The Generalized Multiset Sampler: Theory and Its Application

DISSERTATION

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

By

Hang Joon Kim, M.S.

Graduate Program in Statistics

The Ohio State University

2012

Dissertation Committee:

Steven N. MacEachern, Advisor
Christopher M. Hans
Mario Peruggia
Abstract

The multiset sampler (MSS) proposed by Leman et al. (2009) is a new MCMC algorithm, especially useful to draw samples from a multimodal distribution, and easy to implement. We generalize the algorithm by re-defining the MSS with an explicit description of the link between a target distribution and a limiting distribution. The generalized formulation makes the idea of the multiset (or $K$-tuple) applicable not only to Metropolis-Hastings algorithms, but also to other sampling methods, both static and adaptive. The basic properties of implied distributions and methods are provided. Drawing on results from importance sampling, we also create effective estimators for both the basic multiset sampler and the generalization we propose. Simulation and practical examples confirm that the generalized multiset sampler (GMSS) provides a general and easy approach to dealing with multimodality and improving a chain’s mixing.
This is dedicated to my wife Hyundo, my son Yoobin, and our parents.
Acknowledgments

First, I would like to thank my advisor, Dr. Steven MacEachern, for his guidance, encouragement, and support. Under his direction, I have grown as a Bayesian and an independent thinker. I also thank him for his friendship and care. I was able to freely discuss all my concerns, not just academic topics, with him, and he always encouraged me.

I am deeply grateful to Dr. Mario Peruggia and Dr. Chris Hans, who served on both my dissertation committee and my candidacy examination. Their valuable advice and encouragement helped me complete my dissertation. I also would like to thank my candidacy examination committee members, Dr. Laura Kubatko and Dr. Xinyi Xu, for their very helpful insights, comments, and assistance.

I would like to acknowledge the professors who gave me academic/personal advice and encouragement: Dr. Elizabeth Stasny, Dr. Yoonkyung Lee, Dr. Chris Holloman, Dr. Prem Goel, Dr. Radu Herbei, and Dr. Greg Allenby. I would never have survived the Ph.D. program without their support.

I am also grateful to all the people in the Department of Statistics at the Ohio State University for their support during my graduate study, including all faculty members, staff members, and many friends.
I also wish to thank all my colleagues in Nationwide CACI, especially Dr. Thomas Bishop, Dr. Chris Holloman, and Mr. Yiem Sunbhanich, who opened my eyes to the “real” analytic world. I really enjoyed my work at the center as a research assistant.

I also thank Dr. James Hodges and Dr. Brian Reich, who provided a periodontal data set and introduced me to a HMO example that is used in Chapter 7.

I dedicate this dissertation to my wonderful family: my wife, Hyundo, my son Yoobin, and our parents, all of whom gave me their constant love and support and kept me in their prayers.
Vita

February 2002 ......................... B.A. Business and Applied Statistics, Yonsei University, Korea
February 2006 ......................... M.S. Statistics, Yonsei University, Korea
August 2006 - present ............... Graduate Research Associate, The Ohio State University.

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>iii</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>iv</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
</tbody>
</table>

## 1. Introduction

1.1 Literature Review: Advanced MCMC

1.1.1 Parallel Tempering | 2
1.1.2 Multiple-Try Metropolis-Hastings | 6
1.1.3 The Slice Sampler | 7

1.2 Outline of the Dissertation | 8

## 2. Overview of the Multiset Sampler

2.1 Introduction | 9
2.2 Motivation: The Evolutionary Forest Algorithm | 10
2.3 Construction of the MSS | 13
2.4 Example 1: A Six-Modal Example | 15
2.5 Conclusion | 19
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. The Generalized Multiset Sampler</td>
<td>21</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>21</td>
</tr>
<tr>
<td>3.2 Formulation of the GMSS</td>
<td>21</td>
</tr>
<tr>
<td>3.3 GMSS Algorithms</td>
<td>25</td>
</tr>
<tr>
<td>3.4 Estimation</td>
<td>27</td>
</tr>
<tr>
<td>3.4.1 Revisit: A Six-Modal Example</td>
<td>29</td>
</tr>
<tr>
<td>3.5 Simulation Studies</td>
<td>34</td>
</tr>
<tr>
<td>3.5.1 Example 2: Mixture of Bivariate Normal Distributions</td>
<td>34</td>
</tr>
<tr>
<td>3.5.2 Example 3: Mixture Univariate Normal Example</td>
<td>35</td>
</tr>
<tr>
<td>3.6 Weights, Instrumental Densities, and Permutation Steps</td>
<td>40</td>
</tr>
<tr>
<td>3.7 Convergence of the Algorithm</td>
<td>45</td>
</tr>
<tr>
<td>3.7.1 Type II Update</td>
<td>45</td>
</tr>
<tr>
<td>3.7.2 Type I Update</td>
<td>47</td>
</tr>
<tr>
<td>3.7.3 Type III Update and Other Variants</td>
<td>50</td>
</tr>
<tr>
<td>3.8 Discussion</td>
<td>51</td>
</tr>
<tr>
<td>4. Application to Gene Expression Data</td>
<td>53</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>53</td>
</tr>
<tr>
<td>4.2 Hedenfalk’s Study: Breast Cancer Microarray Data</td>
<td>54</td>
</tr>
<tr>
<td>4.3 Multiple-hypothesis Testing and False Discovery Rate</td>
<td>55</td>
</tr>
<tr>
<td>4.4 Bayesian Approaches to Microarray Experiment</td>
<td>57</td>
</tr>
<tr>
<td>4.4.1 Local False Discovery Rate</td>
<td>57</td>
</tr>
<tr>
<td>4.4.2 The Mixture Density $h$ and the Null Density $h_0$</td>
<td>58</td>
</tr>
<tr>
<td>4.5 Example 4: Breast Cancer Microarray Experiment</td>
<td>60</td>
</tr>
<tr>
<td>4.5.1 Obtain Test Statistics</td>
<td>60</td>
</tr>
<tr>
<td>4.5.2 Bayesian Normal Mixture Model and MCMC Methods</td>
<td>62</td>
</tr>
<tr>
<td>4.5.3 Results</td>
<td>63</td>
</tr>
<tr>
<td>5. Application to Marketing Example: Simultaneous Equation</td>
<td>69</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>69</td>
</tr>
<tr>
<td>5.2 Example 5: Toy Example with Simultaneous Equation</td>
<td>70</td>
</tr>
<tr>
<td>5.3 Example 6: Simultaneous Demand and Supply Example</td>
<td>74</td>
</tr>
<tr>
<td>5.4 Discussion</td>
<td>82</td>
</tr>
<tr>
<td>6. Application to Bayesian Outlier Detection in Regression Analysis</td>
<td>84</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>84</td>
</tr>
<tr>
<td>6.2 Example 7: Rousseeuw Type Data and Bayesian Approach</td>
<td>85</td>
</tr>
<tr>
<td>6.2.1 Gibbs Sampling</td>
<td>88</td>
</tr>
</tbody>
</table>
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Mean estimate and total variation from 10,000 iterations after a burn-in of 1,000 iterations under Gibbs sampling, MH, and the GMSS. The GMSS leads to more accurate estimators.</td>
<td>68</td>
</tr>
<tr>
<td>5.1 Estimates from Metropolis-Hastings and the GMSS with the multiset of size ( K = 2 )</td>
<td>83</td>
</tr>
<tr>
<td>6.1 Least square estimator.</td>
<td>86</td>
</tr>
<tr>
<td>6.2 Posterior results from Gibbs sampling.</td>
<td>89</td>
</tr>
<tr>
<td>6.3 Posterior results from MH.</td>
<td>91</td>
</tr>
<tr>
<td>6.4 Posterior results from the GMSS.</td>
<td>94</td>
</tr>
<tr>
<td>7.1 Comparison of posterior summaries of the Gibbs samplers. Hodges (1998) ran 750 iterations with an additional 250 burn-in iterations. We re-run the Gibbs chain with 10,000 iterations.</td>
<td>107</td>
</tr>
<tr>
<td>7.2 Mean estimate and total variation of 100 independent chains from 2,000 iterations after a burn-in of 500 iterations under Gibbs, MH, and GMSS’s.</td>
<td>114</td>
</tr>
</tbody>
</table>
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Joint density $\pi(x,y)$. The contour lines show the six modes of the density. The density is positive over the entire region.</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Result from (a) ordinary Metropolis-Hastings and (b) multiset sampler with $K = 3$. Top panels: The solid line is the marginal density of $x$ and the histograms show draws from the marginal distribution $\pi(x)$. Bottom panels: The solid line is $\pi(y)$ and the dashed line (on the right panel only) is $\pi_L^*(y_k)$. The histograms show draws of $y$ and $y_k$ from the corresponding stationary distributions.</td>
<td>18</td>
</tr>
<tr>
<td>3.1 Solid curves are desired marginals of $\pi(y)$. Top panel: The attained average marginal of $\pi^*(y_k)$ using the MSS/GMSS with the multiset size $K = 3$. Middle panel: Estimate of $\pi(y)$ using Leman et al.’s method. Bottom panel: Estimate of $\pi(y)$ using the proposed method.</td>
<td>30</td>
</tr>
<tr>
<td>3.2 Top panel: The attained average marginal of $\pi^<em>(y_k)$ using the GMSS with the multiset size $K = 2$. Middle panel: Samples from the distribution for the first element of the multiset, $\pi^</em>(y_1)$, with $\alpha_1 = 0.9$. Bottom panel: Samples from the distribution for the second element, $\pi^*(y_2)$, with $\alpha_2 = 0.1$.</td>
<td>33</td>
</tr>
<tr>
<td>3.3 The twenty simulated mean vectors are drawn as empty circles. Top left: Ordinary MH with proposal scale $\tau = 0.5$, Top right: Ordinary MH with proposal scale $\tau = 1.0$, Bottom left: GMSS with proposal scale $\tau = 0.5$, Bottom right: GMSS with proposal scale $\tau = 1.0$.</td>
<td>36</td>
</tr>
<tr>
<td>3.4 Two algorithms for mixture normal estimation. The Gibbs sampler is used in the top panel; the GMSS is used in the bottom panel. The GMSS uncovers all four means and moves between different three-mean explanations of the data.</td>
<td>38</td>
</tr>
</tbody>
</table>
3.5 The trajectory of the multiset elements, \( \mu_1 \) and \( \mu_2 \), from the GMSS algorithm. .......................................................... 39

3.6 The mixing of GMSS with \( K = 2 \) through the different \( f_k, \alpha_k \), and
the switching step. UU refers to use of the instrumental densities
\( (f_{\text{unif}}, f_{\text{unif}}) \), GU of \( (f_{\text{good}}, f_{\text{unif}}) \), GG of \( (f_{\text{good}}, f_{\text{good}}) \), and BU of \( (f_{\text{bad}}, f_{\text{unif}}) \).
Solid lines are the algorithms with permutation steps and dotted lines
are without permutation steps. The solid circle refers to the MSS as it
has uniform \( f_k \) and equal \( \alpha_k \). .................................................. 42

4.1 Posterior densities from each of 5 replicates of Gibbs sampling, MH
and the GMSS with \( 10^7 \) iterations. The comparability of these density
estimates establishes the variability of this as a target. .................. 64

4.2 (Left) Kernel density estimates of \( \mu \) from 10,000 iterations after a burn-in
of 1,000 iterations under Gibbs sampling, MH, and the GMSS. The
solid line is the average of density estimates from five independent
Gibbs runs, each of \( 10^7 \) iterations. (Right) A trajectory of 11,000
draws from each algorithm. ..................................................... 65

5.1 Samples from ordinary Metropolis-Hastings with different starting points:
(a) \( \phi^{(0)} = (1,1,1), \psi^{(0)} = (3,3,3) \) and (b) \( \phi^{(0)} = (0.5,0.5,0.5), \psi^{(0)} = 
(0.5,0.5,0.5) \). The dotted lines denote the true values of parameters.
Paths beginning at different starting values exhibit different behavior. 72

5.2 Contour plot of the log posterior density as a function of the multiplicities for parameters, \( C_\phi \) and \( C_\psi \), where \( \phi = C_\phi(1,2,3)^T \) and
\( \psi = C_\psi(2,1,4)^T \). The plot shows four local modes. The thick dark
bands are extremely deep valleys. ............................................. 73

5.3 Samples from the GMSS with different starting points: (a) \( \phi^{(0)} =
(1,1,1), \psi^{(0)} = (3,3,3) \) and (b) \( \phi^{(0)} = (2,2,2), \psi^{(0)} = (-1,-1,-1), \)
\( \psi^* = \sum_k w_k \psi_k \prod_{l \neq k} f_l(\psi_l) \). The GMSS finds the global mode in the
both settings. ................................................................. 75

5.4 The trajectories of three elements of the multiset, \( \psi_k, k = 1,2,3 \), from
the GMSS algorithm with the starting point, \( \phi^{(0)} = (2,2,2) \) and \( \psi^{(0)} = 
(-1,-1,-1) \). The global mode (right side of plots) contains nearly all
of the posterior mass. ....................................................... 76
5.5 Relationship between $\gamma$ and $\beta_1$. The range of $\gamma$ is a real line but $\beta_1$ cannot have the value between -1 and 0. 

6.1 Rousseeuw type data. 40 good points from $y \sim N(2 + x, 0.6^2)$ and 10 bad points from $(x, y)^T \sim N((7, 2)^T, 0.1^2I)$. The dashed line is the least squares fit for all observations and the solid line is the least squares fit for the 40 good points only.

6.2 Trajectory of the MH for Rousseeuw type data. (a) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (5.5, -0.4, 1.0)$. (b) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (2.5, 0.8, 0.3)$.

6.3 Trajectory of GMSS estimates for Rousseeuw type data. (a) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (5.5, -0.4, 1.0)$. (b) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (2.5, 0.8, 0.3)$.

6.4 Simulated data for 40 good points and 10 bad points with (a) $(\sigma, \nu) = (0.4, 0.1)$, (b) $(\sigma, \nu) = (0.6, 0.1)$, (c) $(\sigma, \nu) = (0.8, 0.1)$, and (d) $(\sigma, \nu) = (1.0, 0.1)$.

6.5 The GMSS posterior density of $\beta_1$ from a GMSS run of length $10^5$, under the simulated data for 40 good points and 10 bad points with (a) $(\sigma, \nu) = (0.4, 0.1)$, (b) $(\sigma, \nu) = (0.6, 0.1)$, (c) $(\sigma, \nu) = (0.8, 0.1)$, and (d) $(\sigma, \nu) = (1.0, 0.1)$.

6.6 Simulated data under the parameter $(\sigma, \nu) = (0.4, 0.1)$ with (a) 40 good points and 10 bad points, (b) 37 good points and 13 bad points, (c) 34 good points and 16 bad points, and (d) 31 good points and 19 bad points.

6.7 The GMSS posterior density of $\beta_1$ from a GMSS run of length $10^5$, under the simulated data with (a) 40 good points and 10 bad points, (b) 37 good points and 13 bad points, (c) 34 good points and 16 bad points, and (d) 31 good points and 19 bad points.

6.8 The GMSS posterior densities of $\beta_0$ and $\beta_1$ from a GMSS run of length $3 \times 10^6$ (left panels) and the fitted regression lines (right panel), under the simulated data with 34 good points and 16 bad points. The dashed line is the regression fit from the global mode $(\beta_0 = 4.3, \beta_1 = -0.33)$ and the dotted line is the regression fit from the local mode $(\beta_0 = 5.7, \beta_1 = -0.53)$. 

xiii
6.9 The GMSS posterior densities of $\beta_0$ and $\beta_1$ from a GMSS run of length $3 \times 10^6$ (left panels) and the fitted regression lines (right panel), under the simulated data with 31 good points and 19 bad points. The dashed line is the regression fit from the global mode ($\beta_0 = 3.9, \beta_1 = -0.26$) and the dotted line is the regression fit from the local mode ($\beta_0 = 6.0, \beta_1 = -0.56$).

7.1 Trajectories of the Gibbs chain with the HMOs data set. The trajectories show the existence of two modes. The Gibbs chain rarely switches modes.

7.2 Posterior densities from five replicates of the MH algorithm with $5 \times 10^7$ iterations. There is no visible difference between the five replicates. The average of the kernel density estimates is used as the empirical target density to compare the MCMC methods.

7.3 Trajectories for log $\sigma^2$ and log $\tau^2$ from the Gibbs sampler (top panels), the MH (middle panels), and the GMSS (bottom panels). Mixing improves as we move from up to bottom.

7.4 Total variation distance between estimated and target distributions for the log $\tau^2$ under MCMC algorithms with different proposal scales. The dotted line is for MH with 2,000 iterations, the dashed line is for MH with 2,600 iterations, and the solid line is for the GMSS with 2,000 iterations.

7.5 The spatial map of AL measurements when a patient has no missing teeth. The number in a box denotes teeth number. The number around a circle denotes AL measurement site.

7.6 Two types of relations. The solid line is the Type I neighbor relation and the dashed line is the Type II neighbor relation.

7.7 2D and 3D plots of the log-posterior of $z_1$ and $z_2$.

7.8 Trajectory of the MH.

7.9 Trajectory of the GMSS of size $K = 3$. 

xiv
Chapter 1: Introduction

Since the early work of Geman and Geman (1984) and Gelfand and Smith (1990), Markov chain Monte Carlo (MCMC) algorithms have become one of the most popular classes of algorithms used by statisticians to fit models. The algorithms are most commonly used to fit Bayesian models, but they have many other uses. To work well, an algorithm must produce a Markov chain that runs quickly and that mixes well. Consequently, a large amount of research has been done on the construction of algorithms.

In spite of this research, there is still a collection of problems for which satisfactory algorithms have not yet been developed. One problem in this collection is a multimodal, local-trap problem. It is well known that an ordinary MCMC sampler run on a multimodal limiting distribution can suffer from the local-trap problem — that is, the chain is easily trapped in a local mode, essentially rarely or never escaping to visit other modes of the distribution.

This dissertation focuses on the development of an algorithm, or more properly a strategy to produce algorithms, that often enable one to overcome the local-trap problem. Attention is given to construction of the Markov chain, the impact of tuning parameters on mixing of the chain, inference arising from the chain, theoretical properties of the chain, and general implementation issues. The novel algorithms are
implemented, and their performance is illustrated on a number of examples.

1.1 Literature Review: Advanced MCMC

To solve the local-trap problem for a multimodal target density, many advanced MCMC methods have been proposed. One set of algorithms uses a *temperature ladder*: simulated annealing (Kirkpatrick, Gelatt and Vecchi 1983), simulated tempering (Marinari and Parisi 1992), parallel tempering (Geyer 1991), evolutionary Monte Carlo (Liang and Wong 2000), and the equi-energy sampler (Kou, Zhou and Wong 2006). Another class of approaches to improve mixing of the chain relies on directional information from a target distribution: the hit-and-run algorithm (Chen and Schmeiser 1993; Berger 1993), the Langevin algorithm (Grenander and Miller 1994; Besag 1994; Roberts and Starmer 2002), the adaptive direction sampler (Gilks, Roberts and George 1994), and the multiple-try Metropolis algorithm (Liu, Liang and Wong 2000). In addition, the slice sampler (Higdon 1998; Neal 2003) and reversible jump MCMC (Green 1995) can sometimes deal with the local-trap problem. We refer the reader to Liang, Liu and Carroll (2010) for further details about advanced MCMC methods.

1.1.1 Parallel Tempering

*Using a temperature ladder*

Assume that we are interested in a target distribution $\pi(x)$. If $\pi(x)$ has a rugged shape with multiple steep peaks and valleys, a chain from an ordinary MCMC method, such as the Gibbs sampler or the random-walk Metropolis-Hastings (MH) algorithm,
is easily trapped in a local mode. Instead of sampling directly from $\pi(x)$, we can use a sequence of new distributions as follows:

$$P(x, m) = g_m \exp\left(-\frac{H(x)}{T_m}\right), \ m = 1, \ldots, M$$  \hspace{1cm} (1.1)

where $H(x) = -\log \pi(x)$ is called the energy function, $T_m$ is a rung on the temperature ladder, and $g_m$ is a normalizing constant. Conventionally, $T_1 > T_2 > \ldots > T_M = 1$.

Note that the target distribution $\pi(x)$ is the same as the stationary distribution $P(x, m = M)$ at the lowest temperature level $T_M = 1$.

The intuition underlying temperature-based MCMC methods is that a Markov chain at a high temperature moves more freely about the space than at a low temperature. Let us think about the movement of $x$ within the same temperature level $T_m$, from $(x^{(t)}, m)$ to $(x', m)$, with a symmetric proposal density. Then the acceptance ratio is as follows:

$$\frac{P(x', m)}{P(x^{(t)}, m)} = \exp\left(-\frac{H(x') - H(x^{(t)})}{T_m}\right).$$ \hspace{1cm} (1.2)

A higher temperature flattens the surface. The movement at the high temperature will be free as the acceptance ratio under a symmetric proposal distribution is near one, while the movement at the lowest temperature level $T_M = 1$ is just an ordinary MH update.

\textit{Simulated tempering}

The simulation method using a temperature ladder was first introduced by Kirkpatrick et al. (1983). They proposed \textit{simulated annealing}, a heuristic method for a global optimization problem that mimics the physical process of annealing in metallurgy. The idea was then developed in \textit{simulated tempering} proposed by Marinari and
Parisi (1992). Let the current state be \((x(t), T_m(t))\). Then, the simulated tempering algorithm consists of two movements as follows:

- Propose \(x'\) from \(q(|x(t))\) and accept it with the acceptance probability \(\gamma = \min(1, \delta)\) where
  \[
  \delta = \exp \left[ -\frac{H(x') - H(x(t))}{T_m} \right] \frac{q(x(t)|x')}{q(x'|x(t))}.
  \]

- Propose \(m'\) from \(q(|m(t))\) and accept it with the acceptance probability \(\gamma = \min(1, \delta)\) where
  \[
  \delta = \frac{g_{m'}}{g_m} \exp \left[ -H(x^{(t+1)}) \left( \frac{1}{T_{m'}} - \frac{1}{T_{m(t)}} \right) \right] \frac{q(m(t)|m')}{q(m'|m(t))}.
  \]

As the simulated tempering chain runs on \((x, m)\) space, or equivalently \((x, T_m)\) space, the chain at a high temperature level easily jumps to other modes, and the samples are transmitted up to the lowest temperature density, which is a target distribution, through a sequence of temperature updating operations. Especially for distributions whose shape is rugged, these temperature-based-algorithms are known to converge much faster than does the ordinary MH algorithm.

**Parallel tempering**

Geyer (1991) proposed *parallel tempering* (PT), another MCMC method based on a temperature ladder. The main difference between PT and simulated tempering is that PT has multiple chains running simultaneously, with each chain having a different temperature level. Assume that we are interested in a target distribution \(\pi(x)\). PT simulates, in parallel, a sequence of stationary distributions

\[
\pi_m(x) \propto \exp \left( \frac{\log \pi(x)}{T_m} \right), \quad m = 1, \ldots, M,
\]
where $T_m$ is a temperature level associated with the distribution $\pi_m(x)$. Conventionally, $T_1 > T_2 > \ldots > T_{M-1} > T_M \equiv 1$, so $\pi_M(x)$ is the same as the target distribution $\pi(x)$.

Let $x^{(t)} = (x_1^{(t)}, \ldots, x_M^{(t)})$ be the current sample. Then, parallel tempering consists of the following steps:

1. Update $x_m^{(t)}$ to $x_m^{(t+1)}$ via an MH algorithm for $m = 1, \ldots, M$.

2. Exchange $x_m^{(t+1)}$ with its neighbors: Set $m' = m - 1$ or $m + 1$ according to probabilities $q(m, m')$ where $q(m, m + 1) = q(m, m - 1) = 0.5$ for $1 < m < M$ and $q(1, 2) = q(M, M - 1) = 1$. The acceptance probability for the proposal $m'$ (and hence $x_m^{(t')}$) is $\gamma = \min(1, \delta)$ where

\[
\delta = \exp \left( \frac{\log(\pi(x_m)) - \log(\pi(x_{m'}))}{T_m} \right) = \exp \left( \frac{\log(\pi(x_m)) - \log(\pi(x_{m'}))}{T_m} \right) \cdot \exp \left( \frac{-\log(\pi(x_m)) + \log(\pi(x_{m'}))}{T_m} \right) \cdot \left( \frac{1}{T_m} - \frac{1}{T_{m'}} \right).
\]

**Implementation issues**

To run temperature-based methods, including PT, the rungs on the temperature ladder need to be chosen *a priori* and carefully. Liang et al. (2010) recommended to set the temperature levels $T_m$ such that $Var_{\pi_m} \left[ -\log \pi(x) \left( 1/T_{m+1} - 1/T_m \right)^2 \right] = O(1), \ m = 1, \ldots, M - 1$. A preliminary run of the sampler is used at each temperature level $T_m$ to roughly estimate the variance.

In a poorly designed temperature ladder, the acceptance rate of the proposed exchange is often too low, so transition between different temperature levels is difficult. To alleviate this difficulty, the adjacent distributions $f(x, T_i)$ and $f(x, T_{i+1})$ should have considerable overlap (refer to Liang et al. 2010). Having the desirable structure of the temperature ladder increases the efficiency of the algorithm. However,
having too many intermediate temperature levels adversely affects the efficiency of the algorithm. Moreover, Liang et al. (2010) pointed out that, even in the case that the temperature ladder is set well, the expected waiting time for a traversal of the temperature ladder is of the order $O(M^2)$.

### 1.1.2 Multiple-Try Metropolis-Hastings

The ordinary MH algorithm relies on a proposal distribution which, in a typical implementation, does not consider the shape of a target distribution. There are ways to systematically create a proposal such as the hit-and-run algorithm (Chen and Schmeiser 1993; Berger 1993) that creates a proposed jump with a candidate direction and a candidate distance. Besag (1994) also suggested the use of the gradient of $f(x)$ from the Langevin diffusion process in physics to suggest a candidate move for MH. However, gradient information for a target distribution is not easily available in practice.

In an attempt to create a better proposal distribution, Liu et al. (2000) proposed the multiple-try Metropolis (MTM), which approximates the gradient of the target distribution by Monte Carlo samples. Define $w(x, x') = f(x)q(x'|x)\lambda(x, x')$ where $\lambda(x, x')$ is a nonnegative symmetric function in $x$ and $x'$, satisfying $\lambda(x, x') > 0$ whenever $q(x'|x) > 0$. Let the current state of $x$ be $x^{(t)}$. Then MTM consists of the following steps:

1. Draw $M$ i.i.d. trial proposals $x'_1, \ldots, x'_M$ from $q(\cdot|x)$. Compute $w_m = w(x'_m, x)$ for $m = 1, \ldots, M$. 

2. Select $x' = x_j'$ among the trial set \{x_1', \ldots, x_M'\} with probability proportional to $w_m$. Then draw $x_1^*, \ldots, x_{m-1}^*$ from $q(x'|x)$, set $x_m^* = x$, and compute $w_m^* = w(x_m^*, x')$ for $m = 1, \ldots, M$.

3. Accept $x'$ with the probability $\gamma = \min(1, \delta)$ where

$$\delta = \frac{\sum_{m=1}^{M} w_m}{\sum_{m=1}^{M} w_m^*}.$$  

This algorithm is useful in that it allows one to use a large search region around the current state, while the chain still retains a reasonable acceptance rate. However, the performance of the MTM depends heavily on the number of proposals, $M$. In some difficult situations, finding an efficient direction for the next step may require $M$ to be large, resulting in a large computational burden.

1.1.3 The Slice Sampler

To facilitate moves between different modes, the slice sampler (Higdon 1998; Neal 2003) expands the model with the addition of an auxiliary variable $u$. Suppose that one wishes to sample from a target density $\pi(x)$. If we set $\pi(x, u) \propto I[(x, u) : 0 \leq u \leq \pi(x)]$, then

$$\pi(x) = \int \pi(x, u) du = \int_0^{\pi(x)} du.$$  

This implies that the target distribution $\pi(x)$ is regarded as the marginal distribution of the joint distribution 

$$(X, U) \sim Unif[(x, u) : 0 < u < \pi(x)].$$

Alternatively, we can view this method as operating on a uniform distribution over the probability histogram.
The slice sampler is constructed as a Gibbs sampler which runs on \((X, U):\)

1. Draw \(u^{(t)} \sim \text{Unif}[0, \pi(x^{(t-1)})].\)

2. Draw \(x^{(t)} \sim \text{Unif}\{x : \pi(x) \geq u^{(t)}\}.$

This sampler is efficient for multimodal distributions. The chain can move between modes within a slice and we do not need to manually tune the step size of a proposal density. However, the slice sampler is often difficult to implement in practice. Liang et al. (2010) pointed out that sampling uniformly from the region \(\{x : \pi(x) \geq u^{(t)}\}\) is as difficult as sampling from the target distribution \(\pi(x)\) in many practical cases.

### 1.2 Outline of the Dissertation

The dissertation is organized as follows. Chapter 2 reviews the Multiset Sampler (MSS), which is a new MCMC algorithm proposed by Leman, Chen and Lavine (2009) to solve the local-trap problem. It is followed by the introduction of the generalized MSS (GMSS), our proposed method, in Chapter 3. Its formulation, updating algorithms, benefits and theoretical properties are presented, along with some simulation examples. Chapter 4 shows the applicability of the GMSS in practice to a breast cancer microarray study. The use of the GMSS in a Marketing example, for outlier detection and for the variance component model is illustrated in Chapters 5–7. Chapter 8 concludes the dissertation with an overview of the material and directions for future work.
Chapter 2: Overview of the Multiset Sampler

2.1 Introduction

The multiset sampler (MSS) is a MCMC algorithm designed to improve mixing by avoiding the local-trap. The MSS was first introduced by Leman, Uyenoyama, Lavine and Chen (2007) to make inference about an evolutionary process. Leman et al. (2009) described its construction as a general purpose algorithm.

Leman et al. (2009) presents the central idea of the MSS to draw multiple values of $y$’s from a new distribution, $\pi^*_L$, instead of drawing a single $y$ from the original target distribution $\pi$. The relationship between $\pi$ and $\pi^*_L$ is $\pi^*_L(\{y_1, \ldots, y_K\}) \propto \sum_k \pi(y_k)$. Leman et al. showed that this formulation allows flexibility for each $y_k$ to move freely on the entire space of the target distribution. This property works especially well in two main settings: when the posterior distribution (henceforth we describe the application to Bayesian models and adjust our language accordingly) has a valley that is difficult to move across, and when the posterior has a strong ridge that is difficult to move along.

In the previous chapter, we described several MCMC algorithms that are devised to improve a chain’s mixing, but such algorithms require detailed knowledge of the posterior, often need to be hand-tuned, and/or have large computational burdens. In
contrast, Leman et al. (2009) document the ease of use of the MSS and its success for a selection of difficult MCMC problems.

2.2 Motivation: The Evolutionary Forest Algorithm

Population genetics

Population genetics is a field of biology that studies allele frequency distribution in a population and its changes. The allele frequency is mainly affected by genetic drift and gene flow, which have the opposite effects. Genetic drift is an evolutionary process which leads to a gradual change in the allele frequency within a population as a result of chance. It can result in some gene variants being lost from a population and thereby reduces genetic variability within the population. When groups are isolated, their evolutionary paths diverge, the populations become genetically quite different from their ancestors, and so it leads to enhanced divergence. In contrast, gene flow is the exchange of genes between populations which may be accomplished through migration such as the movement of individuals or the exchange of pollen. Gene flow tends to increase genetic variability within the population. However, there is little effect if the allele frequencies in the populations are not much different from each other.

The different nature of the two evolutionary processes makes the study of population divergence difficult because it is challenging to deal simultaneously with a divergence time, \( t \), when the populations become isolated, and migration rate \( m \). Most of traditional models for population subdivision made one of two extreme assumptions: (i) the populations will become increasingly divergent without gene flow
after the populations are separated from a common ancestral population (the isolation model), or (ii) the populations continuously exchange genes at a constant rate \( m \) for a long period of time (the migration model). However, neither approach describes the full complexity of genealogy (refer to Nielsen and Wakeley 2001; Hey and Nielsen 2004).

**MCMC approaches to gene dynamics**

A variety of mathematical models have been developed to explain gene frequency dynamics. Especially, there is a set of methods that use Metropolis-Hastings algorithms (Kuhner, Yarnato and Felsenstein 1995; Beerli and Felsenstein 1999; Nielsen 2000; Nielsen and Wakeley 2001). Among the studies, Nielsen and Wakeley (2001) proposed a framework for obtaining joint estimates of divergence times and migration rates. They assumed that two populations are descended from an ancestral population at some time in the past (the isolation assumption) and that the populations may be connected by migration (the migration assumption). In order to obtain the joint estimates of divergence times and migration rates, they implemented a MCMC approach in a Bayesian framework.

Specifically, they aimed to estimate the demographic parameters

\[
\theta = \{t, m_1, m_2, N_1, N_2, N_A\}
\]

where \( t \) is the divergence time, \( m_j \) is the proportion of descendant population \( j \) that is replaced by migrants from the other population (the migration rates), \( N_j \) is the population size of the population \( j \), and \( N_A \) is the population size of the ancestral population. (Note that they rescaled the parameters to implement their algorithm, but we use these basic demographic parameters for ease of explanation.)
Assuming a prior distribution \(\pi(\theta)\), the posterior distribution used is \(\pi(\theta|D) = p(D|\theta)\pi(\theta)/p(D)\) given the observed sequence data \(D\). To approximate the posterior density \(\pi(\theta|D)\), Nielsen and Wakeley (2001) introduced a nuisance variable \(\eta\), the underlying gene genealogy, consisting of (i) tree topology, (ii) values for all of the branch lengths, and (iii) migration events. Then, the Markov chain is constructed by suggesting updates to the gene genealogy \(\eta\) and to the six demographic parameters \(\theta\). The nuisance parameter \(\eta\) must be integrated out to gain insight into the demographic parameters as

\[
\pi(\theta|D) \propto \pi(\theta) \int_{\eta \in \Omega_{\text{Tree}}} p(D|\theta, \eta) \pi(\eta) d\eta
\]  

where \(\Omega_{\text{Tree}}\) is the set of all possible gene genealogies. Nielsen and Wakeley (2001) pointed out that updates of \(\eta\) are more complicated than in the usual coalescence simulations (see Kingman 1982) because one needs to simulate the structure of all edges in the population at a particular time.

Nielsen and Wakeley’s algorithm does not work especially for data sets with multiple loci. A large number of loci make the mixing/convergence of the chain slow since the update of divergence time \(t\) requires updating the genealogical history \(\eta\), which increases the computational burden (see Nielsen and Wakeley 2001; Leman et al. 2007). As a remedy, Hey and Nielsen (2004) used parallel tempering (PT) that runs multiple chains, each having corresponding divergence time \(t_k\).

**The evolutionary forest algorithm**

Leman et al. (2007) tried the parallel tempering (PT) algorithm to fit the coalescence model of gene histories, but they found that the optimal tuning of PT is a difficult or
time-consuming task. Instead, they proposed the evolutionary forest (EF) algorithm, which is now known as the multiset sampler (MSS).

The EF algorithm updates parameter values based on a probability measure defined on multiple trees, called a forest of genealogies. It defines the space of forests as

\[ \Omega_{\text{Forest}} = \{ \bigcup_{k=1}^{K} \eta_k : \eta_k \in \Omega_T \} \]

where \( \eta_k \) is a tree used in Equation (2.1). The operator \( \bigcup \) denotes a union to form a multiset, an unordered collection which preserves multiplicity: for example, \( \eta_1 \bigcup \eta_2 \bigcup \eta_3 \bigcup \eta_4 = \{ \eta_1, \eta_1, \eta_2, \eta_3, \eta_4 \} \). (The formal notion of a multiset will be introduced in the upcoming Section 2.3.)

Let \( s \) be a multiset such that \( \eta \in s \) and \( s \in \Omega_{\text{Forest}} \). The joint posterior distribution of the multiset and parameter sets is assigned as

\[ \pi^*(s, \theta|D) = C \sum_{\eta \in s} \pi(\eta, \theta|D) \]  

where \( \eta \) is genealogical history, \( \theta \) is the vector of population parameters, and \( C \) is a normalizing constant. Using (2.2), the EF algorithm runs a Gibbs sampler which proposes a move in the parameter values, \( \theta \), and a move in the forest of trees, \( s = \{ \eta_1, \ldots, \eta_K \} \). Leman et al. (2007) studied the difference between PT and the EF algorithm in the sense that PT embodies multiple chains, each specifying a single tree, while a single chain of EF specifies a forest of multiple trees.

### 2.3 Construction of the MSS

The notion of a multiset allows members to appear more than once. Let \( Y \) be a random variable on the support \( \mathcal{Y} \) and \( S \) be a multiset that may contain multiple
copies of elements in $\mathcal{Y}$. For example, $\mathcal{Y}_1 = \{1, 2, 3, 4, 5\}$ and a multiset $S_1 = \{3, 2, 2, 5\}$. In a multiset, the order of elements is irrelevant, so another multiset $S_2 = \{2, 2, 3, 5\}$ is equal to $S_1$. Define $u(S)$ to be the unique elements of $S$ and $m(y, S)$ to be the multiplicity of $y$ in $S$. For example, $u(S_1) = \{2, 3, 5\}$, $m(1, S_1) = m(4, S_1) = 0$, $m(3, S_1) = m(5, S_1) = 1$, and $m(2, S_1) = 2$. The cardinality (roughly size) of the set $\mathcal{Y}_1$ is $|\mathcal{Y}_1| = 5$ and the cardinality of the multiset $S_1$ is $|S_1| = 4$. If $|\mathcal{Y}| = n$ and $|S| = K$, then the number of possible multisets $S$ on $\mathcal{Y}$ is given as $\binom{n + K - 1}{K}$.

By developing the idea of the evolutionary forest (EF) algorithm and adopting the notion of a multiset, Leman et al. (2009) proposed the multiset sampler (MSS). Let $\pi$ be a probability distribution of two random variables $X$ and $Y$ and suppose that we want to sample $x$ from a marginal distribution $\pi(x)$. We construct a multiset $s$ of $K$ values for $y$ with the conditional density

$$\pi^*_L(s|x) = \pi^*_L(\{y_1, y_2, \ldots, y_K\}|x) \equiv C \sum_{y \in s} \pi(y|x),$$

(2.3)

where $C$ is a normalizing constant. Instead of sampling directly from $\pi(x, y)$, the multiset sampler suggests that one sample from $\pi^*_L(x, s)$ defined as $\pi^*_L(x, s) \equiv \pi(x)\pi^*_L(s|x)$.

The MSS is a Metropolis-within-Gibbs sampler that runs on $(X, S)$ instead of $(X, Y)$. It alternates between $X$ moves and $S$ moves. The following steps illustrate the MSS, as implemented via the Metropolis-Hastings algorithm. The acceptance probabilities are calculated in the standard fashion.

1. Accept or reject $x'$ proposed from $g(\cdot|x, s)$ with target density $\pi^*_L(x'|s)$.

2. Randomly select $k \in \{1, \ldots, K\}$ with equal probabilities.
3. Randomly draw \( y'_k \) from a proposal distribution \( g(\cdot|x,s) \). The proposed multiset 
\[ s' = \{ y_1, \ldots, y'_k, \ldots, y_K \} \]

is found by replacing \( y_k \) in the current \( s \) with \( y'_k \). Then accept or reject \( s \) with the target distribution 
\[ \pi^*_L(s|x) \propto \sum_{k=1}^{K} \pi(y_k|x). \]

Since the proposal \( y'_k \) changes only one element of \( s \), only one term in \( \sum_{k=1}^{K} \pi(y_k|x) \) is changed. If \( y_k \) is associated with a relatively small value of \( \pi(y|x) \), it contributes little to \( \pi^*_L(s|x) \), the acceptance probability will be near one, and this component of the multiset can move freely. This component of the multiset can then explore the \( Y \) space, allowing the sampler to escape a local mode. The \( y_k \) associated with a large value of \( \pi(y|x) \) anchors the multiset, ensuring that a region of high probability will be represented by some \((x,y_k)\).

### 2.4 Example 1: A Six-Modal Example

The benefit of MSS is addressed by the following example, which is used by Leman et al. (2009). Assume a two-dimensional density with six well-separated modes is confined to the region \([-10,10] \times [-10,10]\) with the joint density
\[ \pi(x,y) \propto e^{-x^2/2}e^{-(\csc^5 y-x)^2/2}I_{[-10,10],(-10,10]}(x,y). \]

Figure 2.1 presents the joint density of \((X,Y)\).

Following Leman et al. (2009), the ordinary MH and the MSS are run. For the ordinary MH, a normal random-walk proposal distribution with scale \( \tau = 0.5 \) is used for updating \( x \) and \( y \). We draw 100,000 samples from the random-walk (RW) MH with the target distribution \( \pi(x,y) \) in a standard fashion. For the MSS, we construct
Figure 2.1: Joint density $\pi(x, y)$. The contour lines show the six modes of the density. The density is positive over the entire region.
a multiset of size $K = 3$, $s = \{y_1, y_2, y_3\}$, with the joint distribution

$$
\pi^*_L (x, s) \propto \pi(x, y_1) + \pi(x, y_2) + \pi(x, y_3).
$$

We also use the normal RW proposal distribution with scale $\tau = 0.5$ for updating $X$ and each $Y_k$.

Figure 2.2 shows the comparison between samples from the ordinary MH and the MSS. The solid curves represent the true marginals of $x$ and $y$, computed from a grid on the $(X, Y)$ space with a resolution of 0.1 on both axes. Panel (a) is the histogram of 100,000 sample draws from an ordinary Metropolis-within-Gibbs procedure. Clearly, the MH chain gets trapped in a local mode due to the well-separated modes and all of the resulting marginal distributions have failed to converge to their targets. In contrast, panel (b) shows that the MSS chain explores the space much better than the ordinary sampler and that the $x$ marginal converges well to its target. The top row of panel (b) shows the success of the MSS in capturing the distribution of $x$. In the bottom row, the departure of the sampled values from the target distribution of $y$ does not indicate a failure of the method, as the $y_k$ are drawn (marginally) from $\pi^*_L (y_k)$ (dashed line) rather than from $\pi(y)$ (solid line).

The main benefits the MSS from the simulation study is as follows. First, the MSS chain is less likely to be trapped in a local mode. As shown in Figure 2.2b, the MSS flattens the sampling distribution of $Y$ and makes the chain traverse the valleys, while one of the $y_k$’s in $s$ still remains in a mode, so it collect enough information of the target distribution around the local mode.

Second, the MSS is easy to implement. The MSS does not need to tune a parameter $a$ priori or during the runs. While PT needs a preliminary run to determine the structure of the temperature ladder and adaptive MCMC algorithms need to adjust
Figure 2.2: Result from (a) ordinary Metropolis-Hastings and (b) multiset sampler with $K = 3$. Top panels: The solid line is the marginal density of $x$ and the histograms show draws from the marginal distribution $\pi(x)$. Bottom panels: The solid line is $\pi(y)$ and the dashed line (on the right panel only) is $\pi_L^*(y_k)$. The histograms show draws of $y$ and $y_k$ from the corresponding stationary distributions.
proposal scales using the information from the past samples, the MSS does not need a careful parameter tuning. It is also less sensitive to the proposal distribution than are many other MCMC algorithms.

Third, it has been demonstrated empirically that, in some examples, the MSS has better convergence properties than do existing algorithms devised to deal with the local-trap problem. Leman et al. (2007) showed that the multiset sampler is computationally more efficient than is PT, in the example of a genealogy formulation. Leman et al. (2009) also compared the MSS with PT and the MTM using some simulated data sets and found that MSS has the least computation time in some examples.

2.5 Conclusion

Technical challenges remain for the MSS. First, the $y$ values must lie in a bounded region. If this is not the case, $\pi^*_L$ would not be integrable, would not be a probability distribution, and could not be evaluated with MCMC techniques. In practice, Leman et al. (2009) truncate the parameter space to allow use of the MSS.

Second, calculation of the normalizing constant $C$ is complex, due to the definition of the multiset $s$, relying on ordered and unordered versions of the $y_k$’s. The normalizing constant $C$ is expressed differently in the cases of $Y$ being discrete and $Y$ being continuous. If $\mathcal{Y}$ is a finite set, the number of each $y$ appearing in the collection of multisets is the number of possible multisets $S$ on $\mathcal{Y}$ multiplied by the average appearance of each $y$ in a multiset, $\frac{K}{n}$. Then, the multiset distribution is expressed as

\[
\pi^*_L(s|x) = C \sum_{y \in s} \pi(y|x) = C \sum_{y \in u(s)} \pi(y|x) \left( \frac{n + \frac{K - 1}{K} \frac{K}{n} \right). 
\]
Therefore, the normalizing constant in the discrete case is

\[ C = \left( \frac{n + K - 1}{K - 1} \right)^{-1}, \]

since \( \sum_s \pi^*_L(s|x) = 1 \) and \( \sum_s \sum_{y \in \text{u}(s)} \pi(y|x) = 1 \). If \( Y \) is continuous on \([a, b]\), Leman et al. (2009) used the fact that \( m(y, s) = 1 \) almost surely and the number of possible multisets is \( K! / \prod_{y \in \text{u}(s)} m(y, s)! \). Then, the normalizing constant in the continuous case is

\[ C = \frac{(K - 1)!}{(b - a)^{K - 1}}. \]

Third, Leman et al. assume \( X \) is only the parameter of interest. For example, in the coalescence model of Leman et al. (2007), they denote the population parameter as \( X \) and the underlying, unobserved genealogical tree as \( Y \). For their analysis, the only purpose of \( Y \) is to facilitate draws of \( X \). However, in practice there are many cases where parameters with multimodal distributions are also of interest. In Leman et al.’s method, density estimators of \( \pi(y) \) exhibit unusual behavior, even assuming negative values as shown in the following chapter.
Chapter 3: The Generalized Multiset Sampler

3.1 Introduction

Despite the novel idea and its many advantages, the MSS has some disadvantages: the variate $y$ must have bounded support, and estimation of $\pi(y)$ can be poor. The technical challenges for the MSS can be addressed and solved by the generalized multiset sampler (GMSS) proposed by Kim and MacEachern (2011).

We re-portray the MSS as providing a link between the target distribution and the sampling distribution. We then proceed to relax the rather stringent restrictions of the MSS in order to make connections with existing techniques. Eventually, we draw on results from the importance sampling literature to create effective estimators. Moreover, the GMSS can make sampling efficient by using an instrumental density to tune the link between $\pi(x,y)$ and $\pi^*(x,s)$.

3.2 Formulation of the GMSS

The descriptions of the MSS and the GMSS make use of different notation for continuous and discrete distributions. Here, we present the continuous version, noting
that for discrete problems, or partially discrete problems, integrals should be replaced with summations when appropriate.

As with the MSS, we split the parameter vector into two pieces, $X$ and $Y$. These elements may be scalars or vectors. We assume that $(X,Y)$ has a density with respect to Lebesgue measure, $\pi(x,y)$, and that interest centers on this density. The support of $X$ is $\mathcal{X}$, and the support of $Y$ is $\mathcal{Y}$. The joint support of $(X,Y)$ may be a proper subset of $\mathcal{X} \times \mathcal{Y}$. The MSS replaces $y$ with a multiset of cardinality $K$, leading to a density $\pi^*_L(x,y_1,\ldots,y_K)$, where $\{y_1,\ldots,y_K\}$ is a multiset. The relationship between $\pi^*_L$ and $\pi$ is established in (2.3).

In contrast, we replace $y$ with a $K$-tuple—a traditional vector of length $K$—denoted by $s = (y_1,\ldots,y_K)$. The extension presented here constructs the density on $(x,s)$ to be $\pi^*(x,s) = \pi^*(x)\pi^*(s|x)$, where

$$\pi^*(s|x) = \pi^*(y_1,\ldots,y_K|x) \equiv \sum_{k=1}^{K} \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l)$$

(3.1)

and $\pi^*(x) = \pi(x)$. The multiset weights $\alpha_k$ are constants on $[0,1]$ such that $\sum_{k=1}^{K} \alpha_k = 1$. We refer to the $f_k$ as instrumental densities. The density $f_k$ is a probability density, typically with support equal to $\mathcal{Y}$.

The relationship given in (3.1) defines a new object $\pi^*(x,s)$ and we now turn to properties of this object. First, $\pi^*(x,s)$ is a joint probability density function.

**Proposition 1.** Assume that $f_k$, $k = 1,\ldots,K$, is a probability density with support equal to $\mathcal{Y}$. Then the GMSS density $\pi^*(x,s)$ is a probability density function.

**Proof.** The right hand side of expression (3.1) is the sum of the product of non-negative components, so $\pi^*(x,s) = \pi^*(x)\pi^*(s|x) \geq 0$ for all $(x,s)$. Working with the
conditional density of $s|x$, 

$$\int_S \pi^*(s|x)ds = \int_Y \cdots \int_Y \sum_{k=1}^K \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) \, dy_1 \cdots dy_K$$

$$= \sum_{k=1}^K \alpha_k \left[ \int_Y \cdots \int_Y \pi(y_k|x) \prod_{l \neq k} f_l(y_l) \, dy_1 \cdots dy_K \right]$$

$$= \sum_{k=1}^K \alpha_k \left[ \int_Y \pi(y_k|x) \, dy_k \right]$$

$$= 1.$$

Consequently,

$$\int_S \int_X \pi^*(x, s)dx\,ds = \int_S \int_X \pi^*(x) \pi^*(s|x)dx\,ds$$

$$= \int_X \pi^*(x)dx$$

$$= 1.$$

Since $\pi^*(x, s)$ is a probability density, marginal densities exist for each of the components of $(x, s)$. From the definition of $\pi^*$, the marginal distribution for $x$ is just $\pi^*(x) = \pi(x)$. The marginal density for $y_k$ requires calculation.

**Proposition 2.** Assume that $f_k$, $k = 1, \ldots, K$, is a probability density with support equal to $\mathcal{Y}$. Then the marginal density for $Y_k$ is $\pi^*(y_k) = \alpha_k \pi(y_k) + (1 - \alpha_k) f_k(y_k)$, for $y_k \in \mathcal{Y}$. 

23
Proof. Without loss of generality, we take $k = 1$. We begin with the conditional density of $y_1|x$.

$$\pi^*(y_1|x) = \int_{y_1 \in s} \pi^*(s|x) ds$$

$$= \int_{y_1 \in s} \sum_{k=1}^K \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) ds$$

$$= \int_\mathcal{Y} \cdots \int_\mathcal{Y} \alpha_1 \pi(y_1|x) \prod_{l \neq 1} f_l(y_l) dy_2 \cdots dy_K$$

$$+ \sum_{k=2}^K \int_\mathcal{Y} \cdots \int_\mathcal{Y} \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) dy_2 \cdots dy_K$$

$$= \alpha_1 \pi(y_1|x) + \sum_{k=2}^K \alpha_k f_k(y_1)$$

$$= \alpha_1 \pi(y_1|x) + (1 - \alpha_1) f_1(y_1).$$

Passing to the marginal density for $Y_1$,

$$\pi^*(y_1) = \int_\mathcal{X} \pi^*(x) \pi^*(y_1|x) dx$$

$$= \int_\mathcal{X} \pi(x) [\alpha_1 \pi(y_1|x) + (1 - \alpha_1) f_1(y_1)] dx$$

$$= \alpha_1 \pi(y_1) + (1 - \alpha_1) f_1(y_1).$$

$$\square$$

Sliding over the distinction between a multiset and a K-tuple, the MSS uses the specific values $\alpha_k = 1/K$ for each $k$, and takes $f_k$ to be the uniform distribution over the set $\mathcal{Y}$. The limitation to the uniform distribution enforces the constraint found in the MSS of a bounded interval (more appropriately, finite Lebesgue measure) for $\mathcal{Y}$ and, in the discrete case, the finite cardinality of $\mathcal{Y}$.

We note that conditional densities under $\pi^*$ are found in the usual way, whether for $y_k|x$, for $(y_k, y_l|x)$, for $x|s$, or for some other object. Thus, any of the wide variety of MCMC moves and algorithms can be used to obtain samples from $\pi^*$. Moreover,
alternative techniques such as adaptive MCMC, perfect sampling, and particle filtering can also be used to sample \( \pi^* \).

### 3.3 GMSS Algorithms

The basic algorithm that we present is a Metropolis-Hastings algorithm. For each step below, the family of proposal distributions must be specified, the acceptance probability calculated, and acceptance-rejection of the proposed value determined. In all steps, the density which is to be evaluated follows from \( \pi^*(x, s) \). Our investigations suggest three types of sampling algorithms whose effectiveness varies with details of the problem, in particular, the dimension of the multiset.

**Type I Algorithm.**

1. Sample \( x^{(t)} \), conditional on \( s^{(t-1)} \).

2-1. Randomly select \( k \in \{1, \ldots, K\} \) with equal probabilities.

2-2. Sample \( y_k^{(t)} \), conditional on \( x^{(t)} \) and \( y_1^{(t-1)}, \ldots, y_k^{(t-1)}, y_{k+1}^{(t-1)}, \ldots, y_{K}^{(t-1)} \).

To generate \( y_k^{(t)} \) in step 2-2, first propose \( y_k' \sim q(y_k|y_k^{(t-1)}) \). The current multiset is \( s_k^{(t-1)} = (y_1^{(t-1)}, \ldots, y_K^{(t-1)}) \). The proposed multiset is \( s_k' = (y_1', \ldots, y_K') \) where \( y_i' = y_i^{(t-1)}, i \neq k \) and \( y_k^{(t-1)} \) is replaced by \( y_k' \). The acceptance ratio for the proposed \( y_k' \) (and hence \( s_k' \)) is \( \gamma = \min(1, \delta) \) where

\[
\delta = \frac{\pi^*(x, s_k') q(y_k^{(t-1)}|y_k')} {\pi^*(x, s_k^{(t-1)}) q(y_k'|y_k^{(t-1)})} = \frac{\sum_{j=1}^K \left[ \alpha_j \pi(y_j'|x) \prod_{i\neq j} f_i(y_i') \right] q(y_k^{(t-1)}|y_k')} {\sum_{j=1}^K \left[ \alpha_j \pi(y_j^{(t-1)}|x) \prod_{i\neq j} f_i(y_i^{(t-1)}) \right] q(y_k'|y_k^{(t-1)})}
\]
If the proposal density \( q(y_k|y_k^{(t-1)}) \) is symmetric, the acceptance ratio simplifies to

\[
\delta = \frac{\sum_{k=1}^{K} \left[ \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l') \right]}{\sum_{k=1}^{K} \left[ \alpha_k \pi(y_k^{(t-1)}|x) \prod_{l \neq k} f_l(y_l^{(t-1)}) \right]}.
\]

If accepted, set \( s_k^{(t)} = s_k' \); if rejected, set \( s_k^{(t)} = s_k^{(t-1)} \).

As an alternative to updating a single element of the multiset at a time, one can update multiple elements. The extreme is to update all \( K \) elements simultaneously, leading to the second type of algorithm.

**Type II Algorithm.**

1. Sample \( x^{(t)} \), conditional on \( s^{(t-1)} \).

2. Sample \( s^{(t)} = \{ y_1^{(t)}, \ldots, y_K^{(t)} \} \), conditional on \( x^{(t)} \).

Acceptance-rejection under the Type II algorithm is the same as Type I, except the proposed multiset is \( s' \sim q(s|s^{(t-1)}) \) where \( q(s|s^{(t-1)}) = \prod_k q(y_k|y_k^{(t-1)}) \). The acceptance ratio is

\[
\delta = \frac{\sum_{k=1}^{K} \left[ \alpha_k \pi(y_k'|x) \prod_{l \neq k} f_l(y_l') \right] q(s^{(t-1)}|s')}{\sum_{k=1}^{K} \left[ \alpha_k \pi(y_k^{(t-1)}|x) \prod_{l \neq k} f_l(y_l^{(t-1)}) \right] q(s'|s^{(t-1)})}.
\]

If accepted, set \( s^{(t)} = s' \); if rejected, set \( s^{(t)} = s^{(t-1)} \).

Performance of algorithms also varies with the forms of the instrumental densities \( f_k \) and the proposal distribution \( q \). Empirically, it has been found that when a uniform density of \( f_k \) is used, the chain with a Type II algorithm does not work well near the boundary. In such a case, we may update each element of the multiset in turn, leading to the third type of updating method.
Type III Algorithm.

1. Sample $x^{(t)}$, conditional on $s^{(t-1)}$.

2-1. Let $k = 1$.

2-2. Sample $y_k^{(t)}$, conditional on $x^{(t)}$ and $y_1^{(t)}, \ldots, y_{k-1}^{(t)}, y_{k+1}^{(t)}, \ldots, y_K^{(t-1)}$.

2-3. $k = k + 1$ and go to 2-2 if $k < K$.

Additional algorithms abound. In particular, $x$ and $y_k$ will often be multidimensional, and the established MCMC techniques of blocking and collapsing (see Liu 1994; Roberts and Sahu 1997) can be used. Updating a portion of $x$ and a portion of $y_k$ in a single step is an example of the first; marginalizing selected features of $x$ and/or $y$ an example of the second. As evidenced in Section 3.6, we have found that when the $\alpha_k$ differ from one another, mixing is improved by adding a step where $(y_1, \ldots, y_K)$ are permuted.

3.4 Estimation

One of the main benefits of moving from the MSS to the GMSS is that we can draw on established results to estimate summaries of $\pi(x, y)$ that involve $y$. To do so, we cast the problem as an importance sampling problem, where we are sampling from the density $\pi^*(x, s)$ to make inference under the density $\pi(x, y)$. The transformations $g_k(x, s) = g(x, y_k), k = 1, \ldots, K,$ form a one-to-many transformation, as described in MacEachern and Peruggia (2000).
Estimation is accomplished by defining, for a given iterate, the set of weights

\[ w_k = \frac{\alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l)}{\pi^*(s|x)} \]  

(3.2)

The next theorem shows how to use the weights to produce estimates under \( \pi \).

**Theorem 3.** Define \( g^*(x, s) = \sum_{k=1}^{K} w_k g(x, y_k) \). If \( f_k, \ k = 1, \ldots, K, \) is a probability density with support equal to \( \mathcal{Y} \), then \( E_{\pi^*}[g^*] = E_{\pi}[g] \). That is, \( g^*(X, S) \) is an unbiased estimator of \( E_{\pi}[g] \), provided \( (X, S) \sim \pi^* \).

**Proof.**

\[
E_{\pi^*}[g^*(X, S)] = \int_X \int_S \sum_{k=1}^{K} w_k g(x, y_k) \pi^*(s|x) ds \pi(x) dx
\]

\[
= \int_X \int_S \sum_{k=1}^{K} \alpha_k \pi(y_k|x) \prod_{l \neq k} f_l(y_l) g(x, y_k) \pi^*(s|x) ds \pi(x) dx
\]

\[
= \int_X \sum_{k=1}^{K} \alpha_k \left[ \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} \pi(y_k|x) \prod_{l \neq k} f_l(y_l) g(x, y_k) dy_1 \cdots dy_K \right] \pi(x) dx
\]

\[
= \sum_{k=1}^{K} \alpha_k \left[ \int_X \int_{\mathcal{Y}} g(x, y_k) \pi(x, y_k) dy_k dx \right]
\]

\[
= \sum_{k=1}^{K} \alpha_k E_{\pi}[g(X, Y)]
\]

\[
= E_{\pi}[g(X, Y)].
\]

\[\square\]

**Corollary 4.** Define \( g^*(s) = \sum_{k=1}^{K} w_k g(y_k) \). Then \( E_{\pi^*}[g^*(S)] = E_{\pi}[g(Y)] \).

**Corollary 5.** Define \( g^*(x) = \sum_{k=1}^{K} w_k g(x) = g(x) \). Then \( E_{\pi^*}[g^*(X)] = E_{\pi}[g(X)] \).

These results allow us to estimate \( E_{\pi}[g] \) from the the MSS as well as the GMSS. In particular, Corollary 4 remedies shortcomings of the estimator proposed by Leman.
et al. (2009) for functions involving $Y$. Corollary 5 merely extends Leman et al.’s estimator to the GMSS. The formal estimator is

$$E_x[g(X,Y)] = \frac{1}{T} \sum_{t=1}^{T} g^*(x^{(t)}, s^{(t)})$$

$$= \frac{1}{T} \sum_{t=1}^{T} \sum_{k=1}^{K} w_k^{(t)} g(x^{(t)}, y_k^{(t)}),$$

where $t$ indexes the iterate of the sampler. Further adjustment to this estimator is possible, for example by Rao-Blackwellization of the $g(x, y_k)$.

### 3.4.1 Revisit: A Six-Modal Example

To compare the estimator of Leman et al. (2009) to our proposed estimator, the six-modal example in Section 2.4 is revisited. A single MSS/GMSS simulation of $10^5$ iterations was run, leading to the results in Figure 2.2b. Here, we contrast different estimates based on the same MCMC run.

The top panel in Figure 3.1 shows the attained average marginal of $\pi^*(y_k)$. This is computed as $\hat{\pi}^*(y) = \frac{1}{3} \sum_{k=1}^{3} \hat{\pi}^*(y_k)$, where $\hat{\pi}^*(y_k)$ is a histogram of the drawn values of $y_k$. The draws from $\pi^*(y_k)$ fill in the valleys between the modes, as the elements of $s$ explore the space freely. The task of estimation is to use the draws from $\pi^*(y_k)$ to estimate $\pi(y)$ based on the same MCMC run as in the top panel.

The lower two panels show estimates with Leman et al.’s method and our proposed method, based on the same MCMC run as in the top panel. Leman et al.’s density estimate is shown in the middle panel along with the desired target density $\pi(y)$ as a solid line. Following Leman et al. (2009), we first estimate the marginal density of
Figure 3.1: Solid curves are desired marginals of $\pi(y)$. Top panel: The attained average marginal of $\pi^*(y_k)$ using the MSS/GMSS with the multiset size $K = 3$. Middle panel: Estimate of $\pi(y)$ using Leman et al.’s method. Bottom panel: Estimate of $\pi(y)$ using the proposed method.
Y by
\[ \hat{\pi}_{\text{Leman}}(y) = \frac{1}{3} \sum_{k=1}^{3} \hat{\pi}_{\text{Leman}}(y_k) = \frac{1}{3} \sum_{k=1}^{3} 3 \times \left( \hat{\pi}^*(y_k) - \frac{2}{3 \cdot 20} \right). \]

Note that we re-describe the density estimators \( \hat{\pi}_{\text{Leman}}(y) \) in our terminology since we use the notion of a \( K \)-tuple instead of that of a multiset used by Leman et al. (2009). In their original work, the density estimator is \( \hat{\pi}(y) = 3 \times \left( \hat{\pi}^*(y) - \frac{2}{3 \cdot 20} \right) \). However, it is not clear what the estimator means when \( y \) is a continuous variable.

The bottom panel shows the density estimate with our proposed method. To estimate the marginal distribution of \( y \) by our proposed method, we set \( g(X,Y_k) = I[Y_k \leq R_m] \), where \( R_m \) is the right endpoint of the \( m \)-th histogram bin. Appealing to Theorem 3, we construct \( g^*(x,s) = \sum_{k=1}^{3} g(x,y_k) / 3 \), and have, under ergodicity, an unbiased estimator of \( \pi(y) = P_\pi(Y \leq y) = E_\pi(I[Y \leq y]) \). This estimate is then translated into an estimate of the marginal density. The new estimate compares favorably to Leman et al.’s, as it never estimates a probability as negative and is closer to \( \pi(y) \) under a variety of formal measures.

Figure 3.2 shows the relationship between \( \pi(y) \) and \( \pi^*(y_k) \) with different \( \alpha_k \). The setting of the GMSS is the same as above, except the multiset has \( K = 2 \) components whose weights are \( (\alpha_1, \alpha_2) = (0.9, 0.1) \). We draw the marginals of \( y_k \) from the GMSS run. As illustrated by the histograms, the distribution \( \pi^*(y_1) \) of the large weighted element is similar to the target distribution \( \pi \), while \( \pi^*(y_2) \) with the small weight is similar to the uniform instrumental distribution \( f, Unif[-10,10] \). Since the improvements in mixing due to the MSS and the GMSS come from flattening the target distribution, Figure 3.2 shows the possibility of further improvement by using different values of \( \alpha \). Improvement is demonstrated in Section 3.6 for a simpler
example. While the heavily weighted element $y_1$ more often sticks around a mode of target distribution, $y_2$ travels more freely.

By using the GMSS, we no longer need to assume the parameter in a bounded region, as we did in this example. Instead of using the uniform instrumental density implicitly used in the MSS, we can use any continuous instrumental density with adequate tails, for example, the $N(5,4)$ for the six modal density. The assumption of a bounded parameter space makes this six modal example look artificial and limits the application of the MSS in practice.
Figure 3.2: Top panel: The attained average marginal of $\pi^*(y_k)$ using the GMSS with the multiset size $K = 2$. Middle panel: Samples from the distribution for the first element of the multiset, $\pi^*(y_1)$, with $\alpha_1 = 0.9$. Bottom panel: Samples from the distribution for the second element, $\pi^*(y_2)$, with $\alpha_2 = 0.1$
3.5 Simulation Studies

3.5.1 Example 2: Mixture of Bivariate Normal Distributions

This example, previously used by Liang and Wong (2001), illustrates a 20 component location mixture of bivariate normal distributions:

\[
\pi(y) = \frac{1}{2\pi\sigma^2} \sum_{\eta=1}^{20} p_\eta \exp \left[ -\frac{1}{2\sigma^2} (y - \mu_\eta)^T (y - \mu_\eta) \right]
\] (3.3)

where \(\sigma = 0.1\) and \(p_\eta = 1/20\). The simulated mean vectors \(\mu_\eta = (\mu_{\eta,1}, \mu_{\eta,2})^T\) are drawn in Figure 3.3 as twenty empty circles. The exact values of the mean vectors are contained in Liang and Wong (2001). Using (3.3) as a target distribution, we draw \(10^5\) samples of \(y^{(t)}\) under ordinary Metropolis-Hastings (MH) and the GMSS. If an algorithm is mixing well, its estimated distribution of \(y\) will spread equally around the twenty means.

The sampling distribution of the random-walk MH is the same as \(\pi(y)\) in (3.3) and that of the GMSS is set as

\[
\pi^*(y_1, y_2) = \alpha_1 \pi(y_1) f_2(y_2) + \alpha_2 \pi(y_2) f_1(y_1)
\]

where the instrumental densities \(f_k(y_k)\) follow \(Unif[0, 10]^2\). We set the multisets to be equally weighted with \(\alpha_1 = \alpha_2 = 1/2\) and use the Type I update. The bivariate normal proposal density \(y' \sim N(y^{(t-1)}, \tau^2 I)\), is used to run both the ordinary MH algorithm and the GMSS algorithm.

In Figure 3.3, we plot all samples of the ordinary MH and the samples of the GMSS whose weight \(w_k\), as in Equation (3.2), is larger than 0.5. The element in a local mode has a very large weight, most of the time close to the value one, while the other free-moving element has a very small weight, most of times close to zero. From
the graph, we see that the ordinary MH with a small scale, $\tau = 0.5$, samples only four elements in the lower-left corner. The ordinary MH with $\tau = 1.0$ visits all components, but it shows poor mixing, not visiting the lower left corner often enough and, as shown by the sparsity of distinct points in the blobs, accepting proposed values only rarely. In contrast, the GMSS samples all 20 components with both the small and the large scale for proposals. The components are visited approximately equally often. The greater spread of the blobs in the lower left corner panel (the GMSS with $\tau = 0.5$) indicates better local exploration with the small scale for proposals. This result shows that the GMSS is more robust to the scale of the proposal distribution than is ordinary MH.

### 3.5.2 Example 3: Mixture Univariate Normal Example

We construct a mixture normal example with four components:

$$
\pi(\mu, \sigma^2, p; z_j) = \sum_{\eta=1}^{4} p_\eta N(z_j; \mu_\eta, \sigma^2), \ j = 1, \ldots, n.
$$

We generate $n = 100$ observations under the parameter values $\mu = (-4, -3, 3, 4)$, $\sigma = 0.4$ and $p_\eta = 1/4$, $\eta = 1, 2, 3, 4$. For this example, the ordinary Gibbs sampler works fairly well if we know the true number of the mixture components. However, it is often difficult to determine the number of mixture components. In such cases, we may fit a model with too few or too many components.

Here, we consider fitting a mixture model with three components to a simulated data set generated from a four component mixture model. For simplicity, we assume that $\sigma$ is known to be 0.4 and $p = (p_1, p_2, p_3)^T$ is fixed as $(1/3, 1/3, 1/3)^T$, so that we only need to estimate $\mu = (\mu_1, \mu_2, \mu_3)^T$. We use a Bayesian mixture method with a
Figure 3.3: The twenty simulated mean vectors are drawn as empty circles. Top left: Ordinary MH with proposal scale $\tau = 0.5$, Top right: Ordinary MH with proposal scale $\tau = 1.0$, Bottom left: GMSS with proposal scale $\tau = 0.5$, Bottom right: GMSS with proposal scale $\tau = 1.0$. 
prior distribution, \( \mu_\eta \sim N(\bar{\mu}, A) \) where \( \bar{\mu} = 0 \) and \( A = 0.01 \). For the Gibbs sampler, we adopt the latent allocation variable as in Richardson and Green (1997) so that

\[
z_j|\lambda_j \sim N(\mu_{\lambda_j}, \sigma^2) \quad \text{and} \quad \lambda_j \sim \text{mult}(p).
\]

For the GMSS, we use the multiset of two vector elements, \( S = (\mu_1, \mu_2) \), where \( \mu_k = (\mu_{k,1}, \mu_{k,2}, \mu_{k,3})^T \) with the uniform instrumental densities \( f_k(\mu_k) = \prod_{\eta=1}^{3} I[\mu_{k,\eta} \in (-6, 6)] \) and the equal weights \( \alpha_k = 1/2 \) for \( k = 1, 2 \) under the Type I update. The random walk proposal used for the GMSS is \( \mu_k' \sim N(\mu_k^{(t-1)}, \tau^2 I) \) where \( \tau^2 = 0.01 \).

In Figure 3.4, the draws of the ordinary Gibbs sampler and GMSS run for \( 10^6 \) iterations (keeping every 20th iterate) are compared. The values plotted for the GMSS are \( \mu_{\eta}^{*(t)} = w_1\mu_{1,\eta}^{(t)} + w_2\mu_{2,\eta}^{(t)} \), \( \eta = 1, 2, 3 \). The Gibbs sampler finds two modes on the lower side and one mode on the top. The top mode lies in the middle of the two upper means. With a different starting point, the Gibbs sampler can find the two upper modes and a single lower mode. However, the Gibbs sampler does not move between these two explanations for the data. In contrast, the GMSS moves between the pairs of lower means and upper means.

The better movement of the GMSS is explained by the free travel of an element of the multiset (or \( K \)-tuple). Figure 3.5 shows how the elements of the multiset work. First, \( \mu_2 \) finds the two lower means while \( \mu_1 \) moves freely to find other regions of high density. Around the 20,000-th iteration, the free moving \( \mu_1 \) approaches the high density region that \( \mu_2 \) has already found, and \( \mu_2 \) starts its trip to other regions. Finally, \( \mu_2 \) finds the other high density region of the two upper means around the 30,000-th iteration, and this procedure is repeated with \( \mu_1 \) and \( \mu_2 \) changing their roles.
Figure 3.4: Two algorithms for mixture normal estimation. The Gibbs sampler is used in the top panel; the GMSS is used in the bottom panel. The GMSS uncovers all four means and moves between different three-mean explanations of the data.
Figure 3.5: The trajectory of the multiset elements, $\mu_1$ and $\mu_2$, from the GMSS algorithm.
3.6 Weights, Instrumental Densities, and Permutation Steps

We modify the illustrative example of Leman et al. (2009) to show the impact of instrumental densities, weights and a switching step. The assumed target distribution of \((x, y)\) and the various forms of instrumental densities \(f_k\) used in this example are as follows:

\[
\begin{array}{c|ccc}
\pi(x, y) & y & 1 & 2 & 3 \\
\hline
x & \pi(x, y) & 0.0998 & 0.0001 & 0.0001 \\
1 & f_{unif}(y) & 0.333 & 0.333 & 0.333 \\
2 & f_{good}(y) & 0.100 & 0.001 & 0.899 \\
& f_{bad}(y) & 0.899 & 0.001 & 0.100 \\
\end{array}
\]

As Leman et al. (2009) illustrated, the ordinary Metropolis-Hastings algorithm cannot easily move between the global mode \((x = 2, y = 3)\) and the local mode \((x = 1, y = 1)\) since there are big valleys between those two modes.

To solve this issue, we adopt the GMSS with \(K = 2\) and a Type II update. We use three different types of instrumental densities: a uniform instrumental density, a good instrumental density and a bad instrumental density. The good instrumental density is similar to the marginal (target) distribution \(\pi(y)\), and the bad instrumental density flips the good instrumental density. We run a designed experiment with \(4 \times 101 \times 2\) settings. The four combinations of \((f_1(y_1), f_2(y_2))\) are used: (1) two uniform instrumental densities (UU), (2) a good instrumental density and a uniform instrumental density (GU), (3) two good instrumental densities (GG), and (4) a bad instrumental density and a uniform instrumental density (BU). One hundred one equally spaced values of \(\alpha_1\) from 0 to 1 are used. The third factor in the experiment is whether the switching step is included as a final step in each iteration.
As a measure of the chains’ mixing, we use the integrated autocorrelation time (IACT) proposed by Sokal (1989). To implement his “automatic windowing” algorithm, we use the Fourier transform algorithm which is used by Green and Han (1992) to estimate the autocorrelation time. See Appendix A for further details.

Figure 3.6 shows the averages of the IACTs of \( x \) generated by 100 replicates of \( 10^5 \) iterate runs under the GMSS with various settings. The figure shows the impacts of various instrumental densities, different weights and the permutation step.

**Instrumental densities \( f_k \)**

In the GMSS, we no longer need to assume the parameter space is bounded since any form of the instrumental density \( f_k \) on the support \( \mathcal{Y} \) makes \( \pi^* \) integrable and so enables inference. In addition, the instrumental density lets us guide the sampler toward regions of the \( (x,s) \) space suspected to contain substantial probability. As seen in Figure 3.6, the mixing is improved by introducing a good instrumental density \( f_{\text{good}} \) that is combined with a uniform instrumental density \( f_{\text{unif}} \). If we use the GMSS with the pair of \((f_{\text{good}}, f_{\text{unif}})\), an element of the multiset that is connected to \( f_{\text{good}} \) is more likely to visit the state \( y = 3 \), while the other element, connected to \( f_{\text{unif}} \), moves more freely across the space. The different roles of the two elements help the chain mix. As we can see from Figure 3.6, the IACT of GU chain is smaller than those of UU and BU.

Another interesting feature is that the use of two good instrumental densities is not as good as use of GU and UU. If we use two of the same instrumental densities leaning toward the global mode, it is more difficult to accept the movement of any element in the multiset toward the local mode. Therefore, the use of two identical
Figure 3.6: The mixing of GMSS with $K = 2$ through the different $f_k$, $\alpha_k$, and the switching step. UU refers to use of the instrumental densities $(f_{\text{unif}}, f_{\text{unif}})$, GU of $(f_{\text{good}}, f_{\text{unif}})$, GG of $(f_{\text{good}}, f_{\text{good}})$, and BU of $(f_{\text{bad}}, f_{\text{unif}})$. Solid lines are the algorithms with permutation steps and dotted lines are without permutation steps. The solid circle refers to the MSS as it has uniform $f_k$ and equal $\alpha_k$. 
asymmetric densities may lose the primary benefit of the GMSS. The recommended choice of the instrumental densities, from empirical study, is to mix \(f_{\text{unif}}\) and a good instrumental density that is roughly similar to the marginal (target) distribution of \(Y\). If we have little or no information about the marginal target distribution, it might be safe to use the uniform instrumental densities as the comparison between UU and BU in the figure shows.

**Multiset weights \(\alpha_k\)**

The GMSS framework encompasses both the traditional MH algorithm and the MSS as extremes by adjusting the weight \(\alpha_k\). In addition, the weights affect the mixing of the chain as seen in Figure 3.6. When different instrumental densities \(f_k\) are used, the optimal \(\alpha_k\) are not simply \(1/K\), here 0.5. Specifically, if an instrumental density \(f_1\) roughly captures the shape of the target density, i.e., \(f_1 = f_{\text{good}}\), and \(f_2\) is the uniform instrumental density, the optimal weight \(\alpha_1\) is smaller than \(1/K = 0.5\). In contrary, if \(f_1 = f_{\text{bad}}\) and \(f_2 = f_{\text{unif}}\), the optimal weight \(\alpha_1\) is close to the value one. From the result, we see that the mixing of the chain is highly affected by the combination of the instrumental density and the weight: giving higher value to \(\alpha_2\) if \(f_1(y_1)\) is a better instrumental density than \(f_2(y_2)\) improves performance. However, tuning \(\alpha_k\) is difficult as it requires considerable knowledge of \(\pi\) and evaluation of \(f_k\).

**Permutation step**

Consider the case where \(\alpha_1 > \alpha_2 > 0\) and \(f\) is uniform. With this set-up, the role of \(y_2\), the version of \(y\) that receives less weight in \(\pi^*\), is to explore the space, searching for an undiscovered mode. While this is occurring, \(y_1\) will tend to remain in a region of
relatively high probability, so that $\pi^*(x, s)$ is fairly large. Suppose that a new mode is discovered, and that $\pi(x, y_2) > \pi(x, y_1)$. Ideally, the roles of $y_1$ and $y_2$ would switch, with $y_2$ defining the region of high probability and $y_1$ exploring the space. The roles of $y_1$ and $y_2$ can be switched merely by switching their labels. This suggests inclusion of a step in the algorithm where a permutation of the labels is proposed and either accepted or not.

In the toy example above, with a multiset of size $K = 2$, the permutation step merely determines whether $y_1$ and $y_2$ swap each value. Specifically, the permutation step used in the example is as follows:

1. Propose $s' = (y'_1, y'_2)$ where $y'_1 = y_2^{(t)}$ and $y'_2 = y_1^{(t)}$.

2. Accept or reject $s'$ with the acceptance probability

$$
\gamma = \min\left[1, \frac{\alpha_1 \pi(y'_1) f_2(y'_2) + \alpha_2 \pi(y'_2) f_1(y'_1)}{\alpha_1 \pi(y_1^{(t)}) f_2(y_2^{(t)}) + \alpha_2 \pi(y_2^{(t)}) f_1(y_1^{(t)})}\right].
$$

With the GMSS with $K = 2$, the permutation step does not affect the result when we use the same multiset weights $\alpha_1 = \alpha_2$ and the same instrumental densities $f_1 = f_2$. Therefore, GG and UU have the points at the center where the results from the algorithm with the permutation step and without the permutation step are the same. However, except in this specific case, it is shown that the permutation step consistently improves chain’s mixing under a variety of settings.

When the GMSS has multiset size larger than $K = 2$, there are several natural ways to propose a permutation. Among them are a uniform proposal over permutations, a Gibbs sampling step over permutations (for small $K$), a uniform proposal across permutations other than the current one, and a proposal from the conditional distribution across permutations other than the current one. Theoretical results tied
to Peskun’s Theorem (e.g., Peskun 1973; Liu 1996) suggest exclusion of the current state from the proposal set.

### 3.7 Convergence of the Algorithm

Convergence of the algorithm on the $\pi^*$ space, provided the initialization and proposal distributions are adequate, follows from standard results. The strategies used to produce an ergodic Markov chain for $\pi$ can also be used to produce an ergodic chain for $\pi^*$. Here, we follow Robert and Casella (2004) and Roberts and Rosenthal (2006). To streamline the presentation, we will assume that the entire parameter vector has been multisetted.

For simplicity, we drop $x$ and consider the multiset sampling density $\pi^*(s)$. Then, the Type II algorithm is, formally, a full-dimensional Metropolis-Hastings algorithms, the Type I algorithm is a random scan MH algorithm, and the Type III algorithm is a deterministic scan MH algorithm (refer to Roberts and Rosenthal 2006).

#### 3.7.1 Type II Update

The Type II algorithm is driven by a joint proposal of the vector $s = (y_1, \ldots, y_K)$ and it is a full-dimensional Metropolis-Hastings algorithm in the terminology of Roberts and Rosenthal (2006). We presume that $\bigcup_{y \in \mathcal{Y}} \text{supp } q(\cdot | y) \supset \mathcal{Y}$ and $\int \pi(y)dy < \infty$ where $dy$ is Lebesgue measure.
Proposition 6. A Markov chain \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update is a full-dimensional Metropolis-Hastings algorithm whose stationary distribution is \( \pi^* \).

Proof. \( \int \pi^*(s) \lambda(ds) = \int \cdots \int \sum_{k=1}^{K} \alpha_k \pi(y_k) \prod_{l \neq k} f_l(y_l) dy_1 \cdots dy_K = \sum_{k=1}^{K} \alpha_k \int \pi(y_k) dy_k = \int \pi(y) dy. \) Therefore, \( \int \pi^*(s) \lambda(ds) < \infty. \) The acceptance ratio for the proposed \( s' \) of the GMSS algorithm with Type II update is \( \gamma = \min\{1, \delta\} \) where

\[
\delta = \frac{\pi^*(s')}{\pi^*(s^{(t-1)})} \frac{q(s'^{(t-1)}|s')}{q(s^{(t-1)}|s^{(t-1)})},
\]

which is exactly the same as the acceptance ratio used in Roberts and Rosenthal (2006). \( \square \)

The proposal densities (under \( \pi \)) of the Type II algorithm are “rectangular sets” that contain the support of \( y_1 \). We presume that the proposal density under the (original) target density \( \pi \) is positive. That is,

\[
q(y'|y^{(t-1)}) > 0 \quad \text{for every} \quad (y^{(t-1)}, y') \in \mathcal{Y} \times \mathcal{Y}. \tag{3.4}
\]

Proposition 7. A Markov chain \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update is \( \phi \)-irreducible (with respect to \( \pi^* \)).

Proof. Since the support of each individual coordinate \( y_k \) under \( \pi^* \) is contained in the support of \( y \) under \( \pi \), \( q(s'|s^{(t-1)}) = \prod_k q(y_k'|y_k^{(t-1)}) > 0. \) Since the proposal density is positive, which satisfies sufficient condition (7.5) of Robert and Casella (2004), the chain \( \{S^{(t)}\} \) is \( \phi \)-irreducible. \( \square \)

Proposition 8. A Markov chain \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update is Harris recurrent.
Proof. This follows from Theorem 8 of Roberts and Rosenthal (2006) since \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update is a full-dimensional Metropolis-Hastings algorithm by Proposition 6 and is \( \phi \)-irreducible by Proposition 7.

We then have the following result for the \( \pi^* \)-chain.

**Theorem 9.** Consider a Markov chain \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update. If \( g^* \in L^1(\pi^*) \), then

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} g^*(S^{(t)}) = \int g^*(s) \pi^*(s) \, ds \quad \text{a.e. } \pi^*.
\]

Thus, the estimator converges to \( E_{\pi^*}[g^*(S)] \).

**Proof.** \( \{S^{(t)}\} \) is Harris recurrent by Proposition 8. Let \( h(s) = 1 \). By Theorem 6.63 of Robert and Casella (2004),

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} g^*(S_t) = \lim_{T \to \infty} \frac{\sum_{t=1}^{T} g^*(S_t)}{\sum_{t=1}^{T} h(S_t)} = \int g^*(s) d\pi^*(s),
\]

provided \( g^* \in L^1(\pi^*) \).

**Corollary 10.** Consider a Markov chain \( \{S^{(t)}\} \) from the GMSS algorithm with Type II update whose instrumental density \( f_k \) has support equal to \( \mathcal{Y} \). If \( g \in L^1(\pi) \), then the chain \( \{S^{(t)}\} \) yields an estimator which converges to \( E_{\pi}[g(Y)] \).

**Proof.** If \( g \in L^1(\pi) \), then \( g^* \in L^1(\pi^*) \), so the result of Theorem 9 holds. From Theorem 3, \( E_{\pi^*}[g^*(S)] = E_{\pi}[g(Y)] \), so

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} g^*(S^{(t)}) = E_{\pi}[g(Y)].
\]

**3.7.2 Type I Update**

We refer to Roberts and Rosenthal (2006) to prove the ergodicity of the Type I updating chain since it is, formally, a random-scan Metropolis-within-Gibbs algorithm. Let \( \mathcal{S} \) be an open subset of \( \mathcal{R}^K \). We presume that \( \bigcup_{y \in \mathcal{Y}} \text{supp } q(\cdot|y) \supset \mathcal{Y} \) and
\[ \int \pi(y)dy < \infty \] where \( dy \) is one dimensional Lebesgue measure. Then, \( \pi^* \) be a density function with \( \int_S \pi^*(s)\lambda(ds) < \infty \) where \( \lambda \) is \( K \)-dimensional Lebesgue measure where \( s = (y_1, \ldots, y_k) \). Let \( q_k : S \times \mathcal{R} \to [0, \infty) \), \( k = 1, \ldots, K \), be jointly measurable with \( \int_{-\infty}^{\infty} q_k(s, y_k)dy_k = 1 \) for all \( s \in S \). Let \( Q_k(s, \cdot) \) be the Markov kernel on \( \mathcal{R}^K \) defined as
\[
Q_k(s, A'_{k,a,b}) = \int_a^b q_k(s, y'_k)dy'_k
\]
where
\[
A'_{k,a,b} = \{ s' \in S : y'_l = y_l \text{ for } l \neq k \text{ and } a \leq y'_k \leq b \}.
\]
Assume that
\[
q_k((y_1, \ldots, y_{k-1}, y_k, y_{k+1}, \ldots, y_K), y'_k) > 0 \iff q_k((y_1, \ldots, y_{k-1}, y'_k, y_{k+1}, \ldots, y_K), y_k) > 0.
\]
Note that some proofs are very similar to those in the previous section, so we have omitted the duplicated parts.

**Proposition 11.** A Markov chain \( \{S(t)\} \) from the GMSS algorithm with Type I update is a random-scan Metropolis-within-Gibbs chain whose stationary distribution is \( \pi^* \).

**Proof.** The acceptance ratio for the proposed \( y'_k \) of the GMSS algorithm with Type I update is \( \gamma = \text{min}(1, \delta) \) where
\[
\delta = \frac{\pi^*(y_1, \ldots, y_{k-1}, y'_k, y_{k+1}, \ldots, y_K)q(y_k|y'_k)}{\pi^*(y_1, \ldots, y_{k-1}, y_k, y_{k+1}, \ldots, y_K)q(y'_k|y_k)} = \frac{\pi^*(s')q(y_k|y'_k)}{\pi^*(s)q(y'_k|y_k)}.
\]
By denoting the proposal density \( q(y'_k|y_k) \) as \( q_k(s, y'_k) \) for \( k = 1, \ldots, K \), this is equivalent to the acceptance ratio of Metropolis-within-Gibbs chain as in Roberts and Rosenthal (2006).

\[ \square \]
We again presume that the proposal distribution under the (original) target density $\pi$ is positive. That is, 

$$q(y'|y^{(t-1)}) > 0 \text{ for every } (y^{(t-1)}, y') \in \mathcal{Y} \times \mathcal{Y}.$$ 

**Proposition 12.** A Markov chain $\{S^{(t)}\}$ from the GMSS algorithm with Type I update is $\phi$-irreducible.

**Proof.** Since the support of each individual coordinate $y_k$ under $\pi^*$ is contained in the support of $y$ under $\pi$, $q(y'_k|y^{(t-1)}_k) > 0$ for every $(y^{(t-1)}_k, y'_k) \in \mathcal{Y} \times \mathcal{Y}$. Then, every set of $S$ with $\pi^*(s) > 0$ can be reached in a single sweep, which consists of $M(\geq K)$ iterations which update each element of a multiset at least once, with positive probability from multiplication of $q_k$. That is, the chain $\{S^{(t)}\}$ is $\phi$-irreducible with respect to $\pi^*$.

**Proposition 13.** A Markov chain $\{S^{(t)}\}$ from the GMSS algorithm with Type I update is Harris recurrent.

**Proof.** The chain is a random-scan Metropolis-within-Gibbs chain from Proposition 11. Without loss of generality, $\mathcal{Y}^r$ be $r$-dimensional coordinate hyperplane of $S$ for the first $r$ elements in multiset, $(y_1, \ldots, y_r)$. When $r = K$, $\int_S \pi^*(s)ds = 1 < \infty$ from the proof of Proposition 1. When $r < K$, $\int_{\mathcal{Y}^r} \pi^*(s)ds = \sum_{k=1}^s w_k \prod_{i=s+1}^K f_i(y_i) + \sum_{k=s+1}^K w_k \pi(y_k) \prod_{i \geq s+1 \& i \neq k} f_i(y_i) < \infty$. Proposition 12 shows that the chain is $\phi$-irreducible. Then, the result follows from Corollary 19 of Roberts and Rosenthal (2006).

**Theorem 14.** Consider a Markov chain $\{S^{(t)}\}$ from the GMSS algorithm with Type I update. If $g^* \in L^1(\pi^*)$, then 

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T g^*(S^{(t)}) = \int g^*(s) \pi^*(s) ds \quad \text{a.e. } \pi^*.$$
So, the estimation converges to $E_{\pi^*}[g^*(S)]$

**Corollary 15.** Consider a Markov chain $\{S^{(t)}\}$ from the GMSS algorithm with Type I update whose instrumental density $f_k$ has support equal to $\mathcal{Y}$. If $g \in L^1(\pi)$, then the chain $\{S^{(t)}\}$ yields an estimator which converges to $E_{\pi}[g(Y)]$.

### 3.7.3 Type III Update and Other Variants

Dropping $x$, the Type III update is a deterministic scan MH algorithm. Stochastic properties, including convergence (so, its proof), of the deterministic scan MH chain are very similar to those of the random-scan MH chain in the previous subsection.

**Corollary 16.** A Markov chain $\{S^{(t)}\}$ from the GMSS algorithm with Type III update is a deterministic-scan Metropolis-within-Gibbs chain whose stationary distribution is $\pi^*$. So, it is Harris recurrent.

*Proof.* The proof is a straightforward application of Corollary 19 of Roberts and Rosenthal (2006).

In fact, algorithms that do not yield an ergodic Markov chain on the $\pi^*$ space may still yield sufficient convergence and/or mixing for inferential purposes, following the techniques of Section 3.4. The simplest such case relies on a set of non-overlapping instrumental densities which partition the support $\mathcal{Y}$. Such an algorithm bears a close relationship to running an algorithm on one piece of the support and using the importance link function techniques of MacEachern and Peruggia (2000) for estimation.
3.8 Discussion

So far, we have shown that the GMSS has all of the benefits of MSS and adds more advantages over MSS: (i) reliable estimation of the marginal distribution $\pi(Y)$ enables us to deal with situations where $Y$ is also of interest, and (ii) it relaxes the strong assumption that $Y$ is bounded. Along with the benefits, the GMSS also enjoys more advantages by introducing the multiset weight $\alpha_k$ and the instrumental density $f_k$. Especially, (iii) appropriate choices of the various parameters of the GMSS improve the mixing of the chain.

Consideration of the weights and the instrumental densities allows us to place the GMSS into the context of existing algorithms. This also gives insight into the MSS and suggests improvements to the basic sampling algorithm. For specificity, we take $y$ to be a scalar and take $Y$ to be the interval $[a,b]$. As previously noted, setting $\alpha_k = 1/K$ for $k = 1, \ldots, k$ and taking $f_k(y) \sim Unif[a,b]$, for $k = 1, \ldots, K$, yields the MSS with cardinality $k$.

At another extreme, taking $\alpha_1 = 1$ and any $f_k$ with support $Y$ yields the ordinary Metropolis-Hastings algorithm. To see how this happens, we focus on the joint marginal distribution of $(x, y_1)$. The two densities for $x$ are $\pi^*(x) = \pi(x)$. Following the proof of Proposition 2, the conditional density for $y_1$ given $x$ is $\pi^*(y_1|x) = \alpha_1 \pi(y_1|x) + (1-\alpha_1)f_1(y_1) = \pi(y_1|x)$. Multiplying by $\pi(x)$, we have $\pi^*(x, y_1) = \pi(x, y_1)$. Of course, the GMSS engenders a richer distribution on the expanded set of variables $(x, y_1, \ldots, y_K)$. The joint density of this collection of variates is given by $\pi^*(x, y_1, \ldots, y_K) = \pi(x, y_1) \prod_{k=2}^K f_k(y_k)$. This result follows from calculations similar to those in the proof of Proposition 2. For inferential purposes, the variates
\((y_2, \ldots, y_K)\) can be ignored; for computational purposes, one would remove them for a faster implementation of the algorithm.
Chapter 4: Application to Gene Expression Data

4.1 Introduction

In this chapter, the GMSS is applied to a now-famous microarray gene study, the breast cancer complementary DNA (cDNA) microarray experiment of Hedenfalk et al. (2001). Like other microarray experiments, the study is challenging since it deals with simultaneous inference on thousands of genes. We introduce various types of methods to deal with the simultaneous testing problem, specific to this gene study, and show the applicability of the GMSS in this study.

The chapter is organized as follows. Section 4.2 introduces Hedenfalk’s experiment. In Section 4.3, we present the basic concepts of multiple hypotheses testing including the false discovery rate (FDR). This is followed by the introduction of Bayesian approaches to microarray studies, which usually connect to the concept of the local false discovery rate, in Section 4.4. Then, Section 4.5 illustrates the application of the GMSS to the breast cancer microarray study.
4.2 Hedenfalk’s Study: Breast Cancer Microarray Data

BRCA1 and BRCA2 gene mutations are known to increase the risk of breast cancer, as a cell with the mutant genes has decreased ability to repair damaged DNA. Hedenfalk et al. (2001) aimed to find the genes that show different genetic activities between the two type of tumors, BRCA1 mutation related cancer and BRCA2 mutation related cancer. To compare the different genetic activities of each gene, they measured the quantified relative activity of a gene’s messenger RNA (mRNA), in more formal terms the gene expression level. If a gene shows more or less genetic activity in the BRCA1 related cancer compared to the BRCA2 related cancer, the gene has the different levels of expression between the two cancer groups. We call a gene having different expression levels a differentially expressed gene.

The study examined three groups of patients: seven of the women with BRCA1-related cancer, eight with BRCA2-related cancer, and seven with sporadic cases having neither mutation. In this study, the expression level of each woman’s tumor cell is obtained as follows. The DNA base sequences of the 3,226 genes were printed on the microarray plates. When the tumor cells were hybridized with the plate, messenger RNA (mRNA) is produced and we can optically read the gene expression levels with red dye for the tumor cell’s activity and green dye for the background measurement. If a gene is more active in BRCA1-related cancer compared to a base slide, the logarithm of the ratio of red to green intensities has a high value.

As a result, we have a $3,226 \times 7$ data matrix for BRCA1-related cancer, a $3,226 \times 8$ data matrix for BRCA2-related cancer, and a $3,226 \times 7$ data matrix for sporadic cancers. To analyze the gene expression measurements, Hedenfalk et al. (2001) performed an ANOVA test and identified 51 genes as differentially expressed, using a
p-value cutoff of 0.001.

4.3 Multiple-hypothesis Testing and False Discovery Rate

As gene expression data has a collection of null hypotheses and a large number of data points, it raises simultaneous inference issues and needs data reduction techniques. Suppose that we are simultaneously testing $m$ null hypotheses, $H_1, H_2, \ldots, H_m$, with corresponding $p$-values $P_i$, for $i = 1, \ldots, m$.

<table>
<thead>
<tr>
<th># of true null</th>
<th># of true alternative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># of declared negative (not differentially expressed)</td>
<td># of declared positive (differentially expressed)</td>
<td></td>
</tr>
<tr>
<td>$m_0 - F$</td>
<td>$F$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>$m_1 - T$</td>
<td>$T$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$m - S$</td>
<td>$S$</td>
</tr>
</tbody>
</table>

The table above defines some random variables related to the multiple hypothesis tests where $m_0$ is the number of true null hypotheses, $S$ is the number of hypotheses called positive (in our context, the number of genes declared as differentially expressed), $F$ is the number of false positive hypotheses (Type I errors), and $m_1 - T$ is the number of false negative hypotheses (Type II errors). If we call all hypotheses positive whose $p$-value, $P_i$, is less than or equal to some threshold $\alpha$, then the number of false positive is expressed as a random variable, $F(\alpha) = \sum_{i=1}^{m} I[p_i \leq \alpha \text{ and } H_i \text{ is true}]$. 

55
**Per comparison error rate and familywise error rate**

Ignoring any multiple hypothesis testing correction, one uses a per comparison error rate (PCER), the expected rate of false positive hypotheses, i.e., \( E(F/m) \). We can guarantee that \( PCER = E(F/m) \leq \alpha \) if we test each individual hypothesis at the level \( \alpha \). If the conventional p-value threshold of \( \alpha = 0.05 \) is used to control the PCER, the expected number of false positive, \( E[F] \), is less than or equal to \( 0.05m \). This number is too large when we have a large number of hypotheses to be tested simultaneously.

One of the classical multiple testing approaches is to use the familywise error rate (FWER), the probability of making at least one false negative among all hypotheses under consideration, i.e., \( P(F \geq 1) \). Some researchers thought the FWER is useful to control the multiplicity selection effect. However, there are many difficulties which prevent people to use it in practice. We refer the reader to Benjamini and Hochberg (1995) for more details of the difficulties.

**False discovery rate**

Benjamini and Hochberg (1995) suggested a new approach to consider the proportion of false positives. Let \( Q \) be the proportion of the false positive hypotheses among all the hypotheses declared positive, i.e., \( Q = F \) if \( S > 0 \) and 0 otherwise. It is known that controlling \( Q \) itself is difficult or impossible. However, we can instead control the *false discovery rate* (FDR), \( Q_e = E(Q) = P(S > 0)E(F/S|S > 0) \).

Benjamini and Hochberg (1995) proved that a simple sequential Bonferroni-type procedure can control the FDR for independent test statistics. Suppose one wants to simultaneously test \( m \) null hypotheses, \( H_1, \ldots, H_m \), with corresponding independent test statistics, \( y_1, \ldots, y_m \), and p-values, \( P_1, \ldots, P_m \). Let \( P_{(1)} \leq \ldots \leq P_{(m)} \) be the
ordered $p$-values, and let $H_{(i)}$ denote the null hypothesis corresponding to $P_{(i)}$, $P_{(1)}$ being the most significant and $P_{(m)}$ the least significant in the usual terminology. Define the Bonferroni-type multiple-testing procedure, for a specified value $q^*$:

Let $M$ be the largest $i$ for which $P_{(i)} \leq (i/m)q^*$,

then reject all $H_{(i)}$, $\forall i \leq M$.

They proved that this procedure controls the FDR at level $q^*$, for independent test statistics.

4.4 Bayesian Approaches to Microarray Experiment

4.4.1 Local False Discovery Rate

Efron, Tibshirani, Storey and Tusher (2001) proposed a Bayesian version of FDR using a nonparametric empirical Bayes model. They aimed to find the genes which are affected by a treatment of interest. We replace the term ‘affected’ with ‘differentially expressed’ in our context of the BRCA-related cancer study. Let us assume that there are two classes of genes, either differentially or not differentially expressed. Let $p_1$ be the prior probability that a gene is differentially expressed between two groups, and $p_0 = 1 - p_1$ be the prior probability that the gene is not differentially expressed. We have test statistics $Y_i$ where $h_1(y_i)$ denotes the density of $Y_i$ for a differentially expressed gene and $h_0(y_i)$ for a non-differentially expressed gene.

The unconditional density of $Y_i$ is given by the mixture density of the two populations,

$$h(y) = p_0h_0(y) + p_1h_1(y)$$ (4.1)
By using Bayes’ rule, the posterior probabilities that a gene is not differentially expressed given the statistic $y_i$ is

$$\tau_0(y_i) = Pr[\text{non-differentially expressed}|Y_i = y_i] = p_0 \frac{h_0(y_i)}{h(y_i)}. \quad (4.2)$$

Efron et al. (2001) called the Bayesian version of FDR $\tau_0(y_i)$ the local false discovery rate (lFDR). Efron and Tibshirani (2002) showed that the relationship between FDR and lFDR is as follows:

$$FDR(y) = E_h[lFDR(Y)|Y \leq y].$$

Compared to the FDR, the local FDR provides a measure of belief that a gene $i$ is significant based on the corresponding statistic $Y_i$.

### 4.4.2 The Mixture Density $h$ and the Null Density $h_0$

**Nonparametric empirical Bayes approaches**

There have been a number of different approaches to estimate the density $h(y_i)$ and $h_0(y_i)$ in (4.2). Among them is the empirical nonparametric Bayes model which avoids strong parametric assumptions. Efron et al. (2001) calculated the expression scores by data reduction techniques starting from a *difference matrix* and a *null difference matrix*. The difference matrix is the design matrix used to calculate the difference of genes’ expression levels *between* different groups, while the null difference matrix calculates the difference of genes’ expression levels *within* the same group. So, the expression scores from the difference matrix, $Y_i^d$, follow $h(y_i)$, but the expression scores from the null difference matrix, $Y_i^{nd}$, follow $h_0(y_i)$. Then the ratio $h_0(y_i)/h(y_i)$ is estimated from the empirical distributions of $\{Y_i^d\}$ and $\{Y_i^{nd}\}$.
Efron and Tibshirani (2002) used an empirical Bayes method based on Wilcoxon statistics. The expression levels of seven BRCA1 genes and eight BRCA2 genes are ranked to produce the rank sum statistics

\[ Y_i = \text{Sum of BRCA2 ranks}, \; i = 1, \ldots, 3226 \quad (4.3) \]

The lowest value of \( Y_i \) is 36 \((= \sum_{a=1}^{8} a)\) and the largest value is 92 \((= \sum_{a=8}^{15} a)\). If we have a small (or large) value of \( Y_i \), the \( i \)-th gene is called underexpressed (or overexpressed) for BRCA2 tumor compared to BRCA1 tumor. If the null hypotheses, “\( H_i: \) the \( i \)-th gene is not differentially expressed”, are exactly true, the rank sum statistics \( Y_i \) follow the permutation distribution of the numbers 1, 2, \ldots, 15. That is, \( h_0(y_i) \) in (4.2) follows the Wilcoxon Rank-Sum (7,8) distribution. The mixture density \( h(y_i) \) can be estimated by fitting a Poisson regression model to the \( Y_i \) counts.

**Parametric Bayes approaches**

As a parametric Bayesian method, Lee et al. (2000) firstly proposed a normal mixture model for gene expression data. They specifically assumed normal distributions on \( h_0(y) \) and \( h_1(y) \) in (4.1). McLachlan, Bean and Jones (2006) also used the two-component normal mixture model which was fitted via the EM algorithm. Using the breast cancer data above, they calculated the classical pooled \( t \)-statistics \( Y_i \) for each gene \( i \) and computed the two-sided \( p \)-values, \( P_i \), with the null distribution a \( t_{(13)} \) distribution. Then, the observed \( p \)-values were transformed into \( z \)-scores given by

\[ z_i = \Phi^{-1}(1 - P_i), \quad (4.4) \]

where \( \Phi \) is the standard normal distribution. Since the \( p \)-value follows a uniform distribution under the null hypothesis, the statistic \( Z_i \) corresponding to the observed
$z_i$ under the null hypothesis is exactly standard normal. Then the density of the test statistics $Z_i$ is expressed by the two-component mixture normal model

$$h(z_i) = p_0 h_0(z_i) + p_1 h_1(z_i) = p_0 \phi(z_i; 0, 1) + p_1 \phi(z_i; \mu, \sigma^2),$$

where $\phi(z_i; 0, 1)$ is the theoretical null distribution when the $i$-th gene is not differentially expressed.

To estimate $p_0$, $\mu$, and $\sigma^2$, the maximum likelihood via the EM algorithm was used. Starting from the appropriate initial estimate, $\hat{\pi}_0$, McLachlan et al. (2006) obtained

$$\hat{\mu} = \bar{y}/(1 - \hat{\pi}_0)$$

and

$$\hat{\sigma}^2 = \left( s_y^2 - \hat{\pi}_0 - \hat{\pi}_0(1 - \hat{\pi}_0)\hat{\mu}^2_1 \right)/(1 - \hat{\pi}_0),$$

and showed that the mixture model can be fitted very quickly via the EM algorithm to the gene-expression data.

### 4.5 Example 4: Breast Cancer Microarray Experiment

To show the applicability of the GMSS in practice, we fit a parametric Bayesian normal mixture model to the data from Hedenfalk et al. (2001).

#### 4.5.1 Obtain Test Statistics

We use the $p$-values, $P_i$, obtained from two-sample $t$-tests of $m = 3,170$ genes that were used by Storey and Tibshirani (2003). The original data set has 3,226
genes on \( n_1 = 7 \) BRCA1 arrays and \( n_2 = 8 \) BRCA2 arrays and some arrays from sporadic breast cancer. Storey and Tibshirani eliminated 56 genes which do not seem trustworthy. Since the focus is to find differentially expressed genes between BRCA1- and BRCA2-related cancer, they eliminated the data for sporadic cancers and produced two-sample \( t \)-statistics

\[
t_i = \frac{\bar{x}_{i2} - \bar{x}_{i1}}{\sqrt{s^2_{i1}/n_1 + s^2_{i2}/n_2}} \quad \text{for } i = 1, \ldots, m, \tag{4.7}
\]

where \( x_{ij} \) is the log\(_2\) expression value from the \( j \)th array and the \( i \)th gene, \( \bar{x}_{i1} = 1/n_1 \sum_{j \in \text{BRCA1}} x_{ij} \), and \( s^2_{i1} = 1/(n_1 - 1) \sum_{j \in \text{BRCA1}} (x_{ij} - \bar{x}_{i1})^2 \). Then \( p \)-values are computed using the permutation method,

\[
p_i = \frac{B \sum_{b=1}^B \sum_{i=1}^m [t_{i1}^{0b} \geq |t_i|]}{B \cdot m} \tag{4.8}
\]

where a set of null statistics \( t_{i1}^{0b}, \ldots, t_{im}^{0b}, b = 1, \ldots, B \) is recomputed by assigning \( n = 15 \) arrays to two groups in all possible ways. (The \( p \)-values are available in the R package ‘qvalue’, made by Alan Dabney and John D. Storey at “http://CRAN.R-project.org/package=qvalue”.)

Following McLachlan et al. (2006), we transform the \( P \)-value in (4.7) to a \( z \)-score by

\[
z_i = \Phi^{-1}(1 - P_i). \tag{4.9}
\]

After the transformation, the non-null distribution of the \( z \)-score can be expressed as a single normal distribution with a positive mean.
4.5.2 Bayesian Normal Mixture Model and MCMC Methods

Using the transformed $z$-values at in (4.9), we fit a two component mixture model

$$h(z_i) = (1 - \delta)\phi(z_i; 0, 1) + \delta\phi(z_i; \mu, \sigma^2), \quad i = 1, \ldots, m. \quad (4.10)$$

The theoretical null distribution $\phi(z_i; 0, 1)$ assumes that a gene is not differentially expressed between BRCA1- and BRCA2-related cancers.

Motivated by the work of Richardson and Green (1997) on Bayesian mixture models, the prior distributions for $\mu$, $\sigma^2$ and $\delta$ in (4.10) are

$$\mu \sim N(\xi, \kappa^{-1}) \ , \ \sigma^2 \sim IG(a, b) \ , \ \text{and} \ \delta \sim Unif(0, 1)$$

with independence of $\mu$, $\sigma^2$ and $\delta$, where $IG(a, b)$ denotes an inverse gamma distribution with mean $b/(a - 1)$. Weakly informative priors are used: $\xi$ is the midrange, and $\kappa = 1/R^2$ where $R$ is the range. For simplicity, the hyperparameters $a$ and $b$ are fixed as 2.

For comparison of different MCMC methods, the Bayesian mixture model is fitted via the Gibbs sampler, the ordinary MH, and the GMSS. For the Gibbs sampler, we adopt the latent allocation variable in Richardson and Green (1997) so that

$$z_i|\lambda_i \sim N(0, 1)^{1 - \lambda_i}N(\mu, \sigma^2)^{\lambda_i} \ \text{and} \ \lambda_i \sim Ber(\delta),$$

where $\lambda_i$ is the label of the group from which observation $i$ is drawn. For the GMSS, the multiset of size $K = 2$ is composed of vectors $\{\theta_1 = (\mu_1, \sigma_1^2, \delta_1)^T, \theta_2 = (\mu_2, \sigma_2^2, \delta_2)^T\}$ and the instrumental density $f = f_\mu \times f_\sigma \times f_\delta$ is used where

$$f_\mu(\mu) \propto \begin{cases} \frac{1}{0.1}\phi\left(\frac{\mu - 0.5}{0.1}\right) & \text{if } \mu < 0.5 \\ \frac{1}{0.1}\phi(0) & \text{if } 0.5 \leq \mu < 2.0 \\ \frac{1}{0.1}\phi\left(\frac{\mu - 2.0}{0.1}\right) & \text{if } \mu \geq 2.0 \end{cases} \quad (4.11)$$
and \( f_\delta \sim Unif[0, 1] \). The multiset weights are set equally as 0.5. The ordinary MH algorithm and the GMSS use random walk proposals with \( \mu' \sim N(\mu(t-1), 0.2^2) \), \( \sigma^2' \sim \logNormal(\log(\sigma^2(t-1)), 0.25^2) \) and \( \delta' \sim N(\delta(t-1), 0.08^2) \).

Note that we cannot use the MSS in this example since the components of the multisets \( s = \{\theta_1, \theta_2\} \) are not nuisance parameters, but of interest and MSS does not have the reliable estimator for \( \pi(\theta) \) from the sampling distribution \( \pi(\{\theta_1, \theta_2\}) \).

In addition, it is not easy to set the bounded region of parameters, i.e., (bounded) uniform distribution for \( f_k \). If we set too small a boundary for \( \theta \), the Markov chain is not irreducible. On the other hand, too big a boundary for \( \theta \) makes the chain inefficient, as many components of the multiset just wander around the area of near zero mass. However, the GMSS with any continuous instrumental density \( f_k \) is free from the bounded region problem.

### 4.5.3 Results

Figure 4.1 presents posterior kernel density estimates of five independent chains under Gibbs sampling, MH and the GMSS. All runs consist of \( 10^7 \) iterates following an initial burn-in period. The fifteen density estimates are nearly indistinguishable to the eye, indicating that, for a long enough run, all three samples lead to similar inference. This result implies that the three algorithms eventually find the same
Figure 4.1: Posterior densities from each of 5 replicates of Gibbs sampling, MH and the GMSS with $10^7$ iterations. The comparability of these density estimates establishes the variability of this as a target.

However, for shorter runs, differences in mixing among the samplers are apparent.

Figure 4.2 presents a set of density estimates and trace plots from runs of length 10,000 under Gibbs sampling, MH and the GMSS. The substantively greater variation of the density estimates under the Gibbs sampler and the MH are attributable to the poorer mixing of these algorithms. The panels on the right detail the mixing for particular runs. The trace plots for the Gibbs sampler and MH show less frequent visits to the minor mode of the model, while the trace plot of the GMSS (values plotted are $\mu^*(t) = w_1 \mu_1^{(t)} + w_2 \mu_2^{(t)}$) shows more frequent visits. Actual visits of the GMSS chain to the minor mode are a little more frequent than appears from the plot, as the relatively high weight, $w_k$, associated with the major mode impacts $\mu^*(t)$. 
Figure 4.2: (Left) Kernel density estimates of $\mu$ from 10,000 iterations after a burn-in of 1,000 iterations under Gibbs sampling, MH, and the GMSS. The solid line is the average of density estimates from five independent Gibbs runs, each of $10^7$ iterations. (Right) A trajectory of 11,000 draws from each algorithm.
Table 4.1 compares the three algorithms in terms of the mean squared error (MSE) of point estimates and the total variation (TV) of density estimates. For conservative evaluation of the GMSS relative to the other two methods, we use the posterior distribution from the long runs of the Gibbs chain in Figure 4.1 as the desired target distribution. Specifically, the average value of the five posterior means from the Gibbs runs with $10^7$ iterations, whose values are $(\mu, \log \sigma^2, \delta) = (1.457, 0.037, 0.384)$, is used as the parameter of the MSE and the average density estimate from the five Gibbs runs is used as the (empirical) target density for TV.

We sample 200 independent runs with $10^4$ iterations under each of the three algorithms, resulting in 200 posterior summaries from each for calculation of MSE and TV. As seen in the table, the mean estimate from the GMSS, calculated by using the identity function $g(\theta) = \theta$, is unbiased for $E_\pi(\theta)$, while the mean estimators under Gibbs sampling and MH have large biases. Specifically, the point estimators of Gibbs sampling and the MH algorithm are placed toward the mode with small density, implying that the chains under the two algorithms are easily trapped in a local mode. Along with big biases, they also have larger variances, resulting in bigger MSE than does the GMSS.

To avoid the concern in assessing the bimodal, skewed distribution using MSE, we also calculate the total variation (TV) distance based on 50 equal width intervals. For the GMSS to estimate the target density $\pi(\theta)$ from the multiset density $\pi^*(\theta_1, \theta_2)$, the indicator function $g(\Theta) = I[\Theta \leq \theta_r]$ is used. The smaller values of TV for the GMSS in the table show that the GMSS generates the most efficient Markov chain for the breast cancer data set among the three algorithms.
The last column of the table shows the computation time for 10,000 iterations. For a fair comparison, we adjust the different computational burdens by using different numbers of iterations to make the computation time equal for the algorithms. Specifically, we compare the posterior results from 26,100 iterates of Gibbs, 21,582 iterates of MH, 14,876 iterates of the GMSS ($K = 2$, type I), 10,000 iterates of the GMSS ($K = 2$, type II), and 5,119 iterates of the GMSS ($K = 4$, type I). Although the results are not provided here, the results given the same computation time are not much different than the previous results. The performance of Gibbs sampling and MH improves a little as their runs get longer, but those MCMC methods still produce biased posterior means and larger MSE than does the GMSS.
Table 4.1: Mean estimate and total variation from 10,000 iterations after a burn-in of 1,000 iterations under Gibbs sampling, MH, and the GMSS. The GMSS leads to more accurate estimators.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>SD</th>
<th>MSE</th>
<th>Estimate</th>
<th>SD</th>
<th>MSE</th>
<th>Comp. time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(sec/10^4 iter)</td>
</tr>
<tr>
<td>Gibbs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu$</td>
<td>1.437</td>
<td>0.102</td>
<td>1.1E-02</td>
<td>2.74</td>
<td>2.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.049</td>
<td>0.062</td>
<td>4.0E-03</td>
<td>2.63</td>
<td>2.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.394</td>
<td>0.049</td>
<td>2.5E-03</td>
<td>5.47</td>
<td>5.14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu$</td>
<td>1.439</td>
<td>0.099</td>
<td>1.0E-02</td>
<td>3.08</td>
<td>2.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.048</td>
<td>0.061</td>
<td>3.8E-03</td>
<td>2.95</td>
<td>2.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.393</td>
<td>0.047</td>
<td>2.3E-03</td>
<td>5.89</td>
<td>4.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMSS, $K = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Type I)</td>
<td>$\mu$</td>
<td>1.459</td>
<td>0.041</td>
<td>1.7E-03</td>
<td>2.70</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.035</td>
<td>0.027</td>
<td>7.2E-04</td>
<td>2.44</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.383</td>
<td>0.019</td>
<td>3.6E-04</td>
<td>4.94</td>
<td>1.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMSS, $K = 2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Type II)</td>
<td>$\mu$</td>
<td>1.457</td>
<td>0.041</td>
<td>1.7E-03</td>
<td>2.12</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.037</td>
<td>0.027</td>
<td>7.0E-04</td>
<td>1.95</td>
<td>0.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.384</td>
<td>0.019</td>
<td>3.7E-04</td>
<td>3.90</td>
<td>1.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMSS, $K = 4$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Type I)</td>
<td>$\mu$</td>
<td>1.460</td>
<td>0.025</td>
<td>6.2E-04</td>
<td>1.96</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>$\log(\sigma^2)$</td>
<td>0.035</td>
<td>0.017</td>
<td>2.8E-04</td>
<td>1.82</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.383</td>
<td>0.011</td>
<td>1.3E-04</td>
<td>3.30</td>
<td>1.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Chapter 5: Application to Marketing Example:
Simultaneous Equation

5.1 Introduction

Simultaneous equations are widely used and studied in many literatures, including those of economics and marketing. Simultaneity of demand and supply is one example among a vast array of simultaneous relationships. Classical analysis of demand and price using a regression method assumes that price is set outside of the equation, i.e., it is exogenous. However, the endogeneity of price is often observed in practice as firms use a pricing strategy. Price sensitivity of customers affects the sales of a good, and the price is set by sellers who consider the price sensitivity of customers. The existence of endogeneity in price leads to a bias in classical regression analysis of these data, and so one turns to the simultaneous equation modeling as an alternative to the classical approach (refer to Yang, Chen and Allenby 2003).

The complex structure of simultaneous equations makes parameter estimation challenging. Traditional approaches do not consider the common shocks to demand and supply sides to be parameters. Instead, the shocks are considered random effects, and are integrated out of the likelihood function. However, the integration is computationally burdensome and may not be even possible with our current computational
capability. To overcome this challenge, Yang et al. (2003) introduced a Bayesian approach which does not require integration over the demand shocks. Instead, their approach uses MCMC and data augmentation (DA) to generate all unobservable quantities to provide posterior summaries for estimation of quantities of interest.

However, simultaneous equation models can have multiple modes, and a MCMC sampler may get stuck in a local mode. To alleviate this problem, we use the GMSS to fit the simultaneous equation model. Two simulation examples are introduced to show how an ordinary MH chain can become trapped and how the GMSS allows the chain to escape the trap.

5.2 Example 5: Toy Example with Simultaneous Equation

This example is originally from the lecture notes of Allenby (2010). Let us consider the simultaneous system of equations for the observed data \((x_i, y_i)\):

\[
x_i = \phi_1 + \phi_2 g_i + \phi_3 y_i + \zeta_i, \quad \zeta_i \sim N(0, \sigma^2_\zeta)
\]
\[
y_i = \psi_1 + \psi_2 g_i + \psi_3 x_i + \nu_i, \quad \nu_i \sim N(0, \sigma^2_\nu)
\]

where \(g_i\) are exogenous variables and the random variables \(\zeta_i\) and \(\nu_i\) denote shocks to the system.

A thousand simulated data points of \((x_i, y_i)\) are drawn under \(\sigma^2_\zeta = \sigma^2_\nu = 1, \ g_i \sim U(0, 1), \ \phi = (1, 2, 3)^T\), and \(\psi = (2, 1, 4)^T\). To conduct a Bayesian analysis, flat priors are used: \(\phi \sim MVN(0,100I)\) and \(\psi \sim MVN(0,100I)\). For simplicity, we assume that \(\sigma^2_\zeta\) and \(\sigma_\nu\) are known. The likelihood of the observed data \((x_i, y_i)\) can be derived
by the change-of-variable technique as follows:

\[ p(x, y|\phi, \psi) = \prod_{i=1}^{I} p(x_i|\phi, \psi) p(y_i|\phi, \psi)[1 - \phi \psi^3] \]  

(5.1)

where \( x = (x_1, \ldots, x_I) \) and \( y = (y_1, \ldots, y_I) \). We run ordinary MH and the GMSS for comparison.

**Results from ordinary MH**

Figure 5.1 shows the chains consisting of 10,000 draws of MCMC from ordinary MH algorithms under two different starting points. We use the proposal distributions \( \phi' \sim MVN(\phi^{(t-1)}, \tau^2 \phi I) \) and \( \psi' \sim MVN(\psi^{(t-1)}, \tau^2 \psi I) \) with \( \tau^2 = \tau^2 = 0.2 \). From the plots, the performance of MCMC varies with the different starting points: the chain in Panel (a) under the starting points \( \phi^{(0)} = (1, 1, 1) \) and \( \psi^{(0)} = (3, 3, 3) \) successfully converges to a value very close to the true parameter, while the chain in Panel (b) with the starting points \( \phi^{(0)} = (0.5, 0.5, 0.5) \) and \( \psi^{(0)} = (0.5, 0.5, 0.5) \) gets stuck in a local mode which is far from the global mode. Figure 5.2 shows the contour plot of the log posterior density as a function of the multiplicities \( C_\phi \) and \( C_\psi \) where the \( x \) and \( y \) axes refer to \( \phi = C_\phi (1, 2, 3)^T \) and \( \psi = C_\psi (2, 1, 4)^T \), respectively. The plot shows four modes, and deep valleys across which the ordinary MH chain cannot move.

**Results from the GMSS**

To find the global maximum of the log posterior density, we use the GMSS with multiset, \( s = (\psi_1, \psi_2, \psi_3) \), where \( \psi_k \) is a three-dimensional vector. Equal weights \( \alpha_1 = \alpha_2 = 1/2 \) and uniform instrumental densities \( f(\psi_k) = f_k(\psi_k) = \frac{1}{20^3} \mathbb{I}[\psi_k \in (-10, 10)^3] \) are used for \( k = 1, 2, 3 \). With this setting, the conditional distribution of the multiset
Figure 5.1: Samples from ordinary Metropolis-Hastings with different starting points: (a) \( \phi^{(0)} = (1, 1, 1) \), \( \psi^{(0)} = (3, 3, 3) \) and (b) \( \phi^{(0)} = (0.5, 0.5, 0.5) \), \( \psi^{(0)} = (0.5, 0.5, 0.5) \). The dotted lines denote the true values of parameters. Paths beginning at different starting values exhibit different behavior.
Figure 5.2: Contour plot of the log posterior density as a function of the multiplicities for parameters, $C_\phi$ and $C_\psi$, where $\phi = C_\phi(1, 2, 3)^T$ and $\psi = C_\psi(2, 1, 4)^T$. The plot shows four local modes. The thick dark bands are extremely deep valleys.
$s$ can be expressed as
\[
\pi^*(s|\phi) = 1/2 \pi(\psi_1|\phi) f(\psi_2) + 1/2 \pi(\psi_2|\phi) f(\psi_1),
\]
and the posterior (importance) weights from the GMSS chain are
\[
w_k = \frac{\alpha_k \pi(\psi_k|\phi) \prod_{l \neq k} f(\psi_l)}{\pi^*(s|\phi)}, \quad k = 1, 2, 3.
\]

From the trace plots of $\phi$ and $\psi^* (= \sum_k w_k \psi_k \prod f(\psi_i))$ in Figure 5.3, it is shown that the GMSS escapes from the local mode. 5.4 shows how the elements of the multiset work. At the beginning of the runs, the second element $\psi_2$ finds the local mode and explore the region, while the other elements freely move the space. The third element $\psi_3$ finds the global mode at 5,500 iteration by traversing the big valley between the two modes. From experiments with different starting points, we observed that the GMSS always found the global mode, no matter what starting point was given. In this example, the posterior density corresponding to the global mode is much larger than that corresponding to the local mode, so we never see that the GMSS chain returns back to the local mode.

### 5.3 Example 6: Simultaneous Demand and Supply Example

Otter, Gilbride and Allenby (2011) introduced the monopolist pricing example using a Bayesian hierarchical model where the joint likelihood for the data combines a traditional demand model and a term for the endogenous price variable as follows:

\[
\log y_i = \beta_0 + \beta_1 \log p_i + \varepsilon^*_i, \quad \varepsilon^*_i = \varepsilon_i + \eta_i
\]
\[
\log p_i = \log MC_i + \log \left( \frac{\beta_1}{1 + \beta_1} \right) + \nu^*_i, \quad \nu^*_i = \nu_i + \eta_i
\]
\[
\varepsilon_i \sim N(0, \sigma^2_\varepsilon), \quad \nu_i \sim N(0, \sigma^2_\nu), \quad \eta_i \sim N(0, \sigma^2_\eta)
\]
\[
(5.2)
\]
Figure 5.3: Samples from the GMSS with different starting points: (a) $\phi^{(0)} = (1, 1, 1)$, $\psi^{(0)} = (3, 3, 3)$ and (b) $\phi^{(0)} = (2, 2, 2)$, $\psi^{(0)} = (-1, -1, -1)$. $\psi^* = \sum_k w_k \psi_k \prod_{l \neq k} f_l(\psi_l)$. The GMSS finds the global mode in both settings.
Figure 5.4: The trajectories of three elements of the multiset, $\psi_k, k = 1, 2, 3$, from the GMSS algorithm with the starting point, $\phi^{(0)} = (2, 2, 2)$ and $\psi^{(0)} = (-1, -1, -1)$. The global mode (right side of plots) contains nearly all of the posterior mass.
where \( y_i \) is demand, \( p_i \) is price, \( \varepsilon_i \) and \( \nu_i \) are shocks to the individual system, \( \eta_i \) is a latent variable, and \( MC_i \) is the marginal cost. The latent variable \( \eta_i \) explains the common shock and introduces correlation between demand and price. Following the setting of Otter et al. (2011), we generate two simulated data sets, each having 52 samples. For both data sets, parameters are taken to be \( \beta_0 = 1, \sigma^2_\varepsilon = \sigma^2_\nu = 0.1, \sigma^2_\eta = 0.5 \) and \( \log MC_i = 2 + \log \Gamma(30,15) \), where \( \Gamma(a,b) \) denotes a gamma distribution with a mean \( (= a/b) \). \( \beta_1 \) is assumed to have the value \(-2\) for data set I and \(0.5\) for data set II.

**Ordinary MCMC with DA.**

Yang et al. (2003) introduced the Bayesian method of data augmentation (DA) for this problem. DA is used to integrate over the distribution of \( \eta_i \) to obtain the posterior density of interest, \( \pi(\beta_0, \beta_1, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta / \log y_i, \log p_i) \). Following Equation (28) of Yang et al. (2003), the joint posterior density of the model (5.2) can be written as

\[
[\beta_0, \beta_1, \eta, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta / \log y_i, \log p_i] 
\propto [\log y_i, \log p_i, \beta_0, \beta_1, \eta, \sigma^2_\varepsilon][\log p_i, \beta_1, \eta, \sigma^2_\nu][\log y_i, \sigma^2_\eta][\beta_0, \beta_1, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta].
\]

(5.3)

where \( y = (\log y_1, \ldots, \log y_I) \) and \( p = (\log p_1, \ldots, \log p_I) \). From the second equation in (5.2), we can see that values of \( \beta_1 \) between \(-1\) and \(0\) are prohibited, corresponding to a likelihood of \(0\) and a log-likelihood of \(\pm\infty\). Instead of using \( \beta_1 \), we introduce the one-to-one function \( \gamma = \log \left( \frac{\beta_1}{1 - \beta_1} \right) \). Figure 5.5 shows the relationship between \( \gamma \) and \( \beta_1 \).

A new posterior density, \([\beta_0, \gamma, \eta, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta / \log y_i, \log p_i, \log MC] \), is obtained by simply adding the Jacobian, \( |J| = e^\gamma/(1 - e^\gamma)^2 \), to (5.3). Now, assume that the prior distributions are \( \beta_0, \beta_1 \sim N(0,100) \) and \( \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta \sim IG(3,0.2) \), where \( IG(a,b) \) denotes an inverse gamma distribution with mean \( b/(a-1) \). Then, the posterior density
Figure 5.5: Relationship between $\gamma$ and $\beta_1$. The range of $\gamma$ is a real line but $\beta_1$ cannot have the value between -1 and 0.
can be sampled by using the Metropolis-within-Gibbs sampler, with the steps below. For each step, the proposal distributions will be specified, the acceptance probability calculated, and acceptance-rejection of the proposed value determined.

1. Initialize the sampler.

2. Update $\beta_0$ with a normal random-walk proposal distribution

$$
\beta'_0 \sim N(\beta^{(t-1)}_0, \tau^2_{\beta_0}).
$$

3. Update $\gamma$ with a normal random-walk proposal distribution

$$
\gamma' \sim N(\gamma^{(t-1)}, \tau^2_\gamma).
$$

4. Update $\sigma^2_\epsilon$ from a mixture of independent proposal distributions

$$
\sigma'^2_\epsilon \sim IG(3, 5) \text{ with prob 0.5, and } \sigma'^2_\epsilon \sim IG(3, 0.5) \text{ with prob 0.5.}
$$

5. Update $\sigma^2_\nu$ from a mixture of independent proposal distributions

$$
\sigma'^2_\nu \sim IG(3, 5) \text{ with prob 0.5, and } \sigma'^2_\nu \sim IG(3, 0.5) \text{ with prob 0.5.}
$$

6. Update $\sigma^2_\eta$ with a Gibbs proposal

$$
\sigma'^2_\eta \sim IG(A^*, B^*)
$$

where $A^* = A + n/2$ and $B^* = B + \sum_i \eta_i^{(t)}/2$.

7. Generate $\eta^{(t)}$ from a random-walk proposal distribution. To avoid a low acceptance ratio, we only propose a fraction of new $\eta_i$ as follows:

(i) Draw integer change points $(i_1, i_2, \ldots, i_r)$ from the values $i = 1, \ldots, 52$. We
use \( r = 3 \) integer points.

(ii) Propose \( r = 3 \) values as follows:

\[
\eta' = \left( \eta_1^{(t-1)}, \ldots, \eta'_l, \ldots, \eta'_m, \ldots, \eta_5^{(t-1)} \right)
\]

where \( \eta'_l \sim N(\eta_l^{(t-1)}, \tau^2_\eta) \). Accept or reject the proposed value.

(iii) repeat (i)-(ii) for \( R = 10 \) times.

8. Repeat Steps 2 to 7 for \( T \) iterations.

The GMSS algorithm.

Now, we introduce the GMSS of size \( K \), \( s = \{ \beta_{0,k}, \gamma_k \}, \ k = 1, \ldots, K \). Let \( \alpha_k \) be the weight and \( f_k \) be the instrumental density for the \( k \)-th element, \( (\beta_{0,k}, \gamma_k) \) of the multiset. Since the GMSS can be directly applied to a Metropolis-Hastings algorithm, the GMSS algorithm is essentially the same as that of Metropolis-within-Gibbs above, except the GMSS uses the following sampling density

\[
\pi^*(s, \eta, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta|y, p) \propto \sum_{k=1}^{K} \alpha_k \pi(\beta_{0,k}, \gamma_k, \eta, \sigma^2_\varepsilon, \sigma^2_\nu, \sigma^2_\eta|y, p) \prod_{l \neq k} f_l(\beta_{0,l}, \gamma_l)
\]

We use the Type III GMSS algorithm for this example, with the steps below.

1. Initialize the sampler. The starting values of \( \beta_{0,k} \) and \( \gamma_k \) is randomly drawn from the support of the instrumental density \( f_k \), so that \( f_k(\beta_{0,k}^{(0)}, \gamma_k^{(0)}) = 1 \) for all \( k = 1, \ldots, K \).

2. Update \( \beta_{0,k} \), for \( k = 1, \ldots, K \), with a normal random-walk proposal distribution

\[
\beta_{0,k}' \sim N(\beta_{0,k}^{(t-1)}, \tau^2_{\beta_{0,k}}).
\]
3. Update $\gamma_k$, for $k = 1, \ldots, K$, with a normal random-walk proposal distribution

$$
\gamma'_k \sim N(\gamma_k^{(t-1)}, \tau^2_{\gamma}).
$$

Steps 4 to 8 are the same as those of Metropolis-within-Gibbs above.

**Simulation study and results**

We generate two simulated data sets from model 5.3, with different parameter values shown in Table 5.1. Each data set consists of 52 observations of $(\log y_t, \log p_t)$. For estimation, we run MH and the GMSS with two starting points of the $\gamma$, $\gamma^{(0)} = 0.7$ or $-0.1$. The other starting points of the chains are the same through all the settings as $\beta_0^{(0)} = 1$, $\sigma^2_\varepsilon = \sigma^2_\nu = 0.2$, and $\sigma^2_\eta = 0.7$. The combinations of different settings result in a designed experiment with three factors: (i) the parameters of the simulated data sets ($\beta_1 = -2$ or $0.5$), (ii) starting values of $\gamma$ ($\gamma^{(0)} = 0.7$ or $-0.1$), and (iii) estimation methods (MH or the GMSS). For MH and the GMSS, we use the proposal densities as described above in the GMSS algorithm with step sizes $\tau_\gamma = 0.05$, $\tau_{\beta_0} = 0.2$, and $\tau_\eta = 0.5$. For GMSS, we used a multiset of size $K = 2$ with equal weights $\alpha_k = 1/2$ and uniform instrumental densities $f_k(\beta_{0,k}, \gamma_k) \sim I[\beta_{0,k} \in (-10, 10)] \times I[\gamma_k \in (-2, 2)]$.

Table 5.1 shows the summary from eight different runs from the designed experiment. Each run consists of 10,000 burn-in iterations and an additional 10,000 samples for estimation. In the top table for Simulated Data Set I, the MH chain starting from $\gamma^{(0)} = 0.7$ correctly finds the global mode yielding estimates close to the true parameter value, while the MH chain starting from $\gamma^{(0)} = -0.1$ does not. Conversely, the bottom table for Simulated Data Set II shows that the MH chain with $\gamma^{(0)} = 0.7$ does not work, but that with $\gamma^{(0)} = -0.1$ works well. For the GMSS, it is shown that
the chains consistently produce the good estimates, indicating that GMSS runs are effectively free from the starting points in this example.

5.4 Discussion

The model for the simulated example above assumes the situation where the sign of the price coefficient is unknown a priori. Of course, in many practical situations, demand decreases as price rises, so one may add the assumption of a positive price coefficient to remove a portion of the parameter space, thereby creating a unimodal posterior and avoiding the local-trap problem. For example, one may put a prior distribution whose support is confined to positive price coefficients. However, there are special situations where the assumption of negative price coefficient may be true. One of the example is a Giffen good. A second involves situations where price connotes quality.

Previous research in marketing and economics imposes a restriction of negative correlations between demand and price, or simply disregards solutions with a negative price coefficient. However, GMSS makes it possible to consider the usual and unusual situations in a single unified framework. Rather than imposing the strongest imaginable restrictions on a model’s fit, we believe the combination of experimental data and a more modest prior should drive the analysis. The necessary tool that allows this to happen is a MCMC algorithms that moves about the candidate space freely, even when there is a big valley between local modes. The GMSS is just this sort of sampler.
Table 5.1: Estimates from Metropolis-Hastings and the GMSS with the multiset of size $K = 2$

Simulated Data Set I: $\beta_1 = -2$

<table>
<thead>
<tr>
<th>Parameters</th>
<th>True</th>
<th>MH $\gamma^{(0)} = 0.7$</th>
<th>MH $\gamma^{(0)} = -0.1$</th>
<th>GMSS $\gamma^{(0)} = 0.7$</th>
<th>GMSS $\gamma^{(0)} = -0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>1.00</td>
<td>0.87 (0.261)</td>
<td>-7.91 (0.577)</td>
<td>0.89 (0.244)</td>
<td>0.89 (0.373)</td>
</tr>
<tr>
<td>$\beta_1 = \frac{e^{\gamma}}{1+e^{\gamma}}$</td>
<td>-2.00</td>
<td>-1.96 (0.084)</td>
<td>0.58 (0.158)</td>
<td>-1.96 (0.080)</td>
<td>-1.96 (0.122)</td>
</tr>
<tr>
<td>$\sigma^2_\epsilon$</td>
<td>0.10</td>
<td>0.09 (0.029)</td>
<td>2.18 (0.593)</td>
<td>0.09 (0.029)</td>
<td>0.09 (0.035)</td>
</tr>
<tr>
<td>$\sigma^2_\nu$</td>
<td>0.10</td>
<td>0.12 (0.041)</td>
<td>3.77 (0.987)</td>
<td>0.12 (0.040)</td>
<td>0.12 (0.066)</td>
</tr>
<tr>
<td>$\sigma^2_\eta$</td>
<td>0.50</td>
<td>0.48 (0.115)</td>
<td>0.08 (0.065)</td>
<td>0.48 (0.114)</td>
<td>0.49 (0.131)</td>
</tr>
<tr>
<td>corr($\epsilon, \nu$)</td>
<td>0.83</td>
<td>0.83 (0.029)</td>
<td>-0.87 (0.012)</td>
<td>0.84 (0.029)</td>
<td>0.83 (0.049)</td>
</tr>
</tbody>
</table>

Simulated Data Set II: $\beta_1 = 0.5$

<table>
<thead>
<tr>
<th>Parameters</th>
<th>True</th>
<th>MH $\gamma^{(0)} = 0.7$</th>
<th>MH $\gamma^{(0)} = -0.1$</th>
<th>GMSS $\gamma^{(0)} = 0.7$</th>
<th>GMSS $\gamma^{(0)} = -0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>1.00</td>
<td>6.56 (0.224)</td>
<td>0.96 (0.085)</td>
<td>0.95 (0.083)</td>
<td>0.95 (0.086)</td>
</tr>
<tr>
<td>$\beta_1 = \frac{e^{\gamma}}{1+e^{\gamma}}$</td>
<td>0.50</td>
<td>-1.50 (0.172)</td>
<td>0.54 (0.078)</td>
<td>0.56 (0.063)</td>
<td>0.56 (0.065)</td>
</tr>
<tr>
<td>$\sigma^2_\epsilon$</td>
<td>0.10</td>
<td>2.18 (0.611)</td>
<td>0.10 (0.034)</td>
<td>0.09 (0.030)</td>
<td>0.08 (0.035)</td>
</tr>
<tr>
<td>$\sigma^2_\nu$</td>
<td>0.10</td>
<td>0.09 (0.049)</td>
<td>0.11 (0.055)</td>
<td>0.13 (0.041)</td>
<td>0.13 (0.048)</td>
</tr>
<tr>
<td>$\sigma^2_\eta$</td>
<td>0.50</td>
<td>4.92 (1.417)</td>
<td>0.49 (0.115)</td>
<td>0.47 (0.108)</td>
<td>0.47 (0.109)</td>
</tr>
<tr>
<td>corr($\epsilon, \nu$)</td>
<td>0.83</td>
<td>0.97 (0.001)</td>
<td>0.83 (0.031)</td>
<td>0.83 (0.024)</td>
<td>0.83 (0.024)</td>
</tr>
</tbody>
</table>
Chapter 6: Application to Bayesian Outlier Detection in Regression Analysis

6.1 Introduction

It is well known that the ordinary regression method (or the least squares method) is very sensitive to the existence of outliers. To decrease the influence of outliers, Edgeworth (1887) proposed the use of $L^1$ regression in which the squared residual is replaced by the absolute residual. Huber (1973) introduced the $M$ estimator, which suggests the use of a general symmetric function instead of the squared or absolute residual. Later, Rousseeuw (1984) suggested the least median of squares regression which uses the median of the squared residuals instead of the sum of the squared residuals.

Compared to classical regression analysis, Bayesian approaches are flexible in terms of adjusting the influence of outliers on the regression fit by adopting a model which explains the different features of outliers (e.g., Verdinelli and Wasserman 1991; Hurn, Justel and Robert 2003; Peruggia, Santner and Ho 2004; Mohr 2007). However, Bayesian outlier detection is often challenging since the Markov chain used to fit the model can easily fail to converge. Justel and Peña (1996) showed several examples of clustered outliers in regression analysis where the Gibbs sampling algorithm fails.
to converge. Justel and Peña (2001) found that convergence of the Gibbs sampler in outlier detection depends on initial values of latent variables. They proposed good initialization of the sampler as a solution to the convergence problem. However, even a good initialization may not alleviate poor mixing of the chain.

In this chapter, we show how the GMSS can easily replace the Gibbs sampler and demonstrate that it leads to better estimators than does the Gibbs sampler. Moreover, we study how sensitive regression analysis is to outliers in a set of simulated examples with various relative spreads and numbers of outliers.

6.2 Example 7: Rousseew Type Data and Bayesian Approach

We generate simulated data which is similar to the example of Rousseew (1984). Let us assume that we have 40 “inliers” from a linear model $y_i = 2 + x_i + \varepsilon_i$, $i = 1, \ldots, 40$, where $\varepsilon_i \sim N(0, \sigma^2)$, and 10 outliers from a bivariate normal distribution $(X_i, Y_i)^T \sim BVN((7, 2)^T, \nu^2 I)$, $i = 41, \ldots, 50$. In this section, we set $\sigma = 0.6$ and $\nu = 0.1$ and generate the simulated data in Figure 6.1. One aims to find the regression fit that explains the 40 observations with the linear pattern and to flag the 10 outliers. Hereafter, we call the 40 observations from the regression line the good points and the 10 observations from the bivariate normal distribution the bad points. The term outlier will denote a point which is interpreted as having different pattern from main portion of data as a result of an analysis.

If we fit the regression model for the 40 good observations only, the regression line is $\hat{y} = 2.32 + 0.89x$ as shown as solid line in Figure 6.1. However, if we fit the model to all observations, the mass of bad points prevents correct inference with the regression
Table 6.1: Least square estimator.

| Setting            | Parameter | Estimate | Standard Error | t-value | Pr(> |t|) |
|--------------------|-----------|----------|----------------|---------|------|------|
| Using all obs.     | Intercept | 5.40     | 0.32           | 17.05   | < 0.0001 |
|                    | Slope     | -0.38    | 0.08           | -4.76   | < 0.0001 |
| Using good obs. only | Intercept | 2.32     | 0.27           | 8.67    | < 0.0001 |
|                    | Slope     | 0.89     | 0.10           | 8.91    | < 0.0001 |

fit yielding the negative slope of $-0.38$ (see Table 6.1). This phenomenon is called *strong masking* by outliers (see Justel and Peña 1996).

As an alternative to the classic regression analysis, we adopt a Bayesian approach with the scale contaminated normal model (originally introduced by Tukey 1960; later used for Bayesian outlier detection by Justel and Peña 1996). The idea of scale contamination is to allow larger variability for the outliers as follows:

$$y_i = \beta_0 + x_i \beta_1 + \varepsilon_i \quad \text{where} \quad p(\varepsilon_i) = (1 - w) N(0, \sigma^2) + w N(0, c^2 \sigma^2) \quad (6.1)$$

where $w$ is the probability that an observation belongs to the outlier group. We use a nonininformative prior for the regression coefficients and $\sigma$: $p(\beta_0, \beta_1, \sigma) \propto \sigma^{-1}$ (equivalently, $p(\beta_0, \beta_1, \sigma^2) \propto \sigma^{-2}$).

To run the Gibbs sampler, latent variables are introduced to indicate the membership of observations (refer to Verdinelli and Wasserman 1991). Let $\delta_i$ be the indicator variable to identify an outlier: $\varepsilon_i \sim N(0, \sigma^2)$ if $\delta_i = 0$ and $\varepsilon_i \sim N(0, c^2 \sigma^2)$ otherwise. Then, the posterior density including the latent variables $\{\delta_i\}$ is

$$p(\beta_0, \beta_1, \sigma^2, \{\delta_i\} | \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i; \beta_0, \beta_1, \sigma^2, \delta_i) p(\delta_i) \cdot p(\beta_0, \beta_1, \sigma^2) \quad (6.2)$$
Figure 6.1: Rousseeuw type data. 40 good points from $y \sim N(2 + x, 0.6^2)$ and 10 bad points from $(x, y)^T \sim N((7, 2)^T, 0.1^2I)$. The dashed line is the least squares fit for all observations and the solid line is the least squares fit for the 40 good points only.
where \( y = (y_1, \ldots, y_n)^T \) and the prior distribution of \( \delta_i \) is \( p(\delta_i) \sim Ber(w) \).

### 6.2.1 Gibbs Sampling

Let \( X \) be a \( n \times 2 \) matrix whose \( i \)-th row is \( (1, x_i) \). To get samples from the posterior density in (6.2), the Gibbs sampling algorithm is as follows:

1. Initialize the sampler with the starting values \( (\beta_0^{(0)}, \beta_1^{(0)}, \sigma^2(0)) \).

2. Update \( \beta = (\beta_0, \beta_1)^T \) from the Gibbs proposal:

\[
\beta|\sigma^2, \{\delta_i\}, y \sim N \left( \hat{\beta}, \sigma^2 (X^T V X)^{-1} \right)
\]

where \( \hat{\beta} = (X^T V X)^{-1} X^T V y \), and \( V \) is a diagonal matrix with the \( i \)-th diagonal element \( v_{ii} = c^{-2} \) if \( \delta_i = 1 \) and \( v_{ii} = 1 \) otherwise.

3. Update \( \sigma^2 \) from the Gibbs proposal:

\[
\sigma^2 \sim IG \left( \frac{n}{2}, \frac{b}{2} \right)
\]

where \( b = \sum_{\{i: \delta_i = 1\}} (y_i - \beta_0 - x_i \beta_1)^2 / c^2 + \sum_{\{i: \delta_i = 0\}} (y_i - \beta_0 - x_i \beta_1)^2 \). \( IG(a, b) \) denotes an inverse gamma distribution with mean \( b/(a - 1) \).

4. Update \( \delta_i \) for \( i = 1, \ldots, n \) from the Gibbs proposal:

\[
\delta_i \sim Bernoulli(w^*)
\]

where

\[
w^* = \frac{w k^{-1} \exp\left( -\frac{(y_i - \beta_0 - x_i \beta_1)^2}{2c^2\sigma^2} \right)}{w k^{-1} \exp\left( -\frac{(y_i - \beta_0 - x_i \beta_1)^2}{2c^2\sigma^2} \right) + (1 - w) \exp\left( -\frac{(y_i - \beta_0 - x_i \beta_1)^2}{2\sigma^2} \right)}.
\]
5. Repeat steps 2 to 4 for $T$ iterations.

Now, we show that the convergence of the Gibbs chain for the outlier problem is very sensitive to its starting value as Justel and Peña (2001) found. We fix the values of the parameters, $w = 0.2$ and $c = 10.0$, and run the Gibbs sampler for 10,000 iterations after a burn in of 10,000 iterations. Two different starting points are used: (i) all bad points ($i = 41, \ldots, 50$) and some good points start with $\delta_i^{(0)} = 1$, and (ii) a portion of bad points have $\delta_i^{(0)} = 1$. Table 6.2 shows the posterior summary of the Gibbs samplers. Specifically, the Gibbs chain with the first setting for the starting point yields a positive slope coefficient (0.77) while the chain with the second setting yields a negative slope, concluding that the bad points are inliers.

Table 6.2: Posterior results from Gibbs sampling.

<table>
<thead>
<tr>
<th>Starting Value of ${\delta_i}$</th>
<th>Parameters</th>
<th>Posterior Mean</th>
<th>Posterior S.D</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_i = \begin{cases} 0 &amp; \text{for } i = 1, \ldots, 30 \ 1 &amp; \text{for } i = 31, \ldots, 50 \end{cases}$</td>
<td>$\beta_0$</td>
<td>2.60</td>
<td>0.29</td>
<td>(2.41, 2.79)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>0.77</td>
<td>0.11</td>
<td>(0.70, 0.84)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>0.27</td>
<td>0.06</td>
<td>(0.23, 0.30)</td>
</tr>
<tr>
<td>$\delta_i = \begin{cases} 0 &amp; \text{for } i = 1, \ldots, 45 \ 1 &amp; \text{for } i = 46, \ldots, 50 \end{cases}$</td>
<td>$\beta_0$</td>
<td>5.42</td>
<td>0.33</td>
<td>(5.19, 5.63)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>-0.39</td>
<td>0.08</td>
<td>(-0.45, -0.34)</td>
</tr>
<tr>
<td></td>
<td>$\sigma^2$</td>
<td>1.10</td>
<td>0.25</td>
<td>(0.92, 1.24)</td>
</tr>
</tbody>
</table>

To find the ‘good’ starting points, aiming at a regression fit with a positive slope, Justel and Peña (2001) proposed the use of the information derived from the covariance structure of the latent variables $\{\delta_i\}$. We empirically found that the Gibbs
sampler yields a positive slope when most bad points are identified as outliers initially. (i.e., \( \delta_i^{(0)} = 1, \forall i = 41, \ldots, 50 \).) If some bad points have a starting value of \( \delta_i = 0 \), the Gibbs sampler converges to the posterior density whose slope parameter is negative. For runs of this length with these starting values, we have not observed any switches between the positive slope and negative slope portions of the posterior distribution.

### 6.2.2 Metropolis-Hastings

To avoid strong dependence of the Markov chain on the starting values of \( \{ \delta_i \} \) and to improve mixing, we marginalize the \( \{ \delta_i \} \) and run a Metropolis-Hastings (MH) algorithm for the model in (6.1). When the fixed values of \( w = 0.2 \) and \( c = 10.0 \) are assumed, the posterior density is simply proportional to multiplication of the likelihood \( p(y|\beta_0, \beta_1, \sigma^2) \) and the prior density \( p(\beta_0, \beta_1, \sigma^2) \propto \sigma^{-2} \).

Random-walk normal and log-normal proposals are used to update \( \beta^{(t)} \) and \( \sigma^{2(t)} \), respectively:

\[
\beta'|\beta^{(t-1)} \sim BVN(\beta^{(t-1)}, diag(\tau_{\beta_0}^2, \tau_{\beta_1}^2))
\]

and

\[
\sigma^{2'|\sigma^{2(t-1)} \sim logNormal(\log\sigma^{2(t-1)}, \tau_{\sigma^2}^2)}.
\]

With the proposal scale \( (\tau_{\beta_0}, \tau_{\beta_1}, \tau_{\sigma^2}) = (0.2, 0.1, 1.5) \), we run MH for 10,000 burn-in iterations and additional 10,000 iterations for estimation. Two different starting points for \( (\beta_0, \beta_1, \sigma^2) \) are used: \( (5.5, -0.4, 1.0) \) and \( (2.5, 0.8, 0.3) \).

Table 6.3 shows the posterior estimates from MH. Despite omitting the latent variables \( \{ \delta_i \} \), the MH chain yields different results depending on the initial condition \( (\beta_0, \beta_1, \sigma^2)^{(0)} \). Interestingly, the MH chain yields two different explanations which are
Table 6.3: Posterior results from MH.

<table>
<thead>
<tr>
<th>Starting Value ((\beta_0^{(0)}, \beta_1^{(0)}, \sigma^{(0)}))</th>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior S.D.</th>
<th>95 % Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(5.5, -0.4, 1.0)</td>
<td>(\beta_0)</td>
<td>5.43</td>
<td>0.30</td>
<td>(5.21, 5.64)</td>
</tr>
<tr>
<td></td>
<td>(\beta_1)</td>
<td>-0.39</td>
<td>0.08</td>
<td>(-0.45, -0.34)</td>
</tr>
<tr>
<td></td>
<td>(\sigma)</td>
<td>1.09</td>
<td>0.25</td>
<td>(0.92, 1.23)</td>
</tr>
<tr>
<td>(2.5, 0.8, 0.3)</td>
<td>(\beta_0)</td>
<td>2.63</td>
<td>0.26</td>
<td>(2.45, 2.80)</td>
</tr>
<tr>
<td></td>
<td>(\beta_1)</td>
<td>0.76</td>
<td>0.10</td>
<td>(0.70, 0.83)</td>
</tr>
<tr>
<td></td>
<td>(\sigma)</td>
<td>0.27</td>
<td>0.06</td>
<td>(0.23, 0.30)</td>
</tr>
</tbody>
</table>

very similar to the results from the Gibbs sampler in Table 6.2. Figure 6.2 shows the trajectories of the MH chains with the different starting points. From the figure, we strongly suspect the existence of bimodality in the posterior distribution which, in a practical sense, is neither recognized nor is traversed by the MH algorithm (or the Gibbs sampler).

6.2.3 The GMSS

Using the same data set, we run the GMSS with multiset size \(K = 5\), with equal weight for each element of the multiset and uniform instrumental densities: \(f_k(\beta_{0,k}) \propto I[\beta_{0,k} \in (1.5, 7.0)]\), \(f_k(\beta_{1,k}) \propto I[\beta_{1,k} \in (-0.8, 1.2)]\) and \(f_k(\sigma_k^2) \propto I[\sigma_k^2 \in (0.0, 3.0)]\). These instrumental densities more than adequately cover the mass of the posterior distribution. The target distribution and proposal densities for updating parameters are the same as those of the previous subsections implementation of the MH algorithm.
Figure 6.2: Trajectory of the MH for Rousseeuw type data. (a) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (5.5, -0.4, 1.0)$. (b) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (2.5, 0.8, 0.3)$. 

92
Figure 6.3: Trajectory of GMSS estimates for Rousseeuw type data. (a) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (5.5, -0.4, 1.0)$. (b) Chain starting from $(\beta_0, \beta_1, \sigma^2)^{(0)} = (2.5, 0.8, 0.3)$. 

93
Table 6.4: Posterior results from the GMSS.

<table>
<thead>
<tr>
<th>Starting Value $(\beta_0^{(0)}, \beta_1^{(0)}, \sigma^{2(0)})$</th>
<th>Parameter</th>
<th>Posterior Mean</th>
<th>Posterior S.D.</th>
<th>95% Credible Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(5.5, -0.4, 1.0)$</td>
<td>$\beta_0$</td>
<td>5.38</td>
<td>0.64</td>
<td>(5.27, 5.69)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>-0.36</td>
<td>0.25</td>
<td>(-0.47, -0.34)</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>1.04</td>
<td>0.29</td>
<td>(0.88, 1.22)</td>
</tr>
<tr>
<td>$(2.5, 0.8, 0.3)$</td>
<td>$\beta_0$</td>
<td>5.37</td>
<td>0.52</td>
<td>(5.20, 5.63)</td>
</tr>
<tr>
<td></td>
<td>$\beta_1$</td>
<td>-0.37</td>
<td>0.19</td>
<td>(-0.45, -0.34)</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>1.09</td>
<td>0.30</td>
<td>(0.92, 1.23)</td>
</tr>
</tbody>
</table>

Table 6.4 shows that the GMSS yields consistent inference regardless of the stating values of the chain. Trace plots of GMSS estimates (e.g. $\beta_1^* = \sum_k w_k \beta_{1,k} \prod_{l \neq k} f_l(\beta_{1,l})$) in Figure 6.3 clearly show the existence of a bimodal distribution. The GMSS successfully find two explanations (or modes) which are not discovered by the Gibbs sampler and a single MH chain. More frequent visits of $\beta_1$ to the lower side also implies that the explanation corresponding to the negative value of $\beta_1$ (so, larger values of $\beta_0$ and $\sigma^2$) has larger likelihood. If we apply Justel and Peña (2001)’s approach to this data, the Gibbs chain will only find the area with the positive value of $\beta_1$ which is a local mode, having smaller likelihood and posterior density. The overall result can be interpreted as convergence to a local mode, i.e., the MCMC algorithm has fallen into a local trap. The GMSS results also lead us to rethink our description of the data. Under the model we have written, the negative slope mode provides a better explanation of the data than does the positive slope explanation.
6.3 Application of GMSS to Outlier Examples with Different Settings

In this section, we use the GMSS to study sensitivity of the regression model to relative spread and number of outliers.

6.3.1 Relative Spread of Outliers

First, we simulate Rousseeuw-type data with 40 good points from $y_i \sim N(\beta_0 + x_i\beta_1, \sigma^2)$ and 10 bad points from a bivariate normal distribution with mean $(7, 2)^T$ and variance $0.1^2 I$, with four different sizes of $\sigma$: 0.4, 0.6, 0.8 and 1.0. The simulated data sets appear in Figure 6.4.

With the four different data sets, we run the GMSS whose settings are exactly the same as those used in the previous sections. Figure 6.5 shows the GMSS density estimates of $\beta_1$ using $10^5$ samples after a burn-in period of length $10^4$. As expected, the posterior mode of $\beta_1$ has a positive value when the variance of the good points (i.e., $\sigma^2$) is small. As $\sigma^2$ increases, the contribution of the good points to the regression fit diminishes so the posterior mode of $\beta_1$ becomes negative.
Figure 6.4: Simulated data for 40 good points and 10 bad points with (a) \((\sigma, \nu) = (0.4, 0.1)\), (b) \((\sigma, \nu) = (0.6, 0.1)\), (c) \((\sigma, \nu) = (0.8, 0.1)\), and (d) \((\sigma, \nu) = (1.0, 0.1)\).
Figure 6.5: The GMSS posterior density of $\beta_1$ from a GMSS run of length $10^5$, under the simulated data for 40 good points and 10 bad points with (a) $(\sigma, \nu) = (0.4, 0.1)$, (b) $(\sigma, \nu) = (0.6, 0.1)$, (c) $(\sigma, \nu) = (0.8, 0.1)$, and (d) $(\sigma, \nu) = (1.0, 0.1)$. 
6.3.2 Number of Outliers

We next create a simulated data set having different numbers of bad points with fixed variance \((\sigma, \nu) = (0.4, 0.1)\). Figure 6.6 shows the resulting four simulated examples. From the GMSS of size \(K = 5\) with equal weights \(w_k\) and the uniform density used in Section 6.2.3, we get the GMSS density estimates from \(10^5\) samples, appearing in Figure 6.7. The posterior mode of \(\beta_1\) changes as the proportion of bad points changes; more bad points move more posterior mass to the negative value of \(\beta_1\). The dual negative modes in panels (c) and (d) of Figure 6.6 deserve comment. To confirm the patterns in the bottom panels, we re-run the GMSS for \(3 \times 10^6\) iterations after a burn in of \(10^6\) iterations. In Figure 6.6 and Figure 6.9, the modes both correspond to lines passing through the clump of bad points. The larger mode in each plot is for a flatter line and the smaller mode is for a steeper line.

6.4 Conclusion

In practice, it is not easy to find a good starting point for an MCMC chain when outliers are present in a data set. Rather, finding good starting points seems as difficult as finding the outliers which is the goal of the analysis. Although Justel and Peña (2001) proposed use of the probability from the covariance structure of the latent variables, in order to produce a Markov chain that converged to the “good mode”, one might still question the conclusions of such an analysis. Their Markov chain intentionally only explores one local mode, and this seems unstable when we examine different examples, since the Markov chain will converge to a global mode if it is close to the local mode. In our experiments, we found that the global mode
Figure 6.6: Simulated data under the parameter $(\sigma, \nu) = (0.4, 0.1)$ with (a) 40 good points and 10 bad points, (b) 37 good points and 13 bad points, (c) 34 good points and 16 bad points, and (d) 31 good points and 19 bad points.
Figure 6.7: The GMSS posterior density of $\beta_1$ from a GMSS run of length $10^5$, under the simulated data with (a) 40 good points and 10 bad points, (b) 37 good points and 13 bad points, (c) 34 good points and 16 bad points, and (d) 31 good points and 19 bad points.
Figure 6.8: The GMSS posterior densities of $\beta_0$ and $\beta_1$ from a GMSS run of length $3 \times 10^6$ (left panels) and the fitted regression lines (right panel), under the simulated data with 34 good points and 16 bad points. The dashed line is the regression fit from the global mode ($\beta_0 = 4.3, \beta_1 = -0.33$) and the dotted line is the regression fit from the local mode ($\beta_0 = 5.7, \beta_1 = -0.53$).
Figure 6.9: The GMSS posterior densities of $\beta_0$ and $\beta_1$ from a GMSS run of length $3 \times 10^6$ (left panels) and the fitted regression lines (right panel), under the simulated data with 31 good points and 19 bad points. The dashed line is the regression fit from the global mode ($\beta_0 = 3.9, \beta_1 = -0.26$) and the dotted line is the regression fit from the local mode ($\beta_0 = 6.0, \beta_1 = -0.56$).
can have either positive or negative slope coefficients, depending upon details of the example.

By applying the GMSS to the strong-masking outlier problem, we discover, in an automated fashion, the existence of bimodality corresponding to the two different explanation for the observed data. The GMSS density estimates describe both the existence of the modes and their relative importance in the posterior distribution. If the likelihood (hence the posterior density) of the explanation corresponding to the ‘bad’ model (with negative slope parameter) is large enough compared to that of the explanation corresponding to the ‘good’ model, we do not favor focusing exclusively on the ‘good’ model. Instead, we take this as evidence of a deficiency in the model and would attempt to devise a new approach to simultaneously explain the good points and the bad points. One such example is the Bayesian hierarchical model recently introduced by Mohr (2007). In the paper, they suggested a mixture component model assuming that the data set with outliers has subgroups to which the regression lines with different slopes and variances are fitted.
Chapter 7: Application to the Variance Components Model

7.1 Introduction

The variance component model is a hierarchical linear model where different populations have different structures. A simple example is the one-way random effect model,

\[ y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J_i \] (7.1)

where \( \theta_i \sim N(0, \tau^2) \), \( \varepsilon_{ij} \sim N(0, \sigma^2) \), and \( \theta_i \) and \( \varepsilon_{ij} \) are mutually independent. A more complex example is to use a regression model for the group mean, \( \theta_i = \gamma z_i + \delta_i \) where \( \delta_i \sim N(0, \sigma^2_\delta) \).

Bayesian analysis in the variance component model has been widely studied. Hill (1965) proposed the estimation of variance components from a subjective Bayesian point of view and Sun et al. (1996) performed a precise Bayesian analysis by modern computation techniques. However, some difficult problems still remain. Two main topics in this class of the problems have attracted particular attention. They are (i) the choice of proper/improper prior distribution (refer to Hobert and Casella 1996; Lambert, Sutton, Burton, Abrams and Jones 2005; Gelman 2006) and (ii) multimodality of the posterior distribution (refer to Liu and Hodges 2003).
This chapter focuses on the multimodality of the posterior distribution in a variance component model where an ordinary MCMC method suffers from poor mixing. We also discuss another example of bad mixing of Markov chains where a hierarchical model with many precision parameters yields an “irregular” shape for the posterior density. We find that both cases benefit from the use of the GMSS.

7.2 Example 8: Random Effect Model with HMO Data

Despite the prevalent use of variance component models, the multimodality of posterior distribution in the class of the hierarchical models had not been studied much until Liu and Hodges (2003) found a condition that yields bimodality of the posterior joint density of $\sigma^2$ and $\tau^2$. They provided several theorems to characterize the bimodality and found the locations of multiple modes in the setting of the balanced one-way random effect model. They emphasized that the multimodality in a variance component model arises even in the simplest one-way random setting, regardless of whether informative or non-informative priors are used.

If the number of variance components in a hierarchical model increases, the estimation of the precision parameters becomes more difficult. There are a number of examples of the variance component model with three or more precision parameters, such as the conditionally autoregressive (CAR) model with two types of neighbor relations (2NRCAR) and crossed random-effects models (see He, Hodges and Carlin 2007).

In this section, we fit a hierarchical model to empirical data which yields a bimodal posterior density, and we compare the performance of the GMSS with that of
ordinary MCMC methods.

### 7.2.1 Health Maintenance Organization Data

Health maintenance organizations (HMOs) are major providers of health care in the United States. This data set contains data on member health needs in 42 states, the District of Columbia, Guam and Puerto Rico. Hodges (1998) used the Gibbs sampler to analyze each HMO’s monthly premium for the individual subscriber with expenses per admission for each state. Let $y_{ij}$ be the premium for plan $j$ in state $i$. $z_{1,i}$ denotes (centered and scaled) expenses per admission for the $i$th state and $z_{2,i}$ is an indicator for New England. Hodges fitted the random effect model with regressors as

$$y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J_i$$  \hspace{1cm} (7.2)

where $\theta_i = \gamma_0 + \gamma_1 z_{1,i} + \gamma_2 z_{2,i} + \delta_i$, $\delta_i \sim N(0, \tau^2)$, and $\varepsilon_{ij} \sim N(0, \sigma^2)$. The prior distributions are set as $p(\sigma^2) \propto 1$, $p(\gamma_h) \sim N(0, 10^6)$, $h = 0, 1, 2$, and $p(\tau^2) \sim IG(\frac{11}{10}, 10)$, where $IG(a, b)$ denotes an inverse gamma distribution with mean $b/(a - 1)$.

We re-generate Hodges’s results by running the Gibbs sampler for 10,000 iterations with starting values $(\sigma^{(0)}, \tau^{(0)}, \gamma_1^{(0)}, \gamma_2^{(0)}) = (490.2, 141.4, 6.7, 35.0)$. The starting value of $\theta_i$ is set as the average of state $i$'s premia and $\gamma_0$ starts from the average of the starting values of the $\theta_i$'s. Table 7.1 summarizes the posterior. Compared to the result of Hodges (1998), the regression coefficients $\gamma_h$ have similar summaries but the summary for $\sigma^2$ is quite different. As pointed out in Wakefield’s (1998) discussion, we conjecture that the shorter run of Hodges is not long enough to discover the bimodality which is uncovered by our moderate size run.
Table 7.1: Comparison of posterior summaries of the Gibbs samplers. Hodges (1998) ran 750 iterations with an additional 250 burn-in iterations. We re-run the Gibbs chain with 10,000 iterations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Our result with 10000 iter.</th>
<th>Table 1 of Hodges with 750 iter.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Post. Mean</td>
<td>Post. S.D.</td>
</tr>
<tr>
<td>$\gamma_0$</td>
<td>163.2</td>
<td>1.4</td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>4.0</td>
<td>1.4</td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>32.9</td>
<td>4.6</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>565.9</td>
<td>46.7</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>6.1</td>
<td>21.8</td>
</tr>
</tbody>
</table>

Figure 7.1 shows the existence of bimodality in the marginal posterior density in $\sigma^2$ and $\tau^2$. Although our longer run of the Gibbs sampler finds the local mode, the Gibbs sampler easily gets stuck in the local mode, showing poor mixing of the chain. The trace plots show the high dependence between parameters. Specifically, the low (or high) values of $\sigma^2$ correspond to the high (or low) values of $\tau^2$. As shown in the right bottom panel of the log-posterior density, the chain frequently gets stuck in the local mode (the region with the log-posterior around -1750), showing large variability of the regression coefficients $\gamma_h$ which may lead to overestimation of the variance of the parameters.
Figure 7.1: Trajectories of the Gibbs chain with the HMOs data set. The trajectories show the existence of two modes. The Gibbs chain rarely switches modes.
7.2.2 Joint Posterior Density and Marginal Posterior Densities of the Model

Let $\theta = (\theta_1, \ldots, \theta_I)$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2)$. The joint posterior density of $(\theta, \gamma, \sigma^2, \tau^2)$ in model (7.2) can be written as follows:

$$p(\gamma, \sigma^2, \tau^2, \{\theta_i\}|y) \propto (\sigma^2)^{-\frac{N}{2}} \exp\left[\frac{\sum_i \sum_j (y_{ij} - \theta_i)^2}{2\sigma^2}\right] (\tau^2)^{-\frac{l}{2}} \exp\left[\frac{\sum_i (\theta_i - \gamma z_i)^2}{2\tau^2}\right]$$

$$p(\gamma)p(\sigma^2)p(\tau^2)$$

(7.3)

where $N = \sum_{i=1}^I J_i$. Now, if we integrate out the group mean values $\theta$, then the marginal posterior of $(\gamma, \sigma^2, \tau^2)$ can be obtained as

$$p(\gamma, \sigma^2, \tau^2|y) \propto (\sigma^2)^{-\frac{N}{2}} (\tau^2)^{-\frac{l}{2}} \prod_{i=1}^I \left(\frac{J_i}{\sigma^2} + \frac{1}{\tau^2}\right)^{-\frac{1}{2}} p(\gamma)p(\sigma^2)p(\tau^2)$$

$$\exp\left(-\frac{1}{2} \sum_{i=1}^I \left[\frac{\sum_j y_{ij}^2}{\sigma^2} + \frac{(\gamma z_i)^2}{\tau^2} - \left(\frac{J_i}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \left(\frac{\sum_j y_{ij}^2}{\sigma^2} + \frac{\gamma z_i}{\tau^2}\right)\right] \right)$$

(7.4)

where $z_i = (1, z_{1,i}, z_{2,i})^T$. One can directly describe the marginal posterior of $(\gamma, \sigma^2, \tau^2)$ from the following model:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_I \end{pmatrix} \sim N \left[ Z\gamma, \Sigma \right], \quad Z = \begin{pmatrix} Z_1\gamma \\ Z_2\gamma \\ \vdots \\ Z_I\gamma \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_I \end{pmatrix}$$

(7.5)

where

$$y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iJ_i} \end{pmatrix}, \quad Z_i = 1z_i^T, \quad \Sigma_i = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 & \cdots & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 & \cdots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2 + \tau^2 \end{pmatrix}$$

and 1 is an $I \times 1$ vector of ones. The model formulation gives the joint posterior density,

$$p(\gamma, \sigma^2, \tau^2|y) \propto \left|\Sigma\right| \exp\left(-\frac{1}{2} (y - Z\gamma)'\Sigma^{-1}(y - Z\gamma)\right) p(\gamma)p(\sigma^2)p(\tau^2),$$

109
which is equivalent to the equation (7.4).

If we are interested in inference for \((\gamma, \sigma^2, \tau^2)\), we have two options. The first option is to run the MCMC sampler with the joint posterior of \((\theta, \gamma, \sigma^2, \tau^2)\) in (7.3) and look at the posterior summaries of \((\gamma, \sigma^2, \tau^2)\) only. The advantage of this approach is that the closed form of the conditional distributions enables us to use the Gibbs sampler. The second option is to run the MCMC sampler directly with the joint posterior \((\gamma, \sigma^2, \tau^2)\) in (7.4). With this approach, one no longer needs to update the group means \(\{\theta_1, \ldots, \theta_I\}\), so the Markov chain may travel more freely than with the joint posterior density in (7.3). The posterior marginal density of \(\theta\) can be produced by Rao-Blackwellisation (Casella 1996).

### 7.2.3 Comparisons of MCMC samplers

We compare the results of the Gibbs sampler, the random-walk MH, and the GMSS for the random effect model with the HMO data set. For the Gibbs sampler, we use the joint posterior of \((\theta, \gamma, \sigma^2, \tau^2)\) in (7.3). For the random-walk MH and the GMSS, we use the marginal posterior of \((\gamma, \sigma^2, \tau^2)\) in (7.4). For ease of updating, we take a log-transformation of \(\sigma^2\) and \(\tau^2\) for MH and the GMSS, and we use a normal random-walk proposal distribution with the step size of \((1.0, 1.0, 4.0, 0.2, 1.5)\) for \((\gamma_0, \gamma_1, \gamma_2, \log \sigma^2, \log \tau^2)\), respectively. The GMSS uses a multiset of size \(K = 2\) for the variance parameters, \(\log \sigma^2\) and \(\log \tau^2\). We use the equal multiset weights and the following uniform instrumental densities,

\[
f_{\sigma,k}(\log \sigma^2) \propto I \left[ \log \sigma^2 \in (5.9, 6.7) \right] \quad \text{and} \quad f_{\tau,k}(\log \tau^2) \propto I \left[ \log \tau^2 \in (-7, 7) \right], \quad k = 1, 2.
\]

110
Figure 7.2: Posterior densities from five replicates of the MH algorithm with $5 \times 10^7$ iterations. There is no visible difference between the five replicates. The average of the kernel density estimates is used as the empirical target density to compare the MCMC methods.
Figure 7.2 shows the posterior kernel density estimates of five independent chains of MH with $5 \times 10^7$ run after $10^6$ burn-in run. We use the average of the density estimates as the empirical target distributions to compare shorter runs from the three MCMC methods. Form the structure of the model, we might suspect the unimodality of the $\gamma_h$. However, the unimodality of $\log \sigma^2$ looks interesting since its values are closely tied to those of $\log \tau^2$ which shows a bimodal posterior density.

The trace plots in Figure 7.3 shed light on the question. From top to bottom panels, there are trajectories of the Gibbs sampler, the MH, and the GMSS. In addition to the better mixing of the GMSS chain, the plot shows that $\log \sigma^2$ has two components to its distribution. However, the centers of the components are very close and each component has large variability so the (combined) marginal density looks like a unimodal distribution as shown in Figure 7.2.

Table 7.2 is the summary of 100 independent chains with 2,000 iterations for each MCMC method. To adjust for the different run times of the MCMC methods, the summary of longer runs of Gibbs sampling and MH are provided. The point estimates of $\log \sigma^2$ are similar through all of the MCMCs. As explained above, $\log \sigma^2$ has two explanations, but they overlap resulting in a marginally unimodal posterior density. Therefore, finding the posterior mean is not a difficult task for all three MCMC methods. However, the posterior summaries of $\log \tau^2$ show a significant difference between the GMSS and the others, even after adjusting for the different computational burdens.

To check whether the GMSS consistently performs better than MH, we compare multiple chains with different step sizes. Figure 7.4 shows the average of TV over posterior summaries of 500 independent chains for each MCMC methods. Through
Figure 7.3: Trajectories for $\log \sigma^2$ and $\log \tau^2$ from the Gibbs sampler (top panels), the MH (middle panels), and the GMSS (bottom panels). Mixing improves as we move from up to bottom.
Table 7.2: Mean estimate and total variation of 100 independent chains from 2,000 iterations after a burn-in of 500 iterations under Gibbs, MH, and GMSS’s.

<table>
<thead>
<tr>
<th></th>
<th>( \theta )</th>
<th>Total variation</th>
<th>Comp. time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SD</td>
<td>MSE</td>
</tr>
<tr>
<td>Gibbs</td>
<td>( \log \sigma^2 )</td>
<td>6.311</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>( \log \tau^2 )</td>
<td>0.181</td>
<td>1.750</td>
</tr>
<tr>
<td>MH</td>
<td>( \log \sigma^2 )</td>
<td>6.317</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>( \log \tau^2 )</td>
<td>-0.192</td>
<td>0.482</td>
</tr>
<tr>
<td>GMSS</td>
<td>( \log \sigma^2 )</td>
<td>6.316</td>
<td>0.007</td>
</tr>
<tr>
<td></td>
<td>( \log \tau^2 )</td>
<td>-0.124</td>
<td>0.329</td>
</tr>
<tr>
<td>Gibbs w/</td>
<td>( \log \sigma^2 )</td>
<td>6.314</td>
<td>0.025</td>
</tr>
<tr>
<td>3,333 iter</td>
<td>( \log \tau^2 )</td>
<td>0.013</td>
<td>1.469</td>
</tr>
<tr>
<td>MH w/</td>
<td>( \log \sigma^2 )</td>
<td>6.316</td>
<td>0.009</td>
</tr>
<tr>
<td>2,600 iter</td>
<td>( \log \tau^2 )</td>
<td>-0.142</td>
<td>0.435</td>
</tr>
</tbody>
</table>
all of the step sizes, the GMSS outperforms MH with the same number of iterations as well as outperforms MH with more iterations. One interesting feature is that the gap between the GMSS and MH decreases as the overall level of the total variation decreases. This implies that the GMSS is more useful in the case where the appropriate step size is difficult to find.
Figure 7.4: Total variation distance between estimated and target distributions for the log $\tau^2$ under MCMC algorithms with different proposal scales. The dotted line is for MH with 2,000 iterations, the dashed line is for MH with 2,600 iterations, and the solid line is for the GMSS with 2,000 iterations.
7.3 Example 9: 2NRCAR model with Periodontal Data

In this section, we discuss the difficulty of estimating the posterior distribution of variance (or precision) parameters if the number of the variance parameters is more than two. As a practical example, the two neighbor relations conditionally autoregressive model (2NRCAR) is applied to periodontal data.

7.3.1 Attachment Loss Data in Disease Mapping

We used periodontal data, attachment loss measurements. Periodontal attachment loss (AL) is the amount of lost connective tissue fibers that attach a tooth’s root to the alveolar bone. AL is widely used as the reference of the severity of periodontal disease (refer to Darby and Walsh 2010; Reich and Bandyopadhyay 2010). It is usually measured at six sites per tooth as shown in Figure 7.5.

The analysis of periodontal AL is important but it is a difficult task due to several reasons: (i) measurement error is mainly due to inter-examiner variability rather than intra-examiner variability (Osborn et al. 1990), (ii) spatial correlation exists, due to such reasons as transmission of bacterial infection to nearby teeth and spatial heterogeneity (Cui et al. 2010), and (iii) existence of missing teeth with different missing patterns across patients (Reich and Bandyopadhyay 2010).

7.3.2 Conditionally Autoregressive Model with Neighbor Relations

Due to AL’s importance in disease mapping and the complexity of the relationship among measurement sites, there is a literature within statistics that develops effective
Figure 7.5: The spatial map of AL measurements when a patient has no missing teeth. The number in a box denotes teeth number. The number around a circle denotes AL measurement site.
monitoring processes and appropriate methods for statistical analysis. Within this literature, we focus on a series of papers which fit spatial or spatiotemporal models to the AL data (Hodges, Carlin and Fan 2003; Reich, Hodges and Carlin 2007; Reich and Hodges 2008).

Hodges et al. (2003) used a conditionally autoregressive (CAR) prior for AL data to smooth fitted values toward neighboring values. Let $y_i$ be the measured AL at site $i$, $i=1,\ldots,I$. Assume that $y_i|\theta_i \sim N(\theta_i, \frac{1}{\sigma})$ where $\theta_i$ is the true AL and $\sigma$ is the precision parameter for measurement noise. To explain the spatial correlation between measurement sites, one adopts a CAR prior with the $L_2$ norm as a distance measure,

$$p(\theta|\tau) \propto \tau^n e^{-\frac{1}{2}\theta^T \tau_1 Q_1 \theta},$$

where $\theta = (\theta_1, \ldots, \theta_I)$ and $Q$ is the neighbor-relations matrix with nondiagonal entries $q_{i,j} = 1$ if the site $i$ and $j$ are neighbors and 0 otherwise, and diagonal entries $q_{i,i}$ equal to the number of neighbors of site $i$.

Reich et al. (2007) extends the CAR model to explain the different classes of neighbor relations among the measurement sites. Their two neighbor relations CAR (2NRCAR) model is

$$y \sim MVN(\theta, \frac{1}{\sigma}I)$$

$$P(\theta|\tau_1, \tau_2) \propto c(\tau_1, \tau_2)^{1/2} e^{-\frac{1}{2}\theta^T (\tau_1 Q_1 + \tau_2 Q_2) \theta}$$ (7.6)

where $y = (y_1, \ldots, y_I)$. $c(\tau_1, \tau_2)$ is the product of the positive eigenvalues of $\tau_1 Q_1 + \tau_2 Q_2$, i.e., $c(\tau_1, \tau_2) = \prod_{j=1}^{G} (\tau_1 d_{1,j} + \tau_2 d_{2,j})$ where $d_{l,j}$ is the $j$th positive eigenvalue of $Q_l$ and $G$ is the number of islands in the spatial map. The matrix $Q_l$, $l = 1, 2$, has nondiagonal entries $q_{l(i,j)} = 1$ if the sites $i$ and $j$ are type $l$ neighbors and 0 otherwise,
Figure 7.6: Two types of relations. The solid line is the Type I neighbor relation and the dashed line is the Type II neighbor relation.
and diagonal entries \( q_{l(i,i)} \) equal to the number of the type \( l \) neighbors of site \( i \). For our example as shown in Figure 7.6, \( q_{1(3,4)} = 1, q_{1(3,45)} = 0, q_{2(3,2)} = 0, \) and \( q_{2(3,46)} = 1 \).

The posterior density of the 2NRCAR model frequently shows an irregular pattern which cannot be easily explored by ordinary MCMC samplers. To solve this issue, He et al. (2007) compared several parameterizations for variance components: (i) \((\sigma^{-\frac{1}{2}}, \tau_1^{-\frac{1}{2}}, \tau_2^{-\frac{1}{2}})\), (ii) \((\sigma, \frac{\tau_1}{\sigma}, \frac{\tau_2}{\sigma})\), and (iii) \((\log \frac{\tau_1}{\sigma}, \log \frac{\tau_2}{\sigma})\) after integrating out \( \sigma \). They showed that the previously proposed parameterizations do not solve the problem. Instead, they proposed the “simplex” parameterization, in the context of our example \((\frac{\tau_1+\tau_2}{\sigma}, \frac{\tau_1}{\tau_1+\tau_2})\), and claimed that the new parameterization leads to better mixing of Markov chains.

In this section, we focus on the log precision ratios parameterization, \((z_1, z_2) = (\log \frac{\tau_1}{\sigma}, \log \frac{\tau_2}{\sigma})\). We show that the GMSS can estimate the posterior distribution in spite of its irregular pattern.

### 7.3.3 Application of the GMSS to a Posterior Density with an Irregular Pattern

The posterior density of the original parameters in (7.6) is

\[
p(\theta, \sigma, \tau_1, \tau_2 | y) \propto \sigma^{\frac{1}{2}} \exp \left[ -\sigma \sum_{i=1}^{I} (y_i - \theta_i)^2 \right] \prod_{j=1}^{I-G} \left( \tau_1 d_{1,j} + \tau_2 d_{2,j} \right)^{\frac{1}{2}} \exp \left[ -\frac{1}{2} \theta^T (\tau_1 Q_1 + \tau_2 Q_2) \theta \right] p(\sigma)p(\tau_1)p(\tau_2) \tag{7.7}
\]
where \( p(\cdot) \) is the prior distribution of the precision parameters. After integrating out \( \theta \),

\[
p(\sigma, \tau_1, \tau_2 | \mathbf{y}) \propto \sigma^{\frac{I-G}{2}} \prod_{j=1}^{I-G} (\tau_1 d_{1,j} + \tau_2 d_{2,j})^{\frac{1}{2}} |\sigma I + \tau_1 Q_1 + \tau_2 Q_2|^{-\frac{1}{2}} \exp \left[ -\frac{\sigma}{2} (\mathbf{y}^T \mathbf{y} - \sigma \mathbf{y}^T (\sigma I + \tau_1 Q_1 + \tau_2 Q_2)^{-1} \mathbf{y}) \right] p(\sigma) p(\tau_1) p(\tau_2).
\]

Let \((\sigma, \nu_1, \nu_2) = (\sigma, \frac{\tau_1}{\sigma}, \frac{\tau_2}{\sigma})\). Then the posterior density of the transformed parameters is

\[
p(\sigma, \nu_1, \nu_2 | \mathbf{y}) \propto \sigma^{\frac{I-G}{2}} \prod_{j=1}^{I-G} (\nu_1 d_{1,j} + \nu_2 d_{2,j})^{\frac{1}{2}} |I + \nu_1 Q_1 + \nu_2 Q_2|^{-\frac{1}{2}} \exp \left[ -\frac{\sigma}{2} (\mathbf{y}^T \mathbf{y} - \mathbf{y}^T (I + \nu_1 Q_1 + \nu_2 Q_2)^{-1} \mathbf{y}) \right] p(\sigma) p(\nu_1) p(\nu_2).
\]

Now assume \( p(\sigma) \sim \Gamma(a_0, b_0) \), where \( \Gamma(a, b) \) denotes a gamma distribution with a mean \( = ab \). By integrating out \( \sigma \) and using the log precision ratio parameterization \((z_1, z_2) = (\log \nu_1, \log \nu_2)\), the marginal posterior density is

\[
p(z_1, z_2 | \mathbf{y}) \propto \prod_{j=1}^{I-G} (e^{z_1} d_{1,j} + e^{z_2} d_{2,j})^{\frac{1}{2}} |I + \lambda Q_z|^{-\frac{1}{2}} R_z^{\frac{I-G - a_0}{2}} p(z_1) p(z_2)
\]

\[
(7.8)
\]

where \( Q_z = e^{z_1} Q_1 + e^{z_2} Q_2, R_z = b_0 + \frac{1}{2} [\mathbf{y}^T \mathbf{y} - \mathbf{y}^T (I + Q_z)^{-1} \mathbf{y}] \).

We fit the 2NRCAR model with the log precision ratio parameterization in (7.8) to the AL data from Reich and Hodges (2008). Figure 7.7 shows the contour plot of the marginal posterior density \( p(z_1, z_2 | \mathbf{y}) \). There is a long ridge in the posterior which is adjacent to a steep cliff. In spite of the effort of the re-parameterization, the posterior is still difficult to handle with ordinary MH methods. To solve the issue, He et al. (2007) used an adaptive MH method or a slice sampler, depending on which parameterization was used.
We use this example to show another possible use of the GMSS. While the previous chapters focus on examples with a multimodal density, this example shows a different sort of case where an ordinary MH method has trouble finding the modal region of the posterior.

Let us assume conditional independence in the prior and give flat prior distributions to $z_l$, $p(z_l) \sim Unif[-15, 15]$ for $l = 1, 2$. For MH and the GMSS, we use normal random walk proposals, $z'_1 \sim N(z_1, 0.2^2)$ and $z'_2 \sim N(z_2, 0.2^2)$. For GMSS, a multiset of size $K = 3$ with the equal multiset weights is used. Uniform instrumental densities are used as follows: $f_k(z_{1,k}, z_{2,k}) \propto I[z_{1,k} \in (1.5, 4.5)] \times I[z_{2,k} \in (-9, 1)]$.

Figure 7.8 shows the trajectory of a MH chain with the starting points, $(z_1^{(0)}, z_2^{(0)}) = (3.5, -5.0)$. From the bottom panel, the MH sample often shows a sharp drop toward the region with lower log posterior density, and it can stay in this region for many
iterates. We cannot find any evidence of the existence of multimodality form the contour plot in Figure 7.7. Therefore, we conjecture that the poor behavior of the MH chain arises from the irregular shape of the posterior density. It is difficult to move along a long ridge adjacent to a steep cliff.

Figure 7.9 is the trajectory of the GMSS with the same starting points used above. From the bottom panel, it seems that there is no local mode which is consistent with the contour plot in Figure 7.7. As a result of good mixing, the GMSS run shows smaller sample variances of $z_l$ than does the MH run.
Figure 7.8: Trajectory of the MH.
Figure 7.9: Trajectory of the GMSS of size $K = 3$. 
8.1 Discussion

The multimodality of a posterior density is very frequently observed in the various settings of Bayesian analysis. Liu (2001) had a question about the frequency of observing the multimodality in practice, and investigated a number of data sets from Dr. Jim Hodge’s statistical practice in the University of Minnesota School of Dentistry. They analyzed 67 periodontal data sets with balanced one-way random effect models, and found that 47 data sets (70.1 %) meet their condition to yield a bimodal posterior density. When one local mode is located near the other, estimation of the posterior density, which is nearly a unimodal density, is not a difficult task for ordinary MCMC methods such as Gibbs sampling or the Metropolis-Hastings algorithm. However, when the modes are far enough from one another each other, the ordinary MCMC methods easily gets stuck in a local mode.

This dissertation generalizes the idea of the multiset sampler of Leman et al. (2009) to encompass a wide variety of existing algorithms, suitable for different examples. We explicitly describe the relationship between the target distribution and the multiset sampling distribution by introducing multiset weights and instrumental densities. A clean version of an estimator relying on importance sampling is shown to be successful, and the properties and applicability of the GMSS are demonstrated.
From the analyses of empirical examples including gene-expression analysis, outlier
detection, and two variance component models, we show that our proposed method is
an easy-to-implement MCMC algorithm, which is especially useful for drawing sam-
pies from a distribution with multiple modes or having an irregular shape.

8.2 Future Work

Finally, we note that the GMSS is a static MCMC method. Our comparisons
have been to other static methods, notably the Gibbs sampler, the MH algorithm
and the MSS. The explicit description of the GMSS as a static MH method allows
us to modify the algorithms presented here in many ways. A particularly interesting
direction is to use the GMSS as the static core of an adaptive method, where a
variety of features can be adapted – the proposal distributions, the weight vector, the
instrumental densities, and even the size of the multiset. We suspect that combining
a strong static core along with adaptation will provide even greater benefits.

Among many application areas, we expect to find a set of Bayesian marketing
examples implemented by the GMSS. As seen in the simultaneous equations examples
in Chapter 5, marketing models commonly assume many latent variables to explain
the main motivations of humans and their reaction to the market. The analysis
of customer behavior often requires a complex model formulation involving many
endogenous variables. Therefore, we expect that there exists a class of marketing
problems which is complicated enough that the ordinary MCMC approaches fail to
estimate the accurate posterior density. The GMSS may provide us with a robust
and reliable tool to effectively fit these models.
Appendix A: Integrated Autocorrelation Time (IACT)

Let \( \{X_n\} \) be a Markov chain with a stationary distribution \( \pi \). Let \( h \) be a real-valued function defined on the state space, then \( \{h(X_n)\} \) is a stationary Markov process with a mean \( E[h(X_n)] \) and a normalized autocorrelation function

\[
\rho(t) = \frac{Cov[h(X_s), h(X_{s+t})]}{Var[h(X_s)]}.
\]

Sokal (1989) defined the integrated autocorrelation time (IACT),

\[
\tau = \frac{1}{2} \sum_{t=-\infty}^{\infty} \rho(t) = \frac{1}{2} + \sum_{t=1}^{\infty} \rho(t).
\]

Given \( N \) samples of a stationary Markov chain \( \{h(x_1), \ldots, h(x_N)\} \), the sample mean is \( \bar{h} = \frac{1}{N} \sum_{n=1}^{N} h(x_n) \) and the variance of the sample mean is

\[
Var[\bar{h}] \approx \frac{2\tau Var[h(X_n)]}{N} \text{ for } N \gg \tau.
\]

If the samples of a Markov chain \( \{h(Y_n)\} \) are independent, the variance of its sample mean is \( Var[h(Y_n)]/N \). So we can interpret the factor, \( 2\tau \), as the cost of using dependent samples compared to independent samples. In other words, the effective sample size is roughly \( N/(2\tau) \).

To reduce noise present in the sample autocorrelation estimator \( \hat{\rho}(t) \), Sokal proposed the use of a “window” \( \lambda(t) \) in the estimator for IACT

\[
\hat{\tau} = \frac{1}{2} \sum_{t=-(N-1)}^{N-1} \lambda(t) \hat{\rho}(t)
\]
where $\lambda(t) = 1$ if $|t| \leq M$ and 0 if $|t| > M$ where $M$ is a suitably chosen cutoff. Then the IACT estimator has a bias, $E[\hat{\tau} - \tau] = -\frac{1}{2} \sum_{|t|>M} \rho(t) + o(1/N)$, and a variance, $Var(\hat{\tau}) \approx \tau^2 (4M + 2)/N$. We can see the tradeoff between bias and variance as taking a large $M$ makes the variance large, but the bias small.

Sokal suggested an “automatic windowing” algorithm which chooses $M$ to be the smallest number satisfying $M \geq c \hat{\tau}$ with the prespecified value $c$ around $4 \sim 10$. Although they empirically showed the automatic windowing procedure working with large $N$, its theoretical properties remain to be studied.
Bibliography


