsilon|\sigma) \propto \exp\left\{ -\sum_{i \sim j} w_{ij} \Phi(\sigma(\varepsilon_i - \varepsilon_j)) \right\}.
\] (2.1)

In Equation (2.1), the \( w_{ij} \) are a set of weights that are non-zero when \( i \) and \( j \) are neighbours, and \( w_{ij} = w_{ji} \). The summation occurs across all \( i \) and \( j \) pairs where \( i \) and \( j \) are neighbours. \( \Phi(u) \) is a function that will be described shortly.

Specifying the MRF as in Equation 2.1 results in the following distributional form for each of the \( \varepsilon_i \)'s conditional on its neighbourhood,

\[
\pi(\varepsilon_i|\varepsilon_{-i}) \propto \sigma \exp\left\{ -\sum_{j \in \partial i} w_{ij} \Phi(\sigma(\varepsilon_i - \varepsilon_j)) \right\}
\] (2.2)

where \( \partial i \) represents the neighbourhood of location \( i \).

Typically, if the weights, \( w_{ij} \), are positive when two locations are neighbours, and if \( \Phi(u) = \frac{1}{2} u^2 \), then the conditional distribution of any \( \varepsilon_i \) is normal (Besag and Kooperberg, 1995; Besag et al, 1991),

\[
\varepsilon_i|\varepsilon_{-i} \sim N\left( \sum_{j \in \partial i} \frac{w_{ij} \varepsilon_j}{w_{ij}}, \frac{1}{\sigma^2 w_{ij}} \right).
\] (2.3)

Here \( w_{ij} = \sum_{j \in \partial i} w_{ij} \), and if the weights are set to one then \( w_{ij} \) sums to the total number of neighbours for site \( j \).

Models that satisfy Equations like (2.1),(2.2),and (2.3) are often said to have an intrinsic autoregressive structure, and are conditional autoregressive (CAR) models.

Such CAR models have become increasingly popular since their introduction for a variety of reasons. With their conditional form CAR models are well suited to Gibbs sampling.
and other MCMC methods in a Bayesian context (Besag et al, 1995; Besag and Kooperberg, 1995; Besag et al, 1991). Such approaches are also generally described in 1.3.1. This makes them computationally easier to estimate than related frequentist spatial estimation methods. In addition to being suited to standard MCMC methods, the missing values present in the data can be included as parameters in these Bayesian models. This provides a straightforward solution to what is often a tricky problem.

CAR models share information between locations due to the conditional nature of their construction. Even when the sample size for a given location is small, the information ‘borrowed’ from nearby locations results in reduced variability of most estimates. As such, spatial CAR models are used in small area estimations (Ghosh et al, 1998; Ghosh and Rao, 1994; Moura and Migon, 2002; Rao, 2003, p.168-169).

CAR models have also been extended to cases where \( \varepsilon_i \) is a vector of \( d \) response variables (Mardia, 1988; Gelfand and Vounatsou, 2003; Kim et al, 2001). Models are formulated analogous to those shown in equations (2.1), (2.2), and (2.3) where \( \varepsilon_i \) becomes a vector, and \( \sigma \) is replaced with \( \Sigma^{-1} \), a \( d \times d \) inverse variance covariance matrix. The conditionals then become multivariate normal

\[
\varepsilon_i | \varepsilon_{-i} \sim \text{MVN} \left( \sum_{j \in \partial_i} \frac{w_{ij} \varepsilon_j}{w_{i\bullet}} , \frac{\Sigma^{-1}_{\bullet\bullet}}{w_{i\bullet}} \right),
\]

and the model is called a multivariate conditional autoregression (MVCAR) model. Our fisheries data are multivariate because we have observations on multiple stocks. In addition to being multivariate, our data are discrete compositions, imposing additional restrictions on the estimation. There are a few examples of MVCAR models that focus on estimation of such proportions (Besag et al, 1995; Gelfand and Vounatsou, 2003; Vounatsou et al, 2000; Billheimer et al, 1997, 2001). In these, the proportions are estimated using a transformation to the real numbers such as the ALR transformation. Our hierarchy represents an alternative Bayesian approach. We refer to our method as a compositional hierarchical MVCAR model, or a CH MVCAR.
2.2 Model Development

2.2.1 Sampling Protocols

Sampling of sockeye salmon occurs over $T$ days. Multinomial counts are recorded on each day in $D$ component categories where different genetic stock groups make up the individual components. Daily sample sizes are relatively small, introducing variability that needs to be accounted for in modeling. Some days are missed during sampling, even though estimation of the underlying proportions is still desirable. Stocks with zero observed fish commonly occur in component categories, and the proportions underlying the multinomial distribution in adjacent days are thought to be correlated.

2.2.2 A Hierarchical Model For Estimating Stock Proportions

Data Layer

Let the daily fish count in each stock be represented by the vector

$$\mathbf{y}_t = (y_{1t}, \ldots, y_{Dt})$$

where $t$ is the timepoint of interest, and $i = 1, \ldots, D$ represent different stocks. Let

$$n_t = \sum_{i=1}^{D} y_{it} = \mathbf{y}_t \mathbf{1}$$

be the number of fish sampled on day $t$, where $\mathbf{1}$ is a column vector of ones. The underlying proportions can then be written

$$\mathbf{\Pi}_t = (p_{1t}, \ldots, p_{Dt}) \text{ where } \sum_{i=1}^{D} p_{it} = \mathbf{\Pi}_t \mathbf{1} = 1.$$  

The counts in $\mathbf{y}_t$ can then be considered realizations of a multinomial process conditional on the parameters in $\mathbf{\Pi}_t$,

$$P(\mathbf{y}_t | \mathbf{\Pi}_t) = \frac{n_t!}{\prod y_{it}!} \prod_{i=1}^{D} p_{it}^{y_{it}}. \quad (2.5)$$
Middle Layer

Although our data are recorded as observed counts, we are primarily interested in modeling the underlying proportions. We assume a Dirichlet prior on these $\Pi_t$:

$$P(\Pi_t | \alpha_t) = \frac{\Gamma\left(\sum_{i=1}^D \alpha_{it}\right)}{\prod_{i=1}^D \Gamma(\alpha_{it})} \prod_{i=1}^D p_{it}^{\alpha_{it}-1}. \quad (2.6)$$

Ultimately, this prior will be linked to a temporal process layer. Here $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{Dt})$ are the parameters of the Dirichlet prior distribution, with $\alpha_{it}, \ldots, \alpha_{Dt} > 0$. With respect to the posterior, the expected value for a proportion is

$$E(p_{it}) = \frac{\alpha_{it} + y_{it}}{\alpha_t \mathbf{1} + n_t} = q_{nt} \left( \frac{\alpha_{it}}{\alpha_t \mathbf{1}} \right) + (1 - q_{nt}) \left( \frac{y_{it}}{n_t} \right). \quad (2.7)$$

As shown in (2.7), our estimates can be described as a weighted estimate of the prior and likelihood estimates for $p_{it}$, where $q_{nt} = \frac{\alpha_t \mathbf{1}}{\alpha_t \mathbf{1} + n_t}$ defines the weights. If $\alpha_t \mathbf{1}$ is large in magnitude relative to $n_t$, most of the weight for the proportion estimates comes from the temporal process describing $\alpha_t \mathbf{1}$. This MVCAR process layer will be described in Section 2.2.2.

This formulation permits estimation of the proportions, $p_{it}$, by the posterior mean at any given time point, $t$. Because the Dirichlet distribution in (2.6) is conjugate to the multinomial distribution in (2.5), the estimated $p_{it}$s are constrained to sum to one at any given time point. Moreover, the conjugacy of the distributions ensures that the posterior will be proper as long as the elements of all the $\alpha_t = (\alpha_{1t}, \ldots, \alpha_{Dt})$ vectors remain positive (Ord et al, 1993).

Further, it is possible to partition the values in $\alpha_t$ into two parts: $G_t$, where $\sum_i^D G_{it} = 1$; and $\gamma$, is a scaling constant. Consequently,

$$\alpha_t = \gamma G_t. \quad (2.8)$$

$G_t$ represents information on how the relationship between stock proportion estimates changes according to the process layer. If we increase the value of $\gamma$, our value for $\alpha_t \mathbf{1}$ correspondingly increases. A large value for $\gamma$ corresponds to placing high prior information on our belief that a temporal process links observations on adjacent days.
It is difficult computationally to sample from a diffuse gamma distribution, and a Dirichlet is a function of gamma distributions. Component categories with low counts can end up with small $G_{it}$ values which alone would result in a distribution too diffuse to sample from. The partitioning of $\alpha_t$ values as seen in Equation 2.8 can be used to mitigate this computational problem. We can set $\gamma$ as a value large enough to make the $\alpha_{it}$ value less diffuse.

**Process Layer**

In this layer of our hierarchy, we characterize a temporal process governing the observed stock proportions. The parameters, $\alpha_t$ from our middle layer are modeled using an MV-CAR structure through time.

The specific construction of the MVCAR model is as follows:

$$
\log(\alpha_t) = \beta + \varepsilon_t .
$$

In Equation (2.9), $\alpha_t$ is the same as defined in Equation (2.6), $\beta$ is a vector of stock specific intercept terms, $\beta = (\beta_1, \ldots, \beta_D)$, and $\varepsilon_t$ is a vector of time correlated multivariate effect terms, $\varepsilon_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Dt})$. The log-transformation ensures that all elements in $\alpha_t$ remain positive.

The prior for the $\varepsilon_t$ is

$$
\varepsilon_t | \varepsilon_{-t} \sim \left\{ \begin{array}{ll}
MVN(\varepsilon_{t+1}, \Sigma^{-1}) & , t = 1 \\
MVN\left(\frac{\varepsilon_{t-1}+\varepsilon_{t+1}}{2}, \Sigma^{-1}\right) & , t = 2, \ldots, T-1 \\
MVN(\varepsilon_{t-1}, \Sigma^{-1}) & , t = T.
\end{array} \right.
$$

$\Sigma^{-1}$ is the $D \times D$ inverse variance covariance matrix relating the $\varepsilon_{it}$ terms from the different stocks.

**Priors to the Process**

In the final layer of our modeling hierarchy, we assign priors to the parameters $\beta$ and $\Sigma^{-1}$. These priors are selected so that the model is specified as a multivariate intrinsic CAR model (Besag et al, 1991). Elements of the intercept vector $\beta$ are given improper
Uniform\((-\infty, \infty)\) priors. The prior for the elements of \(\Sigma^{-1}\), the inverse variance, are given a Wishart\((V, f)\) prior. Here, \(f\) is the number of degrees of freedom, and \(V\) is a positive definite matrix placing prior information on the inverse variance matrix.

2.2.3 Fitting the Model

This CH MVCAR model can be fit using the WinBUGS software package (Lunn et al, 2000). The code used to fit the model (running WinBUGS from R (Gentleman et al, 1997)) can be found in Appendix A.

2.2.4 Goodness-Of-Fit Assessment

We assess the fit of our model using posterior predictive checks in the form of Bayesian p-values (Gelman et al, 1996, 2003, pp.159-165). Both general and focused measures are used to evaluate differences between the observed data and predicted outcomes based on the model. Bayesian p-values are obtained by comparing these measures based on the observed data, \(x\), and the ‘observed’ data generated based on draws from the posterior, \(x_i\).

Our general test uses a discrepancy measure that is identical to that found in Brooks et al (2000). The discrepancy, \(D(x; \theta)\), is defined as \(D(x; \theta) = \sum_j (\sqrt{x_j} - \sqrt{e_j})^2\). Here, \(x_j\) are the observed counts, and \(e_j\) are the expected counts based on the model taken across each time point and stock. This discrepancy is chosen instead of the more common Pearson’s \(X^2\) statistic due to the stabilizing influence of the square root in the presence of sparse data.

Our focused test is simply defined as the observed data \(x\) compared to the posterior predictive data \(x_i\). Rather than a single overall p-value, such a focused test gives a unique Bayesian p-value for each time point and component stock where observations were taken.

For both the general and focused tests, good model fit is indicated by p-values close to \(\frac{1}{2}\), with model fit becoming more suspect as the p-value moves towards zero or one.
2.3 Example

2.3.1 Motivating Data

This research was motivated by data collected by the Pacific Salmon Commission to help manage the commercial sockeye fishery in the lower Fraser River. Migrating salmon are sampled once daily using nets between July 22 and September 13, 2006. At this time of year, sockeye salmon are migrating from ocean to freshwater in order to spawn. As with other salmon, sockeye have strong homing tendencies to spawn in the rivers/lakes of their birth. Each spawning location is subject to unique temperature, productivity, competition, and other environmental conditions. As a result, sockeye that originate from proximate spawning grounds tend to be more similar than sockeye from more distant spawning grounds. Fish migrating up the Fraser River to spawn can be divided into distinct stock groups based on the physical, behavioural, and genetic characteristics of the salmon reflecting these environmental differences. It is for this reason that the commercial fisheries in the lower Fraser River are referred to as mixed stock fisheries. In the 2006 test fisheries samples, netted fish were divided into five different stock groups. Different groups follow different patterns in the timing of their spawning runs. For example, stock groups that have a similar distance to travel between the river mouth and spawning grounds will exhibit similar run timing patterns, with more disparate stocks showing fewer similarities. We order the 5 stocks from 1 to 5, generally reflecting early run-timing to late run-timing stocks. Managing sockeye in the Fraser River involves maintaining healthy populations in both the river as a whole, and in each of the tributary stocks that contributes to the overall population.

A sample of the data used to model stock proportions in the Fraser River are shown in Table 2.1. In this table, day refers chronologically to the day during sampling that the observation was taken. The number of fish in the sample is denoted $n_t$, and there are a total of $T = 49$ days over which sampling occurs. The remaining columns represent observed stock proportions.

These data are compositional. A finite sample of fish is taken each day, and sample sizes are not reflective of stock abundances. The sample sizes used to determine the observed stock proportions are relatively small, ranging from 16 to 96 fish; it is difficult to achieve uniform daily sample sizes using net collection. There are nine days when test
Table 2.1: Sample of observed data on stock proportions. Days when sampling does not occur are indicated by rows of NAs. Day refers chronologically to the day during sampling that the observation was taken. The number of fish in the sample is denoted \( n_t \), and there are a total of 49 days over which sampling occurs. The remaining columns represent observed stock proportions.

<table>
<thead>
<tr>
<th>Day</th>
<th>( n_t )</th>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Stock 3</th>
<th>Stock 4</th>
<th>Stock 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>37</td>
<td>0.00</td>
<td>0.38</td>
<td>0.35</td>
<td>0.00</td>
<td>0.27</td>
</tr>
<tr>
<td>32</td>
<td>48</td>
<td>0.00</td>
<td>0.27</td>
<td>0.40</td>
<td>0.00</td>
<td>0.33</td>
</tr>
<tr>
<td>33</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

sampling does not occur, constituting missing values. We are interested in estimating stock proportions on these missing days. Such missing values occur on consecutive days only once. There are also numerous days where the observed number of fish from a stock in the sample is zero, as seen in Stock 2 and Stock 4 of Table 2.1. Such observed zero components can occur because a stock is at low abundance, and catching fish is a rare event. Alternatively, zeros in observed components can occur at the beginning or end of a stock’s migration run, before or after the main body of fish from the stock has passed.

2.3.2 Parameter Estimation

Point Estimates

Estimated daily stock proportions are visualized for each stock separately and as a group in Figure 2.1.

A key goal of the model was to induce temporal smoothing in stock proportion estimates. Such smoothing is seen in the Bayesian estimates relative to the observed proportions. The current smooth achieves the goals of fisheries managers to reduce the variability in proportion estimates on adjacent days. The current model also shows how a hierarchy similar to the one demonstrated can be used to address common problems associated with compositional data analysis. In applications where other attributes for the smoothed estimates are desirable, different smoothing properties can be obtained as discussed in Section
Figure 2.1: Graphs of smoothed and frequentist proportions. Grey dots represent observed proportions based on fish counts. The black lines represent estimates based on the mean of the posterior. Hollow black dots represent days where no sampling was taken, and the point estimated has been taken from the MVCAR model. Graphs are ordered in terms of early to late run stocks, with the final graph showing information from all stocks at once.
2.4.

The model also generates sensible estimates on days when sampling did not occur. When observed zeros in components occur in one or more category, our Bayesian estimation generates plausible proportion estimates. In Figure 2.1 daily proportions are estimated for stock 1, despite sampling frequency in this stock being very low. In stock 4, the long left tail of low stock proportions prior to the main body of the stock migrating has also been estimated in a plausible manner.

Interval Estimation

Figure 2.2 shows the relative efficiency of using our highest posterior density (HPD) intervals based on our Bayesian hierarchical model, compared to frequentist confidence intervals based on a multinomial assumption. Frequentist intervals based on multinomial observations were estimated using the exact interval method (Clopper and Pearson, 1934).

The HPD intervals are generally narrower than the frequentist intervals indicating an overall improvement in efficiency. This is confirmed by the histogram of the relative interval widths. Such reductions are achieved because the MVCAR prior allows information to be shared between adjacent days, increasing effective sample sizes.

Despite an overall increase in efficiency, it is interesting to observe that the magnitude of variance reduction caused by our hierarchical model does not appear to be the same in all stocks. For example, relative to frequentist intervals, stock 3 HPD intervals seem to indicate a greater increase in efficiency than stock 5 HPD intervals. This can be explained by the fact that the Wishart prior on the inverse variance covariance matrix has only one degree of freedom parameter, $f$. This can constrain variance estimates (Gelman and Hill, 2007, pp. 284-287). As can be seen in Figure 2.1, different stocks appear to have different levels of variability so such constraint may not be realistic.

Correlation Estimates

The ‘All Stocks’ graph in Figure 2.1 shows estimation from all stocks plotted together. These graphs demonstrate relationships between stocks that we would hope to see recovered in correlation estimates for the model.

Correlation estimates between the $\epsilon_t$s are displayed in Table 2.2, along with estimates
Figure 2.2: Comparison of interval estimates based on CH MVCAR highest posterior densities (HPDs) and frequentist intervals based on a multinomial assumption. Black lines with dots represent the point-wise frequentist 95% confidence intervals based on a multinomial assumption. Gaps between bars indicate days where no sampling occurred. The grey shadow represents the corresponding to 95% highest posterior density estimates. Relative interval widths are calculated as (HPD interval width)/(frequentist interval width).
CHAPTER 2. HIERARCHICAL BAYESIAN MODEL FOR COMPOSITIONS

of the standard deviations in the correlation estimates. Generally, the estimates for correlations between stock parameters seem realistic based on relationships that can be observed in the raw data. For example, the correlation between stocks 2 and 3 is reported as 0.60, which is credible given the positive relationship between observations that could be seen in Figure 2.1. Similarly, a correlation of -0.57 was observed between stock 2 and stock 5. We note that these two stocks had an observed negative correlation between proportions.

Our correlation estimates are calculated from the variance covariance of the multivariate normal distribution at the process level. Coming from a normal model, these are much less constrained than covariances calculated at the level of the multinomial observation model.

Table 2.2: Posterior estimates for standard deviations and correlations of $\epsilon_t$ terms; standard deviations for the point estimates given in brackets

<table>
<thead>
<tr>
<th></th>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Stock 3</th>
<th>Stock 4</th>
<th>Stock 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stock 1</td>
<td>0.56(0.27)</td>
<td>0.30(0.52)</td>
<td>0.12(0.53)</td>
<td>-0.45(0.42)</td>
<td>-0.33(0.54)</td>
</tr>
<tr>
<td>Stock 2</td>
<td>-</td>
<td>0.35(0.15)</td>
<td>0.60(0.34)</td>
<td>-0.31(0.42)</td>
<td>-0.57(0.34)</td>
</tr>
<tr>
<td>Stock 3</td>
<td>-</td>
<td>-</td>
<td>0.27(0.12)</td>
<td>-0.08(0.46)</td>
<td>-0.23(0.48)</td>
</tr>
<tr>
<td>Stock 4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.52(0.17)</td>
<td>0.44(0.36)</td>
</tr>
<tr>
<td>Stock 5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.63(0.18)</td>
</tr>
</tbody>
</table>

2.3.3 Sensitivity Results

In Section 2.2.2 we introduced $\mathbf{V}$ and $f$, the parameters for the Wishart prior on the inverse variance-covariance matrix $\Sigma^{-1}$. We set $f$ to $D + 1$ implying weak prior information. We conducted a sensitivity analysis on $\mathbf{V}$ by adjusting the values on the diagonal of the matrix. Off diagonal elements were set to zero. We tested diagonal values that ranged from 1 to 0.0001. Between values of 0.5 and 0.01 the model gave consistent results in terms of estimated proportions and correlation estimates. Our presented results are based on selecting diagonal values of 0.1, in the middle of this robust results range. In practice, different data will result in a different range of sensible values for this prior. Anyone using this method should conduct a thorough sensitivity analysis (ie. test across a range of values) as is also suggested in Thomas et al (2004).

When diagonal values were between 0.5 and 1, WinBUGS had computational issues
when sampling. When diagonal values were small, such as 0.0001, other problems occurred. Correlation estimates often became ‘stuck’ at unrealistically high or low values near -1 or 1. The absolute magnitudes of the correlation are being increased by such extreme modification of the Wishart prior. Constructing a more informative Wishart prior also had the side-effect of influencing the overall convergence rates. Generally, highly correlated slow to converge chains occurred.

2.3.4 Goodness-Of-Fit Results

The fit of the model was assessed using Bayesian p-values based on posterior predictive data as discussed in section 2.2.4. Our general test gave a Bayesian p-value of 0.1. This p-value is suggestive of some minor lack of fit. The value is better understood after considering p-values generated from the focused tests of individual data points.

We have 5 stocks and 40 days where fish were actually sampled, corresponding to 200 Bayesian p-values when including all stock and time points. For all but one sampling point, the focused Bayesian p-values indicated reasonable model fit. The one exception to this occurred on day 48 which had a Bayesian p-value of 0.036. This appears to be due to a fish being identified as coming from stock 1 when there were numerous observed zeros in that component stock on adjacent days. It should be noted that of the 200 p-values, there are slightly more p-values below 0.5 than above 0.5. The general test represents a summing across these 200 values, likely resulting in its value being 0.1, rather than nearer to 0.5. The full list of Bayesian p-values can be found in Appendix B.

We note a limitation that occurs in calculating the focused Bayesian p-values when there is a zero count in a component stock in the observation model. If our observed count is zero, and all of our predicted observations for this stock are also zero, then we will have a p-value of 1. Normally, such a p-value would indicate that our observed data falls at an extreme of the replicated data. Here that is clearly not the case. The predicted distribution is ‘predicting’ the observed value of zero, indicating good model fit.

2.3.5 Convergence

Evaluating the convergence of posterior model parameters is an important element of Bayesian data analysis using MCMC methods. We use traceplots to help us review and demonstrate
some typical convergence results for the current model in Figure 2.3. Each traceplot shows results from three independent chains. We demonstrate traceplots that show a plot for the deviance of the model; for a proportion estimate at a timepoint where data were collected; for a proportion estimate at a timepoint where data was not collected; and for a correlation parameter where little data was available for estimation.

The first graph in Figure 2.3 shows a traceplot for the overall deviance of the model. Next to this is a traceplot for an estimated stock proportion on a day when net sampling occurred. Both of these traceplots indicate chains that are stationary and mixing well. Convergence can be inferred to have occurred for these chains. The third graph in Figure 2.3 shows a traceplot for a proportion that was estimated on a day when no fish sampling occurred. It is a missing value that is calculated by our model. The overlap of these chains also indicates stationarity and convergence. Traceplots for other stock proportions show similarly convincing evidence of convergence.

The traceplot of the correlations between $\varepsilon_1$ and $\varepsilon_5$ indicates sampling occurring over nearly the entire $-1$ to $1$ range. With few observed fish in stock 1, it is not surprising that the chain for this correlation estimate essentially shows sampling from the uniform $-1$ to $1$ prior that was placed on all correlations.

Convergence can also be assessed using the Gelman Rubin convergence diagnostic (Gelman and Rubin, 1992). When this diagnostic has values near to 1, there is little potential scale reduction associated with continuing MCMC simulations. Values below 1.1 are considered acceptable for most problems (Gelman et al, 2003, pp. 296-298). In this estimation, the Gelman-Rubin convergence diagnostic was computed for all posterior estimates of interest. All were less than 1.1 with the exception of those discussed below.

With compositions, only four proportions need to be specified in order to determine the fifth. At the level of the prior we estimate five stock specific $\alpha_i$ values, an overparameterization. These parameters are unidentifiable and exhibit drift in their traceplots. A similar tendency for drifting was noticed for unidentifiable parameters in a model demonstrated in Besag et al (1995). Similar to them, we recommend that posterior probability statements should be confined to parameters derived from these values such as stock proportions, correlation estimates, etc.
Figure 2.3: Traceplots showing generally good convergence results based on observing three independent MCMC chains. The first graph shows the traces for the overall model deviance. The second plot represents traces from MCMC draws on a proportion estimate on a day when fish were sampled. The third plot represents MCMC chains for proportions estimates on a day when fish were not sampled. The final plot represents chains for correlation estimates between stocks 1 and 5; very few observations were taken from stock 1.
2.4 Discussion

In this paper we developed a novel hierarchical model for successfully estimating temporally or spatially correlated compositional data with missing values (NAs) and observed zeros in component categories. Our model provides an important contribution to compositional data analysis by being able to perform estimation when zeros are observed in any of our $D$ component categories. Another contribution to compositional data analysis was being able to estimate a $D$ by $D$ covariance matrix that was not subject to the limitations induced by multinomial or Dirichlet based models. Although ALR transformed models can utilize the flexibility of normal assumptions, their covariance matrices are $D-1$ by $D-1$.

Zeros observed in components are sampling zeros in our model. This means that we observe a count, $y_{it} = 0$, even though we believe that a non-zero count would be possible if the daily sample size $n_t$ could be increased. Such sampling zeros can be computationally inconvenient in many categorical data analysis problems because they affect the existence of finite likelihoods, and estimable model parameters (Agresti, 2002, p.392). Sampling zeros are also a problem when data are being modeled using the ALR transform because the quantity $\log\left(\frac{y_{it}}{y_1}ight)$ cannot be calculated.

Methods of addressing sampling zeros include collapsing across component categories so that the number of observations per category is non-zero, or to add a small positive constant, e.g. $\frac{1}{2}$, to all counts. In a Bayesian context this is similar to pairing a multinomial observation model with a Dirichlet prior where the parameters for the prior can be described as ‘prior sample sizes’ (Gelman et al, 2003, p.576). Such ideas are so pervasive that similar wording is used when $+4$ confidence intervals for binomial proportions are described in introductory statistics texts (Baldi et al, 2009, p.491). Our model is similar to this prior sample approach, but removes the arbitrariness associated with selecting prior sample sizes based on convenience. Our prior information is assigned based on our MVCAR hyperprior which shares information across all sampling points. Days with no observations will receive information about prior sample sizes from days that do have observations.

Other data concerns were addressed by the model, specifically the variance reduction and smoothing caused by sharing information between adjacent days. By sharing information between adjacent observations, estimation of our 9 missing observations is straightforward. In addition to this, the MVCAR/process level of the hierarchy also smooths propor-
tion estimates through time, reducing the variability of the proportion estimates. Because our MVCAR model occurs at the top level of our hierarchy, the smoothness of estimated proportions is affected by parameters both at the process and other layers in the hierarchy.

At the process level, the degree of smoothing can be influenced by the assignment of the Wishart \((V, f)\) hyperprior on the inverse variance, and also the way the neighbourhood in the MVCAR model is defined. Making adjustments to the Wishart hyperprior has undesirable side-effects of changing prior information on correlations, thus affecting convergence of the model. Consequently, this hyperprior is rarely discussed as a tool for adjusting smoothness of estimates.

More practically, the smooth can be adjusted by changing \(u\) in the function \(\Phi(u)\) in Equations (2.1) and (2.2). Our prior represents a first order random walk. If a prior was chosen to represent a second order random walk, it would reflect a belief that the rate of change in the \(\epsilon\)'s was smooth (i.e. first order differentiable) (Thomas et al, 2004, pp.36-38). A variety of different functions for \(u\), and reasons for their development are given in Besag et al (1995). We recommend adjusting \(u\) as the most promising in terms of generating smooths with properties appropriate to other applications.

Apart from the process layer, the variable \(\gamma\) in the middle layer can also affect the smoothness of our run timing curves. All else held equal, larger values of \(\gamma\) correspond to greater belief in the smoothed MVCAR process relative to the raw observations. Smaller values of \(\gamma\) reflect less belief in this smooth, and put more weight on the data. We chose a value for \(\gamma\) of 10000. This value was selected by manipulating Equation 2.7 such that \(q_{n_t} = 0.99\) when \(n_t = 100\) (approximately our largest sample size). This ensures our parameter estimates are mainly determined by the process layer of our model, consistent with a belief that modeling temporal correlation is important. It is also large enough to avoid computational issues with the estimation.

The data in the motivating example were both compositional, and ‘messy’ due to sampling variability, missing days, and observed zero components. Due to the potential presence of such ‘messy’ compositional data in a wide variety of fields of application, the potential benefits of utilizing this model, or ones based on similar hierarchies are far-reaching.

We believe this paper is the first to attempt to simultaneously address the topics discussed above.

In addition to successfully creating and applying this CH MVCAR model, we identi-
fied areas where improvements and future work are possible. It would be interesting to
determine what effect using variance priors other than the Wishart at the top level of the
hierarchy would have. Although our choice of Wishart hyperprior is consistent with oth-
ers published in the MVCAR literature (Shaddick and Wakefield, 2002; Billheimer et al,
1997), it is suggested that a scaled Wishart as discussed in Gelman and Hill (2007) can be
used to estimate the correlations and standard deviations more realistically. Currently, such
a parameterization is most commonly applied in Bayesian regression applications, but due
to apparent differences in the underlying variability of observations from different stocks,
experimenting with such a scaled Wishart prior in our model could be valuable.

Other important extensions could incorporating covariate information into model pre-
dictions for stock proportions. Additional environmental information such as tidal patterns,
and river flow would likely further inform estimation in our example.

Finally, we are interested in how data collection affects our implementation and under-
standing of the model. When using posterior predictive checks in the form of Bayesian
p-values to evaluate the model, we are hypothesizing that it is possible to replicate the ob-
served data. In part, the observed data are dependent on study design. It is interesting to
contemplate how model estimates and inference might change or improved with different
potential data generated through alternative sampling mechanisms. Improving the reason-
ing used to determine data collection methods for compositional problems, along with the
improved estimation that is offered through our hierarchical model, could greatly improve
the quality of inference that is made using discrete compositional data.
Chapter 3

Exact Approach To Multinomial Sample Size

3.1 Background

For this chapter, we extend the results for an exact highest density probability (HD) interval described for binomial distributions in Rahme and Joseph (1998). Our goal is to select a sample size, $n$, that is smallest for the given interval half-width $d$, and level of significance $\alpha$. We use Rahme and Joseph (1998) approach to identify sample sizes (including worst case) for a variety of combinations of interval widths, levels of significance ($\alpha$), and binomial/multinomial parameters $\theta$. We also extend the definitions for the HD interval to the multinomial case, and use it to evaluate sample sizes under a variety of interval widths, levels of significance, and parameter vectors, $\Pi = (\theta_1, ..., \theta_k)$.

In the motivating example outlined in Chapter 1.1, the data collected on sockeye stock composition constitute a multinomial sample. The multinomial distribution reduces to a binomial distribution when only two groups are present. The binomial distribution is a well known and commonly used statistical distribution. Its basic use is included in most first year statistics textbooks (Baldi et al, 2009; Peck and Devore, 2011). In higher level mathematical statistics courses it is often used to demonstrate characteristics and calculations of discrete distributions (Hogg et al, 2005, pp.133-142). These introductions give practitioners familiarity on how to calculate sample sizes, to form confidence intervals, and to make
other inference using the binomial distribution. Despite an apparent familiarity with the binomial distribution, implications on how the distribution should be used in study design are less clear due to some often unremarked characteristics. There can be confusion about sample size and confidence interval calculations for the binomial and multinomial distributions. Specifically, the discreteness of the distribution can cause ‘jumps’ in the coverage probabilities of confidence intervals, especially when sample sizes are small. The implication is that the common use of normal approximation methods to calculate required sample sizes for multinomial problems is often not appropriate.

**Binomial Review**

Typically, when considering the binomial case, the normal approximation is applied in order to either calculate a required sample size, or to construct a confidence interval. When calculating a sample size using this approximation, it is often suggested to choose a conservative value of $\theta = 0.5$. The variance of the binomial is maximized at this value leading to a conservative sample size calculation. When constructing confidence intervals for binomial parameters, a rule of thumb is that $n\hat{\theta} \geq 10$, or $n(1 - \hat{\theta}) \geq 10$ should be achieved for the normal approximation to be appropriate.

Such intervals based on the normal approximation rarely achieve their stated level of coverage when calculated in terms of either exact confidence coefficients (Wang, 2007) or in terms of average coverage performance (Agresti and Coull, 1998). This problem is worst when $n$ is small, when $\theta$ approaches 0 or 1, or both.

The problems with the normal approximation are twofold. First, the normal distribution is symmetric, and the binomial distribution is only symmetric when $\theta = 0.5$. Second, the normal is a continuous approximation to a discrete distribution.

Using a symmetric confidence interval to represent the uncertainty in $\hat{\theta}$ when $\hat{\theta}$ is far from 0.5 is counterintuitive, because it means that that an (likely) asymmetric area of highest probability is being summarized by something symmetric. Unless $\hat{\theta} = 0.5$, this means that a normal interval cannot possibly be representing the underlying binomial well. Even with the rules of thumb stated above, or corrections that are sometimes introduced when the number of observed events is small, the normal approach is still problematic in terms of confidence coverage.
Even when \( \hat{\theta} \) or \( \theta \) are near to 0.5 when the binomial distribution is (close to) symmetric, there can still be problems with the normal approximation. By the normal approximation, the worst case sample size requirement always occurs at \( \theta = 0.5 \); using exact quantiles of the binomial though, the worst case result can surprisingly occur at other values for \( \theta \). An example from Rahme and Joseph (1998) illustrates that when the desired interval width is 0.2, \( \alpha = 0.4 \), and \( \theta = 0.5 \) then the necessary sample size is \( n = 5 \). If we keep the above, but change to \( \theta = 0.4 \), the required sample size becomes \( n = 12 \). Such quirks of binomial distribution behaviour cannot be replicated by a normal approximation. Problems such as this that occur due to discreteness are more pronounced at smaller \( n \). A question then arises about when \( n \) becomes large enough that this problem can be ignored.

There are relatively few alternatives to the normal approximation for binomial proportions. The best known of these alternatives for confidence intervals is the Clopper-Pearson interval (Clopper and Pearson, 1934). This is an exact interval (meaning it is based on quantiles of a binomial distribution) calculated by inverting equal-tailed binomial tests. While by definition it always meets its nominal confidence level, this interval has been criticized for being too conservative because actual coverage can be higher than desired coverage (Agresti and Coull, 1998). In their own article, Agresti and Coull (1998) suggest an alternative approximate interval to the symmetric normal one. This interval is a score confidence interval which still relies on a normal approximation, but allows the interval to be asymmetric. This asymmetry improves the coverage levels for the interval (Agresti and Coull, 1998; Wang, 2007). Another alternative is an exact interval discussed in Rahme and Joseph (1998). This interval also ensures that nominal confidence levels are always met. The interval is defined over the area of highest density probability (HD) for a binomial distribution. This means that it can be both asymmetric and have unequal tail probabilities. In addition to guaranteeing the correct coverage, the width of this interval will be less than or equal to that of a corresponding Clopper-Pearson interval. Required sample size calculations for the HD interval are also smaller than those required by the normal approximation (Rahme and Joseph, 1998).
Multinomial Review

Because the multinomial and binomial distributions are so similar, many of the problems outlined above for estimating binomial sample sizes and confidence intervals extend to the multinomial case as well. As with the binomial, most methods for estimating multinomial confidence intervals rely on an approximation to a continuous distribution, often a $\chi^2$ or normal distribution. Descriptions of such approximation approaches are well documented and commonly used (Goodman, 1965; Fitzpatrick and Scott, 1987; Quesenberry and Hurst, 1964; Thompson, 1987). Goodman (1965) uses the Bonferroni inequality to control the error rate among all of the intervals being reported. Thompson (1987) identifies the form of the ‘worst case’ parameter vectors, or those requiring the largest sample sizes when interval widths are equal for all component parameters. Fitzpatrick and Scott (1987) independently discovered the same results. Based on an equal widths criteria for all intervals, $\theta = 0.5$ does not necessarily represent a worst case parameter value. Further, they discover that worst case values for $\Pi = (\theta_1, \ldots, \theta_k)$ depend on the level of significance ($\alpha$) and on the interval widths specified. If $k$ is the number of component categories, they identify a worst case parameter vector as being where $\theta_i = 1/m$ for $m$ categories, and 0 for the remaining $k - m$ categories.

Although the above approaches are commonly used for multinomial problems, they are often not appropriate due to their reliance on normal distribution assumptions. As with the binomial, it has been shown that intervals based on these approximations often do not meet their nominal confidence levels (Wang, 2008). Our goal is to explore where worst case parameter values are occurring for the binomial distribution when using HD interval criteria. We also adapt the definitions for simultaneous multinomial approaches to find sample size results for multinomial HD intervals. These results can be used to improve planning and analysis of studies collecting binomial or multinomial data.

3.2 Methods And Notation

3.2.1 Binomial Definitions

Because it is most common to conceptualize sample size criteria in terms of symmetric intervals (Lemeshow et al, 1990), we begin developing our notation based on these standard
forms. In terms of calculating sample size, the objective is to select a sample size, $n$, that is smallest for a given half-interval width $d$, and level of significance $\alpha$. This can be written as

$$P(|\hat{\theta} - \theta| \leq d) \geq 1 - \alpha . \quad (3.1)$$

Here $\hat{\theta} = x/n$ is the maximum likelihood estimate for \theta, and \theta is the population proportion. The probability, $\alpha$, that the estimate $\hat{\theta}$ falls outside Equation 3.1 for a given \theta can be re-written in terms of exact binomial quantiles

$$P(|\hat{\theta} - \theta| \leq d) = \sum_{i=(x-c)}^{(x+c)} p(i; \theta) = \sum_{i=(x-c)}^{(x+c)} \binom{n}{i} \theta^i (1 - \theta)^{n-i} \geq 1 - \alpha . \quad (3.2)$$

Here $n$ is the sample size, and $c = \lfloor nd \rfloor$ is chosen such that the interval is symmetric and at the correct probability level. The normal approximation is often used to create this type of interval. As \theta moves away from 0.5 the binomial distribution is asymmetric, and a symmetric interval will overestimate the level of uncertainty about \theta.

### 3.2.2 Binomial Exact Criteria

Rahme and Joseph (1998) offer a modified criterion where the symmetric interval defined in Equations 3.1 and 3.2 is replaced by an asymmetric interval definition

$$P(-a \leq \hat{\theta} - \theta \leq b) \geq 1 - \alpha . \quad (3.3)$$

In Equation 3.3 $|a + b| \leq 2d$, and the values $a$ and $b$ correspond to potential observations (integers) $na$ and $nb$ allowing us to write the probability content of the HD interval in terms of a binomial distribution

$$\sum_{i=(x-nb)}^{(x+na)} p(i; \theta) = \sum_{i=(x-nb)}^{(x+na)} \binom{n}{i} \theta^i (1 - \theta)^{n-i} \geq 1 - \alpha . \quad (3.4)$$

As before, $n$ is the sample size, and $na$ and $nb$ are chosen to give the HD interval with the correct probability content. In terms of identifying necessary sample sizes, we will be solving Equation 3.4 for $n$ given $d$ the interval half-width, $\alpha$, the level of confidence, and
CHAPTER 3. EXACT APPROACH TO MULTINOMIAL SAMPLE SIZE

\[ \theta, \text{ potential values for the parameter of interest. Note that we continue using } d \text{ to define the interval half-width, even though the interval itself will now be asymmetric in most cases. This keeps notation consistent with previous methods, and simplifies comparison to other methods. We will be using the interval definition in Equation 3.4 to identify worst case sample size results. This is the first time such worst case results have been calculated using either an asymmetric or exact interval criteria.} \]

3.2.3 Multinomial Definitions

The multinomial equivalent of traditional symmetric confidence intervals is shown in Equation 3.1.

\[ P\left( \bigcap_{j=1}^{k} |\hat{\theta}_j - \theta_j| \leq d \right) \geq 1 - \alpha . \]

Here \( \theta_j \) is the population proportion associated with the \( j^{th} \) category; \( \hat{\theta}_j = x_j/n \) is the observed proportion for category \( j \); \( k \) is the total number of component categories being considered; and \( d \) is a specified interval half-width. Note that we specify \( d \) rather than \( d_i \) because we are assuming that the target interval width is the same for all categories.

As with Equations 3.2 and 3.3, this interval is commonly expressed in terms of normal distributions to find a sample size. Alternately, it can also be expressed in terms of sums of binomials in order to determine an appropriate exact sample size for a symmetric interval.

To find sample sizes based on symmetric exact criteria, define the exact \( \alpha_j \) values such that \( \sum_{j=1}^{k} \alpha_j \leq \alpha \). Looking at the binomial marginals of the \( \theta_j \), we can define the probability, \( \alpha_j \), that an observed \( \hat{\theta}_j \) falls outside the given interval as

\[ \sum_{i=(x_j-c)}^{(x_j+c)} p(i; \theta) = \sum_{i=(x_j-c)}^{(x_j+c)} \binom{n}{i} \theta^i(1-\theta)^{n-i} \geq 1 - \alpha_i . \]

As before, \( c \) here is chosen such that the intervals are symmetric, and such that \( \sum_{j=1}^{D} \alpha_j \leq \alpha \).

3.2.4 Multinomial Exact Criteria

The expressions in Equation 3.5 can be adapted so that HD marginals are calculated
Here, $a_j$ and $b_j$ are selected such that $|a_j + b_j| \leq 2d$, and $na_j$ and $nb_j$ are integers. A simultaneous probability statement about the HD marginal intervals can be written as

$$P \left( \bigcap_{j=1}^{k} (-a_j \leq \hat{\theta}_j - \theta_j \leq b_j) \right) \geq 1 - \alpha .$$

(3.7)

Solving Equations 3.6 and 3.7 given $\alpha$, $d$, and $\Pi$ will give $n$, or appropriate HD sample sizes. This is the first time such an exact HD interval criteria has been extended to the multinomial case. As with the binomial case, the individual intervals can be asymmetric, and will require sample sizes equal to or smaller than for Clopper and Pearson (1934) exact interval. We also show that it requires smaller sample sizes than similarly defined intervals based on a normal approximation.

### 3.2.5 Computational Approach To Worst Case Parameter Vectors

Part of the appeal of using normal approximations to determine sample size for binomial and multinomial estimation problems relates to their simplicity and ease of calculation given the prevalence of standard normal probability tables. With faster computation available, it is now also straightforward to solve algorithms that can determine sample sizes based on exact criteria. We perform such a computational search of sample sizes for binomial and multinomial distributions, and base our algorithms for computation on already existing ones.

Similar to steps outlined in Thompson (1987) or Angers (1974), a procedure for evaluating worst case sample size for the HD binomial case is described below.

1. For any possible value of $\theta$, $0 \leq \theta \leq 1$, select an $n$ and compute $\alpha_n$ based on solving Equation 3.4. Here, $\alpha_n$ is the level of confidence for the HD interval for the given values of $\theta$ and $n$.

2. Compare $\alpha_n$ to $\alpha$. If $\alpha_n < \alpha$, step 1 should be repeated for a smaller value of $n$. If $\alpha_n > \alpha$, repeat with a larger value of $n$ until $n$ is the smallest value that can be found
satisfying $\alpha_n \leq \alpha$.

3. Steps 1 through 2 should be repeated for all possible values for $\theta$ in order to determine the worst case parameter value for $\theta$. It should be noted that due to discreteness of the binomial in the exact case, the ‘worst case’ will not be a single value, but a range of values requiring identical worst case $n$.

In practice it is not possible to calculate a required sample size for all possible values of $\theta$, $\theta \in [0, 1]$, but instead sample size requirements across a wide range of parameter values will be considered.

Similarly, this algorithm can be modified to allow for the discovery of worst case parameter vectors in terms of sample size for the multinomial case. Here we are looking for the case such that $\sum_{j=1}^{D} \alpha_j \leq \alpha$ for the specified parameter vector $\Pi = (\theta_1, ..., \theta_k)$.

1. Given a possible value for the parameter vector $\Pi = (\theta_1, ..., \theta_k)$, select a value for $n$, and compute each of the $\alpha_j$ for $j = 1, ..., k$ values based on Equation 3.6.

2. Take the sum $\sum_{j=1}^{k} \alpha_j$.

3. If $\sum_{j=1}^{k} \alpha_j < \alpha$, steps 1-2 should be repeated using a smaller value for $n$. If $\sum_{j=1}^{k} \alpha_j > \alpha$, repeat steps 1-2 with a larger value for $n$ until $n$ is the smallest value that can be found satisfying $\sum_{j=1}^{k} \alpha_j \leq \alpha$.

4. Steps 1-3 should be repeated for all possible parameter vectors $\Pi = (\theta_1, ..., \theta_k)$. As with the binomial, the ‘worst case’ will not be a single $\Pi$, but a range of parameter vectors requiring identical worst case $n$.

In terms of our implementation of the above algorithms, we started with $n = 1$ and progressed upwards in sample size by steps of size 1. Although Rahme and Joseph (1998) suggest a bi-sectional search algorithm to find correct values of $n$, when we compared our results to those in their paper, we found at least one occasion where their algorithm missed the correct minimum value for $n$. 
Binomial HD Parameter Settings

In order to understand the effect of changing the parameter \( \theta \) on sample size, it is only necessary to evaluate between \( 0 \leq \theta \leq 0.5 \) due to the symmetry of the binomial distribution. Sample size calculations are also affected by changes in \( d \), the desired interval half-widths, and changes in \( \alpha \), the level of significance. To understand how sample size can change depending on all these factors, we applied the binomial algorithm to a variety of combinations of levels of parameters \( \theta \), levels of significance, \( \alpha \), and interval widths \( d \).

Specifically, we evaluated required sample sizes for \( \alpha \) values of \((0.4,0.1,0.05,0.01,0.001)\), and for \( d \) of \((0.2,0.1,0.05,0.01)\) on a grid of 10,000 values for \( \theta \) ranging from 0 to 0.5. Sample sizes at each of the \( \alpha \) and \( d \) combinations were stored for each value of \( \theta \). This provides information about how required sample sizes changed as \( \alpha \), \( d \), and \( \theta \) changed. Information on which values of \( \theta \) result in worst case sample size requirements is also apparent in these results.

Multinomial HD Parameter Settings

For the multinomial, we evaluate sample sizes for a variety of values for \( \alpha \), \( d \), and parameters, \( \Pi \). In the multinomial case we must also consider the number of component categories for \( \Pi \). Referring to Chapter 1.2, we need to consider the dimension of the simplex being searched. Due to practical considerations, we restrict our sample size calculations to 3, 4, 5, and 6 simplexes. In Thompson (1987), the theoretical worst case parameter vectors were determined to occur in simplexes of 4 dimensions or smaller.

We evaluate \( \alpha \) at values of \((0.4,0.1,0.05,0.01)\). As in the binomial case, we evaluate at values of \( d \) including \((0.2,0.1,0.05,0.01)\).

In order to generate appropriate parameter values, \( \Pi \), for each of the simplexes, we created a regular cube (or appropriate hypercube) encasing a grid of points where combinations of values between 0 and 1 are considered in each dimension. We then only include points from the cube which fall on a standard \( k \)-simplex (where \( k \) is the dimension of our simplex and cube). This is easily achieved by including only points where the sum of the vector equals 1. For example, if we are considering a 3 dimensional cube, a point such as \((0.5,0.25,0.25)\) would be included.

As the dimension of the simplex being searched increases, so does the computational
time required to calculate sample sizes for parameter vectors contained within it. We can improve the speed of the search by taking advantage of some key features of simplexes. A k-dimensional simplex will have \( k! \) isometries within it. For example, in the binomial case, we only needed to consider values between 0 and 0.5, because these results will be symmetric with those found between 0.5 and 1. Similarly, a 3-simplex will have \( 3! = 3 \times 2 \times 1 = 6 \) isometries. Our 3-simplex triangle defined by the points \([(0,0,1),(0,1,0),(0,0,1)]\) is made up of scalene triangles which will have vertices \([(1,0,0),(1/3,1/3,1/3),(1/2,0,1/2)]\) and all possible permutations thereof. Sample size results from one of these smaller triangles can then be expanded to the full simplex. Similar arguments can be made for higher level simplexes.

A second feature that we are able to take advantage of is that points from the standard simplex (ie. \( \sum \theta_j = 1 \)), can mapped onto an arbitrary n-simplex based using barycentric methods. Specifically, if we have a standard \( k \) simplex with vertices

\[
\begin{align*}
  l_0 &= (1,0,\ldots,0) \\
  l_1 &= (0,1,\ldots,0) \\
  \vdots \\
  l_k &= (0,0,\ldots,1),
\end{align*}
\]

then we can map points from this standard simplex to an arbitrary simplex defined by \((v_1,\ldots,v_k)\) (eg. \([(1,0,0),(1/3,1/3,1/3),(1/2,0,1/2)]\) ) given by the relation

\[
(\theta_1,\ldots,\theta_k) \mapsto \sum_{j=1}^{k} \theta_j v_j.
\]

The positive benefit of these features is that we are able to generate points on a standard \( k \)-simplex, which is relatively easy, and then project them down to an isometric \( k \)-simplex. By permuting points on the isometric \( k \)-simplex, we can see results for the full \( k \)-simplex while searching a much smaller space in much greater detail in less time than it would take to search the full area. This detailed examination can then be reflected to create results for the whole simplicial area, giving us six times the density of points compared to what would have been considered if we had performed the evaluation on a full simplex. It is easier to define points on the standard simplex than on the isometric one, making the mapping
property very useful in examining which parameter values $\Pi = (\theta_1, \ldots, \theta_k)$ lead to worst case sample size results.

### 3.3 Results

In addition to information about where worst case parameter sets are occurring, one of the benefits of searching a full range of parameter values is that we also find information about $\Pi$ where small and medium sample size requirements are occurring.

One of the benefits of searching a wide range of parameter values is that in addition to information about where worst case parameter sets are occurring, we also find information on parameter sets where small or medium size sample requirements are occurring. In Chapter 3.3.1-3.3.2 below, we illustrate a variety of the sample size results for binomial, 3, 4, 5, and 6 simplex parameter spaces in terms of their required sample sizes and worst case parameter vectors. Not all results are shown here, although the complete set of graphs and tables that were generated based on this sample size study can be found in Appendix C.

#### 3.3.1 Binomial Results

In Figure 3.1 below, we show some sample size results for the binomial distribution when using a half interval width of $d = 0.1$. A key feature to notice in this Figure occurs when $\alpha$ is set to a value of 0.4, and sample sizes are evaluated for $0 \leq \theta \leq 0.5$. The worst case sample size does not appear at $\theta = 0.5$. By only evaluating sample size at the hypothetical worst case parameter value based on the normal approximation, this information about required binomial sample sizes would have been lost.

It is also interesting to note that this type of ‘jump’ in required sample size does not just occur when calculating worst case sample sizes, but also occurs at other values for $\theta$. This can also have implications when designing a study.

As $\alpha$ is decreased, the required sample size tends to increase. As the sample sizes increase, the location of the worst case sample size tends to converge on the value we would expect based on the normal approximation, with the worst case appearing at $\theta = 0.5$ (as well as several other values for $\theta$ which are also worst case). The trends observed in Figure 3.1 are fairly typical to those found for other values of $d$. These other results can be
seen in Appendix C in Figures C.1 through C.3.
Figure 3.1: Exact highest probability sample sizes required for binomial distributions when $d=0.1$, values of $\theta$ are between 0 and 0.5, and $\alpha$ values are 0.4, 0.1, 0.05, 0.01, 0.001.
3.3.2 Multinomial Results

3 Simplex Results

As with the binomial results shown in Figure 3.1, it is also possible to graph some of the HD sample size results for the 3-component multinomial distribution. We represent sample sizes on a ternary diagram as described in Chapter 1.2. Here different colour intensities on the ternary diagram are associated with different sample size requirements for the distribution. Similar to Figure 3.1 for the binomial distribution, Figure 3.2 for the three component multinomial distribution also shows sample size results that are not consistent with what we would have expected based on normal approximation results.

Recall Thompson (1987) where the ‘worst case’ values for $\Pi = (\theta_1, \ldots, \theta_k$ were identified as occurring at values of $\theta_i = 1/m$ for $m$ categories, and 0 for the remaining $k - m$ categories. For these normal approximation results, the value of $m$ changed depending on the level of significance, $\alpha$, being considered for the interval.

In our exact results shown in Figure 3.2a, it is clear that the worst case sample size occurs at a value near to $\theta_1 = \theta_2 = \theta_3 = 1/3$, but does not include this parameter vector. In many ways this seems reminiscent of Figure 3.1 where, when $\alpha = 0.4$, the HD worst case occurred near to, but not at, the hypothesized (based on approximation) value of $1/2$. Looking at sub-Figures 3.2b through 3.2d, hypothetical worst case parameter vectors at $1/m$ (in this case $m = 2$ or 3) are seen on the graph. These hypothetical parameter vectors are enclosed in a wide area of equally worst case parameter vectors. When using exact HD criteria, the worst case parameter vector falls in a range, not at a single point. In Figure 3.2, where $d$ is wide, the worst case range of parameter values is also wide. This range of worst case values is different than approximation methods which give only a single worst case point.

Ternary diagrams illustrating the effect of decreasing $\alpha$ and $d$ on sample sizes are shown in Figure 3.3. Here the worst case sample sizes are larger, as is the range of sample sizes required for the parameter vectors studied. Visually these changes in required sample sizes are more difficult to differentiate than in 3.2a. To clarify where worst case parameter vectors are located, Figure 3.4 only shows where the largest 10% of sample sizes are occurring for the given values of $\alpha$ and $d$.

Figures 3.3 and 3.4 indicate that it is possible to identify worst case areas near to and
Figure 3.2: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.2$ at $\alpha = 0.4, 0.1, 0.05,$ and $0.01$ (corresponding to relatively small sample sizes). Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
Figure 3.3: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively large sample sizes). Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
CHAPTER 3. EXACT APPROACH TO MULTINOMIAL SAMPLE SIZE

Figure 3.4: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Figure shows only the largest 10% of sample size requirements. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively large sample sizes). Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
including $1/m$ parameter vectors (the value of $m$ varies with the conditions being considered). Unlike Figure 3.2, the worst case zones in Figures 3.3 and 3.4 actually include the $1/m$ value, as opposed to merely being close to $1/m$. The area representing worst case parameter vectors is still not a single point, but the worst case zone has become more focused. Despite a more clearly identified worst case zone, it is also interesting to note that the difference in sample size between a worst case zone and a next to worst case zone is relatively negligible compared to the total possible range of sample sizes.

Opposite of worst case, it is also interesting to note the points of the ternary diagrams which represent parameter vectors that include very small (or large) values for at least some $\theta_i$. The areas representing such parameter vectors tend to require relatively small sample sizes to reach the specified $d$ and $\alpha$. When using a normal approximation it can be very difficult to determine sample size estimates for $\theta_i \to 0$ or $\theta_i \to 1$.

To more closely observe the effects of $d$ and $\alpha$ on multinomial sample sizes, additional graphs have been included in Appendix C and Figures C.4 - C.6.

All Simplex Results

In simplexes higher than 3, it unfortunately becomes impossible to summarize the results in graphical form. Fortunately, for many combinations of $d$ and $\alpha$ the worst case and almost worst case parameter vectors occur when $m$ is three or less. This leads to similar conclusions as would be gained from observing graphs such as Figures 3.2, 3.3, and 3.4. In fact, the worst case for $m$ is never found to be higher than 4 in the empirical study presented here. To summarize these results where worst case parameter vectors are occurring across all simplexes considered, we present Table 3.1. For comparison, in Table 3.1 we create using the multinomial sample size criterion from Thompson (1987) for normal approximation cases. Tables 3.1 and 3.2 present results based on identical values for $d$ at a variety of values for $\alpha$.

One of the most striking results demonstrated in the tables is that the worst case values for $m$ identified using the normal approximation are also identified as worst case (or at least close to) values for $m$ in the exact case. It should be noted though, that in the exact case $1/m$ forms the nucleus of a worst case region, and a range of parameter vectors can be considered worst case near this nucleus. In the normal approximation, the worst case parameter vector is located at a specific point in the parameter space. Despite this, the
Table 3.1: Exact multinomial sample size results when the half interval width \(d\) is set to 0.05 across all simplexes evaluated. We identify \(m\) near where the worst case sample size results are occurring.

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(n) with (d = 0.05)</th>
<th>(m) (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>190</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>390</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>490</td>
<td>2 or 3</td>
</tr>
<tr>
<td>0.01</td>
<td>770</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 3.2: Normal approximation multinomial sample size results when the half interval width \(d\) is set to 0.05 across all simplexes. We identify \(m\) where the worst case sample size results are occurring. Based on interval definitions from Thompson (1987).

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(n) with (d = 0.05)</th>
<th>(m) (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>203</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>403</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>510</td>
<td>3</td>
</tr>
<tr>
<td>0.01</td>
<td>788</td>
<td>2</td>
</tr>
</tbody>
</table>

3.4 Discussion

Information gained using the HD method to calculate binomial and multinomial exact sample sizes clarifies the location of worst case regions for binomial and multinomial models, and reduces sample size requirements. Such required sample size information is important as it relates to study planning and sample size problems.

For most levels of \(d\) and \(\alpha\), the sets of worst case parameter vectors that were identified were closely related to those that were established in Thompson (1987); Fitzpatrick and Scott (1987). When determining the worst case based on exact criteria we found that even for small \(d\) and \(\alpha\), corresponding to large sample size requirements, the worst case was
never uniquely identified. These results can clearly be seen in the graphs of Chapter 3.3 where worst case parameter areas are identified on the ternary plots.

In terms of study planning, information about the relationship between exact and normal approximation based sample sizes is useful. When we do not have insight about the true values of parameters of interest, it is common to take a conservative approach to selecting a sample size. In the binomial case using the normal approximation this is equivalent to choosing \( \theta = 0.5 \). In the exact case, we can use normal approximation results about which \( m \) will be worst case, and then use exact criteria to evaluate the sample size required for parameter vectors at or near this worst case. Identifying where worst case parameter vectors are likely to occur saves time over searching the entire parameter space. Calculating sample sizes for a worst case parameter vector using exact criteria provides the benefit of reduced sample sizes compared to those based on normal approximations. Given that collecting sample can be the most cost and labour intensive aspect of conducting research, smaller sample sizes that yield intervals of specified width at the correct confidence level are an important contribution.

A caveat to the above occurs when \( \alpha \) and or \( d \) are relatively large. When this occurred, we tended to see increases in sample size required at unexpected values of parameter vectors as a side effect of the discreteness of the binomial distribution. When searching for sample sizes under these conditions, it is recommended to search a greater part of the parameter space to ensure that the true worst case is identified.

Moving away from study designs that are planned based on worst possible cases for parameters, our examination of sample sizes at other possible parameter values is also useful. Many studies are concerned with outcomes that fall near to 0 or 1; examples would be in clinical trials where such low probability outcomes are often associated with serious and undesirable side-effects (Wells, 1999). Where normal approximations perform poorly at such parameter values far from 0.5, there are no such problems using our exact criteria. In all cases, the sample sizes required to estimate such extreme probabilities at specified \( d \) and \( \alpha \) are small relative to their worst case requirements. Using such exact criteria offers great improvement over approximate approaches that tend to perform poorly in this area of the parameter space, and guarantees the correct level of confidence when desired interval widths are equal.

Required sample sizes given a fixed \( d \) and \( \alpha \) depend on the parameter vector being
evaluated; if the underlying parameter vector changes, identical sample sizes will not confer identical precision in estimates. Because sample size requirements change as parameter vectors do suggests that methods to adapt based on previously observed parameter estimates might be useful in application. If we are trying to achieve estimates with similar levels of confidence and interval widths across various strata/clusters/etc., adapting sample size based on anticipated or historic parameter values would likely give improved results.

Viewing estimation involving binomial and multinomial data on exact results, rather than as an application appropriate for normal approximation, has several benefits as outlined above. Using our HD exact approach reduces necessary sample sizes. Considering exact criteria also forces practitioners to think about how study design and estimation should be tailored to give better results based on characteristics of binomial and multinomial distributions rather than approximations. Multinomial inference, especially at smaller possible sample sizes, should not simply be treated as an extension to normal problems.

Use of normal approximations for problems involving binomial and multinomial data is fairly ubiquitous. Such approximations provided a simple approach to solving problems when computers were not as available or fast as they currently are. The normal approximation is also an excellent teaching tool when first introducing normal distributions, central limit theorem, and data in the form of proportions. Despite historic practicality, and value to teaching, our HD exact approach is preferable to an approximation one in many ways. Defining a multinomial data collection problem in terms of the exact approach has the potential to provide improvements to inference for this type of data. Smaller sample size requirements will also reduce sampling costs. Given the availability of fast modern computers, and the relatively simple implementation of an algorithm for the exact approach as outlined, we strongly recommend this approach over normal approximations when considering actual problems involving multinomial data.
Chapter 4

Bayesian Ignorable Sampling

4.1 Background

It is an often repeated maxim in statistics that ‘all models are wrong, but some are useful’ (Box and Draper, 1987, p.424). Ideas in Rubin (1984) reflect this sentiment, and we build on such ideas while developing a simulation approach to evaluating sampling strategies for Bayesian models. Here, we discuss using simulation models that may be an imperfect reflection of reality, but are useful tools when developing a simulation approach for evaluating sampling strategies.

We develop a framework that can be used to explore ways in which ignorable sampling approaches for collecting data, y, can affect the reliability of inference from our model. Based on a simulation approach and definitions of ignorable sampling, we compare sampling strategies for the ecological monitoring example described in Section 1.1.

Derived for this motivating example, we develop a simulation model for Fraser River sockeye salmon. These simulations can be used to emulate data for a number of different ignorable sampling approaches. We develop several such ignorable sampling approaches, describing why they are ignorable in an intuitive way. Our intent is that the strategies should be practical in a real world setting, and that a description of their development allows us to provide some general guidelines to follow in order to develop ignorable sampling strategies for other Bayesian models. The simulation model provides reference distributions to which Bayesian estimates resulting from different sampling approaches can be compared.
4.1.1 Superpopulation Simulations To Examine Operating Characteristics Of Bayesian Models

In Rubin (1984) the author refers to a Bayesian statistician’s model using notation similar to that defined in Chapter 1.3.1, i.e. \( p(\theta)p(y|\theta) \). He then explores the idea that, even for Bayesians, this model is sometimes chosen because it is computationally achievable, and or convenient in some way. This implies that the observed data are generated by an unknown correct model, say \( G(y|\theta)Q(\theta) \). We can make probability statements regarding such models, for example, a 95% Bayes interval \( I_j \) about a parameter(s) of interest for \( \theta_j \). From a frequency standpoint, our goal for a variety \( I_j \) should show that 95% of the intervals cover their unknowns while 5% do not. The article also argues that such frequency based results (which can often be obtained through simulation) can be interpreted under a Bayesian paradigm. Essential to this is the idea that results are interpreted under the premise that data are known values of random variables; that unknown values are treated as unobserved random variables; and that the conditional distribution of the unknowns given data and model specification is obtained based on Bayes’ theorem.

In the context of a statistician’s model, if \( p(\theta)p(y|\theta) = G(y|\theta)Q(\theta) \), then the probability coverage of our intervals, \( PC(I_j) \), should be correct for all \( y \). Often, \( p(\theta)p(y|\theta) \neq G(y|\theta)Q(\theta) \) and \( G(y|\theta)Q(\theta) \) is unknown. The goal is that the statistician’s model \( p(\theta)p(y|\theta) \) be chosen such that when \( I_j \) are chosen, the distribution of their probability coverage will be as close as possible to the stated probability. This should hold for a range of plausible \( G(y|\theta)Q(\theta) \), i.e. such that \( E(PC(I_j)) \approx 0.95 \). Although model specification has an effect on the distribution of \( PC(I_j) \), data collection does also. An example taken from Rosenbaum and Rubin (1984) shows the effect of a data dependent stopping rule. In the example, the posterior calculations for \( \theta \) are the same regardless of whether simple random sampling, or sequential sampling with a stopping rule is used. Despite this, the sampling distribution of \( PC(I_j) \) is affected by the stopping rule applied. Similarly, how sensitive estimates are to departures from the true model can also depend on sampling design.

This example clearly demonstrates that even when ignorable in the context of the model used for inference, data selection decisions can affect the reliability of inference. We are particularly interested in how simulation approaches can be used in a sampling design context for Bayesian models.
4.1.2 Ignorable Sampling

In Chapter 1.3.2 we introduce the idea of ignorable sampling approaches in the context of Bayesian modeling. When ignorable sampling is, or can be achieved, the sampling design can be considered ignorable for further inference. Bayesian models based on observational data are common in the literature, and because data are viewed as fixed, discussions on the effects of sample selection are often neglected. While some suggestions for sampling for observational studies exist, they tend to focus more on cases where study units are assigned to groups and observed (Rubin, 1974) with structures analogous to designed experiments. There is less discussion of sampling for cases when processes in nature are being monitored.

The relationship between observed data and potential/unobserved/missing data is an important one when considering data collection in relation to Bayesian models. Although data are considered fixed in a Bayesian context, Equation 1.4 in Chapter 1 demonstrates how to incorporate information about missing data patterns into Bayesian inference methods. It is often desirable to ignore processes by which missing data occur. We reiterate two key conditions that are required in order for a sample to be considered ignorable. With ignorable samples the process by which missing data occurs does not need to be considered when estimating key parameters of interest. These two criteria are that

1. missing data must be missing at random such that the probability of missing data depends only on \( y_{obs} \) (not \( (y_{obs}, y_{miss}) \)), parameters \( \phi \) associated with missing data, and covariates \( x \); and

2. the condition of distinct parameters such that parameters \( \phi \) associated with the missing data process are independent of, \( \theta \), our parameters of interest (Gelman et al, 2003).

In the Gelman et al (2003) summary on accounting for data collection in a Bayesian context, the authors discuss how the criteria above can be applied in the experimental design context, in the sample survey context, and in experiments where censoring and truncation occur. Experimental design provides an excellent conceptual framework for imagining different potential data sets that could be observed, and for considering the effects of how
data were collected. Essentially, experimentation involves assigning treatments to experimental units. Any given experiment will have a specific allocation of units, providing us with our observed data. We can imagine that treatment allocation could have occurred differently, and can create permutations of all possible treatment allocations of the given units. This provides conceptual potential data. How treatments are allocated to data, and the way in which this affects potential data, determines the ignorability of the design.

Similar arguments about permutations of observations can also be made in the sample survey context (Little and Rubin, 1987). Examples involving censoring and truncation generally describe how the missing data mechanism can be incorporated into an analysis; missing data in such cases cannot be treated as ignorable.

As stated earlier, most of the discussion regarding sampling designs for observational data seems to centre around data that are analyzed as if they came from an experiment (Gelman et al, 2003). There is a gap in the literature for sampling for models used for observational data. Many models for observational data, such as the one we demonstrated in Chapter 2, are modeling a process with the intent to gain inference about that process; there is no intent to treat the observed data as if it came from an experiment. Environmental monitoring problems are a common source of observational data, and inference is commonly achieved through the use of Bayesian models (Shaddick and Wakefield, 2002; Schmidt and O’Hagan, 2003). Despite their prevalence as an inference method, sampling considerations in terms of ignorability of data for the chosen models are rarely explored.

### 4.2 Notation and Simulation Justification

Earlier we defined $p(y|\theta)p(\theta)$ as the statistician’s model, or the model being used for estimation. Subsequently, $G(y|\theta)Q(\theta)$ was used to represent the unknown correct model that actually generated the observed data. Given our lack of knowledge about $G(y|\theta)Q(\theta)$, it can sometimes be useful to create a simulation or reference model that can generate data containing features of interest. Such reference data are created using known distributions and their parameters. Reference models like this could be termed $g(y|\theta)q(\theta)$. Creating such a model in the context of a motivating problem can help to provide information, including in the context of evaluating several different ignorable sampling approaches. We advocate that using a simulation approach to evaluate sampling methods is sensible and should be
adopted in practice.

By using the statistician’s model to generate estimates based on simulated reference data, we can gain an understanding of the robustness of the statistician’s model when estimating the parameters of interest. This is especially useful when comparing how data simulated from $g(y|\theta)q(\theta)$ based on different sampling approaches can affect estimations from the statistician’s model.

Having separate statistician’s and simulation models is particularly useful when applied environmental monitoring problems for several reasons. One is that data collection generally occurs regularly and over long periods of time. This means that sampling and estimation will recur on a regular basis, making the effort of improving sampling and estimates worth the time and effort. Second, especially in environmental monitoring programs, the observations taken from the system can vary greatly from year to year based on processes that we often do not fully understand. By testing the statistician’s model used for inference against simulated reference data, statisticians can work to ensure that the models that they use will be robust and useful, even if not true.

4.3 Simulation Model

In Figure 1.2 from Section 1.1, we illustrated the idea that stock abundances tend to follow bell-shaped curves, starting with a few fish from any given stock migrating up the river, followed by an increase in the number of fish migrating, with abundance of migrants gradually declining through time. Despite this theoretical ‘normal through time’ stock abundance pattern, in modeling the observed data in Chapter 2, we merely assumed that relative stock abundances were correlated through time. We ignored the theory that absolute stock abundances might follow a normal shaped temporal pattern.

The CH MVCAR approach was chosen because in estimating daily stock proportions we were able to share information between days, improving variance estimation. We were also able to estimate how stock specific estimates were correlated. Because no stock abundance information was apparent in the observed data, in building an estimation model we never considered a hierarchy that was based on a theory of stock abundance curves.

In evaluating the usefulness of the statistician’s model, basing observed proportion simulations on the normal through time abundance theory becomes a natural fit as a simulation
model. Simulations based on such a model will contain characteristics similar to those in the observed data. A normal abundance through time simulation approach allows us to generate simulated observations based on different sampling strategies.

When building our simulation model, we make a simplifying assumption that there are only 3 stocks to be considered, rather than 5.

Figure 4.1 mirrors the same ideas as seen in Figure 1.2, but in the 3 stock case that will be used for simulation. The lower graph illustrates hypothetical run-timing curves for our three stocks, and the upper graph shows the total abundance through time. The upper graph is colour coded so that the relative contribution of each stock to the total abundance is visible. Visualizing this relative contribution is important, because in our sample data we see only the relative abundance of each stock, not overall abundance. Our goal is to transform these normal shaped abundance curves into observed proportion data.

Figure 4.1: Theoretical run timing curves in lower graph, added together to conceptually show individual stock contributions to overall abundance in the upper graph

Because the abundance graphs are ‘normal’ in shape, we use quantiles of the normal distribution to help create our relative proportion data. Key to using normal quantiles is knowing the parameter values \( \mu \) and \( \sigma^2 \). Rather than assigning these directly, we connect
the idea of time to the normal by assigning a runlength, or the number of days we want it
to take for 99% of the fish in a stock to pass by the sampling location. Because the normal
is symmetric, the mean (time of maximum abundance) for a given stock can be found as

\[ \mu_{stock} = \frac{\text{runlength}}{2} \]  \hspace{1cm} (4.1)

Note that this runlength is stock specific and does not reference the total run-timing of
all stocks in the river. Incorporating such relational information is discussed after Equation
4.3.

Knowing that 99% of the stock passes over the runlength implies \( Z^* = 3 \). Based on this
we can calculate the appropriate stock specific standard deviation’s

\[ \sigma_{stock} = \frac{x - \mu_{stock}}{Z^*} = \frac{\text{runlength} - \mu_{stock}}{Z^*} \]  \hspace{1cm} (4.2)

Having determined the parameters of a normal through time, and knowing that we intend
to sample on a daily basis, we can calculate the proportion of the stock passing by the
sampling point on any given day. We assume that the proportion of fish passing on day
\( t \), \( p_t \), is represented by the proportion of area under the normal between time \( t \) and \( t - 1 \),
which can be calculated by finding \( Z_t, Z_{t-1} \) and using the cdf of the normal distribution

\[ Z_t = \frac{t - \mu_{stock}}{\sigma_{stock}} \]
\[ Z_{t-1} = \frac{(t - 1) - \mu_{stock}}{\sigma_{stock}} \]
\[ p_t = \Phi(Z_t) - \Phi(Z_{t-1}) \]  \hspace{1cm} (4.3)

We then need to consider the relationship of the different stocks to each other, both
temporally and in magnitude. We assign one of the stocks to be first, and offset the start
times of the rest of the stocks relative to this. For example, in Figure 4.1, our first stock is
represented by the red curve. The second stock to enter the river is represented by the green
curve, and it begins its run past the sampling location 3 days after the red stock. Finally,
the third, or blue stock, migrates past the sampling point 15 days after the first stock, and
12 days after the stock represented by the green curve.
Only relative magnitudes of the different stocks are required for the simulations. We represent them with the variable $M_{stock}$. In Figure 4.1, the relative magnitudes of stocks identified by the colours Red : Green : Blue are 2.5 : 1 : 3. If we consider the total runlength to consist of the time it takes for all stocks to pass by the sampling location based on their individual runlengths and offsets, we can use the notation $T$ to represent the timepoint in the total runlength under consideration.

We can then calculate the relative stock proportion, $\pi_i T$, ($i = 1, ..., D$), for each stock on each day, $T$, of the simulation.

$$\pi_i T = \frac{M_i p_i T}{\sum_{i=1}^{D} M_i p_i T}.$$ 

Transforming the normal data shown in Figure 4.1 into relative proportions based on the calculations outlined, we graph the normal through time relative stock proportions in Figure 4.2.

Figure 4.2: Theoretical relative abundance curves based on converting normal abundances into proportions assuming one sampling event daily.
These curves then represent relative stock proportions based on the normal model, assuming no sampling or other errors. Discontinuities in the curve appear when stocks either enter or leave the river according to our theoretical model. Because error is a realistic part of science, we then introduce some process error to the relative proportions. In our example, this error could represent ways stock timings vary from the underlying model due to environmental factors such as tidal effects, weather, etc. We introduce this error by drawing proportions \((\Pi_1T \ldots \Pi_{DT})\) from a Dirichlet at every time point with parameters \(\gamma \times (\pi_1, \ldots, \pi_{DT})\).

Following this, we add sampling variation to the \(\Pi_{iT}\) by taking multinomial draws. This gives us \(\hat{\pi}_{iT}\), or observed proportions including error. Sample size selection, and days where samples should be collected are determined by the sampling method being considered. Such methods are developed in Section 4.4 below. In Figure 4.3 we demonstrate observed proportion curves where the sampling strategy includes missed days. This gives an idea of the type of data that can be generated using the simulation model described above.

### 4.4 Ignorable Sampling Approaches

In developing ignorable sampling designs, it is important to recognize that just as more than one inference model can be applied to the same data, more than one sampling design can be ignorable in the context of a single model.

Covariates are key to determining whether a sampling approach will be ignorable for a given model. In fact, with enough covariates included in the model, almost any sampling assignment can be treated as ignorable (Gelman et al, 2003). This does not provide statistical modelers license to collect covariates willy-nilly in order to ignore sampling problems! Implicit in the selection of covariates, is the idea that ‘known’ covariates need to be selected from those values that are available, and restricted to those that are worthwhile.

Regardless, we recommend that an important first step in developing ignorable sampling schemes for models is identifying which covariates are implicitly and explicitly incorporated into the model’s structure. Although this seems self-evident, when we created our own model, we were thinking about the model in terms of improving parameter estimates, but not explicitly in terms of how the data were collected. Fortunately, as we shall
Figure 4.3: Theoretical samples based on normal underlying abundance curves converted to proportions. Proportions are then given Dirichlet process error, and multinomial sampling error to reflect real-world causes of variability. Missing days occur with the sampling method illustrated, such days appear as gaps in the lines.

shortly describe, our model was ignorable for the sampling approach that was actually used to collect the data. This will not be the case for all models. Clearly articulating covariates that were conditioned on in the model made identifying ignorable sampling designs a simpler process. We recommend that such a step should always be used when developing sampling methods for hierarchical models.

In Chapter 2, we describe a model for multinomial data where observations are correlated based on a temporal neighbourhood structure. Keywords to notice in the previous sentence are *temporal* and *multinomial*. Because we are considering temporal correlation, time, or the order of collection of samples is known and accounted for in our model. We are conditioning current observations based on ‘neighbour’ observations, and can assume that observations in days adjacent to the current one are known. Due to samples being multinomial, knowledge of both sample size and sample proportions on a given day is implied. Other than time and information about previous days’ observations, the model does
not include any explicitly collected covariates.

Note that in discussing sampling approaches, we are assuming that fish within any given day represent a random sample of fish in the river. Our focus is to consider on which days sampling should occur, and how much sampling should occur on those days.

In-river sampling during fishing season is chaotic. We try to develop ignorable designs that could at least be somewhat plausible in practice. Although our sampling approaches discuss sample size, we recognize that in practice this will translate to fisheries effort, where equal ‘effort’ will result in approximately equal sample sizes. Our approaches cannot be so specific that they would be ruined by imperfect sampling in the field; they are intended to provide a practical guideline.

Conversations with subject area and statistical experts were invaluable in brainstorming and discussing different possible sampling approaches. We found the viewpoints about data collection expressed by these two different groups very diverse; the fisheries biologists often intuitively preferred sampling approaches that statisticians felt would perform poorly. For example, statisticians tended to favour balanced sampling approaches, and fisheries biologists chose to dramatically increase sample sizes after missed days of sampling. Despite the lack of consensus about sampling approaches, this discussion of sampling strategies from different perspectives was a valuable exercise in thinking about sampling in critical and practical ways. Our first recommendation for developing designs was to explicitly list covariates present in the model. Our second recommendation is to have conversations with as many subject area experts as possible. This includes suggestions from both experts in both statistical and scientific application subject area experts. Even if no obvious ‘best’ approach emerges, the process of understanding sampling constraints and identifying potential methods of improvement is well worth the time taken in discussion.

4.4.1 Conventional and Non-Adaptive Sampling Approaches

Below are descriptions of sampling approaches that we identified as being practical based on both statistical understanding and the real world constraints of our motivating example.
Balanced Approach

In what we are terming a balanced sampling approach, the target is to collect samples of equal size during every day of the study period. Balanced designs, where equal sample sizes are collected from different groups, often have desirable statistical properties compared to unbalanced design approaches. In terms of the criteria for ignorable sampling designs, this design can be viewed as ignorable. Missing data are missing at random, because we are assuming that sampling occurs in all days, and no days are missed. In the context of the Fraser River fisheries data, this design is essentially a systematic sampling approach, where samples are taken daily. Because we incorporate the order of data collection into our model, we can consider the selection criteria ($\phi$) fixed, and independent of the parameters of interest, $\theta$.

Balanced Approach With Missed Days Occurring At Random

Above we take a fairly rosy view of sampling, and assume that no days are missed during the period of interest to fisheries managers. Although ideal from a statistical perspective, based on the practical realities of sampling such a belief is impractical. Missing days occur for a variety of reasons, and anecdotal discussions included comments on lack of sample storage equipment, spoilage of samples, insufficient time to collect samples, etc. To reflect this reality, we modify the balanced sampling approach to allow for missing days. We allow missing days that are assumed to occur at random. Similar to above, this also results in an ignorable sampling design. Previously missing days were assumed to occur at random (because there were none); here we assume that they actually occur due to some random chance not associated with the underlying stock proportions that we are trying to estimate. Therefore, our missing data are missing at random, and our parameters of interest, $\theta$, are assumed to be independent from the causes of missed sampling occasions. Note that although we discuss this design where sampling times are missed through accident, they could also be assigned deliberately. Such random assignment, rather than assumption of random occurrence could reduce the likelihood that researchers unintentionally ‘cheat,’ and select missing days based on criteria related to $\theta$ (Gelman et al, 2003, p.225).
Balanced Approach with Missed Days on Weekends

In a second variation of a balanced sampling approach allowing for missing days, here we assume that we deliberately miss sampling on one weekend day. Although many statisticians will intuitively find this approach less appealing than one where missing data are occurring at random, both sampling approaches can be thought of as ignorable. Because we are conditioning on a known covariate, time, this model can be considered ignorable for similar reasons to the above design. Conditional on the covariate, we can consider missing data to be missing at random. We are also assuming that our criteria for missing a day of sampling (it being a weekend) is independent of our parameters of interest, the observed stock proportions. This is actually quite a strong assumption in this context. For example, week to weekend behaviour of human populations is known to be different, and if these anthropomorphistic factors affected fish migration, our assumption that selection parameters were independent of the parameters of interest would no longer be true.

Despite the stronger assumptions that are required to justify this as an ignorable design, we include it to emphasize the importance of the human factor in collecting samples and discussing which design is most appropriate. In a personal conversation with employees of the Pacific Salmon Commission, one of the complaints of a fisherman collecting genetic samples was that he was required to work every single day, missing out on summer vacations with his children and family. When I presented the model of Chapter 2, one of the first questions (semi-jokingly) put to me by one of the scientists there was ‘Oh, your model accounts for missing days, does that mean that we can take weekends off in the summer?’ Given that missing days are bound to occur, including a sampling strategy such as this, allows discussion of ideas of whether missing days should be planned rather than accidental. If planned, should they be planned systematically or by some other method? It is also a reminder that even if a certain sampling method is statistically preferable, the social and economic costs of such a sampling design might make it unattainable.

4.4.2 Response Dependent Adaptive Approaches

The key difference between non-adaptive approaches and adaptive approaches is that the sampling design is at least partially dependent on observations of the variable(s) of interest that have already been collected during sampling. Below, we discuss how such adaptive
designs can be ignorable in a Bayesian context. To the best of our knowledge, this is the first time that adaptive sampling designs have been presented in the context of improving Bayesian inference.

**Double Sampling After Missing Days**

Although initial conversations with fisheries managers indicated that a balanced approach to sampling was their goal, further discussion revealed that the actual sampling method resembles the one described here. The target was to sample approximately 50 fish per day; unfortunately, some days were missed (presumably) at random. In order to maintain an average sample size of 50 fish per day, sampling efforts were doubled after a missed day. The target number of fish to catch after a missing observation was 100. This is an adaptive sampling method, because we are adapting our target sample size after ‘observing’ a missed day. This design is also ignorable for reasons similar to those for the balanced approach with days missing at random. Changing the size of the sample collected after missing days does not change the ignorability of the sampling design because daily sample sizes, and the neighbourhood of samples, are modeled as covariates.

**Sampling Effort Proportional to Multinomial Worst Case**

In Chapter 3, we illustrated sample size requirements for the multinomial distribution based on HD intervals. One of the conclusions of this work was that different parameter vectors had different sample size requirements in order to obtain interval widths of equal precision. Different combinations of interval widths and confidence levels were associated with unique worst case parameter settings. In this sampling approach, we suggest adapting sample sizes so that sampling effort is increased when parameter estimates are near to worst case, and then decreased when estimates of parameters move further from worst case. Considerations for developing guidelines to do this include deciding what parameter vectors should be considered worst case based on the range of sample sizes likely to be collected; defining when parameter estimates are near to worst case; and specifying what sample size targets should be once proximity to the worst case is determined.

In the motivating data set, sample sizes ranged from a minimum of 16 to a maximum of 96 fish sampled. This range of sample sizes corresponds most closely to those seen in
CHAPTER 4. BAYESIAN IGNORABLE SAMPLING

Figure C.4, where $d = 0.1, \alpha = 0.1$. Required sample sizes range between 5 and 95 in this case. Given that the required sample sizes for the identified settings are similar to those observed in the Sockeye fishery, we use these results as a guideline for our adaptive rule.

We intend to adapt the current day’s ($T$) sample size based on what stock proportions were observed on the previous day ($T - 1$). Because the observed proportions are likely to contain sampling error, and the actual worst case parameter vector is not unique, we partition the parameter space into parameter estimates that will trigger low, medium, and high sampling effort. Using the observed proportion from $t - 1$, we calculate the required sample size, $n_{opt}$ based on $d = 0.1$, and $\alpha = 0.1$ as outlined in Chapter 3.

\[
    n_{\text{target}} = \begin{cases} 
    30, & \text{if } 0 < n_{opt} < 65 \\
    50, & \text{if } 65 \leq n_{opt} < 85 \\
    70, & \text{if } 85 \leq n_{opt} \leq 95 
    \end{cases} \tag{4.4}
\]

Other adaptive rules based on the multinomial distribution are possible, but we chose the above as providing a reasonable balance between practical and theoretical considerations.

Having developed an adaptive strategy for sampling based on multinomial results, it is also necessary to discuss its ignorability for our Bayesian model. Because we are assuming no missed days of sampling in this model, missing data can be considered missing at random. Similarly, because sample size and proportions from day $t - 1$ are included in the model as covariates, we can consider that our sampling strategy is independent of our parameters of interest.

**Sampling Effort Proportional to Multinomial Worst Case with Missed Days Occurring at Random**

This strategy is similar to the one outlined above, except that we add more realism to the approach by allowing for days where sampling is missed due to chance circumstances. As with the sampling design we called balanced approach with missing days occurring at random, because the missing data are missing at random, and the underlying sampling design is already ignorable, this sampling design will also be ignorable.

Because day $t - 1$ will sometimes be a missed observation, we base our adaptive protocol based on the proportions estimated on those observed on the nearest previously sampled
day (likely $t - 2$).

4.5 Discussion

There are several key areas where the ideas of this chapter contribute to improving statistical practice. First, we discuss how simulation models can be used to evaluate performance of an estimation model. We demonstrate the development of a simulation model based on sockeye salmon migration and described how it could generate data based on a variety of different sampling approaches.

Second, we discuss the importance of ignorable sampling approaches in the context of Bayesian estimation. We particularly focus on Bayesian models built for observational data such as that collected in environmental monitoring applications. We advocate for the importance of developing sampling strategies in cooperation with both subject area and statistical experts. Statistics is a discipline that supports and guides the practice of science and application of scientific method in many ways; doing this effectively involves recognizing where science has practical limitations in terms of being able to collect data that would be considered ideal statistically.

Finally, in the context of developing potential adaptive sampling approaches, we believe that we are among the first to propose ignorable adaptive sampling methods when the estimation approach being considered is exclusively Bayesian. Additionally, we believe that the adaptive approaches that we proposed based on sample size requirements for multinominal distributions are the first developed where the adaptation results due to the measures of centre and spread in the underlying distribution being related.

Discussions of sampling approaches to collect data exclusively for use in complex Bayesian estimations seem rare in published literature. Our ability to estimate more and more complex Bayesian models has occurred due to improvements in both computer software and hardware. Improved computation also increases our ability to conduct simulation based research. The work conducted in this chapter provides a useful framework for the way such simulation approaches can be used for developing and creating data based on sampling approaches for Bayesian models.
Chapter 5

Optimality For Spatial And Other Complex Models

Data analysis typically has a goal of allowing a researcher to make inferences about a population or process of interest. Good inference often relies on effective sampling strategies and models that are appropriate for the data. In Chapter 1.3.2 we discussed the idea of optimality in the context of developing sampling strategies that support good inference about a parameter of interest, $\theta$. Typically in sampling literature, $\theta$ is a univariate parameter that we wish to estimate, such as a population mean, $\mu$, or total, $\tau$. Optimal sampling then occurs when the mean square error around $\mu$ or $\tau$ is minimized for a given estimation method. In many such cases, the optimal sampling strategy can be determined analytically.

Sometimes though, the response of interest is not a single parameter, but a collection of parameters that collectively represent an underlying surface that we are interested in estimating. This commonly occurs in the spatial modeling context (Cressie, 1993). Such approaches are very common when collecting and analyzing data collected in relation to environmental monitoring or geostatistical data problems where observations occur across space, time, or both. In Thompson et al (1996), the authors consider the challenge of how to add $n_2$ sampling sites to an existing pattern of air pollution sample sites in an optimal way when the population total $\tau$ is being considered. Under the normal model considered, they determine that an optimal design for the $n_2$ new units will be a conventional one. In an analogous problem, Spöck and Pilz (2010) discuss the introduction of new sample sites for monitoring radiation levels. Based on a Bayesian kriging model, their optimizations are
intended to improve either prediction or covariance estimation. Designs similar to space filling models are found to be optimal.

In both of these cases, the authors rely on only one or two criteria/parameters in order to determine optimality. While MSE or alternate variance criteria such as theirs are important, basing the optimization on only one or two strict criteria fails to consider the myriad of ways that a model might be considered optimal for a given set of observed data. For example, Spöck and Pilz (2010) motivate their work with an example focused on monitoring Cesium-137 concentrations near the Chernobyl accident site. We can imagine multiple scientific reasons for monitoring such radioactive particles: perhaps the desire is to monitor ambient radioactivity across a region (in which case the published approach is appropriate); scientists and policy makers might also be interested in establishing the areas where radioactivity levels transition from safe to unsafe (with an eye to resettlement of the newly safe zones); in a different location (where a meltdown has not occurred), we might want monitoring data to identify when leaks or meltdowns might be occurring to facilitate warnings, evacuations, and the establishment of safe zones; and so on. In addition to such scientific considerations, financial considerations might need to be considered due to the limitations they introduce into a sampling plan.

Rather than determining optimal sampling designs for complex models based on narrow criteria such as blanket variance reduction, we suggest and explore a more general approach that allows multiple criteria for optimality to be considered when comparing sampling (or modeling) strategies. Although a unique ‘optimal’ solution to sampling will (can) not be identified with our approach, we show how it is useful to understand the strengths and the weaknesses among viable proposed sampling approaches. The intent is that such information improves scientific practice, and leads to estimates that are robust for a wide range of possible optimality considerations. This is an important idea, because costly monitoring programs are often developed with multiple scientific goals that ought to be considered when determining an ideal sampling approach. In addition to providing an interesting statistical problem, it is also good for the practice of science that we think critically about which estimates in a complex model should be considered to be the most important.

Because complex models often come in a Bayesian hierarchical form, we illustrate our approach using the CH MVCAR model described in Chapter 2. We further argue that the simulation approach we propose should be included with the idea of unified approaches for
exploratory data analysis and for data analysis of complex models (Gelman, 2003, 2004).

5.1 Justification

In Gelman (2003, 2004) the author argues that exploratory data analysis (EDA) and Bayesian inference (aka complex statistical models) should be considered under a unified framework rather than as unrelated paradigms. He particularly focuses on the close relationship between the posterior predictive aspects of Bayesian inference and EDA. He argues the complementary relationship between EDA and modeling.

1. According to Tukey (1972), a definition of exploratory plots is given as “graphs intended to let us see what may be happening over and above what we have already described.” Even though exploratory analysis is often thought to occur prior to modeling data, this is not entirely true. For example, when we draw a scatterplot of two variables \(x\) (a covariate), and \(y\) (a response), we are already presuming something about the relationship between \(x\) and \(y\). We are assuming that a model exists between \(x\) and \(y\) such that \(x\) can help explain the variability in \(y\), as in a linear regression. If the graph shows such a model to be unreasonable, the model will be abandoned prior to any estimation, but there is still a sense in which our data were being compared to an underlying model. If the model is found to be inadequate, we change or improve it.

2. We often use graphical checks to evaluate how well models fit data. Examples could include reviewing residual plots in the regression case, or drawing graphs based on posterior predictive values (Gelman et al, 1996, 2003, pp.157-192). Such graphs may suggest improvements to the model, or an alternate model in the Bayesian modeling context.

Based on the above, whether a graphical check comes before a model is estimated or after, the graph itself provides information on ways that a model fits well, and where it fits poorly and needs to be improved. We believe that our extension of ideas related to prior predictive checking, and Bayesianly justifiable frequency calculations to include
results from simulated data to provide a reference distribution complements and can be incorporated into a broader unified approach to EDA and goodness of fit testing.

Proposing such a unified approach between EDA, posterior predictive methods, and using reference simulations to evaluate sampling approaches, has several advantages. Goodness of fit testing in the Bayesian context relies on posterior predictive simulations from the proposed model given the observed data. Goodness of fit is evaluated based on discrepancies between the observed data, and replicate data generated by the model. It is also possible to evaluate the potential influence of model mis-specification on estimation of parameters of interest, or the development of sampling procedures that make the specified model more robust. Selection of sample units in the Bayesian modeling context is not as well developed as it is in design based, or alternate model based paradigms. Incorporating reference simulations created based on potential sampling approaches has the potential to increase the effectiveness of statistician’s estimation models and sampling procedures prior to money being spent collecting data. Even relatively simple simulation approaches have the potential to provide valuable information about the effects of sample selection on estimation.

Including assessment of sampling methods based on simulations into a unified approach with EDA and goodness of fit in a Bayesian context allow us to use the richness of ideas in literature about EDA and goodness of fit. This literature suggests how issues like assessing for optimality on a surface when multiple definitions of optimality are possible could be addressed.

Earlier, in Chapter 1.3.1, we defined a discrepancy function for use in evaluating goodness of fit in a Bayesian context using posterior predictive values. In Gelman et al (1996), the authors take the test statistic for a discrepancy, $T(y)$, to be a scalar, and the discrepancy between the replicated and observed data to be summarized by a Bayesian p-value. However, in Gelman (2004) the author suggests that when discrepancies of interest form a vector, $T(y)$, that this can visualized graphically, and the graph itself is the test statistic. Being able to view a graph as a test statistic seems particularly useful in the context of spatial and other complex data models where we have a surface of parameters that we are trying to estimate.

Assuming that we are evaluating the performance of estimations from a statistician’s model to data from a simulation model, we can borrow and slightly modify ideas of dis-
crepancies from those used for posterior predictive checks. Rather than defining a discrepancy based on the difference between observed data and posterior predictive data, we can define a discrepancy between parameters from the simulation model to those estimated by the statistician’s model using the simulated data. $T(y)$ becomes $T(\theta)$, and we can define a discrepancy

$$D(\tilde{\theta}^{sim}, \theta^{sim}, y^{sim}) = T(\tilde{\theta}^{sim}, y^{sim}) - T(\theta^{sim}).$$

(5.1)

Here $\theta^{sim}$ are parameter values taken from our simulation model, and $\tilde{\theta}^{sim}$ are parameter values estimated by our statistician’s model based on generated observations $y^{sim}$. Defining the discrepancies as a difference in this way ensures that they will be compared to a value of zero, often simplifying interpretation of visual model checks (Berkhof et al, 2002). As with the posterior predictive case, discrepancies defined could be in the form of a vector, where a graph could be used as a test statistic.

A key to making the idea of graphing discrepancies based on simulated data generated from a reference model work is by creating graphics that visualize the results well. Kerman et al (2008) discuss graphics in a Bayesian context, particularly as they relate to posterior predictive checks and the visualization of discrepancies between simulated and observed data. The ideas expressed are similar to those found in Gelman (2003, 2004). They include the concept that graphs capture subtle patterns in data not seen in scalar summaries. Further, such misfits seen in graphs can also actually suggest useful scalar discrepancies. The value of graphs in representing even scalar summaries of results and data is discussed in Gelman et al (2002). Here the authors change tables from an issue of JASA into effective graphs and provide guidelines for doing so.

In developing ideas using a reference distribution for evaluating a complex statistician’s model for estimation, we hope to capitalize on the use of effective graphics for the presentation of information.
Table 5.1: Settings used to generate proportion data for simulation study

<table>
<thead>
<tr>
<th>Stock</th>
<th>Runlength</th>
<th>Magnitude</th>
<th>Offset Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>2.5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>37</td>
<td>3</td>
<td>15</td>
</tr>
</tbody>
</table>

5.2 Example

5.2.1 Data Simulation

In Chapter 4 we discussed a model for data generation. We also developed six sampling methods which were ignorable for the Bayesian model being used for inference. The data generation model was used to simulate data for each sampling strategy. Here we assess the performance of the different sampling methods by conducting a simulation study. Fifty population realizations of proportion data are created, and the data are analyzed using the Compositional Hierarchical Multivariate Conditional Auto-Regressive (CH MVCAR) model described in Chapter 2. Our goal with the simulated data is to mimic features of the data that were collected during the 2006 migration and fishing season for Fraser River sockeye salmon. To simplify the analysis of simulated data, we assume the presence of only three stock groups rather than five.

For the simulated data sets, we assumed that the observation period for sampling the fisheries occurred over 47 days. This includes days where sampling did not occur in some methods. We based our simulations based on the runlengths, relative stock magnitudes, and time offsets shown in Table 5.1.

Based on these, the appropriate normal quantiles are calculated, and ‘true’ proportions without added error are generated according to Equations 4.1 through 4.3. These settings result in the data used to draw Figure 4.1. The above runlengths and offset times result in 54 days of ‘observations,’ but we only considered that sampling occurred in 47 of them. This reflects that few samples are taken at the beginning or end of the migration season when the commercial fisheries are never opened even though fish are in the river.

Using the underlying proportion vectors for each day $\Pi_t = (\pi_{1t}, \pi_{2t}, \pi_{3t})$, we simulate
50 different population realizations assuming Dirichlet process variation

\[ P_t = (p_{1t}, p_{2t}, p_{3t}) \sim \text{Dirichlet}(\gamma(\pi_{1t}, \pi_{2t}, \pi_{3t})) \quad t=1,...,47 . \]

We select the value \( \gamma = 20 \) which introduces a realistic amount of process variation. Finally, sampling variation is introduced by taking multinomial draws from \( P_t \), giving us observed stock counts,

\[ y_t = (y_{1t}, y_{2t}, y_{3t}) \sim \text{Multinomial}(n_t, (p_{1t}, p_{2t}, p_{3t})) . \]

Here, \( n_t \) is decided based on the sampling method being evaluated. For the non-adaptive method, \( n_t = 50 \) on days that were sampled, and 0 (NA) otherwise. For adaptive methods, \( n_t \) could be 30, 50, 70, or 100 on observed days depending on the sampling protocol that was being considered. It should be noted that in methods allowing missed days of sampling, the rate of days missed is taken to be approximately one in seven, or once a week. This is consistent with the number of missed days in the actual data.

These \( y_t \) are then used as data to generate MCMC estimates of \( \hat{P}_t \) and other parameters using WinBUGS (Lunn et al, 2000). This means that there were 50 different MCMC runs for each of the 6 sampling methods under consideration. Prior settings and constants for the Bayesian estimation are the same as discussed in Chapter 2, except that here we are considering \( D = 3 \) stocks rather than \( D = 5 \) stocks. Each chain was run for 100,000 iterations, and the first 30,000 iterations were discarded as burn-in. The remaining iterations were thinned by a factor of 100, giving 700 draws on which to base posterior estimates. This thinning factor was chosen as a very conservative choice (ie. too much thinning), because it was impossible to examine all traceplots for all simulations individually. The goal is for too much thinning to protect against autocorrelation should any of the simulations generate chains with poor mixing properties.

### 5.2.2 Discrepancy Results

Earlier we argued, particularly for complex models, that it is difficult to define a single scalar criteria to be optimized in order to evaluate the quality of estimations provided by a model using a given sampling method. We discussed the idea of discrepancy functions, and that in the case of vector valued responses that a graph could be viewed as a discrepancy
function. This is particularly true if there is an implicit or explicit model being evaluated using the graph. To successfully convey discrepancies graphically, we emphasized the importance of developing graphs that are effective. In evaluating the sampling methods that are being compared based on our simulation results, we will incorporate such ideas.

The main purpose of our Bayesian statistician’s model in the context of our motivating data is to estimate relative stock proportions while sockeye salmon are migrating upriver to spawn. Although our Bayesian statistician’s model also estimates other potential parameters of interest given the data, we are basing our discrepancy measures on estimated proportions from the statistician’s model to theoretic proportions from the simulation model. We consider the whole surface of proportion estimates as a basis for developing discrepancy measure. We demonstrate the discrepancy as a vector of results that can be visualized graphically.

As with our statistician’s model, our simulation model is hierarchical, and proportions are considered at two levels in this hierarchy. We have a level of proportions in the hierarchy which include Dirichlet process error that we will refer to as process proportions, \( P \). There are 50 different realizations of process proportions against which our estimated proportions based on the statistician’s model \( \tilde{P} \) can be compared.

Moving up a level in our hierarchy, we have underlying or ‘true’ proportions, \( \Pi \), based on an assumption of a normal abundance through time migration pattern. We are terming these \( \Pi \) the underlying trend proportions for the purposes of discussion. There is only one normal model used to generate simulated data. The normal parameters could also be drawn from distributions, but for purposes of demonstration, comparison, and discussion we limited ourselves to one.

In comparing estimated proportions, \( \tilde{P} \), to process proportions, \( P \), we are able to gain an understanding of how well the statistician’s model, in combination with a sampling approach, is able to identify changes in proportions that are analogous to responses that salmon might have to events in the river within a given migration season. For example, stochastic events such as rainy days are thought to trigger influxes of salmon into the river relative to dryer, warmer days.

Contrasting this evaluation of within year proportion trends, we compare estimated proportions, \( \tilde{P} \), to trend proportions, \( \Pi \). This allows us insight into a slightly different problem of interest. Deviations from the trend proportions tell us about run timings in
terms of overall length and magnitudes. Since run timings and magnitudes can change from year to year (and might be changing due to factors such as climate change), these can inform us about how well our model will be able to identify differences in migration trends between migration seasons. In a sense, a superpopulation process is generating different yearly migration patterns. Although between migration season inferences have not been made yet in situ, understanding such changes is part of the long term strategy of the sockeye salmon monitoring program.

A key goal when evaluating discrepancies at both of these levels in the simulation hierarchy is to learn what effects different sampling measures have on the validity of the estimates calculated by the statistician’s model. They also nicely demonstrate how a definition of optimality can be complex; not only do we need to consider that our surface can be divided into sections of greater or less interest, we also need to consider how well that surface compares to conceptually different parts of the underlying processes.

**Process Proportion Results**

Below in Figure 5.1, we show boxplots of the discrepancies between proportions estimated by the statistician’s model, \( \hat{P} \), and the process proportions, \( P \) for the six sampling methods being considered. In Figures 5.1 and 5.3, the 6 different sampling methods are identified by the following abbreviations:

- Adapt - adaptive sampling based on previously observed proportions with no missed sampling days;
- AdaptRand - adaptive sampling based on previously observed proportions where sampling days are missed at random;
- SimpWeek - a balanced approach where 50 fish are collected on days when sampling occurs, missed days occur systematically on a ‘weekend’;
- SimpRan - a balanced approach where 50 fish are collected on days when sampling occurs, and missed sampling days occur at random;
- Simple - a balanced approach where 50 fish are collected daily, and no missing days occur;
• SimDoub - the ‘real life’ approach, where missed sampling days are assumed to occur at random, and the target sample size double from 50 to 100 after a missed day.

All three of the graphs in Figure 5.1 share the same axes, and this is true for all of the Figures 5.1 through 5.4. As suggested in Berkhof et al (2002), we compare our discrepancies to a reference value of zero in all graphs, similar to how residuals are defined. In Figure 5.1, this means that if \( \bar{P} \) perfectly matched \( P \) all of the values would occur at zero. In this figure, we consider only the discrepancies between sampling methods, aggregating the discrepancies across the 47 different observation days. In all three of the boxplots shown in Figure 5.1, the median values occur very close to zero for all of the sampling methods. However, there are marked differences between the sampling methods when we start to consider the tails of the boxplots. The distributions of discrepancies for sampling methods where no days of sampling are missed (Simple, Adapt) have markedly shorter tails than distributions for any of the other sampling approaches. The method, equal sample sizes with missing days at random, appears to result in a distribution of discrepancies with the longest tails. When days are missed in a systematic fashion in SimWeek, tails seem somewhat shorter. Tails are also shorter when random missing days are paired with an adaptive sampling method.

The boxplots in Figure 5.1 provide information about differences between sampling methods as revealed by discrepancies between simulated and estimated parameters. But boxplots aggregate results across the temporal surface we are truly interested in estimating. We illustrate discrepancies across the entire time surface in each of the stocks in Figure 5.2. Again, we use a reference value of zero in the graph, but here each line across time represents the discrepancies between the estimated model and simulated model for one simulation. To increase the readability of the graph, we included the results for only 15 simulations because this provided a balance between having enough data to demonstrate features of interest, but not so much that the results became even scribbled and therefore unreadable.

Generally speaking, Figure 5.2 confirms the conclusions that were made about the different sampling methods based on Figure 5.1; sampling methods which do not allow for missing days perform ‘best’ in terms of having the smallest outlying discrepancies. We do gain important insight into why this is occurring when we consider the discrepancies that
Figure 5.1: Boxplots of discrepancies between $P$ and $\bar{P}$ for 3 stocks and 6 sampling methods.
Figure 5.2: Lineplots of discrepancies between $P$ and $\hat{P}$ for 6 sampling methods demonstrated using Stock 2.
are created in individual simulations. The ‘bad’, or large discrepancy values are not created by entire runs of poorly estimated values, but instead on individual days in individual stocks where the estimated values $\tilde{P}$ are not capturing the process variation in $P$. The fact that the largest discrepancy values are bigger when sampling methods include missing days is due to the fact that the model cannot adjust for days it cannot ‘see’ which have large process variation. Sampling methods with days missing at random performed worst because the several missed days could occur in a row due to randomness, increasing potential discrepancies. The systematic sampling method with days missing on weekends protects against multiple missed days in a row, and therefore, has better overall performance.
Trend Proportion Results

We are interested in discrepancies between estimated values at two different levels of the simulation hierarchy. To capture discrepancies at the second level of possible discrepancy, Figures 5.3 and 5.4 are analogous to Figures 5.1 and 5.2, except that they are based on the trend proportions $\Pi$, of the simulation model and contain no process error. Despite similarities in the graphing approach, the inferences that we gain from the graphs about the sampling approaches being evaluated are distinct.

In the earlier graphs there were clear differences in the discrepancies between the different sampling methods; in the current graphs there are no obvious differences between the sampling methods. Although our goal was to examine discrepancies at the trend (superpopulation) level of the simulation hierarchy due to differences in sampling strategy, the result we have gained is that sampling strategy is not particularly important to these estimates. Such a result suggests that we focus on sampling strategies based on conclusions from discrepancy results from the process level of our simulation hierarchy.
Figure 5.3: Boxplots of discrepancies between $\Pi$ and $\bar{P}$ for 3 stocks and 6 sampling methods.
Figure 5.4: Lineplots of discrepancies between $P_i$ and $\hat{P}$ for 6 sampling methods demonstrated using Stock 2.
5.3 Discussion

In our simulation example, we evaluated discrepancies based on comparison to both a trend, and a process error criteria. This demonstrates that in addition to being able to evaluate an entire surface, we can also choose discrepancies based on different levels of the simulation model. Based on these two levels, we discovered only one of the the process comparison results were sensitive to the ignorable sampling methods being evaluated.

Based on Figures 5.1 and 5.2, we were able to determine that methods that did not have missing days had smaller tail discrepancies than those that did. Methods where missing days could be adjacent tended to have the largest tail discrepancies. Ideally speaking, these results suggest that sampling should occur without missing days; practically speaking, it is unlikely that this can be achieved. As an alternative, we could use these results to create a sampling strategy that is adaptive based on the likelihood of opening the commercial fishery. Commercial fishing is most likely to occur in the middle days of the sampling period because abundances are higher and runs are not contaminated with other salmon species. We suggest that sampling allowing for missed days can be used at the beginning and end of the sampling period; in the middle of the run, greater effort should be made not to miss scheduled sampling. Such a sampling scheme would optimize for estimates of process proportions, while allowing for the cost and personnel savings accrued on missed sampling days. The results of Figures 5.3 and 5.4 indicate that allowing these days would not affect our inference about trends between migration seasons.

In this chapter we have introduced ideas for assessing optimality when the estimates of interest occur across a surface. We particularly focus on how to do this using simulation study in a Bayesian modeling context. This approach was developed with the intent to support the development and the evaluation of ignorable sampling methods for applications where inference from data being collected will be taken across a surface of interest. Such data is often collected in monitoring programs measured through space or time.

We emphasize how simulation studies such as those described can be used and interpreted in a Bayesianly justifiable way by adapting ideas from Box (1980); Rubin (1984). We also recognize that such simulation studies are already commonly used to evaluate Bayesian models are used in many areas of application (Bonner et al, 2010; Juarez-Colunga, 2011; Feng, 2011). We feel it is an important contribution to make the philosophic con-
nection between our application of Bayesianly justifiable simulations, and simulation approaches such as prior and posterior predictive checks. By placing our simulations in a context with other Bayesianly appropriate simulation approaches, we are able to borrow ideas from this literature and apply them here. For our example, ideas from posterior predictive literature and exploratory data analysis allowed us to discuss discrepancies across a surface, and optimality in terms of a vector of responses rather than only as a scalar quantity (Gelman, 2004, 2003). Working with the premise of a unified approach between exploratory data analysis and Bayesian simulation is a valuable tool for developing better sampling approaches for Bayesian models, and thus better inference for Bayesian models.
Chapter 6

Conclusions

Research topics presented in this thesis have all been motivated by a data monitoring program providing information about sockeye salmon stock run timings in the Fraser River. As an economically and socially important ecological system on the West Coast of North America, the problems presented, while important, might seem to be specific in their geographic and social scope. In addition to stakeholders involved with the salmon sampling program, we found that our solutions had general implications for statistical inference and methods beyond this motivating problem. Working closely with this monitoring data has also given us perspectives on the importance of understanding the scientific goals of a research program. Specific scientific goals are not always clear when multiple stakeholders and interest groups are involved with managing a single program. Explicitly discussing the inference desired by different groups is key to achieving statistical solutions that help to support multiple goals in scientific applications.

Presented in Chapter 2, our research began with a focus on creating a Bayesian hierarchical model for inference using the motivating data. The largest practical problem that needed to be addressed using this model was helping to account for variability in the multinomial data. Our multinomial data were collected in order to model relative proportions of different sockeye salmon stock groups. The variability in the data were caused by sampling variability due to small sample sizes; by underlying trends influencing relative stock proportions; missing data; etc. Our goal was the separation of undesirable ‘noise’ variability from ‘signal’ variability which could be used to provide inference about the dynamics of salmon migration. This was achieved by developing a Bayesian CH MVCAR model that in-
corporated prior information about beliefs that observed proportions in adjacent days were correlated; that stocks themselves could be positively or negatively correlated; and that sharing collected information between adjacent days could shrink uncertainty estimates. The model did not specifically address theoretical bell shaped stock abundance curves as they relate to observed sample proportions. Implicitly this information was incorporated by generally modeling temporal correlations. Our temporally correlated model fit well according to posterior predictive goodness-of-fit evaluations, and achieved substantial reductions in interval widths were compared to modeling days independently. The model specification chosen has a general format that means it is be appropriate for use in other applications where spatio-temporally correlated multinomial data occurs.

Some areas identified for future research involving the CH MVCAR model included extending the model to incorporate covariates; including data from more than one sampling location; and improving prior specification of the variance-covariance parameters at the MVCAR level of the hierarchy. Ultimately though, we continued our research motivated by fisheries data by concentrating on sampling approaches for data collection. This includes results for both sample size calculations, and for ignorable sample design. We identified sampling as an area of particular interest, and sampling for complex Bayesian models is also an area of statistical literature that is relatively underserved.

Chapter 3 is addressed to solving sample size requirements for binomial and multinomial problems. Daily observations taken in the Fraser River system are multinomial, and such data are relatively common. While sample size calculations are an important part of study planning, methods and advice for sample size determination were found to be somewhat inconsistent in the binomial and multinomial cases. Sample size methods based on normal approximations are common, but intervals based on this assumption often do not meet stated levels of coverage. This is especially true when some proportion parameters are near zero or one. Approximation methods also fail to account for irregularities in sample size requirements caused by the discreteness of the underlying binomial or multinomial distribution, making it difficult to identify true worst case parameter settings in terms of sample size requirements. To address such problems, we applied an exact method for binomial sample size determination involving highest density probability (HD) criteria for determining sample size. The exact HD criteria was extended to the multinomial case. We comprehensively examined sample size requirements for a variety of interval widths and
levels of significance in both binomial and multinomial cases. These sample size results are used to discuss worst case parameter scenarios. We see the future of this work as being as an extremely useful tool for study planning for scientific research involving binomial or multinomial observations. It is particularly valuable when it is known that sample sizes will be small, or when any proportions are suspected of being near to 0 or 1. We used it as a basis for developing adaptive sampling strategies for the sockeye salmon data where both small sample sizes and proportions near to 0 or 1 occur.

One of the challenges in determining a sampling design for a Bayesian model is that in Bayesian problems the data are considered fixed. This often leads to behaviour where Bayesian models are applied but sample collection is not explicitly considered or modeled, even when necessary. We can only ignore sampling strategy in a Bayesian context when the data being modeled is collected based on an ignorable sampling design in the context of the estimation model. Criteria for ignorability are intuitive, and in Chapter 4 we provide some guidelines for developing ignorable sampling methods. We apply these criteria in the context of our example. The sampling methods we develop acknowledge the challenges of limited resources for sampling, and focus on practical sampling choices. It would be interesting future work to examine sampling costs more explicitly for the different approaches considered.

A second task in Chapter 4 was to discuss Bayesianly appropriate simulation approaches, and how they could be used to mimic data that could potentially be collected by different sampling methods. We then build a hierarchical simulation method that can be used to ‘sample’ data based on different ignorable strategies. Such a simulation approach gives us reference distributions to which we can compare estimation results from complex models. In our example, the complex model used for estimation was the CH MVCAR model.

With the use of Bayesian models becoming more common, having a sample development and simulation framework provides many opportunities for future research. Data collection for many existing models needs to be evaluated as to whether or not it is ignorable. Where ignorability cannot be achieved improvements should be made to models to account for sampling; where sampling is ignorable, different approaches can be compared with a goal of making model estimates more robust.

To evaluate sampling methods for our own model, we encountered the problem of how optimality should be evaluated when the parameters of interest occur across a surface rather
than as a scalar. The solution we propose in Chapter 5 is to unite our idea of using simulation to compare sampling methods with ideas from other Bayesian simulation approaches. We particularly focus on posterior predictive approaches that can be based on discrepancy functions where the discrepancy measure can be either a scalar or a vector. In the vector case, optimality can then be considered based on properties of graphics drawn to represent the discrepancies. Effective communication of information through the use of graphs is a very important consideration, particularly when evaluating complex models.

Using our CH MVCAR model from Chapter 2, and the simulation approach outlined in Chapter 4, we defined two different levels of discrepancies with which to evaluate sampling methods. We then demonstrate the use of graphs to convey information about the effects of sampling over the temporal surface being estimated. This example demonstrates ideas about the use of vector valued discrepancies and visualizing them graphically. It also provides some suggestions and information about potential sampling strategies for the sockeye salmon fishery. Although our adaptive approaches did not work better than conventional alternatives, we continue to be interested in exploring the effectiveness of adaptive sampling approaches for Bayesian inference. With our own approach if the adaptive rule were tweaked, better results might be achieved.

Spatial or temporal data where such ideas about sampling, simulation, and graphics can be applied are relatively common. Although the information conveyed in a graphic can be quite sophisticated and nuanced, our simulation approach for creating graphics to evaluate sampling methods is straightforward to implement. Applications of this easy to apply method for evaluation of sampling strategies will improve sampling strategies for, and ultimately robustness of, Bayesian models.

Because of our focus on graphical displays in Chapter 5, the ways such media are used to convey information in statistics is also an area of interest for future developments and discussion. If video games from the 1970’s and 1980’s are compared to those played today, it obvious that the visuals are much better and more realistic now. Comparing graphs in statistical journals from the 1980’s to those now, the advances are not as obvious. While our ability to estimate complex Bayesian models has increased, our methods of conveying information about such models has remained somewhat static. Statistics is a data to information science. Growing the use of graphics and other electronic media as ways of effectively conveying information about data is an important element of statistical research.
Our work is motivated by a specific scientific problem. The solutions to that problem have resulted in statistical approaches and discussion that can and should be applied to a broader class of problems. As much as statistics is a discipline that supports applied scientific research, working with applied sciences themselves are a rich basis for advancing new statistical methods and ideas such as we did here.
Bibliography


Pearson K (1896-1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation that may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London 60:389–498


Tukey J (1972) Some graphic and semigraphic displays. Statistical papers in honor of George W Snedecor 293:316


Appendix A

R/WinBUGS Code For Running CH MVCAR Model

#This is the R and WinBUGS code needed to run the model
#First create sample data from a multinomial distribution
#similar to the actual data in structure. The generated
#below are very basic and do not generally reflect the complexity
#of real world data
#libraries needed for rmultin to work
library(lattice)
library(Matrix)
library(foreign)
library(deldir)
library(sp)
library(maptools)
library(tripack)
library(boot)
library(spdep)
library(DCluster)
#some libraries that are needed to call WinBUGS from R and analyze the outputs
#(used a bit later in the code)
library(lattice)
library(coda)
library(R2WinBUGS)

#Below takes samples of size 50 from 5 different 'stocks'
makeSample1=function(replicates)
  sample1=rmultin(50,c(10/50,1/5,1/5,1/5,10/50))
  for(i in 1:replicates)
    sample1=rbind(sample1, rmultin(50,c(10/50,1/5,1/5,1/5,10/50)))
  #ends i for
  return(sample1)
#ends makeSample1

#There are 49 days of sampling in the sockeye data
#so we will do the same for the simulation data
Y=makeSample1(49)

#If you want to simulate days where no samples are taken, just change all the
#values in a row to zero

#Based on the data set that you are using (If yours has different dimensions)
nrow=NROW(Y) nstock=NCOL(Y)

#The prior on the inverse variance is Wishart. We recommend using a diagonal matrix
#and being careful to conduct sensitivity checks when using this prior
R=diag(x=0.1,nrow=nstock,ncol=nstock)

#The data items below are needed to make WinBUGS run at all
dataList=list("nrow","Y","nstock","R")

#These are output variables that you want WinBUGS to save information on
#so that they can be used to summarize results. Elements in this list can
#be taken off, or alternatively variables not mentioned here could be included
parameters.to.save=list("prob","sigma1","sigma2","sigma3","sigma4","sigma5","corr12",
  "corr13","corr14","corr15","corr23","corr24","corr25","corr34","corr35","corr45")

#These are starting values for the MCMC chains. WinBUGS can automatically generate

#starting values, but it tends to do a poor job when it comes to this model. Therefore
#I have made a list of 'reasonable' starting values to facilitate convergence and
#prevent the model from crashing due to inappropriate values..
#Note that this sets up initial values for 1 chain at a time. To perform multiple
#chains on the same run this would have to be modified slightly which is not
#difficult
init.vals=function()
#assume that estimates for all time points are at the mean
#value for that stock..
initials=rep(0,245)
init.move=matrix(data=initials,nrow=NCOL(Y),ncol=NROW(Y))
#These are the starting values for the ’mean’ of each stock
#Here I a assuming each model has the same underlying average
#proportion (which is not true but works okay anyways)
intercept=c(log(50),log(50),log(50),log(50),log(50))
list(intercept=intercept,move=init.move)
#ends init.vals
#directory where WinBUGS is installed. This needs to be adjusted
#as per your own computer
bugs.directory = "c:/Program Files/WinBUGS14/"
#Place where you want WinBUGS output to save to. Change
#as per your own preferences
SaveFile="c:/Documents and Settings/MyDocuments/results"
#R command to get the model to run in WinBUGS
modelrun=bugs(
datalist,
inits=init.vals,
parameters.to.save,
model.file="c:/Documents and Settings/MyDocuments/currentmodel.txt",
n.chains=1,
n.iter=1000000, n.burnin=100000,
bugs.directory=bugs.directory,
bugs.seed=6,
working.directory=SaveFile,
debug=TRUE,
n.thin=100)
#currentmodel.txt (The WinBUGS code for the model) created below....
#The sink('filepath') statement is just opening a file and
#writing the text underneath to the designated pilepath. You can also edit the WinBUGs
code
#in a text editor in a separate window, but this just keeps things all nice and together.
sink("c:/Documents and Settings/MyDocuments/currentmodel.txt")
cat(""
model
#the likelihood bit
for(i in 1:nrow)
#note that 1:5 is finding columns containing the proportions
#for different stocks,
nfish[i]<-sum(Y[i,]) #or I could put in a vector/column
Y[i,1:5] dmulti(prob[i,],nfish[i])
for(j in 1:nstock)
prob[i,j]<-delta[i,j]/sum(delta[i,])
#ends j for
#ends i for
#now I am going to set up the adjacency and weights stuff for mv.car
#This would have to be adjusted if you want to define your neighbourhood
differently (ie. spatially vs. temporally)
#special case at the start of chain
for (i in 1:1)
weights[i]<-1
adj[i]<-i+1
num[i]<-i
#ends 1:1 for
#main body of weights and stuff
for (i in 2:(nrow-1))
weights[2+(i-2)*2]<-1
adj[2+(i-2)*2]<-i-1
weights[3+(i-2)*2]<-1
adj[3+(i-2)*2]<-i+1
num[i]<-2
#ends main body for
#special case at end of chain
for (i in nrow:nrow)
weights[(nrow-2)*2+2]<-1
adj[(nrow-2)*2+2]<-i-1
num[i]<-1
# ends nrow:nrow for
#Setting up the Dirichlet prior as a function of gammas
for(i in 1:nrow)
for(j in 1:nstock)
delta[i,j] dgamma(alpha[i,j],1)
#ends second j for
#Want to incorporate a scale parameter to determine how much weight
#in an estimated probability comes from the observed data, and
#how much is coming from the smooth (MVCAR)
#need to calculate alpha as fn of scale*param
for(j in 1:nstock)
alPHA[i,j]<-scale*Comp[i,j]
#ends j
#gives the base parameter (on the simplex)
for(j in 1:nstock)
Comp[i,j]<-CAR[i,j]/sum(CAR[i,])
#ends j for
#needs to be set up in vector form for car.mv.... alphalpha1[i]<-intercept[1]+move[1,i]  
#stock specific weighting for time period i.. etc.
alphalpha2[i]<-intercept[2]+move[2,i]  # ditto I need to make this into my equation with an
alphalpha3[i]<-intercept[3]+move[3,i]  # ditto intercept that has a dflat() prior, and then
CAR[i,1]<-exp(alphalpha1[i])
CAR[i,2]<-exp(alphalpha2[i])
CAR[i,3]<-exp(alphalpha3[i])
CAR[i,4]<-exp(alphalpha4[i])
CAR[i,5]<-exp(alphalpha5[i])
#ends second i for
#This 'scale' parameter is described as gamma in the paper
scale<-10000
#this is setting up the mv.car prior...
move[1:nstock,1:nrow] mv.car(adj[],weights[],num[],omega[,])
#some priors stuff
#degrees of freedom for the wishart prior
dfparam<-nstock+1
omega[1:nstock,1:nstock] dwish(R[,,],dfparam)
for(i in 1:nstock)
intercept[i] dflat()
#ends intercept for
#extracting this to look at my var-cov stuff
sigma[1:nstock,1:nstock]<-inverse(omega[,])
sigma1<-sqrt(sigma[1,1])
sigma2<-sqrt(sigma[2,2])
sigma3<-sqrt(sigma[3,3])
sigma4<-sqrt(sigma[4,4])
sigma5<-sqrt(sigma[5,5])
corr12<-sigma[1,2]/(sigma1*sigma2)
corr13<-sigma[1,3]/(sigma1*sigma3)
corr14<-sigma[1,4]/(sigma1*sigma4)
corr15<-sigma[1,5]/(sigma1*sigma5)
corr23<-sigma[2,3]/(sigma2*sigma3)
corr24<-sigma[2,4]/(sigma2*sigma4)
corr25<-sigma[2,5]/(sigma2*sigma5)
corr34<-sigma[3,4]/(sigma3*sigma4)
corr35<-sigma[3,5]/(sigma3*sigma5)
corr45<-sigma[4,5]/(sigma4*sigma5)
#sample odds and end calculations showing
#how to define parameters of interest
#to potentially use in analysis
ratio2<-intercept[2]/intercept[1]
ratio3<-intercept[3]/intercept[1]
ratio4<-intercept[4]/intercept[1]
ratio5<-intercept[5]/intercept[1]
#and now to convert the ratios back to the original scale
#rather than looking at them in the log scale
Ratio2<-exp(intercept[2])/exp(intercept[1])
Ratio3<-exp(intercept[3])/exp(intercept[1])
Ratio4<-exp(intercept[4])/exp(intercept[1])
Ratio5<-exp(intercept[5])/exp(intercept[1])
#ends model ") sink()
Appendix B

Bayesian P-Values For Focused Tests

Tables of posterior predictive p-values for focused tests based on 1000 posterior predictive draws.
<table>
<thead>
<tr>
<th>day</th>
<th>Stock 1</th>
<th>Stock 2</th>
<th>Stock 3</th>
<th>Stock 4</th>
<th>Stock 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.197</td>
<td>0.483</td>
<td>0.67</td>
<td>1</td>
<td>0.685</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.646</td>
<td>0.225</td>
<td>1</td>
<td>0.511</td>
</tr>
<tr>
<td>4</td>
<td>0.291</td>
<td>0.578</td>
<td>0.621</td>
<td>1</td>
<td>0.34</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.392</td>
<td>0.688</td>
<td>1</td>
<td>0.468</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.487</td>
<td>0.349</td>
<td>1</td>
<td>0.372</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.238</td>
<td>0.708</td>
<td>1</td>
<td>0.581</td>
</tr>
<tr>
<td>10</td>
<td>0.206</td>
<td>0.661</td>
<td>0.628</td>
<td>1</td>
<td>0.261</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.663</td>
<td>0.142</td>
<td>1</td>
<td>0.741</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.365</td>
<td>0.683</td>
<td>1</td>
<td>0.531</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.676</td>
<td>0.287</td>
<td>1</td>
<td>0.414</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.647</td>
<td>0.533</td>
<td>1</td>
<td>0.428</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>0.588</td>
<td>0.199</td>
<td>1</td>
<td>0.737</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>0.321</td>
<td>0.836</td>
<td>1</td>
<td>0.33</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0.698</td>
<td>0.233</td>
<td>1</td>
<td>0.66</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>0.339</td>
<td>0.622</td>
<td>1</td>
<td>0.418</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0.486</td>
<td>0.473</td>
<td>0.272</td>
<td>0.509</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0.346</td>
<td>0.759</td>
<td>1</td>
<td>0.45</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0.937</td>
<td>0.186</td>
<td>0.338</td>
<td>0.579</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>0.378</td>
<td>0.723</td>
<td>1</td>
<td>0.467</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>0.237</td>
<td>0.828</td>
<td>0.457</td>
<td>0.548</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>0.602</td>
<td>0.28</td>
<td>0.333</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td></td>
<td>0.598</td>
<td>0.208</td>
<td>0.34</td>
</tr>
<tr>
<td>26</td>
<td></td>
<td></td>
<td>0.748</td>
<td>0.0833</td>
<td>0.603</td>
</tr>
<tr>
<td>28</td>
<td></td>
<td></td>
<td>0.607</td>
<td>0.18</td>
<td>0.35</td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td>0.34</td>
<td>0.24</td>
<td>0.57</td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td>0.751</td>
<td>0.1</td>
<td>0.403</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td>0.208</td>
<td>0.24</td>
<td>0.752</td>
</tr>
<tr>
<td>32</td>
<td></td>
<td></td>
<td>0.872</td>
<td>0.0625</td>
<td>0.57</td>
</tr>
<tr>
<td>33</td>
<td></td>
<td></td>
<td>0.333</td>
<td>0.28</td>
<td>0.323</td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td>0.367</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td>0.456</td>
<td>0.0233</td>
<td>0.098</td>
</tr>
<tr>
<td>37</td>
<td></td>
<td></td>
<td>0.343</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>38</td>
<td></td>
<td></td>
<td>0.367</td>
<td>0.04</td>
<td>1</td>
</tr>
<tr>
<td>39</td>
<td></td>
<td></td>
<td>0.456</td>
<td>0.0233</td>
<td>0.098</td>
</tr>
<tr>
<td>40</td>
<td></td>
<td></td>
<td>0.182</td>
<td>0.1</td>
<td>0.545</td>
</tr>
<tr>
<td>42</td>
<td></td>
<td></td>
<td>0.383</td>
<td>0.0408</td>
<td>0.582</td>
</tr>
<tr>
<td>43</td>
<td></td>
<td></td>
<td>0.383</td>
<td>0.0408</td>
<td>0.582</td>
</tr>
<tr>
<td>44</td>
<td></td>
<td></td>
<td>0.182</td>
<td>0.1</td>
<td>0.545</td>
</tr>
<tr>
<td>45</td>
<td></td>
<td></td>
<td>0.242</td>
<td>0.0204</td>
<td>0.277</td>
</tr>
<tr>
<td>48</td>
<td></td>
<td></td>
<td>0.392</td>
<td>0.0408</td>
<td>0.169</td>
</tr>
</tbody>
</table>

APPENDIX B. BAYESIAN P-VALUES FOR FOCUSED TESTS
Appendix C

Binomial and Multinomial Sample Sizes

Some binomial distribution results.
  Some 3 component multinomial distribution results.
  Some tables of results for higher dimensional simplex results.
APPENDIX C. BINOMIAL AND MULTINOMIAL SAMPLE SIZES

Figure C.1: Exact highest probability sample sizes required for binomial distributions when $d=0.2$, values of $\theta$ are between 0 and 0.5, and $\alpha$ values are 0.4, 0.1, 0.05, 0.01, 0.001.
APPENDIX C. BINOMIAL AND MULTINOMIAL SAMPLE SIZES

Exact highest probability sample sizes required for binomial distributions when $d=0.1$, values of $\theta$ are between 0 and 0.5, and $\alpha$ values are 0.4, 0.1, 0.05, 0.01, 0.001.
Figure C.2: Exact highest probability sample sizes required for binomial distributions when d=0.05, values of \( \theta \) are between 0 and 0.5, and \( \alpha \) values are 0.4, 0.1, 0.05, 0.01, 0.001.
Figure C.3: Exact highest probability sample sizes required for binomial distributions when \( d=0.01 \), values of \( \theta \) are between 0 and 0.5, and \( \alpha \) values are 0.4, 0.1, 0.05, 0.01, 0.001.
Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.2$ at $\alpha$ values of 0.4, 0.1, 0.05, and 0.01 (corresponding to relatively small sample sizes). Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
Figure C.4: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.1$ at $\alpha = 0.4, 0.1, 0.05,$ and $0.01$. Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
Figure C.5: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.05$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01. Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
Figure C.6: Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Only largest 10% of sample size requirements illustrated. Graphs shown when $d = 0.05$ at $\alpha = 0.4, 0.1, 0.05,$ and $0.01$. Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, 0.1, 0.05, and 0.01 (corresponding to relatively large sample sizes). Largest sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
APPENDIX C. BINOMIAL AND MULTINOMIAL SAMPLE SIZES

Half–interval Width 0.01; $\alpha = 0.4$; Area of Max SS

Half–interval Width 0.01; $\alpha = 0.1$; Area of Max SS

Half–interval Width 0.01; $\alpha = 0.05$; Area of Max SS

Half–interval Width 0.01; $\alpha = 0.01$; Area of Max SS

Exact multinomial sample size requirements displayed on ternary graphs for the three simplex case. Only largest 10% of sample size requirements illustrated. Graphs shown when $d = 0.01$ at $\alpha = 0.4$, $0.1$, $0.05$, and $0.01$ (corresponding to relatively large sample sizes). Largest required sample sizes are shown in violet hues, smallest sample sizes are shown in tan/yellow hues.
APPENDIX C. BINOMIAL AND MULTINOMIAL SAMPLE SIZES

Normal approximation multinomial sample size results when the half interval width (d) is set to 0.05 across all simplexes. We identify $m$ where the worst case sample size results are occurring. Based on interval definitions from Thompson (1987).

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$n$ with $d = 0.05$</th>
<th>$m$ (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>203</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>403</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>510</td>
<td>3</td>
</tr>
<tr>
<td>0.01</td>
<td>788</td>
<td>2</td>
</tr>
</tbody>
</table>

Table C.1: Exact multinomial sample size results when the half interval width (d) is set to 0.2 across all simplexes evaluated. We identify $m$ near where the worst case sample size results are occurring.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$n$ with $d = 0.2$</th>
<th>$m$ (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>10</td>
<td>4 or 3 (not clear)</td>
</tr>
<tr>
<td>0.1</td>
<td>20</td>
<td>(not clear)</td>
</tr>
<tr>
<td>0.05</td>
<td>28</td>
<td>2 or 3 (see Fig. 3.2)</td>
</tr>
<tr>
<td>0.01</td>
<td>45</td>
<td>2</td>
</tr>
</tbody>
</table>

Table C.2: Exact multinomial sample size results when the half interval width (d) is set to 0.1 across all simplexes evaluated. We identify $m$ near where the worst case sample size results are occurring.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$n$ with $d = 0.1$</th>
<th>$m$ (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>45</td>
<td>4 (near to)</td>
</tr>
<tr>
<td>0.1</td>
<td>95</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>120</td>
<td>3</td>
</tr>
<tr>
<td>0.01</td>
<td>190</td>
<td>2</td>
</tr>
</tbody>
</table>
Exact multinomial sample size results when the half interval width (d) is set to 0.05 across all simplexes evaluated. We identify \( m \) near where the worst case sample size results are occurring.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n ) with ( d = 0.05 )</th>
<th>( m ) (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>190</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>390</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>490</td>
<td>2 or 3</td>
</tr>
<tr>
<td>0.01</td>
<td>770</td>
<td>2</td>
</tr>
</tbody>
</table>

Table C.3: Exact multinomial sample size results when the half interval width (d) is set to 0.01 across all simplexes evaluated. We identify \( m \) near where the worst case sample size results are occurring.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n ) with ( d = 0.05 )</th>
<th>( m ) (worst case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>5000</td>
<td>4</td>
</tr>
<tr>
<td>0.1</td>
<td>10000</td>
<td>3</td>
</tr>
<tr>
<td>0.05</td>
<td>12650</td>
<td>3</td>
</tr>
<tr>
<td>0.01</td>
<td>19600</td>
<td>2</td>
</tr>
</tbody>
</table>
Appendix D

Discrepancy Graphics
Boxplots of discrepancies between $P$ and $\tilde{P}$ for 3 stocks
Figure D.1: Lineplots of discrepancies between $P$ and $\tilde{P}$ for Stock 1 (simulated data)
Lineplots of discrepancies between $P$ and $\tilde{P}$ for Stock 2 (simulated data)
Figure D.2: Lineplots of discrepancies between $P$ and $\tilde{P}$ for Stock 3 (simulated data)
Boxplots of discrepancies between \( \Pi \) and \( \tilde{P} \) for 3 stocks
Figure D.3: Lineplots of discrepancies between $\Pi$ and $\tilde{P}$ for Stock 1 (simulated data)
Lineplots of discrepancies between $\Pi$ and $\tilde{P}$ for Stock 2 (simulated data)
Figure D.4: Lineplots of discrepancies between $\Pi$ and $\tilde{P}$ for Stock 3 (simulated data)