When Infinity is Too Long to Wait: On the Convergence of Markov Chain Monte Carlo Methods

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Andrew Nolan Olsen, M.S., B.S.
Graduate Program in Statistics

The Ohio State University

2015

Dissertation Committee:

Radu Herbei, Advisor
Laura Kubatko
Sebastian Kurtek
Abstract

Markov chains are an incredibly powerful tool for statisticians and other practitioners. They allow for random draws, though autocorrelated, to be obtained from a vast array of target distributions, even when the distribution is known only up to a constant. These draws may then be used to answer key questions of interest. Markov chains are used in many settings and are the predominant method for performing inference for Bayesian methods.

The utility of Markov chains lies largely in the simplicity with which they are implemented. The most basic algorithms are easily understood and are not challenging to program. The trade-off with ease of implementation, however, is that issues with Markov chains, particularly with respect to convergence, can occasionally be left undiagnosed. For example, a Markov chain may not have been run long enough to accurately capture the features of the distribution of interest, or perhaps the error of the resulting estimates is grossly underrepresented, if it is considered at all.

The study of Markov chain convergence can be summarized by two main questions:

**Question 1:** Was the simulation run long enough?

**Question 2:** How accurate are the resulting estimates?

While simple and clear, these questions are often left unanswered when Markov chain Monte Carlo methods are implemented. This is largely due to the fact that these
answers require theoretical analysis of the convergence of the Markov chain, which can be challenging.

This dissertation discusses the theory of Markov chains and their convergence, including how to rigorously answer Question 1 and Question 2. A variety of methods are available, and several are illustrated with examples.

One approach answers Question 1 by obtaining draws that approximate the target distribution closely. Markov chains may then be started from these draws, resulting in immediate closeness to the target distribution. Several algorithms for accomplishing this are introduced and developed. Results are provided which quantify the quality of the approximations. A comparison of the efficiency of the algorithms is also provided.

Another approach is the formal establishment of convergence rates. Once these are established, one method to answer Question 1 is to compute the number of iterations required so that the ultimate distribution obtained is close to the target distribution. This approach is also illustrated with examples.

A final approach is to compute standard errors of the resulting estimates, which directly answers Question 2. Question 1, however, is also answered because when estimates are accurate enough, the chain has been run for a sufficient duration. This is similarly illustrated with examples.

Bayesian scale-usage models are used to analyze surveys where individual respondents differ in their use of a rating scale. The convergence rate theory for these models, which guarantees answers to Question 1 and Question 2, is fully established. The methods are then extended to a setting where demographics can govern the way in which respondents differ in their answer styles, and are illustrated with an example using student ratings data for which answers to Questions 1 and 2 are fully provided.
To my amazing wife,

Andrea,

and to our children,

Jake and

McKenna.
Acknowledgments

First, I would like to thank my advisor, Dr. Radu Herbei. A good advisor helps you progress from what you are to what you can become. Radu has done this for me from my first course with him in Probability Theory. Not only is he an incredible instructor, but he instills within his students a passion for what he is teaching. Prior to coming to The Ohio State University, I had never formally studied the convergence of Markov chains. Radu’s masterful mentorship has brought me from that blank space to this dissertation, and I am indebted to him for my success.

Three specific things I’ve learned from Radu have shaped my dissertation and my character.

1. **Don’t give up.** Research isn’t easy, and it takes an incredible advisor to motivate you to keep working when it seems impossible. Radu has never let me give up pursuing a research idea.

2. **Begin with the Simplest Case.** Radu taught me to consider the most basic case before attempting more complicated settings. We can’t solve hard problems without the insights used to solve simple ones.

3. **Don’t Doubt the Impact.** Many graduate students begin research with the intention to revolutionize their field of study. Not every paper is cited thousands of times, but each brings a meaningful contribution to the field.
Without Radu, there would be no dissertation. I would also like to thank my committee, Dr. Laura Kubatko and Dr. Sebastian Kurtek, as well as Dr. Chris Hans from my candidacy committee, for their invaluable research insights.

I would also like to express gratitude to the authors of seminal statistical papers concerning the theory of Markov chain convergence. Their influence on my work is abundantly evident in the citations throughout this dissertation.

My education began long before my graduate studies. I am indebted to the incredible math teachers I had in high school: Jill Corry, Corine Barney, Mary Agrapides, and Tracey Meade. Their passion for math and teaching instilled in me a love and aptitude for the subject, without which I would have never considered statistics as a field of study or have progressed to this level. They remind me of the profound influence a good educator can have on a student.

I am also indebted to the Professors of Statistics at Brigham Young University. Dr. Gilbert Fellingham first introduced me to the wonder of Markov chains in an undergraduate Bayesian course. Dr. Shane Reese, my masters advisor, provided me with my first lessons in statistical research and solidified my understanding of Markov chains and their use for Bayesian inference. Countless are the ways they and the other statistics faculty of Brigham Young University have influenced me as a person and as a statistician.

The Ohio State University statistics faculty have also proved invaluable to my education, providing outstanding instruction and mentorship. I would specifically like to thank Dr. Elizabeth Stasny, the Graduate Chair, for all her tireless and often thankless work.
I would also like to thank my colleagues at JPMorgan Chase, where I served as an intern for two summers, for instilling within me the confidence to be a successful statistician. These mentors have benefited my career and life immensely. Special thanks also goes to Radu for driving to Chase weekly to continue research during my internships.

Most importantly, I would like to thank my wife, Andrea, for her never-ending support. She is the key to my success. She is a constant supply of love, encouragement, and patience. Her accolades during the time of my Ph.D. studies include giving birth to our two children and building a home where they thrive and are growing their young, bright minds. They are a constant source of joy. For me, the secret to surviving graduate school is a supportive spouse who always believes in me and helps me navigate through challenges, all while picking up the slack that my research and dissertation-writing required. My life is forever blessed by having her at my side.

Finally, I am indebted to my parents, Nolan and Jeryl Ann Olsen, for their constant support and for instilling within me the attributes of hard work and determination, without which I would not have been able to complete this Ph.D. I also owe a permanent debt to my Heavenly Father and His Son, Jesus Christ, for all that I am and ever hope to be. I am grateful for my talents and abilities, which I credit to Them completely, and for Their comfort in challenging times. Without Them, this dissertation would not have been possible.
Vita

March 4, 1987 ........................ Born - Davis County, UT

2011 ............................. B.S. Statistical Science,
                               magna cum laude,
                               Brigham Young University

2011 ............................. M.S. Statistics,
                               Brigham Young University

2011–2012 ........................ University Fellow,
                               The Ohio State University.

2012–2015 ........................ Graduate Teaching Associate,
                               The Ohio State University.

Publications

Research Publications


Fields of Study

Major Field: Statistics
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>viii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xvi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Importance of Markov Chains</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Importance of Markov Chain Convergence</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Contributions to the Field</td>
<td>8</td>
</tr>
<tr>
<td>2. Background on Convergence of Markov Chains</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Introduction to Markov Chains</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Geometric Ergodicity</td>
<td>22</td>
</tr>
<tr>
<td>2.2.1 Introduction</td>
<td>22</td>
</tr>
<tr>
<td>2.2.2 Establishing Geometric and Uniform Ergodicity</td>
<td>23</td>
</tr>
<tr>
<td>2.2.3 Examples in the Literature</td>
<td>29</td>
</tr>
<tr>
<td>2.3 Bounds on Total Variation Distance</td>
<td>30</td>
</tr>
<tr>
<td>2.4 Markov Chain Central Limit Theorems</td>
<td>31</td>
</tr>
<tr>
<td>2.4.1 Establishing Existence</td>
<td>32</td>
</tr>
<tr>
<td>2.4.2 Estimation of Asymptotic Variance</td>
<td>34</td>
</tr>
<tr>
<td>2.4.3 Estimation of Quantiles</td>
<td>35</td>
</tr>
<tr>
<td>2.5 Examples of Geometrically Ergodic Markov Chains</td>
<td>37</td>
</tr>
</tbody>
</table>
2.5.1 Bayesian One-Way Random Effects Model using a Gibbs Sampler ........................................... 38
2.5.2 Multivariate Truncated Normal Distribution using a Metropolis-Hastings Sampler .................................. 56

3. Minorization Conditions, Mixture Representations, and Approximate Draws from $\pi$ .................................... 71
   3.1 Minorization Conditions and a Mixture Representation for $\pi$ ................................ 71
   3.2 Obtaining Independent Draws Exactly or Approximately from $\pi$ ................................ 75
      3.2.1 Exact Sample from $\pi$ ........................................................................... 75
      3.2.2 Exact Sample from $\pi$ when the Markov Chain is Geometrically Ergodic .................. 76
      3.2.3 Approximation by Truncation ............................................................................. 79
      3.2.4 Modified Approximation by Truncation ......................................................... 83
      3.2.5 Approximation via Empirical Distribution ..................................................... 84
   3.3 Markov Chains Stationary for $p_n$ ............................................................................. 85
      3.3.1 Reverse Rejection Sampler ............................................................................... 86
      3.3.2 Random Walk Metropolis .............................................................................. 89
      3.3.3 Metropolis-Hastings via The Pseudo-Marginal Approach ......................... 91
   3.4 Examples of Approximate Algorithms ...................................................................... 94
      3.4.1 Known Distribution of $p_n$ ........................................................................... 94
      3.4.2 Bayesian One-Way Random Effects Model ............................................... 100
   3.5 Discussion ............................................................................................................. 110

4. Scale-Usage Models .................................................................................................. 114
   4.1 Introduction ........................................................................................................ 114
      4.1.1 Heterogeneity of Scale Use ............................................................................ 114
      4.1.2 Scale-Usage Data .......................................................................................... 115
   4.2 Models and Inference for Scale-Usage Heterogeneity ............................................ 117
      4.2.1 Motivation ..................................................................................................... 117
      4.2.2 Probit Regression .......................................................................................... 118
      4.2.3 Bayesian Probit Regression .......................................................................... 119
      4.2.4 Bayesian Scale-Usage Models ....................................................................... 122
      4.2.5 Inference Strategies ...................................................................................... 125
   4.3 MCMC Strategies for Scale-Usage Models .......................................................... 127
   4.4 Review of Relevant Matrix and Distributional Properties .................................... 132
      4.4.1 Positive Definite Matrices ............................................................................ 132
      4.4.2 Eigenvalues .................................................................................................. 134
      4.4.3 Determinants ............................................................................................... 135
      4.4.4 Other Matrix Properties ............................................................................... 136
6.2.1 Ensuring Uniform Ergodicity without the Theoretical Legwork ........................................ 267
6.2.2 Nonparametric Bayesian Scale-Usage Models ................................................... 273

Bibliography ................................................................. 285
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Average Absolute Distance of the Monte Carlo Estimate from the Truth</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>Hyper-Parameter Specifications for each Setting</td>
<td>53</td>
</tr>
<tr>
<td>2.2</td>
<td>Convergence Quantities and Iterations Required for TV Distance Less than 0.1 using Bounds by Rosenthal (1995)</td>
<td>55</td>
</tr>
<tr>
<td>2.3</td>
<td>Convergence Quantities and Iterations Required for TV Distance Less than 0.1 using Bounds by Roberts and Tweedie (1999)</td>
<td>55</td>
</tr>
<tr>
<td>2.4</td>
<td>Posterior Medians and 95% Equal-Tail Credible Interval Bounds with Corresponding 95% Estimation bounds</td>
<td>57</td>
</tr>
<tr>
<td>2.5</td>
<td>Generated Multivariate Data Determining the Cutpoints under Setting 1</td>
<td>62</td>
</tr>
<tr>
<td>2.6</td>
<td>Generated Multivariate Data Determining the Cutpoints under Setting 2</td>
<td>63</td>
</tr>
<tr>
<td>2.7</td>
<td>Number of Draws Required for TV Distance to be Below 0.1 for Setting 1</td>
<td>64</td>
</tr>
<tr>
<td>2.8</td>
<td>Number of Draws Required for TV Distance to be Below 0.1 for Setting 2</td>
<td>65</td>
</tr>
<tr>
<td>2.9</td>
<td>Estimated Means for Truncated Multivariate Normal Distributions in Setting 1 with 95% Monte Carlo Margins of Error</td>
<td>69</td>
</tr>
<tr>
<td>2.10</td>
<td>Estimated Means for Truncated Multivariate Normal Distributions in Setting 2 with 95% Monte Carlo Margins of Error</td>
<td>70</td>
</tr>
<tr>
<td>3.1</td>
<td>Iterations Required for TV Distance &lt; 0.005 using a Random Walk Metropolis Markov Chain with Candidate 1</td>
<td>99</td>
</tr>
</tbody>
</table>
3.2 Iterations Required for TV Distance < 0.005 using a Random Walk Metropolis Markov Chain with Candidate 2 ........................................ 99

3.3 Number of Iterations Required for TV Distance < 0.005 .......................... 100

3.4 Number of Approximate Draws Produced using 1 Million Draws of \( \tau \) Under the Different Approximate Algorithms ............................... 106

4.1 Example of Scale-Usage Raw Data from Rossi et al. (2001) .................. 116

4.2 Summary of Geometrically Ergodic Scale-Usage Models ................. 167

5.1 Summary of Demographic-Specific Models ........................................ 198

5.2 Bayes Factor Model Comparison Scales (Kass and Raftery 1995) ...... 200

5.3 Simulation Study Settings for \( \theta \) and \( \sigma_\tau^2 \) .............................. 210

5.4 Percent of Correctly Selected Models using DIC and Bayes Factors . 215

5.5 Distribution of Models Chosen Using DIC for Each Model Setting .... 217

5.6 Ad-Hoc Model Selection Results ....................................................... 219

5.7 Coverage of 95\% Credible Intervals for \( \tau \) Parameters .................. 221

5.8 Correlation Between Posterior Means and Truth for \( \tau \) Parameters .. 222

5.9 Coverage of 95\% Credible Intervals for \( \theta \) Parameters .................. 224

5.10 Correlation Between Posterior Means and Truth for \( \theta \) Parameters . 225

5.11 Coverage of 95\% Credible Intervals for \( \mu \) Parameter ................. 226

5.12 Correlation Between Posterior Means and Truth for \( \mu \) .................. 227

5.13 Coverage of 95\% Credible Intervals for \( \sigma_\tau^2 \) Parameters .......... 228

5.14 Coverage of 95\% Credible Intervals for \( \sigma^2 \) Parameters ............ 229

5.15 Coverage of 95\% Credible Intervals for Free Cutpoint Parameters .. 230
5.16 Summary of Demographic Model Simulation Study Results .................. 231
5.17 Student Ratings Questions .......................................................... 232
5.18 Demographic Information for Student Ratings .............................. 234
5.19 DIC and Bayes Factors for Model Selection Among Demographic-Specific Models ................................................................. 241
5.20 Initial Posterior Summaries for $\theta_t$ for Each Demographic from Model 5.1 which are used for the Ad-Hoc Model Selection Approach .............. 242
5.21 Posterior Summary of $\mu$ .............................................................. 247
5.22 Posterior Medians of Covariance Matrix $\Sigma$ ................................. 248
5.23 Standard Errors for Posterior Medians of Covariance Matrix $\Sigma$ .... 248
5.24 Posterior Summary for Regression Coefficients Predicting Overall Instructor Rating ................................................................. 249
5.25 Posterior Summary of the Cutpoint Parameters ............................... 250
5.26 Posterior Summary for Demographic Level $\theta$ Parameters ............... 251
5.27 Posterior Summary for Demographic Level $\sigma_t^2$ Parameters .......... 252
5.28 Cluster Means for Standardized Scores ........................................... 254
5.29 Cluster Means for Original Data ...................................................... 257
5.30 Summary of Inference for Student Ratings Data ............................. 260
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>6</td>
</tr>
<tr>
<td>1.2</td>
<td>10</td>
</tr>
<tr>
<td>1.3</td>
<td>11</td>
</tr>
<tr>
<td>2.1</td>
<td>68</td>
</tr>
<tr>
<td>3.1</td>
<td>95</td>
</tr>
<tr>
<td>3.2</td>
<td>95</td>
</tr>
<tr>
<td>3.3</td>
<td>97</td>
</tr>
<tr>
<td>3.4</td>
<td>98</td>
</tr>
<tr>
<td>3.5</td>
<td>99</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>3.6</td>
<td>Comparison of the candidate and actual mass functions for the independence Metropolis-Hastings pseudo-marginal algorithm. 101</td>
</tr>
<tr>
<td>3.7</td>
<td>Total variation distance of the pseudo-marginal independence algorithm with varying number of draws utilized to estimate $P(\tau \geq n)$. 101</td>
</tr>
<tr>
<td>3.8</td>
<td>TV Distance and acceptance rates for various $M$ values using Algorithm 3.3 for the Bayesian random effects model. 108</td>
</tr>
<tr>
<td>3.9</td>
<td>Approximation for $p_n$ using Algorithm 3.5 for the Bayesian random effects model. 109</td>
</tr>
<tr>
<td>3.10</td>
<td>Autocorrelation functions for the Markov chains approximating $p_n$ for the Bayesian random effects model. 110</td>
</tr>
<tr>
<td>3.11</td>
<td>Comparison of posterior density estimates using 100,000 approximate draws from Algorithm 3.4 (dashed) with the original Markov chain run for 5 million iterations (solid). 113</td>
</tr>
<tr>
<td>4.1</td>
<td>Several decision paths of a respondent on a given question (left). Continuous distribution of the paths at time $d_T$ (gray, right) along with the corresponding probability distribution of the categories (black, right). 118</td>
</tr>
<tr>
<td>5.1</td>
<td>Original (left) and transformed (right) trace plots for $\tau$, $\theta$, and $\mu$ parameters. 212</td>
</tr>
<tr>
<td>5.2</td>
<td>Trace plots for $\sigma^2_1$, $\Sigma_{1,1}$, $\sigma^2_{1,1}$, $Y_{1,1}$, $c_1$, and $c_4$. No convergence problems are detected. 214</td>
</tr>
<tr>
<td>5.3</td>
<td>Sample scatterplot of the true $\tau$ parameters with the posterior means. 223</td>
</tr>
<tr>
<td>5.4</td>
<td>Bubble chart of scale-usage patterns across all individuals. 234</td>
</tr>
<tr>
<td>5.5</td>
<td>Bubble chart of scale-usage patterns by year in school. 235</td>
</tr>
<tr>
<td>5.6</td>
<td>Bubble chart of scale-usage patterns by cumulative GPA. 237</td>
</tr>
<tr>
<td>5.7</td>
<td>Bubble chart of scale-usage patterns by expected grade. 238</td>
</tr>
<tr>
<td>5.8</td>
<td>Bubble chart of scale-usage patterns by whether or not the course is required. 239</td>
</tr>
</tbody>
</table>
5.9 Posterior means of \( \tau_i \) parameters compared to raw score means.  
5.10 Posterior means of \( \sigma_i^2 \) parameters compared to raw score standard deviations.  
5.11 SSE scree plot for selecting the number of clusters for standardized scores. 
5.12 First two principal components to illustrate cluster separation for standardized scores. 
5.13 SSE scree plot for selecting the number of clusters for raw data. 
5.14 First two principal components to illustrate cluster separation for raw data. 
5.15 Predictive distribution for overall instructor assessment by demographic. 
6.1 Comparison of distributions of ergodic averages after 1 million iterations for both the original chain (left) and the modified, uniformly ergodic chain (right). 
6.2 Comparison of the trace plots for the original Markov chain and the modified chain which is uniformly ergodic.
Chapter 1: Introduction

Markov chains are an incredibly powerful tool for statisticians. With increased computational power, problems which were previously unsolvable are now readily addressed using complex models, many of which rely on Markov chains. The utility of Markov chains lies largely in the simplicity with which they are implemented. The most basic algorithms are easily understood and are not challenging to program. Software packages have increasingly made these tools available to practitioners, even without rigorous statistical training.

The trade-off with ease of implementation, however, is that issues with Markov chains, particularly with respect to convergence, can occasionally be left undiagnosed. For example, a Markov chain may not have been run long enough to accurately capture the features of the distribution of interest, or perhaps the error of the resulting estimates is grossly underrepresented, if it is considered at all. A key purpose of this dissertation is to discuss the theoretical considerations of convergence of Markov chains which remedies these problems. In this chapter, the use of Markov chains is motivated. The importance of Markov chain convergence is also discussed and illustrated with an example. Finally, an outline of the rest of this work is provided.
1.1 Importance of Markov Chains

The primary reason for the widespread use of Markov chain Monte Carlo (MCMC) methods is that Bayesian problems often require Markov chains as a method of approximating the posterior distribution in order to make inference. As Bayesian methods are at the forefront of MCMC use, they are now reviewed and discussed.

The basic premise of Bayesian methodology is that belief about population parameters is updated based upon data. Specifically, Bayesian models combine the data, $Y$, with previously known information or intuition about the parameters $\theta$ which govern $Y$. The data $Y$ are incorporated into the analysis through a likelihood function denoted $f(Y|\theta)$ where $\theta$ is a collection of parameters for which inference is desired. The previous information, if any, is incorporated into the analysis through a prior distribution on the parameters $\theta$. This prior belief may depend upon other parameters, called hyperparameters, and if prior distributions are assigned to these parameters, they are referred to as hyperpriors. For simplicity, let $\theta$ represent all parameters, including any hyperparameters, and denote the entirety of the prior and hyperprior information as the distribution $\pi(\theta)$. Analysis is then focused on the posterior distribution $\pi(\theta|Y)$ which is the updated distribution for $\theta$ given the data $Y$. Bayes’ Theorem, with foundations in Bayes and Price (1763), is the key to how the data likelihood $f(Y|\theta)$ and prior distribution $\pi(\theta)$ are combined to obtain the posterior distribution:

$$\pi(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{\int_\theta f(Y|\theta)\pi(\theta)d\theta}. \quad (1.1)$$

The posterior distribution is the ultimate theoretical underpinning of Bayesian methods. All other Bayesian tools, including MCMC simulation, attempt to understand
the posterior distribution for inference. Ultimately, the answer to any Bayesian problem lies in the posterior distribution.

With the incredible importance of the posterior distribution, it is surprising that, with the exception of conjugate families and a handful of other examples, (1.1) is usually not available in closed form. Specifically, the constant

\[ m(Y) = \int_{\theta} f(Y|\theta)\pi(\theta)d\theta \]

is often difficult, if not impossible, to compute with accuracy. Therefore, computing averages of functions with respect to the posterior distribution in closed form is often futile. It is also often challenging or impractical to sample directly from the posterior distribution.

The benefit of Markov chain Monte Carlo methods, such as Metropolis-Hastings and Gibbs Sampling, is that \( m(Y) \) is not required for implementation. These methods only require the posterior distribution, \( \pi(\theta|Y) \), up to a constant! MCMC methods are then able to obtain an approximate sample, typically with autocorrelation, which nonetheless can be used to compute averages and explore features of the posterior distribution. This is all made possible by the ergodic theorem (see Theorem 2.17) which provides a strong law of large numbers for averages computed from MCMC output when the Markov chain satisfies a few basic and often-true conditions. The question then becomes, given finite time, has the ergodic average converged closely enough to the truth.

### 1.2 Importance of Markov Chain Convergence

The study of Markov chain convergence can be summarized by two main questions:

**Question 1:** Was the simulation run long enough?
Question 2: How accurate are the resulting estimates?

While simple and clear, these questions are often left unanswered when MCMC methods are implemented. For example, Flegal et al. (2008) pointed out that in 2006, of 39 articles using MCMC in the *Journal of the American Statistical Association*, *Biometrika*, and *Journal of the Royal Statistical Society, Series B*, only three of them addressed Question 2.

There are three primary ways to answer Question 1. The first is to compute a bound on the total variation distance (see Definition 2.15) for a given number of iterations. If the number of iterations is sufficiently large so that this bound is small, then the Markov chain has converged closely to the target distribution, and the simulation has been run long enough. Illustrations of this method are provided in Chapter 2.

The second method is to start the Markov chain from an initial distribution which already has a small total variation distance from the target distribution $\pi$. This can typically be done with a number of exact or approximate techniques depending on the properties of the Markov chain, and is discussed in depth in Chapter 3.

The third method to answer Question 1 is also the most commonly used method to answer Question 2: estimate the error associated with the quantity of interest and use a central limit theorem to determine whether the estimate is precise enough for inference. If not, then the simulation should be run longer, answering Question 1. This also answers Question 2 by providing bounds on the estimate and thereby quantifying the accuracy of the results. This method is illustrated in both Chapter 2 and Chapter 5.
Perhaps the biggest reason relatively few practitioners attempt to answer either Question 1 or Question 2 is that each of the three methods discussed here require theoretical analysis of the convergence of the Markov chain, which can be challenging. An introduction to the required theory and examples are provided in Chapter 2; Chapter 4 proves the necessary theory so a Markov chain central limit theorem exists and Markov chain Monte Carlo asymptotic standard errors may be computed for a specific class of Bayesian models. While challenging, this theory provides the tools such that answers to Question 1 and Question 2 are available. Lack of the theory leaves the two questions unanswered.

When Question 1 and Question 2 are not formally answered, a common compensation is to take an incredibly long Markov chain. However, the determination of what is “incredibly long” depends substantially on the problem, and in some cases, the inability to take an infinite sample is paralyzing to inference.

An example is now provided comparing inference when the theoretical properties ensuring answers to Question 1 and Question 2 are satisfied and when they are not satisfied. Note the exponential distribution with scale $\theta$ has probability density function

$$f_\theta(x) = \frac{1}{\theta} \exp \left(-\frac{x}{\theta}\right).$$

Choose as the target distribution

$$\pi(x) = f_1(x) = \exp(-x).$$

Consider several independence Metropolis-Hastings Markov chains (see Algorithm 2.1), with a candidate distribution also exponentially distributed, but with various scale parameters: 0.25, 0.5, 1, and 2. These distributions are shown in Figure 1.1. It is
well known that for $\theta < 1$, the needed convergence properties will not be satisfied for the Markov chain to have an associated central limit theorem (e.g., see Jones and Hobert 2001a). As shown, these distributions have much thinner tails than the target distribution; this causes the Markov chain to explore the space inefficiently in comparison to $\theta = 2$ which has a thicker tail than the target distribution.

Figure 1.1: Comparison of candidate functions for independence Metropolis-Hastings Markov chains stationary for an exponential distribution with scale $\theta = 1$.

To evaluate the behavior of Markov chain Monte Carlo averages after a long chain has been run, begin with an initial value and run the chain for 1 million iterations. This number of iterations is seen relatively often in practice. Then, consider two inferential quantities: (1) $E_\pi(X)$ and (2) $P_\pi(X \geq 2)$. This process is repeated 100,000 times to explore the distribution. Figure 1.2 shows the distribution of the estimates.
for $E_\pi(X)$. Note that the candidates using $\theta = 0.25$ and $\theta = 0.5$ have very skewed distributions, whereas the candidates with $\theta = 1$ (equivalent to i.i.d sampling) and $\theta = 2$ appear to be normally distributed. Similarly, Figure 1.3 shows the distribution of the estimates for $P_\pi(X \geq 2)$. The results using the candidates with the two smaller $\theta$ values are again skewed whereas the results using the two larger candidates appear to be normally distributed. This is consistent with what is expected based on the well-known convergence results for these samplers.

To quantify the error using each of these candidates, the average absolute distance from the truth of the Monte Carlo estimate after 1 million iterations is provided in Table 1.1 for each of the candidates and each of the quantities of interest. The error for $\theta = 0.25$ is substantial for both quantities. The quality of the inference could certainly be inadequate in this case. The error for $\theta = 1$ and $\theta = 2$, however, is quite small.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>$E_\pi(X)$</th>
<th>$P_\pi(X \geq 2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>0.1152</td>
<td>0.0304</td>
</tr>
<tr>
<td>0.50</td>
<td>0.0106</td>
<td>0.0019</td>
</tr>
<tr>
<td>1.00</td>
<td>0.0008</td>
<td>0.0003</td>
</tr>
<tr>
<td>2.00</td>
<td>0.0012</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Thus, even in this one-dimensional example, when the theoretical properties ensuring answers to Question 1 and Question 2 are not satisfied, a finite number of iterations can lead to poor inference. With larger dimensions and more complicated Markov chains, this problem can be magnified. This illustrates the keen importance
of establishing the theory needed to guarantee convergence of Markov chains when infinity is too long to wait. This dissertation outlines the necessary theory and provides several examples of how to apply it in practice.

1.3 Contributions to the Field

Chapter 2 discusses the theory of Markov chains and their convergence, including how to rigorously answer Question 1 and Question 2. Contributions to the field in this chapter include an illustrative proof for showing geometric ergodicity of the Bayesian random effects model and the development of a uniformly ergodic algorithm for sampling from the multivariate truncated normal distribution. Thorough answers to both Question 1 and Question 2 are provided for each of these examples.

Chapter 3 focuses on answering Question 1 by obtaining draws which approximate \( \pi \) closely. Markov chains may then be started from these draws, resulting in immediate closeness to stationarity. The contributions to the field include a modification of the approximate method of Hobert and Robert (2004) which results in an algorithm with finite expected running time. Further, this chapter develops three Markov chains which are stationary for the mixture weights of a mixture distribution equivalent to \( \pi \), hence allowing for approximate draws from \( \pi \) to be obtained.

Chapter 4 introduces a specific type of Bayesian model used to analyze surveys where individual respondents differ in their use of a rating scale. The primary contribution to the field is the establishment of proofs of geometric ergodicity for several algorithms that may be used for fitting various scale-usage models. These theoretical results then allow a practitioner to thoroughly answer Question 1 and Question 2
when making inference. Additionally, a minorization condition is developed for scale-usage models, which may further be used to answer Question 1.

Chapter 5 extends these scale-usage models to a setting where scale-usage patterns are related to demographics. The contributions include a simulation study which compares several potential models to account for demographic effects. These models are then applied to analyze student ratings data. The theory proved in Chapter 4 is used to thoroughly answer Question 1 and Question 2 by estimation of Markov chain Monte Carlo standard errors.

Chapter 6 finally concludes with a summary of the dissertation and the associated conclusions. A discussion of future work is also provided including a method which constructs Markov chains which automatically satisfy the difficult theory necessary to answer Question 1 and Question 2 for a large class of problems. A more flexible, nonparametric model for Bayesian scale-usage models is also motivated and introduced.

The aim of this dissertation is to illustrate how answers to Question 1 and Question 2 may be obtained so their use becomes a common practice among practitioners. This will improve the quality of inference when MCMC methods are employed and allow for increased confidence in the conclusions made based on MCMC simulations.
Figure 1.2: Comparison of the distributions of Monte Carlo averages estimating $E_{\pi}(X)$ when 1 million iterations are obtained using the exponential candidate functions with scale parameter $\theta$ equal to 0.25, 0.5, 1, and 2. The solid gray line indicates the true value.
Figure 1.3: Comparison of the distributions of Monte Carlo averages estimating $P_\pi(X \geq 2)$ when 1 million iterations are obtained using the exponential candidate functions with scale parameter $\theta$ equal to 0.25, 0.5, 1, and 2. The solid gray line indicates the true value.
Chapter 2: Background on Convergence of Markov Chains

In this chapter, an introduction to Markov chain theory is provided, establishing the definitions and notation required for the study of Markov chain convergence. Geometric ergodicity is then introduced along with a discussion about how it may be established and the ways in which Question 1 and Question 2 may be answered given geometric ergodicity. Markov chain Monte Carlo central limit theorems are then discussed, along with tools to estimate the central limit theorem variance. Two examples are then provided which illustrate these methods, providing rigorous answers to Question 1 and Question 2.

2.1 Introduction to Markov Chains

Informally, a discrete-time Markov chain is a collection of random variables $X_0, X_1, X_2, \ldots$, hereafter denoted $\{X_n\}$, whose behavior is governed by a transition kernel such that future states depend on past states only through the current state.

To build the more formal definition of a Markov chain and the related theory, several definitions are required. Note that definitions for this chapter, unless otherwise specified, may be found in Meyn and Tweedie (1993). Robert and Casella (2004) also provide an excellent introduction to Markov chain Monte Carlo theory and many definitions may also be referenced there.
First, an understanding of random variables and transition kernels is required.

**Definition 2.1** (Resnick, 1999). Let \((X, \mathcal{B}(X))\) and \((X', \mathcal{B}(X'))\) be general measurable spaces. A map \(X : X \to X'\) is *measurable* if \(X^{-1}(\mathcal{B}(X')) \subseteq \mathcal{B}(X)\). Further, if \((X', \mathcal{B}(X')) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))\), then the measurable map \(X\) is a *random variable*.

Given the definition of a measurable map, a transition kernel may then be defined.

**Definition 2.2.** If \(K\) is a function defined on \(X \times \mathcal{B}(X)\) where

(i) \(K(x, \cdot)\) is a probability measure on \(\mathcal{B}(X)\) for all \(x \in X\), and

(ii) \(K(\cdot, A)\) is a non-negative measurable function on \(X\) for all \(A \in \mathcal{B}(X)\),

then \(K(\cdot, \cdot)\) is a *transition kernel*. Further, the *n-step transition kernel* is recursively defined as

\[
K^n(x, A) = \int_X K^{n-1}(y, A)K(x, dy).
\]

A Markov chain is then defined in terms of this transition kernel.

**Definition 2.3.** A sequence of random variables \(\{X_n\}\) is a *Markov chain* corresponding to \(K\) if, for any \(t \in \{1, 2, \ldots\}\), and any \(x_0, \ldots, x_t \in \mathbb{R}\)

\[
P(X_{t+1} \in A | X_0 = x_0, \ldots, X_t = x_t) = P(X_{t+1} \in A | X_t = x_t)
= \int_A K(x_t, dx),
\]

where \(K\) is the corresponding transition kernel.

Thus, a Markov chain depends on the past only through the current state. The extension to multiple dimensions is straightforward. Often a transition kernel \(K\) can be defined by a transition density \(k\). To discuss this formally, note the following definition and theorem.
Definition 2.4 (Resnick, 1999). Assume $\lambda$ and $\mu$ are two $\sigma$-finite measures on $(X, \mathcal{B}(X))$. Then $\lambda$ is absolutely continuous with respect to $\mu$ if $\mu(A) = 0$ implies that $\lambda(A) = 0$. This is denoted $\lambda \ll \mu$.

Theorem 2.5 (Radon-Nikodym Theorem, Resnick, 1999). If $\nu$ is a positive $\sigma$-finite measure and $\nu \ll \mu$, then there exists an integrable random variable $f : X \to \mathbb{R}$ such that

$$
\nu(A) = \int_A f \, d\mu, \quad \forall A \in \mathcal{B}(X).
$$

Further, $f$ is almost surely unique with respect to $\mu$ and is written

$$
f = \frac{d\nu}{d\mu} \quad \text{or} \quad d\nu = f \, d\mu.
$$

The random variable $f$ is also called the density of $\nu$ with respect to $\mu$.

Often a Markov transition kernel $K(x, \cdot)$ will be absolutely continuous with respect to some $\sigma$-finite measure $\mu$. In this case, due to the Radon-Nikodym Theorem, there exists a random variable $k(x, \cdot)$ such that for any $A \in \mathcal{B}(X)$

$$
K(x, A) = \int_A k(x, y) \mu(dy).
$$

This random variable $k(x, \cdot)$ is the Radon-Nikodym derivative of $K(x, \cdot)$ with respect to $\mu$ and is denoted

$$
k(x, \cdot) = \frac{dK(x, \cdot)}{d\mu}.
$$

Further, $k(x, \cdot)$ is known as a transition density. The original transition kernel can easily be recovered from the transition density $k(x, \cdot)$ through integration. Throughout this work, when a transition density $k(x, y)$ is provided, it implies the correct transition kernel

$$
K(x, A) = \int_A k(x, y) \mu(dy). \quad (2.1)
$$
In such cases, unless otherwise specified, $\mu$ is assumed to be Lebesgue measure.

**Example 2.1 (Metropolis-Hastings Transition Kernel)**

The Metropolis-Hastings Markov chain transition kernel $K$ is defined by its transition density $k$. Let $\pi$ be the target measure on $(X, \mathcal{B}(X))$ which is absolutely continuous with respect to some measure $\mu$, and, by a common abuse of notation, let $\pi$ also denote its Radon-Nikodym derivative (density) with respect to $\mu$. Let $q(y|x)$ be a conditional density used to make candidate draws. The Metropolis-Hastings transition density is then

$$k(x, y) = q(y|x) \min \left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right) + (1 - r(x))\delta_x(y), \quad (2.2)$$

where $\delta_x(y)$ is the Dirac mass in $x$ and

$$r(x) = \int_x q(y|x) \min \left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right) dy.$$ 

This kernel was originally developed by Metropolis et al. (1953) for the case where $q(x|y) = q(y|x)$ and later by Hastings (1970) for the general case. According to this kernel, the Metropolis-Hastings Algorithm is presented in Algorithm 2.1.

**Algorithm 2.1 (Metropolis-Hastings Algorithm)**

Conditional on $X_{n-1} = x$,

(i) Draw a candidate $y$ from the distribution $q(y|x)$.

(ii) Set

$$X_n = \begin{cases} y & \text{with probability } \min \left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right) \\ x & \text{with probability } 1 - \min \left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right). \end{cases}$$

This algorithm has been widely applied and is part of the arsenal of every practicing Bayesian statistician. Another widely-applied algorithm is the Gibbs sampler.
Example 2.2 (Gibbs Sampler Transition Kernel)

Let \( \pi \) be defined on a multivariate space where the dimensions are broken into \( d \) groups \( \theta = \{ \theta_1, \ldots, \theta_d \} \). Also, assume \( \pi \) is absolutely continuous with respect to a measure \( \mu \) such that a density \( \pi \) (again, abusing notation) is induced. Let the notation \( \theta_{-i} \) denote \( \theta \) without those dimensions contained in \( \theta_i \). Similarly denote \( \theta_{<i} \) to be those dimensions not in \( \theta_i \) yet in some \( \theta_j \) for \( j < i \), and let \( \theta_{>i} \) denote all dimensions not in \( \theta_i \) or \( \theta_{<i} \). Denote the marginal and conditional densities respectively as

\[
\pi_i(\theta_i) = \int \pi(\theta) \mu(d\theta_{-i}) \quad \text{and} \quad \pi_{i|\theta_i}(\theta_{-i}) = \frac{\pi(\theta)}{\pi_{i|\theta_i}(\theta_{-i})}.
\]

A Gibbs sampler is defined by the transition density

\[
k(\theta', \theta) = \pi_{1|\theta_1}(\theta_1|\theta_{-1}') \pi_{2|\theta_1, \theta_2}(\theta_2|\theta_1, \theta_{>2}') \pi_{3|\theta_3}(\theta_3|\theta_{<3}, \theta_{>3}') \times \cdots \times \pi_{d-1|(d-1)}(\theta_{d-1}|\theta_{<d-1}, \theta_{>d-1}') \pi_{d|\theta_d}(\theta_d|\theta_{<d}).
\]

This Markov chain was introduced by Geman and Geman (1984) and was further developed by Gelfand and Smith (1990). The respective algorithm is now defined.

Algorithm 2.2 (Gibbs Sampler Algorithm)

Given \( \theta' = \{ \theta'_1, \ldots, \theta'_d \} \), update to \( \theta \) with the following updates.

(i) Update \( \theta_1 \sim \pi_{1|\theta_1}(\theta_1|\theta_1') \).

(ii) Update \( \theta_i \sim \pi_{i|\theta_i}(\theta_i|\theta_<i, \theta_>i) \) sequentially for \( i = 2, \ldots, d \).

(iii) Set \( \theta = \{ \theta_1, \ldots, \theta_d \} \).

If one or more of the conditional distributions in Algorithm 2.2 is not available in closed form, an acceptable sampling strategy is to replace the update with a Metropolis-Hastings step which is stationary for the desired conditional distribution.
This preserves the stationarity of the algorithm for $\pi$. Such hybrid algorithms are often referred to as Metropolis-within-Gibbs, component-wise Metropolis-Hastings, and conditional Metropolis-Hastings samplers. Altogether, Metropolis-Hastings, Gibbs Samplers, and hybrid algorithms can be applied to an incredibly wide array of problems. The number of unique ways to utilize and improve these methods continues to grow tremendously.

Many Markov chains, including the Metropolis-Hastings algorithm and Gibbs sampler in most cases, share several characteristics which make them stable and well-behaved. These include $\psi$-irreducibility, aperiodicity, Harris recurrence, and invariance. These together imply Harris ergodicity, which provides results which are critical to the use of Markov chains for simulating from intractable probability distributions to estimate quantities of interest. Definitions for each of these properties are now provided.

**Definition 2.6.** For a measure $\varphi$ on $B(X)$, the Markov chain $\{X_n\}$ with transition kernel $K(\cdot, \cdot)$ is $\varphi$-irreducible if for every $A \in B(X)$ such that $\varphi(A) > 0$ there exists an $n$ such that $K^n(x, A) > 0$ for all $x \in X$.

Thus, in essence, $\varphi$-irreducibility implies that the chain can go from any state $x$ to any set $A$ which has non-zero measure according to $\varphi$.

If a chain is $\varphi$-irreducible, then it is also $\psi$-irreducible where $\psi$ is the maximal irreducibility measure (see Proposition 4.2.2 of Meyn and Tweedie (1993) for details and construction of $\psi$). The notation $\psi$ will imply the maximal irreducibility measure throughout. The concept of aperiodicity may then be established for $\psi$-irreducible Markov chains.
**Definition 2.7.** Let \( \{X_n\} \) be a \( \psi \)-irreducible Markov chain on \( X \). Then there exist \( d \) disjoint sets \( D_1, \ldots, D_d \in \mathcal{B}(X) \), called a “d-cycle”, such that

\[
P(x, D_{i+1}) = 1 \text{ for } x \in D_i, \quad i = 1, \ldots, D_{i-1},
\]

and

\[
P(x, D_1) = 1 \text{ for } x \in D_d,
\]

where the set \( N = [\cup_{i=1}^{d} D_i]^c \) is such that \( \psi(N) = 0 \). The largest \( d \) for which a \( d \)-cycle occurs is the period of \( \{X_n\} \). If \( d = 1 \) then the chain is aperiodic (see Theorem 5.4.4 of Meyn and Tweedie (1993) and the surrounding discussion for details including how to find \( d \)).

Aperiodicity thus implies that the chain does not have to follow a specific tour through the space to get from one state to another.

In order to define recurrence for the Markov chain \( \{X_n\} \), for \( A \in \mathcal{B}(X) \) let \( \mathbb{I}_A(x) \) be the indicator function \( \mathbb{I}_A : X \to \{0, 1\} \) where

\[
\mathbb{I}_A(x) = \begin{cases} 
1 & \text{for } x \in A \\
0 & \text{for } x \notin A.
\end{cases} \tag{2.3}
\]

Then, let

\[
\eta_A = \sum_{i=1}^{\infty} \mathbb{I}_{A}(X_n),
\]

be the number of passages of \( \{X_n\} \) into \( A \) where \( A \in \mathcal{B}(X) \). Further, the notation \( P_x \), as utilized by Meyn and Tweedie (1993), denotes the law of the Markov chain transition kernel where \( X_0 = x \). Similarly, the expectation \( E_x \) denotes expectation according to the law of the Markov chain with initial state \( X_0 = x \).

**Definition 2.8.** A Markov chain \( \{X_n\} \) is recurrent if it is \( \psi \)-irreducible and if for every \( A \in \mathcal{B}(X) \) such that \( \psi(A) > 0 \) then \( E_x(\eta_A) = \infty \) for every \( x \in A \).
Thus, for a recurrent chain \( \{X_n\} \), the expected number of passages into the set \( A \) with nonzero \( \psi \)-measure is infinite. A stronger condition called Harris recurrence is now defined.

**Definition 2.9.** A set \( A \) is *Harris recurrent* if

\[
P_x(\eta_A = \infty) = 1
\]

for all \( x \in A \). If \( \{X_n\} \) is \( \psi \)-irreducible and \( A \) is Harris recurrent for all \( A \in \mathcal{B}(X) \) with \( \psi(A) > 0 \), then the chain is Harris recurrent.

Thus, for a Harris recurrent chain, not only is the expected number of passages into the set \( A \) (with nonzero \( \psi \)-measure) infinite, but the number of passages itself is infinite almost surely \( P_x \). This is therefore a much stronger property.

Perhaps the most important purpose of Markov chains is to explore the parameter space according to some invariant distribution. This concept is now formally defined.

**Definition 2.10.** If \( \pi \) is a \( \sigma \)-finite measure and

\[
\pi(A) = \int_X K(x, A)\pi(dx), \quad \forall A \in \mathcal{B}(X),
\]

then \( \pi \) is *invariant* for \( K(\cdot, \cdot) \) and the corresponding Markov chain. If a \( \psi \)-irreducible Markov chain is invariant, then it is *positive*; otherwise, it is *null*.

Thus, a Markov chain is stationary for a distribution \( \pi \) when, if it is distributed according to \( \pi \) at one state, then the transition kernel preserves the distribution such that at the next state it is also distributed according to \( \pi \).

A common method to prove a Markov chain is stationary for \( \pi \) is to establish detailed balance, which is now defined.
Definition 2.11. A Markov chain \( \{X_n\} \) with transition kernel \( K \) and corresponding density \( k \) satisfies the detailed balance condition if there exists a function \( f \) such that

\[
k(y, x)f(y) = k(x, y)f(x).
\]

Before the formal theorem which links detailed balance with stationarity, define reversibility.

Definition 2.12. Define \( \nu \) as the conditional distribution \( X_n | X_{n+1} = x \) for some \( n \) and \( x \). A Markov chain \( \{X_n\} \) stationary for \( \pi \) is reversible if the distribution of \( X_n | X_{n-1} = x \) is also \( \nu \) for all \( n = 1, 2, \ldots \) and \( x \in X \). Thus, the distribution of the next state given the current state \( x \) is the same as the distribution of the previous state given the current state \( x \).

Theorem 2.13. If a Markov chain with transition kernel \( K \) satisfies detailed balance with \( \pi \), a probability density function, then

(i) the density \( \pi \) is the invariant density of the chain, and

(ii) the chain is reversible.

Thus, detailed balance provides the invariant distribution of the chain and guarantees reversibility. The culmination of all these properties is now provided.

Definition 2.14. If \( \{X_n\} \) is \( \psi \)-irreducible, aperiodic, and positive Harris recurrent then it is Harris ergodic.

Note that as \( \psi \)-irreducibility was required in the definitions of aperiodicity and Harris recurrence, it is often implicitly assumed when those conditions are met and therefore omitted from the definition of Harris ergodicity.
Harris ergodic Markov chains enjoy two very important properties which are crucial to their use. First, define the total variation distance between two measures.

**Definition 2.15 (Total Variation Distance).** Let $\mu_1$ and $\mu_2$ be two measures on the measurable space $(X, \mathcal{B}(X))$. Then, the total variation distance between $\mu_1$ and $\mu_2$ is defined to be

$$||\mu_1(\cdot) - \mu_2(\cdot)||_{TV} = \sup_{A \in \mathcal{B}(X)} |\mu_1(A) - \mu_2(A)|.$$  \hspace{1cm} (2.4)

**Theorem 2.16 (Meyn and Tweedie, 1993, Theorem 13.3.3).** If $\{X_n\}$ is Harris ergodic then

$$\lim_{n \to \infty} ||K^n(x_0, \cdot) - \pi(\cdot)||_{TV} = 0,$$

where $x_0 \sim \mu$ and $\mu$ is any initial distribution.

Thus, a Harris ergodic Markov chain converges in total variation distance to its stationary measure regardless of the initial distribution $\mu$. This is critical in practice where initial values are often taken as constants (implying $\mu$ is degenerate) or from distributions far from $\pi$.

For the next ergodic result, recall that $P_x$ denotes the measure of the Markov chain with initial state $x$.

**Theorem 2.17 (Meyn and Tweedie, 1993, Theorem 17.0.1).** If $g$ is a function such that $E_\pi(|g|) < \infty$ and $\{X_n\}$ is Harris ergodic, then

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \to E_\pi(g) \hspace{1cm} a.s. \ [P_x]$$  \hspace{1cm} (2.5)

as $n \to \infty$ for all $x \in X$.

Thus, ergodic averages converge almost surely to the desired expectation with respect to $\pi$. This theorem is the fundamental reason Markov chain Monte Carlo methods are extensively utilized.
For the remainder of this work, all Markov chains are assumed to be Harris ergodic. Due to results in Tierney (1994) (see especially Corollaries 1 and 2), this is true for most Gibbs samplers and Metropolis-Hastings algorithms.

While many Markov chains are Harris ergodic and converge eventually in total variation distance to the invariant distribution $\pi$, these results say nothing concerning the rate at which such convergence occurs. The characterization of this convergence rate is key to the assessment of Markov chain convergence in finite time, and thus to answering Question 1 and Question 2.

2.2 Geometric Ergodicity

This section introduces and discusses the results necessary to establish geometric ergodicity. A review of papers where geometric ergodicity is proved is then provided.

2.2.1 Introduction

One way to characterize the convergence rate of a Markov chain is to establish geometric ergodicity, which implies the chain converges to its stationary distribution at a geometric rate.

Definition 2.18. A Harris ergodic Markov chain is geometrically ergodic if for any $x \in X$, some function $M : X \to \mathbb{R}^+$, and some constant $0 < r < 1$

$$||K^n(x, \cdot) - \pi(\cdot)||_{TV} \leq M(x)r^n,$$

for any positive integer $n$.

Geometric ergodicity implies that the bound on the total variation distance between the $n$-step transition kernel and $\pi$ decreases with each additional iteration by a factor of $r$. A stronger but related form of ergodicity is uniform ergodicity.
Definition 2.19. If a Harris ergodic Markov chain satisfies (2.6) for some bounded function $M$, then the chain is \textit{uniformly ergodic}.

Equipped with the definitions of geometric and uniform ergodicity, methods for establishing these convergence rates may now be established.

2.2.2 Establishing Geometric and Uniform Ergodicity

While Harris ergodicity is very common among Markov chains, geometric ergodicity is far less common and is much more difficult to show. Several conditions describing the Markov chain $\{X_n\}$ which aid in showing geometric ergodicity are now defined.

Definition 2.20. A \textit{minorization condition} holds for a chain $\{X_n\}$ with transition kernel $K(\cdot, \cdot)$ if there exists some set $C \in B(X)$, $m \in \mathbb{N}^+$, $\epsilon > 0$, and a measure $Q$ such that

\[
K^m(x, A) \geq \epsilon Q(A), \quad \forall x \in C, A \in B(X). \tag{2.7}
\]

The set $C$ is called a \textit{small set}.

For a general definition of minorization and a more extensive discussion of the implications of minorization on Markov chain convergence, see Chapter 3.

Before the next definition, denote

\[
E(V(X_{n+1})|X_n = x) = \int_X V(y)K(x, dy).
\]

Also, recall the indicator function $1_A(x)$ defined in (2.3). Now, a drift condition is defined, along with a commonly used alternative definition.

Definition 2.21. Let $C$ be a small set as in Definition 2.20 and $V$ be a function

\[ V : X \to [1, \infty). \]
Also, let $0 < \lambda < 1$ and $b < \infty$. Then, a drift condition is satisfied if

$$E(V(X_{n+1})|X_n = x) \leq \lambda V(x) + \mathbb{1}_C(x)b, \quad \forall x \in X. \quad (2.8)$$

**Definition 2.22** (Rosenthal, 1995). Let $V$ be a function

$$V : X \rightarrow [0, \infty).$$

Also, let $0 < \lambda < 1$ and $b < \infty$. Then, a drift condition is satisfied if

$$E(V(X_{n+1})|X_n = x) \leq \lambda V(x) + b \quad \forall x \in X. \quad (2.9)$$

Jones and Hobert (2001a) commented that the drift condition (2.9) is easier to show than (2.8). Fortuitously, when needed, a conversion from (2.9) to (2.8) can be derived from the ideas in the proof of Lemma 15.2.8 in Meyn and Tweedie (1993). Specifically, assume drift condition (2.9) is satisfied with $\lambda = \lambda_2$, $b = b_2$, and $V = V_2$. Define $V_1(x) = V_2(x) + 1$. Thus $V_1 : X \rightarrow [1, \infty)$, and by assumption

$$E(V_1(X_{n+1})|X_n = x) = E(V_2(X_{n+1})|X_n = x) + 1$$

$$\leq \lambda_2 V_2(x) + 1 + b_2$$

$$= \lambda_2 V_1(x) + (1 - \lambda_2) + b_2$$

$$= \lambda_2 V_1(x) + b_1, \quad (2.10)$$

where $b_1 = (1 - \lambda_2) + b_2$. Further, let

$$C = \{ x \in X : V_1(x) \leq d \},$$

where $d \geq \frac{2b_1}{1 - \lambda_2}$. For $x \notin C$,

$$V_1(x) > d \geq \frac{2b_1}{1 - \lambda_2},$$

24
which implies
\[ b_1 < \frac{(1 - \lambda_2)}{2} V_1(x). \]

Then, with (2.10), when \( x \not\in C \)
\[
E(V_1(X_{n+1})|X_n = x) \leq \lambda_2 V_1(x) + b_1 \\
< \lambda_2 V_1(x) + \frac{1 - \lambda_2}{2} V_1(x) \\
= \frac{\lambda_2 + 1}{2} V_1(x) \\
= \lambda_1 V_1(x) + b_1 1_C(x), \quad \text{for } x \not\in C, \tag{2.11}
\]
where \( \lambda_1 = (\lambda_2 + 1)/2 \) and \( 1_C(x) = 0 \) because \( x \not\in C \). When \( x \in C \), (2.10) implies that
\[
E(V_1(X_{n+1})|X_n = x) \leq \lambda_2 V_1(x) + b_1 \\
\leq \lambda_1 V_1(x) + b_1 1_C(x), \quad \text{for } x \in C, \tag{2.12}
\]
which is true because \( \lambda_2 < \lambda_1 \) and \( 1_C(x) = 1 \) for \( x \in C \). Together, (2.11) and (2.12) imply that
\[
E(V_1(X_{n+1})|X_n = x) \leq \lambda_1 V_1(x) + 1_C(x)b_1, \quad \forall x \in X.
\]
Therefore, the drift condition in (2.8) is satisfied with \( V(x) = V_1(x), \lambda = \lambda_1 \), and \( b = b_1 \).

Another condition needed for many proofs of geometric ergodicity is now defined.

**Definition 2.23.** If \( K(\cdot, O) \) is a lower semi-continuous function for any open set \( O \in \mathcal{B}(X) \), then the chain corresponding to \( K \) is called a *Feller* chain.

This condition is routinely shown for a transition kernel \( K(x, \cdot) \) which has transition density \( k(x, y) \) with respect to Lebesgue measure and where \( k(x, y) \) is a continuous function in \( x \). Recall first Fatou’s Lemma which provides a needed inequality.
Lemma 2.24 (Fatou’s Lemma, Resnick, 1999). Consider the measurable probability space \((X, \mathcal{B}(X), P)\) and a sequence of nonnegative functions

\[ f_n : X \to [0, \infty), \quad \text{for } n = 1, 2, \ldots, \]

and define the function \(f : X \to [0, \infty)\) as

\[ f(x) = \liminf_{n \to \infty} f_n(x), \quad \forall x \in X. \]

Then, for all \(A \in \mathcal{B}(X)\),

\[
\int_A f \, dP = \int_A \liminf_{n \to \infty} f_n \, dP \leq \liminf_{n \to \infty} \int_A f_n \, dP. \tag{2.13}
\]

Fatou’s lemma may then be used in the proof of the following proposition.

Proposition 2.25. Let \(K\) be the kernel of a Markov chain with continuous density \(k\) with respect to Lebesgue measure \(\mu\). Let \(O \in \mathcal{B}(X)\) be any open set. Then, \(K(\cdot, O)\) is a lower semi-continuous function and hence the Markov chain corresponding to \(K\) is a Feller chain.

Proof. Recall that for a function \(f : X \to \mathbb{R}\), \(f\) is lower semi-continuous if

\[ \liminf_{x \to x_0} f(x) \geq f(x_0). \]

Let \(x_m\) be a sequence converging to \(x\) as \(m \to \infty\). Then

\[ \liminf_{m \to \infty} K(x_m, O) = \liminf_{m \to \infty} \int_O k(x_m, y) \, dy \]

by (2.1)

\[ \geq \int_O \liminf_{m \to \infty} k(x_m, y) \, dy \]

due to Fatou’s Lemma

\[ = \int_O k(x, y) \, dy \]

26
due to continuity of $k$

$$= K(x, O),$$

by (2.1), which is the desired inequality. Thus, $K(\cdot, O)$ is lower semi-continuous and
the Markov chain corresponding to $K$ is a Feller chain.

The next key definition is now provided.

**Definition 2.26.** A measurable function $V : X \to \mathbb{R}^+$ is **unbounded off compact sets** for $\{X_n\}$ if for any $d < \infty$ the set $C$ is compact where

$$C = \{x \in X : V(x) \leq d\}.$$

Equipped with these definitions, the two most common ways of establishing geometric ergodicity are now presented.

**Proposition 2.27** (Method 1 for showing geometric ergodicity). If $\{X_n\}$ is Harris ergodic and both the minorization condition (2.7) and drift condition (2.8) are satisfied, then the chain is geometrically ergodic (Meyn and Tweedie 1993, Theorem 15.0.1).

**Proposition 2.28** (Method 2 for showing geometric ergodicity). Let $\{X_n\}$ be Harris ergodic and Feller and assume $\psi$ (the maximal irreducibility measure) has nonempty interior. Then, if the drift condition (2.9) is satisfied and if $V$ is unbounded off compact sets for $\{X_n\}$, then $\{X_n\}$ is geometrically ergodic.

This result is stated concisely in Tan and Hobert (2009), but the proof is due to Meyn and Tweedie (1993) Lemma 15.2.8, Theorem 6.0.1, and Theorem 15.0.1.
Another approach for proving geometric ergodicity, discussed in Roberts and Rosenthal (1997), uses the spectral gap of $K$, which is now defined. First, given the kernel $K$, which is stationary for $\pi$, and $1 \geq p < \infty$, define the norm

$$||\mu||_{L^p(\pi)}^p = \begin{cases} \int_X \left(\frac{d\mu}{d\pi}\right)^p d\pi & \mu \ll \pi \\ \mu^+(X) + \mu^-(X), & p = 1 \\ \infty, & \text{otherwise.} \end{cases}$$

Further, let $L^p(\pi) = \{\mu : ||\mu||_{L^p(\pi)} < \infty\}$. Then $K$ has an $L^2(\pi)$ spectral gap if there exists $\rho < 1$ such that for each signed measure $\mu \in L^2(\pi)$ with $\mu(X) = 0$,

$$||\mu K(\cdot)||_{L^2(\pi)} \leq \rho ||\mu||_{L^2(\pi)}.$$

Then, $K$ is geometrically ergodic if and only if $K$ has an $L^2(\pi)$ spectral gap. See Roberts and Rosenthal (1997) for further details and discussion. See also Choi and Hobert (2013), Roy and Hobert (2007) and Kontoyiannis and Meyn (2012). This method is not pursued further in this work, though it is viable for establishing geometric ergodicity.

Having established these methods for proving geometric ergodicity, now consider how a Markov chain is shown to be uniformly ergodic.

**Proposition 2.29** (Method for showing uniform ergodicity). If the minorization condition (2.7) is satisfied for $C = X$, then the Markov chain is uniformly ergodic with

$$||K^n(x, \cdot) - \pi||_{TV} \leq (1 - \epsilon)^{\lfloor n/m \rfloor},$$

where $\lfloor n/m \rfloor$ is the largest integer smaller than $n/m$ (Rosenthal 1995).

Having now established definitions and methods for proving geometric ergodicity of Markov chains, a review of relevant literature utilizing these methods is now provided.
2.2.3 Examples in the Literature

As Markov chains have become increasingly important and well-utilized, studies of their convergence have also become increasingly prevalent. Many papers address convergence by formally establishing the geometric ergodicity of Markov chains for particular models. Examples utilizing Method 1 include Jones and Hobert (2001a), Marchev and Hobert (2004), and Khare and Hobert (2013). Examples using Method 2 include Hobert and Geyer (1998), Roy and Hobert (2007), Tan and Hobert (2009), Doss and Hobert (2010), Johnson and Jones (2010), Román and Hobert (2012), Khare and Hobert (2012), and Román and Hobert (2014).

Occasionally, one method is not considered in isolation. Roy and Hobert (2007) prove geometric ergodicity of one algorithm with Method 2, used the spectral radius to prove another Markov chain is also geometrically ergodic, and further develop minorization conditions, related to Method 1, for regenerative sampling.

Rather than focus on a particular model, some papers focus on establishing geometric ergodicity for general MCMC algorithms, such as Jarner and Hansen (2000) for random walk Metropolis, Mengersen and Tweedie (1996) for random walk and independence Metropolis-Hastings, Roberts and Polson (1994) for the Gibbs sampler, Johnson et al. (2013) for component-wise methods (e.g. random scan Gibbs), and Fort et al. (2003) and Jones et al. (2014) for conditional Metropolis-Hastings samplers.

Beyond establishing rates of convergence and geometric ergodicity, many methods have been introduced to improve the speed of convergence over traditionally used methods. Adaptive MCMC methods (see Andrieu and Thomas (2008), Roberts and Rosenthal (2009), and Latuszyński et al. (2013)), while often not Markovian and
thus requiring adapted theory, typically offer some improvement over traditional algorithms. Hamiltonian MCMC (Girolami and Calderhead 2011) is a progressing field of relatively fast-converging Markov chains. The partially collapsed Gibbs sampler (Van Dyk and Park 2008) also often speeds up the vanilla Gibbs sampler.

Several MCMC diagnostics have also been developed (see e.g. Gelman and Rubin (1992) and Raftery and Lewis (1992); Cowles and Carlin (1996) provides a relevant review), though many of these methods are ad-hoc and poorly converging chains may unfortunately be determined to converge well. Establishing geometric ergodicity or another convergence rate is strongly preferred when possible.

2.3 Bounds on Total Variation Distance

The definition of geometric ergodicity in (2.6) implies that for a given starting value \( x \), convergence rate \( r \), and number of iterations \( n \), an upper bound on the total variation distance may be computed. Two results that provide such an upper bound dominate the literature, the first by Rosenthal (1995) and the second by Roberts and Tweedie (1999), both of which are now provided.

**Theorem 2.30** (Rosenthal, 1995). Let \( \{X_n\} \) be a Markov chain with kernel \( K \), stationary for \( \pi \), which satisfies the drift condition (2.9) and minorization condition (2.7) for \( C = \{x : V(x) \leq d\} \) where \( d > 2b/(1 - \lambda) \). Let \( X_0 \sim \nu \) and define

\[
\alpha = \frac{1 + d}{1 + 2b + \lambda d} > 1 \quad \text{and} \quad A = 1 + 2(\lambda d + b).
\]

Then, for \( 0 < r < 1 \),

\[
||K^n(x_0, \cdot) - \pi(\cdot)||_{TV} \leq (1 - \epsilon)^{rn} + \left(\frac{A^r}{\alpha^{1-r}}\right)^n \left(1 + \frac{b}{1 - \lambda} + E_{\nu}(V(X_0))\right). \quad (2.15)
\]
Theorem 2.31 (Roberts and Tweedie, 1999). Let \( \{X_n\} \) be a Markov chain with kernel \( K \), stationary for \( \pi \), which satisfies the drift condition (2.8) and associated minorization condition (2.7). Let \( X_0 \sim \nu \) and define

\[
A = \sup_{x \in C} \mathbb{E}[V(X_{i+1}) | X_i = x], \quad \text{and} \quad J = (A - \epsilon) / \lambda,
\]

with

\[
\beta^* = \begin{cases} 
\lambda^{-1} & \text{if } J < 1 \\
\exp \left\{ \frac{\log \lambda \log(1 - \epsilon)}{\log J - \log(1 - \epsilon)} \right\} \leq \lambda^{-1} & \text{if } J \geq 1,
\end{cases}
\]

(2.16)

where \( \epsilon, \lambda \), and the set \( C \) are all provided by the drift and minorization conditions. Choose \( \beta \in (1, \beta^*) \) and define \( \phi(\beta) = \log \beta / \log \lambda^{-1} \). Then

\[
||K^n(x_0, \cdot) - \pi(\cdot)||_{TV} \leq [\mathbb{E}_\nu(V(X_0))]^{\phi(\beta)} \left[ \frac{1 - \beta(1 - \epsilon)}{1 - (1 - \epsilon)(J/(1 - \epsilon))^{\phi(\beta)}} \right] \beta^{-n}. \tag{2.17}
\]

If \( \mathbb{E}_\nu(V(X_0)) \) is difficult to compute, Roberts and Rosenthal (2004) showed that it may be upper bounded by \( b/(\epsilon(1 - \lambda)) \).

Both Theorems 2.30 and 2.31 may then be used to compute the number of iterations required for the total variation distance to be below some threshold. Roberts and Tweedie (1999) showed their bounds are better as \( n \to \infty \), but perhaps the most important difference is the form of the drift condition required for each, though the previously discussed conversion from (2.9) to (2.8) may be used as desired.

### 2.4 Markov Chain Central Limit Theorems

Beyond computable bounds on the total variation distance, one benefit of geometrically ergodic Markov chains is that a central limit theorem (CLT) often holds for ergodic averages. This central limit theorem ensures that

\[
\sqrt{n}(\bar{g}_n - \mathbb{E}_\pi g) \overset{d}{\to} N(0, \sigma_g^2), \tag{2.18}
\]

31
where $g$ is a measurable function, $\overset{d}{\to}$ is convergence in distribution or law,

$$\bar{g}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i),$$

and

$$\sigma_g^2 = \text{var}_\pi\{g(X_0)\} + 2 \sum_{i=1}^{\infty} \text{cov}_\pi\{g(X_0), g(X_i)\} < \infty.$$ 

The specific conditions under which such a CLT exists for a Markov chain are now discussed, followed by a discussion about how to estimate $\sigma_g^2$.

### 2.4.1 Establishing Existence

Jones (2004) provide a theorem which outlines the circumstances under which a Markov chain central limit theorem exists. First, note that a Markov chain has polynomial ergodicity of order $m$ when

$$||K^n(x, \cdot) - \pi(\cdot)|| \leq M(x)n^{-m}, \quad (2.19)$$

where $m \geq 0$. See Jarner and Roberts (2002) for additional details about this type of ergodicity.

**Theorem 2.32** (Jones, 2004). Let $\{X_n\}$ be a positive Harris ergodic Markov chain on $X$ stationary for $\pi$ and let $g : X \to \mathbb{R}$ be a measurable function. Assume one of the following holds:

1. $\{X_n\}$ is polynomially ergodic of order $m > 1$, $\text{E}_\pi M < \infty$ where $M$ is from (2.19) and there exists $B < \infty$ such that $|g(x)| < B$ almost surely;

2. $\{X_n\}$ is polynomially ergodic of order $m$, $\text{E}_\pi M < \infty$ and $\text{E}_\pi |g(x)|^{2+\delta} < \infty$ where $m\delta > 2 + \delta$;

32
3. \( \{X_n\} \) is geometrically ergodic and \( E_\pi |g(x)|^{2+\delta} < \infty \) for some \( \delta > 0 \);

4. \( \{X_n\} \) is geometrically ergodic and \( E_\pi (g^2(x)(\log^+ |g(x)|)) < \infty \);

5. \( \{X_n\} \) is geometrically ergodic, satisfies detailed balance, and \( E_\pi g^2(x) < \infty \), or

6. \( \{X_n\} \) is uniformly ergodic and \( E_\pi g^2(x) < \infty \).

Then, the Markov chain CLT (2.18) is satisfied for any initial distribution.

When a central limit theorem applies, an interval for the desired estimate may be computed as discussed by Flegal and Jones (2010):

\[
\bar{g}_n \pm t^* \frac{\hat{\sigma}_g}{\sqrt{n}},
\]

where \( t^* \) is the appropriate Student’s \( t \) distribution quantile. This provides a rigorous answer to Question 2 about the accuracy of the estimate \( \bar{g}_n \). Typically, \( \sigma_g^2 \) is unknown and is therefore replaced by a consistent estimator, \( \hat{\sigma}_g^2 \); details are provided for one such estimator in the next section. Given the estimator, there are two approaches, discussed by Flegal and Jones (2010), for answering Question 1 about whether the simulation was run long enough.

One approach is the fixed-time method where the Markov chain is run for a specified number of iterations. The interval (2.20) is then computed. If the interval is larger than desired, choose a larger fixed-time until the interval provides satisfactory results.

The adaptive process of increasing the fixed-time until a certain width is achieved introduces the fixed-width approach. With this approach, the half-width of the interval, \( \epsilon \), is specified. Simulation then continues until

\[
t^* \frac{\hat{\sigma}_n}{\sqrt{n}} + p(n) < \epsilon,
\]

(2.21)
where \( p(n) \) is positive and \( p(n) = o(n^{-1/2}) \) as \( n \to \infty \). The purpose of \( p(n) \) is to ensure the simulation effort is large enough. Flegal and Jones (2010) suggest

\[
p(n) = \epsilon \mathbb{1}(n \leq n^*) + n^{-1}
\]

is a useful definition of \( p(n) \) where \( n^* \) is the minimum number of iterations desired. Either approach is reasonable and can lead to satisfactory answers to both Question 1 and Question 2.

2.4.2 Estimation of Asymptotic Variance

In order to implement the above-discussed approaches for estimating \( E_{\pi}g \), an estimate of \( \sigma_g^2 \) is required. As MCMC produces results which are autocorrelated, typically \( \sigma_g^2 \not= \text{Var}_x g \); thus, the simple MCMC sample variance is not a good estimator. Several methods for estimating \( \sigma_g^2 \) include regenerative sampling (see Ripley 1987; Bratley et al. 1987; Mykland et al. 1995; Hobert et al. 2002), non-overlapping batch means (Jones et al. 2006), overlapping batch means (Flegal and Jones 2010), and spectral estimation (Flegal and Jones 2010). Due to ease of implementation and smaller asymptotic variance over non-overlapping batch means, the method of overlapping batch means is now discussed and will be used for the examples within this work.

This method is referenced from Flegal and Jones (2010) who improve the theory regarding overlapping batch means beyond previous work. Consider a simulation of length \( n \). Choose a batch size \( b_n \), guidance for which will later be provided. There are then \( n - b_n + 1 \) batches of length \( n \), each of which may be used to estimate \( E_{\pi}g \) using

\[
\overline{g}_j(b_n) = b_n^{-1} \sum_{i=1}^{b_n} g(X_{j+i}),
\]
where $j$ indexes the batch with $j = 0, \ldots, n - b_n$. The overlapping batch means estimate for $\sigma_g^2$ is then

$$\tilde{\sigma}_{OBM}^2 = \frac{n b_n}{(n - b_n)(n - b_n + 1)} \sum_{j=0}^{n-b_n} (\bar{y}_j(b_n) - \bar{y}_n)^2. \quad (2.22)$$

The following theorem, presented as Corollary 1 in Flegal and Jones (2010), outlines the conditions under which $\tilde{\sigma}_{OBM}^2$ is strongly consistent.

**Theorem 2.33** (Flegal and Jones, 2010). Let $\{X_t\}$ be a geometrically ergodic Markov chain stationary for $\pi$ which satisfies a general minorization condition (see Chapter 3, (3.1)). Assume $g : X \to \mathbb{R}$ is a Borel function with $E^\pi |g|^{2+\delta+\epsilon} < \infty$ where $\delta, \epsilon > 0$. Let $b_n = \lfloor \nu r \rfloor$ with $3/4 > \nu > (1 + \delta/2)^{-1}$. Then $\tilde{\sigma}_{OBM}^2 \to \sigma_g^2$ with probability 1.

Therefore, with a sufficient moment condition and a carefully chosen batch size, $\tilde{\sigma}_{OBM}^2$ is strongly consistent. Thus, the Markov chain central limit theorem may be applied to answer Question 1 and Question 2. Note that the appropriate degrees of freedom for the $t$ distribution in the interval when using the overlapping batch means estimate for $\sigma_g^2$ is $n - b_n$ (Flegal and Jones 2011).

### 2.4.3 Estimation of Quantiles

Often, quantiles of a marginal distribution are desired for inference. For example, a 95% credible interval (see Definition 5.1 for a formal definition and discussion) is often reported for parameters, which entails providing the 0.025 and 0.975 quantiles. As with averages, and perhaps even more often, these estimates are reported without any sense of the Monte Carlo error surrounding them. Doss et al. (2014) discuss a central limit theorem which applies to quantiles and allows for the error about the estimate to be computed and reported.
First, define the $q$ quantile as follows. Let $W \sim \pi$ and set $Y = g(W)$ where $g : X \to \mathbb{R}$ is a measurable function (as in the previous discussion). Then, define

$$
\xi_q := F_Y^{-1}(q) = \inf\{\theta : F_Y(\theta) \geq q\},
$$

where $F_Y$ is the distribution function of $Y$. The following theorem then applies.

**Theorem 2.34** (Doss et al., 2014). If \{$X_n$\} is polynomially ergodic of order $m > 1$ and if $\sigma^2(\xi_q) > 0$, then, as $n \to \infty$,

$$
\sqrt{n}(\widehat{\xi}_{n,q} - \xi_q) \xrightarrow{d} N(0, \gamma^2(\xi_q)),
$$

(2.23)

where

$$
\gamma^2(\xi_q) = \sigma^2(\xi_q)/[f_Y(\xi_q)]^2,
$$

where $f_Y(\xi_q)$ is the density function corresponding to $F_Y$.

Note that, as geometric convergence is faster than polynomial convergence, if \{$X_n$\} is geometrically ergodic, then it is also polynomially ergodic of order larger than 1, and the theorem applies. Note that relatively few restrictions are required for this CLT to exist, which is a desirable property. Indeed, if moment conditions are not readily available for the mean of a particular distribution, perhaps the median should be used as the conditions for establishment of the corresponding central limit theorem are more relaxed.

It remains, of course, to estimate $\gamma^2(\xi_q)$ in order for this central limit theorem to be used. While multiple strategies are possible, consider the subsampling bootstrap, which is found in Politis et al. (1999) and discussed with respect to Markov chains in Flegal (2012) and Flegal and Jones (2011).
As with overlapping batch means, divide the \( n \) iterations into batches, or subsamples, of \( b_n \) iterations. Let \( Y_i = g(X_i) \). Each batch \( j = 1, \ldots, n - b_n + 1 \), consists of the observations \( \{Y_j, \ldots, Y_{j+b_n-1}\} \), which may be ordered as \( \{Y_{(1)}, \ldots, Y_{(b_n)}\} \). The sample quantile for the \( j^{th} \) batch is then defined as

\[
\xi_j^* = Y_{(i)}, \quad \text{where } i - 1 < b_nq \leq i.
\]

Then, define the subsampling bootstrap method (SBM) estimator of \( \gamma^2(\xi_q) \) as

\[
\widehat{\gamma}^2_{SBM} = \frac{b_n}{n - b_n + 1} \sum_{j=1}^{n-b_n+1} (\xi_j^* - \bar{\xi}^*)^2,
\]

where

\[
\bar{\xi}^* = \frac{1}{n - b_n + 1} \sum_{j=1}^{n-b_n+1} \xi_j^*.
\]

Note that \( b_n \) must be chosen so \( b_n \to \infty \) and \( b_n/n \to 0 \) as \( n \to \infty \) (Flegal and Jones 2011). A common choice is \( b_n = \lfloor \sqrt{n} \rfloor \). Then, the following interval can be constructed for the sample quantile \( \widehat{\xi}_{n,q} \) in order to answer Question 2:

\[
\widehat{\xi}_{n,q} \pm t^* \frac{\widehat{\gamma}^2_{SBM}}{\sqrt{n}},
\]

where \( t^* \) is the desired quantile from the Student’s \( t \)-distribution with \( n - b_n \) degrees of freedom.

With these tools in place, including bounds for total variation distance and Markov chain central limit theorems, the theory presented in this chapter may now be applied to establish geometric ergodicity and answer Question 1 and Question 2 for two illustrative examples.

### 2.5 Examples of Geometrically Ergodic Markov Chains

As the Gibbs sampler and Metropolis-Hastings algorithms are the most widely-used MCMC algorithms, illustrations of proving geometric ergodicity are provided
for each. In these examples, the number of iterations resulting in TV distance less than 0.1 is computed, answering Question 1. Monte Carlo standard errors are also computed for parameter estimates, answering Question 2.

2.5.1 Bayesian One-Way Random Effects Model using a Gibbs Sampler

For this example, consider the Bayesian one-way random effects model where for each individual \(i = 1, \ldots, q\) there are \(n_i\) observations, indexed by \(j\), of some continuous measurement. For convenience, let

\[
n^* = \max_{i \in \{1, \ldots, q\}} n_i, \quad \text{and} \quad n_* = \min_{i \in \{1, \ldots, q\}} n_i.\]

Denote \(N = \sum_{i=1}^{q} n_i\). Let the model for the data be

\[
y_{ij} \sim \text{Normal}(\theta_i, \sigma_e^2) \quad i = 1, \ldots, q, \quad j = 1, \ldots, n_j.
\]

The prior and hyperprior distributions are

\[
\theta_i \sim \text{Normal}(\mu, \sigma_{\theta}^2), \quad i = 1, \ldots, q
\]

\[
\mu \sim \text{Normal}(\mu_0, \sigma_{\mu}^2),
\]

\[
\sigma_e^2 \sim \text{Inverse Gamma}(a_e, b_e),
\]

and

\[
\sigma_{\theta}^2 \sim \text{Inverse Gamma}(a_\theta, b_\theta),
\]

where \(\mu_0, \sigma_{\mu}^2, a_e, b_e, a_\theta, \) and \(b_\theta\) are all specified constants. Let \(\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T\) and let \(\mathbf{y}\) be a vector containing the entirety of the data. The posterior density \(\pi_1\) is the target distribution of interest and has the form

\[
\pi_1(\boldsymbol{\theta}, \mu, \sigma_e^2, \sigma_{\theta}^2 | \mathbf{y}) \propto \left(\sigma_e^2\right)^{-N/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \frac{(y_{ij} - \theta_i)^2}{\sigma_e^2} \right)
\]

38
\[
\times (\sigma_\theta^2)^{-q/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \frac{(\theta_i - \mu)^2}{\sigma_\theta^2} \right) \\
\times \exp \left( -\frac{1}{2} \frac{(\mu - \mu_0)^2}{\sigma_\mu^2} \right) \\
\times (\sigma_e^2)^{-a_e - 1} \exp \left( -\frac{b_e}{\sigma_e^2} \right) (\sigma_\theta^2)^{-a_\theta - 1} \exp \left( -\frac{b_\theta}{\sigma_\theta^2} \right).
\]

While algorithms for this model and variations thereof have been proven geometrically ergodic and utilized by others (see e.g. Hobert and Geyer (1998), Jones and Hobert (2001a), Jones and Hobert (2001b), Jones and Hobert (2004), Tan and Hobert (2009), and Flegal and Herbei (2012)) the approach taken here differs somewhat.

For ease of development, it is assumed that \( \mu_0 \) is equal to 0. If not, transform the data and parameters. Specifically, let

\[
y_{ij}^* = y_{ij} - \mu_0,
\]

\[
\theta_i^* = \theta_i - \mu_0,
\]

and

\[
\mu^* = \mu - \mu_0.
\]

Then, the new model is

\[
y_{ij}^* \sim \text{Normal}(\theta_i^*, \sigma_e^2),
\]

\[
\theta_i^* \sim \text{Normal}(\mu^*, \sigma_\theta^2),
\]

and

\[
\mu^* \sim \text{Normal}(0, \sigma_\mu^2),
\]

and the posterior distribution \( \pi_1^* \) is equivalent to \( \pi_1 \):

\[
\pi_1^*(\theta^*, \mu^*, \sigma_e^2, \sigma_\theta^2 | y) \propto (\sigma_e^2)^{-N/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \sum_{j=1}^{n_i} \frac{(y_{ij}^* - \theta_i^*)^2}{\sigma_e^2} \right)
\]
\[
\propto \pi_1(\theta, \mu, \sigma^2_e, \sigma^2_\theta|y).
\]
Then, when draws from \(\pi_i^*\) are obtained, the draws for \(\theta^*\) and \(\mu^*\) can be transformed into draws from \(\theta\) and \(\mu\) by simply adding \(\mu_0\). Thus, without loss of generality, assume \(\mu_0 = 0\).

The Markov chain employed to sample from \(\pi_1\) is a Gibbs sampler. The parameters are divided into two blocks \((\theta, \mu)\) and \((\sigma^2_e, \sigma^2_\theta)\). The algorithm first updates \((\sigma^2_e, \sigma^2_\theta)\) independently conditional on \((\theta, \mu)\). Then \((\theta, \mu)\) is updated conditional on \((\sigma^2_e, \sigma^2_\theta)\) by first updating \(\theta\) with \(\mu\) integrated out, followed by an update of \(\mu\) given \(\theta\). Conditionality on the data is implicitly presumed throughout. To build this algorithm, the conditional densities must be derived. The updates for \(\sigma^2_e\) and \(\sigma^2_\theta\) can be attained by quick inspection of the posterior distribution. The updates for \(\mu\) and \(\theta\) are not as obvious, and are thus derived here. First, notation must be defined.

The typical representation
\[
SSE = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2,
\]
will be used throughout. Note that
\[
\sum_{i=1}^{q} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 = \sum_{i=1}^{q} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + (\theta_i - \bar{y}_i)^2 = SSE + \sum_{i=1}^{q} n_i (\theta_i - \bar{y}_i)^2,
\]
because \(\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) = 0\). Also, denote
\[
\bar{Y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_q)^T, \quad \text{and} \quad N = \text{diag}(n_1, \ldots, n_q),
\]
where \(\text{diag}(\cdot)\) creates a diagonal matrix with its arguments as the diagonal elements.

Let \(I\) be the \(q \times q\) identity matrix with \(1\) defined to be the \(q \times 1\) vector of ones.
Finally, let
\[ \bar{\theta} = \frac{1}{q} \sum_{i=1}^{q} \theta_i = \frac{\theta^T 1}{q}, \]
which aids in development.

The following are additional results which are used in the development of the proofs for this example. These may also be referenced in Chapter 4, and where proof is required, it is provided in Chapter 4.

**Definition 2.35** *(Horn and Johnson, 2013, Definition 7.7.1).* Let \( A \) and \( B \) be \( n \times n \) matrices. Then, write

1. \( A \geq 0 \) if \( A \) is nonnegative definite.

2. \( A > 0 \) if \( A \) is positive definite.

3. \( A \geq B \) if \( A - B \) is nonnegative definite.

4. \( A > B \) if \( A - B \) is positive definite.

**Result 2.36** *(Horn and Johnson, 2013, Corollary 7.7.4).* Let \( A, B, \) and \( C \) be \( n \times n \) matrices.

(1) If \( A > 0 \) and \( B > 0 \) then \( A \geq B \) if and only if \( B^{-1} \geq A^{-1} \).

(2) If \( A > 0 \) and \( B \geq 0 \). Then \( A^{-1} \geq (A + B)^{-1} > 0 \).

(3) If \( A \geq B \) then \( \text{tr}(A) \geq \text{tr}(B) \) with equality if and only if \( A = B \).

(4) If \( B > 0 \) and \( C \geq A \), then \( \text{tr}(AB) \leq \text{tr}(CB) \).

**Result 2.37** *(Sherman-Morrison-Woodbury Formula, Horn and Johnson, 2013).* Let \( A \) and \( R \) be invertible matrices of dimensions \( n \times n \) and \( r \times r \) respectively. Let \( Y \) be
If $A + XRY$ and $R^{-1} + YA^{-1}X$ are invertible, then

$$(A + XRY)^{-1} = A^{-1} - A^{-1}X(R^{-1} + YA^{-1}X)^{-1}YA^{-1}.$$  

**Result 2.38** (Iterative Expectations, Resnick, 1999). Let $X$ and $Y$ be random variables. Then,

$$E(X) = E(E(X|Y)).$$

**Result 2.39.** Let $X$ be a random vector with mean $\mu$ and covariance matrix $\Sigma$. Also, let $c$ be a constant vector and $A$ be a constant matrix. Then,

$$E((X - c)^T A (X - c)) = tr(A \Sigma) + (c - \mu)^T A (c - \mu).$$

**Result 2.40** (Expected Value of Inverse Gamma Distribution). Let $X$ have an inverse gamma distribution with shape $a$ and scale $b$ which corresponds to the probability density function

$$f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp \left( -\frac{b}{x} \right).$$

Then

$$E(X) = \frac{b}{a - 1}, \quad a > 1,$$

$$\text{Var}(X) = \frac{b^2}{(a - 1)^2 (a - 2)}, \quad a > 2,$$

and

$$E(X^{-1}) = \frac{a}{b}.$$  

Now, proceed with the derivation of the conditional distributions of $\theta$ and $\mu$. Begin with $\pi_1$ and take everything not involving $\mu$ or $\theta$ as constant. Then, $\mu$ is integrated out to find the update of $\theta$:

$$\pi_1(\theta|\sigma^2_e, \sigma^2_\theta, y) \propto \int_{\mu} \pi_1(\theta, \mu|\sigma^2_e, \sigma^2_\theta, y) \, d\mu$$
\[
\alpha \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \frac{n_i(\bar{y}_i - \theta_i)^2}{\sigma^2_e} \right) \int_{\mu} \exp \left( -\frac{1}{2} \left[ \sum_{i=1}^{q} \frac{(\theta_i - \mu)^2}{\sigma^2_{\theta}} + \frac{\mu^2}{\sigma^2_{\mu}} \right] \right) d\mu
\]

\[
= \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \frac{n_i(\bar{y}_i - \theta_i)^2 + \beta^2_i}{\sigma^2_e} \right) \int_{\mu} \exp \left( -\frac{1}{2} \left[ \frac{\mu^2 q - 2\mu \bar{\theta} q + \mu^2}{\sigma^2_{\theta}} + \frac{\mu^2}{\sigma^2_{\mu}} \right] \right) d\mu,
\]

by multiplying out the cross product and moving the term without \(\mu\) to the expression before the integral. This indicates that the update for \(\mu\) is

\[
\mu | \theta, \sigma^2_{\theta} \sim \text{Normal} \left( \frac{\bar{\theta} q}{\sigma^2_{\theta}} \left( \frac{1}{\sigma^2_{\mu}} + \frac{q}{\sigma^2_{\theta}} \right)^{-1}, \left( \frac{1}{\sigma^2_{\mu}} + \frac{q}{\sigma^2_{\theta}} \right)^{-1} \right)
\]

which yields, returning to the inequality,

\[
\alpha \exp \left( -\frac{1}{2} \sum_{i=1}^{q} \frac{n_i(\bar{y}_i - \theta_i)^2 + \beta^2_i}{\sigma^2_e} \right) \exp \left( \frac{1}{2} \frac{\bar{\theta} q}{\sigma^2_{\theta}} \left( \frac{1}{\sigma^2_{\mu}} + \frac{q}{\sigma^2_{\theta}} \right)^{-1} \bar{\theta} q \right)
\]

by completing the square to make the appropriate normal distribution kernel for \(\mu\) which then integrates to 1,

\[
= \exp \left( -\frac{1}{2} \left[ \frac{(Y - \theta)^T N(Y - \theta)}{\sigma^2_e} + \frac{\theta^T \theta}{\sigma^2_{\theta}} - \theta^T 1 \left( \frac{1}{\sigma^2_{\mu}} + \frac{1^T 1}{\sigma^2_{\theta}} \right)^{-1} 1^T \theta \right] \right)
\]

by rewriting the expressions in terms of matrices

\[
\alpha \exp \left( -\frac{1}{2} \left[ \frac{\theta^T N \theta - 2 \theta^T N Y}{\sigma^2_e} + \frac{\theta^T \theta}{\sigma^2_{\theta}} - \theta^T 1 \left( \frac{1}{\sigma^2_{\mu}} + \frac{1^T 1}{\sigma^2_{\theta}} \right)^{-1} 1^T \theta \right] \right)
\]

by multiplying out the cross product and removing the term constant in \(\theta\)

\[
= \exp \left( -\frac{1}{2} \left[ \theta^T N \sigma^{-2}_e \theta - 2 \theta^T N \sigma^{-2}_e Y + \theta(\sigma^2_{\theta} I + \sigma^2_{\mu} 11^T)^{-1} \theta \right] \right),
\]

which is true by Result 2.37. This implies that the update for \(\theta\) is Multivariate Normal (MVN),

\[
\theta | \sigma^2_{\theta}, \sigma^2_e \sim \text{MVN}(\Sigma_{\theta} N \sigma^{-2}_e Y, \Sigma_{\theta}),
\]

where

\[
\Sigma_{\theta} = \left( (\sigma^2_{\theta} I + \sigma^2_{\mu} 11^T)^{-1} + N \sigma^{-2}_e \right)^{-1}.
\]
Using these updates, the following algorithm results.

**Algorithm 2.3** (Gibbs Sampler for the Hierarchical Random Effects Model)

Given $\theta', \mu', \sigma_\theta^2$, and $\sigma_e^2$, update to $\theta$, $\mu$, $\sigma_\theta^2$, and $\sigma_e^2$ sequentially as follows.

(i) Update $\sigma_e^2|\theta' \sim$ Inverse Gamma $(a_e + N/2, \sum_{i=1}^q n_i(\bar{y}_i - \theta'_i)^2/2 + SSE/2 + b_e)$

(ii) Update $\sigma_\theta^2|\theta', \mu' \sim$ Inverse Gamma $(a_\theta + q/2, \sum_{i=1}^q (\theta'_i - \mu')^2/2 + b_\theta)$

(iii) Update $\theta|\sigma_e^2, \sigma_\theta^2 \sim$ MVN($\Sigma_{\theta N N^{-1} \Sigma_{\theta}^{-2}}$, $\Sigma_{\theta}$), where

$$\Sigma_{\theta} = \left((\sigma_\theta^2 I + \sigma_\mu^2 11^T)^{-1} + N\sigma_e^{-2}\right)^{-1}.$$  

(iv) Update $\mu|\theta, \sigma_e^2, \sigma_\theta^2 \sim$ Normal $\left(\frac{\bar{\theta}_q}{\sigma_\theta^2} \left(\frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2}\right)^{-1}, \left(\frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2}\right)^{-1}\right)$.

This algorithm is now shown to be geometrically ergodic when certain conditions on the prior distributions are satisfied.

**Proposition 2.41.** The Markov chain stationary for $\pi_1$ induced by Algorithm 2.3 is geometrically ergodic if

$$2a_\theta + q - 2 > 1$$

and

$$2a_e + N - 2 > \text{tr}(N^{-1}) + q.$$  

**Proof.** This proposition may be proved using either Method 1 or Method 2 for proving chains are geometrically ergodic. In either case, begin by establishing the drift condition (2.9). Consider the candidate drift function

$$V(\theta, \mu) = \sum_{i=1}^q (\theta_i - \mu)^2 + \sum_{i=1}^q n_i(\bar{y}_i - \theta_i)^2.$$
The following conditional expectations and associated inequalities will be required for the proof. Recall that expectations are with respect to the transition density. The expectation of the first part of the drift function with respect to the conditional update of \( \mu \) is

\[
E\left( \sum_{i=1}^{q} (\theta_i - \mu)^2 \left| \theta, \sigma_e^2, \sigma_\theta^2 \right. \right)
\]

\[
= E\left( \theta^T \theta - 2\theta^T \mu \left| \theta, \sigma_e^2, \sigma_\theta^2 \right. \right)
\]

\[
= \theta^T \theta + E(q \mu^2 - 2\theta^T \mu | \theta, \sigma_e^2, \sigma_\theta^2)
\]

\[
= \theta^T \theta + q \frac{\theta^T 1}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-2} \frac{1}{\sigma_\theta^2} \frac{1}{\sigma_\theta^2} - 2 \theta^T 1 \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \frac{1}{\sigma_\theta^2}
\]

because

\[
E(\mu^2 | \theta, \sigma_e^2, \sigma_\theta^2) = E(\mu | \theta, \sigma_e^2, \sigma_\theta^2)^2 + \text{Var}(\mu | \theta, \sigma_e^2, \sigma_\theta^2)
\]

\[
= \frac{\theta^T 1}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-2} \frac{1}{\sigma_\theta^2} \frac{1}{\sigma_\theta^2} + \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1}
\]

which, continuing,

\[
\leq \theta^T \theta + \theta^T 1 \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \frac{1}{\sigma_\theta^2} \frac{1}{\sigma_\theta^2} + q \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} - 2 \theta^T 1 \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \frac{1}{\sigma_\theta^2}
\]

because

\[
\left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \leq \frac{\sigma_\theta^2}{q}
\]

\[
= \theta^T \theta - \theta^T 1 \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \frac{1}{\sigma_\theta^2} \frac{1}{\sigma_\theta^2} + q \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1}
\]

by combining like terms

\[
= \sigma_\theta^2 \left[ \frac{\theta^T \theta}{\sigma_\theta^2} - \frac{\theta^T 1}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{1}{\sigma_\theta^2} \right)^{-1} \frac{1}{\sigma_\theta^2} \frac{1}{\sigma_\theta^2} \right] + q \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1}
\]

by factoring out \( \sigma_\theta^2 \) from the first two terms

\[
= \sigma_\theta^2 \theta^T \left( \frac{\sigma_\theta^2 1 + \sigma_\mu^2 \theta^T}{\sigma_\theta^2} \right)^{-1} + q \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1}
\]
by Result 2.37

\[ \leq \theta^\top \theta + \sigma^2_\theta, \] (2.24)

because Result 2.36 implies that

\[ (\sigma^2_\theta I + \sigma^2_\mu \mathbf{1}\mathbf{1}^\top)^{-1} \preceq I \sigma^{-2}_\theta, \]

and again because

\[ \left( \frac{1}{\sigma^2_\mu} + \frac{q}{\sigma^2_\theta} \right)^{-1} \leq \frac{\sigma^2_\theta}{q}. \]

Further, the conditional expectation

\[ \text{E}(\theta^\top \theta | \sigma^2_e, \sigma^2_\theta) = \mathbf{Y}^\top \mathbf{N}^\top \sigma^{-2}_e \Sigma^\top_\theta \Sigma \sigma^{-2}_e \mathbf{N} \mathbf{Y} + \text{tr}(\Sigma_\theta) \]

because

\[ \text{E}(\theta^\top \theta | \sigma^2_e, \sigma^2_\theta) = \text{E}(\theta | \sigma^2_e, \sigma^2_\theta)^\top \text{E}(\theta | \sigma^2_e, \sigma^2_\theta) + \text{tr}(\text{Var}(\theta | \sigma^2_e, \sigma^2_\theta)) \]

\[ = \mathbf{Y}^\top \mathbf{N}^\top \sigma^{-2}_e \Sigma^\frac{1}{2}_\theta \Sigma^\frac{1}{2} \Sigma \sigma^{-2}_e \mathbf{N} \mathbf{Y} + \text{tr}(\Sigma_\theta) \]

because \( \Sigma_\theta \) is symmetric

\[ = \mathbf{Y}^\top \mathbf{N}^\top \sigma^{-2}_e \Sigma^\frac{1}{2}_\theta \Sigma^\frac{1}{2} \Sigma \sigma^{-2}_e \mathbf{N} \mathbf{Y} + \text{tr}(\Sigma_\theta) \]

\[ \leq \mathbf{Y}^\top \mathbf{N}^\top \sigma^{-2}_e \Sigma^\frac{1}{2}_\theta \Sigma^\frac{1}{2} \Sigma \sigma^{-2}_e \mathbf{N} \mathbf{Y} + \text{tr}(\Sigma_\theta) \]

because

\[ \Sigma_\theta = ((\sigma^2_\theta I + \sigma^2_\mu \mathbf{1}\mathbf{1}^\top)^{-1} + \mathbf{N} \sigma^{-2}_e)^{-1} \preceq \mathbf{N}^{-1} \sigma^{-2}_e \]

by Result 2.36,

\[ \leq n^{-1}_s \sigma^{-2}_e \mathbf{Y}^\top \mathbf{N}^\top \Sigma_\theta \mathbf{Y} + \text{tr}(\Sigma_\theta) \]

because \( \mathbf{N}^{-1} \leq n^{-1}_s \mathbf{I} \)

\[ \leq n^{-1}_s \sigma^{-2}_e \mathbf{Y}^\top \mathbf{N}^\top \mathbf{N}^{-1} \sigma^{-2}_e \mathbf{N} \mathbf{Y} + \sigma^2_e \text{tr}(\mathbf{N}^{-1}) \]
again due to $\Sigma_\theta \leq N^{-1}\sigma_e^2$

$$\leq n_*^{-1}\overline{Y}N\overline{Y} + \sigma_e^2 \text{tr}(N^{-1}). \quad (2.25)$$

Additionally, the expectation of the next term of the drift function is

$$E\left(\sum_{i=1}^{q} n_i(y_i - \theta_i)^2 \bigg| \sigma_e^2, \sigma_\theta^2\right) = E((\theta - \overline{Y})^T N(\theta - \overline{Y})|\sigma_e^2, \sigma_\theta^2)$$

by writing the expression in matrix form

$$= (\Sigma_\theta N\sigma_e^{-2}\overline{Y} - \overline{Y})^T N(\Sigma_\theta N\sigma_e^{-2}\overline{Y} - \overline{Y}) + \text{tr}(N\Sigma_\theta)$$

by Result 2.39

$$= \overline{Y}^T (\Sigma_\theta N\sigma_e^{-2} - I)^T N(\Sigma_\theta N\sigma_e^{-2} - I) \overline{Y} + \text{tr}(N\Sigma_\theta). \quad (2.26)$$

Now, note that

$$\Sigma_\theta N\sigma_e^{-2} = ((\sigma_\theta^2 I + \sigma_\mu^2 11^T)^{-1} + N\sigma_e^{-2})^{-1} N\sigma_e^{-2}$$

$$= (N^{-1}\sigma_e^2 - N^{-1}\sigma_e^2(\sigma_\theta^2 I + \sigma_\mu^2 11^T + N^{-1}\sigma_e^2)^{-1} N^{-1}\sigma_e^2) N\sigma_e^{-2}$$

by Result 2.37

$$= I - N^{-1}\sigma_e^2(\sigma_\theta^2 I + \sigma_\mu^2 11^T + N^{-1}\sigma_e^2)^{-1}.$$

This implies that

$$\Sigma_\theta N\sigma_e^{-2} - I = -N^{-1}\sigma_e^2(\sigma_\theta^2 I + \sigma_\mu^2 11^T + N^{-1}\sigma_e^2)^{-1},$$

which, combining with (2.26), yields the continued equality

$$E\left(\sum_{i=1}^{q} n_i(y_i - \theta_i)^2 \bigg| \sigma_e^2, \sigma_\theta^2\right)$$

47
\[
\begin{align*}
\bar{Y}^T (\sigma_\theta^2 I + \sigma_\mu^2 11^T + NN^{-1}\sigma_e^2 N^{-1}NN^{-1}\sigma_e^2 (\sigma_\theta^2 I + \sigma_\mu^2 11^T + NN^{-1}\sigma_e^2)^{-1} \bar{Y} \\
+ \text{tr}(N\Sigma_\theta) \\
\leq \bar{Y}^T (\sigma_\theta^2 I + \sigma_\mu^2 11^T + NN^{-1}\sigma_e^2)^{-1} \bar{Y} + \text{tr}(N\Sigma_\theta)
\end{align*}
\]

because
\[
NN^{-1}\sigma_e^2 \leq NN^{-1}\sigma_e^2 + \sigma_\theta^2 I + \sigma_\mu^2 11^T,
\]

\[
\leq \bar{Y}^T NY + \text{tr}(N\Sigma_\theta),
\]

because
\[
(\sigma_\theta^2 I + \sigma_\mu^2 11^T + NN^{-1}\sigma_e^2)^{-1} \leq N\sigma_e^{-2}
\]

by Result 2.36 (2)

\[
\leq \bar{Y}^T NY + \sigma_e^2 q,
\]

because
\[
\text{tr}(N\Sigma_\theta) \leq \text{tr}(NN^{-1}\sigma_e^2) = \sigma_e^2 \text{tr}(I_q) = \sigma_e^2 q,
\]

by Result 2.36 (4).

Finally, note the two straightforward conditional expectations
\[
E(\sigma_e^2|\mu', \theta') = \frac{1}{2a_e + N - 2} \left( \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i')^2 + SSE + 2b_e \right) \tag{2.28}
\]

and
\[
E(\sigma_\theta^2|\mu', \theta') = \frac{1}{2a_\theta + q - 2} \left( \sum_{i=-1}^{q} (\theta_i' - \mu')^2 + 2b_\theta \right), \tag{2.29}
\]

which follow directly from Result 2.40.

Now, combining (2.24), (2.25), (2.27), (2.28), and (2.29), the expectation of \( V \) satisfies
\[
E(V(\theta, \mu)|\theta', \mu') = E \left( \sum_{i=1}^{q} (\theta_i - \mu)^2 + \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i)^2 \middle| \theta', \mu' \right)
\]

48
\[
\begin{align*}
&= E\left( E\left( \sum_{i=1}^{q} (\theta_i - \mu)^2 + \sum_{i=1}^{q} n_i (y_i - \theta_i)^2 \left\| \theta, \sigma_\varepsilon^2, \sigma_\theta^2 \right\| \theta', \mu' \right) \right)
\end{align*}
\]

by Result 2.38

\[
\leq E\left( \theta^T \theta + \sigma_\theta^2 + \sum_{i=1}^{q} n_i (y_i - \theta_i)^2 \left\| \theta', \mu' \right\| \right)
\]

by (2.24)

\[
= E\left( E\left( \theta^T \theta + \sigma_\theta^2 + \sum_{i=1}^{q} n_i (y_i - \theta_i)^2 \left\| \sigma_\varepsilon^2, \sigma_\theta^2 \right\| \theta', \mu' \right) \right)
\]

by Result 2.38

\[
\leq E\left( (1 + n_*^{-1}) \bar{Y}^T \bar{N} \bar{Y} + \sigma^2_\varepsilon \text{tr}(N^{-1}) + \sigma_\theta^2 + \sigma^2_\varepsilon q \left\| \theta', \mu' \right\| \right)
\]

by (2.25) and (2.27)

\[
= \frac{1}{2a_\theta + q - 2} \left[ \sum_{i=1}^{q} (\theta_i' - \mu')^2 + 2b_\theta \right]
\]

\[
+ \frac{\text{tr}(N^{-1}) + q}{2a_\varepsilon + N - 2} \left[ \sum_{i=1}^{n} n_i (y_i - \theta_i')^2 + SSE + 2b_\varepsilon \right] + (1 + n_*^{-1}) \bar{Y}^T \bar{N} \bar{Y}
\]

by (2.28) and (2.29)

\[
\leq \lambda V(\theta', \mu') + b,
\]

where

\[
\lambda = \max \left( \frac{1}{2a_\theta + q - 2}, \frac{\text{tr}(N^{-1}) + q}{2a_\varepsilon + N - 2} \right)
\]

and

\[
b = \frac{2b_\theta}{2a_\theta + q - 2} + \frac{\text{tr}(N^{-1}) + q}{2a_\varepsilon + N - 2} \left( \frac{SSE + 2b_\varepsilon}{2a_\varepsilon + N - 2} + (1 + n_*^{-1}) \bar{Y}^T \bar{N} \bar{Y} \right).
\]
Note that, by assumption of the proposition, $\lambda < 1$. Therefore, the drift condition (2.9) is satisfied. As discussed, there is a conversion which also implies that the drift condition (2.8) is satisfied.

Given a drift condition, proceed first with Method 2. As the transition densities are all continuous, application of Fatou proves this chain is Feller. The maximal irreducibility measure is Lebesgue measure, and hence has nonempty interior. It remains to show that $V$ is unbounded off compact sets. As $V$ is continuous, the set

$$C = \{(\theta, \mu) : V(\theta, \mu) \leq d\}$$

is closed. It remains to show $C$ is bounded. Consider

$$V(\theta, \mu) = \sum_{i=1}^{q} (\theta_i - \mu)^2 + \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i)^2.$$ 

If $\theta_i$ goes to $\pm\infty$, then the second term in $V$ goes to $\infty$. Thus, $\theta_i$ is bounded when $V(\theta, \mu) \leq d$. Given this, if $\mu$ goes to $\pm\infty$, then the first term in $V$ goes to $\infty$. So, when $V(\theta, \mu) \leq d$, $\mu$ is bounded. Thus, $C$ is bounded and $V$ is unbounded off compact sets for the chain. Note also that as the parameter updates all have continuous densities, by Proposition 2.25, the chain is Feller. Thus, as the drift condition (2.9) is satisfied with $V$, the Markov chain is geometrically ergodic according to Method 2.

As noted abundantly in the literature (e.g. Diaconis et al. 2008; Jones et al. 2014), the Markov chain with the first block integrated out converges at the same speed as the original chain, and thus the drift function $V$ need not contain the parameters $\sigma^2_e$ and $\sigma^2_\theta$ which comprise the first block. It is the marginal chain for which $V$ is unbounded off compact sets, but as the marginal chain is geometrically ergodic, so is the joint chain.
Now, to prove geometric ergodicity with Method 1, a minorization condition (2.7) must be shown. Define
\[
C = \left\{ (\theta, \mu) : \sum_{i=1}^{q} (\theta_i - \mu)^2 + \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i)^2 \leq d \right\}
\]
for some
\[
d > 2b/(1 - \lambda).
\]
It is occasionally more convenient to show (2.7) for a larger set \(C'\) where \(C \subset C'\).
This is acceptable because if (2.7) is satisfied for all \(x \in C'\), then it is also satisfied for all \(x \in C\). This fact was used in Flegal and Herbei (2012) for a highly related Markov chain. Thus, show minorization for the larger set
\[
C' = \left\{ \sum_{i=1}^{q} (\theta_i - \mu)^2 \leq d \right\} \cap \left\{ \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i)^2 \leq d \right\}.
\]
Clearly if \((\theta, \mu) \in C\), then \((\theta, \mu) \in C'\), so \(C \subset C'\).

The transition density of this algorithm is
\[
k(\theta', \mu', (\sigma_e^2)', (\sigma_0^2)'; \theta, \mu, (\sigma_e^2), (\sigma_0^2))
= f_{IG} \left( \sigma_e^2, a_e + N/2, \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i')^2/2 + SSE/2 + b_e \right)
\times f_{IG} \left( \sigma_0^2, a_0 + q/2, \sum_{i=1}^{q} (\theta_i' - \mu')^2/2 + b_0 \right) f_N(\theta'; \Sigma_\theta N \sigma_e^{-2} \overline{Y}, \Sigma_\theta)
\times f_N \left( \mu; \frac{\sigma_0^2}{\sigma_0^2} \left( \frac{1}{\sigma_e^2} + \frac{q}{\sigma_0^2} \right)^{-1} \left( \frac{1}{\sigma_e^2} + \frac{q}{\sigma_0^2} \right)^{-1} \right)
\geq \inf_{(\theta', \mu') \in C'} \left[ f_{IG} \left( \sigma_e^2, a_e + N/2, \sum_{i=1}^{q} n_i (\bar{y}_i - \theta_i')^2/2 + SSE/2 + b_e \right)
\times f_{IG} \left( \sigma_0^2, a_0 + q/2, \sum_{i=1}^{q} (\theta_i' - \mu')^2/2 + b_0 \right) f_N(\theta'; \Sigma_\theta N \sigma_e^{-2} \overline{Y}, \Sigma_\theta)
\times f_N \left( \mu; \frac{\sigma_0^2}{\sigma_0^2} \left( \frac{1}{\sigma_e^2} + \frac{q}{\sigma_0^2} \right)^{-1} \left( \frac{1}{\sigma_e^2} + \frac{q}{\sigma_0^2} \right)^{-1} \right) \right].
\]
51
\[ q'(\theta, \mu, \sigma^2_o, \sigma^2_e), \]

where \( f_{IG}(x; a, b) \) is the inverse gamma density evaluated at \( x \) with parameters \( a \) and \( b \) and \( f_N(x; \mu, \Sigma) \) is the multivariate normal density (perhaps with \( d = 1 \)) evaluated at \( x \) with mean \( \mu \) and covariance matrix \( \Sigma \).

Then, to finish finding the minorization condition, let \( \epsilon \) equal the integral of \( q' \) over its space and let \( q(\cdot) = q'(\cdot)/\epsilon \). Note that the normal densities are already normalized and integrate to 1. Fortunately, the infimum over \( C' \) of the inverse gamma density has a convenient form which is easily integrable (Jones and Hobert 2001a). The general form utilized here is

\[
g(x; a, b, c, d) = \inf_{c \in (0, d)} f_{IG}(x; a, c/2 + b) = \begin{cases} f_{IG}(x; a, d/2 + b) & x < x^* \\ f_{IG}(x; a, b) & x \geq x^* \end{cases}
\]

where \( x^* = d[2a \log(1 + d/(2b))]^{-1} \), which is easily integrated. So,

\[
\epsilon = \int_0^\infty g \left( \sigma^2_e, a_e + N/2, SSE/2 + b_a, \sum_{i=1}^q n_i (\overline{y}_i - \theta_i)^2, d \right) d\sigma^2_e \\
\times \int_0^\infty g \left( \sigma^2_o; a_o + q/2, b_o, \sum_{i=1}^q (\theta_i - \mu)^2, d \right) d\sigma^2_o,
\]

which completes the proof of the minorization condition. Thus, with minorization and a drift condition, the Markov chain is geometrically ergodic via Method 1. \( \square \)

As a specific illustration of this sampler, consider the exposure levels of 13 laminators at a boat-manufacturing facility to the compound styrene. The data were originally published in Lyles et al. (1997) and have since been a common example for Bayesian one-way random effects models (see e.g. Jones and Hobert 2001a; Tan and Hobert 2009; Flegal and Herbei 2012). For this example \( n_i = 3 \) for all individuals \( i = 1, \ldots, 13 \). The total sample size is then 39. Jones and Hobert (2001a) explored
six hyper-parameter settings, shown in Table 2.1. Each reflected different prior assumptions, justification of which is left to Jones and Hobert (2001a). Note that they modeled precisions rather than the variances; as such, one has been added to each of the inverse gamma shape parameters so the prior means match those determined by Jones and Hobert (2001a). The variance parameter priors chosen by Flegal and Herbei (2012) are also shown in Setting 7.

Table 2.1: Hyper-Parameter Specifications for each Setting

<table>
<thead>
<tr>
<th>Setting</th>
<th>$a_\theta$</th>
<th>$b_\theta$</th>
<th>$a_\psi$</th>
<th>$b_\psi$</th>
<th>$\mu_0$</th>
<th>$\sigma^2_{\mu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>61.176</td>
<td>7.7573</td>
<td>4.1237</td>
<td>1.7674</td>
<td>4.809</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>602.760</td>
<td>77.5730</td>
<td>32.2370</td>
<td>17.6740</td>
<td>4.809</td>
<td>10.0</td>
</tr>
<tr>
<td>3</td>
<td>1.100</td>
<td>0.1000</td>
<td>1.1000</td>
<td>0.1000</td>
<td>4.809</td>
<td>10.0</td>
</tr>
<tr>
<td>4</td>
<td>2.000</td>
<td>5.0000</td>
<td>2.0000</td>
<td>1.0000</td>
<td>3.600</td>
<td>1.0</td>
</tr>
<tr>
<td>5</td>
<td>1.600</td>
<td>1.0000</td>
<td>121.0000</td>
<td>16.0000</td>
<td>4.809</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>5.000</td>
<td>80.0000</td>
<td>41.0000</td>
<td>100.0000</td>
<td>4.000</td>
<td>1.0</td>
</tr>
<tr>
<td>7</td>
<td>0.100</td>
<td>10.0000</td>
<td>0.1000</td>
<td>10.0000</td>
<td>4.809</td>
<td>1.0</td>
</tr>
</tbody>
</table>

The convergence results using bounds provided by Rosenthal (1995) and listed in Theorem 2.30 are shown in Table 2.2. The initial value for $\theta$ is taken to be $\vec{Y}$ and the initial value for $\mu$ is taken to be the overall mean $\bar{y}$. For each setting, the number of iterations required for the TV distance to be less than 0.1 is indicated. A grid was used to search for good values of $r$ and $d$. The minimum number of iterations required is for Setting 6 with 214,885 and the maximum required is $6.604e+21$ with Setting 4; thus, the number of iterations required varies wildly depending on the hyperpriors. This illustrates the importance of studying the convergence behavior
of Markov chains; even with the same data and same prior families, Markov chains differ greatly in their convergence.

Similar convergence results using the bounds of Roberts and Tweedie (1999) given in Theorem 2.31 are provided in Table 2.3. A grid was used to somewhat optimally select \( d \). The table does not explicitly state \( \beta; \) as \( \beta^* \) was very close to 1 in each of the settings, \( \beta \) was taken to be \( \beta = (\beta^* + 1)/2 \). As suggested by Roberts and Tweedie (1999), these bounds are all indeed superior to those provided by Rosenthal (1995). For example, Setting 6, the quickest converging chain, only requires 41,074 iterations with these bounds compared to 214,885 with the previous bounds. The order of convergence speed of the models is nearly preserved for both sets of bounds (they disagree on the ranks for Settings 1 and 3).

Additionally, the bounds of Roberts and Tweedie (1999) are a bit easier to use comparatively because once \( d \) and \( \beta \) are selected, the number of iterations required to obtain TV distance less than some threshold may be directly computed. On the other hand, with the bounds of Rosenthal (1995), there are two parts summed together which makes the number of iterations required for a small TV distance more challenging to compute.

One important note is that both sets of these bounds are conservative. The inequalities utilized in the proof of the drift condition are far from tight, indicating that these bounds themselves may be far from tight. As an evaluation of whether or not the large number of iterations required by the bounds are actually required for accurate inference, Markov chain Monte Carlo standard errors are now considered.

Table 2.4 provides posterior medians and 95% equal-tail credible intervals for the parameters \( \sigma^2_{\theta}, \sigma^2_\epsilon, \mu, \) and \( \theta_1 \) (the other \( \theta_i \) are omitted, but are estimated with a
Table 2.2: Convergence Quantities and Iterations Required for TV Distance Less than 0.1 using Bounds by Rosenthal (1995)

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\lambda$</th>
<th>$b$</th>
<th>$d$</th>
<th>$r$</th>
<th>$\epsilon$</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.383</td>
<td>22.340</td>
<td>75.357</td>
<td>5.139e-03</td>
<td>1.137e-16</td>
<td>4.038e+18</td>
</tr>
<tr>
<td>2</td>
<td>0.171</td>
<td>23.912</td>
<td>60.013</td>
<td>7.546e-03</td>
<td>2.676e-07</td>
<td>1.141e+09</td>
</tr>
<tr>
<td>3</td>
<td>0.442</td>
<td>21.842</td>
<td>84.649</td>
<td>8.724e-03</td>
<td>8.181e-17</td>
<td>2.378e+18</td>
</tr>
<tr>
<td>4</td>
<td>0.423</td>
<td>99.053</td>
<td>343.234</td>
<td>9.293e-09</td>
<td>3.757e-17</td>
<td>6.604e+21</td>
</tr>
<tr>
<td>5</td>
<td>0.070</td>
<td>18.277</td>
<td>42.504</td>
<td>0.018</td>
<td>4.934e-09</td>
<td>2.552e+10</td>
</tr>
<tr>
<td>6</td>
<td>0.146</td>
<td>88.214</td>
<td>231.560</td>
<td>0.017</td>
<td>6.220e-04</td>
<td>214,885</td>
</tr>
<tr>
<td>7</td>
<td>0.466</td>
<td>33.193</td>
<td>139.387</td>
<td>0.011</td>
<td>4.979e-06</td>
<td>4.189e+07</td>
</tr>
</tbody>
</table>

Table 2.3: Convergence Quantities and Iterations Required for TV Distance Less than 0.1 using Bounds by Roberts and Tweedie (1999)

<table>
<thead>
<tr>
<th>Setting</th>
<th>$\lambda$</th>
<th>$b$</th>
<th>$d$</th>
<th>$J$</th>
<th>$\epsilon$</th>
<th>$\beta^*$</th>
<th>Iterations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6915</td>
<td>22.96</td>
<td>72.50</td>
<td>20.73</td>
<td>3.704e-16</td>
<td>1+(4.51e+17)</td>
<td>1.302e+17</td>
</tr>
<tr>
<td>2</td>
<td>0.5854</td>
<td>24.74</td>
<td>57.73</td>
<td>9.37</td>
<td>5.828e-07</td>
<td>1+(1.39e+07)</td>
<td>41,141,021</td>
</tr>
<tr>
<td>3</td>
<td>0.7211</td>
<td>22.40</td>
<td>78.39</td>
<td>70.75</td>
<td>2.405e-16</td>
<td>1+(1.85e+17)</td>
<td>3.202e+17</td>
</tr>
<tr>
<td>4</td>
<td>0.7114</td>
<td>99.63</td>
<td>343.54</td>
<td>279.32</td>
<td>3.696e-17</td>
<td>1+(2.23e+18)</td>
<td>2.654e+18</td>
</tr>
<tr>
<td>5</td>
<td>0.5352</td>
<td>19.21</td>
<td>39.36</td>
<td>81.90</td>
<td>1.656e-08</td>
<td>1+(2.35e+09)</td>
<td>2.488e+09</td>
</tr>
<tr>
<td>6</td>
<td>0.5728</td>
<td>89.07</td>
<td>206.72</td>
<td>323.33</td>
<td>1.487e-03</td>
<td>1+(1.43e+04)</td>
<td>41,074</td>
</tr>
<tr>
<td>7</td>
<td>0.7330</td>
<td>33.73</td>
<td>124.43</td>
<td>123.96</td>
<td>1.376e-05</td>
<td>1+(8.86e+07)</td>
<td>6.682,944</td>
</tr>
</tbody>
</table>

A similar level of precision to $\theta_1$ for each of the seven parameter settings. A 95% Monte Carlo margin of error is provided for each of these estimates, computed using the subsampling method discussed in Section 2.4.3, which is justified because the chain is geometrically ergodic. A fixed 1 million iterations are obtained from each Markov chain and the batch size is set to be 1000. Even for Settings 3 and 4, which required a large number of iterations for the total variation distance to be guaranteed to drop below 0.1, the margins of error are all quite small (note that 0.000 doesn’t imply
zero exactly, but that the error rounds down to zero out to three decimal places. This not only answers Question 1 that the simulation has indeed run long enough, but also Question 2 by specifically quantifying the accuracy of each estimate. Had any of these margins of error been too large, additional iterations could be obtained for inference. This suggests that perhaps the number of iterations required using the bounding techniques of Section 2.3 are not always necessary for accurate inference and that Monte Carlo error estimates are a superior method for evaluating the answers to Question 1 and Question 2 for Markov chains.

2.5.2 Multivariate Truncated Normal Distribution using a Metropolis-Hastings Sampler

Many statistical methods require the use of a random draw from a multivariate truncated normal distribution, often as a conditional component of a larger MCMC method. See, for example, scale-usage models in Chapter 4. Several methods exist for drawing from such distributions (see e.g. Geweke 1991). This example is by no means an attempt to improve on these methods, though geometric ergodicity will be proved, a result which may not be possible for some of the methods in the literature.

A thorough description of the target distribution and the corresponding Markov chain is now provided. The probability space is \((\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))\). Let \(\pi_2\) represent both the measure and the density with respect to Lebesgue measure \(\mu\) of the truncated multivariate normal distribution with location parameter \(\mu\) and scale matrix \(\Sigma\). Let \(\mathcal{A} \in \mathcal{B}(\mathbb{R}^d)\) be the set of truncation where \(\mu(\mathcal{A}) > 0\). The density of the truncated multivariate normal distribution is then

\[
\pi_2(y) = P(E \in \mathcal{A})^{-1}(2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y - \mu)^T \Sigma^{-1} (y - \mu)\right) \mathbb{1}_A(y), \quad (2.30)
\]
Table 2.4: Posterior Medians and 95% Equal-Tail Credible Interval Bounds with Corresponding 95% Estimation bounds

<table>
<thead>
<tr>
<th>Setting</th>
<th>Parameter</th>
<th>Median</th>
<th>Lower Bound</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\sigma_0^2$</td>
<td>0.128 ± 0.000</td>
<td>0.101 ± 0.000</td>
<td>0.166 ± 0.000</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>0.544 ± 0.000</td>
<td>0.356 ± 0.000</td>
<td>0.880 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.810 ± 0.000</td>
<td>4.507 ± 0.001</td>
<td>5.112 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>4.184 ± 0.001</td>
<td>3.597 ± 0.002</td>
<td>4.795 ± 0.002</td>
</tr>
<tr>
<td>2</td>
<td>$\sigma_0^2$</td>
<td>0.129 ± 0.000</td>
<td>0.119 ± 0.000</td>
<td>0.140 ± 0.000</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>0.557 ± 0.000</td>
<td>0.424 ± 0.000</td>
<td>0.752 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.809 ± 0.000</td>
<td>4.504 ± 0.001</td>
<td>5.115 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>4.192 ± 0.001</td>
<td>3.616 ± 0.002</td>
<td>4.780 ± 0.002</td>
</tr>
<tr>
<td>3</td>
<td>$\sigma_0^2$</td>
<td>0.109 ± 0.000</td>
<td>0.027 ± 0.000</td>
<td>0.434 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>0.556 ± 0.000</td>
<td>0.346 ± 0.000</td>
<td>0.941 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.809 ± 0.000</td>
<td>4.494 ± 0.000</td>
<td>5.125 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>4.258 ± 0.002</td>
<td>3.421 ± 0.003</td>
<td>4.895 ± 0.002</td>
</tr>
<tr>
<td>4</td>
<td>$\sigma_0^2$</td>
<td>0.943 ± 0.001</td>
<td>0.484 ± 0.001</td>
<td>2.147 ± 0.005</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>0.544 ± 0.000</td>
<td>0.345 ± 0.000</td>
<td>0.922 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.711 ± 0.001</td>
<td>4.111 ± 0.002</td>
<td>5.274 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>3.538 ± 0.001</td>
<td>2.745 ± 0.002</td>
<td>4.351 ± 0.002</td>
</tr>
<tr>
<td>5</td>
<td>$\sigma_0^2$</td>
<td>0.369 ± 0.000</td>
<td>0.179 ± 0.000</td>
<td>0.879 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>0.175 ± 0.000</td>
<td>0.148 ± 0.000</td>
<td>0.208 ± 0.000</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.810 ± 0.000</td>
<td>4.441 ± 0.001</td>
<td>5.178 ± 0.001</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>3.515 ± 0.001</td>
<td>3.050 ± 0.001</td>
<td>3.979 ± 0.001</td>
</tr>
<tr>
<td>6</td>
<td>$\sigma_0^2$</td>
<td>7.903 ± 0.006</td>
<td>4.594 ± 0.006</td>
<td>15.237 ± 0.030</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>1.984 ± 0.001</td>
<td>1.539 ± 0.001</td>
<td>2.618 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.489 ± 0.002</td>
<td>3.219 ± 0.004</td>
<td>5.723 ± 0.003</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>3.396 ± 0.002</td>
<td>1.857 ± 0.004</td>
<td>4.940 ± 0.004</td>
</tr>
<tr>
<td>7</td>
<td>$\sigma_0^2$</td>
<td>2.306 ± 0.002</td>
<td>1.103 ± 0.002</td>
<td>5.911 ± 0.017</td>
</tr>
<tr>
<td></td>
<td>$\sigma_e^2$</td>
<td>1.273 ± 0.001</td>
<td>0.790 ± 0.001</td>
<td>2.231 ± 0.004</td>
</tr>
<tr>
<td></td>
<td>$\mu$</td>
<td>4.809 ± 0.001</td>
<td>3.957 ± 0.002</td>
<td>5.661 ± 0.002</td>
</tr>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>3.546 ± 0.002</td>
<td>2.337 ± 0.004</td>
<td>4.776 ± 0.004</td>
</tr>
</tbody>
</table>
where $E$ is random variable whose distribution is multivariate normal with mean $\mu$ and covariance matrix $\Sigma$.

The Markov chain used to explore $\pi_2$ is a Metropolis-Hastings chain as in Example 2.1 with candidate density $q(y)$ not dependent on the current state, known as an *independence Metropolis-Hastings* sampler. Here the candidate density $q(y)$ is a truncated multivariate normal distribution with location parameter $\mu$ but with scale matrix $D$ which is chosen to be diagonal. The transition density then takes the form

$$k(y',y) = q(y)\alpha(y',y) + r(y')\delta_{y'}(y)$$

where $\delta_{y'}(\cdot)$ is the Dirac mass and

$$r(y') = \int q(y)\alpha(y',y)dy,$$

with

$$\alpha(y',y) = \min \left(1, \frac{\exp\left(-\frac{1}{2}(y' - \mu)^T(\Sigma^{-1} - D^{-1})(y' - \mu)\right)}{\exp\left(-\frac{1}{2}(y - \mu)^T(\Sigma^{-1} - D^{-1})(y - \mu)\right)}\right).$$

The Markov transition kernel $K(\cdot, \cdot)$ may then be obtained by integrating this transition density $k(\cdot, \cdot)$ with respect to Lebesgue measure.

The matrix $D$ is chosen carefully. As mentioned previously, $D$ is diagonal. Thus, each dimension is independent of the others, and the joint draw is a series of univariate truncated normal draws which are readily available (except perhaps for any joint constraints specified in determining the set $A$; in this case, assume there is a viable method for obtaining draws from the candidate density $q$). More than simply diagonal, however, $D$ is chosen such that $\Sigma^{-1} - D^{-1}$ is non-negative definite. To select such a $D$, the decomposition strategy discussed in Hans et al. (2012) is employed. Their paper is discussed further in Chapter 4. For now, it is sufficient to describe the decomposition algorithm.
Algorithm 2.4 (Decomposition of a Positive Definite Matrix)

Let $A$ be a positive definite $d \times d$ matrix. The following algorithm finds matrices $B$ and $E$ such that $A = B + E$ where $B$ is diagonal and $E \geq 0$.

(i) Define $V = \text{diag}(a_{11}, a_{22}, \ldots, a_{dd})$.

(ii) Compute $C = V^{-1/2}AV^{-1/2}$.

(iii) Find the $d$ eigenvalues of $C$. Denote the smallest eigenvalue $\lambda_d$.

(iv) Set $\rho \in (0, 1]$; $\rho = 1$ will produce $E \geq 0$; $\rho \in (0, 1)$ results in $E > 0$.

(v) Define $B = \rho \lambda_d V$. As $V$ is diagonal, so is $B$.

(vi) Define $E = A - B$.

Use Algorithm 2.4 to decompose $\Sigma^{-1}$ into the diagonal matrix $D^{-1}$ and remainder matrix $R = \Sigma^{-1} - D^{-1}$. Then $R$ is nonnegative definite. This matrix $D^{-1}$ is chosen specifically for the convergence properties it induces on the Markov chain. A theorem from Mengersen and Tweedie (1996) is now stated.

Theorem 2.42. The independence Metropolis-Hastings algorithm is uniformly ergodic if there is a $\beta > 0$ such that

$$\frac{q(y)}{\pi(y)} \geq \beta, \quad y \in \mathcal{X}. \quad (2.31)$$

Thus, establishing (2.31) implies uniform (and hence geometric) ergodicity. This is now demonstrated for the current example. Note that

$$\frac{q(y)}{\pi_2(y)} = \frac{P(E \in A)|D|^{-1/2}}{P(D \in A)|\Sigma|^{-1/2}} \exp \left( \frac{1}{2}(y - \mu)^T(\Sigma^{-1} - D^{-1})(y - \mu) \right) \geq \frac{P(E \in A)|D|^{-1/2}}{P(D \in A)|\Sigma|^{-1/2}} \Rightarrow \beta > 0, \quad (2.32)$$
where the first inequality holds because $\Sigma^{-1} - D^{-1}$ is nonnegative definite, which implies that the quadratic form $(y - \mu)^T(\Sigma^{-1} - D^{-1})(y - \mu)$ is nonnegative, which guarantees

$$\exp \left( \frac{1}{2} (y - \mu)^T(\Sigma^{-1} - D^{-1})(y - \mu) \right) \geq 1.$$  

Thus, by Theorem 2.42, the chain is uniformly ergodic. For illustration purposes, this is explicitly shown by finding a minorization condition for the entire space $\mathbb{R}^d$. Consider cases. First, assume $\alpha(y', y) = 1$. Then

$$k(y', y) \geq q(y) \geq \beta \pi_2(y),$$

which is true by (2.32). Now, assume $\alpha(y', y) < 1$. Then

$$k(y', y) \geq q(y) \frac{\pi_2(y)q(y')}{\pi_2(y')q(y)} \geq \beta \pi_2(y),$$

which is true by (2.32). In either case,

$$k(y', y) \geq \beta \pi_2(y)$$

for any $y' \in \mathbb{R}^d$. Integrate both sides of the inequality with respect to Lebesgue measure to obtain the transition kernel and measure $Q$. Thus, as the minorization condition (2.7) is satisfied for the entire space $C = \mathbb{R}^d$ with $\epsilon = \beta$, $Q = \pi_2$, and $m = 1$, the chain induced by the transition density above is uniformly, and therefore geometrically, ergodic according to Method 1b. Further, as discussed in Proposition 2.29, as $m = 1$, the upper bound of the total variation distance is

$$||K^n(x, \cdot) - \pi_2||_{TV} \leq (1 - \beta)^n,$$

which is simple to compute when $\beta$ is known.
This example is now illustrated for $\pi_2$ with specific $\mu$, $\Sigma$, and $\mathcal{A}$. Let the number of dimensions be $d = 10$ and define $\mu = 0$. Let $\Sigma$ have an autoregressive structure such that the $(\Sigma_{ij}) = \sigma^2 \rho^{|i-j|}$; in particular, let $\sigma^2 = 1$ and $\rho = 0.75$. For reference, the first line of $\Sigma$ is

$$\Sigma_{1,1:10} = \begin{pmatrix} 1.00 & 0.75 & 0.56 & 0.42 & 0.32 & 0.24 & 0.18 & 0.13 & 0.10 & 0.08 \end{pmatrix}.$$ 

This $\Sigma$ was chosen so each dimension would be more highly correlated with some other dimensions, and less correlated with others.

To determine the truncation set $\mathcal{A}$, consider the situation of a multivariate ordinal probit model (see Chapter 4) where a conditional update for the latent variable is multivariate truncated normal with cutpoints determining the bounds. Thus, $\mathcal{A}$ is a multivariate rectangle with the cutpoints determining the lower and upper bounds. The cutpoints were chosen such that, for a univariate Normal distribution with mean 0 and variance 1, each region between cutpoints has equal mass. Specifically, let

$$c = \{c_0, c_1, \ldots, c_d\} = \{-\infty, -1.28, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28, \infty\}.$$ 

To explore the effect of the truncation set on the convergence of the Markov chain, 10 random sets of truncation were generated in each of two settings. For each dimension, randomly generated ordinal data $X$ is used to identify the corresponding cutpoints $c_{X_{i-1}}, c_X$. In the first setting, the ordinal data were randomly generated according to a discrete uniform distribution. In the second setting, multivariate normal variates $Y$ were generated with mean $\mu$ and covariance $\Sigma$. Then, where $Y_i \in (c_{X_{i-1}}, c_{X_i})$, $X_i$ was determined for $i = 1, \ldots, 10$. Setting 2 is therefore more realistic in that the cutpoints were randomly generated according to the assumed model of the latent variables. The generated ordinal values for Setting 1 are shown in Table 2.5 and the
ordinal values for Setting 2 are found in Table 2.6. Each row is a different random draw and each column is a different dimension of that draw.

Table 2.5: Generated Multivariate Data Determining the Cutpoints under Setting 1

<table>
<thead>
<tr>
<th>#</th>
<th>Setting 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3 2 6 1 2 7 10 4 10 10</td>
</tr>
<tr>
<td>2</td>
<td>4 7 10 10 9 6 4 6 2 5</td>
</tr>
<tr>
<td>3</td>
<td>4 2 3 7 5 3 9 5 6 3</td>
</tr>
<tr>
<td>4</td>
<td>7 7 5 6 9 4 1 7 7 6</td>
</tr>
<tr>
<td>5</td>
<td>4 10 5 2 8 6 7 1 1 2</td>
</tr>
<tr>
<td>6</td>
<td>7 8 3 4 9 10 2 8 1 3</td>
</tr>
<tr>
<td>7</td>
<td>2 5 8 10 8 4 3 8 2 3</td>
</tr>
<tr>
<td>8</td>
<td>5 3 5 2 10 1 9 1 3 7</td>
</tr>
<tr>
<td>9</td>
<td>9 10 3 10 1 10 1 4 6 6</td>
</tr>
<tr>
<td>10</td>
<td>7 3 9 3 5 6 8 6 4 6</td>
</tr>
</tbody>
</table>

Several convergence quantities are computed for each of these observations in Setting 1 and Setting 2 and are presented in Table 2.7 and Table 2.8 respectively. First, $P(E \in A)$ where $E$ is multivariate normal with mean $\mu$ and covariance $\Sigma$ is shown. This is the acceptance rate of the naive rejection sampler for the truncated multivariate normal distribution which uses a $\text{MVN}(\mu, \Sigma)$ envelope function and accepts with probability 1 if the candidate is in $A$. As all these probabilities are quite small, naive rejection sampling is not practical. Next, the computed

$$\beta = (P(E \in A|D)^{-1/2})/(P(D \in A|\Sigma)^{-1/2})$$

is shown. Finally, the number of iterations $n$ such that the TV distance is below 0.1, computed by rounding up $\log(0.1)/\log(1 - \beta)$, is provided.
Table 2.6: Generated Multivariate Data Determining the Cutpoints under Setting 2

<table>
<thead>
<tr>
<th></th>
<th>Setting 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10 8 10 10 8 8 9 10 9</td>
</tr>
<tr>
<td>2</td>
<td>7 8 6 3 4 3 5 8 9 9</td>
</tr>
<tr>
<td>3</td>
<td>2 1 2 4 5 5 4 5 3</td>
</tr>
<tr>
<td>4</td>
<td>5 2 5 1 2 1 5 3 3 5</td>
</tr>
<tr>
<td>5</td>
<td>5 7 4 3 7 10 8 2 5 9</td>
</tr>
<tr>
<td>6</td>
<td>10 10 7 5 4 4 5 4 1 1</td>
</tr>
<tr>
<td>7</td>
<td>9 7 7 5 3 8 9 8 10 10</td>
</tr>
<tr>
<td>8</td>
<td>5 4 9 7 8 6 10 9 7 7</td>
</tr>
<tr>
<td>9</td>
<td>7 9 9 10 9 6 9 9 8 9</td>
</tr>
<tr>
<td>10</td>
<td>5 3 2 3 5 6 7 4 3 5</td>
</tr>
</tbody>
</table>

Note that the number of iterations required is, in general, less for Setting 2, which is intuitive because Setting 2 was generated with $\pi_2$ in mind. The slowest converging Markov chain is that for Observation 9 of Setting 1, requiring $2.074e+16$ iterations for the TV distance to be below the designated threshold of 0.1. Note the sequence of the draws for Observation 9 of Setting 1. Each successive draw should have 0.75 correlation with the previous draw, thus a sequence from 10 to 3 to 10 to 1 to 10 to 1 has extremely small probability of occurring, which leads to a very poorly mixing chain. Contrast this with observation 10 from Setting 2 where only 12 draws are required for TV distance to be below 0.1. For this observation, all jumps from one time to the next are 3 or less, with 5 of the jumps only moving up or down by 1. Altogether, an overarching conclusion is that for this independence Metropolis-Hastings algorithm for simulating from the multivariate truncated normal distribution $\pi_2$, the convergence speed of the Markov chain is incredibly dependent upon the truncation set $\mathcal{A}$.
Table 2.7: Number of Draws Required for TV Distance to be Below 0.1 for Setting 1

<table>
<thead>
<tr>
<th>#</th>
<th>P((E \in A))</th>
<th>(\beta)</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.631e-13</td>
<td>2.266e-08</td>
<td>101,602,668</td>
</tr>
<tr>
<td>2</td>
<td>5.154e-10</td>
<td>0.001183</td>
<td>1,945</td>
</tr>
<tr>
<td>3</td>
<td>4.148e-11</td>
<td>0.00202</td>
<td>1,139</td>
</tr>
<tr>
<td>4</td>
<td>1.754e-11</td>
<td>0.0003286</td>
<td>7,007</td>
</tr>
<tr>
<td>5</td>
<td>6.599e-13</td>
<td>2.177e-07</td>
<td>10,575,604</td>
</tr>
<tr>
<td>6</td>
<td>9.878e-15</td>
<td>1.162e-08</td>
<td>198,160,564</td>
</tr>
<tr>
<td>7</td>
<td>2.022e-11</td>
<td>0.0001193</td>
<td>19,303</td>
</tr>
<tr>
<td>8</td>
<td>6.775e-19</td>
<td>2.064e-13</td>
<td>1.116e+13</td>
</tr>
<tr>
<td>9</td>
<td>4.643e-21</td>
<td>7.200e-17</td>
<td>2.074e+16</td>
</tr>
<tr>
<td>10</td>
<td>1.097e-11</td>
<td>0.0008366</td>
<td>2,752</td>
</tr>
</tbody>
</table>

To gain a sense for how these levels of convergence look visually, see Figure 2.1 where the first 10,000 iterations of a few of these Markov chains are shown. The first trace plot is the first dimension of the Markov chain associated with Observation 10 in Setting 2 where the TV distance is less than 0.1 after only 12 iterations. The next is the same plot for Observation 6 in Setting 2 where the TV distance is less than 0.1 after 21,585 iterations. The third plot is for Observation 6 of Setting 1 where nearly 200 million iterations are needed for the TV distance to be less than 0.1. The final plot is for Observation 9 of Setting 1 where over 20 quadrillion iterations are required for the TV distance to be less than 0.1. The visual convergence preserves the order of convergence speeds, though it is not obvious that a trace plot that looks as nice as that of Observation 6 of Setting 1 would require over 200 million draws for the TV distance to become reasonably small. One would expect a trace plot as bad as the plot for Observation 9 of Setting 1 would require many iterations, but the practitioner’s estimated number of iterations required would likely be much less than...
Table 2.8: Number of Draws Required for TV Distance to be Below 0.1 for Setting 2

<table>
<thead>
<tr>
<th>#</th>
<th>$P(E \in \mathcal{A})$</th>
<th>$\beta$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.462e-08</td>
<td>0.0003884</td>
<td>5,928</td>
</tr>
<tr>
<td>2</td>
<td>2.663e-09</td>
<td>0.1060000</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>1.081e-08</td>
<td>0.0794800</td>
<td>28</td>
</tr>
<tr>
<td>4</td>
<td>2.393e-10</td>
<td>0.0004553</td>
<td>5,057</td>
</tr>
<tr>
<td>5</td>
<td>1.165e-11</td>
<td>0.0001067</td>
<td>21,585</td>
</tr>
<tr>
<td>6</td>
<td>5.493e-09</td>
<td>0.0003985</td>
<td>5,777</td>
</tr>
<tr>
<td>7</td>
<td>2.614e-09</td>
<td>0.0025030</td>
<td>919</td>
</tr>
<tr>
<td>8</td>
<td>2.938e-10</td>
<td>0.0031620</td>
<td>728</td>
</tr>
<tr>
<td>9</td>
<td>1.233e-08</td>
<td>0.0265200</td>
<td>86</td>
</tr>
<tr>
<td>10</td>
<td>2.325e-09</td>
<td>0.178300</td>
<td>12</td>
</tr>
</tbody>
</table>

20 quadrillion. Thus, looking at a trace plot, while commonly done in practice, does not provide perfect information about the convergence speed of the Markov chain.

While a large number of iterations may be required in some settings for the TV distance to be below a given threshold, fewer iterations may be required for reliable inference. To explore this, each Markov chain was run for 1 million iterations and the output was used to estimate the mean of the given truncated multivariate normal distribution. The overlapping batch means method, discussed in Section 2.4.2, is then used to estimate the CLT variance for these estimates and a Monte Carlo margin of error is computed for each estimate. This method is justified because the Markov chains are uniformly ergodic and the necessary moment condition is satisfied because of the following. Note that for any positive $g$

$$
E_{\pi_2}(g(Y)) = \frac{1}{P(E \in \mathcal{A})} \int_{\mathcal{A}} g(Y)\pi_G(Y) dY
\leq \frac{1}{P(E \in \mathcal{A})} \int_{\mathcal{X}} g(Y)\pi_G(Y) dY
= \frac{1}{P(E \in \mathcal{A})} E_{\pi_G}(g(Y)),
$$
where $E_{\pi_\theta}(g(Y))$ is the desired moment under the truncated multivariate normal distribution and $E_{\pi_\theta^c}(g(Y))$ is the desired moment under the non-truncated multivariate normal distribution. This implies that, as the desired multivariate normal moment is finite and $P(E \in \mathcal{A}) > 0$, then the desired truncated multivariate normal moment is also finite and the moment condition is satisfied. The batch size was taken to be 1000, which satisfies the necessary requirements for strong consistency of the variance estimate.

The results for Setting 1 are provided in Table 2.9, and the results for Setting 2 are provided in Table 2.10. All of the Markov chain Monte Carlo margins of error are small enough for reliable inference. Note that, in general, those settings which required a larger number of iterations for TV distance to be guaranteed to be below 0.1 also have larger Monte Carlo margins of error, but such errors are nevertheless still quite small. This then provides not only the answer to Question 2 about the accuracy of the estimates, but also a good answer to Question 1 that the Markov chains have indeed been run for a sufficient length.

**Conclusions**

This chapter has laid the theoretical foundation for Markov chain convergence. Geometric ergodicity was discussed, along with methods for its establishment. Markov chain central limit theorems were presented, along with methods for estimating the associated asymptotic variances. This convergence theory was illustrated in practice with two examples. Answers to both Question 1 and Question 2 were rigorously obtained by computing the number of iterations required for the TV distance to be below a given threshold and by computing Markov chain Monte Carlo standard errors. This work now switches to another method for answering Question 1: beginning
the Markov chain with a draw from a distribution which is already close to $\pi$ in total variation distance.
Figure 2.1: Trace plots of Markov chains with various convergence speeds.
Table 2.9: Estimated Means for Truncated Multivariate Normal Distributions in Setting 1 with 95% Monte Carlo Margins of Error

<table>
<thead>
<tr>
<th>Obs.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.6837 ± 0.0030</td>
<td>-1.0214 ± 0.0039</td>
<td>0.1021 ± 0.0024</td>
<td>-1.5144 ± 0.0060</td>
<td>-1.0353 ± 0.0042</td>
<td>0.3850 ± 0.0027</td>
<td>1.4595 ± 0.0048</td>
<td>-0.3508 ± 0.0024</td>
<td>1.5567 ± 0.0074</td>
<td>1.7750 ± 0.0117</td>
</tr>
<tr>
<td>2</td>
<td>-0.3794 ± 0.0006</td>
<td>0.3926 ± 0.0006</td>
<td>1.6230 ± 0.0020</td>
<td>1.7076 ± 0.0024</td>
<td>1.0517 ± 0.0010</td>
<td>0.1304 ± 0.0006</td>
<td>-0.3781 ± 0.0006</td>
<td>0.1120 ± 0.0006</td>
<td>-1.0046 ± 0.0009</td>
<td>-0.1343 ± 0.0006</td>
</tr>
<tr>
<td>3</td>
<td>-0.3942 ± 0.0003</td>
<td>-1.0308 ± 0.0004</td>
<td>-0.6724 ± 0.0003</td>
<td>0.3721 ± 0.0003</td>
<td>-0.1270 ± 0.0002</td>
<td>-0.6513 ± 0.0003</td>
<td>0.9868 ± 0.0004</td>
<td>-0.1143 ± 0.0003</td>
<td>0.1172 ± 0.0003</td>
<td>-0.6684 ± 0.0003</td>
</tr>
<tr>
<td>4</td>
<td>0.3876 ± 0.0006</td>
<td>0.3832 ± 0.0006</td>
<td>-0.1200 ± 0.0006</td>
<td>0.1325 ± 0.0006</td>
<td>0.9981 ± 0.0005</td>
<td>-0.3853 ± 0.0006</td>
<td>-1.4564 ± 0.0005</td>
<td>0.3697 ± 0.0006</td>
<td>0.3856 ± 0.0005</td>
<td>0.1289 ± 0.0005</td>
</tr>
<tr>
<td>5</td>
<td>-0.3675 ± 0.0007</td>
<td>1.4352 ± 0.0028</td>
<td>-0.1212 ± 0.0017</td>
<td>-0.9900 ± 0.0025</td>
<td>0.6518 ± 0.0019</td>
<td>0.1343 ± 0.0017</td>
<td>0.3650 ± 0.0017</td>
<td>-1.5387 ± 0.0045</td>
<td>-1.6959 ± 0.0063</td>
<td>-1.0672 ± 0.0030</td>
</tr>
<tr>
<td>6</td>
<td>0.3904 ± 0.0017</td>
<td>0.6591 ± 0.0020</td>
<td>-0.6588 ± 0.0019</td>
<td>-0.3778 ± 0.0014</td>
<td>1.0330 ± 0.0026</td>
<td>1.4602 ± 0.0017</td>
<td>-0.9619 ± 0.0017</td>
<td>0.6320 ± 0.0017</td>
<td>-1.4592 ± 0.0031</td>
<td>-0.6920 ± 0.0020</td>
</tr>
<tr>
<td>7</td>
<td>-1.0269 ± 0.0008</td>
<td>-0.1278 ± 0.0005</td>
<td>0.6825 ± 0.0006</td>
<td>1.5400 ± 0.0006</td>
<td>0.6792 ± 0.0005</td>
<td>-0.3799 ± 0.0005</td>
<td>-0.6590 ± 0.0005</td>
<td>0.6406 ± 0.0005</td>
<td>-1.0049 ± 0.0008</td>
<td>-0.6845 ± 0.0006</td>
</tr>
<tr>
<td>8</td>
<td>-0.1316 ± 0.0004</td>
<td>-0.6678 ± 0.0004</td>
<td>-0.1452 ± 0.0003</td>
<td>-0.9752 ± 0.0048</td>
<td>1.3872 ± 0.0044</td>
<td>-1.3873 ± 0.0043</td>
<td>0.9472 ± 0.0043</td>
<td>-1.4368 ± 0.0066</td>
<td>-0.6805 ± 0.0043</td>
<td>0.3750 ± 0.0037</td>
</tr>
<tr>
<td>9</td>
<td>1.0532 ± 0.0007</td>
<td>1.5027 ± 0.0003</td>
<td>-0.6320 ± 0.0003</td>
<td>1.4026 ± 0.0045</td>
<td>-1.3855 ± 0.0059</td>
<td>1.3765 ± 0.0055</td>
<td>-1.4211 ± 0.0071</td>
<td>-0.3981 ± 0.0044</td>
<td>0.1327 ± 0.0043</td>
<td>0.1197 ± 0.0041</td>
</tr>
<tr>
<td>10</td>
<td>0.3768 ± 0.0003</td>
<td>-0.6449 ± 0.0003</td>
<td>0.9756 ± 0.0003</td>
<td>-0.6518 ± 0.0003</td>
<td>-0.1293 ± 0.0003</td>
<td>0.1293 ± 0.0003</td>
<td>0.6665 ± 0.0003</td>
<td>0.1271 ± 0.0003</td>
<td>-0.3780 ± 0.0003</td>
<td>0.1217 ± 0.0003</td>
</tr>
</tbody>
</table>
Table 2.10: Estimated Means for Truncated Multivariate Normal Distributions in Setting 2 with 95% Monte Carlo Margins of Error

<table>
<thead>
<tr>
<th>Obs.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.6181</td>
<td>0.7094</td>
<td>1.6961</td>
<td>1.8769</td>
<td>1.6895</td>
<td>0.6931</td>
<td>0.6879</td>
<td>1.0633</td>
<td>1.6176</td>
<td>1.0674</td>
</tr>
<tr>
<td></td>
<td>± 0.0062</td>
<td>± 0.0021</td>
<td>± 0.0071</td>
<td>± 0.0090</td>
<td>± 0.0071</td>
<td>± 0.0022</td>
<td>± 0.0023</td>
<td>± 0.0029</td>
<td>± 0.0058</td>
<td>± 0.0031</td>
</tr>
<tr>
<td>2</td>
<td>0.3904</td>
<td>0.6702</td>
<td>0.1244</td>
<td>0.6666</td>
<td>0.3944</td>
<td>0.6703</td>
<td>0.6759</td>
<td>1.0483</td>
<td>1.0514</td>
<td></td>
</tr>
<tr>
<td></td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0003</td>
<td>± 0.0003</td>
</tr>
<tr>
<td>3</td>
<td>0.1067</td>
<td>0.6245</td>
<td>0.1074</td>
<td>0.1041</td>
<td>0.3928</td>
<td>0.3912</td>
<td>0.6747</td>
<td>0.9923</td>
<td>0.1245</td>
<td>1.0204</td>
</tr>
<tr>
<td></td>
<td>± 0.0006</td>
<td>± 0.0013</td>
<td>± 0.0006</td>
<td>± 0.0006</td>
<td>± 0.0004</td>
<td>± 0.0003</td>
<td>± 0.0003</td>
<td>± 0.0003</td>
<td>± 0.0003</td>
<td>± 0.0003</td>
</tr>
<tr>
<td>4</td>
<td>0.1342</td>
<td>0.0110</td>
<td>0.1476</td>
<td>1.5310</td>
<td>1.5278</td>
<td>0.1443</td>
<td>0.6747</td>
<td>0.3831</td>
<td>0.1338</td>
<td>0.6716</td>
</tr>
<tr>
<td></td>
<td>± 0.0009</td>
<td>± 0.0014</td>
<td>± 0.0008</td>
<td>± 0.0025</td>
<td>± 0.0015</td>
<td>± 0.0025</td>
<td>± 0.0008</td>
<td>± 0.0011</td>
<td>± 0.0009</td>
<td>± 0.0009</td>
</tr>
<tr>
<td>5</td>
<td>0.0128</td>
<td>0.3752</td>
<td>0.3837</td>
<td>0.6636</td>
<td>0.3868</td>
<td>1.5165</td>
<td>0.6707</td>
<td>0.9923</td>
<td>0.1245</td>
<td>1.0204</td>
</tr>
<tr>
<td></td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0013</td>
<td>± 0.0006</td>
<td>± 0.0008</td>
<td>± 0.0005</td>
<td>± 0.0008</td>
</tr>
<tr>
<td>6</td>
<td>1.8014</td>
<td>1.6338</td>
<td>0.3955</td>
<td>0.1246</td>
<td>0.3849</td>
<td>0.3864</td>
<td>0.1314</td>
<td>0.3986</td>
<td>0.1641</td>
<td>1.8021</td>
</tr>
<tr>
<td></td>
<td>± 0.0064</td>
<td>± 0.0045</td>
<td>± 0.0013</td>
<td>± 0.0013</td>
<td>± 0.0013</td>
<td>± 0.0013</td>
<td>± 0.0013</td>
<td>± 0.0014</td>
<td>± 0.0046</td>
<td>± 0.0063</td>
</tr>
<tr>
<td>7</td>
<td>1.0332</td>
<td>0.3950</td>
<td>0.3826</td>
<td>0.1263</td>
<td>0.6551</td>
<td>0.6850</td>
<td>1.0391</td>
<td>0.7013</td>
<td>1.6871</td>
<td>1.8123</td>
</tr>
<tr>
<td></td>
<td>± 0.0009</td>
<td>± 0.0006</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
<td>± 0.0007</td>
<td>± 0.0009</td>
<td>± 0.0007</td>
<td>± 0.0023</td>
<td>± 0.0028</td>
</tr>
<tr>
<td>8</td>
<td>0.1285</td>
<td>0.3713</td>
<td>1.0052</td>
<td>0.3979</td>
<td>0.6711</td>
<td>0.1438</td>
<td>1.5277</td>
<td>1.0543</td>
<td>0.3954</td>
<td>0.3874</td>
</tr>
<tr>
<td></td>
<td>± 0.0004</td>
<td>± 0.0005</td>
<td>± 0.0007</td>
<td>± 0.0005</td>
<td>± 0.0006</td>
<td>± 0.0005</td>
<td>± 0.0013</td>
<td>± 0.0008</td>
<td>± 0.0005</td>
<td>± 0.0005</td>
</tr>
<tr>
<td>9</td>
<td>0.3943</td>
<td>1.0413</td>
<td>1.0745</td>
<td>1.6233</td>
<td>1.0495</td>
<td>0.1434</td>
<td>1.0342</td>
<td>1.0486</td>
<td>0.6925</td>
<td>1.0418</td>
</tr>
<tr>
<td></td>
<td>± 0.0004</td>
<td>± 0.0006</td>
<td>± 0.0006</td>
<td>± 0.0013</td>
<td>± 0.0006</td>
<td>± 0.0003</td>
<td>± 0.0006</td>
<td>± 0.0006</td>
<td>± 0.0004</td>
<td>± 0.0006</td>
</tr>
<tr>
<td>10</td>
<td>0.1314</td>
<td>0.6793</td>
<td>1.0387</td>
<td>0.6793</td>
<td>0.1294</td>
<td>0.1266</td>
<td>0.3779</td>
<td>0.3836</td>
<td>0.6702</td>
<td>0.1315</td>
</tr>
<tr>
<td></td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0003</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
<td>± 0.0002</td>
</tr>
</tbody>
</table>
Chapter 3: Minorization Conditions, Mixture Representations, and Approximate Draws from $\pi$

One solution to Question 1 is to start the Markov chain with initial distribution equal to $\pi$. Then, as long as the chain is invariant for $\pi$, the entire Markov chain has zero total variation distance. This is the ideal and, in some settings, is possible to achieve. In cases where such is not yet possible or practical, a draw from a distribution which closely approximates $\pi$ can be sufficient to answer Question 1. Methods for constructing such approximations with small total variation distance from $\pi$ are discussed in this chapter.

3.1 Minorization Conditions and a Mixture Representation for $\pi$

Let $\pi$ be a measure defined on the space $(X, \mathcal{B}(X))$ and let $\{X_n\}$ be an ergodic Markov chain stationary for $\pi$. All the approximations discussed in this chapter have at their core a minorization condition for $\{X_n\}$. Specifically, a general minorization condition is satisfied for $\{X_n\}$ if

$$K^m(x, A) \geq s(x)Q(A), \quad \forall x \in X, A \in \mathcal{B}(X). \quad (3.1)$$

The general minorization condition in (3.1) is equivalent to (2.7), the minorization condition used for proving geometric ergodicity, where $s(x) = \epsilon 1_C(x)$ and $m = 1$. 

71
For this chapter, assume that $m = 1$; if not, take multiple iterations of the chain and consider them as a single iteration.

The minorization condition (3.1) intuitively results in the following representation of the transition kernel

$$K(x, A) = s(x)Q(A) + (1 - s(x))R(x, A),$$

where

$$R(x, A) = \frac{K(x, A) - s(x)Q(A)}{1 - s(x)}$$

is deemed the residual kernel. This then points to another strategy, discussed by Nummelin (1984), for simulating the Markov chain. Given $X_n = x$ is the current state of the Markov chain, simulate a Bernoulli random variable $\delta_n$ with success probability $s(x)$. If $\delta_n = 1$, take $X_{n+1} \sim Q(\cdot)$; otherwise, take $X_{n+1} \sim R(x, \cdot)$. As $Q$ does not depend on $x$, every time $\delta_n = 1$, the chain regenerates or becomes independent of the previous iterations. The Markov chain $\{X_n, \delta_n\}$ is known as a split chain.

Note that it may not be easy, or even possible, to sample from $Q(\cdot)$ or $R(x, \cdot)$ directly. Fortunately, Mykland et al. (1995) devised a method for sampling the split chain $\{X_n, \delta_n\}$ without these. Assume, for convenience, that $K$ and $Q$ are absolutely continuous with respect to some measure $\mu$ and thus have densities $k$ and $q$ respectively. First, obtain the marginal chain $\{X_n\}$. Then, conditional on the entire chain $\{X_n\}$, $\delta_n$ depends only on $X_n$ and $X_{n+1}$ through the formula

$$P(\delta_n = 1|X_n = x, X_{n+1} = y) = \frac{s(x)q(y)}{k(x, y)}. \quad (3.2)$$

Furthermore, an unnormalized version of $Q$, $Q'$, is acceptable because if

$$K(x, A) \geq s'(x)Q'(A)$$
then

\[ K(x, A) \geq s'(x)Q'(X)Q'(A)/Q'(X) = s(x)Q(A) \]

where \( s(x) = s'(x)Q'(X) \) and \( Q(A) = Q'(A)/Q'(X) \). An unnormalized density \( q \) is then also acceptable.

The state of the split chain immediately before regeneration, \( \alpha = \{X \times 1\} \), is called an accessible atom, which is now formally defined.

**Definition 3.1** (Meyn and Tweedie, 1993). A set \( \alpha \in B(X) \) is an atom for \( \{X_n\} \) if

\[ K(x, A) = \nu(A) \quad \forall x \in \alpha, \]

for some measure \( \nu \). If \( \{X_n\} \) is \( \psi \)-irreducible and \( \psi(\alpha) > 0 \), then \( \alpha \) is an accessible atom.

The times when the split chain enters the accessible atom \( \alpha \) and regenerates are called regeneration times, and, as in Hobert and Robert (2004), define the random time between regenerations as

\[ \tau = \min\{n \geq 1 : (X_n, \delta_n) \in \alpha\}, \]

where \( (X_0, \delta_0) \in \alpha \). Note that it is very simple to obtain i.i.d. copies of \( \tau \). Simply run the split chain starting in \( \alpha \) and count the number of iterations until the next regeneration.

The regeneration times \( \tau \) lead to the construction of a mixture representation of \( \pi \). To develop this, note that according to Kac’s Theorem, if \( \{X_n\} \) is positive recurrent, then \( E(\tau) < \infty \) (Meyn and Tweedie 1993). Then, as in Hobert and Robert (2004), define

\[ p_n = \frac{P(\tau \geq n)}{E(\tau)}, \quad (3.3) \]
which defines a probability mass function on $\mathbb{N}^+$ (recall that for a discrete positive random variable $X$, $E(X) = \sum_{t=1}^{\infty} P(X \geq t)$). Also following Hobert and Robert (2004), define

$$Q_n(A) = P(X_n \in A|\tau \geq n),$$

which is conditional on $(X_0, \delta_0) \in \alpha$ and that no regenerations happen before time $n$. Theorem 1 of Hobert and Robert (2004) is now provided.

**Theorem 3.2** (Mixture Representation of $\pi$, Hobert and Robert, 2004). Assume $K$ is the kernel of a Harris ergodic Markov chain $\{X_n\}$ which has $\pi$ as its invariant measure. Assume that the minorization condition (2.7) holds. Then, for all $A \in \mathcal{B}(X)$,

$$\pi(A) = \sum_{n=1}^{\infty} p_n Q_n(A). \tag{3.4}$$

This theorem was also proven in Hobert et al. (2006) for the general minorization condition (3.1). Typically, the general minorization condition in (3.1) results in more frequent regenerations of the chain and a smaller expected value of $\tau$ compared to that in (2.7) (Jones and Hobert 2001a).

Given this mixture representation, several algorithms may be considered that draw from $\pi$ by drawing from $p_n$ and then from $Q_n(\cdot)$. Thus, this mixture representation reduces drawing exactly from $\pi$, regardless of the number of dimensions, to making a draw from a univariate discrete distribution $p_n$ coupled with a draw from $Q_N$. This method may be used to sample from large-dimensional distributions exactly, which is often challenging. The efficiency of such algorithms is based entirely on the distribution of $\tau$; thus, the extent to which a good minorization condition can be found for $\{X_n\}$ determines the efficiency of the algorithm, regardless of the dimension of $\pi$. 

74
3.2 Obtaining Independent Draws Exactly or Approximately from $\pi$

In cases where exact draws from $p_n$ are attainable, exact draws from $\pi$ result, as in the first two algorithms now discussed. When approximate draws from $p_n$ are obtained, as in the remaining algorithms presented in this section, approximate draws from $\pi$ result.

3.2.1 Exact Sample from $\pi$

The first algorithm for exactly sampling from $\pi$ is a direct result of the mixture representation.

**Algorithm 3.1** (Exact Sample from $\pi$)

Assume the conditions of Theorem 3.2 hold.

(i) Sample a variate $N$ according to the distribution $P(N = n) = p_n$ for $n \geq 1$.

(ii) Given $N$, sample $X$ from the distribution $Q_N$.

Step (ii) of Algorithm 3.1, sampling from $Q_N$ given $N$, is straightforward. Simply run the split chain until $\delta_1 = \cdots = \delta_{N-1} = 0$ and then accept $X_N$ (Flegal and Herbei 2012). Blanchet and Meng (2005) showed, however, that this rejection method has infinite expected running time and is therefore not efficient. This is a drawback of any algorithm that requires a sample from $Q_N$ given $N$ which is the case for all but one of the algorithms discussed in this chapter (Algorithm 3.4).

When the underlying Markov chain $\{X_n\}$ is uniformly ergodic, Step (i) of Algorithm 3.1 is simple because $p_n$ has a closed form. Specifically,

$$P(\tau \geq n) = (1 - \epsilon)^n,$$
where $\epsilon$ comes from the minorization condition (2.7) (Hobert and Robert 2004). This implies that $E(\tau) = 1/\epsilon$, and therefore

$$p_n = P(N = n) = \epsilon(1 - \epsilon)^n,$$

and thus $N$ has a geometric distribution from which samples are readily available. Thus, perfect samples may be obtained via Algorithm 3.1. Also in the uniformly ergodic case, Algorithm 3.1 is equivalent to other perfect sampling algorithms including the Multigamma Coupler of Murdoch and Green (1998) and Wilson’s (2000) Read-Once algorithm.

### 3.2.2 Exact Sample from $\pi$ when the Markov Chain is Geometrically Ergodic

If $\{X_n\}$ is not uniformly ergodic, the distribution of $p_n$ is not available in closed form, which limits the direct use of Algorithm 3.1. Despite this, Flegal and Herbei (2012) developed a perfect sampling algorithm for when $\{X_n\}$ is geometrically ergodic satisfying a drift (2.8) and associated minorization (2.7) condition. To accomplish this, they develop an envelope function for $p_n$ and utilize rejection sampling. The envelope is constructed using the bounding argument of Hobert and Robert (2004) which is based on results in Roberts and Tweedie (1999) and was used to bound the total variation distance of a Markov chain in Chapter 2. Specifically, let

$$A = \sup_{x \in C} E[V(X_{i+1})|X_i = x], \quad \text{and} \quad J = (A - \epsilon)/\lambda,$$

with

$$\beta^* = \begin{cases} 
\lambda^{-1} & \text{if } J < 1 \\
\exp \left\{ \log \lambda \log (1 - \epsilon) \over \log J - \log (1 - \epsilon) \right\} & \leq \lambda^{-1} \quad \text{if } J \geq 1,
\end{cases} \quad (3.5)$$
where $\epsilon$, $\lambda$, and the set $C$ are all provided by the drift and minorization conditions. Choose $\beta \in (1, \beta^*)$ and define $\phi(\beta) = \log \beta / \log \lambda^{-1}$. Then

$$P(\tau \geq n) \leq \beta \left[ \frac{b}{\epsilon(1-\lambda)} \right]^{\phi(\beta)} \left[ \frac{1 - \beta(1-\epsilon)}{1 - (1-\epsilon)(J/(1-\epsilon))^{\phi(\beta)}} \right] \beta^{-n} \tag{3.6}$$

$$= Md(n),$$

where $d(n) = \beta^{-n}$ and

$$M = \beta \left[ \frac{b}{\epsilon(1-\lambda)} \right]^{\phi(\beta)} \left[ \frac{1 - \beta(1-\epsilon)}{1 - (1-\epsilon)(J/(1-\epsilon))^{\phi(\beta)}} \right].$$

This inequality provides an envelope function for $P(\tau \geq n)$ and induces a rejection sampling algorithm. A Bernoulli random variable $B$ with success probability

$$P(B = 1) = (M \beta^{-n})^{-1} P(\tau \geq n)$$

is required to determine acceptance of a candidate $n$ from the envelope. If

$$(M \beta^{-n})^{-1} \leq 1,$$

then $B$ is easily drawn by multiplying a Bernoulli random variable with probability $(M \beta^{-n})^{-1}$ by a Bernoulli random variable with probability $P(\tau \geq n)$, both of which are easily obtained. If, however,

$$(M \beta^{-n})^{-1} > 1,$$

a Bernoulli factory is required.

A Bernoulli factory is an algorithm whereby a draw from a Bernoulli($f(p)$) is obtained using draws from a Bernoulli($p$) where $f$ is some function (see Keane and O’Brien (1994) for a discussion of restrictions on $f$ such that a Bernoulli factory exists). In this case, a Bernoulli $f(p) = ap$ is desired where $a > 1$. The algorithm of
Flegal and Herbei (2012) requires that $ap < 1 - \omega$, where $\omega < 1$ is a chosen constant. To ensure this, uniformly decrease the probability of acceptance by noting that

$$P(\tau \geq n) \leq Md(n)\kappa$$

where $\kappa > 1$. Then, choose $\kappa$ such that $\kappa^{-1} < 1 - \omega$, which ensures that

$$(Md(n)\kappa)^{-1}P(\tau \geq n) \leq \kappa^{-1} < 1 - \omega,$$

as desired. The rejection sampling algorithm of Flegal and Herbei (2012), now presented as Algorithm 3.2, produces an exact draw $N$ from $p_n$ and then draws from $Q_N$ to produce a perfect draw from $\pi$. Note that the Geometric$(p)$ distribution in the algorithm has probability mass function

$$P(N = n) = p(1 - p)^{n-1}, \quad n = 1, 2, \ldots$$

with expectation $E(N) = 1/p$.

**Algorithm 3.2** (Exact sample from $\pi$ when drift and minorization are satisfied)

Choose $\omega < 1$ and $\kappa > 1$ such that $\kappa^{-1} < 1 - \omega$. Compute $\beta^*$ as in (3.5) and choose $\beta \in (1, \beta^*)$; compute $M$ as in (3.6).

(i) Draw $N \sim \text{Geometric}(1 - 1/\beta)$. Set $a = [Md(N)\kappa]^{-1}$.

(ii) If $a \leq 1$, draw a $\tau$ random variable. Let $W = 1(\tau \geq N)$ and draw

$$V \sim \text{Bernoulli}(a). \quad \text{Set } B = VW.$$

(iii) If $a > 1$, use the Bernoulli factory (see Flegal and Herbei 2012) to obtain

$$B \sim \text{Bernoulli}(aP(\tau \geq N)).$$

(iv) If $B = 1$, accept $N$; if $B = 0$, return to Step (i).
(v) Draw from $Q_N(\cdot)$.

Exact draws using Algorithm 3.2, however, can come at a computational price. For the one way random effects model, similar to that of the example in Section 2.5.1 with the styrene exposure dataset, an average of 7.2 processor days was required for each i.i.d. exact draw from the posterior distribution (Flegal and Herbei 2012).

This brings to light the trade-off between accuracy and computational speed. The remaining algorithms discussed in this chapter produce approximate draws, but with a generally smaller computational burden. For each situation, this trade-off should be considered. Often, the optimal practice may be to obtain nearly exact draws with smaller computational cost.

### 3.2.3 Approximation by Truncation

The next algorithm considered was established by Hobert and Robert (2004). This algorithm approximates the distribution of $p_n$ with a truncated version $\tilde{p}_n$ supported on the set $\{1, \ldots, M\}$. Specifically,

$$
\tilde{p}_n = \begin{cases} 
  \frac{P(\tau \geq n)}{\sum_{i=1}^{M} P(\tau \geq i)} & \text{if } n \in \{1, \ldots, M\} \\
  0 & \text{otherwise}.
\end{cases} \quad (3.7)
$$

To make a draw from $\tilde{p}_n$, proceed as in Hobert and Robert (2004). Consider the random variable $V$ uniformly distributed on $\{1, \ldots, M\}$ and $W$ independently distributed according to $\tau$. Note that for $n \in \{1, \ldots, M\}$

$$
P(V = n|W \geq V) = \frac{P(V = n, W \geq V)}{P(W \geq V)} = \frac{P(W \geq V|V = n)P(V = n)}{\sum_{i=1}^{M} P(W \geq V|V = i)P(V = i)} = \frac{P(W \geq n)M^{-1}}{\sum_{i=1}^{M} P(W \geq i)M^{-1}}
$$

79
\[ P(\tau \geq n) \] 
\[ \sum_{i=1}^{M} P(\tau \geq i) \]
which matches the probability mass function \( \tilde{p}_n \). This then inspires the rejection sampler where \( V \) and \( W \) are jointly drawn and \( V \) is accepted as a draw from the probability mass function \( \tilde{p}_n \) if \( W \geq V \). The formal algorithm combines this rejection sampler with a draw from \( Q_N \) to result in an approximate draw from \( \pi \).

**Algorithm 3.3** (Approximate Sampler with Truncated Distribution of \( p_n \))

Choose \( M \).

(i) Simulate \( v \sim \text{Uniform}(1, \ldots, M) \). Independently simulate \( w \) according to the distribution of \( \tau \).

(ii) If \( w \geq v \), then accept \( v \) and set \( N = v \). Otherwise, return to step (i).

(iii) Draw from \( Q_N(\cdot) \).

As with any truncation, \( \tilde{p}_n \) becomes closer to \( p_n \) as the set of truncation becomes larger, which corresponds to increasing \( M \). This will now be shown formally by computing the total variation distance between \( p_n \) and \( \tilde{p}_n \). Note that for any \( n \leq M \),

\[
|p_n - \tilde{p}_n| = \left| \frac{P(\tau \geq n)}{E(\tau)} - \frac{P(\tau \geq n)}{\sum_{i=1}^{M} P(\tau \geq i)} \right|
\]

\[
= P(\tau \geq n) \left| \frac{\sum_{i=1}^{M} P(\tau \geq i) - E(\tau)}{E(\tau) \sum_{i=1}^{M} P(\tau \geq i)} \right|
\]

\[
= \frac{P(\tau \geq n) \sum_{i=M+1}^{\infty} P(\tau \geq i)}{E(\tau) \sum_{i=1}^{M} P(\tau \geq i)}. \]

If \( n > M \), then \( \tilde{p}_n = 0 \), which implies

\[
|p_n - \tilde{p}_n| = \frac{P(\tau \geq n)}{E(\tau)}. \]
This implies that

\[ ||p_n - \tilde{p}_n||_{TV} = \frac{1}{2} \sum_{n=1}^{\infty} |p_n - \tilde{p}_n|, \]

which is an alternate definition of TV distance for discrete random variables,

\[
\begin{align*}
&= \frac{1}{2} \left[ \sum_{i=M+1}^{\infty} \frac{P(\tau \geq i)}{E(\tau) \sum_{i=1}^{M} P(\tau \geq i)} \sum_{n=1}^{M} P(\tau \geq n) + \frac{1}{E(\tau)} \sum_{n=M+1}^{\infty} P(\tau \geq n) \right] \\
&= \frac{1}{2} \left[ \frac{1}{E(\tau)} \sum_{i=M+1}^{\infty} P(\tau \geq i) + \frac{1}{E(\tau)} \sum_{n=M+1}^{\infty} P(\tau \geq n) \right] \\
&= \frac{1}{E(\tau)} \sum_{i=M+1}^{\infty} P(\tau \geq i),
\end{align*}
\]

which clearly decreases as \( M \) increases. Note, as in Hobert and Robert (2004), that the total variation distance between \( \pi \) and the output of Algorithm 3.3, denoted by \( \tilde{\pi} \), satisfies

\[
||\pi(\cdot) - \tilde{\pi}(\cdot)||_{TV} = \sup_{A \in \mathcal{B}(\mathcal{X})} |\pi(A) - \tilde{\pi}(A)|
\]

\[
= \sup_{A \in \mathcal{B}(\mathcal{X})} \left| \sum_{n=1}^{\infty} p_n Q_n(A) - \sum_{n=1}^{\infty} \tilde{p}_n Q_n(A) \right|
\]

\[
= \sup_{A \in \mathcal{B}(\mathcal{X})} \left| \sum_{n=1}^{\infty} (p_n - \tilde{p}_n) Q(A) \right|
\]

\[
\leq \sup_{A \in \mathcal{B}(\mathcal{X})} \sum_{n=1}^{\infty} |p_n - \tilde{p}_n| Q(A)
\]

\[
= \sum_{n=1}^{\infty} |p_n - \tilde{p}_n|,
\]

because the \( \sup_{A \in \mathcal{B}(\mathcal{X})} Q(A) = 1 \),

\[
= 2||p_n - \tilde{p}_n||_{TV}
\]

by the definition of TV distance for discrete random variables

\[
= \frac{2}{E(\tau)} \sum_{i=M+1}^{\infty} P(\tau \geq i),
\]
using the result from (3.8). Thus, the TV distance between $\pi$ and $\tilde{\pi}$ is less than twice the TV distance between $p_n$ and $\tilde{p}_n$. This result in (3.9) will apply for all the approximations in this chapter.

Given these results, a key to the successful implementation of Algorithm 3.3 is the selection of a sufficiently large $M$ such that the level of approximation is small enough. When the overall Markov chain is geometrically ergodic and satisfies a drift and minorization condition, Hobert and Robert (2004) discuss a method for selecting a large enough $M$ for the total variation distance to be below a certain threshold $\gamma$. There is, however, no specific guidance when the chain is not geometrically ergodic.

As $M$ increases and the approximation improves, the computational burden also increases. Specifically, the expected number of proposals required to accept a draw from $\tilde{p}_n$ increases as $M$ increases. To see this, define the acceptance probability of Algorithm 3.3, $a_M$, as

$$a_M = P(\tau \geq V) = \frac{1}{M} \sum_{i=1}^{M} P(\tau \geq i),$$

(3.11)

where $V \sim \text{Uniform}(1, \ldots, M)$. It’s simple to see that $a_M$ becomes smaller as $M$ increases because $P(\tau \geq i)$ is nonincreasing with increasing $i$. The expected number of proposals is $a_{M}^{-1}$, which increases with increasing $M$. Therefore, $M$ must be selected to strike a balance between accuracy and computational feasibility.

To aid in assessing the trade-off between total-variation distance and the acceptance rate of Algorithm 3.3, an expression for $||p_n - \tilde{p}_n||_{TV}$ may be obtained using the acceptance probability $a_M$. Combining (3.8) with (3.11), it is observed that

$$||p_n - \tilde{p}_n||_{TV} = \frac{1}{E(\tau)} \sum_{i=M+1}^{\infty} P(\tau \geq i) = \frac{E(\tau) - \sum_{i=1}^{M} P(\tau \geq i)}{E(\tau)}$$

82
Thus the total variation distance is a function of the acceptance rate $a_M$ and the expected value of the waiting times, both of which are easily estimated in the process of running the algorithm. Therefore, (3.12) may be used to approximate a bound for the total variation distance of Algorithm 3.3 from $\pi$, multiplying by 2 as in (3.9).

### 3.2.4 Modified Approximation by Truncation

A small alteration to Algorithm 3.3 eliminates the need to make a draw from $Q_N$, which is highly desirable given that the typical rejection sampler for sampling from $Q_N$ has infinite expected running time as previously discussed.

**Algorithm 3.4** (Modified Approximate Sampler with Truncated Distribution of $p_n$)

Choose $M$.

(i) Simulate $v \sim \text{Uniform}(1,,\ldots,M)$. Independently simulate $w$ according to the distribution of $\tau$ and retain the Markov chain $X_0, X_1, \ldots, X_w$ used to generate $w$. Note that $X_0$ and $X_w$ both lie in the accessible atom $\alpha$.

(ii) If $w \geq v$, set $N = v$. Otherwise, return to step (i).

(iii) Accept $X_N$ as an approximate draw from $\pi$.

Now, to prove that this adjustment maintains the same distribution as Algorithm 3.3, note that

$$
P(X_N \in A, V = N | W \geq V) = \frac{P(X_N \in A, V = N, W \geq V)}{P(W \geq V)}
= \frac{P(X_N \in A, W \geq V | V = N)P(V = N)}{P(W \geq V)}
$$
which is equivalent to using Algorithm 3.3 to draw from \( \tilde{p}_n \) and then follow up with a draw from \( Q_N(\cdot) \). Thus, 3.4 provides an approximation without infinite expected running time. In fact, the computation required is the same as that of Algorithm 3.3 before the draw from \( Q_N(\cdot) \), which had acceptance rate \( a_M \) defined in (3.11). This is the only such approximate algorithm using the mixture representation of \( \pi \) discussed in this chapter which has finite expected running time, making it an excellent candidate for practical use.

### 3.2.5 Approximation via Empirical Distribution

Hobert et al. (2006) provides another approximation for \( p_n \). Let \( \tau_1, \ldots, \tau_m \) be an i.i.d. sample from the distribution of \( \tau \), the waiting times between regenerations. Let \( F_m \) be the empirical distribution function corresponding to \( \tau_1, \ldots, \tau_m \), i.e.

\[
F_m(n) = m^{-1} \sum_{i=1}^{m} \mathbb{1}(\tau_i \leq n).
\]

Then, estimate \( p_n \) with

\[
\hat{p}_n = \frac{1 - F_m(n - 1)}{\bar{\tau}},
\]

(3.13)

where \( \bar{\tau} \) is the sample mean of the \( \tau \) variates. As \( \sum_{i=1}^{\infty} \hat{p}_i = 1 \), \( \hat{p}_n \) is a mass function from which samples may readily be taken. This is summarized in Algorithm 3.5.

**Algorithm 3.5** (Approximate Sampler using Empirical Distribution to Estimate \( p_n \))

Choose an empirical sample size \( m \).
(i) Obtain $m$ i.i.d. samples from $\tau$ and call them $\tau_1, \ldots, \tau_m$.

(ii) Create the empirical distribution $F_m$ based on $\tau_1, \ldots, \tau_m$ and compute the sample average $\bar{\tau}$.

(iii) Compute the empirical estimate for $p_n, \hat{p}_n$, as in (3.13).

(iv) Sample $N$ from $\hat{p}_n$.

(v) Draw from $Q_N$.

(vi) Repeat steps (iv) and (v) as desired.

Intuitively, empirical approximations improve as the number of i.i.d. variates upon which they are based increases. Indeed, Hobert et al. (2006) show that Algorithm 3.5 improves asymptotically as $m$ increases with the total variation distance between $p_n$ and $\hat{p}_n$ converging to 0 as $m \to \infty$. While it is not possible to directly compute $m$ such that the total variation distance between $p_n$ and $\hat{p}_n$ is guaranteed to be below some threshold, Hobert et al. (2006) show that it is possible to compute $m$ such that the total variation distance is below a specified threshold with high probability.

3.3 Markov Chains Stationary for $p_n$

While the previous section described approaches established in the literature for sampling exactly or approximately from $\pi$ using the mixture representation, this section presents three new approaches for obtaining approximate draws from $p_n$ and therefore from the distribution of $\pi$. Each proceeds by constructing a Markov chain stationary for $p_n$. 
3.3.1 Reverse Rejection Sampler

The first new algorithm was devised with a rejection sampler at its foundation. If samples from some density \( f \) are desired, but only samples from a density \( g \) are easily obtained, if \( f(x)/g(x) \leq M \) for some constant \( M < \infty \) and all \( x \in X \), then a rejection sampler may be created. This sampler proposes a state \( x \) from \( g \) and then accepts it with probability \( f(x)/(Mg(x)) \). For the intuition behind the reverse rejection sampler, consider the case where \( X \) is discrete. Before a certain state \( x \) is accepted, it may need to be proposed several times by \( g \). In fact, the random number of times a state \( x \) is proposed by \( g \) before it is accepted as a draw from \( f \) follows a geometric distribution with probability of success \( f(x)/(Mg(x)) \) with expected value \( Mg(x)/f(x) \).

This then leads to a way to sample from \( g \) if only draws from \( f \) are attainable, which may be termed a reverse rejection sampler.

**Algorithm 3.6** (Reverse Rejection Sampler)

Assume \( f(x)/g(x) \leq M \) and draws from \( f \) are available.

(i) Draw a state \( x \) from \( f \).

(ii) Draw \( w \) from a geometric distribution with success probability \( f(x)/(Mg(x)) \).

(iii) Replicate the state \( x \) for \( w \) times.

(iv) Repeat by returning to step (i).

The resulting draws are correlated, but follow the probability law of \( g \) when combined together. This is most easily proved by considering the resulting draws as a Markov chain, which is now done formally. By abuse of notation, let \( f \) and \( g \) each
represent both the probability measures and probability density or mass functions.

Let the Markov transition density, which corresponds to a kernel $K$, be

$$k(x, y) = f(y)\frac{f(x)}{Mg(x)} + \left(1 - \frac{f(x)}{Mg(x)}\right)\delta_x(y). \quad (3.14)$$

The corresponding algorithm follows.

**Algorithm 3.7** (Reverse Rejection Sampler Markov Chain)

From the current state $X_n = x$, update as follows:

(i) Draw a new state $y \sim f$.

(ii) Draw $B \sim \text{Bernoulli}(f(x)/(Mg(x)))$.

(iii) If $B = 1$, accept the candidate and set $X_{n+1} = y$; otherwise, set $X_{n+1} = x$.

It is straightforward to see that Algorithm 3.6 and Algorithm 3.7 are equivalent when $X_0 \sim f$. Thus, to show that Algorithm 3.6 is stationary for $g$, it remains to show that Algorithm 3.7 is stationary for $g$. For this, first recall the property of detailed balance from Definition 2.11 and that its establishment ensures stationarity by Theorem 2.13.

**Proposition 3.3.** Algorithm 3.7 is stationary for $g$.

**Proof.** Proceed by establishing detailed balance. The kernel for the Markov chain corresponding to Algorithm 3.7 is provided in (3.14). Note that

$$g(x)k(x, y) = g(x)f(y)\frac{f(x)}{Mg(x)} + g(x)\left(1 - \frac{f(x)}{Mg(x)}\right)\delta_x(y)$$

$$= f(y)\frac{f(x)}{M} + g(y)\left(1 - \frac{f(y)}{Mg(y)}\right)\delta_y(x)$$

87
because if \( \delta_x(y) = 1 \) then \( g(x) = g(y) \)

\[
= g(y)f(x) \frac{f(y)}{Mg(y)} + g(y) \left( 1 - \frac{f(y)}{Mg(y)} \right) \delta_y(x)
= g(y)k(y, x),
\]

and thus detailed balance is satisfied. Therefore, by Theorem 2.13, \( g \) is the stationary distribution of the Markov chain, as desired. Additionally, due to their equivalence, Algorithm 3.6 is also stationary for \( g \).

Thus far, the discussion of the reverse rejection sampler has applied to general \( g \) and \( f \); a specific case is now provided to draw from \( p_n \). Let

\[
g(n) = p_n = \frac{P(\tau \geq n)}{E(\tau)},
\]

\[
f(n) = P(\tau = n), \text{ and } M = E(\tau).
\]

Then,

\[
\frac{f(n)}{Mg(n)} = \frac{P(\tau = n)}{P(\tau \geq n)} \leq 1,
\]

which implies that the reverse rejection sampler may be used to obtain a Markov chain stationary for \( p_n \) using \( i.i.d \) copies of \( \tau \), which are readily available. The algorithm results in an approximation to \( p_n \) in the same sense that any Markov chain simulates approximately from its stationary distribution. The draws from \( p_n \) may then be combined with updates from \( Q_n \) to construct an approximate algorithm to sample from \( \pi \). Formally, the Markov chain stationary for \( p_n \), \( \{ P_n \} \) is as follows.

**Algorithm 3.8** (Reverse Rejection Sampler Markov Chain Stationary for \( p_n \))

Given \( P_n = n \), do the following:

(i) Draw \( n^* \) according to the distribution of waiting times, \( \tau \).
(ii) Define $\alpha(n)$ as

$$\alpha(n) = \frac{P(\tau = n)}{P(\tau \geq n)}, \quad (3.15)$$

(iii) Set

$$P_{n+1} = \begin{cases} n^* & \text{w.p. } \alpha(n) \\ n & \text{w.p. } 1 - \alpha(n). \end{cases}$$

Note that, while $\alpha(n)$ is not computable in general, a Bernoulli random variable $B$ with probability $\alpha(n)$ is easily obtained. Simply draw $\tau$ until $\tau \geq n$; if $\tau = n$, then $B = 1$; else, $B = 0$.

Also note that the probability of acceptance of a new state $n^*$ does not depend on the new state $n^*$, but only on the old state $n$. This means the new state only needs to be drawn when the old state is rejected. This also implies that a minorization condition (3.1) is easily established with $Q$ as the distribution of waiting times and $s(n) = \alpha(n)$, the acceptance probability. Therefore, the chain regenerates each time a new state is entered. Furthermore, if conditions could be established such that $\inf_n \alpha(n) = \epsilon > 0$ then the chain would be uniformly ergodic and exact samples would be possible.

Other Markov chains stationary for $p_n$ may also be developed as approximations for $p_n$.

### 3.3.2 Random Walk Metropolis

The biggest challenge for constructing a Markov chain stationary for $p_n$ is the fact that $p_n$ is not directly computable. Thus, most acceptance probabilities for new candidate states are not available in closed form, though, as with the reverse rejection sampler Markov chain algorithm, it may be possible to generate a Bernoulli random variable with the desired acceptance probability.
A random walk Metropolis Markov chain stationary for $p_n$ with a discrete symmetric proposal distribution $q$ centered at the current state fits this description. In this case, the acceptance probability of a new state $n^*$ from the current state $n$ is

$$\alpha(n, n^*) = \min\left(1, \frac{P(\tau \geq n^*)}{P(\tau \geq n)}\right).$$

When $n^* \leq n$, then $\alpha(n, n^*) = 1$ because $P(\tau \geq n^*) \geq P(\tau \geq n)$. When $n^* > n$, then $\alpha(n, n^*) \leq 1$, but it is simple to draw a Bernoulli with probability $\alpha(n, n^*)$: draw i.i.d. copies of $\tau$ until $\tau \geq n$ and call this $\tau^*$; if $\tau^* \geq n^*$, then accept the new state $n^*$; otherwise (if $\tau^* < n^*$), keep the current state $n$. Thus, the Markov chain may proceed without directly computing $\alpha(n, n^*)$.

The possibilities for the symmetric candidate distribution abound. Any discrete distribution reflected about 0 such that symmetry results would satisfy the requirements. One example is a uniform distribution $d$ states below and $d$ states above the current state. Specifically, where $n$ is the current state, the mass function would be

$$q_1(n^*|n) = \begin{cases} \frac{1}{2d} & n^* \in \{n - d, n - d + 1, \ldots, n - 1, n + 1, \ldots, n + d\} \\ 0 & \text{otherwise}. \end{cases}$$

Another option is a symmetric mixture of geometric random variables with parameter $p$ defined above and below the current state $n$ as

$$q_2(n^*|n) = \begin{cases} 0.5(1 - p)^{|n^*-n|-1}p & n^* \in \{n \pm 1, n \pm 2, \ldots\} \\ 0 & n^* = n. \end{cases}$$

Either of these candidate distributions, and others that could be proposed, can be tuned by carefully selecting the governing parameters (e.g. $d$ for $q_1$ and $p$ for $q_2$) to result in improved convergence. With a selected candidate distribution, the algorithm may be formally defined.
Algorithm 3.9 (Random Walk Stationary for \( p_n \))

Let \( q \) be a discrete symmetric candidate distribution centered at the current value \( n \).

(i) Draw a candidate \( n^* \) from \( q(n^*|n) \).

(ii) If \( n^* < n \), accept \( n^* \).

(iii) If \( n^* > n \), draw i.i.d. \( \tau \) variates until \( \tau \geq n \) and call it \( \tau^* \). If \( n^* \geq \tau^* \), accept \( n^* \); else, keep the current state \( n \).

A final Markov chain stationary for \( p_n \) now follows.

3.3.3 Metropolis-Hastings via The Pseudo-Marginal Approach

In the event that a Metropolis-Hastings proposal \( q(n^*|n) \) which is not symmetric is advantageous, for an independence sampler for example, the Metropolis Hastings acceptance probability is

\[
\alpha(n, n^*) = \min \left( 1, \frac{P(\tau \geq n^*)q(n|n^*)}{P(\tau \geq n)q(n^*|n)} \right). \tag{3.18}
\]

Suppose also that there is no way to generate a Bernoulli variate with success probability \( \alpha(n, n^*) \). One additional approach, however, may still be employed to obtain a Markov chain to successfully approximate \( p_n \). Consider replacing \( P(\tau \geq n) \) and \( P(\tau \geq n^*) \) in the acceptance probability with unbiased estimates, denoted \( \bar{p}_n \) and \( \bar{p}_{n^*} \), respectively (e.g. \( \bar{p}_n = M^{-1} \sum_{i=1}^{M} \mathbb{1}(\tau_i \geq n) \)) where \( \tau_1, \ldots, \tau_M \) is an i.i.d. sample of the waiting times \( \tau \).

This method has been discussed by Beaumont (2003) and Andrieu and Roberts (2009) and is dubbed pseudo-marginal MCMC. The acceptance probability is thus

\[
\alpha(n, n^*) = \min \left( 1, \frac{\bar{p}_{n^*}q(n|n^*)}{\bar{p}_n q(n^*|n)} \right). \tag{3.19}
\]
The stationary distribution of this Markov chain is now determined. Define

$$\omega = \frac{\bar{p}_n}{P(\tau \geq n)},$$

(3.19)

and let $p(\omega|n)$ be the density of $\omega$ for a particular $n$ such that

$$E(\omega) = \int \frac{\bar{p}_n}{P(\tau \geq n)} p(\omega|n) \, d\omega = \frac{1}{P(\tau \geq n)} E(\bar{p}_n) = 1,$$

(3.20)

which is true because $\bar{p}_n$ is unbiased for $P(\tau \geq n)$. Then,

$$\alpha(n, n^*) = \min \left( 1, \frac{\bar{p}_n q(n|n^*)}{p(n|n)} \right)$$

$$= \min \left( 1, \frac{\omega^* P(\tau \geq n^*) q(n^*|n^*)}{\omega P(\tau \geq n) q(n^*|n)} \right)$$

because of (3.19),

$$= \min \left( 1, \frac{\omega^* P(\tau \geq n^*) q(n^*|n^*)}{\omega P(\tau \geq n) q(n^*|n)} \right)$$

$$= \min \left( 1, \frac{\bar{p}_n q(n^*|n)}{\pi'(n^*, \omega^*) q'(n^*, \omega^*|n, \omega)} \right),$$

where

$$\pi'(n^*, \omega^*) \propto \omega^* P(\tau \geq n^*) p(\omega^*|n^*),$$

and

$$q'(n, \omega|n^*, \omega^*) = q(n^*|n^*) p(\omega|n),$$

which implies that the stationary distribution is $\pi'(n, \omega)$, the joint distribution between $n$ and the Monte Carlo error $\omega$, and the candidate distribution is $q'(n^*, \omega^*|n, \omega)$. Integrating out the Monte Carlo error $\omega$ from the stationary distribution to obtain the marginal distribution for $n$ yields

$$\int \pi'(n, \omega) \, d\omega \propto \int \omega P(\tau \geq n) p(\omega|n) \, d\omega = P(\tau \geq n),$$

(3.21)
as $E(\omega) = 1$, which is proportional to the desired target distribution

$$p_n = P(\tau \geq n)/E(\tau).$$

Thus, the marginal distribution of $n$ for this Markov chain is the desired result.

One consequence of the fact that the stationary distribution is a joint distribution between $\omega$ and $n$ is that $\omega^*$ is accepted or rejected each iteration just like a new state $n^*$. Thus, $\omega$ must remain the same from the end of one iteration to the beginning of the next. This means that the approximation for $P(\tau \geq n)$, $p_n$, is also the same between one iteration and the next. This is crucial as not using the same approximation for the current state results in a different stationary distribution (see Andrieu and Roberts (2009) for additional discussion). Altogether, this algorithm is now presented.

**Algorithm 3.10** (Pseudo-Marginal Metropolis-Hastings Algorithm Marginally Stationary for $p_n$)

Let $q(n^*|n)$ be a non-symmetric proposal distribution. The current state is $n$ with $\overline{p}_n$, an unbiased estimate of $P(\tau \geq n)$.

(i) Draw a candidate $n^*$ from $q(n^*|n)$.

(ii) Compute $\overline{p}_{n^*}$, the unbiased Monte Carlo estimate for $P(\tau \geq n^*)$.

(iii) Compute the acceptance probability

$$\alpha(n, \overline{p}_n; n^*, \overline{p}_{n^*}) = \min \left(1, \frac{\overline{p}_{n^*} q(n|n^*)}{\overline{p}_n q(n^*|n)} \right).$$

(iv) Accept $n^*$ and $\overline{p}_{n^*}$ with probability $\alpha(n, \overline{p}_n; n^*, \overline{p}_{n^*})$; otherwise, retain $n$ and $\overline{p}_n$.

Examples illustrate these Markov chains in practice.
3.4 Examples of Approximate Algorithms

The approximation algorithms, both independent and Markovian, are now illustrated with two examples. The first is where the distribution of \( p_n \) is known, and therefore the total variation distance of each algorithm may be formally computed and compared with that of the other algorithms. The second is a realistic example using the Bayesian one-way random effects model using the styrene data discussed in Section 2.5.1.

3.4.1 Known Distribution of \( p_n \)

For illustration purposes, a known distribution of \( \tau \) is utilized so \( p_n \) may be computed directly to formally assess the approximate algorithms. Select a Poisson distribution with \( \lambda = 5 \) as the distribution of the regeneration times \( \tau \), and shift it forward so it is defined beginning with 1 rather than 0. Figure 3.1 depicts the distribution of \( \tau \) (on the left) and the corresponding probability mass function \( p_n \) (on the right). Each approximate algorithm is now illustrated.

First, consider Algorithm 3.3. As previously discussed, the acceptance rate and the total variation distance have explicit formulas for this sampler. A plot of the TV distance (from \( p_n \)) and the acceptance rate for truncation points \( M = 1, \ldots, 30 \) is shown in Figure 3.2. Note that the total variation distance decreases much more swiftly than the acceptance rate, implying that a small TV distance may be achieved without overwhelming computational requirements. For example, the TV distance from \( p_n \) is below 0.005 with \( M = 11 \) (TV distance from \( p_n \) of less than 0.005 corresponds to a bound on the TV distance from \( \pi \) of 0.01). The corresponding acceptance rate is 0.543, which implies essentially no computational burden. Reducing that threshold
Figure 3.1: Distribution of \( \tau \) (left) and corresponding probability mass function \( p_n \) (right) for the known distribution of \( p_n \) example.

Figure 3.2: TV distance and acceptance rate for several values of \( M \) for Algorithm 3.3.
by a factor of 100 to 0.00005 implies that \(M = 16\) should be used which still has acceptance rate of 0.375. This algorithm therefore appears to be highly efficient computationally. Reducing the TV distance, in this case dramatically, is simply a matter of selecting a larger truncation point, which results in a relatively small increase in computational burden. Note that Algorithm 3.4 would produce the same results.

In contrast to Algorithm 3.3 where the TV distance is fixed given a truncation bound \(M\), the TV distance of Algorithm 3.5 varies based on the individual samples taken given a fixed \(m\). This sample-to-sample variability is shown in Figure 3.3 through 95% bounds based on 10,000 samples (dashed lines) in addition to the average (solid). In this example, \(m = 10,000\) is the smallest value tested which resulted in an average TV distance (from \(p_n\)) below 0.005. In this case, the upper 95\% band is also less than 0.005. Thus, the computational burden for a single good draw is much higher for Algorithm 3.5 compared to Algorithm 3.3. However, an advantage of Algorithm 3.5 is that once the empirical distribution is obtained, additional independent draws are trivial to obtain. Thus, if a large number of independent draws are desired, Algorithm 3.5 may be more efficient than Algorithm 3.3.

The reverse rejection sampler Markov chain, Algorithm 3.8, is now illustrated. The total variation distance from \(p_n\), depicted in Figure 3.4, decreases with increasing iterations, as expected; however, the rate of decline is somewhat slow, particularly when compared to the previous approximations and the other Markov chains which follow in Figures 3.5 and 3.7. The algorithm requires 427 iterations before the TV distance is below the 0.005 threshold.
Figure 3.3: TV distance of Algorithm 3.5 with various sample sizes contributing to the empirical distribution.

The random walk Markov chain, Algorithm 3.9, is now discussed. The uniform candidate $q_1$ as in (3.16) was utilized with $d = 1, \ldots, 6$. Additionally, the geometric-based candidate $q_2$ as in (3.17) was utilized with $p \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. The total variation distance results for both are shown in Figure 3.5. In each case, the TV distance is quickly declining. The iterations required for each TV distance to be below the 0.005 threshold are shown in Tables 3.1 and 3.2 respectively. The worst case is Candidate 1 with $d = 1$, corresponding to a candidate proposing either one above or below the current state randomly, which requires 66 iterations for the TV distance from $p_n$ to be below 0.005. The fastest convergence with Candidate 1 is with $d = 6$ which only requires 10 iterations. The fastest using Candidate 2 with $p = 0.3$ only requires 14 iterations. As each example will differ with respect to the
optimal tuning parameters, it is useful to note that the chains with the smallest TV distance had Markov chain acceptance rates between 0.45 and 0.55, which is close to the optimal result of 44.1% for some univariate random walk Metropolis chains of one-dimension established by Gelman et al. (1996).

Finally, the pseudo-marginal Metropolis-Hastings Markov chain, Algorithm 3.10, is considered. The candidate function $q(n^*)$ was taken to be independent of the current state and is a mixture between an empirical estimate (see Algorithm 3.5) based on 20 random draws of $\tau$ (90% of the mixture) and a somewhat heavy-tailed geometric distribution with success probability 0.01 (10% of the mixture). This is meant to be a rough approximation of the actual distribution with heavier tails. A comparison of the distributions is provided in Figure 3.6.
Figure 3.5: Total variation distance for the random walk Markov chain stationary for $p_n$ applied to the known distribution of $p_n$ example.

Table 3.1: Iterations Required for TV Distance < 0.005 using a Random Walk Metropolis Markov Chain with Candidate 1

<table>
<thead>
<tr>
<th>$d$</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3.2: Iterations Required for TV Distance < 0.005 using a Random Walk Metropolis Markov Chain with Candidate 2

<table>
<thead>
<tr>
<th>$p$</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>14</td>
</tr>
<tr>
<td>0.4</td>
<td>17</td>
</tr>
<tr>
<td>0.5</td>
<td>21</td>
</tr>
<tr>
<td>0.6</td>
<td>26</td>
</tr>
<tr>
<td>0.7</td>
<td>33</td>
</tr>
<tr>
<td>0.8</td>
<td>41</td>
</tr>
</tbody>
</table>
The most engaging question regarding total variation distance is how the accuracy of the estimate for $\bar{p}_n$ changes the total variation distance. As shown in Figure 3.7, even with only five draws of $\tau$ used to estimate $\bar{p}_n$, the total variation distance declines steadily. The estimate using 25 draws has total variation distance decline nearly as quickly as the exact algorithm. There is very little difference between the algorithm estimating with 100 draws compared to the one which computes it exactly. Table 3.3 provides the number of iterations for each approximation to reach a TV distance from $p_n$ below 0.005. Note that $\infty$ in the table implies using the exact probability rather than an estimate (while usually unavailable, the probability is available for this known-truth example). The more draws used in the approximation, the better, though with clearly diminishing returns. This provides useful guidance for the trade-off between increasing the accuracy of the estimates and increasing the computational cost.

<table>
<thead>
<tr>
<th>Draws for Approximation</th>
<th>5</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iterations Required</td>
<td>87</td>
<td>47</td>
<td>25</td>
<td>20</td>
<td>19</td>
<td>20</td>
</tr>
</tbody>
</table>

### 3.4.2 Bayesian One-Way Random Effects Model

These approximation methods are now applied to a more realistic example: the posterior distribution of the Bayesian random effects model discussed in Section 2.5.1 applied to the styrene data. The prior in Setting 1, which was somewhat data-based as in Jones and Hobert (2001a), is considered. As was shown in Example 2.5.1, the
Figure 3.6: Comparison of the candidate and actual mass functions for the independence Metropolis-Hastings pseudo-marginal algorithm.

Figure 3.7: Total variation distance of the pseudo-marginal independence algorithm with varying number of draws utilized to estimate $P(\tau \geq n)$. 
two block Gibbs sampler is geometrically ergodic. Exact samples can therefore be obtained directly using the method of Flegal and Herbei (2012), which is provided in Algorithm 3.2. As this was done with the same dataset for a slightly different prior distribution in Flegal and Herbei (2012), the exact algorithm is not illustrated here.

While the specific minorization condition (2.7) was shown in Example 2.5.1, the general minorization condition (3.1) may be used for this illustration, which will result in faster regenerations. Similar to the strategy of Jones and Hobert (2001a) and Mykland et al. (1995), choose the set \( D = (d_1, d_2) \times (d_3, d_4) = (0.11, 0.14) \times (0.43, 0.63) \) and distinguished points \( \tilde{\mu} = 0.0017 \) and

\[
\tilde{\theta} = (-0.62 -0.09 0.10 0.11 -0.13 0.16 0.05 0.03 0.18 0.08 0.33 -0.19 0.00)^T,
\]

based on an initial posterior sample. Define

\[
V_\theta = \sum_{i=1}^{q} (\theta'_i - \mu')^2,
\]

\[
\tilde{V}_\theta = \sum_{i=1}^{q} (\tilde{\theta}_i - \tilde{\mu})^2,
\]

\[
V_e = \sum_{i=1}^{q} n_i (\bar{y}_i - \theta'_i)^2,
\]

and

\[
\tilde{V}_e = \sum_{i=1}^{q} n_i (\bar{y}_i - \tilde{\theta}_i)^2.
\]

For notational convenience, denote \( \phi = \{ \theta, \mu, \sigma^2_\theta, \sigma^2_e \} \). Then, the general minorization condition (3.1) is satisfied because

\[
k(\phi'; \phi) = f_{IG} \left( \sigma^2_e; a_e + N/2, V_e/2 + SSE/2 + b_e \right)
\]

\[
\times f_{IG} \left( \sigma^2_\theta; a_\theta + q/2, V_\theta/2 + b_\theta \right)
\]

\[
\times f_N (\theta; \Sigma_\theta \Sigma e^{-2} \bar{Y}, \Sigma_\theta) f_N \left( \mu; \frac{\bar{y}q}{\sigma^2_\theta} \left( \frac{1}{\sigma^2_\theta} + \frac{q}{\sigma^2_e} \right)^{-1} \right)
\]

\[
\times \left( \frac{1}{\sigma^2_\theta} + \frac{q}{\sigma^2_e} \right)^{-1}
\]

102
\[
\begin{align*}
&= \frac{f_{IG}(\sigma_e^2; a_e + N/2, V_e/2 + \text{SSE}/2 + b_e)}{f_{IG}(\sigma_e^2; a_e + N/2, \tilde{V}_e/2 + \text{SSE}/2 + b_e)} \\
&\quad \times f_{IG}(\sigma_e^2; a_e + N/2, \tilde{V}_e/2 + \text{SSE}/2 + b_e) \\
&\quad \times \frac{f_{IG}(\sigma_\theta^2; a_\theta + q/2, V_\theta/2 + b_\theta)}{f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta)} f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta) \\
&\quad \times f_N(\theta; \Sigma_\theta \Sigma_e^{-2} \bar{Y}, \Sigma_\theta) f_N \left( \mu; \frac{\bar{y}_q}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} : \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \right) \\
&\geq 1_{D}(\sigma_\theta^2, \sigma_e^2) \inf_{\sigma_e^2 \in (d_3.d_4)} \left\{ \frac{f_{IG}(\sigma_e^2; a_e + N/2, V_e/2 + \text{SSE}/2 + b_e)}{f_{IG}(\sigma_e^2; a_e + N/2, \tilde{V}_e/2 + \text{SSE}/2 + b_e)} \right\} \\
&\quad \times \inf_{\sigma_\theta^2 \in (d_1,d_2)} \left\{ \frac{f_{IG}(\sigma_\theta^2; a_\theta + q/2, V_\theta/2 + b_\theta)}{f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta)} \right\} \\
&\quad \times f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta) \\
&\quad \times f_N(\theta; \Sigma_\theta \Sigma_e^{-2} \bar{Y}, \Sigma_\theta) f_N \left( \mu; \frac{\bar{y}_q}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} : \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \right) \\
&\quad = s'(\phi') q'(\phi)
\end{align*}
\]

where
\[
\begin{align*}
s'(\phi') &= \inf_{\sigma_e^2 \in (d_3,d_4)} \left\{ \frac{f_{IG}(\sigma_e^2; a_e + N/2, V_e/2 + \text{SSE}/2 + b_e)}{f_{IG}(\sigma_e^2; a_e + N/2, \tilde{V}_e/2 + \text{SSE}/2 + b_e)} \right\} \\
&\quad \times \inf_{\sigma_\theta^2 \in (d_1,d_2)} \left\{ \frac{f_{IG}(\sigma_\theta^2; a_\theta + q/2, V_\theta/2 + b_\theta)}{f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta)} \right\}
\end{align*}
\]

and
\[
q'(\phi) = 1_D(\sigma_\theta^2, \sigma_e^2) f_{IG}(\sigma_e^2; a_e + N/2, \tilde{V}_e/2 + \text{SSE}/2 + b_e) \\
\quad \times f_{IG}(\sigma_\theta^2; a_\theta + q/2, \tilde{V}_\theta/2 + b_\theta) \\
\quad \times f_N(\theta; \Sigma_\theta \Sigma_e^{-2} \bar{Y}, \Sigma_\theta) f_N \left( \mu; \frac{\bar{y}_q}{\sigma_\theta^2} \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} : \left( \frac{1}{\sigma_\mu^2} + \frac{q}{\sigma_\theta^2} \right)^{-1} \right),
\]
where \( f_{IG}(x; a, b) \) is the inverse gamma density evaluated at \( x \) with parameters \( a \) and \( b \) and \( f_{N}(x; \mu, \Sigma) \) is the multivariate normal density (perhaps with \( d = 1 \)) evaluated at \( x \) with mean \( \mu \) and covariance matrix \( \Sigma \). Recall that the unnormalized density \( q'(\phi) \) is sufficient because the normalizing constant can be multiplied by \( s'(\phi') \) to obtain \( s(\phi') \) and \( q(\phi) \).

With the minorization condition, it remains to find the probability of regeneration at each step, as in (3.2). First, obtain a closed form expression for \( s'(\phi) \). Note that

\[
\inf_{\sigma_{\theta}^2 \in (d_1,d_2)} \left\{ \frac{f_{IG}(\sigma_{\theta}^2; a_{\theta} + q/2, V_{\theta}/2 + b_{\theta})}{f_{IG}(\sigma_{\theta}^2; a_{\theta} + q/2, \bar{V}_{\theta}/2 + b_{\theta})} \right\}
\]

\[
= \inf_{\sigma_{\theta}^2 \in (d_1,d_2)} \left\{ \frac{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta} (\sigma_{\theta}^2)^{-q/2-a_\theta-1} \exp(-\sigma_{\theta}^{-2}(V_{\theta}/2 + b_{\theta}))}{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta} (\sigma_{\theta}^2)^{-q/2-a_\theta-1} \exp(-\sigma_{\theta}^{-2}(\bar{V}_{\theta}/2 + b_{\theta}))} \right\}
\]

\[
= \inf_{\sigma_{\theta}^2 \in (d_1,d_2)} \left\{ \frac{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta}}{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta}} \exp \left( -\frac{1}{2\sigma_{\theta}^2}(V_{\theta} - \bar{V}_{\theta}) \right) \right\}
\]

\[
= \frac{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta}}{(V_{\theta}/2 + b_\theta)^{q/2+a_\theta}} \exp \left( -\frac{1}{2g_{\theta}}(V_{\theta} - \bar{V}_{\theta}) \right),
\]

where

\[
g_{\theta} = \begin{cases} 
    d_1 & \text{if } V_{\theta} > \bar{V}_{\theta} \\
    d_2 & \text{if } V_{\theta} \leq \bar{V}_{\theta},
\end{cases}
\]

which is easily verified by considering the sign of \( V_{\theta} - \bar{V}_{\theta} \) and minimizing with respect to \( \sigma_{\theta}^2 \) over the set \( (d_1,d_2) \). By the same logic,

\[
\inf_{\sigma_{\epsilon}^2 \in (d_3,d_4)} \left\{ \frac{f_{IG}(\sigma_{\epsilon}^2; a_{\epsilon} + N/2, V_{\epsilon}/2 + SSE/2 + b_{\epsilon})}{f_{IG}(\sigma_{\epsilon}^2; a_{\epsilon} + N/2, \bar{V}_{\epsilon}/2 + SSE/2 + b_{\epsilon})} \right\}
\]

\[
= \frac{(V_{\epsilon}/2 + SSE/2 + b_\epsilon)^{N/2+a_\epsilon}}{(V_{\epsilon}/2 + SSE/2 + b_\epsilon)^{N/2+a_\epsilon}} \exp \left( -\frac{1}{2g_{\epsilon}}(V_{\epsilon} - \bar{V}_{\epsilon}) \right),
\]

where

\[
g_{\epsilon} = \begin{cases} 
    d_3 & \text{if } V_{\epsilon} > \bar{V}_{\epsilon} \\
    d_4 & \text{if } V_{\epsilon} \leq \bar{V}_{\epsilon}.
\end{cases}
\]
Then, note that the probability of regeneration is

\[
P(\delta_n = 1|\phi', \phi) = \frac{s(\phi')q(\phi)}{k(\phi', \phi)} = \frac{(V_\theta/2 + b_\theta)^{q/2+a_\theta} (V_e/2 + SSE/2 + b_e)^{N/2+a_e}}{(\tilde{V}_\theta/2 + b_\theta)^{q/2+a_\theta} (\tilde{V}_e/2 + SSE/2 + b_e)^{N/2+a_e}} \\
\times \exp \left( -\frac{1}{2g_\theta} (V_\theta - \tilde{V}_\theta) - \frac{1}{2g_e} (V_e - \tilde{V}_e) \right) \\
\times 1_D(\sigma_\theta^2, \sigma_e^2) (\tilde{V}_\theta/2 + b_\theta)^{q/2+a_\theta} (\tilde{V}_e/2 + SSE/2 + b_e)^{N/2+a_e} \\
\times (\sigma_\theta^2)^{-q/2-a_\theta-1} (\sigma_e^2)^{-N/2-a_e-1} \\
\times \exp \left( -\frac{1}{\sigma_\theta^2} (V_\theta/2 + b_\theta) - \frac{1}{\sigma_e^2} (V_e/2 + SSE/2 + b_e) \right) \\
\times \left[ (\sigma_\theta^2)^{-q/2-a_\theta-1} (\sigma_e^2)^{-N/2-a_e-1} \right]^{-1} \\
\times \left[ \exp \left( -\frac{1}{\sigma_\theta^2} (V_\theta/2 + b_\theta) - \frac{1}{\sigma_e^2} (V_e/2 + SSE/2 + b_e) \right) \right]^{-1},
\]

where the normal densities from \( q \) and \( k \) cancel each other out

\[
= 1_D(\sigma_\theta^2, \sigma_e^2) \exp \left( -\frac{1}{2g_\theta} (V_\theta - \tilde{V}_\theta) - \frac{1}{2g_e} (V_e - \tilde{V}_e) \right) \\
\times \exp \left( -\frac{1}{2\sigma_\theta^2} (\tilde{V}_\theta/2 + b_\theta) - \frac{1}{\sigma_e^2} (\tilde{V}_e/2 + SSE/2 + b_e) \right) \\
\times \left[ \exp \left( -\frac{1}{\sigma_\theta^2} (V_\theta/2 + b_\theta) - \frac{1}{\sigma_e^2} (V_e/2 + SSE/2 + b_e) \right) \right]^{-1},
\]

by canceling out the constants

\[
= 1_D(\sigma_\theta^2, \sigma_e^2) \exp \left( -\frac{1}{2g_\theta} (V_\theta - \tilde{V}_\theta) - \frac{1}{2g_e} (V_e - \tilde{V}_e) \right) \\
\times \exp \left( -\frac{1}{2\sigma_\theta^2} (\tilde{V}_\theta - V_\theta) - \frac{1}{2\sigma_e^2} (\tilde{V}_e - V_e) \right),
\]

\[
= 1_D(\sigma_\theta^2, \sigma_e^2) \exp \left( \frac{1}{2} \left( \frac{1}{g_\theta} - \frac{1}{\sigma_\theta^2} \right) (V_\theta - \tilde{V}_\theta) + \frac{1}{2} \left( \frac{1}{g_e} - \frac{1}{\sigma_e^2} \right) (V_e - \tilde{V}_e) \right).
\]

Equipped with the probability of regeneration, regeneration times \( \tau \) may be found using the Markov chain, and the approximation methods may be utilized.
To compare the computational efficiency of the various approximation methods, one million i.i.d. draws of \( \tau \) are produced and the number of resulting approximate draws from \( p_n \) are computed. Methods which produce more draws with the same number of \( \tau \) draws are more efficient. The results are provided in Table 3.4. Algorithm 3.3 reports a range of draws obtained with varying \( M \) values; 193,538 draws corresponds to the first \( M \) where the TV distance is less than 0.005; 62,143 corresponds to \( M = \max_i(\tau_i) \). Naturally, more draws would have been attained with a smaller \( M \). Algorithm 3.5 reports \( \infty \) as the number of approximate draws because once the empirical distribution approximating \( p_n \) has been obtained, draws may be freely taken from the approximation requiring no additional \( \tau \) draws. This is desirable if a large number of draws are sought, though without additional \( \tau \) draws, the approximation will not improve.

Table 3.4: Number of Approximate Draws Produced using 1 Million Draws of \( \tau \) Under the Different Approximate Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Algorithm Detail</th>
<th># Approx. Draws</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3</td>
<td>Truncation</td>
<td>62,143–193,538</td>
</tr>
<tr>
<td>3.5</td>
<td>Empirical</td>
<td>( \infty )</td>
</tr>
<tr>
<td>3.6</td>
<td>Reverse Rejection</td>
<td>108,131</td>
</tr>
<tr>
<td>3.9</td>
<td>Random Walk</td>
<td>224,477</td>
</tr>
<tr>
<td>3.10</td>
<td>Pseudo-Marginal</td>
<td>40,000</td>
</tr>
</tbody>
</table>

Comparing the three Markov chains, the random walk of Algorithm 3.9 was most efficient producing 224,477 draws. A larger number of draws are obtained because draws from \( \tau \) are only required half of the time because all proposals smaller than the current value are automatically accepted. The reverse rejection sampler is the next
most efficient with 108,131 draws resulting. Finally, the least efficient algorithm is the pseudo-marginal Metropolis-Hastings chain which requires an unbiased estimate of \(P(\tau \geq N^*)\) for each step of the Markov chain where \(N^*\) is the proposed value. In this case, the unbiased estimate was based on 25 draws, which was shown to produce good results in Section 3.4.1, resulting in a chain with 40,000 draws. If fewer draws were used for the estimation of \(P(\tau \geq N^*)\), then more approximate draws would result, but the TV distance would be higher. On the other hand, if more draws were used to approximate \(P(\tau \geq N^*)\), the TV distance would be lower, but so would the number of draws obtained. Section 3.4.1 illustrated the decreasing marginal benefit of additional draws in the approximation.

Beyond computational efficiency, the results from each of the algorithms are now discussed. Figure 3.8 shows the TV distance and acceptance rates for various truncation values \(M\) using Algorithm 3.3. As in Section 3.4.1, the total variation distance converges to zero much more quickly than the acceptance rate. The TV distance goes below the 0.005 threshold at \(M = 39\) with a large acceptance rate of 19%. Larger \(M\) results in additional accuracy without much loss in computational efficiency.

The empirical distribution used for Algorithm 3.5 is shown in Figure 3.9. The approximation is based on all 1,000,000 draws, which appears to result in a very smooth estimate. The estimate is 0 beyond the maximum \(\tau\) value, 122 in this case. As Algorithm 3.3 perfectly samples from the truncated distribution, it will always result in a smaller TV distance for \(M \geq \max_i(\tau_i)\), however the acceptance rate of Algorithm 3.3 is only 6.21\% for that value of \(M\), whereas the approximation of Algorithm 3.5 directly results in as many draws as desired.
Figure 3.8: TV Distance and acceptance rates for various $M$ values using Algorithm 3.3 for the Bayesian random effects model.

The Markov chain approximations are now compared. The empirical autocorrelation function for each of these chains is found in Figure 3.10. The reverse rejection sampler has the most autocorrelation, followed by the random walk algorithm, and then the pseudo-marginal algorithm. The autocorrelation will be attenuated for each of these methods, however, when looking at the marginal $\{X_n\}$ chain that results when sampling from $Q_N$ after each iteration of the chain. This is because, conditional on the chain of the mixture variables $\{P_n\}$, the draws from $Q_N$ are independent.

Finally, draws are obtained from the $\{X_n\}$ marginal chain using Algorithm 3.4 which has finite expected running time as opposed to the other approximate and exact methods of sampling discussed in this chapter. The truncation is set at $M = 100$ which results in an estimated TV distance bound of 0.0000038 from the $p_n$ chain. This
also implies a TV distance bound of 0.0000077 between the marginal \(\{X_n\}\) chain and \(\pi\). The acceptance rate of the algorithm is 7.58\%, making it quite computationally efficient given the quality of the draws approximating \(\pi\).

To illustrate the accuracy of the approximation visually, Figure 3.11 shows estimates of the posterior densities for \(\sigma^2_{\theta}\), \(\sigma^2_\epsilon\), \(\mu\), and \(\theta_1\) obtained from two sources: (1) 100,000 draws from Algorithm 3.4 and (2) 5 million draws from the original Markov chain \(\{X_n\}\). As shown, for each parameter, the density estimates are incredibly close, indicating that the approximate methods are indeed closely approximating \(\pi\) as expected.

Given all these results, a discussion is now provided about optimal strategies in practice.
Figure 3.10: Autocorrelation functions for the Markov chains approximating $p_n$ for the Bayesian random effects model.

3.5 Discussion

The algorithms presented in this chapter are effective at approximating draws from $p_n$ and thus from $\pi$. Recommendations for best practices are now discussed.

There are two primary purposes for considering these methods. The first is to answer Question 1 by simulating a nearly perfect draw from which the Markov chain $\{X_n\}$ may be started so that the total variation distance is close to zero. For this purpose, perhaps the best algorithm presented here is Algorithm 3.4. The finite expected running time alone is perhaps enough to champion this method over the others. Furthermore, the TV distance decreases quite quickly and with relatively modest increase in computational burden. It is easy to obtain more accurate draws by simply increasing $M$. Furthermore, this is the only method where the total variation distance from $p_n$ may be directly estimated without needing to know the true distribution of $p_n$. These advantages make Algorithm 3.4 an excellent choice for obtaining a single good draw from which to begin a larger Markov chain.
The second primary purpose is to create independent exact draws from \( \pi \). In this case, typical strong laws of large numbers for Monte Carlo averages may be used. In the case where exact draws are not possible or computationally feasible, results may not be guaranteed about convergence of expectations under the true distribution \( \pi \) when approximate draws are used—the strong law of large numbers holds only for expectations computed with respect to the approximate distribution \( \hat{\pi} \).

It is therefore useful to use a Markov chain stationary for \( \pi \) for which ergodic averages follow a strong law of large numbers, and hence converge to the correct expectation under \( \pi \). Thus, a best practice recommendation could be to start one of the Markov chains stationary for \( p_n \), preferably the random walk or pseudo-marginal independence Metropolis-Hastings sampler, with a good draw, perhaps from Algorithm 3.3 with large \( M \), and then continue running the Markov chain stationary for \( p_n \). Ergodic averages will therefore converge to the desired expectation under \( \pi \) rather than under the approximation. This convergence will be faster than for the original Markov chain which may be subject to large autocorrelations; any of the Markov chains from this chapter will have a marginal chain \( \{X_n\} \) with autocorrelation near zero because dependence is only through the mixture weights.

If a very large number of approximate draws are needed but not intended to be used for a central limit theorem, perhaps to begin many parallel chains for example, either Algorithm 3.4 or Algorithm 3.5 could be very useful. Particularly if a large number of independent draws from \( p_n \) are required, Algorithm 3.5 is useful; as many draws as desired may be arbitrarily obtained once enough simulation effort has been spent in constructing the empirical estimate of \( p_n \). The drawback of Algorithm 3.5 is
the infinite expected running time of obtaining the draw from $Q_n$ given draws from $p_n$; Algorithm 3.4 would then again be the superior algorithm.

All these methods allow the total variation distance of Markov chains to start at very small values, answering Question 1, eliminating the need for burn-in, and somewhat reducing the need for extraordinarily long chains for reliable simulation results. Their use can be employed when a general minorization condition is satisfied, even when the original Markov chain is not geometrically ergodic. The approximation methods of this chapter can be one of the best methods for obtaining an approximate distribution for $\pi$ when the distribution is complex, for example, with a large number of dimensions. The limitation, however, is that if an incredibly poor minorization condition is obtained, the use of these approximations may not be practical.
Figure 3.11: Comparison of posterior density estimates using 100,000 approximate draws from Algorithm 3.4 (dashed) with the original Markov chain run for 5 million iterations (solid).
4.1 Introduction

In many surveys where respondents provide ordinal ratings for a number of questions, respondents can fundamentally differ in how they use the provided rating scale. For example, there are yea-sayers who predominately utilize the upper part of the scale and nay-sayers who conversely utilize the lower portion of the scale. Some individuals, on the other hand, use the entirety of the scale or perhaps just the middle of the scale.

4.1.1 Heterogeneity of Scale Use

This heterogeneous use of rating scales is a well-documented phenomenon. Rorer (1965) presented a review of the early literature and credits Lentz (1938) and Lorge (1937) with the original papers which documented the existence of specific response styles. Several other early papers further developed the field. Cronbach (1941) concluded that marking “true” rather than “false” when uncertain on an exam may indeed be a personality trait. This has important consequences when the percentage of true questions is largely different than the percentage of false questions on an exam.
Cronbach (1942) found evidence that the response styles were consistent for individual students across multiple exams, and provided several recommendations for better exam administration. In Cronbach (1946), additional response styles of students on varying types of examinations were discussed. See also Cronbach (1950) for a review of several studies where this phenomenon was observed.

Couch and Keniston (1960) specifically discussed the phenomenon of yea-sayers and nay-sayers. Their study discussed this response style as being an integral part of an individual’s personality. Pettigrew (1958) discusses the phenomenon of some respondents using a small portion of the scale compared to a larger portion of the scale. Hui and Triandis (1985) discuss how, for lengthy surveys, individual response styles can change partway through responding.

Greenleaf (1992) discussed whether these response styles are reflections of attitude alone, or if bias is being conveyed. For example, a yea-sayer is likely to have a positive outlook, but are the responses conveying more positivity than is inherently present in the individual? He also discusses how this bias can impede analyses, with a specific example of its effects on segmentation through clustering. Baumgartner and Steenkamp (2001) similarly cautioned about the spurious results that often result when adjustments are not made for response styles when using inferential methods that depend on correlation. Clarke (2000) discuss how response styles particularly differ across cultures and provide recommendations for international research.

4.1.2 Scale-Usage Data

Before discussing the variety of methods used to account for scale-usage heterogeneity, a discussion of the underlying data is helpful. Consider a survey with $M$
questions, each of which is on an ordinal scale with \( K \) levels. Let there be \( N \) individuals. If individual \( i \) selects the \( k^{th} \) level on question \( j \), this is denoted by \( X_{ij} = k \). The matrix \( \mathbf{X} \) is \( N \times M \) and collects the responses for each of the questions \( j = 1, \ldots, M \) from each individual \( i = 1, \ldots, N \).

As an example, consider the first five observations of the customer satisfaction data discussed and analyzed by Rossi et al. (2001), shown in Table 4.1. For this

<table>
<thead>
<tr>
<th>ID</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>2</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

survey, there are ten questions, each on a ten-point scale. Each row represents a different respondent and each column is a different question. The values inside the matrix vary from 1 to 10 as a ten-point rating scale was used.

Broadly, the source of scale-usage data can be any survey where ratings on an ordinal scale are taken. The applications are extensive. Rossi et al. (2001) analyzed customer satisfaction data on several points from price to effectiveness. Büschken et al. (2013) considered customer satisfaction for hospitals, an undergraduate business program in Germany, smart phones, and Florida vacations. Muthukumarana and Swartz (2014) analyzed student ratings at Simon Fraser University. Beyond customer satisfaction, purchase intentions are often analyzed, e.g. Fong et al. (2010) analyzed
customer purchase intentions for a luxury SUV. Though the applications are many, the basic premise and foundational modeling principles apply.

4.2 Models and Inference for Scale-Usage Heterogeneity

4.2.1 Motivation

To understand the fundamental modeling principle behind scale-usage models, consider the following scenario which is presented graphically in Figure 4.1. A respondent reads a survey question and has an initial reaction. Call this initial reaction $d_0$. As time passes, the respondent drifts up or down in their consideration of responses. Finally, at the decision time $T$, $d_T$ is compared to a set of cutpoints, and the discrete response corresponding to $d_T$ is recorded on the survey. Figure 4.1 shows several possible decision paths in gray with a specific path highlighted in black. For this path, the end result is to respond with a four. Other paths would have resulted in a three; some with a two, and so on. Therefore, more than one discrete response is possible, each with some probability. The right side of Figure 4.1 displays the distribution of the decision paths at time $T$, along with the corresponding distribution of discrete responses. While responding with a three is most likely here, it is also quite likely the response is not a three.

This scenario often happens incredibly quickly with relatively small $T$. The process likely also does not happen consciously, though many survey responders would agree that it requires a bit of thought to nail down a specific choice between two or more options. The statistical framework corresponding to this scenario, which is used for analyzing this type of data, is probit regression for ordered responses.
4.2.2 Probit Regression

Probit regression dates back to Bliss (1935) with toxicology studies. As in McCullagh et al. (1989), let $\pi_x$ denote the probability of survival of the subject with dose $x$. Then,

$$\pi_x = \Phi(\alpha + \beta x),$$

where $\Phi$ is the cumulative distribution function of the normal distribution and $\alpha$ and $\beta$ are regression coefficients. This choice is favorable because it requires $\pi_x$ to lie between 0 and 1. Bliss (1935) further justifies the choice of the normal distribution as an “ideal representation of the variation in susceptibility.” In the context of scale-usage models, the choice of the normal cumulative distribution function as a link implies that the distribution of the decision paths at time $T$ is normally distributed, as in Figure 4.1.
This simple probit regression model with two categories (survival and death in this case), was expanded to include multiple ordered categories. Aitchison and Silvey (1957) developed such a model for insects that progress through six ordered life stages, and Gurland et al. (1960) similarly considered three stages of a housefly after exposure to dosages of a chemical. Extending the simpler model to $K$ ordinal categories, the probability of being in stage $k$ with dosage $x$ is

$$
\pi_k(x) = \begin{cases} 
\Phi(\alpha_1 + \beta x) & k = 1 \\
\Phi(\alpha_k + \beta x) - \Phi(\alpha_{k-1} + \beta x) & 2 \leq k < K \\
1 - \Phi(\alpha_{K-1} + \beta x) & k = K
\end{cases}
$$

where $\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_K - 1$. This model could equivalently be written in terms of cutpoints, as will be done with the remaining models of this chapter:

$$
\pi_k(x) = \Phi(c_k + \alpha + \beta x) - \Phi(c_{k-1} + \alpha + \beta x) \quad k = 1, \ldots, K
$$

where $-\infty = c_0 \leq c_1 \leq c_2 \leq \ldots \leq c_K = \infty$. The equivalence is easily seen with $\alpha = \alpha_1$, $c_1 = 0$, and $c_k = \alpha_k - \alpha_1$ for $k \geq 2$. This approach corresponds directly with the motivating scenario depicted in Figure 4.1 where the distribution of decision paths at time $T$ is normal with mean $-\alpha - \beta x$ and standard deviation equal to one. For additional details on probit and other models for ordinal data using the classical approach, including details about estimation, see McCullagh (1980) and Agresti (2002).

### 4.2.3 Bayesian Probit Regression

With the advent of the Gibbs sampler following Geman and Geman (1984) and Gelfand and Smith (1990), as well as the improvement of other MCMC methods such as Metropolis-Hastings, Bayesian approaches became more feasible and popular, largely due to their flexibility and inferential advantages. Albert and Chib (1993)
introduced a Bayesian model for ordinal probit data. Assume $K$ levels so that $X_i \in \{1, \ldots, K\}$ for each individual $i = 1, \ldots, N$. Also, assume covariates are observed for each individual, denoted $x_i$ (note the distinction between $X_i$, the ordinal response, and $x_i$, the covariates of interest). The Bayesian ordinal probit model is then

$$Y_i \sim \text{Normal}(x_i^T \beta, 1), \quad i = 1, \ldots, N,$$

$$X_i = \{k : c_{k-1} < Y_i \leq c_k\}, \quad i = 1, \ldots, N,$$

where $-\infty = c_0 < c_1 \leq \cdots \leq c_{K-1} < c_K = \infty$, and for identifiability set $c_1 = 0$ (any cutpoint $c_1, \ldots, c_{K-1}$ set to a finite value would suffice; the selection is made arbitrarily). A prior distribution is chosen for $\beta$ and the remaining cutpoint parameters. This is also referred to as a data augmentation model as the $Y_i$ are not observed.

Gibbs samplers are incredibly well-suited for fitting such models.

To create the Gibbs sampler, first identify the complete conditional distribution for each parameter. Assume a diffuse prior is selected so $\pi(\beta, c) \propto 1$. Then, as in Albert and Chib (1993),

$$\beta | Y, x \sim \text{MVN}((x^T x)^{-1} x^T Y, (x^T x)^{-1}),$$

where $x = (x_1, x_2, \ldots, x_N)^T$ and $Y = (Y_1, \ldots, Y_N)^T$. The $Y_i$ latent variables are conditionally independent given $X_i$ and $c$ with distribution

$$Y_i | \beta, c, X, x \sim \text{Truncated Normal}(x_i^T \beta, 1, c_{X_i-1}, c_{X_i}), \quad i = 1, \ldots, N.$$ 

Finally, the update for each cutpoint sequentially is

$$c_k | c_{k-1}, c_{k+1}, Y \sim \text{Uniform}(\max_i\{Y_i : X_i = k\}, \min_i\{Y_i : X_i = k+1\}), \quad k = 2, \ldots, K-1.$$
The Gibbs sampler performs each of these updates sequentially. The algorithm can converge slowly, however, due to the small changes in the cutpoint parameters which often occur.

The extension of this univariate ordinal probit model to a multivariate ordinal probit model is simple. Let $M$ be the dimension of the response vector for each individual $i = 1, \ldots, N$. Throughout, let $j$ index the question, i.e. $j = 1, \ldots, M$. Let

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{iM})^T, \quad i = 1, \ldots, N$$

be the individual ordinal response vector and

$$X = (X_1^T, \ldots, X_N^T)^T$$

collect the ordinal responses of each individual. While it will often be the case that all dimensions are on the same $K$-level ordinal scale, it is possible to construct a model with a distinct $K_j$ for each dimension. Similarly, the same cutpoints $c$ may be used across all dimensions, but they could also be uniquely defined for each dimension as $c_j$. Let $x_{ij}$ denote the vector of $L_j$ covariates for the $i^{th}$ individual and $j^{th}$ dimension of response, and let $\beta_j$ be a vector of coefficients of length $L_j$ for the $j^{th}$ response. The model is then,

$$Y_i \sim MVN(\mu_i, \Sigma), \quad i = 1, \ldots, N$$

$$\mu_i = (x_{i,1}^T \beta_1, x_{i,2}^T \beta_2, \ldots, x_{i,M}^T \beta_M)^T, \quad i = 1, \ldots, N$$

$$X_{ij} = \{k : c_{j,k-1} < Y_{ij} \leq c_{j,k}\} \quad i = 1, \ldots, N; \quad j = 1, \ldots, M, \quad (4.1)$$

where $-\infty = c_{j,0} \leq c_{j,1} \leq \ldots \leq c_{j,K_j} = \infty$ and, for identifiability, set $c_{j,1} = -C$ and $c_{j,K_j-1} = C$ for some constant $C$. As discussed by Jeliazkov et al. (2008), two cutpoints must be restricted due to the unrestricted nature of $\Sigma$. For additional
approaches, see Jeliazkov et al. (2008), Chib and Greenberg (1998), and McCulloch et al. (2000). Prior distributions are then selected for each of the $\beta_j$ coefficients, $j = 1, \ldots, M$, the remaining free cutpoints $c_{j,2}, \ldots, c_{j,K_j-1}$ with $j = 1, \ldots, M$, and the covariance matrix $\Sigma$. This basic model can flexibly be simplified, such as to have constant $\beta$ across all dimensions, or enriched, for example, by adding individual random effects as in Bradlow and Zaslavsky (1999). This model can also be modified to directly account for scale-usage heterogeneity when analyzing survey data.

4.2.4 Bayesian Scale-Usage Models

Rossi et al. (2001) introduced one of the first such models uniting Bayesian ordinal probit methods with scale-usage adjustments. The model in (4.1) is modified by allowing the mean for each observation to be constant except for an individual shift adjustment $\tau_i$. The covariance matrix is also constant except for an individual scale adjustment, $\sigma_i^2$. The cutpoints are also assumed to be constant across dimensions and to take a particular form. Define $\mathbf{1}$ as the $M \times 1$ vector whose entries are all equal to one. Then,

$$
\begin{align*}
Y_i & \sim \text{MVN}_M(\mu + \tau_i \mathbf{1}, \sigma_i^2 \Sigma), \quad i = 1, \ldots, N, \\
X_{ij} & = \{k : c_{k-1} < Y_{ij} \leq c_k\}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M, \\
(\tau_i, \ln \sigma_i)^T & \sim \text{MVN}_2(\phi, \Lambda), \quad i = 1, \ldots, N, \\
c_k & = a + bk + ek^2, \quad k = 1, \ldots, K - 1, \\
\pi(\mu) & \propto 1 \\
\pi(e) & \sim \text{Uniform}(-.2,.2) \\
\Sigma^{-1} & \sim \text{Wishart}_M(\nu_\Sigma, V_\Sigma)
\end{align*}
$$
\[ \Lambda^{-1} \sim \text{Wishart}_2(\nu_\Lambda, V_\Lambda), \]  

(4.2)

where \( \phi, a, b, \nu_\Sigma, V_\Sigma, \nu_\Lambda, \) and \( V_\Lambda \) are all selected constants. For guidance on selection, see Rossi et al. (2001).

A variation of this model was discussed by Hans et al. (2012):

\[ Y_i \sim \text{MVN}_M(\mu + \tau_i 1, \sigma_i^2 \Sigma), \quad i = 1, \ldots, N, \]

\[ X_{ij} | c, Y_{ij} \equiv \begin{cases} k & c_{k-1} < Y_{ij} \leq c_k \end{cases}, \quad i = 1, \ldots, N; \quad j = 1, \ldots, M, \]

\[ \sigma_i^2 \sim \text{Inverse Gamma}(a/2, (a - 2)/2), \quad i = 1, \ldots, N, \]

\[ \tau_i \sim \text{Normal}(0, \sigma_i^2), \quad i = 1, \ldots, N, \]

\[ \Sigma \sim \text{Inverse Wishart}_M(\delta, \Sigma_0) \]

\[ \mu \sim \text{MVN}_M(0, V) \]

\[ \pi(c) \propto \prod_{k=2}^{K-1} (c_k - c_{k-1})^{\alpha_k - 1}, \text{ where } c_1 = -C \text{ and } c_{K-1} = C, \]  

(4.3)

where \( a, \delta, \Sigma_0, V, C, \) and \( \alpha_k \) for \( k = 2, \ldots, K - 1 \) are all constants.

The difference between these two models lies particularly in how the cutpoints are handled. In Rossi et al. (2001), the cutpoints are required to lie on a parabola, with a prior specified on the parameter governing the parabola subject to certain identifiability constraints. In Hans et al. (2012), a more flexible model for the cutpoints is provided which is superior when the cutpoints should not fall according to a given form. Prior values can be selected to favor evenly (large \( \alpha \)) or unevenly (small \( \alpha \)) spaced cutpoints. Selecting \( \alpha_k = 1 \) for all \( k \) results in the uniform prior (Hans et al. 2012).

While the prior structure of Rossi et al. (2001) allows for prior covariance between the \( \tau_i \) and \( \sigma_i^2 \) parameters, the choice of prior in Hans et al. (2012) yields a conditional distribution from which draws can be directly obtained. The Rossi et al. (2001)
parameterization, on the other hand, requires a Metropolis update. Note that not specifying prior correlation does not preclude the posterior distribution from being correlated.

Another widely-used scale-usage strategy is to use respondent-specific cutpoints. Johnson (2003) utilized this method, imposing a constraint that the cutpoints must be symmetric. Other examples include Ying et al. (2006) and Büschken et al. (2013). The predominant modeling choice is to put a prior distribution on the differences between cutpoints, rather than specifying a prior which somehow preserves the ordering of the cutpoints. Several of these models also include random effects for other coefficients (such as regression parameters) coming through the mean (Ying et al. 2006), but heterogeneous scale is predominantly modeled through the cutpoints. Often, each individual is allowed to have their own unique set of cutpoints (Büschken et al. 2013). Jeliazkov et al. (2008) provide an excellent discussion about a variety of related models and the necessary identifiability constraints for the cutpoints.

Several other methods for analyzing multivariate ordinal data with scale-usage heterogeneity include the Bayesian nonparametric approach of Kottas et al. (2005) and the semiparametric methods of Muthukumarana and Swartz (2014) and Kim and Ratchford (2013).

While there are advantages and disadvantages to each approach, the ultimate goal is to adjust the analysis to account for scale-usage heterogeneity so that inference is not obscured by the effects.
4.2.5 Inference Strategies

Several inference strategies may be explored after adjusting for scale-usage heterogeneity including cluster analysis, regression, analysis of predictive distributions, and identification of extreme respondents.

Cluster Analysis

One inferential objective is to split subjects into relevant segments or groups based on the survey questions. This is useful for targeting customers with different products or services which they are then more likely to purchase or utilize. One approach for segmenting is cluster analysis. Greenleaf (1992) discussed how, if clustering is attempted without accounting for heterogeneous scale-usage, drastically different results may be obtained. Cluster analysis is performed on the latent variables which have been standardized by the model for scale-usage heterogeneity; for example, using the model proposed by Rossi et al. (2001), the standardization is

\[
\mathbf{z}_i = \frac{\mathbf{Y}_i - \mathbf{\mu} - \tau_1 \mathbf{1}}{\sigma_i^2}, \quad i = 1, \ldots, N. \tag{4.4}
\]

Typical clustering methods, such as K-means, may then be applied to the \( \mathbf{z}_i \).

Regression

Another common inferential question is to predict one of the survey questions from the others. For example, in customer satisfaction surveys, there is often one question about overall satisfaction, followed by several specific questions about different facets of a product or service. It is useful to determine how these individual components contribute to overall satisfaction. Let \( \mathbf{\mu}_z \) and \( \Sigma_z \) denote the mean and covariance
matrix of the standardized parameters $z_i$. Let $z_1$, the first entry of $z_i$, be the desired target (simply reorder if necessary so the first question corresponds to overall satisfaction). Partition the mean and covariance accordingly so $\mu_z = (\mu_{z,1}, \mu_{z,2:M})^T$ and

$$
\Sigma_z = \begin{pmatrix}
\Sigma_{1,1} & \Sigma_{1,2:M} \\
\Sigma_{2:M,1} & \Sigma_{2:M,2:M}
\end{pmatrix}
$$

Then, the conditional mean for $z_1|Z_{2:M}$ is

$$
E(z_1|Z_{2:M} = z) = \mu_{z,1} + \Sigma_{1,2:M} \Sigma_{2:M,2:M}^{-1} (z - \mu_{z,2:M}),
$$

which implies that the regression coefficients for the other questions are

$$
\beta = \Sigma_{2:M,2:M}^{-1} \Sigma_{2:M,1}
$$

with intercept $\mu_{z,1} - \Sigma_{1,2:M} \Sigma_{2:M,2:M}^{-1} \mu_{z,2:M}$. As discussed in Rossi et al. (2001), the standardized latent variables $z_i$, defined in (4.4), have MVN($0, \Sigma$) distribution. Therefore the intercept is 0 and the covariance matrix $\Sigma_z$ is equal to $\Sigma$ of the model. Thus, the regression estimates are directly obtained from the covariance matrix. The estimates vary dramatically from those obtained by regression using the originally unscaled data, as shown in Rossi et al. (2001). Thus, adjusting for scale-usage heterogeneity is critical for regression analysis.

**Predictive Distributions**

Another inferential question is to find a predictive distribution of the survey questions. One important predictive distribution is the distribution of $X$, the ordinal data, with no scale-usage effects (in the case of Rossi et al. (2001), i.e. $\sigma_i^2 = 1$ and $\tau_i = 0$). This distribution paints a good picture for the general feeling among survey respondents after adjusting for their heterogeneous biases. Predictive distributions incorporating different values of $\tau_i$ and $\sigma_i^2$ can also be informative.
Extreme Respondents

The identification of extreme respondents, for example those who are either extremely satisfied or extremely dissatisfied, is also often of interest. This could be based on a number of quantities, but Rossi et al. (2001) suggest that perhaps the best indicator of true satisfaction is a high standardized latent score on the overall satisfaction question, $z_1$. Another argument could be made that high $\tau_i$ values correspond to large levels of satisfaction, but caution needs taken as high values of $\tau_i$ often correspond with yea-sayers whose true satisfaction may not be expressed in the data. Perhaps individuals with higher $\tau_i$ but also higher $\sigma_i^2$ are good candidates for highly satisfied respondents because their use of the scale is broader (high $\sigma_i^2$), yet they are quite satisfied (high $\tau_i$). There is much less information about a respondent with high $\tau_i$ and low $\sigma_i^2$. If on a ten-point scale a person provides all tens, the only information that can be ascertained is that the relative level of satisfaction across all dimensions is the same (Rossi et al. 2001).

With the models and inferential strategies laid out, a discussion of the MCMC strategies to sample from these models is required.

4.3 MCMC Strategies for Scale-Usage Models

MCMC methods are now developed for the model discussed by Hans et al. (2012) which was defined in (4.3). This model is selected over the model in Rossi et al. (2001) for its computational advantages and the flexibility of the structure of the cutpoint parameters.

The overall MCMC strategy to sample from the posterior distribution of this model will be to utilize a Gibbs sampler due to the convenient form of the complete
conditional distributions. To identify the complete conditional distributions, note
that the full posterior distribution has the following form:
\[
\pi(Y, \mu, \Sigma, \tau, \sigma^2, c | X) \propto |\Sigma|^{-N/2} \prod_{i=1}^{N} (\sigma_i^2)^{-M/2} \prod_{j=1}^{M} (c_{X_{i,j-1}} < Y_{ij} \leq c_{X_{i,j}}) \\
\times \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_i^{-2}(Y_i - \mu - \tau_i 1)^T \Sigma^{-1}(Y_i - \mu - \tau_i 1) \right) \\
\times \prod_{i=1}^{N} (\sigma_i^2)^{-a/2-1} \exp \left( -\frac{(a-2)}{2\sigma_i^2} \right) \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \frac{\tau_i^2}{\sigma_i^2} \right) \\
\times |\Sigma|^{-(\delta+M+1)/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma_0 \Sigma^{-1}) \right) \exp \left( -\frac{1}{2} \mu^T V^{-1} \mu \right) \\
\times \prod_{i=2}^{K-1} (c_k - c_{k-1})^{\alpha_k-1}. \quad (4.5)
\]

One intuitive sampler simply circulates among the complete conditional distributions
of each parameter.

**Algorithm 4.1** (Naive Gibbs Sampler)

Given \(Y', \mu', \Sigma', \tau', \sigma'^2\), and \(c'\), update

(i) \(\mu|Y', \tau', \Sigma', \sigma'^2 \sim \text{MVN}_M(\gamma, \Psi)\), where

\[
\Psi = \left( \Sigma'^{-1} \sum_{i=1}^{N} \sigma_i'^{-2} + V^{-1} \right)^{-1}, \quad \text{and}
\]

\[
\gamma = \Psi \Sigma'^{-1} \sum_{i=1}^{N} \sigma_i'^{-2}(Y_i' - \tau_i' 1_M).
\]

(ii) \(\Sigma|Y', \mu, \tau', \sigma'^2 \sim \text{Inverse Wishart}(N + \delta, S_N + \Sigma_0)\) where

\[
S_N = \sum_{i=1}^{N} (Y_i' - \mu - \tau_i' 1)(Y_i' - \mu - \tau_i' 1)^T \sigma_i'^{-2}.
\]

(iii) \(\tau_i|Y_i', \mu, \Sigma, \sigma_i^2 \sim \text{Normal}(\mu_\tau, s_\tau^2)\) where

\[
s_\tau^2 = (\sigma_i^{-2} 1^T \Sigma^{-1} 1 + \sigma_\tau^{-2})^{-1}, \quad \text{and}
\]

\[
\mu_\tau = s_\tau^2 \sigma_i^{-2} (Y_i' - \mu)^T \Sigma^{-1} 1.
\]

128
(iv) $\sigma_i^2|Y_i', \mu, \tau_i, \Sigma \sim \text{Inverse Gamma}(a_\sigma, b_\sigma)$

$$a_\sigma = \frac{a + M}{2}, \quad \text{and} \quad b_\sigma = \frac{1}{2} \left[ (Y'_i - \mu - \tau_i 1)^T \Sigma^{-1} (Y'_i - \mu - \tau_i 1) + a - 2 \right].$$

(v) $Y_i|\mu, \tau_i, \Sigma, \sigma_i^2, c' \sim \text{Multivariate Truncated Normal}(\mu + \tau_i 1, \sigma_i^2 \Sigma, c_{X_{i,-1}, c_{X_i}'}).$

(vi) $c_k|X, Y \sim \text{Uniform}(a_{c_k}, b_{c_k})$, where

$$a_{c_k} = \max_{(i,j):X_{i,j}=k} Y'_{ij},$$

$$b_{c_k} = \min_{(i,j):X_{i,j}=k+1} Y'_{ij}.$$

Note this applies to the situation where $\alpha_1 = \ldots = \alpha_k = 1$, which corresponds to a uniform prior for the distance between free cutpoints. Otherwise, the uniform update is replaced by an update from a properly shifted and scaled beta distribution truncated by $a_{c_k}$ and $b_{c_k}$.

This Gibbs sampler is less than ideal, however, because $a_{c_k}$ and $b_{c_k}$ are often extremely close together, resulting in very slow convergence of the cutpoint parameters (Hans et al. 2012). A solution to this problem is to integrate out some of the $Y_i$, thus reducing the constraints on the cutpoints. This, however, also creates challenges as this integration requires the evaluation of multivariate normal integrals across the respective cutpoint intervals. Unfortunately, even in low-dimensional settings, this often cannot be done with perfect precision. These small errors in the Markov chain acceptance probability for $c_k$ no longer guarantee that the Markov chain is stationary for the target distribution and can often lead to substantially different results (Hans et al. 2012).
Another problem of this algorithm is that obtaining exact draws from a truncated multivariate normal distribution with a non-diagonal covariance matrix is very computationally expensive (and in some cases effectively impossible) for more than a few dimensions. One solution by Geweke (1991) remedying this problem is to update each dimension of $Y_i$ one at a time conditional on the others, yielding a series of univariate truncated normal draws. This, however, increases computational cost and autocorrelation compared to a single joint update.

Fortunately, Hans et al. (2012) propose a method that solves both these problems. They decompose the covariance matrix $\Sigma$ into $\Sigma = D + R$ where $D$ is diagonal and $R$ is either nonnegative definite or positive definite depending on the decomposition used; assume $R$ is positive definite for the remainder of the chapter. For their decomposition algorithm, see Algorithm 2.4. The model is then modified such that

\[ Y_i \sim \text{MVN}_M(\mu + \tau_i 1 + Z_i, \sigma_i^2 D), \quad i = 1, \ldots, N, \]
\[ Z_i \sim \text{MVN}_M(0, \sigma_i^2 R), \quad i = 1, \ldots, N, \]

and the remaining prior distributions are unchanged. The marginal distribution of $Y_i$ with $Z_i$ integrated out is the same as that of the original model in (4.3) for $i = 1, \ldots, N$.

As $D$ is diagonal, the updates which previously were multivariate truncated normal have become univariate truncated normal, for which efficient sampling algorithms exist. Similarly, the once-multivariate integral which would be required to update $c$ and $Y$ together in the same block by integrating out $Y$ before updating $c$ is reduced to a series of univariate normal integrals, which are also readily available at essentially perfect accuracy. Their method does require updating $c$ using a Metropolis-Hastings step rather than directly from its complete conditional distribution, but such is much
more efficient than the original Gibbs updates which were constrained by the $Y$ vari-
ables to explore the parameter space slowly (for illustration and discussion, see Hans et al. (2012)). The algorithm is now provided.

**Algorithm 4.2** (Block Gibbs Sampler)

Proceed as in Algorithm 4.1, replacing steps (v) and (vi) with the following:

(v) Decompose $\Sigma$ into $D$ and $R$ according to Algorithm 2.4.

(vi) Update $Z_i|Y', \mu, \tau, \sigma^2, \Sigma \sim \text{MVN}_M(\mu_Z, \Sigma_Z)$ where

$$
\mu_Z = (D^{-1} + R^{-1})^{-1}D^{-1}(Y'_i - \mu - \tau_i 1), \quad \text{and} \quad \Sigma_Z = \sigma_i^2(D^{-1} + R^{-1})^{-1}.
$$

(vii) Update $c_k|c_{k-1}, c_{k+1}', \mu, \tau, Z, \Sigma$ sequentially with a Metropolis-Hastings step utilizing some candidate distribution $g(\cdot|c_k)$ to generate a candidate $c_k^*$ which is accepted with probability

$$
\alpha_k = \min \left\{ 1, \left(\frac{c_{k+1} - c_k^*}{c_{k+1} - c_k} \right)^{\alpha_{k+1}-1} \left(\frac{c_k^* - c_{k-1}}{c_k - c_{k-1}} \right)^{\alpha_k-1} \frac{w_k(c_k^*)g(c_k^*|c_k)}{w_k(c_k)g(c_k^*|c_k)} \right\},
$$

where

$$
w_k(c_k) = \prod_{\ell = k}^{k+1} \prod_{(i,j):X_{ij} = \ell} \Phi \left( \frac{c_{\ell} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right).
$$

(viii) Update $Y_{ij}|\mu, \tau_i, \Sigma, \sigma_i^2, c \sim \text{Truncated Normal}(\mu_j + \tau_i + Z_{ij}, \sigma_i^2 D_{jj}, c_{X_{ij},-1}, c_{X_{ij}})$.

This algorithm was used by Hans et al. (2012) and appears to have good convergence properties. After a review of mathematical properties required for proving drift and minorization, formal arguments are made establishing that this Markov chain, when slightly modified, is geometrically ergodic.
4.4 Review of Relevant Matrix and Distributional Properties

Several matrix and mathematical properties are required for the proofs of geometric ergodicity including results about positive definite matrices, eigenvalues, determinants, and expected values. Some of these results were utilized in Chapter 2 as well, but are presented here again for convenience.

4.4.1 Positive Definite Matrices

First, the definition of positive definite matrices is given, along with some notation. Several properties of these matrices are then provided.

**Definition 4.1** (Horn and Johnson, 2013, Chapter 7). Let \( x \) be an \( n \times 1 \) vector whose entries are not all zero. An \( n \times n \) matrix \( A \) is *positive definite* if

\[
x^T A x > 0,
\]

for all \( x \). Similarly, an \( n \times n \) matrix \( A \) is *nonnegative definite* if

\[
x^T A x \geq 0,
\]

for all \( x \).

**Result 4.2** (Horn and Johnson, 2013, Observation 7.1.4). Each eigenvalue of a positive definite matrix is positive; each eigenvalue of a nonnegative definite matrix is nonnegative.

**Result 4.3** (Horn and Johnson, 2013, Observation 7.1.8). Let \( A \) be an \( n \times n \) matrix with \( C \) an \( n \times m \) matrix.

1. If \( A \) is nonnegative definite, then \( C^T A C \) is nonnegative definite.
2. If $A$ is positive definite, then $C^TAC$ is positive definite if and only if $\text{rank}(C) = m$.

**Definition 4.4** (Horn and Johnson, 2013, Definition 7.7.1). Let $A$ and $B$ be $n \times n$ matrices. Then, write

1. $A \geq 0$ if $A$ is nonnegative definite.

2. $A > 0$ if $A$ is positive definite.

3. $A \geq B$ if $A - B$ is nonnegative definite.

4. $A > B$ if $A - B$ is positive definite.

**Result 4.5** (Horn and Johnson, 2013, Corollary 7.7.4). Let $A$, $B$, and $C$ be $n \times n$ matrices.

1. If $A > 0$ and $B > 0$ then $A \geq B$ if and only if $B^{-1} \geq A^{-1}$.

2. If $A > 0$ and $B \geq 0$. Then $A^{-1} \geq (A + B)^{-1} > 0$.

3. If $A \geq B$ then $\text{tr}(A) \geq \text{tr}(B)$ with equality if and only if $A = B$.

4. If $B > 0$ and $C \geq A$, then $\text{tr}(AB) \leq \text{tr}(CB)$.

**Proof.** Corollary 7.7.4 in Horn and Johnson (2013) directly states (1) and (3). Then, (2) follows directly from (1). To prove (4), note first that $C \geq A$ implies $C - A \geq 0$. As $C - A \geq 0$, by Result 4.3, $B^{1/2}(C - A)B^{1/2} \geq 0$ and thus

$$B^{1/2}AB^{1/2} \leq B^{1/2}CB^{1/2}. \quad (4.7)$$

It then follows that

$$\text{tr}(AB) = \text{tr}(AB^{1/2}B^{1/2})$$
because as $B > 0$, a square root matrix exists,

$$= \text{tr}(B^{1/2}AB^{1/2})$$

by Result 4.11

$$\leq \text{tr}(B^{1/2}CB^{1/2})$$

by (4.7) and statement (3) of this result

$$= \text{tr}(CB^{1/2}B^{1/2})$$

by Result 4.11

$$= \text{tr}(CB),$$

as desired. \hfill \Box

### 4.4.2 Eigenvalues

The following basic results about eigenvalues are useful.

**Result 4.6** *(Ravishanker and Dey, 2002, Result 1.2.14)*. Assume $\lambda$ is an eigenvalue of $A$ and $c$ is a real scalar. Then

1. $\lambda^k$ is an eigenvalue of $A^k$, for any integer $k$.

2. $c\lambda$ is an eigenvalue of $cA$.

3. $\lambda + c$ is an eigenvalue of $A + cI$

**Result 4.7** *(Ravishanker and Dey, 2002, Result 1.2.15)*. Let $A$ be an $n \times n$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then

1. $\text{tr}(A) = \sum_{i=1}^{n} \lambda_i$.

2. $|A| = \prod_{i=1}^{n} \lambda_i$. 

134
**Result 4.8.** Let $A$ be a symmetric $n \times n$ matrix with ordered eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_n,$$

and let $x$ be any $n \times 1$ vector. Then

$$\lambda_n x^T x \leq x^T Ax \leq \lambda_1 x^T x.$$

**Proof.** Let $A$ be a symmetric $n \times n$ matrix. Then, by Result 2.3.4 of Ravishanker and Dey (2002),

$$A = Q\Lambda Q^T,$$

where $Q$ is an orthogonal matrix (i.e. $Q^T = Q^{-1}$) and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n$ are the ordered eigenvalues of $A$. Let $x$ be any $n \times 1$ vector. Then,

$$x^T Ax = x^T Q\Lambda Q^T x$$

$$\leq \lambda_1 x^T QQ^T x$$

$$= \lambda_1 x^T x.$$

Similarly,

$$x^T Ax = x^T Q\Lambda Q^T x$$

$$\geq \lambda_n x^T QQ^T x$$

$$= \lambda_n x^T x,$$

which provides the desired inequality. \qed

**4.4.3 Determinants**

The following properties of determinants are also important.
Result 4.9 (Ravishanker and Dey, 2002, Result 1.2.9). Let $A$ be any $n \times n$ matrix, $c$ be any real scalar, and $k$ be any integer. Then

1. $|cA| = c^n|A|$.
2. $|A^k| = |A|^k$.

Result 4.10 (Horn and Johnson, 2013, p. 511). If $A$ and $B$ are positive definite matrices, then

$$|A + B| \geq |A| + |B|.$$ 

4.4.4 Other Matrix Properties

The following are miscellaneous matrix properties which are useful for the proofs in this chapter.

Result 4.11 (Cyclical Trace Property, Ravishanker and Dey, 2002, Result 1.2.8). Assume $A$ is an $n \times m$ matrix, $B$ is an $m \times \ell$ matrix, and $C$ is an $\ell \times n$ matrix. Then

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB).$$ 

Result 4.12 (Sherman-Morrison-Woodbury Formula, Horn and Johnson, 2013). Let $A$ and $R$ be invertible matrices of dimensions $n \times n$ and $r \times r$ respectively. Let $Y$ be $r \times n$ and $X$ be $n \times r$. If $A + XRY$ and $R^{-1} + YA^{-1}X$ are invertible, then

$$(A + XRY)^{-1} = A^{-1} - A^{-1}X(R^{-1} + YA^{-1}X)^{-1}YA^{-1}.$$ 

Result 4.13 (Bounding the Elements of $A$). Assume $A$ is a symmetric positive definite $n \times n$ matrix and that $\text{tr}(A) \leq d$ where $d < \infty$. Then all the elements of $A$ are bounded.
Proof. Assume \( \text{tr}(A) \leq d \). Note from Result 4.7 that

\[
\text{tr}(A) = \sum_{i=1}^{n} \lambda_i,
\]

where \( \lambda_i \) are the eigenvalues of \( A \). As \( A \) is positive definite, we have that the eigenvalues are all positive by Result 4.2, which then implies that all the eigenvalues are bounded by \( d \) as the trace is bounded by \( d \). This implies, by Result 4.6, that

\[
\text{tr}(AA) = \sum_{i=1}^{n} \lambda_i^2 < nd^2,
\]

and is thus bounded. Note that \( \text{tr}(AA) = \text{tr}(A^T A) \) by symmetry and

\[
\text{tr}(A^T A) = \text{tr} \left( \begin{bmatrix} A_{11} & A_{21} & A_{31} & \ldots & A_{n1} \\ \vdots & \vdots & \vdots & & \vdots \\ A_{1n} & A_{2n} & A_{3n} & \ldots & A_{nn} \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & A_{13} & \ldots & A_{1n} \\ \vdots & \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \ldots & A_{nn} \end{bmatrix} \right)
= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2,
\]

which, being bounded by \( nd^2 \), implies that each element of the matrix \( A \) is bounded.

\[\square\]

Result 4.14 (Marchev and Hobert, 2004, Lemma 1). Let \( A \) be a positive definite \( n \times n \) matrix and \( x \) be an \( n \times 1 \) vector, then

\[x^T (xx^T + A)^{-1} x < 1.\]

4.4.5 Expected Values

The following are results relating to expectations which are used throughout the proofs of geometric ergodicity.
Result 4.15 (Iterative Expectations, Resnick, 1999). Let $X$ and $Y$ be random variables. Then,

$$E(X) = E(E(X|Y)).$$

Definition 4.16 (Expected Value of a Matrix). Let $X$ be an $n \times m$ random matrix. Then,

$$E(X) = \begin{pmatrix} E(X_{1,1}) & E(X_{1,2}) & \cdots & E(X_{1,m}) \\ E(X_{2,1}) & E(X_{2,2}) & \cdots & E(X_{2,m}) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_{n,1}) & E(X_{n,2}) & \cdots & E(X_{n,m}) \end{pmatrix}.$$

Result 4.17 (Properties of Matrix Expected Values). Let $X$ be an $n \times n$ random matrix and $A$ be an $m \times n$ constant matrix. Then, the following properties hold:

1. $E(\text{tr}(X)) = \text{tr}(E(X))$.
2. $E(AX) = AE(X)$.

Proof. Result (1) is easily seen from Definition 4.16 and the definition of the trace of a matrix (the sum of the diagonal elements of an $n \times n$ matrix). Result (2) is also easily seen by observing that the inner product of the first row of $A$ and the first column of $X$ is

$$E(A_{1,1}X_{1,1} + A_{1,2}X_{2,1} + \cdots + A_{1,n}X_{n,1}) = A_{1,1}E(X_{1,1}) + A_{1,2}E(X_{2,1}) + \cdots + A_{1,n}E(X_{n,1}),$$

which is the inner product of the first row of $A$ and the first column of $E(X)$. As this is satisfied for every such row and column combination, the property is satisfied.

Result 4.18. Let $X$ be an $n \times 1$ random vector with mean $\mu$ and covariance matrix $\Sigma$. Also, let $c$ be a constant $n \times 1$ vector and $A$ be a constant $n \times n$ matrix. Then,

$$E((X - c)^T A(X - c)) = \text{tr}(A\Sigma) + (c - \mu)^T A(c - \mu).$$
Proof. Note that

\[ E((X - c)^t A(X - c)) = E[(X - \mu - (c - \mu))^t A(X - \mu - (c - \mu))] \]

\[ = E[(X - \mu)^t A(X - \mu)] - 2(c - \mu)^t A E[X - \mu] + (c - \mu)^t A(c - \mu) \]

by Result 4.17 (2)

\[ = E[tr((X - \mu)^t A(X - \mu))] - 2(c - \mu)^t A0 + (c - \mu)^t A(c - \mu) \]

because \( E(X) = \mu \) and the trace of a \( 1 \times 1 \) matrix is equal to the matrix itself

\[ = E[tr(A(X - \mu)(X - \mu)^t)] + (c - \mu)^t A(c - \mu) \]

by Result 4.11

\[ = tr(A E[(X - \mu)(X - \mu)^t]) + (c - \mu)^t A(c - \mu) \]

Result 4.17 (1) and (2)

\[ = tr(A \Sigma) + (c - \mu)^t A(c - \mu) \]

by definition of the covariance matrix, as desired. \( \square \)

The next results are concerning expectations for particular distributions and are used in the proofs of geometric ergodicity.

Result 4.19 (Expected Value of Inverse Gamma Distribution). Let \( X \) have an inverse gamma distribution with shape \( a \) and scale \( b \) which corresponds to the probability density function

\[ f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp \left( -\frac{b}{x} \right). \]

Then

\[ E(X) = \frac{b}{a - 1}, \quad a > 1, \]

\[ \text{Var}(X) = \frac{b^2}{(a - 1)^2(a - 2)}, \quad a > 2, \]

\[ 139 \]
\[ E(X^{-1}) = \frac{a}{b}. \]

**Result 4.20** (Expected Value of an Inverse Wishart Distribution). Let \( X \), a \( p \times p \) matrix, have an inverse Wishart distribution with \( \nu \) degrees of freedom and scale matrix \( \Psi \) with probability density function

\[ f(X|\nu, \Psi) \propto |X|^{-\frac{\nu+p+1}{2}} \exp\left(-\frac{1}{2} \text{tr}(\Psi X^{-1})\right). \]

Then,

\[ E(X) = \frac{\Psi}{\nu - p - 1} \]

and

\[ E(X^{-1}) = \nu \Psi^{-1}. \]

### 4.5 Geometric Ergodicity of Scale-Usage Models

Geometric ergodicity is now proved for Markov chains stationary for several specific models related to (4.3) and its decomposition extension in (4.6). A summary of the modifications for all models include a flat prior for \( \mu \) as in Rossi et al. (2001), a prior for \( \tau_i \) with bounded support, and setting \( c_0 \) along with \( c_k \) to be finite constants. Intuitive justification for this last restriction is that there are an additional two levels of the measurement scale which are so extreme they are never observed, one above and one below the current scale. The purpose of this restriction is to bound all of the \( Y \) latent variables. Formal model definitions, Markov chains, and proofs of geometric ergodicity follow for each model. For notation throughout, define the entire collection of parameters as \( \theta = \{ Y, \mu, Z, \tau, \Sigma, \sigma, c \} \), and let \( q_i = \sigma_i^{-2} \), where \( i \) is any index.
4.5.1 Model 4.1: No Scale-Usage Effects

To introduce the process of proving geometric ergodicity for scale-usage models, the first model considered is simple with no scale-usage effects and does not use the decomposition sampler. Formally, the model is

\[ Y_i \sim \text{MVN}_M(\mu, \Sigma), \quad i = 1, \ldots, N \]

\[ X_{ij} | c, Y_{ij} = \{k : c_{k-1} < Y_{ij} \leq c_k\}, \quad i = 1, \ldots, N; \ j = 1, \ldots, M \]

\[ \Sigma \sim \text{Inverse Wishart}_M(\delta, \Sigma_0) \]

\[ \pi(\mu) \propto 1 \]

\[ \pi(c) \propto \prod_{k=1}^{K} (c_k - c_{k-1})^{\alpha_k-1}, \text{ where } c_0 = -C \text{ and } c_K = C, \tag{4.8} \]

where \( \delta, \Sigma_0, \) and \( \alpha_k \) for \( k = 1, \ldots, K \) are all constants. The sampler for this model is provided in Algorithm 4.3.

**Algorithm 4.3** (MCMC Sampler with No Scale-Usage Effects)

From the current \( \theta' \), update to \( \theta \) according to the following steps.

(i) \( \Sigma | Y', \mu' \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0) \) where

\[ S' = \sum_{i=1}^{N} (Y'_i - \mu') (Y'_i - \mu')^T. \]

(ii) \( Y_i | \mu', \Sigma, c' \sim \text{Multivariate Truncated Normal}(\mu', \Sigma, c'_{X_{i-1}}, c'_{X_i}). \)

(iii) \( c_k | X, Y \sim \text{Uniform}(a_{c_k}, b_{c_k}) \), where

\[ a_{c_k} = \max_{(i,j): X_{i,j} = k} Y_{ij} \]

\[ b_{c_k} = \min_{(i,j): X_{i,j} = k+1} Y_{ij}. \]
(iv) Update $\mu|\mathbf{Y}, \Sigma \sim \text{MVN}_M(\mathbf{Y}, \Sigma/N)$, where

$$\mathbf{Y} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i.$$ 

The proof for geometric ergodicity for this sampler will proceed as in Method 2 by showing a drift condition and unboundedness off compact sets. As the transition densities are all continuous, the chain is Feller by Proposition 2.25. Note also that $\psi$ is Lebesgue measure, and thus has nonempty interior. These two conditions are similarly satisfied for the remaining models in this chapter and are hereafter assumed without mention. The candidate drift function $V(\theta)$ is

$$V(\theta) = \sum_{i=1}^{N} (\mathbf{Y}_i - \mu)^T (\mathbf{Y}_i - \mu) + \text{tr}(\Sigma).$$

First, a proof that $V$ is unbounded off compact sets for the Markov chain is given. As typical, $V$ is continuous, and thus $C$ is closed; it remains to show that $\{\theta : V(\theta) \leq d\}$ is bounded. Assume $V(\theta) \leq d$; then the following statements are true.

- The parameters $\mathbf{Y}_i, i = 1, \ldots, N$ and $\mathbf{c}$ are bounded by the model itself because

$$c_0 = -C \leq c_1 \leq \cdots \leq c_K = C,$$

and because $C < \infty$, this implies that $c_k$ is bounded for $k = 1, \ldots, K - 1$. Accordingly, as each $Y_{ij} \in (c_{X_{ij}-1}, c_{X_{ij}})$, then $Y_{ij}$ must be between $-C$ and $C$ for $i = 1, \ldots, N$ and $j = 1, \ldots, M$. As $C < \infty$, each $Y_{ij}$ and hence $\mathbf{Y}_i$ is bounded for all $i = 1, \ldots, N$.

- As

$$\sum_{i=1}^{N} (\mathbf{Y}_i - \mu)^T (\mathbf{Y}_i - \mu) \leq d,$$
each component of the sum is also less than \( d \),

\[
(Y_i - \mu)^T(Y_i - \mu) \leq d, \quad i = 1, \ldots, N.
\]

As \( Y_i \) is bounded by the model (for any \( i \)), this implies that \( \mu \) is bounded.

- As \( \text{tr}(\Sigma) \) is bounded, every element of \( \Sigma \) is bounded according to Result 4.13.

Therefore, \( \theta \) is bounded if \( V(\theta) \leq d \), and \( V \) is unbounded off compact sets for the Markov chain.

To show that a drift condition is satisfied, begin with the first part of \( V \). Recall that

\[
\mu|Y, \Sigma \sim \text{MVN}_M(\bar{Y}, \Sigma/N).
\]

Then,

\[
\mathbb{E}\left( \sum_{i=1}^{N} (Y_i - \mu)^T(Y_i - \mu) \bigg| Y, \Sigma \right)
\]

\[
= \sum_{i=1}^{N} (Y_i - \bar{Y})^T(Y_i - \bar{Y}) + N \text{tr} \left( \frac{1}{N} \Sigma \right)
\]

by Result 4.18

\[
= \sum_{i=1}^{N} Y_i^T Y_i - 2\bar{Y}^T \sum_{i=1}^{N} Y_i + N\bar{Y}^T \bar{Y} + \text{tr}(\Sigma)
\]

\[
\leq \sum_{i=1}^{N} Y_i^T Y_i + \text{tr}(\Sigma) \quad (4.9)
\]

which is true because

\[
-2\bar{Y}^T \sum_{i=1}^{N} Y_i + N\bar{Y}^T \bar{Y} = -2N\bar{Y}^T \bar{Y} + N\bar{Y}^T \bar{Y} = -N\bar{Y}^T \bar{Y} < 0.
\]

Continuing the inequality expression, it is observed that

\[
\mathbb{E}\left( \sum_{i=1}^{N} (Y_i - \mu)^T(Y_i - \mu) \bigg| Y, \Sigma \right) \leq \sum_{i=1}^{N} Y_i^T Y_i + \text{tr}(\Sigma)
\]

143
by Result 4.20 and Result 4.17

\begin{equation}
\frac{\sum_{i=1}^{N} (Y_i' - \mu')^T (Y_i' - \mu')}{N + \delta - M - 1} + \frac{\text{tr}(\Sigma_0)}{N + \delta - M - 1},
\end{equation}

by substituting $S'$. Then, it is observed that

\begin{equation*}
E(V(\theta)|\theta') = E \left( \sum_{i=1}^{N} (Y_i - \mu)^T (Y_i - \mu) + \text{tr}(\Sigma) \bigg| \theta' \right) = E \left( E \left( \sum_{i=1}^{N} (Y_i - \mu)^T (Y_i - \mu) + \text{tr}(\Sigma) \bigg| Y, \Sigma \right) \bigg| \theta' \right)
\end{equation*}

by Result 4.15

\begin{equation*}
\leq E \left( 2 \text{tr}(\Sigma) + NMC^2 \bigg| \theta' \right)
\end{equation*}

by (4.10)

\begin{equation*}
\leq \frac{2}{N + \delta - M - 1} \sum_{i=1}^{N} (Y_i' - \mu')^T (Y_i' - \mu')
\end{equation*}
by (4.11)

\[
\leq \lambda V(\theta') + b,
\]

where

\[
\lambda = \frac{2}{N + \delta - M - 1}
\]

and

\[
b = \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + NMC^2.
\]

Thus, when \( \lambda < 1 \), a drift condition (2.9) is satisfied for the Markov chain induced by Algorithm 4.3, and coupled with the unboundedness off compact sets which was already established, the chain is geometrically ergodic. This requires

\[N + \delta - M - 1 > 2,\]

but this is nearly always true in practice. Now, continue to a model which allows for scale-usage effects.

### 4.5.2 Model 4.2: Constrained Scale-Usage Effects

This model expands Model 4.1 by adding the scale-usage effects \( \tau_i \) and \( \sigma_i^2 \). As discussed previously, the parameter space for \( \tau_i \) is constrained. This model also constrains the parameter space for \( \sigma_i^2 \) with upper and lower (non-zero) bounds. Formally, the model is

\[
Y_i \sim \text{MVN}_M(\mu + \tau_i \cdot 1, \sigma_i^2 \Sigma), \quad i = 1, \ldots, N
\]

\[
X_{ij} | c, Y_{ij} = \{k : c_{k-1} < Y_{ij} \leq c_k\}, \quad i = 1, \ldots, N; j = 1, \ldots, M
\]

\[
\sigma_i^2 \sim \text{Truncated Inverse Gamma}(a/2, a/2 - 1, a_{\sigma_i^2}^{-1}, a_{\sigma^2}), \quad i = 1, \ldots, N
\]
\[ \tau_i \sim \text{Truncated Normal}(0, \sigma^2_{\tau_i}, -a_\tau, a_\tau), \quad i = 1, \ldots, N \]

\[ \Sigma \sim \text{Inverse Wishart}_M(\delta, \Sigma_0) \]

\[ \pi(\boldsymbol{\mu}) \propto 1 \]

\[ \pi(\mathbf{c}) \propto \prod_{k=1}^{K} (c_k - c_{k-1})^{a_k - 1}, \quad \text{where} \quad c_0 = -C \text{ and } c_K = C, \]

(4.12)

where \( a, a_\sigma^2, \sigma^2_\tau, a_\tau, \delta, \Sigma_0, \) \( \text{and} \ a_k \) for \( k = 1, \ldots, K \) are all constants. The corresponding Markov chain is now defined as Algorithm 4.4.

**Algorithm 4.4** (MCMC Sampler with Constrained Scale-Usage Effects)

From the current \( \theta' \), update to \( \theta \) according to the following steps.

(i) \( \Sigma | Y', \mu', \sigma^2_i \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0) \) where

\[ S' = \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i \mathbf{1}) (Y'_i - \mu' - \tau'_i \mathbf{1})^T q_i'. \]

(ii) \( \sigma^2_i | Y'_i, \mu', \tau'_i, \Sigma \sim \text{Truncated Inverse Gamma}((a + M)/2, b_{\sigma^2_i}, a_{\sigma^2_i}, a_{\sigma^2_i}) \) where

\[ b_{\sigma^2_i} = \frac{(Y'_i - \mu' - \tau'_i \mathbf{1})^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i \mathbf{1}) + a - 2}{2}. \]

(iii) \( Y_i | \mu', \sigma^2_i, \tau'_i, c' \sim \text{Multivariate Truncated Normal}(\mu' + \tau'_i \mathbf{1}, \sigma^2_i \Sigma, \mathbf{c}'_{X_i - 1}, \mathbf{c}'_{X_i}). \)

(iv) \( c_k | \mathbf{X}, Y \sim \text{Uniform}(a_{ck}, b_{ck}) \), where

\[ a_{ck} = \max_{(i,j):X_{i,j}=k} Y_{ij} \]

\[ b_{ck} = \min_{(i,j):X_{i,j}=k+1} Y_{ij}. \]

(v) \( \tau_i | Y_i, \mu', \Sigma, \sigma^2_i \sim \text{Truncated Normal}(\mu_{\tau_i}, s^2_{\tau_i}, -a_\tau, a_\tau), \) where

\[ s^2_{\tau_i} = (q_i \mathbf{1}^T \Sigma^{-1} \mathbf{1} + q_\tau)^{-1} \]

and

\[ \mu_{\tau_i} = s^2_{\tau_i} q_i (Y_i - \mu')^T \Sigma^{-1} \mathbf{1}. \]
(vi) \( \mu | Y, \tau, \Sigma, \sigma^2 \sim \text{MVN}_M(\hat{Y}, \Sigma / q^*) \), where

\[
q^* = \sum_{i=1}^{N} q_i, \quad \text{and} \quad \hat{Y} = \frac{1}{q^*} \sum_{i=1}^{N} (Y_i - \tau_i 1) q_i.
\]

The proof for geometric ergodicity for this sampler will also proceed as in Method 2 by showing a drift condition and unboundedness off compact sets. The candidate drift function \( V(\theta) \) is

\[
V(\theta) = \sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) + \text{tr}(\Sigma).
\]

First, consider a proof that \( V \) is unbounded off compact sets for the Markov chain. Assume \( V(\theta) \leq d \); then the following statements are true.

- The parameters \( Y_i, i = 1, \ldots, N, \) and \( c \) are bounded by the model itself, as with Model 4.1.

- The parameters \( \tau_i, \sigma^2_i \) are bounded directly by the truncated prior distributions for \( i = 1, \ldots, N. \)

- As

\[
\sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) \leq d,
\]

each component of the sum is also smaller than \( d \) which implies that

\[
(Y_i - \mu - \tau_i)^T(Y_i - \mu - \tau_i) \leq d \sigma^2_i \leq da \sigma^2, \quad i = 1, \ldots, N.
\]

As \( Y_i \) and \( \tau_i \) are both bounded by the model (for any \( i \)), this implies that \( \mu \) is bounded.

- As \( \text{tr}(\Sigma) \) is bounded, every element of \( \Sigma \) is bounded according to Result 4.13.
Therefore, $\theta$ is bounded if $V(\theta) \leq d$ and $V$ is unbounded off compact sets for the Markov chain.

To show that a drift condition is satisfied, begin with the first part of $V$. Recall that

$$
\mu|Y, \tau, \Sigma, \sigma^2 \sim \text{MVN}_M(\bar{Y}, \Sigma/q^*),
$$

where

$$
q^* = \sum_{i=1}^{N} q_i, \quad \text{and} \quad \bar{Y} = \frac{1}{q^*} \sum_{i=1}^{N} (Y_i - \tau_i 1) q_i.
$$

Then, note that

$$
E\left( \sum_{i=1}^{N} q_i(Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) \right| Y, \tau, \Sigma, \sigma^2)
$$

$$
= \sum_{i=1}^{N} q_i(Y_i - \tau_i 1 - \bar{Y})^T(Y_i - \tau_i 1 - \bar{Y}) + \text{tr} (\Sigma)(q^*)^{-1} \sum_{i=1}^{N} q_i
$$

by Result 4.18

$$
= \sum_{i=1}^{N} q_i(Y_i - \tau_i 1)^T(Y_i - \tau_i 1) - 2\bar{Y}^T \sum_{i=1}^{N} q_i(Y_i - \tau_i 1) + \bar{Y}^T q^* \bar{Y} + \text{tr} (\Sigma)
$$

$$
\leq \sum_{i=1}^{N} q_i(Y_i - \tau_i 1)^T(Y_i - \tau_i 1) + \text{tr} (\Sigma)
$$

which is true because

$$
-2\bar{Y}^T \sum_{i=1}^{N} q_i(Y_i - \tau_i 1) + \bar{Y}^T q^* \bar{Y} = -\bar{Y}^T q^* \bar{Y} < 0
$$

which then yields

$$
\leq MK_1 \sum_{i=1}^{N} q_i + \text{tr} (\Sigma) \quad (4.13)
$$

where $K_1$ is the maximum squared distance between any $Y_{ij}$ and $\tau_i$, which is finite and computable because $Y_{ij} \in (c_0, c_K)$ and $\tau_i \in (-a_\tau, a_\tau)$,

$$
\leq NMK_1 a_\sigma^2 + \text{tr} (\Sigma), \quad (4.14)
$$

148
because $q_i \leq a_{\sigma^2}$.

Continuing, recall that

$$\Sigma|Y', \mu' \sim \text{Inverse Wishart}_M(N + \delta, S' + \Sigma_0)$$

where

$$S' = \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^T q'_i.$$ 

Then, note that

$$E(\text{tr}(\Sigma)|Y', \mu', \tau') = \frac{\text{tr}(S' + \Sigma_0)}{N + \delta - M - 1}$$

by Result 4.20 and Result 4.17

$$= \sum_{i=1}^{N} q'_i (Y'_i - \mu' - \tau'_i 1)^T (Y'_i - \mu' - \tau'_i 1) \frac{\text{tr}(\Sigma)}{N + \delta - M - 1} + \frac{\text{tr}(\Sigma_0)}{N + \delta - M - 1}, \quad (4.15)$$

by substituting $S'$.

Then, it is observed that

$$E(V(\theta)|\theta') = E \left( \sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T (Y_i - \mu - \tau_i 1) + \text{tr}(\Sigma) \bigg| \theta' \right)$$

$$= E \left( E \left( \sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T (Y_i - \mu - \tau_i 1) + \text{tr}(\Sigma) \bigg| Y, \tau, \sigma^2, \Sigma \right) \bigg| \theta' \right),$$

by Result 4.15

$$\leq E \left( 2 \text{tr}(\Sigma) + N M K_1 a_{\sigma^2} \bigg| \theta' \right),$$

by (4.14)

$$= \frac{2}{N + \delta - M - 1} \sum_{i=1}^{N} q'_i (Y'_i - \mu' - \tau'_i 1)^T (Y'_i - \mu' - \tau'_i 1)$$

$$+ \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + N M K_1 a_{\sigma^2}$$

149
by (4.15)

\[ \leq \lambda V(\boldsymbol{\theta}') + b, \]

where

\[ \lambda = \frac{2}{N + \delta - M - 1} \]

and

\[ b = \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + NMK_1a_\sigma^2. \]

Thus, when \( \lambda < 1 \), a drift condition (2.9) is satisfied for the Markov chain induced by Algorithm 4.4, and coupled with the unboundedness off compact sets which was already established, the chain is geometrically ergodic. Again, typically

\[ N + \delta - M - 1 > 2, \]

so \( \lambda < 1 \) with no difficulty.

### 4.5.3 Model 4.3: Constrained Scale-Usage Effects with Decomposition

This model keeps the same constrained prior distributions on \( \sigma_i^2 \) and \( \tau_i \), but utilizes the decomposition strategy of Hans et al. (2012). Formally, the model is

\[ Y_i \sim \text{MVN}_M(\mu + \tau_i 1 + Z_i, \sigma_i^2 D), \quad i = 1, \ldots, N \]

\[ X_{ij} | c, Y_{ij} = \{ k : c_{k-1} < Y_{ij} \leq c_k \}, \quad i = 1, \ldots, N; j = 1, \ldots, M \]

\[ Z_i \sim \text{MVN}_M(0, \sigma_i^2 R), \quad i = 1, \ldots, N \]

\[ \sigma_i^2 \sim \text{Truncated Inverse Gamma}(a/2, a/2 - 1, a_\sigma^{-1}, a_\sigma^2), \quad i = 1, \ldots, N \]

\[ \tau_i \sim \text{Truncated Normal}(0, \sigma_\tau^2, -a_\tau, a_\tau), \quad i = 1, \ldots, N \]

\[ \Sigma \sim \text{Inverse Wishart}_M(\delta, \Sigma_0) \]
\[ \pi(\mu) \propto 1 \]
\[ \pi(c) \propto \prod_{k=1}^{K} (c_k - c_{k-1})^{\alpha_k - 1}, \text{ where } c_0 = -C \text{ and } c_K = C, \] (4.16)
where \( \alpha, \sigma_\tau^2, \sigma_\tau, \delta, \Sigma_0, \) and \( \alpha_k \) for \( k = 1, \ldots, K \) are all constants. The algorithm is provided in Algorithm 4.5.

**Algorithm 4.5** (MCMC Sampler with Constrained Scale-Usage Effects and Covariance Decomposition)

From the current \( \theta' \), update to \( \theta \) according to the following steps.

(i) \( \Sigma|Y', \mu', \tau', \sigma^2 \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0) \) where
\[ S' = \sum_{i=1}^{N} (Y_i' - \mu' - \tau'_i 1)(Y_i' - \mu' - \tau'_i 1)^T q_i'. \]

(ii) \( \sigma^2_i|Y_i', \mu', \tau', \Sigma \sim \text{Truncated Inverse Gamma}((a + M)/2, b_{\sigma^2_i}, a_{\sigma^2_i}^{-1}, a_{\sigma^2}), \) where
\[ b_{\sigma^2_i} = \frac{(Y_i' - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau'_i 1) + a - 2}{2}. \]

(iii) \( Z_i|Y', \mu', \tau', \sigma^2, \Sigma \sim \text{MVN}_M(\mu_Z, \Sigma_Z) \) where
\[ \mu_Z = (D^{-1} + R^{-1})^{-1} D^{-1} (Y_i' - \mu' - \tau'_i 1), \quad \text{and} \]
\[ \Sigma_Z = \sigma^2_i (D^{-1} + R^{-1})^{-1}. \]

(iv) Update \( c_k|c_{k-1}, c_{k+1}', \mu', \tau', Z, \Sigma \) sequentially with a Metropolis-Hastings step utilizing some candidate distribution \( g(\cdot|c_k) \) to generate a candidate \( c_k' \) which is

\[ \alpha_k = \min \left\{ 1, \left( \frac{c_k + c_k'}{c_{k+1} - c_k} \right)^{\alpha_k + 1 - 1} \left( \frac{c_k' - c_{k-1}}{c_k - c_{k-1}} \right)^{\alpha_k - 1} \frac{w_k(c_k') g(c_k|c_k')}{w_k(c_k) g(c_k'|c_k')} \right\}, \]

where
\[ w_k(c_k) = \prod_{\ell=k}^{k+1} \prod_{(i,j):X_{ij} = \ell} \left\{ \Phi \left( \frac{c_{\ell} - \mu'_j - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_i D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu'_j - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_i D_{jj}}} \right) \right\}. \]
(v) $Y_{ij} | \mu', \tau_j', \Sigma, \sigma^2_i, c \sim \text{Truncated Normal}(\mu_j' + \tau_j' + Z_{ij}, \sigma^2_i D_{jj}, c_{X_{i,j-1}, c_{X_{i,j}}})$.

(vi) $\tau_i | Y_i, \mu', \Sigma, \sigma^2_i \sim \text{Truncated Normal}(\mu_{\tau_i}, s_{\tau_i}^2, -a_{\tau}, a_{\tau})$, where

$\begin{align*}
  s_{\tau_i}^2 &= (q_i \mathbf{1}^T \Sigma^{-1} \mathbf{1} + q_{\tau})^{-1} \\
  \mu_{\tau_i} &= s_{\tau_i}^2 q_i (Y_i - \mu')^T \Sigma^{-1} \mathbf{1}.
\end{align*}$

(vii) $\mu | Y, \tau, \Sigma, \sigma^2 \sim \text{MVN}_M(\hat{Y}, \Sigma/q^*)$, where

$q^* = \sum_{i=1}^N q_i$, and $\hat{Y} = \frac{1}{q^*} \sum_{i=1}^N (Y_i - \tau_i \mathbf{1}) q_i$.

Consider the modified drift function

$V(\theta) = \sum_{i=1}^N q_i (Y_i - \mu - \tau_i \mathbf{1})^T (Y_i - \mu - \tau_i \mathbf{1}) + \sum_{i=1}^N Z_i^T \Sigma^{-1} Z_i + \text{tr}(\Sigma)$.

For the proof of unboundedness off compact sets, note the following:

- The parameters $Y_i, i = 1, \ldots, N$, and $c$ are bounded by the model itself, as previously observed.

- The parameters $\tau_i$ and $\sigma^2_i$ for $i = 1, \ldots, N$ are bounded by the truncated prior distributions.

- As

$\sum_{i=1}^N q_i (Y_i - \mu - \tau_i \mathbf{1})^T (Y_i - \mu - \tau_i \mathbf{1}) \leq d$,

each component of the sum is also smaller than $d$ which implies that

$(Y_i - \mu - \tau_i)^T (Y_i - \mu - \tau_i) \leq d \sigma_i^2 \leq d a_{\sigma^2}, \quad i = 1, \ldots, N$.

As $Y_i$ and $\tau_i$ are both bounded by the model (for any $i$), this implies that $\mu$ is bounded.
As 
\[ \mathbf{Z}_i^T \Sigma^{-1} \mathbf{Z}_i \leq d, \quad i = 1, \ldots, N, \]
it is also true by Result 4.8 that 
\[ (\lambda_1)^{-1} \mathbf{Z}_i^T \mathbf{Z}_i \leq d, \quad i = 1, \ldots, N, \]
where \((\lambda_1)^{-1}\) is the smallest eigenvalue of \(\Sigma^{-1}\) (\(\lambda_1\) is the largest eigenvalue of \(\Sigma\)). This implies that 
\[ \mathbf{Z}_i^T \mathbf{Z}_i \leq d \lambda_1 \leq d \text{tr}(\Sigma) \leq d^2, \quad i = 1, \ldots, N, \]
and therefore every element of \(\mathbf{Z}_i\) is bounded for \(i = 1, \ldots, N\).

As \(\text{tr}(\Sigma)\) is bounded, every element of \(\Sigma\) is bounded according to Result 4.13.

To show the drift condition, recall the update for \(\mathbf{Z}_i\) is 
\[ \mathbf{Z}_i | \mathbf{Y}', \mu', \tau', \sigma^2, \Sigma \sim \text{MVN}_M(\mu_Z, \Sigma_Z) \]
where
\[ \mu_Z = (\mathbf{D}^{-1} + \mathbf{R}^{-1})^{-1} \mathbf{D}^{-1} (\mathbf{Y}_i' - \mu' - \tau'_1), \quad \text{and} \]
\[ \Sigma_Z = \sigma_i^2 (\mathbf{D}^{-1} + \mathbf{R}^{-1})^{-1}. \]

Then, note that 
\[
\begin{align*}
\mathbb{E} \left( \sum_{i=1}^{N} \mathbf{Z}_i^T \Sigma^{-1} \mathbf{Z}_i \bigg| \sigma^2, \Sigma, \mathbf{Y}', \mu', \tau' \right) \\
&= \sum_{i=1}^{N} (\mathbf{Y}_i' - \mu' - \tau'_1)^T \mathbf{D} (\mathbf{Y}_i' - \mu' - \tau'_1) \\
&\quad + \sum_{i=1}^{N} \sigma_i^2 \text{tr}(\Sigma^{-1} (\mathbf{D}^{-1} + \mathbf{R}^{-1})^{-1}) \quad (4.17)
\end{align*}
\]
by Result 4.18 where

\[
\tilde{D} = D^{-1}(D^{-1} + R^{-1})^{-1}\Sigma^{-1}(D^{-1} + R^{-1})^{-1}D^{-1}
\]

\[
= D^{-1}(D^{-1} + R^{-1})^{-1}D^{-1}\Sigma^{-1}DD^{-1}(D^{-1} + R^{-1})^{-1}D^{-1}
\]

\[
= (D^{-1} - \Sigma^{-1})D\Sigma^{-1}D(D^{-1} - \Sigma^{-1})
\]

by Result 4.12 because

\[
\Sigma^{-1} = (D + R)^{-1} = D^{-1} - D^{-1}(D^{-1} + R^{-1})^{-1}D^{-1},
\]

which continuing provides

\[
\tilde{D} = D^{-1}D\Sigma^{-1}DD^{-1} - D^{-1}D\Sigma^{-1}D\Sigma^{-1}
\]

\[
- \Sigma^{-1}D\Sigma^{-1}DD^{-1} + \Sigma^{-1}D\Sigma^{-1}D\Sigma^{-1}
\]

by multiplying out the cross product

\[
= \Sigma^{-1} - 2\Sigma^{-1}D\Sigma^{-1} + \Sigma^{-1}D\Sigma^{-1}D\Sigma^{-1}
\]

because \(DD^{-1} = I\)

\[
\leq \Sigma^{-1} - 2\Sigma^{-1}D\Sigma^{-1} + \Sigma^{-1}DD^{-1}D\Sigma^{-1}
\]

because

\[
D^{-1} \geq (D + R)^{-1} = \Sigma^{-1}
\]

by Result 4.5 because \(D > 0\) and \(R > 0\),

\[
= \Sigma^{-1} - 2\Sigma^{-1}D\Sigma^{-1} + \Sigma^{-1}D\Sigma^{-1}
\]

\[
\leq \Sigma^{-1},
\]

(4.18)

because \(\Sigma^{-1}D\Sigma^{-1}\) is positive definite. Combining this inequality for \(\tilde{D}\) with what was shown previously in (4.17),

\[
E\left(\sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i \mid \sigma^2, \Sigma, \mu', \nu'\right)
\]

154
\[
\leq \sum_{i=1}^{N} (Y' - \mu' - \tau'_i 1)^t \Sigma^{-1} (Y' - \mu' - \tau'_i 1) \\
+ \sum_{i=1}^{N} \sigma_i^2 \text{tr}(\Sigma^{-1}(D^{-1} + R^{-1})^{-1})
\]  
(4.19)

by (4.18)

\[
\leq \sum_{i=1}^{N} (Y' - \mu' - \tau'_i 1)^t \Sigma^{-1} (Y' - \mu' - \tau'_i 1) + NM \sigma^2,
\]  
(4.20)

which is true because \( \sigma_i^2 \leq \sigma^2 \) and

\[
\Sigma^{-1} = (D + R)^{-1} \leq D^{-1} \leq D^{-1} + R^{-1},
\]

by Result 4.5 (2) and \( R^{-1} > 0 \); then, by Result 4.5 (4),

\[
\text{tr}(\Sigma^{-1}(D^{-1} + R^{-1})^{-1}) \leq \text{tr}((D^{-1} + R^{-1})(D^{-1} + R^{-1})^{-1}) = \text{tr}(I) = M. \quad (4.21)
\]

Before continuing with the first part of (4.20), recall

\[
\Sigma|Y', \mu', \tau', \sigma^2 \sim \text{Inverse Wishart}_M(N + \delta, S' + \Sigma_0)
\]

where

\[
S' = \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^t q'_i.
\]

Then,

\[
\mathbb{E}\left( \sum_{i=1}^{N} (Y' - \mu' - \tau'_i 1)^t \Sigma^{-1} (Y' - \mu' - \tau'_i 1) \bigg| Y', \mu', \tau', \sigma^2 \right)
= (N + \delta) \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)^t \sigma_i^2 q'_i
\]
\[
\left( \sum_{j=1}^{N} q'_j (Y'_j - \mu' - \tau'_j 1)(Y'_j - \mu' - \tau'_j 1)^t + \Sigma_0 \right)^{-1} (Y' - \mu' - \tau'_i 1)
\]

by Result 4.20

\[
\leq (N + \delta) \sum_{i=1}^{N} \sigma_i^2
\]  
(4.22)
by Result 4.14 where
\[ \mathbf{x} = \sqrt{q_i}(\mathbf{Y}_j' - \mu' - \tau'_j \mathbf{1}) \]
and
\[ \mathbf{A} = \sum_{j \neq i} q'_j(\mathbf{Y}_j' - \mu' - \tau'_j \mathbf{1})(\mathbf{Y}_j' - \mu' - \tau'_j \mathbf{1})^T + \Sigma_0, \]
which is positive definite. Continuing with the inequality,
\[
E \left( \sum_{i=1}^{N} (\mathbf{Y}' - \mu' - \tau'_i \mathbf{1})^T \Sigma^{-1}(\mathbf{Y}' - \mu' - \tau'_i \mathbf{1}) \right) \leq (N + \delta) \sum_{i=1}^{N} \sigma_i^{2\nu} \]
from (4.22)
\[
\leq (N + \delta)Na_{\sigma^2}, \tag{4.23} \]
because \( \sigma_i^{2} \leq a_{\sigma^2} \).

Note that for Model 4.3, (4.14) and (4.15) still hold because the updates for \( \mu \) and \( \Sigma \) are identical in Model 4.2 and Model 4.3. Then,
\[
E(V(\theta)|\theta') = E \left( \sum_{i=1}^{N} q_i(\mathbf{Y}_i - \mu - \tau_i \mathbf{1})^T(\mathbf{Y}_i - \mu - \tau_i \mathbf{1}) + \sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i + \text{tr}(\Sigma) \right| \theta') \\
= E \left( \sum_{i=1}^{N} q_i(\mathbf{Y}_i - \mu - \tau_i \mathbf{1})^T(\mathbf{Y}_i - \mu - \tau_i \mathbf{1}) \right| \mathbf{Y}, \mathbf{Z}, \Sigma, \tau, \sigma) \\
+ \sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i + \text{tr}(\Sigma) \right| \theta') \\
\]
by Result 4.15,
\[
\leq E \left( \sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i + 2 \text{tr}(\Sigma) + NMK_1a_{\sigma^2} \right| \theta') \\
\leq E \left( \sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i \right| \mathbf{Y}', \mu', \Sigma, \tau', \sigma^2 \right) + 2 \text{tr}(\Sigma) + NMK_1a_{\sigma^2} \right| \theta') \]
by Result 4.15

\[
E \left( 2 \text{tr}(\Sigma) + NM K_1 a_{\sigma^2} + NM a_{\sigma^2} + (N + \delta) Na_{\sigma^2} \bigg| \theta' \right)
\]

by (4.20) and (4.23)

\[
\leq \frac{2}{N + \delta - M - 1} \sum_{i=1}^{N} q_i(Y' - \mu' - \tau'_i 1) (Y' - \mu' - \tau'_i 1)^T + \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + NM K_1 a_{\sigma^2} + NM a_{\sigma^2} + (N + \delta) Na_{\sigma^2}
\]

by (4.15)

\[
\leq \lambda V(\theta') + b,
\]

where

\[
\lambda = \frac{2}{N + \delta - M - 1}, \quad \text{and} \quad b = \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + NM K_1 a_{\sigma^2} + NM a_{\sigma^2} + (N + \delta) Na_{\sigma^2}.
\]

Therefore, combining unboundedness off compact sets and this drift condition, Method 2 shows that this Markov chain is geometrically ergodic when

\[N + \delta - M - 1 > 2,\]

which will typically hold.

4.5.4 Model 4.4: Unconstrained Scale-Usage Effects with Decomposition

The final model considered relaxes the constraint on \(\sigma_i^2\) that the parameter space must be bounded. Formally,

\[
Y_i \sim \text{MVN}_M(\mu + \tau_i 1 + Z_i, \sigma_i^2 D), \quad i = 1, \ldots, N
\]
\[
X_{ij} \mid \mathbf{c}, Y_{ij} = \{k : c_{k-1} < Y_{ij} \leq c_k\}, \quad i = 1, \ldots, N; j = 1, \ldots, M
\]

\[
\mathbf{Z}_i \sim \text{MVN}_M(\mathbf{0}, \sigma_i^2 \mathbf{R}), \quad i = 1, \ldots, N
\]

\[
\sigma_i^2 \sim \text{Inverse Gamma}(a/2, a/2 - 1), \quad i = 1, \ldots, N
\]

\[
\tau_i \sim \text{Truncated Normal}(0, \sigma_i^2, -a_\tau, a_\tau), \quad i = 1, \ldots, N
\]

\[
\Sigma \sim \text{Inverse Wishart}_M(\delta, \Sigma_0)
\]

\[
\pi(\mu) \propto 1
\]

\[
\pi(\mathbf{c}) \propto \prod_{k=1}^K (c_k - c_{k-1})^{\alpha_k - 1}, \quad \text{where } c_0 = -C \text{ and } c_K = C,
\]

where \(a, \sigma_i^2, a_\tau, \delta, \Sigma_0,\) and \(\alpha_k\) for \(k = 1, \ldots, K\) are all constants. The following algorithm results.

\textbf{Algorithm 4.6 (Decomposition Sampler)}

From the current \(\theta'\), update to \(\theta\) according to the following steps.

(i) \(\Sigma' | \mathbf{Y}', \mu', \sigma'^2 \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0)\) where

\[
S' = \sum_{i=1}^N (Y_i' - \mu' - \tau_i' \mathbf{1})(Y_i' - \mu' - \tau_i' \mathbf{1})^T q_i'.
\]

(ii) \(\sigma_i^2' | \mathbf{Y}_i', \mu', \tau_i', \Sigma \sim \text{Inverse Gamma}((a + M)/2, b_{\sigma_i^2})\) where

\[
b_{\sigma_i^2} = \frac{(Y_i' - \mu' - \tau_i' \mathbf{1})^T \Sigma^{-1} (Y_i' - \mu' - \tau_i' \mathbf{1}) + a - 2}{2}.
\]

(iii) \(\mathbf{Z}_i | \mathbf{Y}_i', \mu', \tau', \sigma^2, \Sigma \sim \text{MVN}_M(\mu_Z, \Sigma_Z)\) where

\[
\mu_Z = (D^{-1} + R^{-1})^{-1} D^{-1} (Y_i' - \mu' - \tau_i' \mathbf{1}), \quad \text{and}
\]

\[
\Sigma_Z = \sigma_i^2(D^{-1} + R^{-1})^{-1}.
\]

(iv) Update \(c_k | c_{k-1}, c'_{k+1}, \mu', \tau', \mathbf{Z}, \Sigma\) sequentially with a Metropolis-Hastings step utilizing some candidate distribution \(g(\cdot | c_k)\) to generate a candidate \(c_k^*\) which is
accepted with probability
\[ \alpha_k = \min \left\{ 1, \left( \frac{c_{k+1} - c_k^*}{c_{k+1} - c_k} \right)^{\alpha_k+1} \left( \frac{c_k^* - c_{k-1}}{c_k - c_{k-1}} \right)^{\alpha_k-1} \frac{w_k(c_k^*)g(c_k|c_k^*)}{w_k(c_k)g(c_k^*|c_k)} \right\}, \]

where
\[ w_k(c_k) = \prod_{\ell=k}^{k+1} \prod_{(i,j): X_{ij} = \ell} \left\{ \Phi \left( \frac{c_{\ell-1} - \mu_j' - \tau_i' - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu_j' - \tau_i' - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) \right\}. \]

(v) \( Y_{ij} | \mu', \tau_i', \Sigma, \sigma_i^2, c \sim \text{Truncated Normal}(\mu_j' + \tau_i' + Z_{ij}, \sigma_i^2 D_{jj}, c_{X_{ij}-1}, c_{X_{ij}}) \).

(vi) \( \tau_i | Y_i, \mu', \Sigma, \sigma_i^2 \sim \text{Truncated Normal}(\mu_{\tau_i}, s_{\tau_i}^2, -a_{\tau}, a_{\tau}) \), where
\[ s_{\tau_i}^2 = (q_i 1^T \Sigma^{-1} 1 + q_{\tau})^{-1} \]
and
\[ \mu_{\tau_i} = s_{\tau_i}^2 q_i (Y_i - \mu')^T \Sigma^{-1} 1. \]

(vii) \( \mu | Y, \tau, \Sigma, \sigma^2 \sim \text{MVNM}(\tilde{Y}, \Sigma/q^*), \) where
\[ q^* = \sum_{i=1}^N q_i, \quad \text{and} \quad \tilde{Y} = \frac{1}{q^*} \sum_{i=1}^N (Y_i - \tau_i 1)q_i. \]

The proof for geometric ergodicity for this sampler will proceed as the others in using Method 2 by showing that a drift condition and unboundedness off compact sets hold. The candidate drift function \( V(\theta) \) is
\[ V(\theta) = \sum_{i=1}^N q_i (Y_i - \mu - \tau_i 1)^T (Y_i - \mu - \tau_i 1) + \sum_{i=1}^N q_i Z_i^T \Sigma^{-1} Z_i + \sum_{i=1}^N \sigma_i^2 + \text{tr}(\Sigma). \]

First, a proof that \( V \) is unbounded off compact sets for the Markov chain. As typical, \( V \) is continuous, so it remains to show that \( \{ \theta : V(\theta) \leq d \} \) is bounded. Assume \( V(\theta) \leq d; \) then the following statements are true.

- The parameters \( Y_i, i = 1, \ldots, N, \) and \( c \) are bounded by the model itself, as previously observed.
• The parameters \(\tau_i, i = 1, \ldots, N,\) are bounded by the truncated prior distributions.

• The parameters \(\sigma_i^2\) are bounded directly by \(V(\theta) \leq d\) for \(i = 1, \ldots, N.\)

• As

\[
\sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T (Y_i - \mu - \tau_i 1) \leq d,
\]

each component of the sum is less than or equal to \(d\), which implies that

\[
(Y_i - \mu - \tau_i)^T (Y_i - \mu - \tau_i) \leq d \sigma_i^2 \leq d^2, \quad i = 1, \ldots, N.
\]

As \(Y_i\) and \(\tau_i\) are both bounded by the model (for any \(i\)), this implies that \(\mu\) is bounded.

• As

\[
q_i Z_i^T \Sigma^{-1} Z_i \leq d, \quad i = 1, \ldots, N,
\]

it is also true by Result 4.8 that

\[
(\lambda_1)^{-1} q_i Z_i^T Z_i \leq d, \quad i = 1, \ldots, N
\]

where \((\lambda_1)^{-1}\) is the smallest eigenvalue of \(\Sigma^{-1}\) (\(\lambda_1\) is the largest eigenvalue of \(\Sigma\)). This implies that

\[
Z_i^T Z_i \leq d \lambda_1 \sigma_i^2 \leq d \text{tr}(\Sigma) \sigma_i^2 \leq d^3, \quad i = 1, \ldots, N,
\]

and thus every element of \(Z_i\) is bounded for \(i = 1, \ldots, N.\)

• As \(\text{tr}(\Sigma)\) is bounded, every element of \(\Sigma\) is bounded according to Result 4.13.

Therefore, \(\theta\) is bounded if \(V(\theta) \leq d\) and \(V\) is unbounded off compact sets for the Markov chain.
To show that a drift condition is satisfied, begin with the first part of $V$. Note that

$$
E\left(\sum_{i=1}^{N} q_i (Y_i - \mu - \tau_i 1)^T (Y_i - \mu - \tau_i 1) \bigg| Y, \tau, Z, \Sigma, \sigma^2\right) 
\leq MK_1 \sum_{i=1}^{N} q_i + \text{tr}(\Sigma)
$$

(4.25)

from (4.13). Recall that

$$
\sigma^2_i | Y', \mu', \tau_i, \Sigma \sim \text{Inverse Gamma}((a + M)/2, b_{\sigma_i^2})
$$

where

$$
b_{\sigma_i^2} = \frac{(Y_i' - \mu' - \tau_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau_i 1) + a - 2}{2}.
$$

Then,

$$
E\left(\sum_{i=1}^{n} q_i \left| \Sigma, Y', \mu', \tau'\right\right) 
= \sum_{i=1}^{N} \frac{M + a}{(Y_i' - \mu' - \tau_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau_i 1) + a - 2}
$$

by Result 4.19, and

$$
\leq \frac{N(M + a)}{a - 2}
$$

(4.26)

because

$$
(Y_i' - \mu' - \tau_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau_i 1) \geq 0,
$$

and therefore

$$
\frac{M + a}{(Y_i' - \mu' - \tau_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau_i 1) + a - 2} \leq \frac{M + a}{a - 2}.
$$

Continuing,

$$
E\left(\sum_{i=1}^{N} q_i Z_i^T \Sigma^{-1} Z_i \bigg| \sigma^2, \Sigma, Y', \mu', \tau'\right) 
$$
\[
\leq \sum_{i=1}^{N} q_i (Y' - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y' - \mu' - \tau'_i 1) + N \text{tr}(\Sigma^{-1}(D^{-1} + R^{-1})^{-1})
\]

by (4.19) where \( q_i \) has been multiplied to each part of each sum

\[
\leq \sum_{i=1}^{N} q_i (Y' - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y' - \mu' - \tau'_i 1) + NM \quad (4.27)
\]

by (4.21). Further,

\[
E \left( \sum_{i=1}^{N} q_i (Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1) \right| \Sigma, Y', \mu', \tau') = (M + a) \sum_{i=1}^{N} \frac{(Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1)}{(Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1) + a - 2}
\]

by Result 4.19

\[
\leq N(M + a), \quad (4.28)
\]

because

\[
\frac{(Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1)}{(Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1) + a - 2} \leq 1,
\]

assuming \( a > 2 \), which will typically be satisfied. Additionally,

\[
E \left( \sum_{i=1}^{N} \sigma_i^2 \right| \Sigma, Y', \mu', \tau') = \frac{1}{a + M - 2} \left( \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1) + a - 2 \right), \quad (4.29)
\]

by Result 4.19, and

\[
E \left( \frac{1}{a + M - 2} \left( \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y'_i - \mu' - \tau'_i 1) + a - 2 \right) \right| Y', \mu', \tau', \sigma^2)
\]

\[
= \frac{N + \delta}{a + M - 2} \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)^T (S' + \Sigma_0)^{-1} (Y'_i - \mu' - \tau'_i 1) + \frac{N(a - 2)}{a + M - 2}
\]

by Result 4.20

\[
\leq \frac{N + \delta}{a + M - 2} \sum_{i=1}^{N} \sigma_i^2 + \frac{N(a - 2)}{a + M - 2} \quad (4.30)
\]
which is due to Result 4.14 as in (4.22). Then, with these results it is observed that

\[
E(V(\theta)|\theta') = E \left( \sum_{i=1}^{N} q_i(y_i - \mu - \tau_i 1)^T (y_i - \mu - \tau_i 1) + \sum_{i=1}^{N} q_i z_i^T \Sigma^{-1} z_i \right.
\]

\[
+ \sum_{i=1}^{N} \sigma_i^2 + \text{tr}(\Sigma) \left| \theta' \right|
\]

\[
= E \left( \sum_{i=1}^{N} q_i(y_i - \mu - \tau_i 1)^T (y_i - \mu - \tau_i 1) \left| y, z, \Sigma, \tau, \sigma^2 \right. \right)
\]

\[
+ \sum_{i=1}^{N} q_i z_i^T \Sigma^{-1} z_i + \sum_{i=1}^{N} \sigma_i^2 + \text{tr}(\Sigma) \left| \theta' \right|
\]

by Result 4.15

\[
\leq E \left( \sum_{i=1}^{N} q_i z_i^T \Sigma^{-1} z_i + \sum_{i=1}^{N} \sigma_i^2 + 2 \text{tr}(\Sigma) + MK_1 \sum_{i=1}^{n} q_i \left| \theta' \right) \right.
\]

by (4.25)

\[
= E \left( \sum_{i=1}^{N} q_i z_i^T \Sigma^{-1} z_i \left| y', \mu', \tau', \sigma^2, \Sigma \right. \right)
\]

\[
+ \sum_{i=1}^{N} \sigma_i^2 + 2 \text{tr}(\Sigma) + MK_1 \sum_{i=1}^{n} q_i \left| \theta' \right)
\]

by Result 4.15

\[
\leq E \left( \sum_{i=1}^{N} q_i(y' - \mu' - \tau'_i 1)^T \Sigma^{-1} (y' - \mu' - \tau'_i 1) \right.
\]

\[
+ \sum_{i=1}^{N} \sigma_i^2 + 2 \text{tr}(\Sigma) + MK_1 \sum_{i=1}^{n} q_i + NM \left| \theta' \right)
\]

by (4.27)

\[
= E \left( \sum_{i=1}^{N} q_i(y' - \mu' - \tau'_i 1)^T \Sigma^{-1} (y' - \mu' - \tau'_i 1) \right.
\]

\[
+ \sum_{i=1}^{N} \sigma_i^2 + MK_1 \sum_{i=1}^{n} q_i \left| \Sigma, \theta' \right) + 2 \text{tr}(\Sigma) + NM \left| \theta' \right)
\]

163
by Result 4.15

\[
\begin{align*}
\mathbf{\text{by (4.26)}} \\
E \left( \sum_{i=1}^{N} q_i \mathbf{y}_i' \right) & = \mathbf{b,} \\
\mathbf{\text{where}} \\
\lambda & = \max \left( \frac{2}{N + \delta - M - 1}, \frac{N + \delta}{a + M - 2} \right)
\end{align*}
\]
and
\[ b = \frac{2 \text{tr}(\Sigma_0)}{N + \delta - M - 1} + \frac{N M K_1 (M + a)}{a - 2} + N M + N (M + a) + \frac{N (a - 2)}{a + M - 2}. \]

Thus, when \( \lambda < 1 \), a drift condition (2.9) is satisfied for the Markov chain induced by Algorithm 4.6, and coupled with the unboundedness off compact sets which was already established, the chain is geometrically ergodic.

Regarding the conditions under which \( \lambda < 1 \), note that
\[ N + \delta - M - 1 > 2 \]
typically with no challenge. However, the constraint that
\[ a + M - 2 > N + \delta \]
is much more restrictive, forcing
\[ a > N + \delta - M + 2. \]

Recall that the prior variance for \( \sigma_i^2 \) is
\[ \text{Var}_{\pi(\sigma_i^2)}(\sigma_i^2) = \frac{\left( \frac{a}{2} \right)^2}{\left( \frac{a}{2} - 1 \right)^2 \left( \frac{a}{2} - 2 \right)}, \]
by Result 4.19, which is \( O(a^{-1}) \). Thus, for this particular drift condition to be satisfied, for even moderately sized \( N \), given fixed \( M \) and \( \delta \), the prior for \( \sigma_i^2 \) is constrained to have very small variance, which decreases at rate \( N^{-1} \). This somewhat limits the ability of the model to account for varying scales among the respondents, which negates a key purpose of the scale-usage model. While this model could still be useful for small samples or a correspondingly small value of \( C \) to scale the problem down, Model 4.3 will typically be preferred. Bounds on \( \sigma_i^2 \), which may be set arbitrarily,
will typically not constrain Model 4.3 substantially. Note that this discussion has assumed the use of these drift functions established in this chapter. Other drift functions would not necessarily impose the same constraint in order to ensure geometric ergodicity, and alternatives could be sought if desired.

4.5.5 Summary of Models and Extensions

This section has established the geometric ergodicity for four Bayesian scale-usage models. A summary of these models, including the drift function $V(\theta)$ and the drift constants $\lambda$ and $b$, is provided in Table 4.2. One note is that as the model becomes more complicated, so do $V(\theta)$ and $b$.

The following recommendations are made assuming the drift functions provided in this chapter are used to ensure geometric ergodicity. Other drift functions would result in different guidance. With this understanding, it is suggested, in the presence of scale-usage heterogeneity, that Model 4.3 is likely the best choice among these four models. Use of Model 4.1 is precluded because it lacks the scale-usage framework. Model 4.2 suffers from the computational disadvantages inherent in not using the decomposition sampler, as discussed previously in this chapter. Model 4.4 requires a highly constrained prior distribution for moderately-sized $N$ on the $\sigma_i^2$ parameters in order to ensure $\lambda < 1$. Thus, Model 4.3 is likely the best option, given these drift functions, for flexible modeling of scale-usage data with Markov chains which are geometrically ergodic.

While these models can successfully be used in a wide array of scale-usage applications, additional model flexibility may be desired.
Table 4.2: Summary of Geometrically Ergodic Scale-Usage Models

<table>
<thead>
<tr>
<th>Model 4.1</th>
<th>No scale-usage</th>
<th>$V(\theta)$</th>
<th>$\sum_{i=1}^{N} (Y_i - \mu)^T(Y_i - \mu) + \text{tr}(\Sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\frac{2}{N+\delta - M - 1}$</td>
<td>$b$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 4.2</th>
<th>Bounded $\sigma_i^2$</th>
<th>$V(\theta)$</th>
<th>$\sum_{i=1}^{N} q_i(Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) + \text{tr}(\Sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\frac{2}{N+\delta - M - 1}$</td>
<td>$b$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 4.3</th>
<th>Decomposition, bounded $\sigma_i^2$</th>
<th>$V(\theta)$</th>
<th>$\sum_{i=1}^{N} q_i(Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) + \sum_{i=1}^{N} Z_i^T \Sigma^{-1} Z_i + \text{tr}(\Sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\frac{2}{N+\delta - M - 1}$</td>
<td>$b$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model 4.4</th>
<th>Decomposition, unbounded $\sigma_i^2$</th>
<th>$V(\theta)$</th>
<th>$\sum_{i=1}^{N} q_i(Y_i - \mu - \tau_i 1)^T(Y_i - \mu - \tau_i 1) + \sum_{i=1}^{N} q_i Z_i^T \Sigma^{-1} Z_i + \sum_{i=1}^{N} \sigma_i^2 + \text{tr}(\Sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\lambda$</td>
<td>$\max\left(\frac{2}{N+\delta - M - 1}, \frac{N + \delta}{a + M - 2}\right)$</td>
<td>$b$</td>
</tr>
</tbody>
</table>
Additional Models A number of simpler models are geometrically ergodic by virtue of the proofs of geometric ergodicity for Models 4.1–4.4. For example, if \( \tau_i \) is defined to be 0 in a scenario where no location adjustment is made for each of the respondents, geometric ergodicity still holds for the models presented here. Similarly, if \( \sigma_i^2 \) is defined to be 1, geometric ergodicity still holds using Model 4.2 or Model 4.3 (i.e. \( a_{\sigma^2} = 1 \)).

Note that the complete conditional distribution for \( \tau_i \) was never used in any of the geometric ergodicity proofs. The only requirement is that \( \tau_i \) is bounded above and below by some constant. A uniform prior could easily be substituted. More complicated effects governing the distribution of \( \tau \) could also be added, provided the new parameters do not effect any of the parameters whose complete conditional distributions were utilized in the proofs of geometric ergodicity. It is also necessary that any new parameters are either bounded or that terms may be added to the drift function which maintain unboundedness off compact sets yet preserve the drift condition. This is formalized in a proposition.

Proposition 4.21. Geometric Ergodicity of Other Scale-Usage Models. Assume one of the Models 4.1–4.4 for which algorithms exist which are geometrically ergodic. Let a modification to the model be made such that

(i) none of the complete conditional distributions used in establishing the drift conditions are changed,

(ii) the order of updating the new parameters is considered so as to not invalidate the previously established proofs, and
(iii) either (a) the new parameters are bounded or (b) additional terms may be added to the drift function such that the drift condition is still established while maintaining unboundedness off compact sets.

Then, the corresponding algorithm is geometrically ergodic.

Proof. Conditions (i), (ii), and (iii) were chosen so that a drift condition holds which is unbounded off compact sets for the Markov chain. Thus, the chain is geometrically ergodic.

This proposition will be used in Chapter 5 to develop a demographic-specific model for the scale-usage effect:

$$\tau_i \sim \text{Normal}(\theta_{d_i}, \sigma_{\tau, d_i}^2), \quad i = 1, \ldots, N,$$

where $d_i$ is the demographic index of respondent $i$. Prior distributions are provided for $\theta_{\ell}$ and $\sigma_{\tau, \ell}^2$ for each demographic $\ell$. It’s clear this adjustment does not change the updates of $\mu, \Sigma, Z$, or $\sigma^2$, all of which were utilized for one model or another in proving geometric ergodicity. Choosing bounded priors for $\theta_{\ell}$ and $\sigma_{\tau, \ell}^2$ then ensures geometric ergodicity of the resulting algorithm.

The other parameters beyond $\tau$ whose distributions were not required in the proof of geometric ergodicity are $c$, $Y$, and $\sigma_i^2$ (for all but the final model). The update mechanisms or prior distributions for these parameters could then be adjusted without impacting geometric ergodicity as long as the basic assumptions are fulfilled (e.g. bounded support for $\sigma_i^2$ and $Y_i$).

As a final matter of discussion, it is unsurprising that constraints must be made on the parameters of this model in order to obtain a geometrically ergodic Markov

169
chain with the specific drift conditions presented in this chapter. Considering the latent distribution of $Y_i$, there is nonidentifiability among the parameters. For example, multiply each $\sigma_i^2$ by a constant and divide $\Sigma$ by the same constant. Similarly, subtract a constant from $\mu$ and add the same constant to $\tau$. The same latent variable density results, though the parameters differ. While from a Bayesian perspective the model is identifiable, this type of nonidentifiability can make the establishment of convergence rates more challenging. Nonidentifiable parameters will often take large tours about the parameter space which are due to the nonidentifiability, rather than the exploration of the actual distribution of interest (see e.g. Figure 5.1). This lack of convergence can be paralyzing for inference when a Markov chain has a finite number of steps. The parameter constraints force the parameters to stay within a prescribed area, allowing for improved exploration of the parameter space according to the target distribution, and thus, improved convergence.

Note also that the truncation constraints imposed are entirely flexible in that bounds may be freely chosen to be as large as desired. Indeed, a way to trivially create a geometrically ergodic algorithm is to truncate every parameter, resulting in uniform ergodicity. In this model, many of the parameters were still unconstrained including $\mu$, $\Sigma$, and the decomposition latent variables $Z$.

Therefore, despite the model complexity, it is possible to show that variants of the Bayesian scale-usage model are geometrically ergodic. This fact may then be used to establish an ergodic average central limit theorem and estimate the asymptotic variance so confidence intervals can be constructed for inference as discussed in Chapter 2. This will be illustrated in Chapter 5. First, however, a minorization condition for scale-usage models is developed.
4.6 Minorization Condition for Scale-Usage Models

While proving geometric ergodicity by unboundedness of compact sets induces a minorization condition for any appropriately large choice of \( d \) (i.e. \( d \geq 2b/(1 - \lambda) \)) with the small set \( C = \{ \theta : V(\theta) \leq d \} \) (Meyn and Tweedie 1993, Theorem 15.0.1), working through the formal establishment of a minorization condition is often useful. For example, minorization is required to apply the approximate sampling methods in Chapter 3 and to compute the number of iterations required for the total variation distance to be below a certain threshold, as in Chapter 2. The hope is that a “good” minorization condition may be obtained, for example, where Markov chain regenerations happen frequently. Occasionally, however, a minorization condition may prove too inefficient for practical use. In this case, additional minorization conditions could be sought, or other methods for answering Question 1 and Question 2, such as estimation of Markov chain Monte Carlo errors, could be used. This section establishes a minorization condition for scale-usage models. The quality of the minorization, however, will largely be problem specific, depending on the size of the dataset, the data values themselves, and the prior distributions.

Recall that for a general minorization condition to be satisfied, the transition kernel \( K \) must satisfy

\[
K(x, A) \geq s(x)Q(A), \quad \forall x \in X, A \in \mathcal{B}(X).
\]  

(4.32)

As all the transition kernel updates have densities, it is sufficient to show

\[
k(x, y) \geq s(x)q(y), \quad \forall x \in X, y \in X,
\]

for some density \( q \). As in Chapter 3, an unnormalized version of \( q \) suffices.
The scale-usage model under consideration is that presented by
Hans et al. (2012) listed in (4.3) with the decomposition modification in
(4.6). The algorithm used is found in Algorithm 4.2 with small but
important modifications, which are provided in Algorithm 4.7.

Algorithm 4.7 (Scale-Usage for Minorization)

Referencing Algorithm 4.2, perform the following updates:

(i) Update, $\sigma^2$, $\Sigma$, $\mu$, and $\tau$ sequentially according to the complete conditional
   distributions in Algorithm 4.1.

(ii) Decompose $\Sigma$ and update $Z$ as in Algorithm 4.2.

(iii) Update $c|\mu, \tau, Z, \Sigma$ using a single independence Metropolis-Hastings step. Denote
     the candidate distribution with $g(\cdot)$. A reasonable candidate distribution
     can be obtained from initial posterior simulations using Algorithm 4.2. The
     acceptance probability is then

     $$\alpha_k = \min \left\{ 1, \prod_{k=2}^{K-1} \left( \frac{c^*_k - c^*_i}{c_k - c_{k-1}} \right)^{\alpha_k-1} \frac{w(c^*)g(c)}{w(c)g(c^*)} \right\},$$

     where

     $$w(c) = \prod_{i=1}^N \prod_{j=1}^M \left\{ \Phi \left( \frac{c_{X_{ij}} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma^2_i D_{jj}}} \right) - \Phi \left( \frac{c_{X_{ij}^{-1}} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma^2_i D_{jj}}} \right) \right\}.$$ 

(iv) Update $Y_{ij}|\mu, \tau, \Sigma, \sigma^2_i, c \sim$ Truncated Normal($\mu_j + \tau_i + Z_{ij}, \sigma^2_i D_{jj}, c_{X_{i,j}^{-1}}, c_{X_{i,j}}$).

The approach to show (4.32) and thereby establish a minorization condition
will be to consider each transition density alone, combining the results at the end for the
overall minorization condition.
Update for $\sigma_i^2$. The first update in Algorithm 4.7 is $\sigma_i^2$, which conditionally is distributed

$$
\sigma_i^2|\theta' \sim \text{Inverse Gamma}(a_1, b_{1i})
$$

where

$$
a_1 = \frac{a + M}{2}
$$

and

$$
b_{1i} = \frac{1}{2} \left[ (Y'_i - \mu' - \tau'_i 1)^T (\Sigma')^{-1} (Y'_i - \mu' - \tau'_i 1) + a - 2 \right].
$$

The transition density is then

$$
k_{1i}(\sigma_i^2|Y', \mu', \Sigma', \tau') \propto b_{1i}^{a_1}(\sigma_i^2)^{-a_1-1} \exp \left( -\frac{b_{1i}}{\sigma_i^2} \right).
$$

For this density, take the approach used in Chapter 3. Choose points $\bar{Y}$, $\bar{\mu}$, $\bar{\Sigma}$, and $\bar{\tau}$ with an interval $D_{1i} = (l_{1i}, u_{1i})$. Then,

$$
k_{1i}(\sigma_i^2|Y', \mu', \Sigma', \tau') \geq \inf_{\sigma_i^2 \in D_{1i}} \left\{ \frac{k_{1i}(\sigma_i^2|Y', \mu', \Sigma', \tau')}{k_{1i}(\sigma_i^2|\bar{Y}, \bar{\mu}, \bar{\Sigma}, \bar{\tau})} \right\} k_{1i}(\sigma_i^2|\bar{Y}, \bar{\mu}, \bar{\Sigma}, \bar{\tau}) \mathbb{I}_{D_{1i}}(\sigma_i^2)
$$

imposing the constraints on $\sigma_i^2$ and taking the infimum over the constraints

$$
= \inf_{\sigma_i^2 \in D_{1i}} \left\{ \frac{b_{1i}^{a_1}(\sigma_i^2)^{a_1-1} \exp \left( -\frac{b_{1i}}{\sigma_i^2} \right)}{\tilde{b}_{1i}^{a_1}(\sigma_i^2)^{a_1-1} \exp \left( -\frac{\tilde{b}_{1i}}{\sigma_i^2} \right)} \right\} k_{1i}(\sigma_i^2|\bar{Y}, \bar{\mu}, \bar{\Sigma}, \bar{\tau}) \mathbb{I}_{D_{1i}}(\sigma_i^2)
$$

substituting the transition density

$$
= \inf_{\sigma_i^2 \in D_{1i}} \left\{ \left( \frac{b_{1i}}{\tilde{b}_{1i}} \right)^{a_1} \exp \left( -\frac{1}{\sigma_i^2} (b_{1i} - \tilde{b}_{1i}) \right) \right\} k_{1i}(\sigma_i^2|\bar{Y}, \bar{\mu}, \bar{\Sigma}, \bar{\tau}) \mathbb{I}_{D_{1i}}(\sigma_i^2)
$$

by simplification

$$
= \left( \frac{b_{1i}}{\tilde{b}_{1i}} \right)^{a_1} \exp \left( -\frac{b_{1i} - \tilde{b}_{1i}}{d_1} \right) k_{1i}(\sigma_i^2|\bar{Y}, \bar{\mu}, \bar{\Sigma}, \bar{\tau}) \mathbb{I}_{D_{1i}}(\sigma_i^2)
$$
which is true where

\[
d_{1i} = \begin{cases} 
  l_{1i} & \text{if } b_{1i} > \tilde{b}_{1i} \\
  u_{1i} & \text{if } b_{1i} \leq \tilde{b}_{1i}; 
\end{cases}
\]

which continuing is

\[
= s_{1i}(Y', \mu', \Sigma', \tau')q_{1i}(\sigma_i^2)
\]

where

\[
s_{1i}(Y', \mu', \Sigma', \tau') = \left( \frac{b_{1i}}{b_{1i}} \right)^{a_1} \exp \left( -\frac{b_{1i} - \tilde{b}_{1i}}{d_{1i}} \right),
\]

and

\[
q_{1i}(\sigma_i^2) = k_{1i}(\sigma_i^2|\tilde{Y}, \tilde{\mu}, \tilde{\Sigma}, \tilde{\tau})\mathbb{I}_{D_{1i}}(\sigma_i^2).
\]

**Update for \( \Sigma \).** Next, consider the update for \( \Sigma \), which conditionally is

\[
\text{Inverse } \text{Wishart}_M(N + \delta, S_N + \Sigma_0),
\]

where

\[
S_N = \sum_{i=1}^{N} (Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^T \sigma_i^{-2}.
\]

Thus, the kernel is proportional to

\[
k_2(\Sigma|Y', \mu', \tau', \sigma^2) \propto |S_n + \Sigma_0|^{(N+\delta)/2}|\Sigma|^{-(N+\delta+M+1)/2}\exp \left( -\frac{1}{2} \text{tr}((S_n + \Sigma_0)\Sigma^{-1}) \right).
\]

Begin first by considering \(|S_n + \Sigma_0|^{(N+\delta)/2}\). The challenge is to separate \( \sigma^2 \) from the rest of the parameters as it has already been updated. Note that

\[
|S_n + \Sigma_0| \geq |S_n| + |\Sigma_0|
\]

by Result 4.10

\[
\geq \sum_{i=1}^{N} |\sigma_i^{-2}(Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^T| + |\Sigma_0|
\]
again by Result 4.10
\[
= \sum_{i=1}^{N} \sigma_i^{-2M}|(Y'_i - \mu' - \tau'_i1)(Y'_i - \mu' - \tau'_i1)^T| + |\Sigma_0|
\]
by Result 4.9
\[
\geq \mathbb{I}_{D_1}(\sigma^2_i) \sum_{i=1}^{N} u_{1i}^{-2M}|(Y'_i - \mu' - \tau'_i1)(Y'_i - \mu' - \tau'_i1)^T| + |\Sigma_0|,
\]
which is true because, by restricting \( \sigma^2_i \) to fall in \( D_{1,i} \) and the fact that all the determinants are positive, the minimum occurs when \( \sigma^2_i \) is largest at \( u_{1i} \). This implies that
\[
|S_n + \Sigma_0^{(N+\delta)/2}
\geq \mathbb{I}_{D_1}(\sigma^2_i) \left( \sum_{i=1}^{N} u_{1i}^{-2M}|(Y'_i - \mu' - \tau'_i1)(Y'_i - \mu' - \tau'_i1)^T| + |\Sigma_0| \right)^{(N+\delta)/2},
\]
which, aside from the indicator function, is free of \( \sigma^2_i \), as desired.

Next, define \( D_2 = \{ \Sigma : \lambda_2 \leq \lambda_M \leq \cdots \leq \lambda_1 \leq u_2 \} \), where \( \lambda_1 \) and \( \lambda_M \) are the largest and smallest eigenvalues of \( \Sigma \) respectively. Assume that \( \Sigma \in D_2 \) and \( \sigma^2_i \in D_{1,i} \) for all \( i = 1, \ldots, N \). Then,
\[
\text{tr}(S_n \Sigma^{-1}) = \text{tr} \left( \sum_{i=1}^{N} \sigma_i^{-2}(Y'_i - \mu' - \tau'_i1)(Y'_i - \mu' - \tau'_i1)^T \Sigma^{-1} \right)
\geq \sum_{i=1}^{N} \sigma_i^{-2}(Y'_i - \mu' - \tau'_i1)^T \Sigma^{-1}(Y'_i - \mu' - \tau'_i1)
\leq \sum_{i=1}^{N} l_{ii}^{-1}(Y'_i - \mu' - \tau'_i1)^T \Sigma^{-1}(Y'_i - \mu' - \tau'_i1),
\]
which is true because
\[
(Y'_i - \mu' - \tau'_i1)^T \Sigma^{-1}(Y'_i - \mu' - \tau'_i1) > 0,
\]
so \( l_{ii} \) replaces \( \sigma^2_i \) to maximize the quantity subject to the constraint that \( \sigma^2_i \in D_{1,i} \)
\[
\leq \sum_{i=1}^{N} \lambda_i^{-1} l_{ii}^{-1}(Y'_i - \mu' - \tau'_i1)^T (Y'_i - \mu' - \tau'_i1),
\]
175
by Result 4.8 where $\lambda^{-1}_M$ is the largest eigenvalue of $\Sigma^{-1}$

$$
\leq \sum_{i=1}^{N} l^{-1}_{2i} l^{-1}_{i} (Y'_i - \mu' - \tau'_i 1)^T(Y'_i - \mu' - \tau'_i 1),
$$

(4.34)
because $\lambda^{-1}_M \leq \ell^{-1}_2$ when $\Sigma \in D_2$. This expression is free of both $\Sigma$ and $\sigma^2_i$, beyond the assumption that $\Sigma \in D_2$ and $\sigma^2_i \in D_{1,i}$.

Altogether, (4.33) and (4.34) imply that

$$
k_2(\Sigma|Y', \mu', \tau', \sigma^2) \propto |S_n + \Sigma_0|^{(N+\delta)/2} |\Sigma|^{-(N+\delta+M+1)/2} \exp\left(-\frac{1}{2} \text{tr}((S_n + \Sigma_0)\Sigma^{-1})\right)
\geq \mathbb{1}_{D_2}(\Sigma) \prod_{i=1}^{N} \mathbb{1}_{D_{ii}}(\sigma_i^2) |\Sigma|^{-(N+\delta+M+1)/2} \exp\left(-\frac{1}{2} \text{tr}(\Sigma_0\Sigma^{-1})\right)
\times \left(\sum_{i=1}^{N} u^{-2M}_{1i} |(Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^T| + |\Sigma_0|\right)^{(N+\delta)/2}
\times \exp\left(-\frac{1}{2} \sum_{i=1}^{N} l^{-1}_{2i} l^{-1}_{i} (Y'_i - \mu' - \tau'_i 1)^T(Y'_i - \mu' - \tau'_i 1)\right)
$$

since

$$
\text{tr}((S_n + \Sigma_0)\Sigma^{-1}) = \text{tr}(S_n\Sigma^{-1}) + \text{tr}(\Sigma_0\Sigma^{-1})
$$

$$
= s_2(Y', \mu', \tau')q_2(\Sigma, \sigma^2)
$$

where

$$
s_2(Y', \mu', \tau') = \left(\sum_{i=1}^{N} u^{-2M}_{1i} |(Y'_i - \mu' - \tau'_i 1)(Y'_i - \mu' - \tau'_i 1)^T| + |\Sigma_0|\right)^{(N+\delta)/2}
\times \exp\left(-\frac{1}{2} \sum_{i=1}^{N} l^{-1}_{2i} l^{-1}_{i} (Y'_i - \mu' - \tau'_i 1)^T(Y'_i - \mu' - \tau'_i 1)\right)
$$

and

$$
q_2(\Sigma, \sigma^2) = \mathbb{1}_{D_2}(\Sigma) \prod_{i=1}^{N} \mathbb{1}_{D_{ii}}(\sigma_i^2) |\Sigma|^{-(N+\delta+M+1)/2} \exp\left(-\frac{1}{2} \text{tr}(\Sigma_0\Sigma^{-1})\right).
$$
Update for $\mu$. Now, consider the update for $\mu$, which is

$$\text{MVN}_M(\hat{Y}, \Sigma/q^*)$$

where

$$q^* = \sum_{i=1}^{N} q_i,$$
and

$$\hat{Y} = \frac{1}{q} \sum_{i=1}^{N} (Y_i' - \tau_i'1)q_i.$$

Define the set $D_{3i} = \{Y_i, \tau_i : |Y_i - \tau_i1| \leq d_{3i}\}$. Then,

$$k_3(\mu|Y', \tau', \Sigma, \sigma^2) \propto q^{M/2}|\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (\mu - \hat{Y})^T \Sigma^{-1} q^*(\mu - \hat{Y}) \right)$$

$$\geq q^{M/2}|\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (\mu - \hat{Y})^T \lambda_M^{-1} q^*(\mu - \hat{Y}) \right)$$

by Result 4.8 where $\lambda_M^{-1}$ is the largest eigenvalue of $\Sigma^{-1}$

$$\geq 1_{D_2}(\Sigma)q^{M/2}|\Sigma|^{-1/2} \exp \left( -\frac{1}{2} (\mu - \hat{Y})^T l_2^{-1} q^* \mu \right)$$

$$\times \exp \left( -\frac{1}{2} \hat{Y}^T l_2^{-1} q^* \hat{Y} \right) \exp \left( \mu^T l_2^{-1} q^* \hat{Y} \right)$$

by multiplying out the cross product

$$= 1_{D_2}(\Sigma)q^{M/2}|\Sigma|^{-1/2} \exp \left( -\frac{1}{2} \mu^T l_2^{-1} q^* \mu \right)$$

$$\times \exp \left( -\frac{1}{2} \left( \frac{1}{q^*} \sum_{i=1}^{N} (Y_i' - \tau_i'1)q_i \right)^T l_2^{-1} q^* \left( \frac{1}{q^*} \sum_{i=1}^{N} (Y_i' - \tau_i'1)q_i \right) \right)$$

$$\times \exp \left( \mu^T l_2^{-1} q^* \left( \frac{1}{q^*} \sum_{i=1}^{N} (Y_i' - \tau_i'1)q_i \right) \right)$$

by substituting $\hat{Y}$

$$\geq 1_{D_2}(\Sigma)q^{M/2}|\Sigma|^{-1/2} \left[ \prod_{i=1}^{n} 1_{D_{3i}}(Y_i', \tau_i') \right] \exp \left( -\frac{1}{2} \mu^T l_2^{-1} q^* \mu \right)$$

177
\begin{align*}
\times \exp \left( -\frac{1}{2} \left( \frac{1}{q^*} \sum_{i=1}^{N} d_{3i} q_i \right)^T l_2^{-1} q^* \left( \frac{1}{q^*} \sum_{i=1}^{N} d_{3i} q_i \right) \right) \\
\times \exp \left( -|\mu|^T l_2^{-1} q^* \left( \frac{1}{q^*} \sum_{i=1}^{N} d_{3i} q_i \right) \right)
\end{align*}

by making the sums as large as possible subject to the constraint that \((Y_i, \tau_i) \in D_{3i}\)

\begin{align*}
= \mathbb{1}_{D_2}(\Sigma)q^* M/2|\Sigma|^{-1/2} \left[ \prod_{i=1}^{n} 1_{D_{3i}}(Y'_i, \tau'_i) \right] \exp \left( -\frac{1}{2} \mu^T l_2^{-1} q^* \mu \right) \\
\times \exp \left( -\frac{1}{2} \widehat{d}_3^T l_2^{-1} q^* \widehat{d}_3 \right) \exp \left( -|\mu|^T l_2^{-1} q^* \widehat{d}_3 \right)
\end{align*}

defining

\[ \widehat{d}_3 = \left( \frac{1}{q^*} \sum_{i=1}^{N} d_{3i} q_i \right) \]

which is then

\[ = s_3(Y', \tau') q_3(\mu, \Sigma, \sigma^2) \]

where

\[ s_3(Y', \tau') = \prod_{i=1}^{n} 1_{D_{3i}}(Y'_i, \tau'_i), \]

and

\[ q_3(\mu, \Sigma, \sigma^2) = \mathbb{1}_{D_2}(\Sigma)q^* M/2|\Sigma|^{-1/2} \exp \left( -\frac{1}{2} \mu^T l_2^{-1} q^* \mu \right) \\
\times \exp \left( -\frac{1}{2} \widehat{d}_3^T l_2^{-1} q^* \widehat{d}_3 \right) \exp \left( -|\mu|^T l_2^{-1} q^* \widehat{d}_3 \right). \]

**Update for \(\tau_i\).** The update for \(\tau_i\) is Normal(\(\mu_{\tau_i}, s_{\tau_i}^2\)) where

\[ s_{\tau_i}^2 = (\sigma_i^{-2} 1^T \Sigma^{-1} 1 + \sigma_{\tau}^{-2})^{-1}, \]

and

\[ \mu_{\tau_i} = \frac{s_{\tau_i}^2}{\sigma_i^{-2}} (Y'_i - \mu)^T \Sigma^{-1} 1. \]

Define \(D_{4i} = \{ Y_i : |Y_i| \leq d_{4i} \}\). The transition kernel is then

\[ k_{4i}(\tau_i|\mu, \Sigma, \sigma, Y') \propto (s_{\tau_i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau_i}^2}(\tau_i - \mu_{\tau_i})^2 \right) \]

\[ = (s_{\tau_i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau_i}^2} \left( \tau_i + \frac{s_{\tau_i}^2}{\sigma_i^{-2}} \mu^T \Sigma^{-1} 1 - \frac{s_{\tau_i}^2}{\sigma_i^{-2}} (Y'_i)^T \Sigma^{-1} 1 \right)^2 \right) \]

178
by substituting \( \mu_{ri} \)

\[
= (s_{\tau i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau i}^2} \left( \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

\[
\times \exp \left( -\frac{1}{s_{\tau i}} \left( \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right) \left( s_{\tau i}^2 \sigma_i^{-2} \left( Y_i^\prime \right)^T \Sigma^{-1} \mathbf{1} \right) \right)
\]

by multiplying out the cross product

\[
\geq (s_{\tau i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau i}^2} \left( \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

\[
\times \exp \left( -\frac{1}{s_{\tau i}} \left| \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right| \left( s_{\tau i}^2 \sigma_i^{-2} \left| Y_i^\prime \right| \Sigma^{-1} \mathbf{1} \right) \right)
\]

\[
\times \exp \left( -\frac{1}{2s_{\tau i}^2} \left( s_{\tau i}^2 \sigma_i^{-2} \left| Y_i^\prime \right| \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

which is true because all the absolute values result in larger quantities

\[
\geq \mathbb{1}_{D_{4i}} (Y_i^\prime) (s_{\tau i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau i}^2} \left( \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

\[
\times \exp \left( -\frac{1}{s_{\tau i}} \left| \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right| \left( s_{\tau i}^2 \sigma_i^{-2} \mathbf{d}_{4i}^T \Sigma^{-1} \mathbf{1} \right) \right)
\]

\[
\times \exp \left( -\frac{1}{2s_{\tau i}^2} \left( s_{\tau i}^2 \sigma_i^{-2} \mathbf{d}_{4i}^T \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

by applying the constraints

\[
= s_{4i}(Y_i^\prime) \rho_{4i}(\mu, \Sigma, \sigma^2)
\]

where

\[
s_{4i}(Y_i^\prime) = \mathbb{1}_{D_{4i}} (Y_i^\prime),
\]

and

\[
\rho_{4i}(\mu, \Sigma, \tau_i, \sigma_i^2) = (s_{\tau i}^2)^{-1/2} \exp \left( -\frac{1}{2s_{\tau i}^2} \left( \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right)^2 \right)
\]

\[
\times \exp \left( -\frac{1}{2s_{\tau i}} \left| \tau_i + s_{\tau i}^2 \sigma_i^{-2} \mu^T \Sigma^{-1} \mathbf{1} \right| \left( s_{\tau i}^2 \sigma_i^{-2} \mathbf{d}_{4i}^T \Sigma^{-1} \mathbf{1} \right) \right)
\]

\[
\times \exp \left( -\frac{1}{2s_{\tau i}^2} \left( s_{\tau i}^2 \sigma_i^{-2} \mathbf{d}_{4i}^T \Sigma^{-1} \mathbf{1} \right)^2 \right).
\]
Update for $Z_i$. The update for $Z_i$ is $\text{MVN}_M(\mu_{Z_i}, \Sigma_{Z_i})$ where

$$
\mu_{Z_i} = (D^{-1} + R^{-1})^{-1}D^{-1}(Y_i' - \mu - \tau_i1), \quad \text{and}
$$

$$
\Sigma_{Z_i} = \sigma_i^2(D^{-1} + R^{-1})^{-1}.
$$

The transition density for this update is then

$$
k_5(Z_i|\mu, \Sigma, \tau, \sigma^2, Y') \propto |\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2}(Z_i - \mu_{Z_i})^t \Sigma_{Z_i}^{-1}(Z_i - \mu_{Z_i}) \right)
$$

$$
= |\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( \frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
= |\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( \frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( -\frac{1}{2}(Y_i')^t(D^{-1} + R^{-1})^{-1}\Sigma_{Z_i}^{-1}(D^{-1} + R^{-1})^{-1}Y_i' \right)
$$

by substituting

$$
\hat{\mu}_i = (D^{-1} + R^{-1})^{-1}D^{-1}Y_i'
$$

and by multiplication of the cross product

$$
\Sigma_{Z_i}^{-1} = \sigma_i^{-2}(D^{-1} + R^{-1})
$$

$$
\geq |\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( \frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( -\frac{1}{2}\sigma_i^{-2}(Y_i')^t(Y_i') \right)
$$

which is true because $(D^{-1} + R^{-1})^{-1} \leq D$ by Result 4.5 (2),

$$
\geq 1_{D_{4i}}(Y_i')|\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2}(Z_i + \hat{\mu}_i)\Sigma_{Z_i}^{-1}(Z_i + \hat{\mu}_i) \right)
$$

$$
\times \exp \left( -|(Z_i + \hat{\mu}_i)|\sigma_i^{-2}D^{-1}d_{4i} \right)
$$

180
which is true by applying the constraint, recalling that \( D \) is diagonal with all positive entries,

\[
= s_5(Y'_i)q_5(\mu, \Sigma, \tau, \sigma_i^2, Z_i)
\]

where

\[
s_5(Y'_i) = \mathbb{1}_{D_4}(Y'_i)
\]

and

\[
q_5(\mu, \Sigma, \tau, \sigma^2) = |\Sigma_{Z_i}|^{-1/2} \exp \left( -\frac{1}{2} (Z_i + \hat{\mu}_i)^T \Sigma_{Z_i}^{-1} (Z_i + \hat{\mu}_i) \right) \\
\times \exp \left( -|Z_i + \hat{\mu}_i| \sigma_i^{-2} D^{-1} d_{4i} \right) \\
\times \exp \left( -\frac{1}{2} \sigma_i^2 d_{4i}^T D^{-1} d_{4i} \right)
\]

**Update for \( c \).** The final transition density requiring manipulation is the independence Metropolis-Hastings update for \( c \). Recall that \( g(c) \) is the candidate distribution and define

\[
h(c) = \prod_{k=2}^{K-1} (c_k - c_{k-1})^{\alpha_k - 1} \frac{w(c)}{g(c)}.
\]

Then,

\[
k_6(c|c', \mu, \Sigma, \tau, \sigma^2, Z) \geq g(c) \min \left( 1, \frac{h(c)}{h(c')} \right) \\
\geq g(c) \min \left( 1, \frac{h(c)}{d_6} \right) \min \left( 1, \frac{d_6}{h(c')} \right)
\]

since for any positive \( a \) and \( b \),

\[
\min(1, ab) \geq \min(1, a) \min(1, b),
\]

which is easily proved by considering cases;

\[
= g(c) \min \left( 1, \frac{h(c)}{d_6} \right) \min \left( 1, \prod_{k=2}^{K-1} (c'_k - c'_{k-1})^{-\alpha_k - 1} \frac{d_6 g(c')}{w(c')} \right)
\]
\[ g(c) \min \left( 1, \frac{h(c)}{d_0} \right) \min \left( 1, \prod_{k=2}^{K-1} (c'_k - c'_{k-1})^{-(\alpha_k-1)} d_0 g(c') \right), \]

which is true because \( w(c') \leq 1 \) by construction recalling

\[ w(c) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left\{ \Phi \left( \frac{c_{X_{ij}} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) - \Phi \left( \frac{c_{X_{ij-1}} - \mu_j - \tau_i - Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) \right\}. \]

Note that this inequality is required because \( w(c') \) is a function of the previously updated parameters \( \mu, \Sigma, \tau, \sigma^2, \) and \( Z. \) Continuing,

\[ k_6(c|c', \mu, \Sigma, \tau, \sigma^2, Z) = s_6(c')q_6(c, \mu, \Sigma, \tau, \sigma^2, Z), \]

where

\[ s_6(c') = \min \left( 1, \prod_{k=2}^{K-1} (c'_k - c'_{k-1})^{-(\alpha_k-1)} d_0 g(c') \right), \] and

\[ q_6(c, \mu, \Sigma, \tau, \sigma^2, Z) = g(c) \min \left( 1, \frac{h(c)}{d_0} \right). \]

Note that \( d_6 > 0 \) may be chosen to be any convenient constant.

**Update for Y.** Finally, the update for each \( Y_{ij} \) is free of any of the original variables, and is distributed

\[ \text{Truncated Normal}(\mu_j + \tau_i + Z_{ij}, \sigma_i^2 D_{jj}, c_{X_{i,j-1}}, c_{X_{i,j}}). \]

The density is

\[ k_{\tau ij}(Y_{ij}, \mu_j, \tau_i, D_{jj}, \sigma_i^2, Z_{ij}) \propto d_{\tau ij}^{-1/2} \exp \left( -\frac{1}{2\sigma^2} (Y_{ij} - \mu_j - \tau_i - Z_{ij})^2 \right), \]

where

\[ d_{\tau ij} = \Phi \left( \frac{c_{X_{ij}} - \mu_j + \tau_i + Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right) - \Phi \left( \frac{c_{X_{ij-1}} - \mu_j + \tau_i + Z_{ij}}{\sqrt{\sigma_i^2 D_{jj}}} \right). \]
**Combined Minorization Condition.** Combining all of these results together, and denoting the totality of the parameters as \( \theta \), the minorization condition is

\[
k(\theta', \theta) \geq s(\theta')q(\theta),
\]

where

\[
s(\theta') = \prod_{i=1}^{N} s_{1i}(Y_i', \mu', \Sigma', \tau_i') s_{4i}(Y_i') s_{5i}(Y_i')
\]
\[
\times s_2(Y', \mu', \tau') s_3(Y', \tau') s_6(c')
\]

and

\[
q(\theta) = \prod_{i=1}^{N} q_{1i}(\sigma_i^2) q_{4i}(\mu, \Sigma, \tau_i, \sigma_i^2) q_{5i}(\mu, \Sigma, \tau_i, \sigma_i^2, Z_i)
\]
\[
\times q_2(\Sigma, \sigma^2) q_3(\mu, \Sigma, \sigma^2) q_6(c, \mu, \Sigma, \tau, \sigma^2, Z)
\]
\[
\times \prod_{i=1}^{N} \prod_{j=1}^{M} k_{ij}(Y_{ij}, \mu_j, \tau_i, D_{jj}, \sigma_i^2, Z_{ij}).
\]

This minorization condition could then be used for regenerative sampling methods, including the approximate methods of Chapter 3. For use with Chapter 2 in computing bounds for total variation distance, the function \( s(\cdot) \) will need to be minimized over the small set, adding another layer of complexity to such implementations. While perhaps not efficient, this minorization condition nonetheless makes methods which require its use available to scale-usage problems.

**Conclusions**

This chapter provided several results for the convergence analysis of Bayesian scale-usage models. First, the motivation behind probit regression and Bayesian scale-usage models was discussed. MCMC strategies for these models were provided. Then, four models were discussed for which the stationary Markov chains were shown...
to be geometrically ergodic. Finally, a minorization condition was proved for scale-usage models. These results aid in guaranteeing the existence of a Markov chain central limit theorem, which can then be used to thoroughly answer Question 1 and Question 2 for Bayesian scale-usage models in practice. This dissertation now shifts to discuss an extension of Bayesian scale-usage models when scale-usage heterogeneity differs due to demographics. The results in Chapter 4 will allow for the two key convergence questions to then be answered for these modified Bayesian scale-usage models for a particular dataset and corresponding inference.
Chapter 5: Scale-Usage and the Effects of Demographics

As widely established in the literature, scale-usage patterns differ across cultures. For a review and example where this phenomenon is established, see Clarke (2000). The assertion therein is that without adjusting for differences in cultures, statistically significant results may be obtained which are due to response style alone, rather than the true question of interest.

With such meaningful differences in scale-usage across cultures, it is plausible that demographic differences could be important when analyzing data with scale-usage heterogeneity. Section 5.1 discusses models which may be useful in accounting for demographic effects. Section 5.2 discusses how model selection between these and other scale-usage models may be performed. Section 5.3 provides a simulation study where the models from Section 5.1 and the model selection procedures from Section 5.2 are evaluated. Finally, Section 5.4 applies the previous sections to an analysis of student ratings data.
5.1 Scale-Usage Models Incorporating Demographic Effects

Simple modifications of the third geometrically ergodic scale-usage model in Chapter 4, Model 4.3, allow for the incorporation of demographic-specific scale-usage effects. Let the demographic of interest have $L$ levels, and let $d_i$ denote the demographic for person $i$, $i = 1, \ldots, N$. Define $D_\ell = \{ i : d_i = \ell \}$ for $\ell = 1, \ldots, L$, and let $n_\ell = |D_\ell|$, the cardinality of $D_\ell$. Note that for computational simplicity, all the models in Chapter 5 include the covariance decomposition, proposed by Hans et al. (2012) and discussed in Chapter 4, which provides several computational advantages over a model without such a decomposition.

5.1.1 Model 5.1: Demographic-Based Scale-Usage Shift

The following model specifically accommodates differences in shift for each demographic. Let $X$ be an $N \times M$ matrix ($N$ individuals, $M$ questions) such that for all $i = 1, \ldots N$, $j = 1, \ldots, M$, and $\ell = 1, \ldots, L$

$$X_{ij} | c, Y_{ij} = \{ k : c_{k-1} < Y_{ij} \leq c_k, \quad k = 1, \ldots K \},$$

$$Y_i \sim \text{Multivariate Normal}(\mu + \tau_i 1 + Z_i, \sigma_i^2 D)$$

$$\tau_i \sim \text{Truncated Normal}(\theta_{d_i}, \sigma_{\tau_i}^2, -a_\tau, a_\tau)$$

$$\theta_\ell \sim \text{Uniform}(-a_\theta, a_\theta)$$

$$\sigma_{\tau_i}^2 \propto (\sigma_{\tau_i}^2)^{-1} \mathbb{1}(\sigma_{\tau_i}^2 \in (a^{-1}_{\sigma_\tau}, a_{\sigma_\tau}^2))$$

$$\sigma_i^2 \sim \text{Truncated Inverse Gamma}(a, a - 1, 1/a_{\sigma^2}, a_{\sigma^2}^2)$$

$$Z_i \sim \text{Multivariate Normal}(0, \sigma_i^2 R)$$

$$\Sigma \sim \text{Inverse Wishart}(\delta, \Sigma_0)$$

$$\pi(\mu) \propto 1$$
\[
\pi(c) \propto \prod_{k=1}^{K} (c_k - c_{k-1})^{\alpha_{k-1}}, \quad \text{where } c_0 = -C \text{ and } c_K = C, \tag{5.1}
\]

where \(a_r, a_\sigma, a_{\sigma^2}, a, a_{\sigma^2}, \delta, \Sigma_0, \) and \(\alpha_k\) for \(k = 1, \ldots, K\) are all constants. This model differs from Model 4.3 in Chapter 4 by allowing each \(\tau_i\) to be centered \emph{a priori} at a demographic-specific mean \(\theta_{d_i}\), rather than at 0. A demographic-specific variance for the shift parameter for each demographic, \(\sigma^2_{\tau_i}\), is also modeled.

The full posterior distribution for this model has the following form:

\[
\pi(Y, Z, \mu, \Sigma, \tau, \sigma^2_\tau, \theta, \sigma_\tau^2, c | X) \propto |D|^{-N/2} \prod_{i=1}^{N} (\sigma_i^2)^{M/2} \prod_{j=1}^{M} \mathbb{I}(c_{X_{i,j-1}} < Y_{ij} \leq c_{X_{i,j}})
\]

\[
\times \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_i^{-2}(Y_i - \mu - \tau_i 1 - Z_i)^T D^{-1}(Y_i - \mu - \tau_i 1 - Z_i) \right)
\]

\[
\times |R|^{-N/2} \prod_{i=1}^{N} (\sigma_i^2)^{-M/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_i^{-2} Z_i^T R^{-1} Z_i \right)
\]

\[
\times \prod_{i=1}^{N} \left( (\sigma_i^2)^{-a-1} \exp \left( -\frac{(a-1)}{\sigma_i^2} \right) \mathbb{I}(a^{-1}, a; \sigma_i^2) \right)
\]

\[
\times \left[ \prod_{\ell=1}^{L} (\sigma_{\tau_\ell}^2)^{-n_{\ell}/2} \right] \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \frac{(\tau_i - \theta_{d_i})^2}{\sigma_{\tau_i}^2} \right) \prod_{i=1}^{N} \mathbb{I}(a_r, a_\sigma; \tau_i)
\]

\[
\times \prod_{\ell=1}^{L} (\sigma_{\tau_\ell}^2)^{-1} \mathbb{I}(a_r^{-1}, a_\sigma; \sigma_{\tau_\ell}^2) \mathbb{I}(a_\sigma, a_\sigma; \theta_\ell)
\]

\[
\times |\Sigma|^{-(\delta+M+1)/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma_0 \Sigma^{-1}) \right)
\]

\[
\times \prod_{i=1}^{K} (c_k - c_{k-1})^{\alpha_{k-1}}. \tag{5.2}
\]

The updating scheme for the Markov chain stationary for this posterior distribution follows the general updating scheme of Model 4.3 with additional updates for the added parameters. The algorithm is now provided. Denote the entire collection of parameters \(\Theta = \{Y, Z, \mu, \Sigma, \tau, \sigma^2_\tau, \theta, \sigma_\tau^2, c\}\).
**Algorithm 5.1** (Demographic-Specific Effects Scale-Usage Model)

From the current $\Theta'$ parameters, update to $\Theta$ according to the following steps.

(i) $\Sigma|Y', \mu', \tau', \sigma^2 \sim$ Inverse Wishart$(N + \delta, S' + \Sigma_0)$ where

$$S' = \sum_{i=1}^{N} \frac{1}{\sigma_i^2} (Y_i' - \mu' - \tau'_i 1)(Y_i' - \mu' - \tau'_i 1)^T.$$  

(ii) $\sigma_i^2|Y', \mu', \tau_i, \Sigma \sim$ Truncated Inverse Gamma$(M/2 + a, b_{\sigma_i^2}, a_{\sigma_i^2}, a_{\sigma_i^2})$ where

$$b_{\sigma_i^2} = \frac{1}{2} (Y_i' - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau'_i 1) + a - 1.$$  

(iii) $Z_i|Y', \mu', \tau', \sigma^2, \Sigma \sim$ Multivariate Normal$(\mu_{Z_i}, \Sigma_{Z_i})$ where

$$\mu_{Z_i} = (D^{-1} + R^{-1})^{-1} D^{-1} (Y_i' - \mu'_i - \tau'_i 1), \quad \text{and}$$

$$\Sigma_{Z_i} = \sigma_i^2 (D^{-1} + R^{-1})^{-1}.$$  

(iv) Update $c_k|c_{k-1}, c_{k+1}', \mu', \tau', Z, \Sigma$ sequentially with a Metropolis-Hastings step utilizing some candidate distribution $g(\cdot|c_k)$ to generate a candidate $c^*_k$ which is accepted with probability

$$\alpha_k = \min \left\{ 1, \left( \frac{c_{k+1} - c^*_k}{c_{k+1} - c_k} \right)^{\alpha_{k+1} - 1} \left( \frac{c^*_k - c_{k-1}}{c_k - c_{k-1}} \right)^{\alpha_{k-1} - 1} \frac{w_k(c^*_k) g(c_k|c^*_k)}{w_k(c_k) g(c^*_k|c_k)} \right\},$$

where

$$w_k(c_k) = \prod_{\ell = k}^{k+1} \prod_{(i,j): X_{ij} = \ell} \left\{ \Phi \left( \frac{c_{\ell} - \mu_j' - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_{\ell} D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu_j' - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_{\ell} D_{jj}}} \right) \right\}.$$  

(v) $Y_{ij}|\mu', \tau'_i, \Sigma, \sigma^2, c \sim$ Truncated Normal$(\mu_j' + \tau'_i + Z_{ij}, \sigma^2_{\ell} D_{jj}, c_{X_{ij}, -1}, c_{X_{ij}})$.

(vi) $\sigma^2_{\tau' \ell}|\tau'_\ell, \theta'_\ell \sim$ Truncated Inverse Gamma$(n_\ell/2, b_{\sigma_{\tau' \ell}^2}, a_{\sigma_{\tau' \ell}^2}, a_{\sigma_{\tau' \ell}^2})$ where

$$b_{\sigma_{\tau' \ell}^2} = \frac{1}{2} \sum_{i \in D_\ell} (\tau'_i - \theta'_\ell)^2.$$
(vii) \( \tau_i | Y_i, \mu', \theta_d, \sigma_i^2, \sigma_{\tau_i}^2, \Sigma \sim \text{Truncated Normal}(\mu_{\tau_i}, s_{\tau_i}^2, -a_\tau, a_\tau) \) where
\[
\mu_{\tau_i} = s_{\tau_i}^2 \left[ \sigma_i^{-2} (Y_i - \mu')^T \Sigma^{-1} 1 + \sigma_{\tau_i d_i}^{-2} \theta_d \right], \quad \text{and}
\]
\[
s_{\tau_i}^2 = (\sigma_i^{-2} 1^T \Sigma^{-1} 1 + \sigma_{\tau_i}^{-2})^{-1}.
\]

(viii) \( \theta | \tau, \sigma_{\tau}^2 \sim \text{Truncated Normal}(\bar{\tau}_\ell, \sigma_{\tau_i}^2 n_\ell^{-1}, -a_\theta, a_\theta) \) where
\[
\bar{\tau}_\ell = n_\ell^{-1} \sum_{i \in D_\ell} \tau_i.
\]

(ix) \( \mu | Y, \tau, \sigma^2, \Sigma, \sim \text{Multivariate Normal}(\mu_{\mu}, \Sigma_{\mu}) \), where
\[
\mu_{\mu} = \left( \sum_{i=1}^N \sigma_i^{-2} \right)^{-1} \sum_{i=1}^N \sigma_i^{-2} (Y_i - \tau_i 1), \quad \text{and}
\]
\[
\Sigma_{\mu} = \left( \sum_{i=1}^N \sigma_i^{-2} \right)^{-1} \Sigma.
\]

This model therefore accounts for shifts in scale-usage based on demographics and further characterizes the variability within each demographic more appropriately. It does not, however, address the differences in scale between demographics, at least explicitly. As all new parameters introduced to the model have compact support, this Markov chain will be geometrically ergodic due to Proposition 4.21, a fact which will be employed in inference to compute Markov chain Monte Carlo standard errors.

5.1.2 Model 5.2: Demographic-Based Scale-Usage Shift and Scale

This model is identical to Model 5.1 with the exception that rather than having individual scale-usage variability \( \sigma_i^2 \), it is instead assumed that the entire demographic uses scale similarly with one common \( \sigma_\ell^2 \). Thus, Model 5.2 is Model 5.1 where \( \sigma_i^2 = \sigma_\ell^2 \) for each \( i \in D_\ell \). This may be a better model when culture rather than individual variability largely determines how much of the scale a respondent uses.
Explicitly written, for \( i = 1, \ldots N, j = 1, \ldots, M, \) and \( \ell = 1, \ldots, L, \) let

\[
X_{ij} | c, Y_{ij} = \{ k : c_{k-1} < Y_{ij} \leq c_k, \; k = 1, \ldots K \},
\]

\[
Y_i \sim \text{Multivariate Normal}(\mu + \tau_i 1 + Z_i, \sigma_{d_i}^2 D)
\]

\[
Z_i \sim \text{Multivariate Normal}(0, \sigma_{d_i}^2 R)
\]

\[
\sigma_{\ell}^2 \sim \text{Truncated Inverse Gamma}(a, a_{-1}^2, a_{+2}^2),
\]

with the remaining parameters and constants defined as in Model 5.1. The resulting posterior distribution is then

\[
\pi(Y, Z, \mu, \Sigma, \tau, \sigma^2, \theta, \sigma_{\tau}^2, c | X) \propto |D|^{-N/2} \prod_{i=1}^{N} \left( \sigma_{d_i}^2 \right)^{-M/2} \prod_{j=1}^{M} \mathbb{I}(c_{X_{i,j-1}} < Y_{ij} \leq c_{X_{i,j}}) \times \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_{d_i}^2 (Y_i - \mu - \tau_i 1 - Z_i)^\top D^{-1} (Y_i - \mu - \tau_i 1 - Z_i) \right)
\]

\[
\times |R|^{-N/2} \prod_{i=1}^{N} (\sigma_{d_i}^2)^{-M/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_{d_i}^{-2} Z_i^\top R^{-1} Z_i \right)
\]

\[
\times \prod_{\ell=1}^{L} \left[ (\sigma_{\ell}^2)^{-a_{-1}^2} \exp \left( -\frac{a_{-1}^2}{\sigma_{\ell}^2} \right) \mathbb{I}(a_{-1}^2, a_{+2}^2)(\sigma_{\ell}^2) \right] \times \prod_{\ell=1}^{L} (\sigma_{\tau_{\ell}}^2)^{-n_{\ell}/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \frac{(\tau_{\ell} - \theta_{d_{\ell}})^2}{\sigma_{\tau_{d_{\ell}}}^2} \right) \prod_{i=1}^{N} \mathbb{I}(a_{\tau}, a_{\tau})(\tau_{\ell}) \times \prod_{\ell=1}^{L} (\sigma_{\tau_{\ell}}^{-1} \mathbb{I}(a_{\tau}, a_{\tau})(\sigma_{\tau_{\ell}}^2) \mathbb{I}(a_{\theta}, a_{\theta})(\theta_{\ell}) \times |\Sigma|^{-\left(\delta + M + 1\right)/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma_0 \Sigma^{-1}) \right)
\]

\[
\times \prod_{k=1}^{K} \mathbb{I}(c_k - c_{k-1})^{\alpha_k-1}. \tag{5.3}
\]

The Markov chain stationary for this posterior distribution is similar to Algorithm 5.1, though with a different update for \( \sigma_{\ell}^2. \)

**Algorithm 5.2** (Demographic-Specific Scale-Usage Model with Demographic Scales)

From the current \( \Theta', \) update to \( \Theta \) according to the following steps.
(i) \( \Sigma | \mathbf{Y}', \mathbf{\mu}', \mathbf{\tau}', \sigma^{2g} \sim \text{Inverse Wishart}(\mathbf{N} + \delta, \mathbf{S}' + \Sigma_0) \) where
\[
\mathbf{S}' = \sum_{i=1}^{N} \frac{1}{\sigma_{d_i}^g} (\mathbf{Y}_i' - \mathbf{\mu}' - \mathbf{\tau}'_i)(\mathbf{Y}_i' - \mathbf{\mu}' - \mathbf{\tau}'_i)^T.
\]

(ii) \( \sigma^{2g}_i | \mathbf{Y}', \mathbf{\mu}', \mathbf{\tau}', \Sigma \sim \text{Truncated Inverse Gamma}(n_{\ell} M / 2 + a, b_{\sigma^2_i}, a_{\sigma^2}, a_{\sigma^2}) \) where
\[
b_{\sigma^2_i} = \frac{1}{2} \sum_{i \in D_{\ell}} \left[ (\mathbf{Y}_i' - \mathbf{\mu}' - \mathbf{\tau}'_i)^T \Sigma^{-1} (\mathbf{Y}_i' - \mathbf{\mu}' - \mathbf{\tau}'_i) + a - 1. \right]
\]

(iii) \( \mathbf{Z}_i | \mathbf{Y}', \mathbf{\mu}', \mathbf{\tau}', \sigma^{2g}_{d_i}, \Sigma \sim \text{Multivariate Normal}(\mathbf{\mu}_{Z_i}, \Sigma_{Z_i}) \) where
\[
\mathbf{\mu}_{Z_i} = (\mathbf{D}^{-1} + \mathbf{R}^{-1})^{-1} \mathbf{D}^{-1} (\mathbf{Y}_i' - \mathbf{\mu}'_i - \mathbf{\tau}'_i) \text{, and}
\]
\[
\Sigma_{Z_i} = \sigma^{2g}_{d_i} (\mathbf{D}^{-1} + \mathbf{R}^{-1})^{-1}.
\]

(iv) Update \( c_k | c_{k-1}, c'_{k+1}, \mathbf{\mu}', \mathbf{\tau}', \mathbf{Z}, \Sigma \) sequentially with a Metropolis-Hastings step utilizing some candidate distribution \( g(\cdot | c_k) \) to generate a candidate \( c'_k \) which is accepted with probability
\[
\alpha_k = \min \left\{ 1, \left( \frac{c_{k+1} - c'_k}{c_{k+1} - c_k} \right)^{\alpha_{k+1} - 1} \left( \frac{c'_k - c_k}{c'_k - c_{k-1}} \right)^{\alpha_k - 1} \frac{w_k(c'_k)g(c'_k)}{w_k(c_k)g(c_k)} \right\},
\]
where
\[
w_k(c_k) = \prod_{i=k}^{k+1} \prod_{(i,j): \mathbf{X}_{ij} = \ell} \left\{ \Phi \left( \frac{c_{\ell} - \mathbf{\mu}'_j - \mathbf{\tau}'_i - \mathbf{Z}_{ij}}{\sqrt{\sigma^{2g}_{d_i} D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mathbf{\mu}'_j - \mathbf{\tau}'_i - \mathbf{Z}_{ij}}{\sqrt{\sigma^{2g}_{d_i} D_{jj}}} \right) \right\}.
\]

(v) \( \mathbf{Y}_{ij} | \mathbf{\mu}', \mathbf{\tau}', \Sigma, \sigma^{2g}_c \sim \text{Truncated Normal}(\mathbf{\mu}'_j + \mathbf{\tau}'_i + \mathbf{Z}_{ij}, \sigma^{2g}_{d_i} D_{jj}, \mathbf{c}_{X_{i,j-1}, c_{X_{i,j}}} \).}

(vi) \( \sigma^{2g}_{\mathbf{\tau}_c} | \mathbf{\tau}', \theta^{\tau}_c \sim \text{Truncated Inverse Gamma}(n_{\ell} / 2, b_{\sigma^2_{\tau_c}}, a_{\sigma^2}, a_{\sigma^2}) \) where
\[
b_{\sigma^2_{\tau_c}} = \frac{1}{2} \sum_{i \in D_{\ell}} (\mathbf{\tau}'_i - \theta^{\tau}_c)^2.
\]

(vii) \( \mathbf{\tau}_i | \mathbf{Y}_i, \mathbf{\mu}', \theta_{d_i}, \sigma^{2g}_{d_i}, \Sigma \sim \text{Truncated Normal}(\mu_{\tau_i}, s^{2g}_{\tau_i}, -a_{\tau}, a_{\tau}) \) where
\[
\mu_{\tau_i} = s^{2g}_{\tau_i} \left[ \sigma^{-2}_{d_i} (\mathbf{Y}_i - \mathbf{\mu}')^T \Sigma^{-1} 1 + \sigma^{-2}_{\tau_i} \right], \text{ and}
\]
\[
s^{2g}_{\tau_i} = (\sigma^{-2}_{d_i} 1^T \Sigma^{-1} 1 + \sigma^{-2}_{\tau_i})^{-1}.
\]

191
(viii) $\theta_\ell | \tau, \sigma^2_{\tau_\ell} \sim \text{Truncated Normal}(\bar{\tau}_\ell, \sigma^2_{\tau_\ell} n_{\ell}^{-1}, -a_\theta, a_\theta)$ where

$$\bar{\tau}_\ell = n_{\ell}^{-1} \sum_{i \in D_\ell} \tau_i.$$ 

(ix) $\mu | Y, \tau, \sigma^2, \Sigma, \sim \text{Multivariate Normal}(\mu_\mu, \Sigma_\mu)$, where

$$\mu_\mu = \left( \sum_{i=1}^{N} \sigma_{d_i}^{-2} \right)^{-1} \sum_{i=1}^{N} \sigma_{d_i}^{-2} (Y_i - \tau_i 1), \quad \text{and}$$

$$\Sigma_\mu = \left( \sum_{i=1}^{N} \sigma_{d_i}^{-2} \right)^{-1} \Sigma.$$ 

Geometric ergodicity for this sampler is provided by Proposition 4.21 thanks to the compact support of the $\theta_\ell$ and $\sigma^2_{\tau_\ell}$ parameters. The next model addresses the situation where all of the $\sigma^2_{\tau_\ell}$ are close to one.

5.1.3 Model 5.3: Demographic-Based Scale-Usage Shift with No Scale Adjustment

This model is a special case of Model 5.1 where $\sigma^2_i = 1$ for all $i = 1, \ldots, N$. Thus, scale-usage heterogeneity is accounted for in the mean, but not in the variance. The full model and posterior follow from Model 5.1 where $\sigma^2_i = 1$. The only change in the algorithm for this model from Algorithm 5.1 is that $\sigma^2_i$ is not updated, but rather stays constant equal to 1. This algorithm is similarly geometrically ergodic by Proposition 4.21.

Two other models are considered in this chapter as a comparison when demographic effects are not required.

5.1.4 Model 5.4: No Demographic Effects

This model is similar to Model 4.3, but with a hyperprior on the parameter $\sigma^2_{\tau}$, rather than fixing it as constant. The model is equivalent to Model 5.1 where $\theta_\ell = 0$. 

192
for all $\ell = 1, \ldots, L$ and $\sigma_{\tau_i}^2 = \cdots = \sigma_{\tau_L}^2$. The model is therefore, for $i = 1, \ldots, N$ and $j = 1, \ldots, M$,

$$X_{ij} | c, Y_{ij} = \{k : c_{k-1} < Y_{ij} \leq c_k, \quad k = 1, \ldots, K\},$$

$$Y_i \sim \text{Multivariate Normal}(\mu + \tau_i 1 + Z_i, \sigma_i^2 D)$$

$$\tau_i \sim \text{Truncated Normal}(0, \sigma_{\tau_i}^2, -a_{\tau}, a_{\tau})$$

$$\sigma_{\tau_i}^2 \propto (\sigma_i^2)^{-1} 1(\sigma_i^2 \in (a_{-1}, a_1))$$

with the prior distributions for $\sigma_i^2, Z_i, \Sigma, \mu$, and $c$ unchanged from Model 5.1. The full posterior distribution follows:

$$\pi(Y, Z, \mu, \Sigma, \tau, \sigma^2, \theta, \sigma^2, c | X) \propto |D|^{-N/2} \prod_{i=1}^{N} (\sigma_i^2)^{-M/2} \prod_{j=1}^{M} 1(c_{X_{i,j}-1} < Y_{ij} \leq c_{X_{i,j}})$$

$$\times \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_i^{-2} (Y_i - \mu - \tau_i 1 - Z_i)^T D^{-1} (Y_i - \mu - \tau_i 1 - Z_i) \right)$$

$$\times |R|^{-N/2} \prod_{i=1}^{N} (\sigma^2_i)^{-M/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \sigma_i^{-2} Z_i^T R^{-1} Z_i \right)$$

$$\times \prod_{i=1}^{N} \left( (\sigma_i^2)^{-a-1} \exp \left( -\frac{(a-1)}{\sigma_i^2} \right) 1_{(a_{-1}, a_1)}(\sigma_i^2) \right)$$

$$\times (\sigma_{\tau_i}^2)^{-N/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \tau_{\tau_i}^2 \right) \prod_{i=1}^{N} 1_{(-a_\tau, a_\tau)}(\tau_i)$$

$$\times \sigma_{\tau_i}^{-2} 1_{(a_{-1}, a_1)}(\sigma_{\tau_i}) 1_{(-a_\theta, a_\theta)}(\theta)$$

$$\times |\Sigma|^{-(d+M+1)/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma \Sigma^{-1}) \right)$$

$$\times \prod_{i=1}^{K} (c_k - c_{k-1})^{\alpha_k-1}, \quad (5.4)$$

which induces the following Markov chain, similar to Algorithm 5.1.

**Algorithm 5.4** (Standard Scale-Usage Model)

From the current $\Theta'$, update to $\Theta$ according to the following steps.
(i) $\Sigma|Y', \mu', \tau', \sigma^2 \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0)$ where

$$S' = \sum_{i=1}^{N} \frac{1}{\sigma^2_i} (Y_i' - \mu' - \tau'_i 1) (Y_i' - \mu' - \tau'_i 1)^T.$$  

(ii) $\sigma^2_i|Y', \mu', \tau_i, \Sigma \sim \text{Truncated Inverse Gamma}(M/2 + a, b_{\sigma^2_i}, a_{\sigma^2_i}, a_{\sigma^2_i})$ where

$$b_{\sigma^2_i} = \frac{1}{2} (Y_i' - \mu' - \tau'_i 1)^T \Sigma^{-1} (Y_i' - \mu' - \tau'_i 1) + a - 1.$$  

(iii) $Z_i|Y', \mu', \tau', \sigma^2 \sim \text{Multivariate Normal}(\mu_{Z_i}, \Sigma_{Z_i})$ where

$$\mu_{Z_i} = (D^{-1} + R^{-1})^{-1} D^{-1} (Y_i' - \mu'_i - \tau'_i 1), \quad \text{and}$$  $$\Sigma_{Z_i} = \sigma^2_i (D^{-1} + R^{-1})^{-1}.$$  

(iv) Update $c_k|c_{k-1}, c'_{k+1}, \mu', \tau', Z, \Sigma$ sequentially with a Metropolis-Hastings step utilizing some candidate distribution $g(.|c_k)$ to generate a candidate $c^*_k$ which is accepted with probability

$$\alpha_k = \min\left\{ 1, \frac{(c_{k+1} - c_k)}{c_{k+1} - c_k} \right\} \frac{w_k(c^*_k)g(c^*_k|c_k)}{w_k(c_k)g(c_k|c_k)},$$

where

$$w_k(c_k) = \prod_{\ell = k}^{k+1} \prod_{i,j} \left\{ \Phi \left( \frac{c_{\ell'} - \mu'_i - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_{ij} D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu'_i - \tau'_i - Z_{ij}}{\sqrt{\sigma^2_{ij} D_{jj}}} \right) \right\}.$$  

(v) $Y_{ij}|\mu', \tau_i, \Sigma, \sigma^2, c \sim \text{Truncated Normal}(\mu'_j + \tau'_i + Z_{ij}, \sigma^2_i D_{jj}, c_{X_{i,j-1}}, c_{X_{i,j}}).$  

(vi) $\sigma^2_i|\tau' \sim \text{Truncated Inverse Gamma}(N/2, b_{\sigma^2_i}, a_{\sigma^2_i}, a_{\sigma^2_i})$ where

$$b_{\sigma^2_i} = \frac{1}{2} \sum_{i=1}^{N} (\tau'_i)^2.$$  

(vii) $\tau_i|Y_i, \mu', \sigma^2_i, \sigma^2, \Sigma \sim \text{Truncated Normal}(\mu_{\tau_i}, s_{\tau_i}^2, -a_{\tau}, a_{\tau})$ where

$$\mu_{\tau_i} = s_{\tau_i}^2 \left[ \sigma^{-2}_i (Y_i - \mu')^T \Sigma^{-1} 1 \right], \quad \text{and}$$  $$s_{\tau_i}^2 = (\sigma^{-2}_i 1^T \Sigma^{-1} 1 + \sigma^{-2}_2)^{-1}.$$  

194
(viii) \( \mu | \mathbf{Y}, \tau, \sigma^2, \Sigma, \sim \text{Multivariate Normal}(\mu_{\mu}, \Sigma_{\mu}) \), where
\[
\begin{align*}
\mu_{\mu} &= \left( \sum_{i=1}^{N} \sigma_i^{-2} \right)^{-1} \sum_{i=1}^{N} \sigma_i^{-2} (Y_i - \tau_i \mathbf{1}), \quad \text{and} \\
\Sigma_{\mu} &= \left( \sum_{i=1}^{N} \sigma_i^{-2} \right)^{-1} \Sigma.
\end{align*}
\]

Proposition 4.21 also ensures the geometric ergodicity of this sampler as the prior for \( \sigma_i^2 \) has compact support.

The final model considers the situation where there are in fact no scale-usage effects.

5.1.5 Model 5.5: No Demographic or Scale-Usage Effects

This model is a special case of Model 5.1 (or any of the previous models) where \( \sigma_i^2 = 1 \) and \( \tau_i = 0 \) for all \( i = 1, \ldots, N \). All associated hyperpriors are therefore unnecessary. Thus, only a mean vector and covariance matrix are utilized to fit the latent structure behind the ordinal data.

Formally, the model is, for \( i = 1, \ldots, N \) and \( j = 1, \ldots, M \),
\[
X_{ij} | \mathbf{c}, Y_{ij} = \{ k : c_{k-1} < Y_{ij} \leq c_k, \quad k = 1, \ldots K \},
\]
\[
Y_i \sim \text{Multivariate Normal}(\mu + Z_i, D)
\]
\[
Z_i \sim \text{Multivariate Normal}(\mathbf{0}, \mathbf{R})
\]
\[
\Sigma \sim \text{Inverse Wishart}(\delta, \Sigma_0)
\]
\[
\pi(\mu) \propto 1
\]
\[
\pi(\mathbf{c}) \propto \prod_{k=1}^{K} (c_k - c_{k-1})^{\alpha_k - 1}, \text{ where } c_0 = -C \text{ and } c_K = C,
\]
where \( c_k = C \), (5.5)

The full posterior distribution is then:
\[
\pi(\mathbf{Y}, \mathbf{Z}, \mathbf{\mu}, \Sigma, \mathbf{c} | \mathbf{X}) \propto |\mathbf{D}|^{-N/2} \prod_{i=1}^{N} \prod_{j=1}^{M} \mathbb{1}(c_{X_{i,j}-1} < Y_{ij} \leq c_{X_{i,j}})
\]

195
\begin{align*}
&\times \exp \left( -\frac{1}{2} \sum_{i=1}^{N} (Y_i - \mu_i - Z_i)^T D^{-1} (Y_i - \mu_i - Z_i) \right) \\
&\times |R|^{-N/2} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} Z_i^T R^{-1} Z_i \right) \\
&\times |\Sigma|^{-(\delta+M+1)/2} \exp \left( -\frac{1}{2} \text{tr}(\Sigma_0 \Sigma^{-1}) \right) \\
&\times \prod_{i=1}^{K} (c_k - c_{k-1})^{\alpha_k-1}.
\end{align*}

(5.6)

The corresponding algorithm follows directly.

**Algorithm 5.5** (Standard Model (no scale-usage))

From the current $\Theta'$, update to $\Theta$ according to the following steps.

(i) $\Sigma|Y', \mu' \sim \text{Inverse Wishart}(N + \delta, S' + \Sigma_0)$ where

\[ S' = \sum_{i=1}^{N} (Y'_i - \mu')(Y'_i - \mu')^T. \]

(ii) $Z_i|Y', \mu', \Sigma \sim \text{Multivariate Normal}(\mu_{Z_i}, \Sigma_{Z_i})$ where

\[ \mu_{Z_i} = (D^{-1} + R^{-1})^{-1} D^{-1} (Y'_i - \mu'_i), \quad \text{and} \]
\[ \Sigma_{Z_i} = (D^{-1} + R^{-1})^{-1}. \]

(iii) Update $c_k|c_{k-1}, c'_{k+1}, \mu', Z, \Sigma$ sequentially with a Metropolis-Hastings step utilizing some candidate distribution $g(\cdot|c_k)$ to generate a candidate $c^*_k$ which is accepted with probability

\[
\alpha_k = \min \left\{ 1, \left( \frac{c_{k+1} - c_k}{c_{k+1} - c_k} \right)^{\alpha_{k+1}-1} \left( \frac{c_k^* - c_{k-1}}{c_k^* - c_{k-1}} \right)^{\alpha_{k-1}} \frac{w_k(c_k^*) g(c_k^*|c_k^*)}{w_k(c_k) g(c_k|c_k)} \right\},
\]

where

\[
w_k(c_k) = \prod_{\ell=k}^{k+1} \prod_{(i,j): X_{ij} = \ell} \left\{ \Phi \left( \frac{c_{\ell} - \mu'_j - Z_{ij}}{\sqrt{D_{jj}}} \right) - \Phi \left( \frac{c_{\ell-1} - \mu'_j - Z_{ij}}{\sqrt{D_{jj}}} \right) \right\},
\]

196
(iv) \( Y_{ij} | \mu', \Sigma, c \sim \text{Truncated Normal}(\mu'_j + Z_{ij}, D_{jj}, c_{X_{ij}}, c_{X_{ij}}) \).

(v) \( \mu | Y, \Sigma, \sim \text{Multivariate Normal}(\mu, \Sigma) \), where

\[
\mu = N^{-1} \sum_{i=1}^{N} Y_i, \quad \text{and}
\]

\[
\Sigma = N^{-1} \Sigma.
\]

This model is also geometrically ergodic via the results in Chapter 4.

### 5.1.6 Summary of Demographic-Specific Models

A summary of these five potential models is provided in Table 5.1. Note that for the purposes of this table, \( TN(\mu, \sigma^2) \) indicates a truncated normal distribution with shift \( \mu \), scale \( \sigma^2 \), and assumed truncation points \(-a_\tau\) and \( a_\tau\). Similarly, \( TIG(a, b) \) indicates a truncated inverse gamma distribution with shape \( a \), scale \( b \), and assumed truncation points \( a_{\sigma^2}^{-1} \) and \( a_{\sigma^2} \).

In order to make a selection from Model 5.1 through Model 5.5, model selection tools are required.

### 5.2 Model Selection for Scale-Usage Models

With several potential models, a discussion about model selection in the Bayesian framework is prudent. Bayes factors and DIC are now introduced, followed by a discussion of their application in the context of Bayesian scale-usage models. An ad-hoc approach is then provided.

#### 5.2.1 Bayes Factors

The Bayes factor was introduced in a hypothesis testing framework by Jeffreys (1935, 1961). The idea was modernized and popularized by Kass and Raftery (1995).
Table 5.1: Summary of Demographic-Specific Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
<th>Shift Prior / Scale Prior</th>
</tr>
</thead>
</table>
| 5.1   | o Individual shift adjustment has a common demographic mean  
      o Individual scale adjustment | \( \tau_i \sim TN(\theta_d, \sigma^2_{\tau_d}) \)  
                      \( \sigma_i^2 \sim TIG(a, a - 1) \) |
| 5.2   | o Individual shift adjustment has a common demographic mean  
      o Common demographic scale adjustment | \( \tau_i \sim TN(\theta_d, \sigma^2_{\tau_d}) \)  
                      \( \sigma_d^2 \sim TIG(a, a - 1) \) |
| 5.3   | o Individual shift adjustment has a common demographic mean  
      o No scale adjustment | \( \tau_i \sim TN(\theta_d, \sigma^2_{\tau_d}) \)  
                      \( \sigma_i^2 = 1 \) |
| 5.4   | o Individual shift adjustment  
      o Individual scale adjustment | \( \tau_i \sim TN(0, \sigma^2_{\tau}) \)  
                      \( \sigma_i^2 \sim TIG(a, a - 1) \) |
| 5.5   | o No individual shift adjustment  
      o No individual scale adjustment | \( \tau_i = 0 \)  
                      \( \sigma_i^2 = 1 \) |

According to their notation and definitions, let \( \mathbf{D} \) represent data which has arisen according to one of two hypothesized models, \( H_1 \) or \( H_2 \), which correspond to the likelihood functions \( P(\mathbf{D}|H_1) \) and \( P(\mathbf{D}|H_2) \). Let \( P(H_1) \) and \( P(H_2) \) be prior probabilities of the models, and \( P(H_1|\mathbf{D}) \) and \( P(H_2|\mathbf{D}) \) be the posterior probabilities of the models. Using Bayes’ Theorem, the posterior probability of each model is

\[
P(H_k|\mathbf{D}) = \frac{P(\mathbf{D}|H_k) P(H_k)}{P(\mathbf{D}|H_1) P(H_1) + P(\mathbf{D}|H_2) P(H_2)}, \quad k = 1, 2,
\]

which implies that the posterior odds are

\[
\frac{P(H_1|\mathbf{D})}{P(H_2|\mathbf{D})} = \frac{P(\mathbf{D}|H_1) P(H_1)}{P(\mathbf{D}|H_2) P(H_2)}.
\]

Thus, to arrive at the posterior odds from the prior odds, simply multiply the prior odds by

\[
B_{12} = \frac{P(\mathbf{D}|H_1)}{P(\mathbf{D}|H_2)}, \quad (5.7)
\]

198
which is the Bayes factor. Note that the $P(D|H_k)$ is found through integration

$$P(D|H_k) = \int \pi(\theta_k|H_k) P(D|\theta_k, H_k) d\theta_k \quad k = 1, 2$$

where $\theta_k$ are the parameters governing the model $H_k$ and $\pi(\theta_k|H_k)$ is the prior distribution of $\theta_k$ under model $H_k$ (Kass and Raftery 1995). This integral is often not directly computable, which often prohibits the usage of exact Bayes factors for model selection. Several methods for approximating $P(D|H_k)$ are available; see Kass and Raftery (1995) for a tutorial to several of these approximations. The present application will use the approximation provided by (11) of Kass and Raftery (1995) as

$$\hat{P}(D|H_k) = \left\{ \frac{1}{m} \sum_{i=1}^{m} P(D|\theta^{(i)})^{-1} \right\}^{-1}, \quad (5.8)$$

where $\theta^{(i)}$ is the $i^{th}$ posterior sample of model $H_k$ obtained from Markov chain Monte Carlo methods. This estimator is based on an importance sampling estimate using the posterior distribution as the importance sampling function. Kass and Raftery (1995) state that this estimate converges almost surely to $P(D|H_k)$ as $m \to \infty$, but that a normal central limit theorem will likely not be satisfied. While somewhat unstable, Kass and Raftery (1995) allude that the results are accurate enough for interpretation using the logarithm scale they provide, which is now discussed.

A scale is provided Kass and Raftery (1995), somewhat adapted from that provided by Jeffreys (1961), which is helpful in understanding the interpretation of a Bayes factor in the context of comparing two models, say $H_1$ to a null model $H_0$. This guidance is found in Table 5.2. Thus, when $2 \log(B_{10}) > 10$ there is very strong evidence in favor of $H_1$ over $H_0$, and so on.
While accepted as one of the best model selection criteria, the heavy computation required (or the level of approximation inherent in simpler solutions) occasionally precludes the use of Bayes factors in practice. Other model selection criteria are therefore also used abundantly.

5.2.2 Deviance Information Criterion

The deviance information criterion (DIC) has been an incredibly popular Bayesian model selection tool since its introduction by Spiegelhalter et al. (2002). Traditional model comparisons such as AIC and BIC require the number of parameters in the model as a penalty. One purpose of DIC is to act as a model comparison tool where the number of true parameters is challenging to define, such as with hierarchical models where parameters borrow strength from one another. Hence, to begin the definition of DIC, first define deviance and then the effective number of parameters.

Spiegelhalter et al. (2002) define the Bayesian deviance function $D(\theta)$ as

$$D(\theta) = -2 \log\{P(D|\theta)\} + 2 \log\{f(D)\},$$

(5.9)

where $p(D|\theta)$ is the likelihood of the Bayesian model for the data $D$ and $f(D)$ is a standardizing constant which is a function of $D$ alone, which therefore cancels out of any model comparisons and is hence not required. The effective number of parameters

<table>
<thead>
<tr>
<th>$2\log(B_{10})$</th>
<th>$B_{10}$</th>
<th>Evidence against $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 to 2</td>
<td>1 to 3</td>
<td>Not worth more than a bare mention</td>
</tr>
<tr>
<td>2 to 6</td>
<td>3 to 20</td>
<td>Positive</td>
</tr>
<tr>
<td>6 to 10</td>
<td>20 to 150</td>
<td>Strong</td>
</tr>
<tr>
<td>$&gt; 10$</td>
<td>$&gt; 150$</td>
<td>Very Strong</td>
</tr>
</tbody>
</table>

Table 5.2: Bayes Factor Model Comparison Scales (Kass and Raftery 1995)
is then defined as

\[ p_D = \overline{D(\theta)} - D(\bar{\theta}). \]  

(5.10)

The DIC is then

\[ DIC = D(\bar{\theta}) + 2p_D = \overline{D(\theta)} + p_D. \]  

(5.11)

Spiegelhalter et al. (2002) describe \( \overline{D(\theta)} \) as a measure of model adequacy or fit, and thus, observing the second expression in (5.11), DIC is a measure of fit penalized by the number of effective parameters.

Note that both \( \overline{D(\theta)} \) and \( D(\bar{\theta}) \) may be estimated directly from MCMC output (assuming that \( D(\theta) \) may be evaluated). This simplicity, yet flexibility, for a vast array of models are the crowning features of DIC. Many MCMC software packages report DIC routinely, which has also increased its popularity.

Numerous discussions concerning DIC and whether it is the best Bayesian model comparison tool, or even a good model comparison tool, are found in the literature. The original authors provide discussion in Spiegelhalter et al. (2014). The main criticisms listed therein are that \( p_D \) is not reparameterization invariant, there is a lack of consistency in identifying a “true model”, DIC is not based on the predictive distribution, and DIC has weak theoretical justification. They also discuss what has been done to improve DIC to address some of these concerns.

No defense of DIC as the best criterion is given by Spiegelhalter et al. (2002) or Spiegelhalter et al. (2014), and modest surprise about the popularity of the method is expressed. Despite some inadequacies, DIC remains a valuable Bayesian model
comparison tool and hence its use is explored here. As with AIC and BIC, smaller values of DIC are better.

5.2.3 Defining the Data D for Scale-Usage Models

For both Bayes factors and DIC, D plays an important role in defining the model selection criterion. As discussed by Celeux et al. (2006), with missing (or latent) data there are three primary considerations of what D can be: (1) the observed data, (2) the complete data, and (3) the conditional data. In the context of scale-usage models, the observed data is X, the ordinal scale ratings from each respondent. The complete data adds the latent variables and is X, Y. In the case of the decomposition model, Z is also added. The third paradigm, conditional data, is X|Y; this paradigm, however, is not meaningful for scale-usage models as discussed here because the Y are constrained such that \( P(X|Y) = 1 \). This then yields four main definitions for D under scale usage: (1) \( D_1 = \{X, Z\} \), (2) \( D_2 = \{X, Y\} \), (3) \( D_3 = \{X, Y, Z\} \), and finally (4) \( D_4 = X \).

The first definition, \( D_1 \), considers the data to be X combined with the latent variables Z, while integrating out the Y latent variables. This definition may suffer from lack of intuition because it is not very sensible to consider Z as data and not Y as the use of Z is purely for computational convenience. However, as in the MCMC algorithms, this method does have computational advantages. For notational convenience, let \( \theta \) denote the other parameters of the model. Then, the data likelihood which is required for both DIC and Bayes factor computations is

\[
P(D_1|\theta) = P(X, Z|\theta) = P(X|Z, \theta)P(Z|\theta) = \int_Y P(X, Y|Z, \theta)P(Z|\theta)\,dY,
\]
which is available in closed form because the joint distribution of \(X\) and \(Y\) given \(Z\) and \(\theta\) is the product of independent univariate normal distributions, constrained by \(X\), which are easily computed and integrated. The DIC and Bayes factor corresponding to \(D_1\) and its likelihood definition will be denoted DIC\(_1\) and BF\(_1\) respectively.

The second definition, \(D_2\), considers the data to be \(X\) and \(Y\) with \(Z\) integrated out. This approach is more intuitive as the role of \(Z\) is computational efficiency; if multivariate Gaussian integrals were accurately computable and simulation of truncated multivariate Gaussian random variables were efficient, the \(Z\) would not be included in the model. Therefore, it is sensible to integrate out \(Z\) when defining the data. In this case, the likelihood is

\[
P(D_2|\theta) = P(X, Y|\theta) = \int \! P(X, Y|Z, \theta)P(Z|\theta)dZ,
\]

which is available in closed form as the density function of a multivariate normal distribution with a non-diagonal covariance matrix. The corresponding DIC and Bayes factors are denoted DIC\(_2\) and BF\(_2\).

The third definition \(D_3\) integrates out neither \(Z\) nor \(Y\) but considers these together with \(X\) as the data. The likelihood is straightforward and requires no integration

\[
P(D_3|\theta) = P(X, Y, Z|\theta) = P(X, Y|Z, \theta)P(Z|\theta).
\]

The corresponding DIC and Bayes factors are DIC\(_3\) and BF\(_3\).

The final definition \(D_4\) integrates out both \(Y\) and \(Z\); i.e. \(Y\) and \(Z\) are considered parameters rather than data. While this consideration perhaps most adequately describes the modeling context, it is, however, not practical for many scale-usage applications because of the difficulty of accurately integrating the multivariate normal
distribution with a moderate number of dimensions. The likelihood would be

\[ P(D_4|\theta) = \int_Y \int_Z P(X, Y|Z, \theta)P(Z|\theta)dZdY = \int_Y P(X, Y|\theta)dY, \]

and, as discussed by Hans et al. (2012), these integrals are computationally expensive and are subject to a relatively large error. Hans et al. (2012) showed that this error compounded across many observations led to large cumulative MCMC error. This error will likely plague the approximations of both Bayes factors and DIC similarly. Thus, when the number of questions on a survey exceeds two or three, this definition of \( D \) will not be feasible; it is therefore no longer considered as a viable definition for \( D \) in the context of the scale-usage models in this chapter.

Note that with each of these data definitions \( D_1 \) through \( D_3 \), as discussed by Celeux et al. (2006), one of the only reasonable choices for \( D(\overline{\theta}) \) in estimating DIC is the likelihood evaluated at the posterior mode. The simulation-based estimate for the posterior mode, the Markov chain values for \( \theta \) which provide the highest posterior value among the sampled values, is therefore utilized for each of these (Celeux et al. 2006).

A simulation study will be conducted to evaluate each of these data definitions, along with the corresponding DIC and Bayes factors, to see if there is an optimal model selection criterion to be used to evaluate different scale-usage models. These model selection methods are also compared to an ad-hoc approach which uses intuition to select from Model 5.1 through Model 5.5.

### 5.2.4 Ad-Hoc Model Selection

Ad-hoc model selection is incredibly flexible based on the models which are being tested. Numerous ways to proceed are possible. While the methods discussed here
are specific to the scale-usage models under present consideration, the general principles are widely applicable for similarly-nested models. Before describing the overall strategy, the Bayesian analog to a confidence interval for estimation must be defined.

**Definition 5.1** (Carlin and Louis, 2009). A $100 \times (1 - \alpha)\%$ credible set for a parameter $\theta$ defined on $\Theta$ is a subset $C \subset \Theta$ such that

$$1 - \alpha \leq P(C|y) = \int_C p(\theta|y) d\theta.$$ 

If $C$ is an interval, it is referred to as a credible interval.

There are two most common methods for creating a credible set. An equal-tail credible interval is the interval $[c_1, c_2]$ such that

$$P(\theta \leq c_1|y) \leq \frac{\alpha}{2} \quad \text{and} \quad P(\theta \geq c_2|y) \leq \frac{\alpha}{2}.$$ 

To compute a marginal equal-tail credible interval from simulations, simply take the $\alpha/2$ and $1 - (\alpha/2)$ sample quantiles.

The other common method for creating a credible set $C$ is obtaining the highest posterior density (HPD) credible set, which satisfies

$$C = \{\theta \in \Theta : p(\theta|y) \geq k(\alpha)\}$$ 

where $k(\alpha)$ is the largest constant such that

$$P(C|y) \geq 1 - \alpha,$$ 

as described by (Carlin and Louis 2009). This credible set groups together values of $\theta$ which have the highest posterior density. When distributions are non-symmetric, this interval better describes the regions of high posterior probability.
credible set is an interval, but it is possible to have more complicated circumstances, e.g. bimodal or mixture distributions. One disadvantage to the HPD credible set, however, is additional computation to find \( k(\alpha) \) and the corresponding \( \theta \) values. It is for this reason that equal-tail credible intervals, which are trivial to compute, are used here unless otherwise specified. Given the ability to construct credible intervals, the overall ad-hoc strategy is now described.

Consider the main difference between the demographic-specific Models 5.1 through 5.3 and the non-demographic-specific Models 5.4 and 5.5. Perhaps the most important difference is that the distribution of \( \tau \) has demographic-specific mean \( \theta_\ell \), rather than mean zero. Based on this fact, one way to choose between Models 5.1 through 5.3 and Models 5.4 and 5.5 is to fit Model 5.1 and create a credible interval for each \( \theta_\ell \). If more than a certain number of these intervals does not include zero (around 5% or so), choose one of the demographic models; otherwise, choose one of the non-demographic models. Consideration may also be given to the different \( \sigma^2_\tau \) parameters for each demographic, but comparison of the demographic-level means will likely suffice for deciphering between demographic and non-demographic models.

Next, given that a model is demographic-specific, a choice must be made between Models 5.1, 5.2, and 5.3. One clear distinction of Model 5.2 is that there are demographic-specific \( \sigma^2_\tau \) parameters. Thus, one indicator that this model is preferred is if the demographics are estimated to have different \( \sigma^2_\tau \) parameters. Therefore, fit Model 5.2, and consider all the pairs of demographic scale parameters \((\sigma^2_{\ell_1}, \sigma^2_{\ell_2})\) for \( \ell_1 \neq \ell_2 \) with \( \ell_1, \ell_2 \in \{1, \ldots, L\} \). Count the number of these pairs such that the associated credible intervals do not intersect, which implies that the demographics differ
with respect to $\sigma_i^2$. If more than a certain percent of these intervals do not overlap, select Model 5.2. Otherwise, continue selecting between Model 5.1 and Model 5.3.

To select between Model 5.1 and Model 5.3, a determination must be made as to whether the individual $\sigma_i^2$ parameters are close to one, as they are defined to be in Model 5.3, whereas Model 5.1 allows them to take a broad range of values. It will often be the case, under either model, that the credible interval contains one due to the relatively small amount of data estimating $\sigma_i^2$; therefore, simply tallying the number of credible intervals containing one is not sufficient. Rather, consider the distance of the posterior median for each $\sigma_i^2$ parameter from one. Average the absolute value of these to obtain the average absolute distance of the posterior median of $\sigma_i^2$ from 1 across all individuals. If this is less than some threshold, choose Model 5.3; if it is greater than the threshold, select Model 5.1.

The remaining decision is then among non-demographic models, between Model 5.4 and Model 5.5. Perhaps the easiest difference to detect is whether or not it is plausible that each of the $\tau_i$ is equal to 0. Therefore, fit Model 5.4 and compute the credible interval for each $\tau_i$. If a certain proportion of these intervals contain zero, select Model 5.5; otherwise, if a large percent of the intervals differ from zero, select Model 5.4.

It is important to note that this approach has not been optimized. Many other possibilities abound, but this approach is evaluated and compared to DIC and Bayes factors as to its accuracy in model selection in the simulation study.
5.3 Simulation Study: Models for Demographic Effects and Scale-Usage

A simulation study is conducted to explore the utility of the demographic-specific scale-usage models. The three main purposes of this simulation study are as follows:

1. Fit the demographic-specific models in a variety of settings to ascertain that the Markov chains are converging appropriately and to identify any potential concerns (e.g. identifiability problems, parameters that do not converge, etc.).

2. Evaluate DIC, Bayes factors, and the ad-hoc model selection approaches with the different data definitions to identify guidance for future model selection.

3. Ensure that the models are accurate, to some degree, at estimating the true parameters.

The setup of the simulation study is now discussed, followed by an exploration of the output of the Markov chains to identify any concerns. The results of the model selection procedure are then provided and discussed, followed by an evaluation of how well the models capture the true parameters.

5.3.1 Simulation Setup

The setup governing the data generation for the simulation study is now provided. Each dataset contains the responses of 100 subjects to eight questions which are evaluated on a five-point ordinal scale. For data generation models including demographics, each had five demographic levels, \( L = 5 \), with 20 respondents per demographic, \( n_\ell = 20 \) for \( \ell = 1, \ldots, L \).
In order to provide a somewhat realistic relationship among the questions, the first eight dimensions of the posterior mean estimates for \( \mu \) (shifted to have mean zero) and \( \Sigma \) from Rossi et al. (2001) were adopted as the truth for all simulation datasets. For reference,

\[
\mu = \begin{pmatrix} -0.31 & -0.58 & -0.27 & -0.74 & -0.28 & 0.65 & 0.76 & 0.76 \end{pmatrix}^T
\]

and

\[
\Sigma = \begin{pmatrix}
4.13 & 0.31 & 0.25 & 0.55 & 0.29 & 0.39 & 0.15 & 0.05 \\
1.50 & 5.70 & 0.65 & 0.61 & 0.16 & -0.09 & -0.11 & -0.11 \\
1.33 & 4.07 & 6.93 & 0.53 & 0.13 & -0.08 & -0.05 & -0.03 \\
2.79 & 3.70 & 3.49 & 6.34 & 0.31 & 0.29 & 0.04 & 0.05 \\
1.36 & 0.87 & 0.82 & 1.81 & 5.44 & 0.38 & 0.22 & 0.21 \\
1.55 & -0.42 & -0.39 & 1.42 & 1.73 & 3.89 & 0.20 & 0.12 \\
0.77 & -0.67 & -0.34 & 0.43 & 1.31 & 1.00 & 6.49 & 0.49 \\
0.24 & -0.60 & -0.15 & 0.26 & 1.10 & 0.56 & 2.84 & 5.29 \\
\end{pmatrix},
\]

where the upper-triangle italicized values show the correlations rather than the covariances. The true cutoff points for each model were selected to be

\[ c = \begin{pmatrix} -20 & -6 & -2 & 2 & 6 & 20 \end{pmatrix}, \]

recalling that with Model 5.1 through Model 5.5, the smallest and largest cutpoints are chosen to be finite to ensure geometric ergodicity.

For Model 5.1 through Model 5.3, demographic specific \( \theta \) and \( \sigma_\tau^2 \) values are required. Six settings were explored, and are provided in Table 5.3. These settings were selected to represent an array of possibilities for demographic variability. For example, in Setting 1, each demographic level has an increasing mean but decreasing variance, a relationship which is often observed with scale-usage models. Settings 2 and 5 illustrate the situation when two demographics are starkly different from the rest; Settings 3 and 6 show a similar effect, though with less drastic differences. The
\( \tau_i \) parameter for each individual is then randomly generated according to the normal distribution with the correctly specified mean and variance for the demographic setting shown in Table 5.3. Note that for Model 5.4, the mean of the distribution of each \( \tau_i \) parameter is 0 with standard deviation 6, which was chosen to maintain a similar level of variability as the parameter settings shown in Table 5.3.

Table 5.3: Simulation Study Settings for \( \theta \) and \( \sigma^{2}_{\tau} \)

<table>
<thead>
<tr>
<th>Setting</th>
<th>( \theta_1 )</th>
<th>( \theta_2 )</th>
<th>( \theta_3 )</th>
<th>( \theta_4 )</th>
<th>( \theta_5 )</th>
<th>( \sigma^{2}_{\tau,1} )</th>
<th>( \sigma^{2}_{\tau,2} )</th>
<th>( \sigma^{2}_{\tau,3} )</th>
<th>( \sigma^{2}_{\tau,4} )</th>
<th>( \sigma^{2}_{\tau,5} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-8</td>
<td>-4</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0.67</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>-8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0.67</td>
<td>0.5</td>
</tr>
<tr>
<td>3</td>
<td>-4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
<td>0.67</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>-8</td>
<td>-4</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>-8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>-4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

For Model 5.1, Model 5.3, and Model 5.4, the \( \sigma^{2}_{\tau} \) parameters were drawn independently from an Inverse Gamma(5.5 - 1) distribution, which has mean equal to 1. For Model 5.2, the demographic-specific \( \sigma^{2}_{\tau} \) parameters were set to be 2, 1.5, 1, 0.67, and 0.5 for \( \ell = 1, \ldots, 5 \) respectively.

This produced a total of 20 different models under which the data were generated: six settings for each of the first three models, plus Model 5.4 and Model 5.5. For each of these 20 settings, 200 datasets were generated, resulting in 4,000 raw datasets. Each of these datasets was then fit with a Markov chain stationary for each of Models 5.1 through 5.5 for variable selection. The Markov chains were each run with 100,000 iterations after discarding a preliminary burn-in of 10,000 iterations beginning with initial values equal to the true parameters generating the dataset.
5.3.2 MCMC Output and Identifiability

The first key purpose of this simulation study is to evaluate the demographic Models 5.1 through 5.3 to ensure that the additional model structure beyond Model 5.4 to account for demographics does not pose any sampling issues. Though the chains are geometrically ergodic, visual inspection is useful to determine whether the chain is behaving as expected. Trace plots of the output for $\tau_1, \mu_1, \text{ and } \theta_1$ using the Markov chain stationary for Model 5.1 with data generated under Setting 1 of the same model are shown in Figure 5.1 on the left. A problem with convergence, likely due to non-identifiability, is clearly present. Upon inspection of the model, note that

$$\mu + \tau_1 \mathbf{1} = (\mu + \bar{\tau}) \mathbf{1} + (\tau_1 - \bar{\tau}) \mathbf{1},$$

which implies a lack of identifiability. This lack of identifiability doesn’t adversely effect the sampling of $Y$ and other parameters because the sum is identifiable. Comparing $\tau$ with $\theta$, note that

$$\tau_i - \theta_{d_i} = (\tau_i - \bar{\tau}) - (\theta_{d_i} - \bar{\tau}).$$

Therefore, for each iteration, compute transformed parameters

$$\mu^* = \mu + \bar{\tau} \mathbf{1}$$

$$\tau_i^* = \tau_i - \bar{\tau}, \text{ and}$$

$$\theta_{\ell}^* = \theta_{\ell} - \bar{\tau}.$$ 

Trace plots of these transformed parameters are provided on the right side of Figure 5.1. The convergence is largely improved by considering the transformed variables. While the original variables are used in the simulation, the transformed variables are used for inference.
Figure 5.1: Original (left) and transformed (right) trace plots for $\tau$, $\theta$, and $\mu$ parameters.
Trace plots for several of the other parameters, $\sigma^2_1, \Sigma_{1,1}, \sigma^2_{\tau_1}, Y_{1,1}, c_1$, and $c_4$ are provided in Figure 5.2. None of these plots indicate problems with convergence, and no adjustments or transformations are necessary.

With successful visual convergence checks and having shown geometric ergodicity, model selection is now explored.

### 5.3.3 Model Selection

The second key purpose of this simulation study is to evaluate the ability of Bayes factors, DIC, and an ad-hoc approach to select the correct model. The proportion of correct model identifications using the DIC and Bayes factor criteria based on the three considerations of the data $D$ are presented in Table 5.4. The best model selection criteria for each model and setting is provided on the right of the table. The first important note from these results is that Model 5.2 and Model 5.5 are never selected correctly. The inability to select these models using DIC and Bayes factors is present throughout the simulation study. Possible reasons for this are that these models are restricted cases of other models: Model 5.2 is Model 5.3 restricted to the same scale coefficient across the entire demographic, and Model 5.5 is Model 5.4 restricted to no scale-usage heterogeneity. The preference of the non-restricted models suggests that these restrictions are not optimal, even when the data were generated under such settings. Also, particularly in the framework of DIC, the improved fit outweighs the additional parameters required for the non-restricted models.

Another key observation from the results in Table 5.4 is that DIC outperforms Bayes factors in every setting where a best method is present. Specifically, DIC$_1$ is most frequently the best criteria. This suggests that if consistent results are not
Figure 5.2: Trace plots for $\sigma_1^2$, $\Sigma_{1,1}$, $\sigma_{\tau_1}^2$, $Y_{1,1}$, $c_1$, and $c_4$. No convergence problems are detected.
present among the model selection criteria, perhaps the DIC criteria, particularly DIC$_1$, should be given additional weight beyond the Bayes factors. In most of the settings, however, Bayes factors perform \textit{nearly} as well as DIC. Therefore, Bayes factors cannot be completely counted out of the model selection process.

Table 5.4: Percent of Correctly Selected Models using DIC and Bayes Factors

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>DIC$_1$</th>
<th>DIC$_2$</th>
<th>DIC$_3$</th>
<th>BF$_1$</th>
<th>BF$_2$</th>
<th>BF$_3$</th>
<th>Best</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.575</td>
<td>0.550</td>
<td>0.620</td>
<td>0.465</td>
<td>0.440</td>
<td>0.530</td>
<td>DIC$_3$</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.160</td>
<td>0.055</td>
<td>0.125</td>
<td>0.040</td>
<td>0.040</td>
<td>0.045</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.345</td>
<td>0.235</td>
<td>0.310</td>
<td>0.245</td>
<td>0.250</td>
<td>0.290</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.555</td>
<td>0.635</td>
<td>0.630</td>
<td>0.565</td>
<td>0.525</td>
<td>0.495</td>
<td>DIC$_2$</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.130</td>
<td>0.020</td>
<td>0.110</td>
<td>0.040</td>
<td>0.025</td>
<td>0.050</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.405</td>
<td>0.340</td>
<td>0.360</td>
<td>0.305</td>
<td>0.320</td>
<td>0.350</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.000</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>DIC$_2$</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.065</td>
<td>0.025</td>
<td>0.010</td>
<td>0.005</td>
<td>0.015</td>
<td>0.010</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.010</td>
<td>0.000</td>
<td>0.015</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>DIC$_3$</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.035</td>
<td>0.005</td>
<td>0.020</td>
<td>0.005</td>
<td>0.005</td>
<td>0.000</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.025</td>
<td>0.035</td>
<td>0.030</td>
<td>0.025</td>
<td>0.005</td>
<td>0.015</td>
<td>DIC$_2$</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.005</td>
<td>0.000</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>DIC$_1$&amp;$3$</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.005</td>
<td>0.020</td>
<td>0.020</td>
<td>0.010</td>
<td>0.000</td>
<td>0.010</td>
<td>DIC$_2$&amp;$3$</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.585</td>
<td>0.525</td>
<td>0.545</td>
<td>0.520</td>
<td>0.535</td>
<td>0.510</td>
<td>DIC$_1$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>None</td>
</tr>
</tbody>
</table>

While Table 5.4 described the proportion of times the correct model was selected, Table 5.5 illustrates the full distribution of models chosen to determine if there is a systematically favored model or models when an incorrect choice is made. The most commonly favored criterion from Table 5.4, DIC$_1$, is used to produce Table
5.5, though similar patterns are attributed to the other criteria. The most commonly selected models are Model 5.1 and Model 5.4. Models 5.2 and 5.5 were never chosen here and Model 5.3 is rarely chosen. This again points to the conclusion that these models which restrict Model 5.1 and Model 5.4 are perhaps too restrictive. Another observation is that the balance between Model 5.1 and Model 5.4 is somewhat based on the setting. For example, Settings 2 and 5 strongly favor Model 5.4. Similarly, for Settings 3 and 6, Model 5.4 is favored a majority of the time. This phenomenon is likely due to the fact that three of the five \( \theta \) values for these four settings are equal to 0, and Model 5.4 is similar to Model 5.1 when all values of \( \theta \) are zero. Conversely, Settings 1 and 4, where four of five \( \theta \) values differ from 0, favor Model 5.1. This leads to the conclusion that Model 5.4 is preferred when less demographic variability is present, but Model 5.1 is preferred when more demographic variability is present, which is precisely the desired property. While the model selection criteria may not perfectly select models, they favor models based on the level of demographic variability, and the least constrained model is typically preferred.

While these results are promising, if the goal is to correctly identify the original data model, DIC and Bayes factors certainly leave room for improvement. Perhaps an ad-hoc approach can achieve higher accuracy. The specific ad-hoc approach was outlined generally earlier, and is now provided in detail. First, a decision between the non-demographic models and the demographic models is made by fitting Model 5.1 and computing credible intervals for each \( \theta \). If two or more of the five intervals do not contain 0, conclude a demographic-specific model; otherwise, choose a non-demographic model. Then, among demographic-specific models, conclude Model 5.2
Table 5.5: Distribution of Models Chosen Using DIC\(_1\) for Each Model Setting

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
<th>Model 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.575</td>
<td>0.000</td>
<td>0.005</td>
<td>0.420</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.160</td>
<td>0.000</td>
<td>0.000</td>
<td>0.840</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.345</td>
<td>0.000</td>
<td>0.015</td>
<td>0.640</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.555</td>
<td>0.000</td>
<td>0.000</td>
<td>0.445</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.130</td>
<td>0.000</td>
<td>0.005</td>
<td>0.865</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.405</td>
<td>0.000</td>
<td>0.010</td>
<td>0.585</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.500</td>
<td>0.000</td>
<td>0.035</td>
<td>0.465</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.135</td>
<td>0.000</td>
<td>0.005</td>
<td>0.860</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.285</td>
<td>0.000</td>
<td>0.020</td>
<td>0.695</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.610</td>
<td>0.000</td>
<td>0.045</td>
<td>0.345</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.160</td>
<td>0.000</td>
<td>0.000</td>
<td>0.840</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.370</td>
<td>0.000</td>
<td>0.030</td>
<td>0.600</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.635</td>
<td>0.000</td>
<td>0.065</td>
<td>0.300</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.140</td>
<td>0.000</td>
<td>0.010</td>
<td>0.850</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.360</td>
<td>0.000</td>
<td>0.035</td>
<td>0.605</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.585</td>
<td>0.000</td>
<td>0.025</td>
<td>0.390</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.090</td>
<td>0.000</td>
<td>0.005</td>
<td>0.905</td>
<td>0.000</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.320</td>
<td>0.000</td>
<td>0.005</td>
<td>0.675</td>
<td>0.000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.405</td>
<td>0.000</td>
<td>0.010</td>
<td>0.585</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.305</td>
<td>0.000</td>
<td>0.045</td>
<td>0.650</td>
<td>0.000</td>
</tr>
</tbody>
</table>
if more than two of the pairwise-comparisons of the credible intervals of $\sigma_i^2$ parameters from fitting Model 5.2 overlap. Finally, if the average absolute distance of the posterior median from one when fitting Model 5.1 is less than 0.2, conclude Model 5.3; otherwise, conclude Model 5.1. Among non-demographic models, if the proportion of credible intervals for the $\tau_i$ parameters including 0 exceeds 0.5, conclude Model 5.5; otherwise conclude Model 5.4.

The proportions of choosing each of the five models based on this ad-hoc model selection procedure are shown in Table 5.6 for each model and setting. The column second from the right provides the proportion of correct separations into demographic and non-demographic models. The ad-hoc method is incredibly accurate at such distinctions, only erring occasionally in favor of a demographic model when a non-demographic model generated the data. The rightmost column provides the proportion of correct decisions for each setting using the full ad-hoc method. The distinction between Model 5.4 and Model 5.5 is incredibly accurate, with no mistakes for the generated data given the correct demographic/non-demographic decision was made. It is a greater challenge to distinguish between Models 5.1, 5.2, and 5.3, though the correct decision is typically made well over a majority of the time. Comparing these results to those in Table 5.5, it is clear that an ad-hoc strategy provides superior accuracy at selecting the data-generation model compared to DIC and Bayes factors. Thus, an ad-hoc approach is important to any model selection process for scale-usage models. The typical model-selection criteria may not provide results with the same accuracy.

Having explored the model-selection process for these models, the final purpose of this simulation study, assessing parameter estimation, is now addressed.
Table 5.6: Ad-Hoc Model Selection Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Selected Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.565</td>
<td>0.095</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.605</td>
<td>0.160</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.830</td>
<td>0.075</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.640</td>
<td>0.130</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.610</td>
<td>0.120</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.785</td>
<td>0.120</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.115</td>
<td>0.505</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.065</td>
<td>0.730</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.050</td>
<td>0.915</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.085</td>
<td>0.500</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.100</td>
<td>0.730</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.040</td>
<td>0.920</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.090</td>
<td>0.055</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.100</td>
<td>0.100</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.240</td>
<td>0.010</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.090</td>
<td>0.070</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.110</td>
<td>0.130</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.340</td>
<td>0.015</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.010</td>
<td>0.000</td>
</tr>
</tbody>
</table>
5.3.4 Estimation Accuracy

The third key purpose of this simulation study is to evaluate the accuracy of parameter estimates. Coverage for each of the parameters of interest is now discussed. Coverage is defined as the proportion of 95% credible intervals for a parameter which contain the true value of the parameter. In an ideal setting, the coverage should be close to 0.95. In the case of scale-usage models, however, smaller coverage proportions are not alarming as only ordered categories are observed. With richer data, a larger sample size, and more questions per respondent, coverage probabilities would increase toward 0.95.

Table 5.7 shows the coverage probabilities of the $\tau$ parameters for each of the model settings (rows) and each of the fit models (columns). While these coverage probabilities are a bit lower than desired, ranging from 0.60 to 0.88, there are only $M = 8$ questions per respondent from which to estimate $\tau$, beyond borrowing strength within a demographic. Thus, smaller coverage probabilities are expected. One positive indication that the structure of the $\tau$ parameters is being captured accurately is by considering the correlation between the posterior means and the true values of $\tau$. An illustration of why this is useful is shown in Figure 5.3. There is a strong, linear, positive association between the true values and the posterior means. Correlation is therefore a useful metric for comparison between the posterior distributions and the truth. Table 5.8 provides the average correlation for the $\tau$ parameters for each of the model settings fit with each of the models. Each correlation is quite large, indicating that the structure of the $\tau$ parameters is being preserved despite lower than ideal coverage probabilities.
Table 5.7: Coverage of 95% Credible Intervals for \( \tau \) Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.65</td>
<td>0.78</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.75</td>
<td>0.81</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.69</td>
<td>0.80</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.65</td>
<td>0.78</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.74</td>
<td>0.80</td>
<td>0.72</td>
<td>0.68</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.69</td>
<td>0.80</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.64</td>
<td>0.79</td>
<td>0.66</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.69</td>
<td>0.81</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.68</td>
<td>0.80</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.62</td>
<td>0.77</td>
<td>0.65</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.69</td>
<td>0.82</td>
<td>0.68</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.67</td>
<td>0.80</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.61</td>
<td>0.75</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.73</td>
<td>0.80</td>
<td>0.72</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.67</td>
<td>0.79</td>
<td>0.67</td>
<td>0.65</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.60</td>
<td>0.75</td>
<td>0.63</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.75</td>
<td>0.82</td>
<td>0.73</td>
<td>0.69</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.66</td>
<td>0.78</td>
<td>0.66</td>
<td>0.65</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.71</td>
<td>0.77</td>
<td>0.71</td>
<td>0.70</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.69</td>
<td>0.88</td>
<td>0.69</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Table 5.8: Correlation Between Posterior Means and Truth for $\tau$ Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.980</td>
<td>0.980</td>
<td>0.979</td>
<td>0.967</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.973</td>
<td>0.975</td>
<td>0.973</td>
<td>0.959</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.920</td>
<td>0.929</td>
<td>0.920</td>
<td>0.900</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.981</td>
<td>0.981</td>
<td>0.980</td>
<td>0.967</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.974</td>
<td>0.976</td>
<td>0.973</td>
<td>0.959</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.919</td>
<td>0.928</td>
<td>0.919</td>
<td>0.898</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.975</td>
<td>0.979</td>
<td>0.974</td>
<td>0.963</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.967</td>
<td>0.973</td>
<td>0.967</td>
<td>0.954</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.912</td>
<td>0.922</td>
<td>0.912</td>
<td>0.890</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.976</td>
<td>0.980</td>
<td>0.975</td>
<td>0.963</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.968</td>
<td>0.975</td>
<td>0.967</td>
<td>0.954</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.910</td>
<td>0.922</td>
<td>0.909</td>
<td>0.886</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.979</td>
<td>0.981</td>
<td>0.979</td>
<td>0.967</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.974</td>
<td>0.976</td>
<td>0.973</td>
<td>0.960</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.919</td>
<td>0.929</td>
<td>0.920</td>
<td>0.900</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.980</td>
<td>0.982</td>
<td>0.980</td>
<td>0.967</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.975</td>
<td>0.978</td>
<td>0.974</td>
<td>0.960</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.914</td>
<td>0.925</td>
<td>0.915</td>
<td>0.894</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.962</td>
<td>0.963</td>
<td>0.963</td>
<td>0.963</td>
</tr>
</tbody>
</table>
Table 5.9 shows the coverage probabilities for the $\theta$ parameters. The coverage appears to be closely related to the setting. For example, Settings 1 and 4 where $\theta = (-8, -4, 0, 4, 9)$, fitted with Model 5.1, have much lower coverage than for the other settings. As will be later seen in Table 5.15, the cutpoint coverage probabilities are also low for these settings. Perhaps some subtle non-identifiability between the cutpoints and the $\theta$ parameters is causing this. Or perhaps, as these settings have the most non-zero $\theta$ values, they are simply more difficult to estimate. In any case, the other settings have fairly good coverage probabilities. The correlations between the posterior means and truth are provided in Table 5.10. These correlations are all incredibly strong, indicating that the overall structure of $\theta$ is being captured.

Table 5.11 shows the coverage probabilities for the $\mu$ parameter. With a few exceptions, such as Settings 1 and 4 with true Model 5.2 fit by Models 5.1, 5.3, and
Table 5.9: Coverage of 95% Credible Intervals for $\theta$ Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.63</td>
<td>0.83</td>
<td>0.73</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.89</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.92</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.62</td>
<td>0.85</td>
<td>0.71</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.85</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.94</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.61</td>
<td>0.85</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.79</td>
<td>0.92</td>
<td>0.78</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.92</td>
<td>0.94</td>
<td>0.92</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.59</td>
<td>0.81</td>
<td>0.64</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.78</td>
<td>0.93</td>
<td>0.76</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.93</td>
<td>0.95</td>
<td>0.93</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.60</td>
<td>0.79</td>
<td>0.68</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.91</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.54</td>
<td>0.78</td>
<td>0.63</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.93</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.94</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>
Table 5.10: Correlation Between Posterior Means and Truth for $\theta$ Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.991</td>
<td>0.991</td>
<td>0.992</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.997</td>
<td>0.997</td>
<td>0.998</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.994</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.994</td>
<td>0.997</td>
<td>0.994</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.993</td>
<td>0.997</td>
<td>0.993</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.989</td>
<td>0.992</td>
<td>0.989</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.994</td>
<td>0.997</td>
<td>0.993</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.992</td>
<td>0.997</td>
<td>0.992</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.991</td>
<td>0.992</td>
<td>0.990</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.992</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.997</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.993</td>
<td>0.993</td>
<td>0.993</td>
</tr>
</tbody>
</table>
5.4, the coverage probabilities are quite good for \( \mu \). The models are therefore able to capture the mean structure inherent between the questions. This is further confirmed in Table 5.12 with strong correlations between true parameters and posterior means. Thus, \( \mu \) is captured quite well by each of these models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
<th>Model 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.90</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.92</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.89</td>
<td>0.91</td>
<td>0.90</td>
<td>0.89</td>
<td>0.99</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.91</td>
<td>0.93</td>
<td>0.91</td>
<td>0.92</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.92</td>
<td>0.93</td>
<td>0.91</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.39</td>
<td>0.89</td>
<td>0.33</td>
<td>0.37</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.69</td>
<td>0.90</td>
<td>0.67</td>
<td>0.71</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.82</td>
<td>0.90</td>
<td>0.80</td>
<td>0.78</td>
<td>0.96</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.41</td>
<td>0.90</td>
<td>0.34</td>
<td>0.42</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.66</td>
<td>0.91</td>
<td>0.65</td>
<td>0.74</td>
<td>0.99</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.82</td>
<td>0.92</td>
<td>0.82</td>
<td>0.81</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.89</td>
<td>0.93</td>
<td>0.90</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.91</td>
<td>0.92</td>
<td>0.90</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.91</td>
<td>0.93</td>
<td>0.91</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.88</td>
<td>0.92</td>
<td>0.89</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.91</td>
<td>0.93</td>
<td>0.91</td>
<td>0.91</td>
<td>1.00</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.89</td>
<td>0.91</td>
<td>0.90</td>
<td>0.90</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.89</td>
<td>0.91</td>
<td>0.89</td>
<td>0.88</td>
<td>1.00</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.89</td>
<td>0.92</td>
<td>0.91</td>
<td>0.89</td>
<td>0.93</td>
</tr>
</tbody>
</table>

The coverage probabilities for the \( \sigma^2 \) parameters fit by Models 5.1, 5.2, and 5.3 are shown in Table 5.13. The coverage probabilities are reasonable, particularly considering the complex relationship between the \( \sigma^2 \) parameters and the \( \sigma^2_i \) parameters. Each contributes additional variation to \( \mathbf{Y} \), one in an additive sense, and the other
Table 5.12: Correlation Between Posterior Means and Truth for $\mu$

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
<th>Model 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.939</td>
<td>0.934</td>
<td>0.936</td>
<td>0.939</td>
<td>0.936</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.939</td>
<td>0.932</td>
<td>0.934</td>
<td>0.941</td>
<td>0.933</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.946</td>
<td>0.943</td>
<td>0.942</td>
<td>0.945</td>
<td>0.941</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.937</td>
<td>0.931</td>
<td>0.934</td>
<td>0.938</td>
<td>0.936</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.939</td>
<td>0.934</td>
<td>0.935</td>
<td>0.940</td>
<td>0.935</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.944</td>
<td>0.941</td>
<td>0.942</td>
<td>0.944</td>
<td>0.942</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.934</td>
<td>0.936</td>
<td>0.933</td>
<td>0.935</td>
<td>0.934</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.932</td>
<td>0.932</td>
<td>0.929</td>
<td>0.934</td>
<td>0.929</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.939</td>
<td>0.945</td>
<td>0.935</td>
<td>0.939</td>
<td>0.934</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.934</td>
<td>0.938</td>
<td>0.932</td>
<td>0.935</td>
<td>0.933</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.934</td>
<td>0.938</td>
<td>0.930</td>
<td>0.935</td>
<td>0.930</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.945</td>
<td>0.949</td>
<td>0.941</td>
<td>0.945</td>
<td>0.941</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.931</td>
<td>0.928</td>
<td>0.932</td>
<td>0.933</td>
<td>0.933</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.936</td>
<td>0.933</td>
<td>0.937</td>
<td>0.937</td>
<td>0.938</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.944</td>
<td>0.945</td>
<td>0.946</td>
<td>0.944</td>
<td>0.946</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.933</td>
<td>0.932</td>
<td>0.936</td>
<td>0.935</td>
<td>0.936</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.933</td>
<td>0.930</td>
<td>0.933</td>
<td>0.935</td>
<td>0.932</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.946</td>
<td>0.947</td>
<td>0.947</td>
<td>0.946</td>
<td>0.948</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.936</td>
<td>0.932</td>
<td>0.934</td>
<td>0.936</td>
<td>0.932</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.945</td>
<td>0.946</td>
<td>0.946</td>
<td>0.945</td>
<td>0.946</td>
</tr>
</tbody>
</table>
in a multiplicative sense. This complex relationship is a likely culprit for less than ideal coverage probabilities.

Table 5.13: Coverage of 95% Credible Intervals for $\sigma_i^2$ Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.74</td>
<td>0.81</td>
<td>0.71</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.64</td>
<td>0.77</td>
<td>0.63</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.53</td>
<td>0.72</td>
<td>0.53</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.81</td>
<td>0.82</td>
<td>0.74</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.63</td>
<td>0.78</td>
<td>0.64</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.60</td>
<td>0.73</td>
<td>0.59</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.71</td>
<td>0.85</td>
<td>0.72</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.63</td>
<td>0.76</td>
<td>0.66</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.56</td>
<td>0.70</td>
<td>0.57</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.76</td>
<td>0.83</td>
<td>0.73</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.59</td>
<td>0.74</td>
<td>0.58</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.49</td>
<td>0.70</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.74</td>
<td>0.82</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.59</td>
<td>0.74</td>
<td>0.58</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.48</td>
<td>0.67</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.80</td>
<td>0.84</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.60</td>
<td>0.74</td>
<td>0.59</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.50</td>
<td>0.69</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.92</td>
<td>0.97</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Relatedly, Table 5.14 shows the coverage probabilities for the $\sigma_i^2$ parameters. Note that for Model 5.2, there is a $\sigma_i^2$ parameter common to each demographic. This appears to adversely affect the ability of the model to capture the true parameter. The coverage probabilities for Models 1 and 4 are quite good. In looking at the credible intervals for these parameters, they are often quite large, aiding the high coverage, but also resulting in imprecise estimates. This is largely due to the $\sigma_i^2$ parameters being influenced by relatively few questions per respondent. The prior
distribution then carries substantial weight. The difficulty of precisely estimating $\sigma_i^2$ is further exacerbated by the fact that only ordinal categories are observed. Larger scales (e.g. a ten-point scale instead of a five-point scale) and more questions per respondent can attenuate this problem, though for many inference questions, precise estimation of $\sigma_i^2$ may not be necessary.

Table 5.14: Coverage of 95% Credible Intervals for $\sigma^2$ Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.953</td>
<td>0.591</td>
<td>0.954</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.955</td>
<td>0.546</td>
<td>0.953</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.955</td>
<td>0.490</td>
<td>0.954</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.954</td>
<td>0.587</td>
<td>0.957</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.954</td>
<td>0.545</td>
<td>0.950</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.956</td>
<td>0.484</td>
<td>0.956</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.962</td>
<td>0.468</td>
<td>0.982</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.968</td>
<td>0.416</td>
<td>0.985</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.947</td>
<td>0.349</td>
<td>0.944</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.957</td>
<td>0.465</td>
<td>0.982</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.961</td>
<td>0.415</td>
<td>0.984</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.942</td>
<td>0.342</td>
<td>0.944</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.995</td>
<td>0.840</td>
<td>0.996</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.996</td>
<td>0.828</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.996</td>
<td>0.937</td>
<td>0.996</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.995</td>
<td>0.847</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.996</td>
<td>0.856</td>
<td>0.997</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.996</td>
<td>0.932</td>
<td>0.996</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.950</td>
<td>0.516</td>
<td>0.950</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.995</td>
<td>0.929</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Finally, the coverage probabilities of the free cutpoint parameters are provided. These coverage probabilities are largely lower than ideal. However, when examining the credible intervals, many of them are not far from covering the truth. Furthermore,
the model can often adjust for slightly different cutpoints by making minute adjustments to other parameters, such as $\theta$ or $\tau$. Therefore, lower coverage probabilities here are not alarming; the models should be able to adjust as needed to accommodate the latent variable structure despite the slight error in estimating cutpoint parameters.

Table 5.15: Coverage of 95% Credible Intervals for Free Cutpoint Parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Setting</th>
<th>Model 5.1</th>
<th>Model 5.2</th>
<th>Model 5.3</th>
<th>Model 5.4</th>
<th>Model 5.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.33</td>
<td>0.71</td>
<td>0.39</td>
<td>0.30</td>
<td>0.15</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.73</td>
<td>0.74</td>
<td>0.55</td>
<td>0.20</td>
<td>0.09</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.69</td>
<td>0.80</td>
<td>0.61</td>
<td>0.58</td>
<td>0.38</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>0.36</td>
<td>0.73</td>
<td>0.41</td>
<td>0.32</td>
<td>0.19</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>0.72</td>
<td>0.69</td>
<td>0.54</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>0.65</td>
<td>0.82</td>
<td>0.58</td>
<td>0.57</td>
<td>0.40</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.45</td>
<td>0.75</td>
<td>0.45</td>
<td>0.44</td>
<td>0.38</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.61</td>
<td>0.88</td>
<td>0.50</td>
<td>0.18</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.56</td>
<td>0.91</td>
<td>0.50</td>
<td>0.50</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.40</td>
<td>0.70</td>
<td>0.43</td>
<td>0.37</td>
<td>0.33</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>0.65</td>
<td>0.90</td>
<td>0.54</td>
<td>0.20</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>0.54</td>
<td>0.95</td>
<td>0.49</td>
<td>0.53</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.28</td>
<td>0.71</td>
<td>0.33</td>
<td>0.27</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.88</td>
<td>0.89</td>
<td>0.80</td>
<td>0.41</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0.78</td>
<td>0.92</td>
<td>0.75</td>
<td>0.67</td>
<td>0.52</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.28</td>
<td>0.72</td>
<td>0.34</td>
<td>0.26</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.90</td>
<td>0.91</td>
<td>0.83</td>
<td>0.38</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>0.74</td>
<td>0.88</td>
<td>0.68</td>
<td>0.64</td>
<td>0.46</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.40</td>
<td>0.68</td>
<td>0.42</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.90</td>
<td>0.91</td>
<td>0.87</td>
<td>0.90</td>
<td>0.83</td>
</tr>
</tbody>
</table>

5.3.5 Summary of Simulation Study

A summary of the results of this simulation study is provided in Table 5.16. This simulation study has established Model 5.1 as a very viable demographic model.
Some parameter adjustments are helpful for inference and visual convergence inspection. While DIC and Bayes factors are useful, perhaps the most important model selection tool is an ad-hoc comparison of posterior distributions among demographic parameters. Finally, estimation of the true parameters is quite accurate, especially when considering that the parameters are removed from the original data by one hierarchical level—the latent variables. These demographic models are now applied to a real data scenario where demographics are likely to change the scale-usage analysis.

Table 5.16: Summary of Demographic Model Simulation Study Results

<table>
<thead>
<tr>
<th>Reference</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 5.1</td>
<td>○ Adjustment should be made for inference to improve convergence due to the likelihood nonidentifiability among the $\mu$, $\tau$, and $\theta$ parameters.</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>○ Using DIC and Bayes Factors model selection tools, Model 5.1 and Model 5.4 are most accurately identified compared to the other models. ○ DIC often outperforms Bayes Factors, though frequently not by much.</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>○ Model 5.1 and Model 5.4 are the most commonly selected models by DIC with rare selections of the other models. ○ This leads to the conclusion that the model constraints of Models 5.2, 5.3, and 5.5 may be unrealistic.</td>
</tr>
<tr>
<td>Table 5.6</td>
<td>○ Ad hoc model selection is nearly perfect at choosing between demographic and nondemographic models. ○ Ad hoc selection of the correct model among the five choices is also quite accurate.</td>
</tr>
<tr>
<td>Tables 5.7–5.12</td>
<td>○ The coverage is reasonable, and the correlation between posterior means and true values is very high for $\tau, \theta$, and $\mu$ parameters.</td>
</tr>
<tr>
<td>Tables 5.13–5.15</td>
<td>○ Coverage of other parameters is also reasonable. ○ More questions per respondent or a richer rating scale would likely improve the estimation.</td>
</tr>
</tbody>
</table>
5.4 Application: Student Ratings Data

5.4.1 Data Summary

Demographic effects on scale-usage are now studied using a dataset consisting of student evaluations of a course and instructor. The data were provided by an anonymous professor of an introductory statistics course who is interested in a more thorough analysis of student ratings beyond the brief overview often provided by a university at the end of a semester. The results are course and instructor specific and are not generalizable to an overall population, though the methods discussed could be applied in a larger study with such intentions (e.g. a university-wide analysis of student ratings).

Ratings are provided by 68 students on eight facets, which are shown in Table 5.17, encompassing attributes of both the course and instructor. Each question was evaluated on a five-point scale with the labels “poor,” “fair,” “good,” “very good,” and “excellent.”

<table>
<thead>
<tr>
<th>Table 5.17: Student Ratings Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Description of course objectives and assignments.</td>
</tr>
<tr>
<td>2. Communication of ideas and information.</td>
</tr>
<tr>
<td>3. Expression of expectations for performance in class.</td>
</tr>
<tr>
<td>4. Availability to assist students in or out of class.</td>
</tr>
<tr>
<td>5. Respect and concern for students.</td>
</tr>
<tr>
<td>6. Stimulation of interest in the course.</td>
</tr>
<tr>
<td>7. Facilitation of learning.</td>
</tr>
<tr>
<td>8. Overall assessment of instructor.</td>
</tr>
</tbody>
</table>
Demographic information was also obtained from each respondent as shown in Table 5.18. It is possible that students could differ in scale-use across any of these demographic variables. With the understanding that demographic-specific scale-usage patterns are most likely to occur when populations differ substantially, overall academic performance through GPA is likely the best differentiator among different types of students. Year in school is less likely because scale-usage patterns have been shown to often be constant across time for individuals who will eventually pass through all the years of school. The grade in this course may differentiate students, but it represents a single data-point of academic performance whereas the cumulative GPA represents the entire body of academic work for the student at the university. While the students who are not required to take a statistics course yet nevertheless enroll are likely different from their counterparts, those who are required to take the course are from such widely varied disciplines that pinpointing a cultural pattern among them would be challenging. As part of the model-selection process, in addition to selecting from Model 5.1 through Model 5.5, each of these demographics will be considered with Model 5.1 through Model 5.3 to see which, if any, demographic effects are important.

As a summary of the data and the scale-usage patterns therein, Figure 5.4 provides a bubble chart comparing the median and range across the eight questions for each individual. The most common median/range combination is a median of five and a range of zero, i.e. the individuals who responded with “excellent” to each of the questions. Similarly common patterns are a median of four or five and a range of one, i.e. individuals who use only fours and fives in their rankings. Other response
patterns are present, but much less likely. No individual has a median less than 2.5, and only a few have a range greater than two.

Figure 5.4: Bubble chart of scale-usage patterns across all individuals.

To gain a sense for the demographic variability, a similar bubble plot is now provided for each of the levels of each demographic. Figure 5.5 categorizes the students
according to class rank. There are subtle differences, but no striking discrepancies in the scale-usage patterns. Juniors appear to respond with a median of five and a range of one more than any other group, with very few students using a range of zero. Nearly all the seniors have a median of five, though there are many fewer of them compared to the other years, so this pattern may be due to chance.

Figure 5.5: Bubble chart of scale-usage patterns by year in school.
Figure 5.6 classifies students according to their cumulative GPAs. With this demographic, more differences are evident as previously postulated. Those with GPA less than 2.49 are quite scattered. Only one individual has median of four and one has median of five. Three individuals have a median of three, much more than any other group relatively. Those with GPA between 2.5 and 2.99 have all individuals respond with a median of four or five, most with range of either zero or one. Thus, the within-group variation for this group is quite small. For those with GPA from 3.0 to 3.49, most individuals respond with a median of five. The group with GPA from 3.5 to 4.0 has the largest range of both the ranges and medians. While several respond with median five and range zero, more respond with median four and range one, a response which is barely present in the 2.49 or less and 3.00-3.49 groups. These differences may be important enough to warrant the use of a demographic-specific scale-usage model.

Figure 5.7 provides the plots categorizing students by their expected grades in the course. The first note is that the majority of students expect to receive an A in the course, with the remaining students mostly expecting a B. Students expecting less than a C were combined with those expecting a C due to small numbers. With regards to scale usage, the overall patterns do not appear to starkly contrast between the groups.

Similarly, in Figure 5.8, there are many more students required to take the course. Their pattern doesn’t seem to differ much from those who are not required to take the course.

Having explored the scale-usage patterns in the data, the attention now turns to the formal fitting and selection of a model for inference.
Figure 5.6: Bubble chart of scale-usage patterns by cumulative GPA.
Figure 5.7: Bubble chart of scale-usage patterns by expected grade.
5.4.2 Model Selection

To determine which model should be utilized for inference with these data, Models 5.1 through 5.3 were fit using each of the four demographics, hereafter indexed by the respective question number listed in Table 5.18. Models 5.4 and 5.5, which are not demographic specific, were also fit to the data. Markov chains were run for 10,000 burn-in iterations from a fixed initial distribution and thereafter run for one million iterations. The tuning parameters for the Metropolis-Hastings steps for the cutpoint parameters were updated during the burn-in period to produce acceptance rates within the range of 0.3 to 0.5. After burn-in, the tuning parameters were held constant.

DIC and Bayes factors were computed for each of the models under each of the three previously-discussed data definitions and are provided in Table 5.19. Note that the quantity listed under each Bayes factor is the log of the estimated $P(D|M)$ where
D is the data and M is the model. Therefore, to compute the actual Bayes factor between Model i and Model j, simply subtract the value provided in Table 5.19 for Model j from that for Model i and exponentiate the difference. Hence, models with larger values in the table (smaller absolute values as all are negative) are preferred. With DIC, smaller values are preferred.

For each of the selection criterion, it is clear that Model 5.1 and Model 5.4 are superior to Model 5.2, Model 5.3, and Model 5.5, which was largely anticipated from the results of the simulation study. It remains then to choose between Model 5.1 with each of the demographics and Model 5.4. For all six model selection criteria (save DIC1 for which it is tied), at least one of the Model 5.1 fits is preferred over Model 5.4, leading to the conclusion that a demographic-based model is likely better. The criteria, however, don’t perfectly agree as to which demographic is best used with Model 5.1. Demographic 1 is preferred by DIC1 and DIC2; Demographic 2 is preferred by BF2; Demographic 3 is preferred by DIC3; and finally, Demographic 4 is preferred by BF1 and BF3. Given the lack of a clearly superior criterion from the simulation study, the question of which demographic should be fit with Model 5.1 may be left to the ad-hoc approach. This was illustrated to be superior at distinguishing models compared to the traditional approaches of DIC and Bayes factors in the simulation study.

The ad hoc method is applied to each of the four demographics. Recall the first step is to fit Model 5.1 and compare the credible intervals of each of the $\theta$ parameters to zero. As there are a relatively small number of levels per demographic, any interval which does not contain zero results in selecting a demographic-specific model; if all contain zero, choose a non-demographic model. The 95% equal-tail credible bounds
Table 5.19: DIC and Bayes Factors for Model Selection Among Demographic-Specific Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Dem.</th>
<th>DIC_1</th>
<th>DIC_2</th>
<th>DIC_3</th>
<th>BF_1</th>
<th>BF_2</th>
<th>BF_3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1590</td>
<td>1795</td>
<td>2152</td>
<td>-765</td>
<td>-884</td>
<td>-1099</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1675</td>
<td>1835</td>
<td>2220</td>
<td>-742</td>
<td>-873</td>
<td>-1148</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1677</td>
<td>1802</td>
<td>2102</td>
<td>-757</td>
<td>-891</td>
<td>-1095</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>1654</td>
<td>1899</td>
<td>2155</td>
<td>-739</td>
<td>-880</td>
<td>-1071</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1982</td>
<td>2202</td>
<td>3145</td>
<td>-981</td>
<td>-1082</td>
<td>-1529</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2078</td>
<td>2320</td>
<td>3383</td>
<td>-1043</td>
<td>-1154</td>
<td>-1654</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>2070</td>
<td>2316</td>
<td>3352</td>
<td>-997</td>
<td>-1105</td>
<td>-1624</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>2253</td>
<td>2310</td>
<td>3668</td>
<td>-1024</td>
<td>-1135</td>
<td>-1687</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1769</td>
<td>1895</td>
<td>2340</td>
<td>-820</td>
<td>-945</td>
<td>-1162</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1853</td>
<td>1913</td>
<td>2310</td>
<td>-821</td>
<td>-950</td>
<td>-1174</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1701</td>
<td>1904</td>
<td>2328</td>
<td>-806</td>
<td>-943</td>
<td>-1178</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>1741</td>
<td>1904</td>
<td>2358</td>
<td>-817</td>
<td>-949</td>
<td>-1191</td>
</tr>
<tr>
<td>4</td>
<td>N/A</td>
<td>1590</td>
<td>1852</td>
<td>2120</td>
<td>-744</td>
<td>-894</td>
<td>-1106</td>
</tr>
<tr>
<td>5</td>
<td>N/A</td>
<td>2027</td>
<td>2166</td>
<td>2568</td>
<td>-927</td>
<td>-1063</td>
<td>-1320</td>
</tr>
</tbody>
</table>

are provided in Table 5.20. Note this represents the fitting of four separate models, one for each demographic. The only credible interval which does not contain zero is that for GPA from 3.00 to 3.49, indicating that, under the ad-hoc approach, the only demographic which requires a demographic-specific model is based on cumulative GPA. The posterior means of $\theta_t$ also vary the most for this demographic, from -2.32 to 1.35, indicating a demographic-specific model is useful.

Finishing with the ad-hoc procedure for the cumulative GPA demographic, all the credible intervals for $\sigma_i^2$ fit using Model 5.2 overlap with at least one other interval, eliminating Model 5.2. The average absolute distance between 1 and the posterior median for each $\sigma_i^2$ is larger than 0.2 for this demographic, so Model 5.1 is ultimately selected.
Based on the ad-hoc procedure, Model 5.4 is unsurprisingly selected over Model 5.5 for each of the other demographics where demographic effects are not used in modeling. Thus, according to these results, Model 5.4 would be the best alternative if the demographic model for cumulative GPA were not used.

Combining these results with those from DIC and Bayes factors where the determination was that Model 5.1 would be a good model choice, the ultimately selected model is Model 5.1 with the cumulative GPA demographic. This matches the intuition and exploratory analysis at the beginning of this section. While other models would also be useful and justification could be made for a different demographic selection, the inference here will be carried out with this model.

Table 5.20: Initial Posterior Summaries for $\theta_t$ for Each Demographic from Model 5.1 which are used for the Ad-Hoc Model Selection Approach

| Demographic                  | $E(\theta_t|X)$ | CI Lower Bound | CI Upper Bound |
|------------------------------|-----------------|----------------|---------------|
| 1st Year                     | -0.14           | -2.34          | 2.24          |
| Sophomore                    | -0.33           | -1.84          | 1.28          |
| Junior                       | -0.50           | -1.61          | 0.64          |
| Senior                       | 2.49            | -0.36          | 6.07          |
| GPA 2.49 or Less             | -2.32           | -6.87          | 2.46          |
| GPA 2.50–3.00                | 0.26            | -0.77          | 1.41          |
| GPA 3.00–3.49                | 1.35            | 0.10           | 2.85          |
| GPA 3.50–4.00                | -1.00           | -2.80          | 0.85          |
| Expected A                   | 0.50            | -0.59          | 1.67          |
| Expected B                   | -0.68           | -2.14          | 0.81          |
| Expected C or less           | -0.90           | -5.14          | 3.66          |
| Required                     | -0.06           | -0.94          | 0.82          |
| Not Required                 | 0.35            | -1.94          | 2.99          |

242
5.4.3 Inference

For inference, the Markov chain was run for 1 million iterations. With respect to an initial distribution, the final iteration of the model-selection chain was used as the initial value, providing over 1 million draws as burn-in for this ultimate chain. As discussed by Flegal and Jones (2010), burn-in is not required for the theory to guarantee a Markov chain central limit theorem; practically, however, a closer initial distribution will effect the results when the sample is finite, as in this case.

The methods of Chapter 3 were attempted for this Markov chain in order to obtain an approximate draw to use as the initial value. The minorization condition tried, however, proved insufficient for practical application; several of the regeneration times exceeded 15 million iterations, leaving the methods discussed in Chapter 3 unsuitable for this problem until a better minorization condition is obtained. As will be seen, the Markov chain Monte Carlo standard errors are small enough for this simulation to successfully answer Question 1 and Question 2 without the need to implement the Chapter 3 methods.

Posterior Summaries

A discussion of posterior summaries is now provided. As the chain is geometrically ergodic, Markov chain CLTs exist for averages, provided the necessary moment conditions are satisfied as discussed in Section 2.4. Similarly, CLTs exist for quantiles according to the discussion in Section 2.4.3. The most important requirement is polynomial ergodicity of order larger than 1, which is satisfied as the algorithms are geometrically ergodic (geometric convergence is faster than polynomial convergence).
Begin with the individual-level parameters. In lieu of providing a table with 68 posterior summaries, Figure 5.9 plots the posterior mean for each \( \tau_i \) against the raw data mean. There is a strong relationship between these two quantities, indicating that \( \tau_i \) is performing the desired scale-usage adjustments.

Note that as the \( \tau_i \) parameters have compact support, the moment conditions on \( \mathbb{E}_\pi|\tau_i|^{2+\delta+\epsilon} \) are satisfied for any \( \delta > 0 \), guaranteeing the strong consistency of the estimate of the CLT variance as discussed in Section 2.4.2 with the batch size equal to \( \lfloor n^\nu \rfloor \) for any \( \nu < 3/4 \). Choose a batch size of \( \lfloor n^{1/2} \rfloor = 1000 \) as suggested by results in Flegal and Jones (2010) and utilize overlapping batch means. The median Markov chain Monte Carlo standard error (MCSE) for the \( \tau_i \) parameters was 0.013 with a maximum value of 0.056. Thus, with respect to these parameters, the chain appears to have been run a sufficient number of iterations.

Figure 5.9: Posterior means of \( \tau_i \) parameters compared to raw score means.
Similarly, Figure 5.10 plots the posterior mean of each $\sigma_i^2$ parameter against the raw data standard deviation. While the relationship is not as strong as with the $\tau_i$ parameters, it is clear that larger raw standard deviations correspond with larger $\sigma_i^2$ values. The MCSEs were computed in the same fashion as with the $\tau_i$ parameters with the same moment condition satisfied due to the compact support. The median MCSE was 0.0016 with maximum 0.036. Thus, the chain has also been run long enough for accurate inference on these parameters.

Figure 5.10: Posterior means of $\sigma_i^2$ parameters compared to raw score standard deviations.

Table 5.21 provides the posterior summary for each of the dimensions of $\mu$. The posterior median is provided, along with the 0.025 and 0.975 quantiles, providing a 95% equal-tail credible interval for each dimension marginally. The estimates for these quantiles are also presented with MCSEs parenthetically, computed with the
subsampling bootstrap method as discussed in Section 2.4.3 with batch size $|n^{1/2}| = 1000$. Note that throughout the following tables, an MCSE of 0.00 is not precisely zero, but simply implies that the MCSE is below 0.005, and thus rounds to 0.00 when taken to two decimal places. All the MCSEs are below 0.01 for the medians and credible interval estimates for $\mu$, indicating that the simulation effort is enough for suitable accuracy.

Examination and comparison of the posterior medians is informative in comparing the eight dimensions considered in this survey. Recall that the cutpoint between responses of four and five is set to six. The highest posterior median corresponds to overall assessment of the instructor, question (8), on the survey. The entire credible interval for this question falls above the cutpoint of 6, implying “excellent” overall ratings on average. The next highest median corresponds to question (5) about respect and concern for the students.

The lowest median corresponds to question (6) with respect to stimulation of interest in the course. As any undergraduate introductory statistics course instructor understands, this is no easy task! While lowest, the median is still high and corresponds with “very good” on the rating scale. The next lowest is question (1), description of course objectives and assignments, but the median is quite close to the “excellent” threshold. The remaining evaluation attributes have median implying “excellent” ratings. This information can then be used by the interested faculty member to evaluate his or her strengths and set one or two goals for improvement as desired.

Next, consider the posterior summary for $\Sigma$, provided in Table 5.22. The lower triangle provides the marginal posterior medians of $\Sigma$. The upper triangle, italicized,
Table 5.21: Posterior Summary of $\mu$

<table>
<thead>
<tr>
<th>Question</th>
<th>Lower Bound</th>
<th>Median</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.37 (0.00)</td>
<td>5.94 (0.00)</td>
<td>6.55 (0.01)</td>
</tr>
<tr>
<td>2</td>
<td>5.61 (0.00)</td>
<td>6.10 (0.00)</td>
<td>6.64 (0.01)</td>
</tr>
<tr>
<td>3</td>
<td>5.72 (0.00)</td>
<td>6.21 (0.00)</td>
<td>6.75 (0.01)</td>
</tr>
<tr>
<td>4</td>
<td>5.69 (0.00)</td>
<td>6.35 (0.00)</td>
<td>7.15 (0.01)</td>
</tr>
<tr>
<td>5</td>
<td>6.60 (0.01)</td>
<td>7.15 (0.01)</td>
<td>7.83 (0.01)</td>
</tr>
<tr>
<td>6</td>
<td>4.37 (0.01)</td>
<td>5.06 (0.00)</td>
<td>5.73 (0.00)</td>
</tr>
<tr>
<td>7</td>
<td>5.76 (0.00)</td>
<td>6.37 (0.00)</td>
<td>7.06 (0.01)</td>
</tr>
<tr>
<td>8</td>
<td>6.65 (0.01)</td>
<td>7.20 (0.01)</td>
<td>7.85 (0.01)</td>
</tr>
</tbody>
</table>

provides the marginal posterior medians of the corresponding correlation matrix for reference. The strongest correlation (0.66, CI from 0.15 to 0.88) is positive between overall satisfaction (question 8) and facilitation of learning (question 7). There is an equally strong ($-0.65$, CI from $-0.87$ to $-0.20$) negative correlation between facilitation of learning (question 7) and availability to assist students in or out of class (question 4), which is an incredibly curious result. These are the only two correlations which have credible bounds which do not cover zero. The only other correlation strong enough to mention (0.50, CI from $-0.13$ to 0.83) is between availability to assist in or out of class (question 4) and respect and concern for students (question 5), which is an intuitive relationship, though the credible interval does contain 0.

To evaluate Question 1 and Question 2 with respect to the $\Sigma$ parameters, a matrix of the MCSEs is provided in Table 5.23. None of the MCSEs exceed 0.03, with most of them smaller than 0.01. Those for correlations, shown again in the upper triangle of the matrix, italicized, are all 0.003 or below. Thus, the median estimates provided in Table 5.22 are quite accurate.
Table 5.22: Posterior Medians of Covariance Matrix $\Sigma$

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.65</td>
<td>0.20</td>
<td>0.30</td>
<td>-0.27</td>
<td>-0.19</td>
<td>-0.42</td>
<td>-0.13</td>
<td>-0.05</td>
</tr>
<tr>
<td>0.21</td>
<td>0.72</td>
<td>0.03</td>
<td>-0.18</td>
<td>-0.18</td>
<td>0.07</td>
<td>0.05</td>
<td>0.08</td>
</tr>
<tr>
<td>0.31</td>
<td>0.02</td>
<td>0.74</td>
<td>0.30</td>
<td>0.32</td>
<td>-0.37</td>
<td>-0.36</td>
<td>-0.28</td>
</tr>
<tr>
<td>-0.56</td>
<td>-0.25</td>
<td>0.43</td>
<td>3.17</td>
<td>0.50</td>
<td>0.03</td>
<td>-0.65</td>
<td>-0.44</td>
</tr>
<tr>
<td>-0.20</td>
<td>-0.12</td>
<td>0.23</td>
<td>0.74</td>
<td>0.80</td>
<td>-0.05</td>
<td>-0.17</td>
<td>-0.17</td>
</tr>
<tr>
<td>-0.84</td>
<td>0.09</td>
<td>-0.48</td>
<td>0.09</td>
<td>-0.07</td>
<td>2.74</td>
<td>0.23</td>
<td>0.24</td>
</tr>
<tr>
<td>-0.24</td>
<td>0.06</td>
<td>-0.44</td>
<td>-1.66</td>
<td>-0.21</td>
<td>0.54</td>
<td>2.32</td>
<td>0.66</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.06</td>
<td>-0.22</td>
<td>-0.73</td>
<td>-0.14</td>
<td>0.36</td>
<td>0.96</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 5.23: Standard Errors for Posterior Medians of Covariance Matrix $\Sigma$

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.010</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>0.004</td>
<td>0.004</td>
<td>0.002</td>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>0.004</td>
<td>0.002</td>
<td>0.005</td>
<td>0.003</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>0.006</td>
<td>0.005</td>
<td>0.007</td>
<td>0.025</td>
<td>0.003</td>
<td>0.002</td>
<td>0.001</td>
<td>0.002</td>
</tr>
<tr>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
<td>0.009</td>
<td>0.006</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>0.008</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
<td>0.005</td>
<td>0.017</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>0.005</td>
<td>0.004</td>
<td>0.005</td>
<td>0.014</td>
<td>0.004</td>
<td>0.008</td>
<td>0.017</td>
<td>0.002</td>
</tr>
<tr>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.008</td>
<td>0.003</td>
<td>0.006</td>
<td>0.010</td>
<td>0.007</td>
</tr>
</tbody>
</table>
As discussed in Section 4.2.5, regression estimates predicting overall assessment of the instructor from the other survey questions can be computed using the covariance matrix $\Sigma$. This is done for each iteration in order to construct a posterior distribution of regression coefficients, a summary of which is provided in Table 5.24. All of the credible intervals include zero, indicating that none of the other questions are particularly important when predicting overall assessment of the instructor. While the credible interval contains zero, the most important predictor is question (7), facilitation of learning.

The lack of a strong relationship between questions may indicate that the survey questions are in fact assessing largely independent facets of the assessment of the course and instructor, which is often the purpose of such surveys. A more detailed survey is likely required for a study of what specific attributes caused students to highly regard this professor. Note that the parenthetical MCSEs are all 0.002 or less, indicating an excellent level of precision in estimation.

Table 5.24: Posterior Summary for Regression Coefficients Predicting Overall Instructor Rating

<table>
<thead>
<tr>
<th>Question</th>
<th>Lower Bound</th>
<th>Median</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.42 (0.002)</td>
<td>0.01 (0.001)</td>
<td>0.41 (0.002)</td>
</tr>
<tr>
<td>2</td>
<td>-0.55 (0.002)</td>
<td>-0.04 (0.001)</td>
<td>0.46 (0.002)</td>
</tr>
<tr>
<td>3</td>
<td>-0.62 (0.002)</td>
<td>-0.10 (0.001)</td>
<td>0.44 (0.002)</td>
</tr>
<tr>
<td>4</td>
<td>-0.45 (0.002)</td>
<td>-0.05 (0.001)</td>
<td>0.36 (0.002)</td>
</tr>
<tr>
<td>5</td>
<td>-0.58 (0.002)</td>
<td>-0.08 (0.001)</td>
<td>0.42 (0.002)</td>
</tr>
<tr>
<td>6</td>
<td>-0.34 (0.002)</td>
<td>0.03 (0.001)</td>
<td>0.37 (0.001)</td>
</tr>
<tr>
<td>7</td>
<td>-0.15 (0.002)</td>
<td>0.33 (0.001)</td>
<td>0.74 (0.002)</td>
</tr>
</tbody>
</table>
The posterior summary of the free cutpoints is provided in Table 5.25. Recall the minimum cutpoint is $-20$, the cutpoint between ratings 1 and 2 is $-6$, and the cutpoint between ratings 4 and 5 is 6, with maximum cutpoint 20. The free cutpoints in Table 5.25, $c_2$ and $c_3$ are between ratings 2 and 3 and 3 and 4 respectively. These summaries are particularly useful in understanding how to interpret the posterior summary of $\mu$. For example, the lowest lower credible bound for $\mu$ is 4.37, which is larger than the upper bound for $c_3$, implying that all the $\mu$ ratings correspond to ratings of 4 or 5 on average (depending on the dimension). Again, all of the MCSEs are small enough, even for these parameters which were updated using a Metropolis-Hastings step which are typically more prone to autocorrelation. This implies that enough simulations of the Markov chain were obtained for trustworthy inference.

Table 5.25: Posterior Summary of the Cutpoint Parameters

<table>
<thead>
<tr>
<th>Cutpoint</th>
<th>Lower Bound</th>
<th>Median</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_2$</td>
<td>$-2.99 (0.03)$</td>
<td>$-1.08 (0.02)$</td>
<td>$0.70 (0.02)$</td>
</tr>
<tr>
<td>$c_3$</td>
<td>$1.37 (0.01)$</td>
<td>$2.48 (0.01)$</td>
<td>$3.39 (0.01)$</td>
</tr>
</tbody>
</table>

Now, consider the demographic-specific parameters $\theta$ and $\sigma_\tau^2$ whose posterior summaries are provided in Tables 5.26 and 5.27 respectively. The first demographic, with GPA less than 2.49, has a wide credible interval for both $\theta$ and $\sigma_\tau^2$, but this is intuitive as only five students fit this demographic (compared to 20, 21, and 22 for the other demographics respectively). The posterior median for $\theta$ for this group is lowest among the four demographics, indicating more nay-saying tendencies for this demographic
overall. The next most critical group, interestingly, is those with demographic corresponding to GPAs between 3.50 and 4.00. The students with GPAs between 3.00 and 3.49 give the most favorable reviews overall with a credible interval that is entirely above zero. Those with GPAs between 2.50 and 2.99 are most neutral with posterior median equal to 0.23.

The MCSEs for the $\theta$ parameters are all quite small, 0.03 or smaller. The MCSEs for the $\sigma^2_i$ parameters are the highest of any parameters considered thus far, but they also have the highest scale of any parameter such that the larger MCSE isn’t a concern. For example, the highest MCSE is 0.33, which is still quite small, especially when the respective value estimated is 80.40. Thus, Question 1 and Question 2 have been answered for each of the parameters of interest, and it appears that the simulation has been run long enough and that the estimates are sufficiently accurate.

Table 5.26: Posterior Summary for Demographic Level $\theta$ Parameters

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Lower Bound</th>
<th>Median</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-6.89 (0.02)</td>
<td>-2.38 (0.01)</td>
<td>2.41 (0.03)</td>
</tr>
<tr>
<td>2</td>
<td>-0.78 (0.00)</td>
<td>0.23 (0.01)</td>
<td>1.43 (0.01)</td>
</tr>
<tr>
<td>3</td>
<td>0.09 (0.00)</td>
<td>1.32 (0.01)</td>
<td>2.91 (0.02)</td>
</tr>
<tr>
<td>4</td>
<td>-2.83 (0.01)</td>
<td>-1.00 (0.01)</td>
<td>0.86 (0.01)</td>
</tr>
</tbody>
</table>

Clustering

Inference will now shift focus to consider if the means of the standardized scores $z_i$, defined in (4.4), may be used to effectively cluster the respondents. Note that $z_i$ defined in (4.4) is not equivalent to the $Z_i$ parameters used to facilitate MCMC
Table 5.27: Posterior Summary for Demographic Level $\sigma^2$ Parameters

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Lower Bound</th>
<th>Median</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.74 (0.07)</td>
<td>17.43 (0.22)</td>
<td>80.40 (0.33)</td>
</tr>
<tr>
<td>2</td>
<td>1.49 (0.03)</td>
<td>4.23 (0.05)</td>
<td>12.88 (0.12)</td>
</tr>
<tr>
<td>3</td>
<td>2.59 (0.05)</td>
<td>7.08 (0.09)</td>
<td>20.01 (0.18)</td>
</tr>
<tr>
<td>4</td>
<td>6.34 (0.08)</td>
<td>15.33 (0.14)</td>
<td>37.53 (0.28)</td>
</tr>
</tbody>
</table>

sampling of multivariate truncated normal random variables; this is a slight abuse of notation, attenuated by using lowercase $z_i$ when referring to the standardized scores and uppercase $Z_i$ when referring to the $Z_i$ used in the decomposition sampler. While clustering students based on how they rate a professor may not be a necessary exercise compared to segmenting a customer base for a business, it is nevertheless pursued here to illustrate the importance of adjusting for scale-usage heterogeneity before clustering.

The $K$-means algorithm is used to separate the respondents into clusters. For an accessible introduction, see Hastie et al. (2009). In general, the algorithm creates $K$ groups which seek to minimize the within-group sum of squared errors. Answers are not unique, however, and can vary based on the starting values of the algorithm. Variants of this procedure and many other clustering algorithms are available if this is the focus of inference.

One important part of $K$-means clustering is to identify the number of clusters to use as the vanilla $K$-means algorithm has no mechanism for automatic selection of $K$. A scree plot using the sum of squared errors is often used for this, and is shown for this problem in Figure 5.11. The idea is to select the smallest $K$ such that the
marginal decrease in SSE for an additional cluster is small. For this plot, it appears that five would be a good number of clusters.

To evaluate the performance of the clustering algorithm, each of the clusters is shown in Figure 5.12, plotted with the first principal component against the second principal component. It appears that distinct clusters are captured quite well. To gain a sense for the practical differences in clusters, the cluster means are provided in Table 5.28. The first cluster highly rates the description of course objectives and assignments (question 1), but has lower-than-average ratings on questions 4, 6, 7, and 8. Cluster 3 poorly rates question 6, stimulation of interest in the course, but more highly rates description of course objectives (question 1) and expression of expectations for performance in class (question 3). Cluster 4 highly rates availability to assist students in or out of class (question 4), but has low ratings for facilitation of learning (question 7). The fifth cluster takes the opposite perspective by highly rating facilitation of learning (question 7), but having a low rating for availability to assist students in or out of class (question 4). Again, with student ratings, it’s not clear how this information might be used to cater to different subgroups of students, but for segmentation analysis for a product, certain features could be marketed to different segments of the population.

For comparison, the same analysis is performed on the original data X to illustrate the importance of scale-usage adjustments for clustering purposes. The scree plot is provided in Figure 5.13. For these data, three clusters appear sufficient. The $K$-means is then fit with three clusters, and the results according to the first two principal components are presented in Figure 5.14. The separation is superb, but appears to only require the first principal component. A look at the cluster means
Table 5.28: Cluster Means for Standardized Scores

<table>
<thead>
<tr>
<th>Question</th>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.89</td>
<td>-0.22</td>
<td>0.82</td>
<td>-0.45</td>
<td>-0.06</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-0.06</td>
<td>0.07</td>
<td>-0.10</td>
<td>-0.32</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
<td>-0.16</td>
<td>0.51</td>
<td>0.41</td>
<td>-0.35</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-0.72</td>
<td>0.11</td>
<td>0.03</td>
<td>2.08</td>
<td>-2.25</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-0.15</td>
<td>-0.10</td>
<td>0.16</td>
<td>0.60</td>
<td>-0.38</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.52</td>
<td>0.61</td>
<td>-2.15</td>
<td>-0.17</td>
<td>0.20</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>-0.73</td>
<td>0.08</td>
<td>-0.28</td>
<td>-1.56</td>
<td>1.73</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>-0.47</td>
<td>0.11</td>
<td>-0.23</td>
<td>-0.70</td>
<td>0.68</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.11: SSE scree plot for selecting the number of clusters for standardized scores.
Figure 5.12: First two principal components to illustrate cluster separation for standardized scores.
in Table 5.29 reveals how the clusters were selected: average-responders, yeah-sayers, and nay-sayers respectively. Thus, most of the interesting cluster structure has been lost by not accounting for scale-usage heterogeneity before clustering. Accounting for scale-usage heterogeneity is therefore paramount to clustering of ordinal survey data.

![Figure 5.13: SSE scree plot for selecting the number of clusters for raw data.](image)

**Predictive Distributions**

The next inferential question of interest is to compare the predictive distributions across the different demographics. For these predictive distributions, individual $\sigma_i^2$ parameters were set equal to 1, though there is little difference in the distributions if $\sigma_i^2$ is drawn from the prior distribution. The $\tau$ parameters were distributed as in the prior with the specific $\theta_t$ and $\sigma_{i,t}^2$ for that demographic coming from the MCMC
Figure 5.14: First two principal components to illustrate cluster separation for raw data.

Table 5.29: Cluster Means for Original Data

<table>
<thead>
<tr>
<th>Question</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.12</td>
<td>4.83</td>
<td>2.67</td>
</tr>
<tr>
<td>2</td>
<td>4.08</td>
<td>4.89</td>
<td>3.00</td>
</tr>
<tr>
<td>3</td>
<td>4.17</td>
<td>4.91</td>
<td>3.00</td>
</tr>
<tr>
<td>4</td>
<td>4.04</td>
<td>4.94</td>
<td>3.11</td>
</tr>
<tr>
<td>5</td>
<td>4.42</td>
<td>5.00</td>
<td>3.33</td>
</tr>
<tr>
<td>6</td>
<td>3.62</td>
<td>4.77</td>
<td>2.56</td>
</tr>
<tr>
<td>7</td>
<td>4.00</td>
<td>4.91</td>
<td>3.22</td>
</tr>
<tr>
<td>8</td>
<td>4.29</td>
<td>5.00</td>
<td>3.33</td>
</tr>
</tbody>
</table>
sample. The distributions were created specifically for the overall instructor assessment, and are shown in Figure 5.15. Error bars are provided using the Markov chain CLT discussed in Section 2.4 (a moment condition is guaranteed because each is a probability between 0 and 1), but, as is shown, the Monte Carlo error is essentially negligible. While each demographic is most likely to rate the overall quality of the instructor with a 5, those with the lowest and the highest GPAs are more likely than the middle GPAs to evaluate something lower than a 5. In fact, those with middle GPAs are very unlikely to rate below a 4, and are almost completely unlikely to rate below a 3. These predictive distributions certainly affirm that demographic-specific effects are helpful in this analysis.

The final inferential question of interest is to identify extreme respondents. With such high responses overall as the default, a good way to do this as suggested in Chapter 4 is to order by the standardized score on the overall satisfaction question. After doing this, those with the lowest standardized scores on \( z_8 \) rate quite highly on the other attributes, but low on question 8; conversely those with the highest standardized scores rate lower on the other attributes but high on overall instructor assessment. The definition of “extreme-respondents” should depend on the context, and in the case of student ratings, such identification is likely unnecessary as anonymity is one of the key facets of such evaluations (and hence the extreme respondents are not listed in a table here). Among customers of a product, however, it may be helpful to understand extreme attitudes and their prevalence.
Figure 5.15: Predictive distribution for overall instructor assessment by demographic.
Table 5.30: Summary of Inference for Student Ratings Data

<table>
<thead>
<tr>
<th>Reference</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 5.19</td>
<td>○ Models 5.1 and 5.4 are preferred over the other models using DIC and Bayes Factors.</td>
</tr>
</tbody>
</table>
| Table 5.20 | ○ The ad-hoc approach concludes that the only demographic which should be used in adjusting for scale-usage heterogeneity is cumulative GPA.  
○ This conclusion is sensible because cumulative GPA captures the most information about overall academic performance for a given student.  
○ Model 5.1 is therefore selected for these data using cumulative GPA for the demographic effects. |
| Table 5.21 | ○ According to the posterior medians for $\mu$, the students rated this professor best on overall assessment of instructor.  
○ The next best attribute was respect and concern for students.  
○ The other attributes were all relatively similar. |
| Table 5.22 | ○ Overall assessment has moderate positive correlation with facilitation of learning.  
○ Overall assessment has moderate negative correlation with availability to assist students in or out of class. |
| Table 5.24 | ○ None of the regression coefficients predicting overall assessment are important. |
| Table 5.26 | ○ The most complementary demographic group has GPA 3.00–3.49.  
○ The most critical demographic group has GPA less than 2.49. |
| Table 5.27 | ○ Demographic groups with the most varied response styles have GPA less than 2.49 or greater than 3.50. |
| Tables 5.28–5.29 | ○ Cluster analysis is much more meaningful after accounting for scale-usage heterogeneity. |
| Figure 5.15 | ○ Predictive distributions differ among the demographics. |
| Tables 5.21–5.27 | ○ Small MCMC standard errors show that 1 million iterations are sufficient for accurate inference. |
5.4.4 Summary of Inference for Student Ratings Data

In all of these inference approaches, the common theme has been that accounting for scale-usage heterogeneity is important, and that demographic-specific effects are often useful in helping to account for this heterogeneity. A more complete summary of the inferential conclusions is provided in Table 5.30.

The Markov chain Monte Carlo standard errors were all small for the estimated quantities. This answers Question 1 that the simulation was indeed run long enough and provides a direct answer to Question 2 about the accuracy of the estimates. Geometric ergodicity, which was established in Chapter 4, allowed for the establishment of the appropriate Markov chain central limit theorems and the estimation of the respective asymptotic standard errors.

Once geometric ergodicity is established, the implementation of computing Markov chain Monte Carlo standard errors is incredibly simple and quick. The burden for using these methods is then the establishment of geometric or polynomial ergodicity to guarantee existence of the CLTs and the asymptotic standard errors.
Chapter 6: Conclusions and Future Work

Markov chains are an incredibly powerful tool for statisticians. Often, their implementation, even for complicated problems, is very simple and straightforward. Several Markov chain properties, including irreducibility, aperiodicity, recurrence, and stationarity, are routinely satisfied. These conditions guarantee a strong law of large numbers for ergodic averages such that the estimates converge to the truth almost surely. The challenge, however, is that infinity is typically too long to wait for inference.

Given that finite time is a reality for any Markov chain in practice, two very important questions result.

**Question 1:** Was the simulation run long enough?

**Question 2:** How accurate are the resulting estimates?

This dissertation has provided a discussion for how these questions may be answered in a variety of settings.

### 6.1 Summary and Conclusions

Chapter 2 provided the theory behind Markov chain convergence. At the core of this theory is the establishment of convergence rates. Uniform, geometric, and polynomial convergence rates were defined and discussed. Two commonly used methods
for establishing geometric ergodicity were detailed: (1) proving drift and minorization, and (2) proving drift and unboundedness on compact sets. A discussion was then provided about how bounds for the total variation distance from \( \pi \) can be computed for uniformly or geometrically ergodic Markov chains, providing one answer to Question 1 as the number of iterations can be obtained such that the TV distance is below some threshold. Markov chain central limit theorems, which require establishment of a convergence rate, are then detailed, both for averages and quantiles, along with strategies for estimating asymptotic variances. These Markov chain CLTs and corresponding asymptotic variances can also be used to effectively answer both Question 1 and Question 2 by estimating the accuracy of the estimates.

Two examples are then provided. The first considers a Bayesian random effects model with proper prior distributions. Both methods of establishing geometric ergodicity are illustrated. Bounds are computed such that the TV distance is guaranteed to be below 0.1 using two different bounding inequalities. The required number of iterations is heavily dependent upon the prior distribution. On the other hand, a Markov chain Monte Carlo central limit theorem was used to create error bounds for the posterior medians and credible interval bounds after 1 million iterations of each chain. The estimates were all very accurate, which provides the answer to Question 2 while simultaneously answering Question 1 that the chain had been run long enough for accurate inference.

The second example develops an independence Metropolis-Hastings algorithm for sampling from a truncated multivariate normal distribution. This chain is shown to be uniformly ergodic, and bounds such that the TV distance is less than 0.1 are also provided. Again, the number of iterations required for this guaranteed precision
varied drastically, yet the Monte Carlo standard errors were appropriately small for inference in estimation of the means of the truncated multivariate normal distributions using 1 million iterations. The overall conclusion is that Markov chain CLTs and the corresponding asymptotic variances are quite effective at answering both Question 1 and Question 2, once the required theory has been established which guarantees their existence.

Chapter 3 discussed another approach for answering Question 1: starting the Markov chain with an observation which closely approximates the target distribution $\pi$. A general minorization condition leads to a mixture representation equivalent to $\pi$. Several approaches were then discussed for drawing from the mixture weights, which is the biggest challenge for using the mixture representation to sample from $\pi$. When the chain is uniformly or geometrically ergodic, this may be done exactly. The remainder of the chapter was dedicated to situations where the chain is not necessarily geometrically or uniformly ergodic, and thus approximations are obtained rather than exact draws.

The first approximation simply truncated the distribution of the mixture weights at some value $M$. The larger the $M$, the better the approximation, which was shown by obtaining a formula for the TV distance which decreases with increasing $M$. A modified algorithm was developed which samples jointly from both the mixture weights and the distribution given the mixture weights with finite expected running time, which is an improvement over the other algorithms discussed in this chapter. Further, the total variation distance from the target distribution can easily be estimated for the truncation approximation without knowledge of the true target distribution.
The second approximation uses an empirical estimate of the distribution of the mixture weights. The approximation clearly improves as a larger sample is used to construct the empirical estimate. This method has the advantage that, once a suitable empirical estimate is constructed, an arbitrarily large number of draws from the distribution of mixture weights may be obtained.

The next approximations constructed Markov chains which are stationary for the distribution of mixture weights. The challenge is computing acceptance probabilities for Metropolis-Hastings algorithms when the target distribution is not directly computable, even up to a constant. The approaches developed include a reverse rejection sampler, a random walk, and a pseudo-marginal independence chain.

Two examples were provided for each of the approximations discussed in this chapter. The first was with a known distribution of the mixture weights so performance could be directly compared between the approximations. Overall, the truncation approximation appeared to be the best performing method. The best Markov chain seemed to be the random walk, though the pseudo-marginal independence algorithm also performed quite well. The second example was the Bayesian random effects model considered in Chapter 2. Overall, these approximations can be incredibly useful for beginning a Markov chain close to the target distribution, answering Question 1, and they may be used whenever a general minorization condition is satisfied.

Further work with these approximate methods, particularly the Markov Chains, will hopefully yield a method for drawing exactly from the target distribution \( \pi \), even when an overall chain is neither geometrically nor uniformly ergodic but satisfies a minorization condition.
Chapter 4 provided an introduction to Bayesian scale-usage models, including a discussion on basic methods for analyzing ordinal categorical responses and how they can be adjusted for scale-usage heterogeneity. Several scale-usage models were discussed, along with computational strategies for each. The decomposition method of Hans et al. (2012) provides several computational advantages over traditional methods.

Proofs of geometric ergodicity of several scale-usage models with various model assumptions were then established. Equipped with a proof of geometric ergodicity, answers to Question 1 and Question 2 may readily be obtained for these models. Additionally, a minorization condition was proved for a scale-usage model which would enable the methods of Chapter 3 to be explored.

Chapter 5 provides an extension to the traditional scale-usage models which account for scale-usage heterogeneity patterns within a demographic. Several models were proposed, and a simulation study was conducted to compare them and to evaluate proposed model selection criteria.

These results were then used to analyze a real dataset composed of student evaluations of an instructor. A model was selected following the best practices established in the simulation study. The demographic with scale-usage heterogeneity differing most across demographic levels was cumulative GPA. Several inference strategies were then explored for these data using the ultimately selected model. Thorough answers to Question 1 and Question 2 were provided for the inference, made possible by the theory established in Chapter 4.

The overarching conclusions of this dissertation are that, while rigorous theory is required to establish answers to Question 1 and Question 2, the effort is worthwhile
because inference based on Markov chain Monte Carlo methods may then be trusted. As alluded to in Jones and Hobert (2001a), this is the “honest” way to perform inference. Answers to Question 1 and Question 2 allow conclusions to be made in finite time when infinity is too long to wait to guarantee Markov chain and ergodic average convergence.

6.2 Future Work

Two extensions of this work are now discussed. The first somewhat relieves the theoretical burden required in this dissertation in order to answer Question 1 and Question 2. The second provides a more flexible modeling framework for Bayesian scale-usage models.

6.2.1 Ensuring Uniform Ergodicity without the Theoretical Legwork

One observation from this dissertation (see Chapter 2 and Chapter 4) is that the proof of geometric or uniform ergodicity for a particular MCMC sampler is often nontrivial. Establishing the required theory is typically the biggest hurdle when answering Question 1 and Question 2 for a particular sampler. As such, a widely applicable method for ensuring these convergence rates are satisfied without the theoretical burden would be immensely useful. Consider the following proposition.

**Proposition 6.1.** Consider a Bayesian model with data $Y$ and parameters $\theta$. Let the data likelihood $f(Y|\theta)$ be bounded above in $\theta$ by a constant $M$, and assume that it is possible to sample from the prior distribution $\pi(\theta)$. Also assume that the
marginal probability of $Y$, $m(Y)$, is strictly greater than zero. Then, the independence Metropolis-Hastings algorithm stationary for $\pi(\theta|Y)$ using candidate distribution $\pi(\theta)$ is uniformly ergodic.

**Proof.** Let the assumptions of Proposition 6.1 hold. Then, note that

$$\frac{\pi(\theta)}{\pi(\theta|Y)} = \frac{\pi(\theta)m(Y)}{f(Y|\theta)\pi(\theta)} \geq \frac{m(Y)}{M} := \beta > 0.$$ 

Then, by Theorem 2.42 which was provided by Mengersen and Tweedie (1996), the corresponding Markov chain is uniformly ergodic.

Proof. Let the assumptions of Proposition 6.1 hold. Then, note that

$$\frac{\pi(\theta)}{\pi(\theta|Y)} = \frac{\pi(\theta)m(Y)}{f(Y|\theta)\pi(\theta)} \geq \frac{m(Y)}{M} := \beta > 0.$$ 

Then, by Theorem 2.42 which was provided by Mengersen and Tweedie (1996), the corresponding Markov chain is uniformly ergodic.

Note that if a maximum likelihood estimate $\hat{\theta}$ exists, then $M = f(Y|\hat{\theta})$ satisfies the condition. Further, note that if $\beta$ were computable, then exact samples could be obtained from the posterior distribution using the methods of Chapter 3; additionally, a rejection sampler would be available. Most of the time, however, $\beta$ will be unavailable directly. Note that $\beta$ is never required to run the independence Metropolis-Hastings chain. The acceptance probability of this chain for a new candidate $\theta$ drawn from the prior distribution $\pi(\theta)$, given the current draw $\theta'$ is

$$\alpha(\theta', \theta) = \min \left( 1, \frac{\pi(\theta|Y) \pi(\theta')}{\pi(\theta|Y) \pi(\theta)} \right) = \min \left( 1, \frac{f(Y|\theta) f(Y|\theta')}{f(Y|\theta')} \right),$$

which is free of $\beta$.

One remark about this Markov chain is that, though uniformly ergodic, it can nevertheless converge somewhat slowly compared to other Markov chains. In these situations, it can be helpful to consider this Markov chain, which is easily shown to be uniformly ergodic, in addition to a more quickly converging chain for which the convergence properties are not readily available. Mykland et al. (1995) discuss
two intuitive methods for combining these Markov chains such that the overarching
Markov chain is still uniformly ergodic.

**Proposition 6.2** (Mykland et al., 1995, Proposition 2). Let $K_1$ and $K_2$ be transition
kernels with $\pi$ as their invariant distribution. Suppose that the kernel $K_1$ satisfies a
minorization condition (3.1) with function $s(x)$ and measure $Q$. Then, the cycle kernel
$K_1K_2$ satisfies a minorization condition with function $s(x)$ and measure $QK_2$, and
the mixture kernel $\alpha K_1 + (1 - \alpha)K_2$ satisfies a minorization condition with function
$\alpha s(x)$ and measure $Q$ where $0 < \alpha \leq 1$.

This proposition also guarantees the cycle kernel and mixture kernel are uniformly
ergodic if $K_1$ is uniformly ergodic because the minorization condition is satisfied for
the entire space and $\inf_X s(x) := \epsilon > 0$.

Therefore, any Markov chain stationary for $\pi(\theta|Y)$ for which the conditions of
Proposition 6.1 are satisfied may be made to be uniformly ergodic by either cycling or
mixing with the independence Metropolis-Hastings sampler using $\pi(\theta)$ as a candidate
function. This then ensures uniform ergodicity without the theoretical rigor typically
required in order to guarantee the existence of a Markov chain central limit theorem
and estimate the asymptotic variance.

Note that the conditions of Proposition 6.1 preclude the use of improper prior dis-
tributions as samples may not be taken directly from $\pi(\theta)$ to use as candidates for the
independence Metropolis-Hastings algorithm. Future work includes the consideration
of adaptations for use when improper priors are desired.

**Example 6.1**
A simple example is now provided illustrating that a Markov chain which does not
satisfy a Markov chain central limit theorem can be mixed with the uniformly er-
godic independence Metropolis-Hastings chain to ensure a CLT exists. Consider the
following model. Let the data $y_i$ for $i = 1, \ldots, n$ satisfy

$$y_i \sim \text{Normal}(\mu, 1)$$

with

$$\mu \sim \text{Normal}(\mu_0, 1).$$

The posterior distribution is then

$$\pi(\mu|y) \propto \exp \left( -\frac{1}{2} \sum_{i=1}^{n} (y_i - \mu)^2 \right) \exp \left( -\frac{1}{2} (\mu - \mu_0)^2 \right),$$

which implies that

$$\mu|y \sim \text{Normal}\left((n + 1)^{-1}(n\bar{y} + \mu_0), (n + 1)^{-1}\right).$$

Clearly, the posterior distribution may be sampled from directly in this case. Con-
sider, however, an independence Metropolis-Hastings chain with a normal candidate
distribution with the same variance as the posterior distribution, but with a shifted
mean $(n + 1)^{-1}(n\bar{y} + \mu_0) + 1$. Mengersen and Tweedie (1996) showed that this Markov
chain is neither uniformly nor geometrically ergodic.

Now, consider the independence Metropolis-Hastings chain with Normal($\mu_0, 1$)
candidate distribution. By Proposition 6.1, this Markov chain is uniformly ergodic.
Consider then the modified kernel which mixes the original chain with the uniformly
ergodic chain with probability $\alpha = 0.01$ of using the uniformly ergodic chain. This
modified kernel is then uniformly ergodic.

To illustrate that a CLT exists for the modified mixture chain but not for the
original chain alone, consider Figure 6.1 which provides a density estimate—based on
a sample of 100,000 chains started in stationarity—for the Markov chain Monte Carlo average after 1 million iterations for both the original chain (left) and the modified chain (right). The distribution for the original chain is clearly skewed whereas the distribution for the modified chain appears to be Gaussian, as expected by the theory.

![Graph](image)

**Figure 6.1:** Comparison of distributions of ergodic averages after 1 million iterations for both the original chain (left) and the modified, uniformly ergodic chain (right).

To visualize the effect of the uniformly ergodic chain on the original chain in the mixture, consider the trace plots in Figure 6.2. The trace plot for the original chain is on the left, with the similar trace plot for the mixture chain on the right. It is clear that the original chain occasionally gets stuck for many iterations on the same value. The mixture chain with occasional updates from a uniformly ergodic chain greatly attenuates this problem, as seen on the right, even with a small mixing probability. This improved mixing behavior is enough to guarantee the existence of a central limit theorem.
Figure 6.2: Comparison of the trace plots for the original Markov chain and the modified chain which is uniformly ergodic.

This example has illustrated that mixing an inferior chain with a uniformly ergodic chain produces a uniformly ergodic chain for which a Markov chain CLT exists. The true value of this method, however, is that the theoretical justifications for a CLT are not required on the original chain.

This method could have been used to create a mixture chain with the original scale-usage algorithm proposed by Hans et al. (2012). The mixture chain is then assured to be uniformly ergodic without the theory discussed in Chapter 4 or the corresponding model restrictions (e.g. truncating certain parameters) required for the proof of geometric ergodicity. Similarly, the example in Section 2.5.1 would not have required drift and minorization proofs using this approach, and the answers to Question 1 and Question 2 could have been directly obtained. This approach could be used by software packages and general practitioners so that answers to Question 1
and Question 2 could always be provided when the assumptions of Proposition 6.1 are met without additional theoretical considerations or substantial increase in computational effort. Future work includes relaxing the assumptions of Proposition 6.1 and abundantly implementing this approach in practice so that answers to Question 1 and Question 2 accompany Bayesian inference on a more regular basis.

6.2.2 Nonparametric Bayesian Scale-Usage Models

Kottas et al. (2005) develop a nonparametric Bayesian model for the analysis of multivariate ordinal data. First, recall the definition and representations of a Dirichlet process, which are referenced from Müller and Rodriguez (2013). From Ferguson (1973) who originally developed the Dirichlet process, consider the probability space \((\mathcal{X}, \mathcal{B}(\mathcal{X}), G)\) and let \(\{A_1, \ldots, A_k\}\) be any finite partition of \(\mathcal{X}\). Then \(G\) satisfies a Dirichlet process prior with baseline measure \(G_0\) and mass \(M_0\), denoted by \(G \sim \text{DP}(M_0, G_0)\), if

\[
(G(A_1), \ldots, G(A_k)) \sim \text{Dirichlet}(M_0G_0(A_1), \ldots, M_0G_0(A_k)).
\]

The stick-breaking representation of a Dirichlet process was introduced by Sethuraman (1994). Specifically,

\[
G(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{\tilde{\theta}_h}(\cdot),
\]

where \(\delta_{\theta}\) is the Dirac mass at \(\theta\), \(\tilde{\theta}_h\) are \(i.i.d.\) samples from \(G_0\), and

\[
w_h = v_h \prod_{k<h} \{1 - v_k\}
\]

where

\[
v_h \overset{i.i.d.}{\sim} \text{Beta}(1, M).
\]

From this representation, it is clear to see that \(G\) is discrete.
Another popular and intuitive representation is the Pólya urn. Let there be an urn with \( M \) black balls and one colored ball distributed according to \( G_0 \). Then, draw a single ball from the urn. If the ball is a color, add back the ball with an additional ball of the same color. If the ball is black, then return the ball to the urn in addition to a ball of a new color distributed according to \( G_0 \). This corresponds to the following predictive representation. Let \( \theta_1, \theta_2, \ldots \) be an i.i.d. sample from \( G \). As \( G \) is almost surely discrete, there are ties among the sample. Define \( k_n \) as the number of unique \( \theta \) values among \( \theta_1, \ldots, \theta_n \) and denote these unique values as \( \{ \theta_{k_n}^*, \ldots, \theta_{k_n}^* \} \). Then, as discussed by Blackwell and MacQueen (1973), the predictive probability is

\[
p(\theta_{n+1}|\theta_1, \ldots, \theta_n) \propto \sum_{j=1}^{k_n} n_{n_j} \delta_{\theta_j^*} + MG_0,
\]

where \( n_{n_j} \) is the number of draws among the \( \theta_1, \ldots, \theta_n \) equal to \( \theta_j^* \). For additional discussion and representations of the Dirichlet process prior, see Müller and Rodriguez (2013).

With an understanding of Dirichlet processes, the ordinal categorical model of Kottas et al. (2005) may now be defined. As with the methods of Chapter 4 and Chapter 5, let \( X \) be the \( N \times M \) matrix of the \( N \) individual responses on \( M \) questions satisfying

\[
X_{ij} = \{ k : c_{j,k-1} < Y_{ij} \leq c_{j,k} \},
\]

where \( c_{j,0} < \ldots < c_{j,K_j} \) are the cutpoints for the \( j^{th} \) question. Then, define the distribution of the latent variables and the remainder of the model as

\[
Y_i|\theta_i \overset{\text{ind}}{\sim} \text{Normal}(\mu_i, \Sigma_i), \quad i = 1, \ldots, N, \quad (6.1)
\]

\[
\theta_i = (\mu_i, \Sigma_i)
\]

\[
\theta_1, \ldots, \theta_n|G \overset{\text{iid}}{\sim} G
\]
\[ G | M, m, V, S \sim DP(M_0, G_0) \]
\[ G_0(m, S) = \text{Normal}(\mu | m, V) \text{Inverse Wishart}(\Sigma | \delta, S) \]
\[ m \sim \text{Normal}(q, Q) \]
\[ V \sim \text{Inverse Wishart}(b, B) \]
\[ S \sim \text{Wishart}(c, C) \]
\[ M_0 \sim \text{Gamma}(a_0, b_0), \]

where \( \delta, q, b, c, a_0, \) and \( b_0 \) are all fixed constants and \( Q, B, \) and \( C \) are fixed \( M \times M \) matrices. For details on constructing a Markov chain stationary for the posterior distribution of this model, see Kottas et al. (2005) and, as a reference for sampling from Dirichlet processes, MacEachern and Müller (1998). Kottas et al. (2005) show that this flexible model may be used for any probability distribution of the raw ordinal responses. Furthermore, this is accomplished with arbitrarily fixed cutoffs, eliminating the need to sample the cutoffs which are often quite autocorrelated.

The parameter \( M \) controls the discreteness of the Dirichlet process. As \( M \to 0 \), the model with a single common \( \theta \) results (i.e. \( \theta_i = \theta \) for all \( i = 1, \ldots, N \)). As \( M \to \infty \), the model with each \( \theta_i \overset{i.i.d.}{\sim} G_0 \) results. As some clustering is likely among the latent variables, the Dirichlet process prior with finite and positive \( M \) is likely to outperform both of these special cases.

Consider now a case of scale-usage heterogeneity. Clearly, the model by Kottas et al. (2005) is flexible enough to account for scale usage heterogeneity with distinct \( \mu_i \) and \( \Sigma_i \) values. Yet, it might be beneficial to directly account for scale-usage heterogeneity by replacing (6.1) with

\[ \mathbf{Y}_i | \theta_i \overset{\text{ind}}{\sim} \text{Normal}(\mu_i + \tau_i, \sigma_i^2 \Sigma_i), \quad i = 1, \ldots, N, \]
where

\[ \tau_i \sim \text{Normal}(0, \sigma^2_\tau), \]

\[ \sigma^2_i \sim \text{Inverse Gamma}(a, a - 1). \]

Thus, in situations where, for some \(i\) and \(j\), \(\mu_i\) and \(\mu_j\) differ only by an additive constant (i.e. \(\mu_i = \mu_j + \ell_1\) for some \(\ell_1\)) and \(\Sigma_i\) and \(\Sigma_j\) differ only by a multiplicative constant (i.e. \(\Sigma_i = \ell_2 \Sigma_2\) for some \(\ell_2\)), as is the general premise with scale-usage models, the same discrete \(\theta^*\) can be utilized for both observations. This would then likely lead to fewer distinct values of \(\theta\) and allow for clustering after adjusting for scale-usage, which was discussed to be critical in Chapter 4 and Chapter 5, while still providing the model flexibility discussed by Kottas et al. (2005).

Future work includes exploration of this strategy including running a simulation study to evaluate its performance over the model by Kottas et al. (2005) and the traditional scale-usage models in Chapter 4. This method may then be applied to real data and evaluated in terms of performance in a realistic setting.
Bibliography


277


