Efficient Algorithms for Fitting Bayesian Mixture Models

DISSERTATION

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

By

Xiuyun Zhang, M.S.

* * * * *
The Ohio State University

2009

Dissertation Committee:

Mario Peruggia, Adviser
Steven N. MacEachern
Chris Hans

Approved by

[Signature]
Adviser
Graduate Program in
Statistics
Mixture distributions have been given considerable attention due to their flexible form and convenience of use. Markov Chain Monte Carlo (MCMC) methods enable us to generate samples from a target distribution from which it is difficult to sample directly by simulating a Markov chain. However, practical difficulties arise when MCMC methods are implemented to fit mixture distributions with several isolated modes. Most MCMC sampling methods have difficulties transitioning between the isolated modal regions and the inferences based on the samples generated by these methods can be unreliable. This motivated us to develop efficient algorithms for fitting Bayesian mixture models. Our approach hinges on the premise that a preliminary understanding of some essential features of the posterior distribution is needed to make sampling more efficient.

In this thesis we introduce two algorithms that rely on an initial identification of possible isolated modes of the mixture distribution. The algorithms are applied to fit four different models: a Bayesian univariate normal mixture model; a Bayesian univariate outlier accommodation model; a Bayesian linear regression model; and a hierarchical Bayesian regression model for repeated measures data. Their performance is compared to that of other methods including the Gibbs sampler and an MCMC tempering transition method by examining the accuracy of inferences and the ease of transition between isolated modal regions of the posterior distributions for the
Bayesian models. The results show that the proposed algorithms outperform the Gibbs sampler and the tempering transition method.
This work is dedicated to my parents, Longshan and Genxian, my husband, Mansen, and my son, Leonard.
ACKNOWLEDGMENTS

My deepest gratitude goes to my advisor, Professor Mario Peruggia. He gave me invaluable guidance over the years and encouraged me when I had moments of crisis with my research. Without his guidance and consistent help this work could not have been completed. I am very appreciative of Professor Steven MacEachern for serving on my general examination and dissertation committee and acting as a reference on my behalf. I am very grateful to Professor Chris Hans for giving me the opportunity to learn from him, serving on my general examination and dissertation committee and acting as a reference on my behalf. I would like to thank Professor Peter Craigmile for serving on my general examination committee. I thank all of my professors, fellow students, and department staff for their kind support.

I would like to express my sincere appreciation to Professor Giovanni Petris for his invaluable advice and consistent encouragement and acting as a reference on my behalf. I also thank Professor Edward Gbur for carefully reading one chapter of my dissertation manuscript, giving very useful comments and acting as a reference on my behalf.

This work was supported in part by NSF Awards No. SES-0214574, SES-0437251 and DMS-0605052.

Finally, I thank my husband, my parents and my sister for their love and consistent support.
VITA

1998 ........................................ B.S. in Chemical Engineering, Shandong University of Technology
1998-2000 ................................. Technician, Hangzhou Oil-Chemical Plant
2000-2001 ................................. Graduate Student, Zhejiang University
2004 ........................................... M.S. in Statistics, University of Arkansas
2004-present ............................. Graduate Teaching and Research Associate, The Ohio State University

FIELDS OF STUDY

Major Field: Statistics
TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapters:</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Maximum Likelihood Estimation Method via the EM Algorithm</td>
<td>3</td>
</tr>
<tr>
<td>1.2 Bayesian Estimation</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Markov Chain Monte Carlo Methods</td>
<td>8</td>
</tr>
<tr>
<td>1.4 Issues with Sampling from a Posterior Distribution with Multimodality</td>
<td>11</td>
</tr>
<tr>
<td>1.5 Trans-Dimensional Markov Chain Monte Carlo Methods</td>
<td>14</td>
</tr>
<tr>
<td>1.6 Dirichlet Process Mixtures</td>
<td>16</td>
</tr>
<tr>
<td>1.7 Organization of This Dissertation</td>
<td>17</td>
</tr>
<tr>
<td>2. Standard MCMC Methods for Fitting Mixture Models</td>
<td>19</td>
</tr>
<tr>
<td>2.1 Metropolis-Hastings Algorithm</td>
<td>20</td>
</tr>
<tr>
<td>2.2 Gibbs Sampler</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Tempering MCMC with an Up-and-Down Power Scheme</td>
<td>26</td>
</tr>
</tbody>
</table>

Abstract . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . ii
Dedication . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . iv
Acknowledgments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . v
Vita . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . vi
List of Tables . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . x
List of Figures . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . xii
3. New MCMC Methods for Fitting Mixture Models
   3.1 Algorithm Based on a Mixture of Modal Proposals
   3.2 Hybrid Metropolis-Hastings-Gibbs Algorithm with Clustering Initialization
     3.2.1 The Combination of Two Proposal Kernels
     3.2.2 Building the Metropolis-Hastings Kernel of Step (2)
     3.2.3 Acceptance Probability
   3.3 Application to a Finite Univariate Normal Mixture Model
     4.1 A Finite Univariate Normal Mixture Model
     4.2 Full Conditional Distributions
     4.3 Posterior Distributions of $\mu$, $\sigma^2$ and $w$
     4.4 The Marginal Posterior Distribution of the Allocation Variables
     4.5 Example
   4. Application to an Outlier Accommodation Model
     5.1 An Outlier Accommodation Model
     5.2 Full Conditional Distributions
     5.3 Conditional Distributions of $\mu$ and Reparameterized $\sigma^2$
     5.4 The Marginal Posterior Distribution of the Allocation Variables
     5.5 Example
   5. Application to an Outlier Accommodation Bayesian Regression Model
     6.1 An Outlier Accommodation Bayesian Regression Model
     6.2 Full Conditional Distributions
     6.3 Conditional Distributions of $\beta$ and Reparameterized $\sigma^2$
     6.4 The Marginal Posterior Distribution of the Allocation Variables
     6.5 Example
   6. Application to an Outlier Accommodation Hierarchical Bayesian Linear Model of Repeated Measures Data
     7.1 An Outlier Accommodation Hierarchical Bayesian Linear Model of Repeated Measures Data
     7.2 Full Conditional Distributions
     7.3 Conditional Distributions of the Parameters without the Stage I and II Outlier Indicators $\delta_i^y$ and $\delta_i^\beta$
     7.4 Example
7.4.1 Case 1 ......................................................... 106
7.4.2 Case 2 ......................................................... 112

8. Discussion ......................................................... 116
   8.1 Conclusions .................................................. 116
      8.1.1 Choice of the Tuning Parameters ....................... 117
   8.2 Future Work ................................................ 122

Bibliography ...................................................... 125
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Exact posterior probabilities of $Z_1$, $Z_2$, $Z_3$ and $Z_4$</td>
<td>52</td>
</tr>
<tr>
<td>4.2 RMSE’s for the marginal posterior estimates on the configurations $Z_1 = {(1, \ldots, 5), (6, \ldots, 10)}$, $Z_2 = {(1, \ldots, 4), (5, \ldots, 10)}$, $Z_3 = {(1, \ldots, 6), (7, \ldots, 10)}$, $Z_4 = {1, \ldots, 10}$ based on 100 batches with $r_0 = 0.99$, each batch with length 10,000 iterations.</td>
<td>61</td>
</tr>
<tr>
<td>4.3 RMSE’s for the marginal posterior estimates on the configurations $Z_1 = {(1, \ldots, 5), (6, \ldots, 10)}$, $Z_2 = {(1, \ldots, 4), (5, \ldots, 10)}$, $Z_3 = {(1, \ldots, 6), (7, \ldots, 10)}$, $Z_4 = {1, \ldots, 10}$ based on 100 batches with $r_0 = 0.95$, each batch with length 10,000 iterations.</td>
<td>62</td>
</tr>
<tr>
<td>5.1 True posterior outlier probabilities $p_i$, the estimates using the Gibbs sampler $\hat{p}_i^G$, the estimates using the MCMC tempering with two starting points $\hat{p}_i^{T_1}$ and $\hat{p}_i^{T_2}$, the estimates using our proposed algorithms $\hat{p}_i^M$ and $\hat{p}_i^H$. The estimates are based on 50,000 iterations with first 10,000 iterations as burn-in.</td>
<td>77</td>
</tr>
<tr>
<td>5.2 Geometric means of ratios of true outlier posterior probabilities $p_i$ to the estimates $\hat{p}_i$ by each method for outlying cases and non-outlying cases</td>
<td>78</td>
</tr>
<tr>
<td>6.1 True posterior outlier probabilities, $p_i$, estimated by: the Gibbs sampler, $\hat{p}_i^G$; MCMC tempering with two starting points, $\hat{p}_i^{T_1}$ and $\hat{p}_i^{T_2}$; our proposed mixture-based algorithm, $\hat{p}_i^M$; and our proposed hybrid algorithm, $\hat{p}_i^H$. Standard errors are given in parentheses. The estimates are based on 90,000 iterations after a burn-in of size 10,000.</td>
<td>94</td>
</tr>
<tr>
<td>6.2 RMSEs and geometric means of ratios of true outlier posterior probabilities $p_i$ to the estimates $\hat{p}_i$ by each method for outlying cases and non-outlying cases</td>
<td>95</td>
</tr>
</tbody>
</table>
6.3 The marginal posterior means of the model parameters and their estimates (with RMSEs) ................................................. 95

7.1 Estimates of $\hat{p}_i^{\beta}$: $\hat{p}_i^{\beta,G}$ by the Gibbs sampler, $\hat{p}_i^{\beta,T}$ by the tempering transition method, and $\hat{p}_i^{\beta,M}$ by the proposed mixture-based algorithm, $\hat{p}_i^{\beta,H_a}$, $\hat{p}_i^{\beta,H_b}$ and $\hat{p}_i^{\beta,H_c}$ by the proposed hybrid algorithms. Standard errors are given in parentheses. The estimates are based on 45,000 iterations after a burn-in of size 5,000 iterations. .............................. 115

8.1 The values of $\varphi_l$ for the example in Chapter 5. .......... 119

8.2 The values of $\varphi_l$ for the example in Chapter 6. .......... 120

8.3 The values of $\varphi_l$ for the example in Section 7.4.2. .......... 121
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Scatter plot of the data.</td>
</tr>
<tr>
<td>4.2</td>
<td>Trace plots of the sampled parameters by the Gibbs sampler. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).</td>
</tr>
<tr>
<td>4.3</td>
<td>Trace plots of the sampled parameters by the Gibbs sampler for iterations between 882,800 and 883,000. (a) and (b): the output of $\mu$; (c) and (d): the output of $\sigma^2$.</td>
</tr>
<tr>
<td>4.4</td>
<td>Trace plots of the sampled parameters by the tempering transition method. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).</td>
</tr>
<tr>
<td>4.5</td>
<td>Trace plots of the sampled parameters by the proposed algorithm in Section 3.1. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).</td>
</tr>
<tr>
<td>4.6</td>
<td>Trace plots of the sampled parameters by the proposed algorithm in Section 3.2. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).</td>
</tr>
<tr>
<td>5.1</td>
<td>Scatter plot of the data.</td>
</tr>
<tr>
<td>5.2</td>
<td>Trace plots of the sampled parameters by the Gibbs sampler. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth).</td>
</tr>
</tbody>
</table>
5.3 Trace plots of the sampled parameters by the tempering transition method. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth). 73

5.4 Trace plots of the sampled parameters by the proposed algorithm in Chapter 3.1. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth). 74

5.5 Trace plots of the sampled parameters by the proposed algorithm in Chapter 3.2. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth). 76

6.1 Scatter plot of the data. 87

6.2 Trace plots of the sampled parameters by the Gibbs sampler for the last 2,000 iterations. (a): the output of $\beta_0$; (b): the output of $\beta_1$; and (c): the output of $\sigma^2$. 87

6.3 Trace plots of the sampled parameters by the tempering transition method for the last 2,000 iterations. (a) and (d): the output of $\beta_0$; (b) and (e): the output of $\beta_1$; and (c) and (f): the output of $\sigma^2$. 89

6.4 Trace plots of the sampled parameters by the proposed algorithm in Section 3.1 for the last 2,000 iterations. (a): the output of $\beta_0$; (b) the output of $\beta_1$; and (d): the output of $\sigma^2$. 90

6.5 Trace plots of the sampled parameters by the proposed algorithm in Section 3.2 for the last 2,000 iterations. (a): the output of $\beta_0$; (b) the output of $\beta_1$; and (d): the output of $\sigma^2$. 92

7.1 Scatter plot of the data. 107

7.2 (a): Scatter plot of the transformed data; (b): Least-squares lines. 108
7.3 Trace plots of the samples for $\gamma_2$ by each method based on the output for the last 10,000 iterations (every fifth). (a) the Gibb sampler; (b) the MCMC tempering transition method; (c) the proposed mixture-based algorithm; (d) the proposed hybrid algorithm. 111

7.4 Trace plots of the samples for $\gamma_2$ by each method based on the output for the last 10,000 iterations (every fifth). (a) the Gibb sampler; (b) the MCMC tempering transition method; (c) the proposed mixture-based algorithm; (d) the proposed hybrid algorithm. 114
CHAPTER 1

INTRODUCTION

For a heterogeneous data set with homogeneous subgroups, if the group membership of each observation is unknown, it is appropriate to specify a model based on a mixture distribution of the form

\[ p(y|\theta, w) = \sum_{k=1}^{K} w_k p(y|\theta_k), \]  

where \( K \) is the (possibly unknown) number of mixture components, \( \theta = (\theta_1, \ldots, \theta_K) \) is a vector of component specific parameters, and \( w = (w_1, \ldots, w_K) \) is a vector of non-negative weights with \( \sum_{k=1}^{K} w_k = 1 \). McLachlan and Peel (2001) and Frühwirth-Schnatter (2006) provide detailed illustrations of mixture modeling. Mixture distributions have been given considerable attention due to their flexible form and convenience of use. For example, in the dairy field, somatic cell score (SCS) is an important trait used in a genetic selection program. An appropriate genetic SCS model should be built to predict SCS. Considering that it is usually unknown whether cows are intramammary infection absent (IMI-) or intramammary infection present (IMI+), Detilleux and Leory (2000) use a two-component normal mixture model to estimate genetic parameters for SCS. One component represents SCS data from IMI- cows and another component represents SCS data from IMI+ cows. A similar situation
happens in nuclear testing (Wang et al., 1997). A treaty about a worldwide ban on nuclear testing was approved in 1996. It is known that underground nuclear tests produce seismic signals. A training sample of previous seismic activity in the area is used to determine whether a new observed seismic event is a nuclear explosion or not. Since no historical data on nuclear explosions is available in most areas of the world, Wang et al. (1997) assume that the previous seismic activities are non-nuclear (earthquakes or mining explosions). And they use hypothesis testing to determine whether a new event belongs to the distribution of non-nuclear data or appears to be an outlier. Since for some observations in the training sample it is unknown whether they are earthquakes or mining explosions, the authors use a two-component mixture model to handle these unlabeled observations.

Assume that in the mixture distribution $p(y|\theta, w)$ the true number $K$ of components is known and the vector of component parameters $\theta$ and the vector of weights $w$ are unknown. The $n$ observations $y_1, \ldots, y_n$ are assumed to be independently drawn from the mixture distribution $p(y|\theta, w)$. Let $y = (y_1, \ldots, y_n)$. To estimate $\theta$ and $w$ in the mixture distribution, an estimation technique must be chosen. In the frequentist approach, maximum likelihood estimation is the most commonly used method for mixture models. In recent years, the Bayesian approach has received more and more attention due to the emergence of Markov Chain Monte Carlo (MCMC) methods. In the following sections I briefly review the maximum likelihood estimation method, introduce Bayesian estimation, and illustrate the idea of MCMC simulation. I then elaborate on the sampling issues for a posterior distribution with multimodality. Lastly, I review sampling algorithms when the true number of components is unknown and describe the organization of this dissertation.
1.1 Maximum Likelihood Estimation Method via the EM Algorithm

With the invention of powerful computers and the development of numerical methods during the 1960’s, the method of maximum likelihood estimation became the preferred approach to estimate the parameters for finite mixture models. Redner and Walker (1984) provide a detailed review of maximum likelihood (ML) estimation for finite mixture distributions. In an early reference, Rao (1948) uses the ML estimation method to estimate the parameters for a univariate normal mixture distribution with two components with common variance. The method of maximum likelihood estimates the parameters $\theta$ and $w$ by maximizing the likelihood $p(y, S|\theta, w)$ or equivalently, by maximizing the log likelihood $\log p(y, S|\theta, w)$. For some simple cases, the maximization equations have an analytically tractable form. However, for complex distributions such as mixture distributions, they will not have a closed form solution. Fortunately, this problem can be dealt with using numerical optimization methods. Among numerical optimization methods, the most commonly used method is the EM algorithm, which was introduced by Dempster et al. (1977) to obtain the ML estimator for a finite mixture distribution. Frühwirth-Schnatter (2006) gives the details on ML estimation via the EM algorithm.

Let $s_i$ indicate the group membership of the observation $y_i$, treated as missing data and let $S = (s_1, \ldots, s_n)$. Precisely, $s_i = k$ indicates that $y_i$ belongs to the $k^{th}$ component of the mixture. The complete-data likelihood function is written as

$$p(y, S|\theta, w) = \prod_{i=1}^{n} \prod_{k=1}^{K} (w_k p(y_i|\theta_k))^{{s_ik}},$$

3
where $e_{ik}$ is an indicator of $s_i$; $e_{ik} = 1$ if $s_i = k$; otherwise, $e_{ik} = 0$. Then the logarithm of the complete-data likelihood function is

$$\log p(y, S|\theta, w) = \sum_{i=1}^{N} \sum_{k=1}^{K} e_{ik} \log(w_k p(y_i|\theta_k)).$$

The EM algorithm starts from an arbitrary point $(\hat{\theta}^{(0)}, \hat{w}^{(0)})$, then alternates between two steps: the E-step for expectation and the M-step for maximization. In the E-step, the conditional expectation of $\log p(y, S|\theta, w)$ is computed given the data and the current parameters $(\hat{\theta}^{(t)}, \hat{w}^{(t)})$. In the M-step, the expectation obtained from the E-step is maximized with respect to $(\theta, w)$ to yield the new estimate $(\hat{\theta}^{(t+1)}, \hat{w}^{(t+1)})$.

Given the data and the current parameters $(\hat{\theta}^{(t)}, \hat{w}^{(t)})$, the conditional expectation of $e_{ik}$ is

$$E(e_{ik}|y, \hat{\theta}^{(t)}) = P(e_{ik} = 1|y, \hat{\theta}^{(t)}) \equiv \hat{e}_{ik}^{(t+1)}.$$

Then the estimate of $e_{ik}$ is

$$\hat{e}_{ik}^{(t+1)} = \frac{\hat{w}_k^{(t)} p(y_i|\hat{\theta}_k^{(t)})}{\sum_{j=1}^{K} \hat{w}_j^{(t)} p(y_i|\hat{\theta}_j^{(t)})}.$$

Substituting $\hat{e}_{ik}^{(t+1)}$ into the log of the complete-data likelihood function, we obtain

$$\sum_{i=1}^{N} \sum_{k=1}^{K} \hat{e}_{ik}^{(t+1)} \log(w_k p(y_i|\theta_k)).$$

Maximizing the above function in the M-step with respect to $\theta$ and $w$ yields new estimates $(\hat{\theta}^{(t+1)}, \hat{w}^{(t+1)})$. 


For the important special case of univariate normal mixture distributions with
\( \theta_k = (\mu_k, \sigma_k^2) \), the M-step yields

\[
\hat{w}_k^{(t+1)} = \frac{n_k}{n}, \\
\hat{\mu}_k^{(t+1)} = \frac{1}{n_k} \sum_{i=1}^{n} \hat{e}_{ik}^{(t+1)} y_i, \\
(\hat{\sigma}_k^2)^{(t+1)} = \frac{1}{n_k} \sum_{i=1}^{n} \hat{e}_{ik}^{(t+1)} (y_i - \hat{\mu}_k^{(t+1)})^2,
\]

where \( n_k = \sum_{i=1}^{n} \hat{e}_{ik}^{(t+1)} \).

The maximum likelihood estimator is consistent and invariant to transformation. Under certain regularity conditions, the ML estimator is asymptotically normal

\[
\sqrt{n}(\hat{\vartheta}_{ML} - \vartheta^{true}) \xrightarrow{d} N(0, I(\vartheta^{true})^{-1}),
\]

where \( \hat{\vartheta}_{ML} \) is the ML estimator of \( \vartheta = (\theta, w) \), \( \vartheta^{true} \) is the true value of \( \vartheta \), and \( I(\vartheta^{true}) \) is the expected Fisher information matrix evaluated at \( \vartheta^{true} \). However, as mentioned in Frühwirth-Schnatter (2006), in practice it may be hard for the numerical methods to locate the global maximum of the likelihood function. ML estimation via EM algorithm may be sensitive to the choice of starting values. In addition, in some instances the likelihood function is unbounded and the ML estimator does not exist.

For example, for a K-component univariate normal mixture model
\[ p(y|\mu, \sigma^2, w) = \sum_{k=1}^{K} w_k f_N(y|\mu_k, \sigma_k^2), \]
where \( \mu = (\mu_1, \ldots, \mu_K) \) and \( \sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2) \), the mixture likelihood function
\[ p(y|\mu, \sigma^2, w) = \prod_{i=1}^{n} \sum_{k=1}^{K} w_k f_N(y|\mu_k, \sigma_k^2) \]
is unbounded.
1.2 Bayesian Estimation

In the mixture distribution \( p(y|\theta, w) \) the unknown parameters \( \theta \) and \( w \) need to be estimated. In the Bayesian approach, the data-dependent likelihood function of the mixture distribution and the prior distribution of \( (\theta, w) \) are combined to obtain the posterior distribution, which is regarded as containing all information in the data about the unknown parameters \( \theta \) and \( w \). Using Bayes theorem, the posterior distribution \( p(\theta, w|y) \) of \( \theta \) and \( w \) satisfies

\[
p(\theta, w|y) \propto p(y|\theta, w)p(\theta, w),
\]

where \( p(y|\theta, w) = \prod_{i=1}^{n} p(y_i|\theta, w) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} w_k p(y_i|\theta_k) \right\} \) is the likelihood, and \( p(\theta, w) \) is the prior distribution of \( \theta \) and \( w \).

Alternatively, one can treat the allocation \( S \) of the observations as missing. An observation \( y_i \) is independently allocated to the \( k^{th} \) component with probability \( w_k \); i.e., \( p(s_i = k) = w_k \). Given \( s_i = k \), the observation \( y_i \) follows the individual distribution from the \( k^{th} \) component. Using Bayes theorem, the complete-data posterior distribution satisfies

\[
p(S, \theta, w|y) \propto p(y|S, \theta, w)p(S|\theta, w)p(\theta, w),
\]

where \( p(S|\theta, w) \) is the joint distribution of \( S \) before having observed the data. Given the values of \( S \), we have \( p(y|S, \theta, w) = p(y|S, \theta) \). Since the allocations of all the observations are independent with \( p(s_i = k) = w_k \), we have \( p(S|\theta, w) = p(S|w) \). Thus the complete-data posterior distribution can be simplified as

\[
p(S, \theta, w|y) \propto p(y|S, \theta)p(S|w)p(\theta, w).
\]
The posterior distribution \( p(\theta, w|y) \) and the complete-data posterior distribution \( p(S, \theta, w|y) \) satisfy the following relationship

\[
p(\theta, w|y) \propto \sum_{S \in \Omega_s} p(S, \theta, w|y),
\]

where summation is taken over the set of all possible allocations of the observations, \( \Omega_s \). If one is interested in the allocation of the observations, the marginal posterior distribution of the allocations \( S \) is given by

\[
p(S|y) \propto \int p(S, \theta, w|y) d\theta dw.
\]

One problem is to choose an appropriate prior distribution for \((\theta, w)\). It is common to assume that the component parameters \( \theta \) and the allocation weights \( w \) are independent, which means that \( p(\theta, w) = p(\theta)p(w) \). A standard form of \( p(w) \) is a Dirichlet distribution with the same hyperparameters so that the prior distribution is invariant to relabeling of the indices. The component parameters \( \theta_1, \ldots, \theta_K \) are usually assumed to be independent given a hyperparameter \( \zeta \):

\[
p(\theta_1, \ldots, \theta_k|\zeta) = \prod_{k=1}^{K} p(\theta_k|\zeta).
\]

Two common choices of \( p(\theta_k|\zeta) \) are a conditionally conjugate prior and a hierarchical prior (Frühwirth-Schnatter, 2006). The former prior is used for the complete-data likelihood \( p(y, S|\theta, w) \). Frühwirth-Schnatter (2006) states that if each component density \( p(y|\theta_k) \) is from an exponential family and for each component the prior \( p(\theta_k|\zeta) \) is from the exponential family with fixed \( \zeta \), then the conditional posterior \( p(\theta_k|S, y) \) of \( \theta_k \) will also be in the exponential family. Frühwirth-Schnatter (2006) points out that in some situations the posterior distribution may be sensitive to specific choices of \( \zeta \). To reduce the sensitivity, a common procedure is to use a hierarchical prior by
treated as unknown with a hyperprior distribution \( p(\zeta) \). Thus, we have

\[
p(\theta_1, \ldots, \theta_K) = \int p(\theta_1, \ldots, \theta_K | \zeta) \ p(\zeta) \ d\zeta.
\]

Note that \( \theta_1, \ldots, \theta_K \) are dependent now. Richardson and Green (1997) apply such prior to finite normal mixture models.

Due to the mixture likelihood form, no matter what prior distribution is chosen, the posterior distribution \( p(\theta, w | y) \) does not have a closed form density which can be sampled directly. Therefore, some sampling method should be implemented to draw samples from the posterior distribution for making inferences on the unknown parameters in the mixture distribution.

### 1.3 Markov Chain Monte Carlo Methods

Quite often a target distribution \( \pi \), which has a density with respect to a measure \( \mu \), can be evaluated but is difficult to sample directly. Markov Chain Monte Carlo (MCMC) sampling methods enable us to generate samples from the target distribution \( \pi \) by simulating a Markov chain. Let \( \chi \) denote the state space of \( \pi \) and \( A \) be any measurable set of a countably generated \( \sigma \)-algebra on \( \chi \). A transition kernel \( P(x, A) \) is a probability measure in the second argument. It gives the probability of moving from \( x \) into the set \( A \). According to Tierney (1994) and Chib and Greenberg (1995), the MCMC simulation method amounts to constructing a Markov chain \( \{ x^{(i)} \} \), \( i = 0, 1, 2, \ldots \), which has stationary and limiting distribution \( \pi \). The chain is simulated through a transition kernel \( P(x, dy) \), which satisfies the invariance condition

\[
\pi(dy) = \int P(x, dy) \pi(x) \mu(dx).
\]
The $n$th iterate of the kernel $P(x, dy)$ is written as

$$P^n(x, A) = \int \chi P^{n-1}(x, dy) P(y, A),$$

and $P(y, A)$ is constructed so that $P^n(x, A)$ converges to $\pi(A)$ as $n \to \infty$.

Tierney (1994) and Chib and Greenberg (1995) give conditions on $P$ so that $\pi$ is invariant. For example, suppose that, for some function $p(x, y)$, the transition kernel $P(x, dy)$ can be expressed as

$$P(x, dy) = p(x, y)\mu(dy) + r(x)\delta_x(dy),$$

where $p(x, x) = 0$, $\delta_x$ is a point mass at $x$, and

$$r(x) = 1 - \int p(x, y)\mu(dy),$$

which is the probability that the chain remains at $x$.

If $p(x, y)$ satisfies the reversibility condition

$$\pi(x)p(x, y) = \pi(y)p(y, x), \quad (1.2)$$

then $\pi$ is the invariant distribution of $P$. The following proof is from Tierney (1994).

Proof.

$$\int P(x, A)\pi(x)\mu(dx) = \int [\int_A p(x, y)\mu(dy)] \pi(x)\mu(dx) + \int r(x)\delta_x(A)\pi(x)\mu(dx)$$

$$= \int_A [\int p(x, y)\pi(x)\mu(dx)] \mu(dy) + \int r(x)\pi(x)\mu(dx)$$

$$= \int_A [\int p(y, x)\pi(y)\mu(dx)] \mu(dy) + \int r(x)\pi(x)\mu(dx)$$

$$= \int (1 - r(y))\pi(y)\mu(dy) + \int r(x)\pi(x)\mu(dx)$$

$$= \int \pi(y)\mu(dy).$$
Tierney (1994) provides sufficient conditions on \( P \) to guarantee that the stationary distribution \( \pi \) is also the limiting distribution with respect to the total variation norm and that empirical averages follow a strong law of large numbers. The latter is important for estimation. Suppose in fact that \( x^{(1)}, x^{(2)}, \ldots, x^{(N)} \) is an MCMC sample generated from a chain with limiting distribution \( \pi \). If we want to estimate the expectation \( \mu_g \) of an arbitrary function \( g \) with respect to \( \pi \), the estimate can be obtained as follows:

\[
\hat{\mu}_g = \frac{1}{N} \sum_{i=1}^{N} g(x^{(i)}).
\]

Assuming that the conditions in Tierney (1994) are satisfied, as \( N \rightarrow \infty \), \( \hat{\mu}_g \rightarrow \mu_g = \int g(x) \pi(x) \, dx \) by the strong law of large numbers. Assume \( g(X)^2 \) has finite expectation under \( \pi \). Then, under stronger conditions also given in Tierney (1994), a central limit theorem holds and \( \hat{\mu}_g \) has an approximate normal distribution for large \( N \). Thus, we can make statistical inference and create approximate confidence interval for \( \mu_g \).

Due to the emergence of MCMC methods, Bayesian statistics has developed tremendously during the recent years. For finite mixture models, the target distribution is the posterior distribution \( p(\theta, w | y) \). Two important and popular MCMC methods are the Gibbs sampling and the Metropolis-Hasting (MH) algorithm. Geman and Geman (1984) published an early paper using the Gibbs sampler to fit image-processing models. After the appearance of that paper, the Gibbs sampler became popular and was studied extensively. Examples of its application include Gelfand and Smith (1990) and Smith and Roberts (1993). Casella and George (1992) provide an intuitive explanation of the Gibbs sampler.
The MH algorithm was first introduced by Metropolis et al. (1953) and later generalized by Hastings (1970). The algorithm is a general and flexible MCMC method. Gelman (1992) points out that the Gibbs sampler is a special case of the MH algorithm. Tierney (1994) and Chib and Greenberg (1995) have detailed illustrations of the MH algorithm.

1.4 Issues with Sampling from a Posterior Distribution with Multimodality

Fitting Bayesian mixture models by MCMC may be computationally challenging when the posterior is multimodal. One source of multimodality is the fact that the likelihood of the mixture distribution $p(y_i|\theta, w)$ is invariant with respect to relabeling of the components, which is also a source of unidentifiability (Frühwirth-Schnatter, 2006). If an invariant prior distribution is used to specify a fully Bayesian model, the posterior distribution will then have $K!$ equivalent modes for each genuinely different mode. This identifiability problem can be dealt with by imposing an ordering constraint on the model parameters, for example, $w_1 < w_2 < \ldots < w_K$ (Richardson and Green, 1997) or $\theta_{1,l} < \theta_{2,l} < \ldots < \theta_{K,l}$ for any element $\theta_{k,l}$ of $\theta_k$. Stephens (2000b) mentions that considerable care must go into the choice of constraints to prevent label switching from occurring when the model is fit by MCMC.

Alternatively, one may fit an unconstrained model and postprocess the MCMC output using some relabeling algorithm (Stephens, 1997; Celeux, 1998; Stephens, 2000b). Stephens (1997) suggests a relabeling algorithm so that the parameters of interest have unimodal distributions. Celeux (1998) proposes an on-line relabeling algorithm by attempting to make the marginal posterior distributions of the parameters as close to normal and independent as possible. Stephens (2000b) states that the
classification of observations can be treated as an action choice problem. Setting an action space of interest and choosing an appropriate loss function, the relabeling can be determined by minimizing the posterior expected loss. He claims that relabeling methods introduced in Stephens (1997) and Celeux (1998) can be viewed as versions of his algorithm with certain action spaces and loss functions.

However, the issue of relabeling is somewhat of a red herring and exploration of the $K!$ equivalent modal regions is redundant (Celeux et al., 2000). As stated in Frühwirth-Schnatter (2006), in a marked point process representation of a mixture distribution introduced in Stephens (2000a), the $K!$ symmetric modes constitute a single point. Thus, to explore all of the $K!$ equivalent modes, the author suggests adding a Metropolis-Hastings (MH) step at each iteration to propose a random permutation of the indices.

In light of these considerations, we are mainly interested in those situations when the posterior is genuinely multimodal, which can happen, for example, if two or more of the mixture components collapse into a single component. When the posterior distribution has several isolated modes, most MCMC methods have difficulties in sampling. For example, the Gibbs sampler is unable to transition between modal regions. In many cases, if a chain is initialized in the neighborhood of a mode, the chain will be trapped near this region and will not move to the regions surrounding the other modes for very long periods of time. This results in an imprecise (and often incomplete) exploration of the posterior distribution. The masking phenomenon described in Justel and Peña (1996) in the context of linear regression models results in Gibbs samplers which are affected by this sort of problem. Specifically, the authors adopt a scale contaminated normal model to detect outliers in some data sets using the
Gibbs sampler. They find that, if a data set has many outliers that mask each other or swamp other data points, the Gibbs sampler will fail and the Gibbs output will provide poor estimates. The outliers may not be identified and some good observations may be misidentified. A similar difficulty occurs in Bayesian variable selection (George and McCulloch, 1993). In this paper $\beta_i$ denotes the regression coefficient for the predictor $x_i$ and the 0-1 latent variable $\gamma_i$ is introduced so that $\beta_i$ can be modeled as a mixture of two normal distributions with different variances

$$
\beta_i|\gamma_i \sim (1 - \gamma_i)N(0, \tau_i^2) + \gamma_i N(0, c_i^2 \tau_i^2)
$$

and

$$
P(\gamma_i = 1) = 1 - P(\gamma_i = 0) = p_i.
$$

George and McCulloch (1993) stress that when the prior of $\beta_i$ is chosen close to a “spike and slab” mixture, then it is difficult for the Gibbs sampler to move the latent variable $\gamma_i$ from $\gamma_i = 0$ to $\gamma_i = 1$ or in the reverse direction. They point out that this situation will become worse if there exist more than one model with high posterior probability and it is difficult to transition between one and the other. In this case the Gibbs sampler will take a very long time to converge.

Tailor-made methods have been developed to deal with multimodal distributions, including the MCMC tempering method (Neal, 1996; Celeux et al., 2000; Jasra et al., 2005), evolutionary Monte Carlo (Liang and Wong, 2001) and the use of mode jumping proposals (Tjelmeland and Hegstad, 2001). These are all ingenious and useful methods that need to be carefully tweaked to optimize performance. For the tempered transition method, the number of steps and a temperature parameter for each step need to be specified in advance. For some complex distributions this may be
difficult to do effectively. The use of evolutionary Monte Carlo also requires choosing
the number of distributions and their temperature parameters. For the method pro-
posed by Tjelmeland and Hegstad (2001), the jump size needs to be selected and the
jumping proposals must be created via local optimization. Since the move may not
lead the current state to be within the neighborhood of another mode, the proposed
state may fail to be accepted, with negative impact on the efficiency of the method.

These shortcomings provide the motivation for developing efficient algorithms for
fitting Bayesian mixture models. My approach hinges on the premise that a prelimi-
nary understanding of some essential features of the posterior distribution is needed
to make sampling more efficient. I have developed, fine-tuned, and tested two algo-
rithms which rely on an initial identification of possible isolated modes of the posterior
distribution. I will introduce these algorithms in Chapter 3.

1.5 Trans-Dimensional Markov Chain Monte Carlo Methods

In many practical situations, the true number of components for a finite mixture
distribution is unknown. It is then important to determine the number of components
from the data since the definition of the mixture model is based on it. Some trans-
dimensional MCMC algorithms have been developed to deal with mixture models
with an unknown number of components. A popular one is a reversible jump MCMC
algorithm (Green, 1995; Richardson and Green, 1997). The model parameters and
allocations are updated through the full conditional distributions. Two more steps
are added to perform a dimension-changing move: splitting one mixture component
into two components or combing two components into one component, and the birth
or death of an empty component. Richardson and Green (1997) point out that we
need to be more careful when calculating the acceptance probability for dimension-
changing moves. Suppose that a move type is to move from $x$ to a point $y$ in a 
higher dimensional space. To obtain $y$, a vector of continuous random variables $u$
is drawn independent of $x$ and $y$ is determined by using an invertible deterministic
function $f(x, u)$. Then the calculation of the acceptance probability must include the
term of a Jacobian, which arises from the change of variable from $(x, u)$ to $y$. The
details about these two dimension-changing move steps can be found in Richardson
and Green (1997).

An alternative trans-dimensional MCMC method is a birth and death MCMC,
which was used to deal with univariate and bivariate mixtures of normal and t-
distributions by Stephens (2000a). This method was later applied to mixtures in
regression by Hurn et al. (2003). Stephens (2000a) notes that a finite mixture distri-
bution $p(y|\theta, w)$ can be viewed as a marked point process, with each point $\theta_k$
having an associated mark $w_k$ and the marks being constrained to sum to unity. To simulate
from the posterior distribution of a finite mixture model with an unknown number
of components, Stephens (2000a) modifies methods that were developed to simulate
realizations of marked point processes and are discussed in Ripley (1977). These
methods are used to construct a continuous time spatial birth and death Markov pro-
cess with a spatial point process as the invariant distribution. Stephens (2000a) claims
that his approach appears more natural and elegant than the reversible jump MCMC
method since his method avoids the need for calculation of the complicated Jacobian,
which arises from dimension-changing moves in the reversible jump MCMC algorithm.
Stephens (2000a) contains the details of the birth and death MCMC method. Cappé
et al. (2003) demonstrate that the reversible jump and birth-and-death sampler are
similar. In addition, they show that a sequence of reversible jump samplers converges to the birth-and-death process.

1.6 Dirichlet Process Mixtures

An additional way to deal with the situation of an unknown number of components is the Dirichlet process mixture model which has become popular in recent years due to its ability to fit a countably infinite number of components. The idea of a Dirichlet process traces back to the pioneering work of Ferguson (1973), Blackwell and MacQueen (1973), Ferguson (1974) and others. The Dirichlet process mixture model specifies the following conditionally independent hierarchical structure for the data \( y = (y_1, \ldots, y_n) \):

\[
\begin{align*}
y_i | \theta_i & \sim p(y_i | \theta_i) \\
\theta_i | F(\theta_i) & \sim F(\theta_i) \\
F(\theta) & \sim DP(F_0(\theta), \alpha),
\end{align*}
\]  

(1.3)

where \( p(y | \theta) \) is a known distribution parametrized by \( \theta \) and \( DP(F_0(\theta), \alpha) \) is a Dirichlet process with base distribution \( F_0(\theta) \) and concentration parameter \( \alpha > 0 \).

Much recent research work has been concentrated on developing techniques for fitting the Dirichlet process mixture models. Escobar (1994) uses a Gibbs sampler to sample from the posterior distribution of Dirichlet process mixture models. For conjugate Dirichlet process models, the model parameters of each component can be integrated out. Thus, MacEachern (1994) and Neal (1992a) propose algorithms by updating only the indicator variable that allocates observations into mixture components. To deal with nonconjugate Dirichlet process models, Gibbs-like Markov chain sampling methods have been proposed by Damien et al. (1999), MacEachern and
Müller (2000) and Neal (1992b). These methods update the allocation for one data point at a time. Jain and Neal (2007) point out that these methods may get trapped around isolated modes when the mixture components are nearby or overlapping.

To solve the deficiencies of the methods based on Gibbs sampler, some advanced methods have been developed that update the allocation of all observations into components in a single iteration, for instance, the Green and Richardson (2001) method based on the reversible-jump procedure, and the Jain and Neal (2004) method for conjugate Dirichlet process models which uses a split-merge technique. Recently, Jain and Neal (2007) extend their method (Jain and Neal, 2004) to nonconjugate Dirichlet process mixture models. Dahl (2003) adopts ideas from sequential importance sampling (Liu, 1996; MacEachern et al., 1999) and proposes the sequentially-allocated merge-split sampler.

West (1997) states that the use of Dirichlet process mixtures is more geared towards density estimation or related objectives than parameter estimation.

1.7 Organization of This Dissertation

Two efficient MCMC algorithms are developed to deal with posterior distributions with several isolated modal regions. A review of some standard MCMC methods and of the related literature appears in Chapter 2. A detailed description of the proposed algorithms is presented in Chapter 3. In the subsequent four chapters the proposed algorithms are implemented to fit different models and are compared to other MCMC methods including the Gibbs sampler and the MCMC tempering transition methods by examining the accuracy of inferences and the ease of transition between isolated modal regions of the posterior distributions. In Chapter 4, the proposed algorithms
are implemented to fit a finite univariate normal mixture modal. In Chapter 5, the proposed algorithms are applied to a univariate outlier accommodation model. In Chapter 6, I fit the algorithms to a Bayesian regression model, which is the extension of the univariate outlier accommodation model. In Chapter 7, the algorithms are applied to fit a hierarchical Bayesian regression model for repeated measures data. All the results indicate that the proposed algorithms are superior to the Gibbs sampler and the MCMC tempering transition method. Finally, Chapter 8 provides a discussion of the work presented and makes suggestions for future extensions.
CHAPTER 2

STANDARD MCMC METHODS FOR FITTING MIXTURE MODELS

MCMC sampling methods enable us to generate samples successively from a target distribution $\pi$ and have been widely used in Bayesian analysis. As I noted in Chapter 1, mixture models are very popular due to their flexible form. However, some practical difficulties arise when MCMC methods are implemented to fit mixture distributions (Celeux et al., 2000). Some advanced algorithms have been developed to deal with these difficulties. In this chapter I review two standard MCMC methods and an advanced algorithm for fitting Bayesian mixture models.

Given a sample $y_1, \ldots, y_n$ from a mixture distribution and a prior distribution on the model parameters, I consider MCMC algorithms for generating approximate samples from the posterior distribution of the model parameters. I find it useful to distinguish between methods that operate on the original parameter space and those that work on an augmented state space where the unobserved group membership indicators for the observations $s_1, \ldots, s_n$ are added to the collection of model parameters. Note that if one uses an algorithm that operates on the original parameter space and one is interested in investigating the allocation of the observations to the various mixture components, at each MCMC iteration group membership indicators
can be generated independently conditional on the current values of the other model parameters.

### 2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm was initially introduced by Metropolis et al. (1953) for investigating properties of substances consisting of interacting individual molecules. This algorithm was later generalized by Hastings (1970) and has become better known to statisticians in recent years. The papers by Tierney (1994) and Chib and Greenberg (1995) and the book by Robert and Casella (2004) provide concise and excellent illustrations of the MH algorithm. This algorithm is versatile and most advanced MCMC methods are developed based on it.

To construct an MH-based Markov chain, we need to have a density that can generate candidate states. Let \( q(x, y) \) denote a candidate-generating density, where a new state \( y \) is generated from the density \( q \) given the current point \( x \). Initialized with an arbitrary value \( x^{(0)} \), the MH algorithm generates realizations of a Markov chain \( \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \) as follows:

1. Repeat for \( i = 1, 2, \ldots, N \).

2. Generate a candidate \( y \) from \( q(x^{(i)}, \cdot) \);

3. Compute the acceptance probability

\[
\alpha(x^{(i)}, y) = \min \left\{ \frac{\pi(y)q(y, x^{(i)})}{\pi(x)q(x^{(i)}, y)} 1 \right\}.
\]

Note that the normalizing constant of \( \pi \) is not necessary since the acceptance probability \( \alpha(x^{(i)}, y) \) only depends on \( \pi \) through the ratio \( \pi(y)/\pi(x^{(i)}) \). Also,
α(x(i), y) is always defined since y can be proposed only when both π(x) and q(x(i), y) are greater than 0;

4. Accept y as x(i+1) with probability α(x(i), y); otherwise, set x(i+1) = x(i).

5. Return the values \{x(1), x(2), \ldots, x(N)\}.

The MH algorithm is a very flexible method for constructing a Markov chain sampling scheme because it allows various choices for the candidate-generating density q(x, y). The following are some examples of this density choice presented in Tierney (1994) and Chib and Greenberg (1995).

A first class of candidate-generating densities is given by q(x, y) = f(y − x). The candidate y is generated by the process y = x + z, where z is an increment variable and is independently drawn from the distribution f. This process is called a random walk. Natural choices of the distribution f include a uniform distribution, a multivariate normal distribution and a multivariate t distribution. The random walk sampler is simple and popular in applications. However, one needs to be careful in setting the variance of z. If it is too large, the proposed state will be rejected with high probability and the chain will remain at the same state for a while. If the variance is too small, the chain will tend to make small moves and thus the chain has to be run for long periods before it can move through the whole support of the target distribution. When f is symmetric about the origin (i.e., f(∆) = f(−∆)), the acceptance probability α(x, y) can then be simplified to

$$\alpha(x, y) = \min\left\{\frac{\pi(y)}{\pi(x)}, 1\right\}.$$ 

A second choice is to select a candidate y by generating it from some density f, independently of the current point x. In this case, the candidate-generating density
is given by \( q(x, y) = f(y) \) and the acceptance probability \( \alpha(x, y) \) can be written as

\[
\alpha(x, y) = \min\left\{ \frac{\pi(y)f(x)}{\pi(x)f(y)}, 1 \right\}.
\]

Common choices of \( f \) are a multivariate normal distribution and a multivariate \( t \) distribution.

A third method for candidate generation is to choose a candidate-generating density based on the known form of the target distribution \( \pi \). For example, suppose that \( \pi \) can be expressed as \( \pi(z) \propto \rho(z)h(z) \), where \( h(\cdot) \) is a density that can be sampled from and \( \rho(\cdot) \) is uniformly bounded, then set \( q(x, y) = h(y) \) to generate a candidate \( y \). In this case, the acceptance probability \( \alpha(x, y) \) reduces to

\[
\alpha(x, y) = \min\left\{ \frac{\rho(y)}{\rho(x)}, 1 \right\}.
\]

A fourth method of drawing a candidate is to use a modified version of the acceptance-rejection (AR) method. This method was developed by Tierney (1994). Assume that \( \pi(x) \) is dominated by \( ch(x) \) for all \( x \), where \( c \) is a positive constant and \( h(x) \) is a density that can be sampled from. The steps of the AR method to generate \( y \) from the target distribution \( \pi \) are:

1. Generate \( z \) from \( h(\cdot) \) and \( u \) from the uniform distribution on \([0, 1]\).

2. If \( u \leq f(z)/ch(z) \), set \( y = z \); otherwise, return to step 1.

Note that the rejection rate tends to be very large if \( c \) is chosen excessively large. In practice it is often very difficult to choose an appropriate \( c \) such that \( \pi \) can be dominated by \( ch \). Chib and Greenberg (1995) show that for an arbitrary \( c > 0 \) the variable \( y \) generated through the AR method has density \( q(y) = \min\{\pi(y), ch(y)\} \).
Let \( C = \{ x : \pi(x) \leq ch(x) \} \). To make \( y \) have density \( \pi \), Tierney (1994) and Chib and Greenberg (1995) show that the MH acceptance probability is given by

\[
\alpha(x, y) = \begin{cases} 
1 & \text{if } x \in C,
\frac{ch(x)}{\pi(x)} & \text{if } x \notin C, y \in C,
\min\left\{ \frac{\pi(y)h(x)}{\pi(x)h(y)}, 1 \right\} & \text{if } x \notin C, y \notin C.
\end{cases}
\]

A final example of a candidate-generating density is based on an autoregressive strategy. An increment \( z \) is drawn from a density \( f \) and a candidate \( y \) is created by \( y = a + b(x - a) + z \), where \( a \) is a fixed vector and \( b \) is either a real constant or a fixed matrix. Clearly, both \( a \) and \( b \) must be conformable with \( x \). Thus, \( q(x, y) = f(y - a - b(x - a)) \). One can see that the random walk process is a special case of this family when \( b = 1 \). If \( b = -1 \), the candidate \( y \) is generated by reflecting the current state \( x \) about the point \( a \) and adding the increment \( z \). This is a natural strategy for introducing negative correlation between successive states in the chain. Tierney (1994) states that the strategy with reflection about \( a \) is most effective when the target distribution \( \pi \) is approximately symmetric about \( a \).

When one uses the MH algorithm to simulate from a Bayesian model based on a mixture distribution likelihood \( p(y|\theta, w) \) with a known number of components, usually the posterior distribution \( p(\theta, w|y) \) is taken as the target distribution and the allocations \( s_i \) of the observations to the various components are not simulated. The parameters \( \vartheta = (\theta, w) \) can be updated simultaneously or can be updated separately from the full conditionals. If the posterior distribution has several modal regions, the steps taken by a random walk may not be big enough to transition between different regions. Further, it is in general difficult to utilize any other of the aforementioned choices of a candidate-generating density to sample from the posterior distribution of the mixture model. For instance, for the last choice of \( q(x, y) \) it is very hard to
find an appropriate value of \( a \) since the locations of the modal regions are unknown. To transition effectively between different modal regions, one strategy is to combine a random walk MH algorithm with some other algorithm, such as the tempered transition method, which is able to transition between different modal regions. This approach is described in Section 2.3.

2.2 Gibbs Sampler

The Gibbs sampling algorithm was introduced by Geman and Geman (1984) to perform simulations on image-processing models. This method is easy to implement typically with simple computations and has since been extensively applied (for instance, Gelfand and Smith, 1990; Smith and Roberts, 1993). The paper by Casella and George (1992) provides an intuitive explanation of the Gibbs sampler and the book by Givens and Hoeting (2005) includes concise and excellent illustrations of this method.

Suppose that \( x \) is decomposed into \( x_1, \ldots, x_d \). Let \( x_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_d) \) and let \( \pi(x_j|x_{-j}) \) denote the full conditional density of \( x_j \). Starting at \( x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)}) \) by setting \( i = 0 \), the Gibbs sampler generates realizations of a Markov chain \( \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \) as follows:

1. Generate

\[
x_1^{(i+1)} \sim \pi(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_d^{(i)})
\]

\[
x_2^{(i+1)} \sim \pi(x_2|x_1^{(i+1)}, x_3^{(i)}, \ldots, x_d^{(i)})
\]

\[
\vdots
\]

\[
x_d^{(i+1)} \sim \pi(x_d|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{d-1}^{(i+1)})
\]
2. Set $i = i + 1$ and return to step 1.

Note that the Gibbs sampler requires that the conditional distributions can be sampled directly. An important feature of the Gibbs sampler is that each proposed parameter value is always accepted.

Gelman (1992) points out that the Gibbs sampler can be viewed as a special case of the Metropolis-Hastings algorithm. Each Gibbs update in step 1 can be interpreted as a Metropolis-Hastings step. Thus, a Gibbs cycle consists of $d$ MH steps. Note that the $j$th Gibbs update step generates $x^c = (x_1^{(i+1)}, \ldots, x_j^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_d^{(i)})$ given $x^t = (x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_j^{(i)}, \ldots, x_d^{(i)})$. This $j$th step can be realized by a MH step with the following candidate-generating density

$$q_j(x^t, x^c) = \begin{cases} 
\pi(x_j^c|x_{-j}^t) & \text{if } x_{-j}^c = x_{-j}^t, \\
0 & \text{otherwise.}
\end{cases}$$

Under this proposal density, the acceptance probability is

$$\alpha(x^t, x^c) = \frac{\pi(x^c) q_j(x^c, x^t)}{\pi(x^t) q_j(x^t, x^c)} = \frac{\pi(x^c) \pi(x_j^c|x_{-j}^c)}{\pi(x^t) \pi(x_j^t|x_{-j}^t)} = 1,$$

and thus $x^c$ is always accepted.

The Gibbs sampler is very easy to implement and it is widely used. When a posterior distribution is unimodal, the Gibbs sampler can be an efficient sampling method since the proposed parameter values are always accepted. However, when the posterior distribution has multiple isolated modes, the Gibbs sampler is not appropriate since the chain is easily attracted to a local modal region and may not move easily to other modal regions. This phenomenon is well illustrated in Celeux et al. (2000).

If this situation occurs, the inference made from the Gibbs output will be very poor.
To use the Gibbs sampler to generate samples from a Bayesian finite mixture model \( p(y|\theta, w) \), data augmentation can be applied by treating the allocations \( s_i \) of the observations as missing (\( s_i = k \) if \( y_i \) belongs to component \( k \)). Recall that \( S = (s_1, s_2, \ldots, s_n) \). The augmented parameter \((S, \theta, w)\) is then estimated by sampling from the complete-data posterior distribution \( p(S, \theta, w|y) \). By Bayes’ theorem, \( p(S, \theta, w|y) \) satisfies

\[
p(S, \theta, w|y) \propto p(y|S, \theta, w)p(S|\theta, w)p(\theta, w).
\]

### 2.3 Tempering MCMC with an Up-and-Down Power Scheme

Neal (1996) developed the tempered transition method to sample from a target distribution \( \pi \) with multiple isolated modes. This method has since been widely applied to fit mixture models (e.g., Celeux et al. 2000, and Jasra et al. 2005). In the implementation of the method presented in Jasra et al. (2005), a set of constants \( 0 < \gamma_m < \ldots < \gamma_1 < 1 \) (the temperatures) is first specified. Starting from the current MCMC state \( x^{(i)} \), the next state \( x^{(i+1)} \) is generated by implementing intermediate MCMC steps with an up-down power scheme, which is described as follows:

1. Generate \( \hat{x}_1 \) from \( x^{(i)} \) using an MCMC step with respect to \( \pi^{\gamma_1} \)

2. Generate \( \hat{x}_2 \) from \( \hat{x}_1 \) using an MCMC step with respect to \( \pi^{\gamma_2} \)

\[\vdots\]

m. Generate \( \hat{x}_m \) from \( \hat{x}_{m-1} \) using an MCMC step with respect to \( \pi^{\gamma_m} \)

m+1. Generate \( \hat{x}_{m+1} \) from \( \hat{x}_m \) using an MCMC step with respect to \( \pi^{\gamma_{m-1}} \)

\[\vdots\]
2m-1. Generate $\hat{x}_{2m-1}$ from $\hat{x}_{2m-2}$ using an MCMC step with respect to $\pi^{\gamma_1}$.

The new state $x^{(i+1)}$ is set equal to $\hat{x}_{2m-1}$ with probability

$$\min\left\{1, \frac{\pi^{\gamma_1}(x^{(i)})}{\pi(x^{(i)})} \frac{\pi^{\gamma_{m-1}}(\hat{x}_{m-1})}{\pi^{\gamma_{m-1}}(\hat{x}_{m-2})} \cdots \frac{\pi^{\gamma_1}(\hat{x}_{2m-1})}{\pi^{\gamma_1}(\hat{x}_{2m-2})} \right\},$$

and set equal to $x^{(i)}$ with the remaining probability. The acceptance probability can be rewritten as

$$\min\left\{1, \left(\frac{\pi(\hat{x}_{2m-1})}{\pi(x^{(i)})}\right)^{1-\gamma_1} \left(\frac{\pi(\hat{x}_{2m-2})}{\pi(\hat{x}_{1})}\right)^{\gamma_1-\gamma_2} \cdots \left(\frac{\pi(\hat{x}_{m})}{\pi(\hat{x}_{m-1})}\right)^{\gamma_{m-1}-\gamma_m} \right\}. \quad (2.1)$$

Note that the acceptance probability only depends on the ratios of the flattened-out distributions $\pi^{\gamma_j}$ and does not depend on the normalizing constants. One can see that the first $m$ steps flatten out the target by raising $\pi$ to a sequence of monotonically decreasing powers ($\gamma_1 > \ldots > \gamma_m$) thus facilitating MCMC transitions. The next $m-1$ simulations help to return to a state that receives high posterior probability. Equation (2.1) shows that the acceptance probability should be large when the differences between consecutive temperatures $\gamma_i - \gamma_{i+1}$ are small. A satisfactory performance of this up-and-down power algorithm rests on a judicious choice of the number of intermediate distributions and of the temperature schedule: $\pi^{\gamma_m}$ ought to be flat enough to favor good mixing and the intermediate steps ought to be close enough to ensure good acceptance probabilities (but, obviously, the computational burden increases as the number of intermediate steps increases).

The tempering transitions facilitate global moves and can be combined with MH steps for local exploration. Jasra et al. (2005) consider the model

$$y_i|\theta, K \sim \sum_{k=1}^{K} w_k N(\mu_k, \lambda_k^{-1}),$$

27
with the priors specified as

\[ \mu_k \sim N(\xi, \kappa^{-1}), \]
\[ \lambda_k | \beta \sim \text{Gamma}(\alpha, \beta), \]
\[ \beta \sim \text{Gamma}(g, h), \]
\[ w \sim \text{Dirichlet}(\delta, \ldots, \delta), \]

where \( w = (w_1, \ldots, w_K) \) and the priors are the same for each component \( k = 1, \ldots, K \). The tempering method is combined with Metropolis-Hastings to simulate from the model above by implementing MH steps with a specified probability \( u \) and performing the tempering steps with probability \( 1 - u \). An important modification in this paper is that the component mean vector \( \mu = (\mu_1, \ldots, \mu_K) \), precision vector \( \lambda = (\lambda_1, \ldots, \lambda_K) \) and reparametrized allocation weights \( v = (v_1, \ldots, v_K) \) with \( w_k = v_k / \sum_{k=1}^{K} v_k \) are simulated from their own full conditionals, i.e, \( p(\mu|\cdot), p(\lambda|\cdot), \) and \( p(v|\cdot) \), respectively instead of simulating them from the full posterior distribution \( p(\mu, \lambda, v|\cdot) \) simultaneously. The authors argue that one may achieve higher acceptance probabilities by reducing the dimensions of the sampled parameters.
CHAPTER 3

NEW MCMC METHODS FOR FITTING MIXTURE MODELS

As I noted in Chapter 1, some MCMC methods have been developed to deal with a posterior distribution for a Bayesian mixture model with multiple isolated modes. However, these methods have intrinsic shortcomings. This gave me the motivation to develop efficient algorithms for fitting Bayesian mixture models. My approach is based on the premise that a preliminary understanding of some essential features of the posterior distribution is helpful for drawing samples that are reflective of the posterior distribution. I have developed two algorithms which are based on an initial identification of possible isolated modes of the posterior distribution. The first algorithm has three basic steps and does not simulate the group membership indicators for the observations. Implementation of the second algorithm also follows three basic steps. This algorithm operates on an augmented state space where group membership indicators for the observations are generated together with model parameters. In this chapter, I describe these two algorithms.
### 3.1 Algorithm Based on a Mixture of Modal Proposals

The first algorithm relies on a preliminary identification of the possible modes of the posterior distribution through some MCMC method which is capable, albeit typically inefficiently, of exploring the whole support. The algorithm works on the original parameter space and consists of the following three basic steps:

1. Run an MCMC algorithm to obtain preliminary estimates and locate the possible modes of the posterior distribution.

2. Build an MH transition kernel capable of proposing each possible modal region with high probability.

3. Combine the MH sampler developed in Step (2) with a random walk sampler and generate an MCMC sample to compute the final estimates.

The algorithm introduced in Ibrahim and Jennison (2007) has some features in common with the above algorithm. With regard to Step (1), it is known that some MCMC algorithms such as the tempered transition method reviewed in Chapter 2 are capable of exploring the whole posterior region when the generated Markov chain sample path is long enough, although their mixing properties may be poor. Thus, these MCMC methods can be used to perform an initial exploration of the posterior distribution. The possible modes can be located from the samples simulated in Step (1) by using some clustering algorithm. In Step (2), for each identified mode, a local Gaussian proposal distribution is built and centered at the mode. The center and variance covariance matrix of each local proposal are estimated based on those samples generated in Step (1) which can be associated with the corresponding modal
region. The local proposals are then combined to form a mixture distribution capable of proposing model parameters from the various modal regions. The mixture components can be weighted equally or more appropriate weights can be estimated based on the samples from Step (1).

Suppose that $M$ possible modes are detected by Step (1). Let $q_m$ denote the local normal distribution centered at the $m^{th}$ mode with mean $\mu_m$ and variance covariance matrix $\Sigma_m$ derived in Step 2 and let $v_m$ denote the corresponding weight used to form a mixture distribution. A candidate $\theta^c$ of the mixture parameters $\theta$ can then be generated from the mixture proposal distribution

$$q(\theta) = \sum_{m=1}^{M} v_m q_m(\theta; \mu_m, \Sigma_m). \quad (3.1)$$

The above proposal does not depend on the current state $\theta^t$ and allows the chain to move freely across all the modal regions identified in Step (1), thus enhancing mixing. The random walk sampler added in Step (3) enables the chain to fully explore each local modal region.

Under the above proposal, the acceptance probability can then be expressed as

$$\alpha = \min \left\{ \frac{p(\theta^c | y)}{p(\theta^t | y)} \frac{q(\theta^t)}{q(\theta^c)} \right\}.$$  

If one is interested in investigating the allocation of the observations to the various mixture components, at each iteration of the sampler group membership indicators can be generated independently conditional on the current values of the other model parameters.
3.2 Hybrid Metropolis-Hastings-Gibbs Algorithm with Clustering Initialization

Recall that \( s_i \) denotes the mixture component membership of each observation \( y_i \), where \( s_i = k \) if \( y_i \) belongs to component \( k \), \( i = 1, \ldots, n \). My second algorithm works on an augmented state space in which the indicators \( s_i \) are simulated together with the model parameters. In this algorithm, clustering is first performed so that the observations can be grouped into homogeneous subsets which possibly correspond to nearly disjoint posterior regions centered around different modes. An appropriate MH kernel is then developed to facilitate transitions between the isolated modal regions. This is later combined with the Gibbs sampler (or some other MCMC algorithm) that facilitates local moves. This algorithm also consists of three basic steps that can be summarized as follows:

1. Identify nearly disjoint posterior regions based on a preliminary clustering of the observations;

2. Build an MH transition kernel to propose these nearly disjoint regions with high probability;

3. Construct a hybrid algorithm that combines a basic Gibbs sampler with MH steps based on the kernel in Step (2).

Model-based clustering is a comprehensive clustering strategy of choosing the “optimal” number of clusters via the Bayesian Information Criterion (BIC) and the classification of the observations is determined from the EM algorithm. Detailed and concise descriptions of this approach can be found in the papers by Fraley and Raftery (1998) and Fraley and Raftery (2002). The examples in Dasgupta and Raftery (1998)
indicate that this clustering method works well for Gaussian mixture models. The R package *mclust* has been developed to implement model-based clustering for normal mixtures.

In Step (1), model-based clustering is appropriate if the observations follow approximately a normal mixture distribution, but clustering methods based on other distance measures might be more appropriate if the observations substantially violate the normal mixture assumption. The successful execution of this step is very important for the overall performance of the algorithm. Step (2) is central to the successful performance of the algorithm and guidelines about construction of efficient transition kernels are detailed in the following subsection. To complete Step (3), a probability \( p_{\text{MH}} \) must be specified in advance. Then, at each iteration, with probability \( p_{\text{MH}} \) the algorithm updates the parameters using the MH transition kernel of Step (2) and with probability \( 1 - p_{\text{MH}} \) the update is realized via the Gibbs sampler. Like the random walk in Step (3) of the first algorithm, the Gibbs sampler added in Step (3) is used to fully explore each local modal region. When \( p_{\text{MH}} = 0 \), the algorithm becomes a Gibbs sampler which is typically unable to transition easily between the various isolated modes. When \( p_{\text{MH}} = 1 \), the simulation of the membership parameters is only operated by using the MH kernel of Step (2). This implementation enables a chain to move between the isolated modes, but each local modal region may not be well explored. The performance of the algorithm can be fine-tuned by varying the value of the parameter \( p_{\text{MH}} \).
3.2.1 The Combination of Two Proposal Kernels

The combination of two proposal kernels introduced in Tjelmeland and Hegstad (2001) to facilitate jumping between different modes is used there for a continuous state space. Here, I extend that approach to a general state space. Let \( x \) denote the state variable. In my application \( x \) is the complete set of parameters \((\theta, S)\). Suppose that we have a posterior distribution \( \pi \) on \( x \) with values in a state space \( \chi \) and that \( \pi \) has a density with respect to a measure \( \mu \). Let \( \Omega \) be a measurable space on \( \chi \). As we learned in Chapter 1, given any measurable set \( A \in \Omega \), if a transition kernel \( P \) is chosen to satisfy

\[
\int_A \pi(x)\mu(dx) = \int_{\chi} P(x, A)\pi(x)\mu(dx),
\]

then the Markov Chain defined by the transition kernel \( P \) admits \( \pi \) as its invariant distribution.

Assume a proposal kernel \( Q \) to have a density \( q(x, y) \) with respect to \( \mu \). This means that

\[
Q(x, dy) = q(x, y)\mu(dy).
\]

Then the Metropolis transition kernel \( P \) is given by

\[
P(x, A) = \int_A q(x, y)\alpha(x, y)\mu(dy) + \delta_x(A)r(x),
\]

where \( \delta_x(A) = 1 \) if \( x \in A \) and 0 otherwise, \( \alpha(x, y) \) is an acceptance probability with the following expression

\[
\alpha(x, y) = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\},
\]

and \( r(x) \) is the probability of staying at \( x \), which is

\[
r(x) = \int_{\chi} q(x, y)(1 - \alpha(x, y))\mu(dy).
\]
Note that, in the notation of Section 1.3, \( q(x, y)\alpha(x, y) \) corresponds to \( p(x, y) \).

Consider two proposal kernels \( Q_0 \) and \( Q_1 \) with corresponding proposal densities \( q_0 \) and \( q_1 \) respectively. A Markov transition kernel \( P \) can be defined as

\[
P(x, A) = \frac{1}{2} \int_A q_0(x, y)\alpha_0(x, y)\mu(dy) + \frac{1}{2} \int_A q_1(x, y)\alpha_1(x, y)\mu(dy) + \delta_x(A)r(x),
\]

where the acceptance probabilities \( \alpha_0(x, y) \) and \( \alpha_1(x, y) \) are

\[
\alpha_i(x, y) = \min \left\{ 1, \frac{\pi(y)}{\pi(x)} \frac{q_{1-i}(y, x)}{q_i(x, y)} \right\} \quad i = 0, 1,
\]

and \( r(x) \) is given by

\[
r(x) = \frac{1}{2} \int_X q_0(x, y) (1 - \alpha_0(x, y)) \mu(dy) + \frac{1}{2} \int_X q_1(x, y) (1 - \alpha_1(x, y)) \mu(dy).
\]

It turns out that \( \pi \) is an invariant distribution for the above transition kernel \( P \).

**Proof of the claim that \( \pi \) is invariant for \( P \):**

\[
\int_X P(x, A)\pi(x)\mu(dx)
= \int_X \left\{ \frac{1}{2} \int_A q_0(x, y)\alpha_0(x, y)\mu(dy) + \frac{1}{2} \int_A q_1(x, y)\alpha_1(x, y)\mu(dy) \right\} \pi(x)\mu(dx) + \\
\int_A r(x)\pi(x)\mu(dx)
= \frac{1}{2} \int_X \int_A \pi(x)q_0(x, y)\alpha_0(x, y)\mu(dy)\mu(dx) + \frac{1}{2} \int_X \int_A \pi(x)q_1(x, y)\alpha_1(x, y)\mu(dy)\mu(dx) + \\
\frac{1}{2} \int_A \int_X \pi(x)q_0(x, y) (1 - \alpha_0(x, y)) \mu(dy)\mu(dx) + \\
\frac{1}{2} \int_A \int_X \pi(x)q_1(x, y) (1 - \alpha_1(x, y)) \mu(dy)\mu(dx)
= I_1 + I_2 + I_3 + I_4
\]
and

\[ I_1 = \frac{1}{2} \int_A \int_X \min\{\pi(x)q_0(x, y), \pi(y)q_1(y, x)\} \mu(dy)\mu(dx) \]
\[ = \frac{1}{2} \int_A \int_X \min\{\pi(x)q_0(x, y), \pi(y)q_1(y, x)\} \mu(dx)\mu(dy) \]
\[ = \frac{1}{2} \int_A \int_X \min\{\pi(y)q_0(y, x), \pi(x)q_1(x, y)\} \mu(dy)\mu(dx) \]
\[ = \frac{1}{2} \int_A \int_X \pi(x)q_1(x, y)\alpha_1(x, y)\mu(dy)\mu(dx). \]

Thus, we have

\[ I_1 + I_4 = \frac{1}{2} \int_A \int_X \pi(x)q_1(x, y)\alpha_1(x, y)\mu(dy)\mu(dx) + \]
\[ \frac{1}{2} \int_A \int_X \pi(x)q_1(x, y)(1 - \alpha_1(x, y)) \mu(dy)\mu(dx) \]
\[ = \frac{1}{2} \int_A \int_X \pi(x)q_1(x, y)\mu(dy)\mu(dx) \]
\[ = \frac{1}{2} \int_A \pi(x)\mu(dx). \]

Similarly, we have

\[ I_2 = \frac{1}{2} \int_A \int_X \pi(x)q_0(x, y)\alpha_0(x, y)\mu(dy)\mu(dx) \]
and

\[ I_2 + I_3 = \frac{1}{2} \int_A \pi(x)\mu(dx). \]

Therefore, \( I_1 + I_2 + I_3 + I_4 = \int_A \pi(x)\mu(dx). \)

The potential new state \( y \) is proposed from the density \( q_i(x, y), i = 0 \) or 1, with 50-50 chance and accepted with probability \( \alpha_i \). Note that the strategy of the combination of two proposal kernels randomly picks one proposal to draw a potential new state and uses the other proposal to move back to the current state. If we used a mixture proposal instead, we would employ the same distribution to propose a candidate and go back to the current state.
3.2.2 Building the Metropolis-Hastings Kernel of Step (2)

Two terms are first defined before describing the construction of the MH transition kernel. Recall that $s_i, s_i \in \{1, \ldots, K\}$, denotes the mixture component membership indicator of each observation $y_i$. The vector $S = (s_1, \ldots, s_n)$ of membership indicators is called a *configuration*. The clustering operation of Step (1) assigns the observations to various homogeneous subsets. The observations in each subset (or at least most of them) are likely to belong to the same mixture component. These assignments are called *preferred configurations* and denoted by $S_j^*, j = 1, \ldots, J$, where $J$ is the number of assignments. For example, suppose that two clusters (called $A$ and $B$) are identified by the clustering method in Step (1), which suggests the existence of $K = 2$ mixture components. If we want to fit a model without identifiability constraints and consider the possible collapse of two mixture components into one, then $J$ would be set equal to 4. The first preferred configuration $S_1^* = (1, \ldots, 1)$ indicates that all observations belong to component 1 and component 2 is empty. Relabeling $S_1^*$ produces the second preferred configuration $S_2^* = (2, \ldots, 2)$ which means that all observations belong to component 2 and component 1 is empty. The third preferred configuration is $S_3^* = (s_{31}^*, \ldots, s_{3n}^*)$ with $s_{3i}^* = 1$ if observation $y_i$ is assigned to cluster $A$ and $s_{3i}^* = 2$ if observation $y_i$ is assigned to cluster $B$. Relabeling $S_3^*$ leads to $S_4^* = (s_{41}^*, \ldots, s_{4n}^*)$ with $s_{4i}^* = 1$ if observation $y_i$ is assigned to cluster $B$ and $s_{4i}^* = 2$ if observation $y_i$ is assigned to cluster $A$.

Considering that the $J$ preferred configurations identified in the clustering step are likely to be tied to nearly disjoint posterior regions, it will be helpful, in an effort to build an effective MH kernel, to have some preliminary knowledge about what values of the parameters are in agreement with these $J$ preferred configurations. To this end,
preliminary samples of $\theta$ for each preferred configuration are generated by running an MCMC method (typically the Gibbs sampler) conditional on the assignment of the observations to the various mixture components that is implied by that preferred configuration. Note that, up to relabeling, some preferred configurations essentially correspond to the same assignment of the observations. Thus, we need only generate one set of MCMC draws for all preferred configurations that correspond to the same assignment. For example, in the previous example configurations $S^*_1$ and $S^*_2$ correspond to the same assignment, and configurations $S^*_3$ and $S^*_4$ correspond to another one. Once the draws of $\theta$ are generated for $S^*_1$, the corresponding draws for $S^*_2$ can be obtained by relabeling. Let $\theta_j^{(1)}, \ldots, \theta_j^{(L)}$ denote the simulated samples corresponding to the preferred configuration $S^*_j$, where $L$ is the sample size.

I build the transition kernel by specifying rules for proposing a new configuration and collection of model parameters $(S^c, \theta^c)$, given the current state $(S^t, \theta^t)$ of the simulated chain at time $t$. As stated in Tjelmeland and Hegstad (2001), to obtain frequent transitions between isolated modes, the transition kernel must be constructed so that (1) the proposed states are located in modal regions with high posterior probability; (2) the acceptance probabilities must be high, so that the proposed states are accepted frequently. Given our preliminary understanding of the salient features of the posterior distribution, task (1) can be accomplished by proposing a configuration in the neighborhood of one of the preferred configurations and a compatible set of parameters $\theta$. However, the acceptance probability in a MH step, derived to achieve the detailed balance condition, balances the relative sizes of the posterior probabilities and of the transition probabilities for the proposed and the current states. If a state is proposed which has a configuration that is too close to one of the preferred

38
configurations, it may not be accepted because the chance of returning to the current state might be too small.

This problem is dealt with by adapting to my setting the method introduced in Tjelmeland and Hegstad (2001), which is based on the combination of two proposal densities \( q_\ell, \ell = 0, 1 \). In my application the proposal densities \( q_\ell \) are built as follows. One of the preferred configurations is first picked with some probability \( p_{\delta j} \) and denoted by \( S_p \). Then, \( q_\ell(S_t, S_c) \) is built by setting

\[
s_i^c = \begin{cases} 
  s_i^p & \text{with prob. } r_\ell \\
  s_{i,1}^p, \ldots, s_{i,K-1}^p & \text{with prob. } (1 - r_\ell)/(K - 1)
\end{cases}, \quad \text{for } i = 1, \ldots, n,
\]

where \( s_{i,j}^p \) denotes a label selected at random among those different from \( s_i^p \) and \( r_\ell \) is a kernel-specific parameter value. Typically, the values of \( r_0 \) and \( r_1 \) should be close to one. They can be tuned to achieve good performance and obtain reasonably high acceptance probabilities by balancing the attractiveness of the proposed state with the likelihood of transitioning back to the current one. Note that the new potential configuration \( S^c \) generated by the above proposal is independent of the current configuration \( S_t \). Thus, \( q_\ell(S_t, S^c) \) can be rewritten as \( q_\ell(S^c) \).

To complete the construction of the transition kernel, a candidate model parameter vector \( \theta^c \) must be proposed conditional on the proposed configuration \( S^c \). Suppose that \( S^*_j \) is the preferred configuration with the minimum number of switches (i.e., non-matching labels) from the proposed configuration \( S^c \). The ratio of the posterior of \( \theta \) conditional on \( S^c \) to that conditional on \( S^*_j \) is evaluated at each of the preliminary Gibbs sampler draws \( \theta_j^{(l)} \) corresponding to \( S^*_j \), yielding

\[
w_j^{(l)} = \frac{p\left( \theta_j^{(l)} | y, S^c \right)}{p\left( \theta_j^{(l)} | y, S^*_j \right)},
\]

39
for \( l = 1, \ldots, L \). The normalized ratio \( \bar{w}^{(l)}_{j^*} \) are then computed as:

\[
\bar{w}^{(l)}_{j^*} = \frac{w^{(l)}_{j^*}}{\sum_{l=1}^{L} w^{(l)}_{j^*}} \quad l = 1, \ldots, L.
\]

To generate a coordinate or a group of coordinates, \( \theta \), of \( \theta \), an effective approach is to construct a proposal distribution whose shape is comparable with that of the full posterior distribution conditional on \( S^c \) and having first and second moments matching those of \( \theta \) given \( S^c \). These moments are estimated by importance sampling using the preliminary MCMC draws according to the following expressions

\[
E[\theta] \approx \sum_{l=1}^{L} \bar{w}^{(l)}_{j^*} \theta^{(l)}_{j^*},
\]

\[
E[\theta^2] \approx \sum_{l=1}^{L} \bar{w}^{(l)}_{j^*} (\theta^{(l)}_{j^*})^2.
\]

For example, for a univariate normal mixture model we have the component parameters \((\mu_k, \sigma^2_k)\), where \( \mu_k \) is the mean of component \( k \) and \( \sigma^2_k \) is its variance, for \( k = 1, \ldots, K \). Under conjugate priors the full conditional distribution of \( \mu_k \) is normal and the full conditional of \( \sigma^2_k \) is Inverse-Gamma. We can use these two distributional forms as proposal for generating candidate values of \( \mu_k \) and \( \sigma^2_k \). The parameters in each proposal distribution can be estimated by the method of moments combined with importance sampling. Specifically, the first and second moments of \( \mu_k \) given the proposed configuration \( S^c \) can be estimated as:

\[
m_k = \sum_{l=1}^{L} \bar{w}^{(l)}_{j^*} \mu^{(l)}_{j^*,k},
\]

\[
u_k = \sum_{l=1}^{L} \bar{w}^{(l)}_{j^*} (\mu^{(l)}_{j^*,k})^2.
\]

Thus, the variance parameter of the proposal distribution for \( \mu_k \) can be obtained as:

\[
v_k = u_k - (m_k)^2.
\]
At this point, we can generate $\mu_k$ from the distribution $N(m_k, v_k)$. The same procedure can be applied to estimate the shape and rate parameters for the proposal of $\tau_k = 1/\sigma_k^2$. First, we note that the first and second moments given $S^c$ can be approximated as:

$$g_k = \sum_{l=1}^{L} \sum_{j} \sum_{j^*} \sum_{\ell} \frac{w_j^{(l)}(\tau_{j^*,k}^{(l)})}{\tau_{j^*,k}^{(l)}}$$

$$h_k = \sum_{l=1}^{L} \sum_{j} \sum_{j^*} \sum_{\ell} \frac{w_j^{(l)}(\tau_{j^*,k}^{(l)})^2}{\tau_{j^*,k}^{(l)}}$$

and, the parameter rate $b_k$ and shape $a_k$ can be estimated by:

$$b_k = \frac{g_k}{h_k} - (g_k)^2$$

$$a_k = b_k \times g_k.$$

Therefore, we can generate $\sigma_k^2$ from the Inverse-Gamma$(a_k, b_k)$ distribution.

### 3.2.3 Acceptance Probability

From the two proposal densities $q_\ell(S^t, \cdot)$ described above, one proposal is randomly selected to propose a potential new configuration $S^c$. Assuming that $q_\ell$ is chosen and denoting by $q(\theta|S)$ the proposal density for $\theta$ conditional on $S$, following Tjelmelend and Hegstad (2001) the acceptance probability $\alpha_\ell$ can be expressed as:

$$\alpha_\ell = \frac{p(S^c, \theta^c|y) \cdot q_{1-\ell}(S^t) \cdot q(\theta^t|S^t)}{p(S^t, \theta^t|y) \cdot q_\ell(S^c) \cdot q(\theta^c|S^c)}, \quad \ell = 0, 1,$$

where

$$q_\ell(S^c) = \sum_{j=1}^{J} \left\{ P(S_j^c) \prod_{i=1}^{n} \left[ I(s_i^e = s_{j,i}^c) r_\ell + I(s_i^e \neq s_{j,i}^c) (1 - r_\ell) / (K - 1) \right] \right\},$$

$I(\cdot)$ is an indicator variable, and $r_\ell$ is the kernel-specific parameter value corresponding to the proposal $q_\ell$. 

41
CHAPTER 4

APPLICATION TO A FINITE UNIVARIATE NORMAL MIXTURE MODEL

In this chapter I apply the algorithms described in Chapter 3 to a finite univariate mixture model. One goal of the analysis is to estimate the posterior probabilities of some specific allocations of the observations to the mixture components. The proposed algorithms are compared to other MCMC methods, such as the Gibbs sampler and MCMC tempering, by evaluating the accuracy of estimation of those posterior probabilities and by examining the ease of transition between isolated modal regions of the posterior distribution.

4.1 A Finite Univariate Normal Mixture Model

A common special case of Model (1.1) is the univariate normal mixture given by

\[ p(y_i | \theta, w) = \sum_{k=1}^{K} w_k N(y_i | \mu_k, \sigma_k^2), \quad \text{independently for } i = 1, \ldots, n, \quad (4.1) \]

with \( \theta_k = (\mu_k, \sigma_k^2) \).

The mixture model (4.1) arises if the group membership of an observation \( y_i \) is unknown. Let \( s_i \) denote the group membership indicator of the observation \( y_i \), where \( s_i \in \{1, \ldots, K\} \) and \( s_i = k \) if \( y_i \) belongs to component \( k \). An observation \( y_i \)
is assumed to be independently allocated to component $k$ with probability $w_k$, i.e., $p(s_i = k) = w_k$. Given the indicator $s_i = k$, $y_i$ follows the distribution $N(y_i|\mu_k, \sigma_k^2)$.

Then, the marginal distribution of $y_i$ can be written as

$$p(y_i|\theta, w) = \sum_{k=1}^{K} p(y_i|\theta, w, s_i = k)p(s_i = k) = \sum_{k=1}^{K} w_k N(y_i|\mu_k, \sigma_k^2).$$

Here, I assume that the number of components is known and use the prior distributions described in Richardson and Green (1997). Specifically, I let

$$w \sim \text{Dirichlet}(\eta, \ldots, \eta)$$

$$\mu_k \sim \text{N}(\xi, \psi)$$

$$\sigma_k^{-2}|\beta \sim \text{Gamma}(\alpha, \beta)$$

$$\beta \sim \text{Gamma}(g, h).$$  \hspace{1cm} (4.2)

The prior on the allocation vector $w$ is taken to be a symmetric Dirichlet distribution with parameter $\eta$, which is appropriate since the model is invariant with respect to relabeling of the components. To make the prior on $\mu_k$ flat over the range of the data, Richardson and Green suggest that $\xi$ can be set equal to the midpoint of the data and $\psi^{-1}$ equal to a small multiple of $1/R^2$, where $R$ is the length of the data interval. Since it is hard to capture the size of $\sigma_k^2$ from the range of the data, Richardson and Green propose a hyper prior for $\beta$, which is a gamma distribution with parameters $g$ and $h$.

### 4.2 Full Conditional Distributions

To use the Gibbs sampler to simulate posterior samples from Model (4.1), a data augmentation strategy is applied by treating the $s_i$ as unknown. Thus, the unknown parameters are $\mu = (\mu_1, \ldots, \mu_k)$, $\sigma^2 = (\sigma^2_1, \ldots, \sigma^2_K)$, $w$, $S = (s_1, \ldots, s_n)$ and $\beta$.  

43
Conditional on the number of components, the complete-data posterior distribution of these parameters satisfies

\[ p(w, \mu, \sigma^2, \beta, S | y) \propto p(y | \mu, \sigma^2, S) p(w) p(S | w) p(\mu) p(\sigma^2 | \beta) p(\beta). \]

Under the priors (4.2), the full conditional distribution of \(w\) satisfies

\[ p(w | \mu, \sigma^2, \beta, S, y) \propto p(w) p(S | w) \]

\[ \propto K \prod_{k=1}^{K} w^{\eta_{n_k} - 1} \prod_{i=1}^{n} w^I_{s_i = k} \]

\[ \propto K \prod_{k=1}^{K} w^{\eta + n_k - 1}, \]

where \(n_k = \# \{i : s_i = k\}\). Thus, the full conditional distribution of \(w\) is Dirichlet with parameters \(\eta + n_1, \ldots, \eta + n_K\).

Let \( \mu_{-k} = (\mu_1, \ldots, \mu_{k-1}, \mu_{k+1}, \ldots, \mu_K) \). For \(\mu_k\) we have

\[ p(\mu_k | \mu_{-k}, \sigma^2, \beta, S, y) \propto \prod_{i : s_i = k} \left\{ \frac{1}{\sqrt{2\pi \sigma_k}} \exp \left[ - \frac{(y_i - \beta_k)^2}{2\sigma_k^2} \right] \right\} \]

\[ \propto \sqrt{2\pi \psi} \exp \left[ - \frac{(\mu_k - m_k)^2}{2\psi} \right], \]

where \(m_k = (\sigma_k^{-2} \sum_{i : s_i = k} y_i + \psi^{-1} \beta_k) / (\sigma_k^{-2} n_k + \psi^{-1})\) and \(v_k^2 = (\sigma_k^{-2} n_k + \psi^{-1})^{-1}\). Thus, the full conditional distribution of \(\mu_k\) is normal with mean \(m_k\) and variance \(v_k^2\).

Let \(\sigma_{-k}^2 = (\sigma_1^2, \ldots, \sigma_{k-1}^2, \sigma_{k+1}^2, \ldots, \sigma_K^2)\). The full conditional of \(\sigma_k^2\) satisfies

\[ p(\sigma_k^{-2} | \sigma_{-k}^2, \mu, w, \beta, S, y) \propto \prod_{i : s_i = k} \left\{ \sigma_k^{-1} \exp \left[ - \frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right] \right\} (\sigma_k^{-2})^{n_k-1} \exp(-\beta / \sigma_k^2) \]

\[ \propto (\sigma_k^{-2})^{\alpha_n + n_k/2 - 1} \exp \left\{ - \frac{1}{\sigma_k^2} \left[ \beta + \frac{1}{2} \sum_{i : s_i = k} (y_i - \mu_k)^2 \right] \right\}. \]
Thus, the full conditional of $\sigma^2_k$ is Inverse-Gamma with shape $\alpha + n_k/2$ and rate $\beta + \sum_{\{i : s_i = k\}} (y_i - \mu_k)^2/2$.

For the allocation variables $s_i$ we have

$$p(s_i = k|\mu, \sigma^2, w, \beta, y) \propto w_k \exp\left\{ -\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right\}.$$  

Finally, for the hyperparameter $\beta$ we have

$$p(\beta|\mu, \sigma^2, S, y) \propto \beta^{g-1} \exp\left( -h\beta \right) \prod_{k=1}^K (\sigma_k^{-2})^{\alpha-1} \exp\left( -\beta/\sigma_k^2 \right)$$

$$\propto \beta^{(g+K\alpha)} \exp\left[ -(h + \sum_{k=1}^K \sigma_k^{-2})\beta \right].$$

Thus, the full conditional distribution of $\beta$ is a Gamma distribution with shape $g+K\alpha$ and rate $h + \sum_{k=1}^K \sigma_k^{-2}$.

### 4.3 Posterior Distributions of $\mu$, $\sigma^2$ and $w$

For the MCMC tempering transition method reviewed in Section 2.2, the allocation variables $s_i$, $i = 1, \ldots, n$, are not simulated together with the other model parameters $\mu$, $\sigma^2$ and $w$. Rather, samples are generated from the joint posterior distribution $p(\mu, \sigma^2, w|y)$. Clearly, $\sigma^2_k$ is constrained to be positive, and $w_k$ is non-negative with $\sum_{k=1}^K w_k = 1$, $k = 1, \ldots, K$. Jasra et al. (2005) point out that the Metropolis-Hastings algorithm may not perform well on a constrained space and tempering MCMC essentially uses Metropolis-Hasting kernels to simulate from the posterior distribution. Thus, $\sigma^2_k$, for $k = 1, \ldots, K$, is reparameterized as $u_k = -\log(\sigma^2_k)$ and $w_k$, for $k = 1, \ldots, K$, is reparameterized as $v_k = \log(w_k/(1 - w_k))$. Jasra et al. (2005) note that acceptance rates may be improved by reducing the dimensionality of the simulated parameters. Let $u = (u_1, \ldots, u_K)$ and $v = (v_1, \ldots, v_K)$. Then, to
reduce the dimension of the simulated parameters, $\mu$, $u$ and $v$ are generated through tempering with intermediate steps (see Section 2.3 for details) from their full conditional distributions $p(\mu|u, v, y)$, $p(u|\mu, v, y)$ and $p(v|\mu, u, y)$, respectively. To ensure that local modal regions are fully explored, a MH step is added. With probability $q$ the candidate values for $\mu$, $u$ and $v$ are generated through additive random walks. With probability $1 - q$ the simulation is executed via the tempering method.

The likelihood $p(y|\mu, u, v)$ is given by

$$p(y|\mu, u, v) \propto n \prod_{i=1}^{n} \sum_{k=1}^{K} \left\{ \frac{e^{u_k} \sqrt{e^{u_k}}}{1 + e^{u_k}} \exp \left[ -\frac{(y_i - \mu_k)^2}{2/e^{u_k}} \right] \right\}.$$ 

Under the priors (4.2), the full conditional of $\mu$ satisfies

$$p(\mu|u, v, y) \propto p(y|\mu, u, v)p(\mu)$$

$$\propto \prod_{i=1}^{n} \sum_{k=1}^{K} \left\{ \frac{e^{u_k} \sqrt{e^{u_k}}}{1 + e^{u_k}} \exp \left[ -\frac{(y_i - \mu_k)^2}{2/e^{u_k}} \right] \right\}$$

$$\times \prod_{k=1}^{K} \left\{ \frac{1}{\sqrt{2\pi \psi}} \exp \left[ -\frac{(\mu_k - \xi)^2}{2\psi} \right] \right\}.$$ 

In the priors (4.2), $\sigma_k^2$, for $k = 1, \ldots, K$, depends on the hyperparameter $\beta$. Integrating $\beta$ out yields the prior for $(\sigma_1^{-2}, \ldots, \sigma_K^{-2})$, which is given by

$$p(\sigma_1^{-2}, \ldots, \sigma_K^{-2}) = \int \prod_{k=1}^{K} p(\sigma_k^{-2} | \beta) p(\beta) d\beta$$

$$\propto \int \prod_{k=1}^{K} \left[ \beta^\alpha (\sigma_k^{-2})^{\alpha-1} \exp(-\beta \sigma_k^{-2}) \right] \beta^{\alpha-1} e^{-h \beta} d\beta$$

$$\propto \prod_{k=1}^{K} (\sigma_k^{-2})^{\alpha-1} \left( \sum_{k=1}^{K} \sigma_k^{-2} + h \right)^{-g-K\alpha}.$$ 

By a change of variables argument, $p(u)$ can be seen to satisfy

$$p(u) \propto \exp \left( \alpha \sum_{k=1}^{K} u_k \left( \sum_{k=1}^{K} e^{u_k} + h \right)^{-g-K\alpha} \right).$$
Thus, the full conditional of \( u \) satisfies

\[
p(u|\mu, v, y) \propto p(y|\mu, u, v)p(u) \\
\propto \prod_{i=1}^{n} \sum_{k=1}^{K} \left\{ \frac{e^{v_k}}{1 + e^{v_k}} \sqrt{\frac{e^{u_k}}{2\pi}} \exp \left[ -\frac{(y_i - \mu_k)^2}{2/e^{u_k}} \right] \right\} \\
\exp \left( \alpha \sum_{k=1}^{K} u_k \right) \left( \sum_{k=1}^{K} e^{u_k} + h \right)^{-G-K\alpha}.
\]

Similarly, the prior of \( v \) satisfies

\[
p(v) \propto \left\{ \prod_{k=1}^{K} \frac{e^{v_k}}{1 + e^{v_k}} \right\}^{\eta-1}.
\]

Thus, for \( v \) we have

\[
p(v|\mu, u, y) \propto p(y|\mu, u, v)p(v) \\
\propto \prod_{i=1}^{n} \sum_{k=1}^{K} \left\{ \frac{e^{v_k}}{1 + e^{v_k}} \sqrt{\frac{e^{u_k}}{2\pi}} \exp \left[ -\frac{(y_i - \mu_k)^2}{2/e^{u_k}} \right] \right\} \left\{ \prod_{k=1}^{K} \frac{e^{v_k}}{1 + e^{v_k}} \right\}^{\eta-1}.
\]

At each iteration once the samples for \( \mu, \sigma^2 \) and \( w \) are simulated, \( s_i \), for \( i = 1, \ldots, n \), can be generated independently conditional on the current values of \( \mu, \sigma^2 \) and \( w \).

### 4.4 The Marginal Posterior Distribution of the Allocation Variables

One goal of the analysis is to estimate the posterior probabilities that the observations are allocated to some specific configurations of interest. By Bayes theorem, the marginal posterior density \( p(S|y) \) satisfies

\[
p(S|y) \propto p(y|S)p(S).
\]
The prior distribution \( p(S) \) can be written as

\[
p(S) = \int p(S|\mathbf{w})p(\mathbf{w})d\mathbf{w} = \int \frac{\Gamma(K\eta)}{\Gamma(\eta)^K} \prod_{k=1}^{K} \left\{ w_k^{n_k} w_k^{\eta-1} \right\} d\mathbf{w} = \frac{\Gamma(K\eta)}{\Gamma(\eta)^K} \prod_{k=1}^{K} \frac{\Gamma(\eta + n_k)}{\Gamma(K\eta + n)} ,
\]

where \( n_k = \# \{ i : s_i = k \} \). The integrated likelihood \( p(y|S) \) is equal to

\[
p(y|S) = \int p(y|\mu, \sigma^2, S)p(\mu, \sigma^2|\beta)p(\beta)d\mu d\sigma^2 d\beta = \int \prod_{k=1}^{K} \left\{ \prod_{i:s_i = k} p(y_i|\mu_k, \sigma_k^2) \right\} p(\mu_k)p(\sigma_k^2|\beta) p(\beta)d\mu d\sigma^2 d\beta.
\]

Under the priors (4.2) we have

\[
\int \prod_{k=1}^{K} \left[ p(\sigma_k^2|\beta) \right] p(\beta)d\beta = \int \prod_{k=1}^{K} \left[ \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_k^2)^{-(\alpha+1)} \exp(-\beta/\sigma_k^2) \right] \frac{h^g}{\Gamma(g)} \beta^{(g-1)} \exp(-h\beta) d\beta
\]

\[
= \frac{\Gamma(K\alpha + g)h^g}{\Gamma(\alpha)^K \Gamma(g)} \left( \prod_{k=1}^{K} \sigma_k^2 \right)^{-(\alpha+1)} \left( \sum_{k=1}^{K} \sigma_k^{-2} + h \right)^{-(K\alpha + g)}
\]

and

\[
\int \prod_{i:s_i = k} p(y_i|\mu_k, \sigma_k^2)p(\mu_k)d\mu_k \propto \int \sigma_k^{-n_k} \exp \left\{ -\frac{1}{2\sigma_k^2} \sum_{i:s_i = k} (y_i - \mu_k)^2 \right\} \exp \left\{ -\frac{(\mu_k - \xi)^2}{2\psi} \right\} d\mu_k \propto \frac{\sigma_k^{-n_k}}{\sqrt{a_k}} \exp(c_k),
\]

48
where

\[ a_k = \frac{n_k}{\sigma_k^2} + \psi^{-1}, \]
\[ b_k = \frac{n_k \bar{y}_k/\sigma_k^2 + \xi \psi^{-1}}{n_k/\sigma_k^2 + \psi^{-1}}, \]
\[ c_k = \frac{(n_k \bar{y}_k/\sigma_k^2 + \xi \psi^{-1})^2}{2(n_k/\sigma_k^2 + \psi^{-1})} - \frac{1}{2\sigma_k^2} \sum_{\{i:s_i=k\}} y_i^2 - \frac{\xi^2}{2\psi}, \]

with

\[ \bar{y}_k = \frac{1}{n_k} \sum_{\{i:s_i=k\}} y_i. \]

Thus, the integrated likelihood \( p(y|S) \) satisfies

\[ p(y|S) \propto \int f(\sigma^2) d\sigma^2, \]

where

\[ f(\sigma^2) = \left( \sum_{k=1}^{K} \sigma_k^{-2} + h \right)^{-(2\alpha+\rho)} \prod_{k=1}^{K} \left[ (\sigma_k^2)^{-(n_k/2+\alpha+1)}a_k^{-\frac{1}{2}} \right] \exp \left( \sum_{k=1}^{K} c_k \right). \]

The above derivation shows that, under the priors (4.2), the marginal posterior distribution of the allocation variables \( S \) does not have a closed form. For a small number of components and a small sample size, numerical methods can be applied to approximate the exact marginal posterior probability of a specific \( S \) (for example, the adaptive numerical method implemented by the function adapt in the R library adapt). Note that, when the sample size \( n \) is large, we have too many possible allocations to add up to obtain the normalizing constant.

For the case of a large number of components or a large data set, samples can be generated from the model using MCMC algorithms and Monte Carlo integration can be used to approximate \( p(S|y) \). Suppose that \( S^{(1)}, \ldots, S^{(N)} \) are the simulated
samples for $S$. Then, for a specific allocation $S_l$, $p(S = S_l | y)$ can be estimated as

$$p(S = S_l | y) = \frac{1}{N} \sum_{j=1}^{N} I(S^{(j)} = S_l),$$

where $I(\cdot)$ is a 0-1 indicator.

### 4.5 Example

Here I consider a data set with 10 observations generated as, $y_i \sim N(-1.4, 0.2^2)$, for $i = 1, \ldots, 5$ and $y_i \sim N(0.5, 0.1^2)$, for $i = 6, \ldots, 10$. The scatter plot of the data set is presented in Figure 4.1. There are two reasonable explanations for the data set, which may correspond to two posterior modal regions:

**Region 1.** All the observations are in the same group. This means that one component is empty.

**Region 2.** The first five observations and the remaining five observations are in different different mixture components.

A univariate two-component mixture model with $f(y_i) = wN(y_i; \mu_1, \sigma_1^2) + (1 - w)N(y_i; \mu_2, \sigma_2^2)$ was fitted to this data set. The prior distributions in (4.2) were used and I specified $\eta = 1$, $\xi$ equal to the mean of the data, $\psi$ equal to $R^2$, where $R$ is the length of the data interval, and $\alpha = 2$. The hyperparameter $\beta$ was given a $G(g, h)$ prior with $g = 0.2$ and $h = 10/R^2$. One goal of the analysis is to estimate the posterior probabilities of some interesting allocations, such as $Z_1 = \{(1, \ldots, 5), (6, \ldots, 10)\}$, which indicates that the observations $y_1, \ldots, y_5$ are allocated to one mixture component and the observations $y_6, \ldots, y_{10}$ are allocated to the other component. Other allocations of interest are $Z_2 = \{(1, \ldots, 4), (5, \ldots, 10)\}$, $Z_3 = \{(1, \ldots, 6), (7, \ldots, 10)\}$.
and $Z_4 = \{(1, \ldots, 10)\}$. The exact posterior probabilities of these allocations were evaluated by using an adaptive numerical method and recorded in the row labeled $p_Z$ of Table 4.1. This table shows that $Z_1$ takes up almost all the posterior probabilities and $Z_4$ has tiny posterior probability. Thus, Region 1 will be hard to explore for most MCMC algorithms.

To examine the accuracy of the inferences concerning the posterior probabilities of these allocations, for each MCMC method under consideration I generated a long chain and subdivided each chain into 100 batches with 10,000 iterations each, after discarding the first 50,000 iterations. The root mean squared errors (RMSE) were then computed based on the squared deviations of the batch estimates from the true posterior probabilities.

Figure 4.1: Scatter plot of the data.
I ran the Gibbs sampler and the Gibbs samples are plotted in Figure 4.2. The plots show that the Gibbs sampler had difficulty with transitioning between the two symmetric modal regions of Region 2 which correspond to a relabeling of the observations in each of the two mixture components. The RMSEs of the Gibbs sampler estimates of each $p_Z$ are recorded in the rows of Table 4.2 labeled “Gibbs.” Although the Gibbs sampler cannot transition easily between the two symmetric modal regions of Region 2, this does not adversely impact the estimate of the probability of $Z_1$. The reason is that such probability only accounts for the fact that the first five observations are assigned to one mixture component and the remaining five to the other component, irrespective of which label is assigned to which component. The same is also true for the estimates of the probabilities of $Z_2$ and $Z_3$. The RMSE of the estimate of the probability of $Z_4$ is, however, quite large. This is a consequence of the slow mixing of the Gibbs sampler. Specifically, as seen in Figure 4.3, many consecutive samples consistent with Region 1 were generated around iteration 882,900, leading to a gross overestimate of the probability of $Z_4$ based on the batch containing that stretch of Gibbs iterates.

To implement the MCMC tempering sampler of Chapter 2, I set the number $m$ of intermediate steps at 17 and the temperature parameter at each step was given by $\gamma_j = 1/(2^j)$, $j = 1, \ldots, m$. The MCMC samples are summarized in Figure 4.4, where

<table>
<thead>
<tr>
<th>$Z$</th>
<th>${(1, \ldots, 5), (6, \ldots, 10)}$</th>
<th>${(1, \ldots, 4), (5, \ldots, 10)}$</th>
<th>${(1, \ldots, 6), (7, \ldots, 10)}$</th>
<th>${(1, \ldots, 10)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Z$</td>
<td>0.999329</td>
<td>0.00016</td>
<td>0.00013</td>
<td>0.000033</td>
</tr>
</tbody>
</table>

Table 4.1: Exact posterior probabilities of $Z_1$, $Z_2$, $Z_3$ and $Z_4$. 


Figure 4.2: Trace plots of the sampled parameters by the Gibbs sampler. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).
Figure 4.3: Trace plots of the sampled parameters by the Gibbs sampler for iterations between 882,800 and 883,000. (a) and (b): the output of $\mu$; (c) and (d): the output of $\sigma^2$. 
we can find that this method can transition between the two symmetric modal regions corresponding to Region 2 better than the Gibbs sampler. However, it still cannot mix well between these two symmetric modal regions. The RMSEs of the tempering estimates of each $p_Z$ are recorded in the rows of Table 4.2 labeled “Tempering.” The relatively smaller RMSE for $Z_4$ implies that this method can move between different posterior regions better than the Gibbs sampler.
The implementation of the algorithm in Section 3.1 is based on the preliminary identification of possible isolated modal regions by some method such as the tempering transition method. The trace plots of the tempering samples indicate that there exist two possible isolated symmetric modal regions. Local proposals with a multivariate normal distribution for the parameter vector \((\mu_1, \mu_2, -\log(\sigma_1^2), -\log(\sigma_2^2), \log\left(\frac{w}{1-w}\right))\) were then built with center and variance covariance matrices estimated based on the tempering samples within each identified modal region. Equal weights were assigned to these local proposals for building a mixture proposal, which is used in Step (2) of the algorithm of Section 3.1. The generated samples are summarized in Figure 4.5, which clearly indicates that this algorithm mixes very well between the two isolated modal regions of Region 2. The RMSEs of estimates of each \(p_Z\) are recorded in the rows of Table 4.2 labeled “Mixture.”

To implement the algorithm of Section 3.2, the model-based clustering method mentioned in Section 3.2 was first used to identify the possible isolated modal regions. This method separates \(\{y_1, \ldots, y_5\}\) and \(\{y_6, \ldots, y_{10}\}\), which indicates two possible clusters. Considering the possible collapse of two mixture components into one, four important preferred configurations were then used. They are \(S_1^* = (1, \ldots, 1)\), \(S_2^* = (2, \ldots, 2)\), \(S_3^* = (1, 1, 1, 1, 2, 2, 2, 2)\), and \(S_4^* = (2, 2, 2, 2, 1, 1, 1, 1)\). The procedure of Section 3.2.2 was used to generated candidates for \(\mu_k\) and \(\sigma_k^2\), for \(k = 1, \ldots, K\). A candidate \(w\) was generated from the full conditional distribution \(B(\delta + n_1, \delta + n_2)\), where \(n_k = \#\{i : s_i = k\}\). Finally, a candidate for the hyperparameter \(\beta\) was drawn from its full conditional distribution \(G(g + 2\alpha, h + \sum_{k=1}^{2} \sigma_k^{-2})\).

To examine how the tuning parameters \(r_0, r_1\) and \(p_{\text{MH}}\) affect the algorithm’s performance, different sets of \((r_0, r_1, p_{\text{MH}})\) were chosen (see Tables 4.2 and 4.3 for the values
Figure 4.5: Trace plots of the sampled parameters by the proposed algorithm in Section 3.1. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).
Figure 4.6: Trace plots of the sampled parameters by the proposed algorithm in Section 3.2. (a) and (b): the output of $\mu$ for the last 10,000 iterations (every tenth); (c) and (d): the output of $\sigma^2$ for the last 10,000 iterations (every tenth).

of these tuning parameters). The samples with $(r_0 = 0.95, r_1 = 0.90, p_{MH} = 0.30)$ are summarized in Figure 4.6. These plots show that this algorithm mixes quite well between the two isolated modal regions of Region 2. The RMSEs of the estimates of each $p_Z$ are recorded in the rows of Tables 4.2 and 4.3 labeled “Hybrid”.

58
As noted earlier, Table 4.2 shows that the Gibbs estimate for the posterior probability of configuration $Z_4 = \{1, \ldots, 10\}$ has large RMSE due to slow mixing. However, Tables 4.2 and 4.3 show that the estimates for the posterior probability of configuration $Z_4 = \{1, \ldots, 10\}$ obtained using the algorithm proposed in Section 3.2 have almost always smaller RMSE than the Gibbs’ estimate. This indicates that the added MH steps in this proposed algorithm can effectively facilitate transitions between different posterior regions. The tempering estimates of the probability of $Z_4 = \{1, \ldots, 10\}$ are much more accurate than those produced by the Gibbs sampler. However, the trace plots of the tempering samples indicate that this method does not transition well between the two symmetric isolated modal regions of Region 2. The proposed mixture-based algorithm, on the other hand, can transition well between different posterior regions, including the two symmetric modal regions of Region 2.

The estimates for configurations $Z_2 = \{(1, \ldots, 4), (5, \ldots, 10)\}$ and $Z_3 = \{(1, \ldots, 6), (7, \ldots, 10)\}$ obtained using the proposed algorithms are comparable (but slightly worse) to the Gibbs’ estimates. However, the gain for configuration $Z_4 = \{1, \ldots, 10\}$ obtained using the proposed algorithm is much larger than the loss for configurations $Z_2 = \{(1, \ldots, 4), (5, \ldots, 10)\}$ and $Z_3 = \{(1, \ldots, 6), (7, \ldots, 10)\}$ since the true posterior probabilities of the latter configurations are around 5 and 4 times larger than that of configuration $Z_4 = \{1, \ldots, 10\}$.

With regard to the implementation of the algorithm in Section 3.2, I noticed that, for given $r_0$ and $r_1$, the RMSEs of the estimates for the posterior probabilities for configurations $Z_2 = \{(1, \ldots, 4), (5, \ldots, 10)\}$ and $Z_3 = \{(1, \ldots, 6), (7, \ldots, 10)\}$ tend to increase when $p_{\text{MH}}$ increases, which means that the chain can move around the preferred configurations $S_3^*$ and $S_4^*$ better if the frequency of Gibbs steps is increased.
The RMSEs of the estimates for $Z_4 = \{1, \ldots, 10\}$ tend to be large when $p_{\text{MH}}$ is set at 0.3. This is due to the fact that most of the steps in the chain are updated via the Gibbs sampler. Overall, this algorithm performs better when the tuning parameters $(r_0, r_1)$ are set at $(r_0, r_1) = (0.99, 0.85)$ and $(r_0, r_1) = (0.95, 0.85)$. Thus, it is effective to use different values for $r_0$ and $r_1$ in this example.
<table>
<thead>
<tr>
<th>Algorithm</th>
<th>RMSE $\times 10^6$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>${(1,\ldots,5), (6,\ldots,10)}$</td>
</tr>
<tr>
<td>Gibbs</td>
<td>66.11</td>
</tr>
<tr>
<td>Tempering</td>
<td>76.51</td>
</tr>
<tr>
<td>Mixture</td>
<td>94.92</td>
</tr>
<tr>
<td>Hybrid</td>
<td></td>
</tr>
<tr>
<td>$(r_0, r_1, p_{MH}) =$</td>
<td></td>
</tr>
<tr>
<td>(0.99, 0.99, 0.30)</td>
<td>61.35</td>
</tr>
<tr>
<td>(0.99, 0.99, 0.50)</td>
<td>54.14</td>
</tr>
<tr>
<td>(0.99, 0.99, 0.70)</td>
<td>65.95</td>
</tr>
<tr>
<td>(0.99, 0.99, 0.80)</td>
<td>67.29</td>
</tr>
<tr>
<td>(0.99, 0.99, 0.90)</td>
<td>119.89</td>
</tr>
<tr>
<td>(0.99, 0.90, 0.30)</td>
<td>75.28</td>
</tr>
<tr>
<td>(0.99, 0.90, 0.50)</td>
<td>59.82</td>
</tr>
<tr>
<td>(0.99, 0.90, 0.70)</td>
<td>73.32</td>
</tr>
<tr>
<td>(0.99, 0.90, 0.80)</td>
<td>98.53</td>
</tr>
<tr>
<td>(0.99, 0.90, 0.90)</td>
<td>104.86</td>
</tr>
<tr>
<td>(0.99, 0.85, 0.30)</td>
<td>57.07</td>
</tr>
<tr>
<td>(0.99, 0.85, 0.50)</td>
<td>53.33</td>
</tr>
<tr>
<td>(0.99, 0.85, 0.70)</td>
<td>65.30</td>
</tr>
<tr>
<td>(0.99, 0.85, 0.80)</td>
<td>96.92</td>
</tr>
<tr>
<td>(0.99, 0.85, 0.90)</td>
<td>98.26</td>
</tr>
<tr>
<td>(0.99, 0.80, 0.30)</td>
<td>59.66</td>
</tr>
<tr>
<td>(0.99, 0.80, 0.50)</td>
<td>58.98</td>
</tr>
<tr>
<td>(0.99, 0.80, 0.70)</td>
<td>67.35</td>
</tr>
<tr>
<td>(0.99, 0.80, 0.80)</td>
<td>74.26</td>
</tr>
<tr>
<td>(0.99, 0.80, 0.90)</td>
<td>87.52</td>
</tr>
</tbody>
</table>

Table 4.2: RMSE’s for the marginal posterior estimates on the configurations $Z_1 = \{(1,\ldots,5), (6,\ldots,10)\}$, $Z_2 = \{(1,\ldots,4), (5,\ldots,10)\}$, $Z_3 = \{(1,\ldots,6), (7,\ldots,10)\}$, $Z_4 = \{1,\ldots,10\}$ based on 100 batches with $r_0 = 0.99$, each batch with length 10,000 iterations.
| Algorithm | \[ (r_0, r_1, p_{MH}) = \]
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>((0.95, 0.95, 0.30))</td>
<td>((0.95, 0.95, 0.50))</td>
<td>((0.95, 0.95, 0.70))</td>
<td>((0.95, 0.95, 0.80))</td>
</tr>
<tr>
<td>((r_0, r_1, p_{MH}) = )</td>
<td></td>
<td>((0.95, 0.90, 0.30))</td>
<td>((0.95, 0.90, 0.50))</td>
<td>((0.95, 0.90, 0.70))</td>
</tr>
<tr>
<td>Hybrid</td>
<td>71.48</td>
<td>68.89</td>
<td>63.64</td>
<td>62.06</td>
</tr>
<tr>
<td></td>
<td>18.81</td>
<td>21.76</td>
<td>27.21</td>
<td>23.47</td>
</tr>
<tr>
<td></td>
<td>16.63</td>
<td>18.31</td>
<td>19.30</td>
<td>13.07</td>
</tr>
<tr>
<td></td>
<td>24.75</td>
<td>7.37</td>
<td>6.69</td>
<td>7.33</td>
</tr>
<tr>
<td></td>
<td>((0.95, 0.85, 0.30))</td>
<td>((0.95, 0.85, 0.50))</td>
<td>((0.95, 0.85, 0.70))</td>
<td>((0.95, 0.85, 0.80))</td>
</tr>
<tr>
<td>((r_0, r_1, p_{MH}) = )</td>
<td></td>
<td>((0.95, 0.80, 0.30))</td>
<td>((0.95, 0.80, 0.50))</td>
<td>((0.95, 0.80, 0.70))</td>
</tr>
<tr>
<td>Hybrid</td>
<td>47.99</td>
<td>62.00</td>
<td>71.53</td>
<td>100.14</td>
</tr>
<tr>
<td></td>
<td>17.57</td>
<td>31.76</td>
<td>34.16</td>
<td>36.60</td>
</tr>
<tr>
<td></td>
<td>16.62</td>
<td>20.36</td>
<td>32.05</td>
<td>22.58</td>
</tr>
<tr>
<td></td>
<td>5.80</td>
<td>7.70</td>
<td>4.93</td>
<td>7.44</td>
</tr>
<tr>
<td></td>
<td>((0.95, 0.95, 0.30))</td>
<td>((0.95, 0.95, 0.50))</td>
<td>((0.95, 0.95, 0.70))</td>
<td>((0.95, 0.95, 0.80))</td>
</tr>
<tr>
<td>((r_0, r_1, p_{MH}) = )</td>
<td></td>
<td>((0.95, 0.80, 0.30))</td>
<td>((0.95, 0.80, 0.50))</td>
<td>((0.95, 0.80, 0.70))</td>
</tr>
<tr>
<td>Hybrid</td>
<td>51.93</td>
<td>61.97</td>
<td>69.11</td>
<td>87.96</td>
</tr>
<tr>
<td></td>
<td>18.87</td>
<td>23.49</td>
<td>25.64</td>
<td>26.54</td>
</tr>
<tr>
<td></td>
<td>15.79</td>
<td>16.80</td>
<td>33.00</td>
<td>27.64</td>
</tr>
<tr>
<td></td>
<td>10.08</td>
<td>14.74</td>
<td>11.84</td>
<td>10.18</td>
</tr>
</tbody>
</table>

Table 4.3: RMSE’s for the marginal posterior estimates on the configurations \(Z_1 = \{(1, \ldots, 5), (6, \ldots, 10)\} \}, \(Z_2 = \{(1, \ldots, 4), (5, \ldots, 10)\} \}, \(Z_3 = \{(1, \ldots, 6), (7, \ldots, 10)\} \), \(Z_4 = \{1, \ldots, 10\} \) based on 100 batches with \(r_0 = 0.95\), each batch with length 10,000 iterations.
CHAPTER 5

APPLICATION TO AN OUTLIER ACCOMMODATION MODEL

In this chapter I apply the algorithms described in Chapter 3 to fit the univariate outlier accommodation model discussed in Verdinelli and Wasserman (1991). To examine the performance of these algorithms against that of other algorithms, the Gibbs sampler and the MCMC tempering transition method reviewed in Chapter 2 are also implemented to fit the model. One goal of the analysis is to estimate the posterior probability that an observation $y_i$ is an outlier. The exact posterior probabilities can be evaluated by numerical quadratures. At the same time, these probabilities can also be estimated from the output of each implemented method. Thus, by comparing the estimates with the exact posterior probabilities we can observe the performance of each method and tell how well the proposed algorithms work.

5.1 An Outlier Accommodation Model

The outlier accommodation model proposed by Verdinelli and Wasserman (1991) states that

$$ y_i = \mu + \delta_i c_i + \varepsilon_i, \quad i = 1, \ldots, n, $$

(5.1)
where \( \varepsilon_i \sim iid \sim \mathcal{N}(0, \sigma^2_\varepsilon) \). With the assumption of conditional independence, the remaining pieces of the model are specified as \( \mu \sim \mathcal{N}(\xi, \sigma^2_\mu) \), \( c_i \sim \mathcal{N}(0, \sigma^2_c) \), for \( i = 1, \ldots, n \), \( \delta_i \sim \text{Bernoulli}(p_0) \), for \( i = 1, \ldots, n \), and \( \sigma^2_\varepsilon \sim IG(a, b) \). The hyperparameter values \( \xi \), \( \sigma^2_\mu \), \( \sigma^2_c \), \( p_0 \), \( a \) and \( b \) are prespecified. The unknown parameters are \( \{\delta_i\}_{i=1}^n \), \( \{c_i\}_{i=1}^n \), \( \mu \), and \( \sigma^2_\varepsilon \).

The parameter \( \delta_i \) is an indicator variable; \( \delta_i = 1 \) means that the observation \( y_i \) is an outlier and a term \( c_i \) is added to inflate the variance of \( y_i \) so that \( y_i \) can be accommodated. An observation \( y_i \) is identified as an outlier if the posterior probability of \( \{\delta_i = 1\} \) is large.

### 5.2 Full Conditional Distributions

To use the Gibbs sampler to generate samples for the model (5.1), the indicators \( \delta_i \), for \( i = 1, \ldots, n \), are treated as missing and generated together with other model parameters. The full conditional distributions of the augmented parameter \( (\mu, \sigma^2_\varepsilon, \{\delta_i\}_{i=1}^n, \{c_i\}_{i=1}^n) \) are required to implement the Gibbs sampler. Let \( \delta = (\delta_1, \ldots, \delta_n) \) and \( c = (c_1, \ldots, c_n) \). The complete-data posterior distribution satisfies

\[
p(\mu, \sigma^2_\varepsilon, \delta, c | y) \propto p(y | \delta, c, \mu, \sigma^2_\varepsilon) \ p(\delta) \ p(c) \ p(\mu) \ p(\sigma^2_\varepsilon)
\]

\[
\propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi} \sigma_\varepsilon} \exp \left[ -\frac{(y_i - \mu - \delta_i c_i)^2}{2\sigma^2_\varepsilon} \right] \right\} \\
\prod_{i=1}^n \left\{ p_0^{\{\delta_i = 1\}} (1 - p_0)^{\{\delta_i = 0\}} \right\} \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi} \sigma_c} \exp \left[ -\frac{c_i^2}{2\sigma^2_c} \right] \right\} \\
\frac{1}{\sqrt{2\pi} \sigma_\mu} \exp \left[ -\frac{(\mu - \xi)^2}{2\sigma^2_\mu} \right] \frac{\Gamma(a)}{\Gamma(a+c)} \left( \frac{\sigma^2_\varepsilon}{\sigma^2_\mu} \right)^{(a+c+1)/2} e^{-\frac{a+c}{\sigma^2_\mu}}.
\]  

(5.2)
The full conditional distribution of \( \mu \) satisfies

\[
p(\mu | \delta, c, \sigma^2, y) \propto \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi} \sigma_{\epsilon}} \exp \left[ -\frac{(y_i - \mu - \delta_i c_i)^2}{2\sigma^2_{\epsilon}} \right] \right\} \frac{1}{\sqrt{2\pi} \sigma_{\mu}} \exp \left[ -\frac{(\mu - \xi)^2}{2\sigma^2_{\mu}} \right] \propto \frac{1}{\sqrt{2\pi} v} \exp \left[ -\frac{(\mu - m)^2}{2v^2} \right],
\]

where \( v^2 = (n/\sigma^2_{\epsilon} + 1/\sigma^2_{\mu})^{-1} \) and \( m = \left( \sum_{i=1}^{n} (y_i - \delta_i c_i)/\sigma^2_{\epsilon} + \xi/\sigma^2_{\mu} \right)/v^2 \). Thus, the full conditional of \( \mu \) is normal with mean \( m \) and variance \( v^2 \).

Let \( c_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, c_n) \). Then, for the full conditional distribution of the location shifts \( c_i \) we have

\[
p(c_i | \mu, \sigma^2_{\epsilon}, \delta, c_{-i}, y) \propto \frac{1}{\sqrt{2\pi} \sigma_{\epsilon}} \exp \left[ -\frac{(y_i - \mu - \delta_i c_i)^2}{2\sigma^2_{\epsilon}} \right] \frac{1}{\sqrt{2\pi} v} \exp \left( -\frac{c_i^2}{2v^2} \right) \propto \left\{ \begin{array}{ll}
\frac{1}{\sqrt{2\pi} \sigma_{\epsilon}} \exp \left[ -\frac{c_i^2}{2\sigma^2_{\epsilon}} \right] & \text{if } \delta_i = 0, \\
\frac{1}{\sqrt{2\pi} v} \exp \left[ -\frac{(c_i - m_{ci})^2}{2v^2} \right] & \text{if } \delta_i = 1, 
\end{array} \right.
\]

where \( m_{ci} = \sigma^2_{\epsilon}(y_i - \mu)/(\sigma^2_{\epsilon} + \sigma^2_{\mu}) \) and \( v^2 = (1/\sigma^2_{\epsilon} + 1/\sigma^2_{\mu})^{-1} \).

The full conditional distribution of the error variance \( \sigma^2_{\epsilon} \) satisfies

\[
p(\sigma^2_{\epsilon} | \mu, \sigma^2_{\epsilon}, \delta, c, y) \propto \frac{1}{\sigma^2_{\epsilon}} \exp \left[ -\frac{\sum_{i=1}^{n} (y_i - \mu - \delta_i c_i)^2}{2\sigma^2_{\epsilon}} \right] (\sigma^2_{\epsilon})^{-a_{\epsilon} + n/2 - 1} e^{-b_{\epsilon} / \sigma^2_{\epsilon}} \propto (\sigma^2_{\epsilon})^{-a_{\epsilon} + n/2 + 1} \exp \left\{ -\left[ b_{\epsilon} + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu - \delta_i c_i)^2 \right] / \sigma^2_{\epsilon} \right\}.
\]

Thus, the full conditional distribution of \( \sigma^2_{\epsilon} \) is Inverse-Gamma with shape \( a_{\epsilon} + n/2 \) and scale \( b_{\epsilon} + \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu - \delta_i c_i)^2 \).

For the outlier indicators \( \delta_i \) we have

\[
p(\delta_i = 0 | \mu, \sigma^2_{\epsilon}, c, y) \propto (1 - p_o) \exp \left[ -\frac{(y_i - \mu)^2}{2\sigma^2_{\epsilon}} \right], \\
p(\delta_i = 1 | \mu, \sigma^2_{\epsilon}, c, y) \propto p_o \exp \left[ -\frac{(y_i - \mu - c_i)^2}{2\sigma^2_{\epsilon}} \right].
\]

65
All the above conditional distributions have a closed form and can be sampled directly. Thus, the Gibb sampler can be implemented by generating samples of each parameter from their full conditionals.

### 5.3 Conditional Distributions of $\mu$ and Reparameterized $\sigma_\epsilon^2$

As in Section 4.3, for the tempering method of Section 2.2, the outlier indicators $\delta_i$ and location shifts $c_i, i = 1, \ldots, n$, are not simulated together with the other model parameters $\mu$ and $\sigma^2$, and $\sigma^2$ is reparameterized as $v_\epsilon = -\log(\sigma_\epsilon^2)$. Then, $\mu$ and $v_\epsilon$ are simulated through tempering from their full conditionals $p(\mu|v_\epsilon, y)$ and $p(v_\epsilon|\mu, y)$, respectively.

To obtain the conditional distributions $p(\mu|v_\epsilon, y)$ and $p(v_\epsilon|\mu, y)$ of $\mu$ and $v_\epsilon$, we first need to derive the joint posterior distribution of $(\mu, \sigma^2)$. From model (5.1), we have

$$p(y_i|\mu, \sigma^2, c_i) = p_0 N(y_i|\mu + c_i, \sigma^2) + (1 - p_0) N(y_i|\mu, \sigma^2),$$

and

$$p(y_i|\mu, \sigma^2) = \int p(y_i|\mu, \sigma^2, c_i) p(c_i) dc_i = p_0 N(y_i|\mu, \sigma^2 + \sigma^2_c) + (1 - p_0) N(y_i|\mu, \sigma^2).$$

Then $p(\mu, \sigma^2|y)$ satisfies

$$p(\mu, \sigma^2|y) \propto \prod_{i=1}^n p(y_i|\mu, \sigma^2) p(\mu) p(\sigma^2)$$

$$\propto \prod_{i=1}^n \left\{ \frac{p_0}{\sqrt{2\pi(\sigma^2_c + \sigma^2)}} \exp \left[ -\frac{(y_i - \mu)^2}{2(\sigma^2_c + \sigma^2)} \right] + \frac{1 - p_0}{\sqrt{2\pi \sigma_\epsilon}} \exp \left[ -\frac{(y_i - \mu)^2}{2\sigma^2_\epsilon} \right] \right\} \frac{1}{\sqrt{2\pi \sigma_\mu}} \exp \left[ -\frac{(\mu - \xi)^2}{2\sigma^2_\mu} \right] \frac{b_{\sigma^2_\epsilon}^{a_{\sigma^2_\epsilon}}}{\Gamma(a_{\sigma^2_\epsilon})} \left(\sigma^2_\epsilon\right)^{-(a_{\sigma^2_\epsilon}+1)} e^{-b_{\sigma^2_\epsilon}/\sigma^2_\epsilon}. \right.$$
It follows that the conditional distribution $p(\mu|v, y)$ of $\mu$ satisfies

$$p(\mu|v, y) \propto \prod_{i=1}^{n} p(y_i|\mu, v) p(\mu)$$

$$\propto \prod_{i=1}^{n} \left\{ \frac{p_0}{\sqrt{2\pi}(\exp(-v_i) + \sigma_i^2)} \exp \left[ -\frac{(y_i - \mu)^2}{2(\exp(-v_i) + \sigma_i^2)} \right] + \frac{1 - p_0}{\sqrt{2\pi \exp(-v_i)}} \exp \left[ -\frac{(y_i - \mu)^2}{2 \exp(-v_i)} \right] \right\} \frac{1}{\sqrt{2\pi \sigma}} \exp \left[ -\frac{(\mu - \xi)^2}{2\sigma^2} \right],$$

and the conditional distribution $p(v|\mu, y)$ of $v$ satisfies

$$p(v|\mu, y) \propto \prod_{i=1}^{n} p(y_i|\mu, v) p(v)$$

$$\propto \prod_{i=1}^{n} \left\{ \frac{p_0}{\sqrt{2\pi}(\exp(-v_i) + \sigma_i^2)} \exp \left[ -\frac{(y_i - \mu)^2}{2(\exp(-v_i) + \sigma_i^2)} \right] + \frac{1 - p_0}{\sqrt{2\pi \exp(-v_i)}} \exp \left[ -\frac{(y_i - \mu)^2}{2 \exp(-v_i)} \right] \right\} \frac{b_c^a}{\Gamma(a_c)} e^{a \cdot v} e^{-b_c \exp(v)},$$

At each iteration, once samples of $\mu$ and $\sigma_i^2$ have been simulated by the tempering transition method, $\delta_i$, can be generated independently, $i = 1, \ldots, n$, conditional on the current values of $\mu$ and $\sigma_i^2$.

### 5.4 The Marginal Posterior Distribution of the Allocation Variables

One goal of the analysis is to estimate the posterior probability that an observation $y_i$ is an outlier. Let $p_i$ denote the posterior probability that $\delta_i = 1$. The marginal posterior distribution $p(\delta|y)$ of $\delta$ is given by

$$p(\delta|y) = \int p(\mu, \sigma^2, \delta, c|y) \, d\mu \, d\sigma^2 \, dc.$$
Substituting \( p(\mu, \sigma^2, \delta, c|y) \) from Expression (5.2) into the above equation and integrating \( c \) out, we obtain

\[
p(\delta|y) \propto \int p_0^n (1 - p_0)^{n - n_1} \prod_{\{i: \delta_i = 1\}} \frac{1}{\sqrt{\sigma^2 + \sigma^2_c}} \exp \left[ -\frac{(y_i - \mu)^2}{2(\sigma^2 + \sigma^2_c)} \right] \\
\prod_{\{i: \delta_i = 0\}} \frac{1}{\sigma_c} \exp \left[ -\frac{(y_i - \mu)^2}{2\sigma^2} \right] \exp \left[ -\frac{(\mu - \xi)^2}{2\sigma^2_c} \right]
\]

\[
\frac{b^\sigma_c}{\Gamma(a)} (\sigma^2_c)^{(a+1)} e^{-b/\sigma^2_c} d\mu d\sigma^2_c,
\]

where \( n_1 = \sum_{i=1}^n \delta_i \). Integrating \( \mu \) out in the above equation yields

\[
p(\delta|y) \propto \int \sqrt{\frac{\nu^2}{2}} \frac{\sigma^2}{\sigma^2 + \sigma^2_c}^{n-n_1} (\sigma^2 + \sigma^2_c)^{-\nu/2} (\sigma^2_c)^{-(a+1)/2} d\sigma^2_c
\]

\[
\exp \left\{ \frac{m^2}{\nu^2} - \frac{2b_c + \sum_{\{\delta_i = 0\}} y_i^2}{2\sigma^2_c} - \frac{\sum_{\{\delta_i = 1\}} y_i^2}{2(\sigma^2_c + \sigma^2)} - \frac{\xi^2}{\sigma^2_c} \right\} d\sigma^2_c,
\]

(5.3)

where

\[
v^2 = (n_1/\sigma^2 + \sigma^2_c) + (n - n_1)/\sigma^2_c + 1/\sigma^2_c^{-1},
\]

\[
m = (v^2)^{-1} \left[ \sum_{\{\delta_i = 1\}} y_i^2 \sigma^2_c + \frac{\sum_{\{\delta_i = 0\}} y_i^2}{\sigma^2_c} \sigma^2_c + \frac{\xi^2}{\sigma^2_c} \right].
\]

The marginal distribution \( p(\delta|y) \) in Expression (5.3) is an integral only with respect to \( \sigma^2_c \). Thus, numerical quadratures can be used to approximate the exact unnormalized values of \( p(\delta|y) \). For a data set of size \( n \), however, the total number of possible \( \delta \)s is \( 2^n \). Thus, when \( n \) is large, it is impossible to evaluate the exact values of \( p_i \) due to the computational effort that would be required to normalize the values of \( p(\delta|y) \).

We can also use Monte Carlo methods to estimate \( p_i \) from the MCMC draws of \( \delta_i \). Suppose that \( \delta_i^{(1)} , \ldots , \delta_i^{(N)} \) are the simulated samples for \( \delta_i \). Then \( p_i \) can be estimated as

\[
\hat{p}_i = \frac{1}{N} \sum_{j=1}^N I(\delta_i^{(j)} = 1),
\]

68
where $I(\cdot)$ is a 0-1 indicator. Note that this estimation method requires that the simulated samples of the parameters truly reflect the posterior distribution. Otherwise, the estimates will be poor. The standard errors of the estimates can be estimated by using the overlapping batch statistics method of Chen and Schmeiser (1993). Specifically, the variance of the point estimator $\hat{p}_i$ can be estimated by the overlapping batch statistics estimator

$$\hat{V}(m) = \frac{m}{N - m} \sum_{j=1}^{N-m+1} (\hat{p}_{i,j} - \hat{p}_i)^2 \left( \frac{N}{N-m+1} \right),$$

where $\hat{p}_{i,j}$ is defined analogously to $\hat{p}_i$ but is a function of only $\delta_i^{(j)}, \ldots, \delta_i^{(j+m-1)}$, the data in the $j^{th}$ batch of size $m$. Chen and Schmeiser (1993) note that in many situations a reasonable choice of $m$ is such that $10 \leq \frac{N}{m} \leq 20$.

### 5.5 Example

I fit the outlier accommodation model (5.1) to the data set considered by Peruggia and Santner (1998). The data set consists of 10 observations, which are shown in Figure 5.1. The first five observations (called Group A) were drawn from a $N(-1, 1)$ distribution and the remaining five observations (called Group B) were generated from a $N(1, 0.01^2)$ distribution. There are two reasonable explanations for the data set, which may correspond to two posterior regions:

Region 1. There are no outliers, i.e., $\delta_1 = 0$. This implies that the measurement error variability is large enough to cover all the observations.

Region 2. The observations in Group A are outliers and the observations in Group B are not, i.e., $\delta_2 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$. This implies that the measurement error variability is small and the location shift is needed to account for the large variability in Group A.
To fit the outlier accommodation model (5.1) to this data set, I specified $\xi = 0$, $\sigma_\mu^2 = 100$, $\sigma_c^2 = 100$, $p_o = 0.1$, $a = 0.000001$ and $b = 0.000001$. The unknown parameters are $(\delta_1, \ldots, \delta_{10})$, $(c_1, \ldots, c_{10})$, $\mu$ and $\sigma_\varepsilon^2$. The exact values of the posterior probability that $\delta_i = 1$ are evaluated by numerical quadratures and recorded in the row labeled $p_i$ of Table 5.1.

To implement the Gibbs sampler, two starting points were considered, one starting point chosen from Region 1 and another from Region 2. Two Gibbs chains were then generated. The Gibbs samples are plotted in Figure 5.2, which shows that the Gibbs chain, once trapped in one region, fails to move to the other for the entire length of the simulation. Gibbs sampler estimates of $p_i$ are recorded in the rows of Table 5.1 labeled $\hat{p}_i^{G_1}$ and $\hat{p}_i^{G_2}$. They compare poorly with the exact values of $p_i$ and depend heavily on the posterior region that is being explored. For example, the chain with
the starting point drawn from Region 2 was trapped in this region and never moved to Region 1. Thus, the estimates are essentially 0 or 1.

To implement the MCMC tempering transition method of Chapter 2, the number \( m \) of intermediate steps was set at 17 and the temperature parameter of each step was chosen to be \( \gamma_j = 1/(2^j) \), \( j = 1, \ldots, m \). Two MCMC tempering chains were
generated with the same starting points used in the implementation of the Gibbs sampler. The MCMC samples are summarized in Figure 5.3. From the trace plots of $\mu$ we can find that two isolated posterior modal regions corresponding to $\delta_1$ and $\delta_2$ are clearly discovered. However, the tempering chains stay in Region 2 for many iterations before moving to Region 1. The posterior probabilities that an observation is an outlier were estimated from the two generated chains and are recorded in the rows of Table 5.1 labeled as $\hat{p}_i^{T1}$ and $\hat{p}_i^{T2}$. The estimates appear to be very poor.

Two possible isolated modal regions are identified by the MCMC tempering sampler. The simulated values within each identified modal region were used to build local Gaussian proposals for the parameter vector $(\mu, -\log(\sigma_2^2))$. These local proposals were assigned equal weights to create the mixture proposal used in Step (2) of the algorithm in Section 3.1. Two chains were generated with the same initial points used in the implementation of the tempering sampler. The simulated samples are summarized in Figure 5.4, which shows that this algorithm can transition easily between the two isolated modal regions and that the choice of starting point does not affect the output. The estimates of $p_i$ are recorded in the row labeled $\hat{p}_i^M$ of Table 5.1 and they appear to be very precise.

To implement the algorithm in Chapter 3.2, a model-based clustering method was first applied to identify the possible modal regions. This method separates $\{y_1, \ldots, y_5\}$ and $\{y_6, \ldots, y_{10}\}$ correctly. Thus, the configurations $\delta_1$, $\delta_2$ and $\delta_3 = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$ were taken as the preferred configurations. Gibbs samples were drawn for each preferred configuration. The procedure of Section 3.2.2 was used to generate candidate values for $\delta$, $\mu$ and $\sigma_2^2$. Candidate values for the location shift
Figure 5.3: Trace plots of the sampled parameters by the tempering transition method. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth).
Figure 5.4: Trace plots of the sampled parameters by the proposed algorithm in Chapter 3.1. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth).
$c_i$ were generated from their full conditionals, which are given by

$$c_i = \begin{cases} N(0, k_1 \sigma^2) & \text{if } \delta_i = 0 \\ N(y_i - \mu, k_2 \sigma^2) & \text{if } \delta_i = 1, \end{cases}$$

where $k_1$ and $k_2$ are set equal to 1. Through preliminary runs I discovered that the region corresponding to $\delta_3 = (0, 0, 0, 0, 1, 1, 1, 1, 1, 1)$ was never accepted. Thus, I removed $\delta_3$ from the list of preferred configurations. The tuning parameters $r_0$ and $r_1$ need to be set high to obtain a high acceptance probability of jumping from one posterior region to the other. To examine the performance of the algorithm for different settings of the tuning parameters $r_0$ and $r_1$, three sets of $(r_0, r_1)$ pairs were used with $(r_0 = 0.98, r_1 = 0.95)$, $(r_0 = 0.98, r_1 = 0.98)$ and $(r_0 = 0.95, r_1 = 0.95)$. Also, $p_{\text{MH}}$ was set at 0.1. Two MCMC chains were generated for each setting with the same initial points used in the implementation of the Gibbs sampler. Samples with $r_0 = r_1 = 0.95$ are summarized in Figure 5.5, which shows that the proposed method can transition effectively between the two isolated modal regions and that there is no effect due to the choice of the starting point on the output. The estimates of $p_i$ for each set of $(r_0, r_1)$ are recorded in the rows labeled $\hat{p}_i^H$ of Table 5.1. We can see that these estimates are quite precise.

Table 5.1 shows that the first five observations are identified as having high posterior outlier probabilities. The observations can then be divided into two groups: outlying cases and non-outlying cases. In an attempt to derive an assessment of across-the-board performance, the last column in Table 5.1 records the root mean squared errors (RMSE) of estimation of $p_i$ by each method for outlying cases and non-outlying cases. It is clear that the estimates by the new algorithms are very precise with small deviation from the true values. However, the estimates by the Gibbs sampler and the tempering transition method are, overall, quite different from the
Figure 5.5: Trace plots of the sampled parameters by the proposed algorithm in Chapter 3.2. The output is: (a) and (b) the last 40,000 iterations with the initial point in Region 1 (every fortieth), (c) and (d) the last 40,000 iterations with the initial point in Region 2 (every fortieth).
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y_i$</td>
<td>$p_i$</td>
<td>$\hat{p}_i^{G}$</td>
<td>$\hat{p}_i^{T}$</td>
<td>$\hat{p}_i^{M}$</td>
</tr>
<tr>
<td></td>
<td>-1.8036</td>
<td>0.4833</td>
<td>0.0486 (0.0014)</td>
<td>0.9438 (0.0310)</td>
<td>0.4790 (0.0235)</td>
</tr>
<tr>
<td></td>
<td>-0.0974</td>
<td>0.4657</td>
<td>0.0167 (0.0007)</td>
<td>0.9417 (0.0321)</td>
<td>0.4588 (0.0243)</td>
</tr>
<tr>
<td></td>
<td>-2.1559</td>
<td>0.5000</td>
<td>0.0782 (0.0018)</td>
<td>0.9460 (0.0296)</td>
<td>0.4938 (0.0243)</td>
</tr>
<tr>
<td></td>
<td>-0.8950</td>
<td>0.4680</td>
<td>0.0185 (0.0008)</td>
<td>0.9418 (0.0320)</td>
<td>0.4607 (0.0243)</td>
</tr>
<tr>
<td></td>
<td>-0.7698</td>
<td>0.4673</td>
<td>0.0178 (0.0006)</td>
<td>0.9419 (0.0320)</td>
<td>0.4614 (0.0244)</td>
</tr>
<tr>
<td></td>
<td>RMSE $\hat{p}_i^{G}$</td>
<td>0.4410</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 1 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{G}$</td>
<td>0.5233</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{T}$</td>
<td>0.0486 (0.0014)</td>
<td>0.9438 (0.0310)</td>
<td>0.4790 (0.0235)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0782 (0.0018)</td>
<td>0.9460 (0.0296)</td>
<td>0.4938 (0.0243)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0185 (0.0008)</td>
<td>0.9418 (0.0320)</td>
<td>0.4607 (0.0243)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0178 (0.0006)</td>
<td>0.9419 (0.0320)</td>
<td>0.4614 (0.0244)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0225 (0.0008)</td>
<td>0.0233 (0.0009)</td>
<td>0.0237 (0.0009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 0 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{T}$</td>
<td>0.0129 (0.0008)</td>
<td>0.0205 (0.0122)</td>
<td>0.0230 (0.0122)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0225 (0.0008)</td>
<td>0.0233 (0.0009)</td>
<td>0.0237 (0.0009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 1 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{T}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 1 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{T}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 1 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{T}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{M}$</td>
<td>0.0143 (0.0010)</td>
<td>0.0124 (0.0010)</td>
<td>0.0130 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\hat{p}_i^{H}$</td>
<td>0.0117 (0.0008)</td>
<td>0.0120 (0.0009)</td>
<td>0.0124 (0.0010)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\approx$ 1 for all $y_i$ (≈ 0) for all $y_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: True posterior outlier probabilities $p_i$, the estimates using the Gibbs sampler $\hat{p}_i^{G}$, the estimates using the MCMC tempering with two starting points $\hat{p}_i^{T_1}$ and $\hat{p}_i^{T_2}$, the estimates using our proposed algorithms $\hat{p}_i^{M}$ and $\hat{p}_i^{H}$. The estimates are based on 50000 iterations with first 10000 iterations as burn-in.

77
Table 5.2: Geometric means of ratios of true outlier posterior probabilities $p_i$ to the estimates $\hat{p}_i$ by each method for outlying cases and non-outlying cases.

<table>
<thead>
<tr>
<th>geometric mean of $\hat{p}_i/p_i$</th>
<th>Outlying Cases</th>
<th>Non-outlying Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_1$</td>
<td>0.0611</td>
<td>1.8349</td>
</tr>
<tr>
<td>$G_2$</td>
<td>2.0978</td>
<td>0</td>
</tr>
<tr>
<td>$T_1$</td>
<td>1.9783</td>
<td>0.1404</td>
</tr>
<tr>
<td>$T_2$</td>
<td>1.8716</td>
<td>0.2361</td>
</tr>
<tr>
<td>$M$</td>
<td>0.9871</td>
<td>1.0406</td>
</tr>
<tr>
<td>$H$ $(r_0, r_1) = (0.98, 0.95)$</td>
<td>1.0329</td>
<td>0.9497</td>
</tr>
<tr>
<td>$(0.98, 0.98)$</td>
<td>0.9997</td>
<td>0.9668</td>
</tr>
<tr>
<td>$(0.95, 0.95)$</td>
<td>1.002</td>
<td>0.9736</td>
</tr>
</tbody>
</table>

true values. Table 5.2 summarizes the geometric mean values of $\hat{p}_i/p_i$ by each method for outlying cases and non-outlying cases. A geometric mean value close to 1 implies that the estimates are precise. From this table, it is apparent that the estimates produced by the proposed algorithms are very reliable compared to those produced by the competing algorithms.
CHAPTER 6

APPLICATION TO AN OUTLIER ACCOMMODATION
BAYESIAN REGRESSION MODEL

In this chapter I apply the algorithms described in Chapter 3 to a Bayesian regression model and examine their performance by comparing their inferences for the model parameters and the posterior probability that a given observation \( y_i \) is an outlier with the inferences based on the output from the Gibbs sampler and the MCMC tempering transition method.

6.1 An Outlier Accommodation Bayesian Regression Model

The outlier accommodation Bayesian regression model is given by

\[
y_i = x_i' \beta + \delta_i c_i + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{6.1}
\]

where \( x_i = (1, x_i)' \) and the \( \epsilon_i \) are independent and identically distributed according to a \( N(0, \sigma_\epsilon^2) \) distribution. Assuming conditional independence, the remaining parts of the model are given by \( \beta = (\beta_0, \beta_1)' \sim N(0, \Psi) \), \( c_i \sim N(0, \sigma_c^2) \), for \( i = 1, \ldots, n \), \( \delta_i \sim \text{Bernoulli}(p_0) \), for \( i = 1, \ldots, n \), and \( \sigma_\epsilon^2 \sim \text{Inverse-Gamma}(a_\epsilon, b_\epsilon) \). Here, \( \delta_i \) is an indicator variable which equals either 0 or 1. The quantities \( \Psi, \sigma_c^2, p_0, a_\epsilon \) and \( b_\epsilon \) are constants and need to be specified. The parameters \( \{\delta_i\}_{i=1}^n, \{c_i\}_{i=1}^n, \beta \) and \( \sigma_\epsilon^2 \) are unknown.
This model is an extension of the outlier accommodation model discussed in Chapter 5 with \( \delta_i = 1 \) indicating that the observation \( y_i \) is an outlier. When \( \delta_i = 1 \), a term \( c_i \) is added to inflate the variance of \( y_i \) so that \( y_i \) can be accommodated. To identify whether an observation \( y_i \) is an outlier or not, we can look at the posterior probability that \( \delta_i = 1 \). If the probability is large, the observation \( y_i \) is identified as an outlier.

### 6.2 Full Conditional Distributions

To use the Gibbs sampler to generate samples for Model (6.1), a data augmentation strategy is applied by treating the \( \delta_i \) as missing and generating them together with the other model parameters. Before implementing the Gibbs sampler, it is necessary to obtain the full conditional distributions of the augmented set of parameters \((\beta, \sigma^2, \{\delta_i\}_{i=1}^n, \{c_i\}_{i=1}^n)\). Let \( \delta = (\delta_1, \ldots, \delta_n) \) and \( c = (c_1, \ldots, c_n) \). The complete-data posterior distribution satisfies

\[
p(\beta, \sigma^2, \delta, c | y) \propto p(y | \delta, c, \beta, \sigma^2) \ p(\delta) \ p(c) \ p(\beta) \ p(\sigma^2)
\]

\[
\propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left[ -\frac{(y_i - (x_i'\beta + \delta_ic_i))^2}{2\sigma^2_\epsilon} \right] \right\}
\]

\[
\prod_{i=1}^n \left\{ p_0^{I(\delta_i=1)} (1 - p_0)^{I(\delta_i=0)} \right\} \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma_c} \exp \left[ -\frac{c_i^2}{(2\sigma^2_c)} \right] \right\}
\]

\[
\frac{|\Psi|^{-1/2}}{2\pi} \exp \left(-\frac{1}{2}\beta'\Psi^{-1}\beta\right) \cdot \frac{b^{a_\epsilon}}{\Gamma(a_\epsilon)} \left(\sigma^2_\epsilon\right)^{-(a_\epsilon+1)} e^{-b_\epsilon/\sigma^2_\epsilon}.
\]  

(6.2)

For the full conditional distribution of \( \beta \) we have

\[
p(\beta | \delta, c, \sigma^2, y) \propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \exp \left[ -\frac{(y_i - (x_i'\beta + \delta_ic_i))^2}{2\sigma^2_\epsilon} \right] \right\}
\]

\[
\frac{|\Psi|^{-1/2}}{2\pi} \exp \left(-\frac{1}{2}\beta'\Psi^{-1}\beta\right)
\]

\[
\propto \frac{|V|^{-1/2}}{2\pi} \exp \left\{ -\frac{1}{2}(\beta - \hat{\beta})'V^{-1}(\beta - \hat{\beta}) \right\} \frac{\Psi^{-1/2}}{2\pi} \exp \left(-\frac{1}{2}\beta'\Psi^{-1}\beta\right),
\]

80
where \( \hat{\beta} = (X'X)^{-1}X'Z \) and \( V = \sigma^2_\varepsilon (X'X)^{-1} \) with \( X = (x_1, \ldots, x_n)' \) and \( Z = (y_1 - \delta_1 c_1, \ldots, y_n - \delta_n c_n)' \). Therefore, the full conditional of \( \beta \) is normal with mean \((\Psi^{-1} + V^{-1})^{-1}V^{-1}\hat{\beta} \) and variance covariance matrix \((\Psi^{-1} + V^{-1})^{-1}\).

Let \( c_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, c_n) \). The full conditional distribution of a shift \( c_i \) satisfies

\[
p(c_i|\beta, \sigma^2_\varepsilon, \delta, c_{-i}, y) \propto \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp \left[ -\frac{(y_i - (x_i'\beta + \delta_i c_i))^2}{2\sigma^2_\varepsilon} \right] \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp \left( -\frac{c_i^2}{2\sigma^2_\varepsilon} \right)
\]

\[
\times \left\{ \begin{array}{ll}
\frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp \left( -\frac{c_i^2}{2\sigma^2_\varepsilon} \right) & \text{if } \delta_i = 0 \\
\frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp \left( -\frac{(c_i - m_{c,i})^2}{2\sigma^2_\varepsilon} \right) & \text{if } \delta_i = 1
\end{array} \right.,
\]

where \( m_{c,i} = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \sigma^2_\varepsilon} (y_i - x_i'\beta) \) and \( v^2_{c,i} = \frac{\sigma^2_\varepsilon}{\sigma^2_\varepsilon + \sigma^2_\varepsilon} \). This shows that the full conditional distribution of \( c_i \) is the same as its prior distribution when \( \delta_i = 0 \). This is logical since there is no information available about the shift \( c_i \) when the observation \( y_i \) is identified as a non-outlier. When \( \delta_i = 1 \), the full conditional of \( c_i \) is normal with mean \( m_{c,i} \) and variance \( v^2_{c,i} \).

The full conditional of the error variance \( \sigma^2_\varepsilon \) satisfies

\[
p(\sigma^2_\varepsilon|\beta, \sigma^2_\varepsilon, \delta, c, y) \propto \frac{1}{\sigma^2_\varepsilon^n} \exp \left[ -\sum_{i=1}^n (y_i - (x_i'\beta + \delta_i c_i))^2 \right] \sigma^2_\varepsilon^{-(a_\varepsilon + 1)} e^{-b_\varepsilon/\sigma^2_\varepsilon}
\]

\[
\times \sigma^2_\varepsilon^{-(a_\varepsilon + n/2 + 1)} \exp \left\{ -\left[ b_\varepsilon + \frac{1}{2} \sum_{i=1}^n (y_i - x_i'\beta - \delta_i c_i)^2 \right] / \sigma^2_\varepsilon \right\},
\]

which is the kernel of an Inverse-Gamma distribution with shape \( a_\varepsilon + n/2 \) and scale \( b_\varepsilon + \frac{1}{2} \sum_{i=1}^n (y_i - x_i'\beta - \delta_i c_i)^2 \).

For the outlier indicators \( \delta_i \) we have

\[
p(\delta_i = 0|\beta, \sigma^2_\varepsilon, c, y) \propto \frac{1 - p_o}{\sigma_\varepsilon} \exp \left[ -\frac{(y_i - x_i'\beta)^2}{2\sigma^2_\varepsilon} \right],
\]

\[
p(\delta_i = 1|\beta, c, \sigma^2_\varepsilon, y) \propto \frac{p_o}{\sigma_\varepsilon} \exp \left[ -\frac{(y_i - x_i'\beta - c_i)^2}{2\sigma^2_\varepsilon} \right]
\]

81
6.3 Conditional Distributions of $\beta$ and Reparameterized $\sigma_\epsilon^2$

As in Section 4.3, in the implementation of the tempering transition method the outlier indicators $\delta_i$ and the shifts $c_i$ are not simulated together with the model parameters $\beta$ and $\sigma^2_\epsilon$. Also, $\sigma^2_\epsilon$ is reparameterized as $v_\epsilon = -\log(\sigma^2_\epsilon)$ and the parameters $\beta$ and $\sigma^2_\epsilon$ are simulated through tempering from their full conditionals $p(\beta|v_\epsilon, y)$ and $p(v_\epsilon|\beta, y)$ respectively.

To obtain the full conditionals, we first need to derive the joint posterior distribution of $(\beta, \sigma^2_\epsilon)$. From Model (6.1), we have

$$p(y_i|\beta, \sigma^2_\epsilon, c_i) = p_0 N(y_i; x'_i\beta + c_i, \sigma^2_\epsilon) + (1 - p_0) N(y_i; x'_i\beta, \sigma^2_\epsilon).$$

and

$$p(y_i|\beta, \sigma^2_\epsilon) = \int p(y_i|\beta, \sigma^2_\epsilon, c_i) p(c_i) dc_i$$

$$= p_0 N(y_i; x'_i\beta, \sigma^2_\epsilon + \sigma^2_\epsilon) + (1 - p_0) N(y_i; x'_i\beta, \sigma^2_\epsilon).$$

Then $p(\beta, \sigma^2_\epsilon|y)$ satisfies

$$p(\beta, \sigma^2_\epsilon|y) \propto \prod_{i=1}^n p(y_i|\beta, \sigma^2_\epsilon) p(\beta) p(\sigma^2_\epsilon)$$

$$\propto \prod_{i=1}^n \left\{ \frac{p_0}{\sqrt{\sigma^2_\epsilon + \sigma^2_\epsilon}} \exp \left[ -\frac{(y_i - x'_i\beta)^2}{2(\sigma^2_\epsilon + \sigma^2_\epsilon)} \right] + \frac{1 - p_0}{\sigma_\epsilon} \exp \left[ -\frac{(y_i - x'_i\beta)^2}{2\sigma^2_\epsilon} \right] \right\}$$

$$|\Psi|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \beta' \Psi^{-1} \beta \right) \frac{b^a}{\Gamma(a)} (\sigma^2_\epsilon)^{-a} e^{-b/\sigma^2_\epsilon}.$$

For the conditional distribution of $\beta$ we have

$$p(\beta|v_\epsilon, y) \propto \prod_{i=1}^n p(y_i|\beta, v_\epsilon) p(\beta)$$

$$\propto \prod_{i=1}^n \left\{ \frac{p_0}{\sqrt{e^{-v_\epsilon} + \sigma^2_\epsilon}} \exp \left[ -\frac{(y_i - x'_i\beta)^2}{2(e^{-v_\epsilon} + \sigma^2_\epsilon)} \right] + \frac{1 - p_0}{\sqrt{e^{-v_\epsilon}}} \exp \left[ -\frac{(y_i - x'_i\beta)^2}{2e^{-v_\epsilon}} \right] \right\}$$

$$|\Psi|^{-\frac{1}{2}} \exp \left( -\frac{1}{2} \beta' \Psi^{-1} \beta \right).$$
And the conditional distribution \( p(v|\beta, y) \) of \( v \) satisfies
\[
p(v|\beta, y) \propto \prod_{i=1}^{n} p(y_i|\beta, v) p(v)
\]
\[
\propto \prod_{i=1}^{n} \left\{ \frac{p_o}{\sqrt{e^{-v} + \sigma^2_c}} \exp \left[ -\frac{(y_i - x_i'\beta)^2}{2(e^{-v} + \sigma^2_c)} \right] + \frac{1 - p_o}{\sqrt{e^{-v}}} \exp \left[ -\frac{(y_i - x_i'\beta)^2}{2e^{-v}} \right] \right\}
\]
\[
\frac{b^a}{\Gamma(a)} \exp \left( -b e^{a} \right).
\]

Suppose \( \theta \) is a parameter in \( \{\beta_0, \beta_1, \sigma^2_c\} \). The posterior mean \( \mu_\theta \) of \( \theta \) can be written as
\[
\mu_\theta = \int \theta p(\beta, \sigma^2_c|y) d\beta d\sigma^2_c.
\]

The exact value of \( \mu_\theta \) can be obtained using a numerical quadrature method, but it may be hard to evaluate \( \mu_\theta \) numerically when \( \beta \) is high dimensional. For such situations Monte Carlo methods can be used to do the estimation. Suppose \( \theta^{(1)}, \ldots, \theta^{(N)} \) are the generated MCMC samples of \( \theta \), then the posterior mean of \( \theta \) can be approximated by
\[
\hat{\mu}_\theta = \frac{1}{N} \sum_{j=1}^{N} \theta^{(j)}.
\]

Note that the simulated samples of \( \theta \) must represent accurately the marginal posterior distribution of \( \theta \). Otherwise, the estimate \( \hat{\mu}_\theta \) will be poor.

### 6.4 The Marginal Posterior Distribution of the Allocation Variables

One goal of the analysis is to estimate the posterior probability that \( \delta_i = 1 \), i.e., the posterior probability that observation \( y_i \) is an outlier. Let \( p_i \) denote the outlyingness posterior probability. To obtain the exact values of \( p_i \), we first need to obtain the marginal posterior distribution \( p(\delta|y) \) of \( \delta \), which is given by
\[
p(\delta|y) = \int p(\beta, \sigma^2_c, \delta, c|y) d\beta d\sigma^2_c dc.
\]
Substituting expression (6.2) for \( p(\beta, \sigma^2, \delta, c|y) \) and integrating \( c \) out yields

\[
p(\delta|y) \propto \int p^n y (1 - p_0)^{n - n_1} \prod_{\{i: \delta_i = 1\}} \frac{1}{\sqrt{\sigma^2 + \sigma^2_c}} \exp \left[ -\frac{(y_i - \beta\sigma^2)^2}{2(\sigma^2 + \sigma^2_c)} \right] \prod_{\{i: \delta_i = 0\}} \frac{1}{\sigma} \exp \left[ -\frac{(y_i - \beta\sigma^2)^2}{2\sigma^2_c} \right] \left| \Psi^{-1/2} \right| \exp \left( -\frac{1}{2} \beta' \Psi^{-1} \beta \right) \frac{b_c^2}{\Gamma(a_c)} (\sigma^2_c)^{-(a_c + 1)} e^{-b_c/\sigma^2_c} d\beta d\sigma^2_c,
\]

where \( n_1 = \sum_{i=1}^n \delta_i \). Let \( A = \prod_{\{i: \delta_i = 1\}} \frac{1}{\sqrt{\sigma^2 + \sigma^2_c}} \exp \left[ -\frac{(y_i - \beta\sigma^2)^2}{2(\sigma^2 + \sigma^2_c)} \right] \) and \( B = \prod_{\{i: \delta_i = 0\}} \frac{1}{\sigma} \exp \left[ -\frac{(y_i - \beta\sigma^2)^2}{2\sigma^2_c} \right] \). Let \( X_1 \) denote the \((n_1 \times 2)\) matrix with rows equal to \( \{X_i\}_{\delta_i=1} \), \( X_0 \) the \(((n - n_1) \times 2)\) matrix with rows equal to \( \{X_i\}_{\delta_i=0} \), \( y_1 \) the column vector with elements equal to \( \{y_i\}_{\delta_i=1} \) and \( y_0 \) the column vector with elements equal to \( \{y_i\}_{\delta_i=0} \). Then, \( A \) and \( B \) can be rewritten as

\[
A = (\sigma^2 + \sigma^2_c)^{-n/2} \exp \left\{ \frac{\beta' X_1' X_1 \beta - 2y_1' X_1 \beta + y_1'y_1}{2(\sigma^2 + \sigma^2_c)} \right\},
\]

\[
B = (\sigma_c)^{-(n-n_1)} \exp \left\{ \frac{\beta' X_0' X_0 \beta - 2y_0' X_0 \beta + y_0'y_0}{2\sigma^2_c} \right\}.
\]

Thus, we have

\[
AB \exp \left( -\frac{1}{2} \beta' \Psi^{-1} \beta \right)
=
(\sigma^2 + \sigma^2_c)^{-n/2}(\sigma_c)^{-(n-n_1)} \exp \left\{ -\frac{1}{2} \left[ \beta' \left( \frac{X_1' X_1}{\sigma^2 + \sigma^2_c} + \frac{X_0' X_0}{\sigma^2_c} \right) + \Psi^{-1} \right] \beta - 2 \left( \frac{y_1' X_1}{\sigma^2 + \sigma^2_c} + \frac{y_0' X_0}{\sigma^2_c} \right) \beta + \frac{y_1'y_1}{\sigma^2 + \sigma^2_c} + \frac{y_0'y_0}{\sigma^2_c} \right\}
=
(\sigma^2 + \sigma^2_c)^{-n/2}(\sigma_c)^{-(n-n_1)} \exp \left\{ -\frac{1}{2} (\beta - \tilde{\beta})' \Psi^{-1} (\beta - \tilde{\beta}) + C \right\},
\]

84
where

\[ V^{-1} = X_1'X_1/(\sigma_e^2 + \sigma_c^2) + X_0'X_0/\sigma_c^2 + \Psi^{-1}, \]

\[ \tilde{\beta} = V(X_1'y_1/(\sigma_e^2 + \sigma_c^2) + X_0'y_0/\sigma_c^2), \]

\[ C = (\tilde{\beta}'V^{-1}\tilde{\beta} - y_1'y_1/(\sigma_e^2 + \sigma_c^2) + y_0'y_0/\sigma_c^2)/2. \]

Therefore, \( p(\delta|y) \) satisfies

\[ p(\delta|y) \propto \int u p_{01}^{n_1}(1 - p_0)^{n-n_1}(\sigma_e^2)^{-(a+1)}e^{-b_e/\sigma_e^2} d\sigma_e^2, \]

where \( u = |V|^{1/2}(\sigma_e^2 + \sigma_c^2)^{-n_1/2}(\sigma_e^2)^{-(n-n_1)/2} \exp(C) \). Note that \( p(\delta|y) \) is proportional to an integral with respect to \( \sigma_e^2 \) only, which can be evaluated numerically. All the possible \( \delta \)s need be considered when we compute the exact posterior probability that an observation \( y_i \) is an outlier. For a data set with \( n \) observations, the total number of the possible \( \delta \)s is \( 2^n \), which is very large for large \( n \). Because of computational limitations it is therefore impossible to complete the evaluation when the sample size is large. In such cases the value \( p_i \) can be estimated by Monte Carlo methods using the MCMC draws of \( \delta_i \). See Section 5.4 for details about the estimation.

### 6.5 Example

Justel and Peña (1996) show that the Gibbs sampler can fail to provide good estimates for linear regression models when a data set has many outliers that mask each other. To demonstrate this point, they used data from Rousseeuw (1984) that were generated from two groups: one group with 30 observations from the linear model \( y_i = 2 + x_i + \epsilon_i \) with error standard deviation 0.2, and another group with 20 observations following a bivariate normal distribution with mean \((7, 2)\) and covariance matrix \(0.5I\). Justel and Peña (1996) regarded the observations in the second group as
outliers. Since the evaluation of the exact posterior outlier probability using numerical quadratures is only possible for a small data set, I hereby considered a subset of size 10 generated from the data set. Seven data points were randomly selected from the first group and three points were randomly selected from the second group. The observations are recorded in the top two rows of Table 6.1 on page 94 and plotted in Figure 6.1.

To fit the Bayesian regression model to the above data, I specified the constants $\Psi$, $\sigma^2_c$, $a_e$ and $b_e$ as $\Psi = \text{diag}_2(100)$, $\sigma^2_c = 500$, $p_o = 0.3$, $a_e = 0.000001$ and $b_e = 0.000001$. The unknown parameters are $(\delta_1, \ldots, \delta_{10})$, $(c_1, \ldots, c_{10})$, $\beta$ and $\sigma^2_\varepsilon$. Model (6.1) was fit to the data set. A posterior regression line conditional on a specific $\delta$ can be obtained with intercept $E(\beta_0 | y, \delta)$ and slope $E(\beta_1 | y, \delta)$, which can be evaluated using numerical quadratures. The dashed posterior regression line in Figure 6.1 corresponds to the configuration $\delta_1 = (0, \ldots, 0)$ and the solid posterior regression line corresponds to the configuration $\delta_2 = (0, 0, 0, 0, 0, 0, 1, 1, 1)$ in which the last three observations are outliers. The exact values of the posterior outlier probabilities are calculated by using numerical quadratures and recorded in the row labeled $p_i$ of Table 6.1. The last three observations have large posterior outlier probabilities around 0.73.

I ran the Gibbs sampler and trace plots of the Gibbs samples are plotted in Figure 6.2. The plots show that the Gibbs sampler only explored the posterior region corresponding to $\delta_1 = (0, \ldots, 0)$. The Gibbs sampler estimates of $p_i$ are recorded in the row labeled $\hat{p}_i^G$. Comparing them with the exact values of $p_i$ shows that the Gibbs sampler estimates are very poor. In particular, the Gibbs sampler completely fails to recognize that the last three observations are unusual, because the estimated posterior probabilities of being an outlier are small. Further, as a consequence of
Figure 6.1: Scatter plot of the data.

Figure 6.2: Trace plots of the sampled parameters by the Gibbs sampler for the last 2,000 iterations. (a): the output of $\beta_0$; (b): the output of $\beta_1$; and (c): the output of $\sigma^2$.

extremely slow mixing, the use of different initial points does not solve the problem and does not improve the estimates.
To implement the MCMC tempering sampler of Chapter 2, I set the number $m$ of intermediate steps at 20, with the temperature parameter at each step was given by $\gamma_j = 1/(2^j)$, $j = 1, \ldots, m$. The MCMC samples are summarized in Figure 6.3, where two isolated posterior modal regions corresponding to $\delta_1$ and $\delta_2$ are clearly identified, especially based on the trace plot and histogram for $\beta_1$. (A scatter plot of $\beta_1$ vs. $\beta_0$—not shown—also clearly displays the two modal regions.) However, it is clear that the tempering sampler has difficulty transitioning between these two isolated modal regions. Two MCMC tempering chains were generated with different starting points, the first initial point in one of the two modal regions and the second point in the another region. The posterior probabilities of outlyingness were estimated from these two chains and are recorded in the rows of Table 6.1, labeled as $\hat{p}_{i1}^{T}$ and $\hat{p}_{i2}^{T}$. The differences between $\hat{p}_{i1}^{T}$ and $\hat{p}_{i2}^{T}$ indicate that the estimates are dependent upon the initial conditions due to the slow mixing of the tempering sampler.

To implement the algorithm of Section 3.1, possible isolated modal regions were first identified based on the output by the tempering transition method. As noted above, the trace plots of the tempering samples show us the existence of two possible isolated modal regions. Local proposals with a multivariate normal distribution for the parameter vector $(\beta_0, \beta_1, -\log(\sigma^2))$ were then built with center and variance covariance matrix estimated based on the values of the tempering samples within each identified modal region. Equal weights were assigned to these local proposals for building a mixture proposal, which is used in Step (2) of the algorithm in Section 3.1. The generated samples are summarized in Figure 6.4, which clearly indicates that this algorithm mixes well between the two isolated modal regions. However, the behavior of the trace plots in Figure 6.4 around iteration 600 suggests that mixing within one
Figure 6.3: Trace plots of the sampled parameters by the tempering transition method for the last 2,000 iterations. (a) and (d): the output of $\beta_0$; (b) and (e): the output of $\beta_1$; and (c) and (f): the output of $\sigma^2$. 
of the modal regions is slightly suboptimal. The use of heavier tailed components in the mixture proposal might alleviate this problem. The estimates of \( p_i \) are recorded in the row labeled \( \hat{p}^M_i \) of Table 6.1 and they appear to be quite precise.

To implement the algorithm of Section 3.2, a model-based clustering method, which was described in Section 3.2 and can be implemented using the R package \textit{MCLUST}, was first used to identify the possible isolated modal regions. As it turns out, the observations generated by the linear regression model are correctly separated from those generated by the bivariate normal distribution. Two important preferred configurations of the observations were then used, corresponding to the two modal regions. The first configuration is \( \delta_1 = (0, \ldots, 0) \), which means that there are no outliers. This corresponds to the dashed posterior regression line in Figure 6.1 and implies that the measurement error variability is large enough to accommodate all the observations. The second preferred configuration is \( \delta_2 = (0, 0, 0, 0, 0, 1, 1, 1) \),
i.e., the three observations from the bivariate normal distribution are outliers and
the observations from the linear regression model are not. This corresponds to the
solid posterior regression line in Figure 6.1 and implies that the measurement error
variability is small and the location shift is needed to account for the unusual nature
of the observations from the bivariate normal distribution. The procedure of Section
3.2.2 was used to generate candidates for $\beta$ and $\sigma^2$. Candidates for the location shifts
$c_i$ were generated from their full conditional distributions given by

$$c_i \sim \begin{cases} N(0, k_1\sigma^2_c) & \text{if } \delta_i = 0 \\ N(y_i - x_i'\beta, k_2\sigma^2_c) & \text{if } \delta_i = 1, \end{cases}$$

where $k_1$ and $k_2$ were set equal to 1. To examine how the tuning parameters $r_0$ and
$r_1$ may affect the algorithm’s performance three sets of $(r_0, r_1)$ values were chosen:
$(r_0 = 0.99, r_1 = 0.99)$, $(r_0 = 0.99, r_1 = 0.95)$ and $(r_0 = 0.95, r_1 = 0.95)$. The tuning
parameter $p_{\text{MH}}$ was set at 0.5. The samples with $r_0 = r_1 = 0.95$ are summarized
in Figure 6.5, showing fast mixing. The estimates of $p_i$ for each set of $(r_0, r_1)$ are
recorded in the rows labeled $\hat{p}_i^H$ of Table 6.1. We can see that these estimates are
also quite precise. The estimates for the last three observations with $r_0 = r_1 = 0.99$
may be a little better than those corresponding to the other two settings.

As shown in Table 6.1, the last three observations from the bivariate normal
distribution are identified as having high posterior outlier probabilities. Thus, the
data can be divided into two groups: outlying cases and non-outlying cases. Table
6.2 summarizes the information in Table 6.1 for these two groups, examining the root
mean squared errors (RMSE) of estimation and geometric mean values of $p_i/\hat{p}_i$ by
each method. We can see that the estimates by the new algorithms are very precise
with small values for the RMSEs and geometric mean values of $p_i/\hat{p}_i$ close to one.
The Gibbs estimates are very poor with large values for the RMSEs and geometric
mean values of \( p_i / \hat{p}_i \) far away from one. The tempering estimates are close to the true values. However, they are dependent on the initial points.

In addition, I examined the accuracy of the inferences for the model parameters \( \beta \) and \( \sigma^2 \). For each method (except the Gibbs sampler), I generated a long chain and each chain was subdivided into 10 batches with 60,000 iterations each after discarding the first 10,000 iterations. The true marginal posterior mean of each model parameter was evaluated using a numerical method. The point estimates were obtained by averaging the batch estimates and are summarized in Table 6.3. The root mean squared errors were computed based on the squared deviations of the batch estimates from the true posterior mean and are also summarized in Table 6.3. We can see that the point estimates based on our proposed algorithms are quite close to the exact values and exhibit high reliability with small root mean squared errors. However, the
tempering estimates are inaccurate and not reliable with relatively large root mean squared errors.
Table 6.1: True posterior outlier probabilities, $p_i$, estimated by: the Gibbs sampler, $\hat{p}_G^i$; MCMC tempering with two starting points, $\hat{p}_T^1_i$ and $\hat{p}_T^2_i$; our proposed mixture-based algorithm, $\hat{p}_M^i$; and our proposed hybrid algorithm, $\hat{p}_H^i$. Standard errors are given in parentheses. The estimates are based on 90,000 iterations after a burn-in of size 10,000.
<table>
<thead>
<tr>
<th>RMSE</th>
<th>Non-outlying Cases</th>
<th>Outlying Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>0.0732</td>
<td>0.6374</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.0073</td>
<td>0.0671</td>
</tr>
<tr>
<td>$T_2$</td>
<td>0.0073</td>
<td>0.0840</td>
</tr>
<tr>
<td>$M$</td>
<td>0.0020</td>
<td>0.0104</td>
</tr>
</tbody>
</table>

$H \begin{array}{l} (r_0, r_1) = (0.99, 0.99) \\ (0.99, 0.95) \\ (0.95, 0.95) \end{array}$

<table>
<thead>
<tr>
<th>geometric mean of $\hat{p}_i/p_i$</th>
<th>Non-outlying Cases</th>
<th>Outlying Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>1.9442</td>
<td>0.1244</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.8527</td>
<td>1.0922</td>
</tr>
<tr>
<td>$T_2$</td>
<td>1.1431</td>
<td>0.8847</td>
</tr>
<tr>
<td>$M$</td>
<td>1.0204</td>
<td>0.9857</td>
</tr>
</tbody>
</table>

$H \begin{array}{l} (r_0, r_1) = (0.99, 0.99) \\ (0.99, 0.95) \\ (0.95, 0.95) \end{array}$

Table 6.2: RMSEs and geometric means of ratios of true outlier posterior probabilities $p_i$ to the estimates $\hat{p}_i$ by each method for outlying cases and non-outlying cases.

<table>
<thead>
<tr>
<th>parameter</th>
<th>true</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>true</td>
<td>tempering</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>3.1063</td>
<td>3.0412 (0.3231)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.5771</td>
<td>0.6064 (0.1426)</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.7388</td>
<td>0.7058 (0.2391)</td>
</tr>
</tbody>
</table>

Table 6.3: The marginal posterior means of the model parameters and their estimates (with RMSEs).
Hierarchical Bayesian linear models are widely used to analyze repeated measures data from multiple subjects. Peruggia et al. (2004) propose outlier detection methods for a three-stage hierarchical Bayesian linear model. Stage I of the model involves a regression with curve-specific coefficients. Stage II of the model specifies distributions of the curve-specific regression coefficients used in Stage I and may contain additional regression variables. To identify deviations in measurement and regression coefficients, Peruggia et al. (2004) introduce two location-shift outlier indicators for each subject, one at each stage. The Stage I outlier indicator equals one when there exists measurement error and the Stage II outlier indicator equals one when a regression coefficient error exists.

In this chapter the algorithms proposed in Chapter 3 are implemented to fit the multi-stage hierarchical Bayesian regression model proposed by Peruggia et al. (2004) for outlier detection. At the same time, the Gibbs sampler and the tempering transition method are also applied to generate samples for the model. One goal of the analysis is to estimate stage II posterior outlier probabilities to examine if any of the
regression lines exhibit an unusual slope. The estimation is based on the samples generated by each method. The samples of several model parameters are also monitored to examine the mixing properties of these sampling methods.

7.1 An Outlier Accommodation Hierarchical Bayesian Linear Model of Repeated Measures Data

To detect Stage I and II location-shift outliers for repeated measures data, Perugia et al. (2004) propose to use the following multi-stage model

Stage I.

\[ y_i = x_i \beta_i + \delta_i^y c_i + \eta_i, \quad \text{for } i = 1, \ldots, I, \]

where \( I \) denotes the number of subjects, \( y_i = (y_{i,1}, \ldots, y_{i,J_i})' \) is the vector of measurements for the \( i \)-th subject, \( x_i \) is a \( J_i \times L \) design matrix, \( \beta_i \) is an \( L \times 1 \) vector of parameters, \( \delta_i^y \) is an indicator with value 0 or 1, \( c_i \) is a \( J_i \times 1 \) measurement-shift vector for the \( i \)-th subject, \( \eta_i = (\eta_1, \ldots, \eta_{J_i})' \) is the vector of measurement errors with

\[ \eta_{i,j} = \phi_1 \eta_{i,j-1} + \cdots + \phi_{\min(j,P)} \eta_{i,-\min(j,P)} + \varepsilon_{i,j}, \quad \text{for } j = 1, \ldots, J_i, \]

where \( \varepsilon_{i,j} \sim N(0, \sigma_{\varepsilon,j}^2) \), for \( i = 1, \ldots, I \) and \( j = 1, \ldots, P \), and \( \varepsilon_{i,j} \sim N(0, \sigma_{\varepsilon}^2) \), for \( i = 1, \ldots, I \) and \( j = P+1, \ldots, J_i \).

Stage II. Let \( \gamma = (\gamma_1', \ldots, \gamma_L')', \ d = (d_{11}, \ldots, d_{1L}, \ldots, d_{I1}, \ldots, d_{IL})' \), with \( d_i \) standing for a location shift in the regression coefficient \( \beta_i \), \( \delta^\beta = (\delta_1^\beta, \ldots, \delta_L^\beta) \), where \( \delta_i^\beta \) is an indicator with value 0 or 1, \( \sigma_{\beta,i}^2 = (\sigma_{\beta,1}^2, \ldots, \sigma_{\beta,L}^2)' \), and \( \phi = (\phi_1, \ldots, \phi_P) \).
Then,

\[ \beta_i|\gamma, d, \delta, \sigma_\beta^2 \sim N(z_i'\gamma + \delta_id, \sigma_\beta^2[l]), \quad \text{for } i = 1, \ldots, I \text{ and } l = 1, \ldots, L, \]

\[ c_{ij} \sim N(\mu_c, \sigma_c^2), \quad \text{for } i = 1, \ldots, I \text{ and } j = 1, \ldots, J_i, \]

\[ \delta_i^y|p_y \sim Ber(p_y), \quad \text{for } i = 1, \ldots, I, \]

\[ \phi_j \sim N(0, \sigma_\phi^2[j]), \quad \text{for } j = 1, \ldots, P, \]

\[ \sigma_c^2 \sim IG(a_c, b_c), \]

\[ \sigma_{\epsilon_j}^2 \sim IG(a_{\epsilon_j}, b_{\epsilon_j}), \quad \text{for } j = 1, \ldots, P. \]

The hyperparameters \( \mu_c, \sigma_c^2, a_c, b_c, \sigma_\phi^2[j], a_{\epsilon_j} \) and \( b_{\epsilon_j} \), for \( j = 1, \ldots, P \), are assumed to be known.

**Stage III.**

\[ \gamma_l \sim N_{K_i}(\mu_{\gamma}[l], \text{diag}(\sigma_{\gamma}^2[l, 1], \ldots, \sigma_{\gamma}^2[l, K_i])), \quad \text{for } l = 1, \ldots, L, \]

\[ d_it \sim N(\mu_d[l], \sigma_d^2[l]), \quad \text{for } i = 1, \ldots, I \text{ and } t = 1, \ldots, L, \]

\[ \delta_i^d|p_\beta \sim Ber(p_\beta), \quad \text{for } i = 1, \ldots, I, \]

\[ \sigma_{\beta l}^2 \sim IG(a_l, b_l), \quad \text{for } l = 1, \ldots, L, \]

where the hyperparameters \( \mu_{\gamma}[l], \mu_d[l], \sigma_d^2[l], a_l, \) and \( b_l \), for \( l = 1, \ldots, L, \) and \( \sigma_{\gamma}^2[l, k], \)

\( \sigma_{\beta}^2[l, k], \) for \( l = 1, \ldots, L \text{ and } k = 1, \ldots, K_l, \) are assumed to be known.

**Stage IV.**

\[ p_y \sim Beta(u_y, v_y), \quad p_\beta \sim Beta(u_\beta, v_\beta), \]

where the hyperparameters \( u_y, v_y, u_\beta \) and \( v_\beta \) are assumed to be known.

In the model above, a value of the parameter \( \delta_i^y \) equal to one means that the \( i \)-th subject has a Stage I deviation with one or more abnormal observations and the shift \( c_{ij} \) can be used to identify the anomalous observations. Similarly, a value of the
parameter \( \delta_i^\beta \) equal to one implies that the \( i \)-th curve has a Stage II deviation with one or more outlying regression coefficients and the shift \( d_{il} \) can reveal which regression coefficient(s) have an anomaly. A useful diagnostic is to examine the posterior probabilities \( p_i^y = P(\delta_i^y = 1|y) \) and \( p_i^\beta = P(\delta_i^\beta = 1|y) \). Curve \( i \) is identified as having a Stage I deviation when \( p_i^y \) is high; curve \( i \) is identified as having a Stage II outlier when \( p_i^\beta \) is high.

A simple and common case of the above model is the model with \( P = 1 \) and \( \beta_{il}|\gamma, d, \delta^\beta, \sigma_{\beta}^2 \sim N(\gamma_l + \delta_i^\beta d_{il}, \sigma_{\beta}^2[l]) \). The conditional distributions of the parameters for this model are derived in the following two sections.

### 7.2 Full Conditional Distributions

To implement the Gibbs sampler, the full conditional distributions of the unknown parameters need to be derived. Let \( \Delta_{i,j} = y_{i,j} - x_{i,j}\beta_i - \delta_i^\gamma c_{i,j} \). Then the likelihood for the \( i \)-th subject is given by

\[
p(y_i|\beta_i, c_i, \phi, \sigma_{\epsilon_1}^2, \sigma_{\epsilon}^2) = p(y_{i,1}|\beta_i, \delta_i^\gamma, c_{i,1}, \phi, \sigma_{\epsilon_1}^2) \prod_{j=2}^{J_i} p(y_{i,j}|\beta_i, \delta_i^\gamma, c_{i,j}, \phi, \sigma_{\epsilon}^2, y_{i,j-1})
\]

\[
= \frac{1}{\sqrt{2\pi}\sigma_{\epsilon_1}^2} \exp \left\{ -\frac{\Delta_{i,1}^2}{2\sigma_{\epsilon_1}^2} \right\} \prod_{j=2}^{J_i} \frac{1}{\sqrt{2\pi}\sigma_{\epsilon}^2} \exp \left\{ -\frac{(\Delta_{i,j} - \phi\Delta_{i,j-1})^2}{2\sigma_{\epsilon}^2} \right\}.
\]

Considering the high dimensionality of the unknown vector of parameters, in what follows I will use “\(|\cdot|\)” to denote conditioning on all the remaining parameters not appearing on the right of the conditioning bar. The full conditional distribution of
Thus, we have

\[
p(\beta | \cdot) \propto p(y_i | \beta, c_i, \phi, \sigma^2_{\varepsilon}, \sigma^2_\varepsilon) p(\beta)
\]

\[
= \frac{1}{\sqrt{2\pi \sigma^2_{\varepsilon}}} \exp \left\{ - \frac{\Delta^2_{i,1}}{2\sigma^2_{\varepsilon}} \right\} \prod_{j=2}^{J_i} \frac{1}{\sqrt{2\pi \sigma^2_\varepsilon}} \exp \left\{ - \frac{(\Delta_{i,j} - \phi \Delta_{i,j-1})^2}{2\sigma^2_\varepsilon} \right\}
\]

\[
\prod_{l=1}^{L} \frac{1}{\sqrt{2\pi \sigma^2_\beta[l]}} \exp \left\{ - \frac{(\beta_{il} - \gamma_l - \delta^2_{\beta} d_{il})^2}{2\sigma^2_\beta[l]} \right\}.
\]

Let \( \Sigma = \text{diag}(\sigma^2_\beta[1], \ldots, \sigma^2_\beta[L]) \), \( u_{x,i} = ((x_{i,2} - \phi x_{i,1}), \ldots, (x_{i,J_i} - \phi x_{i,J_i-1}))' \) and \( u_{y,i} = (y_{i,2} - \delta^y c_{i,2} - \phi(y_{i,1} - \delta^y c_{i,1}), \ldots, y_{i,J_i} - \delta^y c_{i,J_i} - \phi(y_{i,J_i-1} - \delta^y c_{i,J_i-1}))' \). Then, from the above equation we can easily obtain that the full conditional distribution of \( \beta_i \) is \( N_L(\bar{\beta}_i, V) \), where \( V^{-1} = \Sigma^{-1} + x'_{i} x_{i}/\sigma^2_{\varepsilon} + u'_{x,i} u_{x,i}/\sigma^2_\varepsilon \) and \( \bar{\beta}_i = V[\Sigma^{-1} (\gamma + \delta^2_{\beta} d_{i}) + x'_{i} (y_{i,1} - \delta^y c_{i,1})/\sigma^2_{\varepsilon} + u'_{x,i} u_{y,i}/\sigma^2_\varepsilon] \).

The full conditional distribution of \( \delta^y_i \) satisfies

\[
p(\delta^y_i | \cdot) \propto p(y_i | \beta, c_i, \phi, \sigma^2_{\varepsilon}, \sigma^2_\varepsilon) p(\delta^y_i | p_y)
\]

\[
= \frac{1}{\sqrt{2\pi \sigma^2_{\varepsilon}}} \exp \left\{ - \frac{\Delta^2_{i,1}}{2\sigma^2_{\varepsilon}} \right\} \prod_{j=2}^{J_i} \frac{1}{\sqrt{2\pi \sigma^2_\varepsilon}} \exp \left\{ - \frac{(\Delta_{i,j} - \phi \Delta_{i,j-1})^2}{2\sigma^2_\varepsilon} \right\}
\]

\[
p_y^{I(\delta^y_i=1)} (1 - p_y)^{I(\delta^y_i=0)}.
\]

Thus, we have

\[
p(\delta^y_i = 1 | \cdot) \propto \frac{p_y}{\sigma_{\varepsilon}} \exp \left\{ - \frac{(y_{i,1} - x_{i,1} \beta_i - c_{i,1})^2}{2\sigma^2_{\varepsilon}} \right\}
\]

\[
\frac{1}{\sigma_\varepsilon} \exp \left\{ - \frac{\sum_{j=2}^{J_i} [y_{i,j} - x_{i,j} \beta_i - c_{i,j} - \phi(y_{i,j-1} - x_{i,j-1} \beta_i - c_{i,j-1})]^2}{2\sigma^2_{\varepsilon}} \right\},
\]

and

\[
p(\delta^y_i = 0 | \cdot) \propto \frac{1 - p_y}{\sigma_{\varepsilon}} \exp \left\{ - \frac{(y_{i,1} - x_{i,1} \beta_i)^2}{2\sigma^2_{\varepsilon}} \right\}
\]

\[
\frac{1}{\sigma_\varepsilon} \exp \left\{ - \frac{\sum_{j=2}^{J_i} [y_{i,j} - x_{i,j} \beta_i - \phi(y_{i,j-1} - x_{i,j-1} \beta_i)]^2}{2\sigma^2_{\varepsilon}} \right\}.
\]
For $c_{i,1}$ we have

$$p(c_{i,1}|\cdot) \propto \frac{1}{\sqrt{2\pi}\sigma_{\varepsilon_1}} \exp \left[ -\frac{(y_{i,1} - x_{i,1}\beta_i - \delta_1^y c_{i,1})^2}{2\sigma_{\varepsilon_1}^2} \right] \frac{1}{\sqrt{2\pi}\sigma_c} \exp \left( -\frac{(c_{i,1} - \mu_c)^2}{2\sigma_c^2} \right)$$

$$\propto \frac{1}{\sqrt{2\pi}\sigma_c} \exp \left[ -\frac{(c_{i,1} - \mu_c)^2}{(2\sigma_c^2)} \right] \text{ if } \delta_1^y = 0,$$

$$\frac{1}{\sqrt{2\pi}\sigma_c} \exp \left[ -\frac{(c_{i,1} - m_{i,1})^2}{(2v_{c,1}^2)} \right] \text{ if } \delta_1^y = 1,$$

where $m_{i,1} = \left[ \sigma_c^2(y_{i,1} - x_{i,1}\beta_i) + \sigma_{\varepsilon_1}^2 \mu_c \right] / (\sigma_c^2 + \sigma_{\varepsilon_1}^2)$ and $v_{c,1}^2 = (1/\sigma_{\varepsilon_1}^2 + 1/\sigma_c^2)^{-1}$.

Similarly, for $c_{i,j}$ we have

$$p(c_{i,j}|\cdot) \propto \frac{1}{\sqrt{2\pi}\sigma_c} \exp \left[ -\frac{(c_{i,j} - \mu_c)^2}{(2\sigma_c^2)} \right] \text{ if } \delta_i^y = 0,$$

$$\frac{1}{\sqrt{2\pi}\sigma_c} \exp \left[ -\frac{(c_{i,j} - m_{i,j})^2}{(2v_c^2)} \right] \text{ if } \delta_i^y = 1,$$

where $m_{i,j} = \left[ \sigma_c^2(y_{i,j} - x_{i,j}\beta_i - \phi(y_{i,j-1} - x_{i,j-1}\beta_i - c_{i,j-1})) + \sigma_c^2 \mu_c \right] / (\sigma_c^2 + \sigma_c^2)$ and $v_c^2 = (1/\sigma_c^2 + 1/\sigma_c^2)^{-1}$.

The full conditional distribution of $\phi$ satisfies

$$p(\phi|\cdot) \propto \exp \left\{ -\frac{\sum_{i=1}^I \sum_{j=2}^{J_i} (\Delta_{i,j} - \phi \Delta_{i,j-1})^2}{2\sigma_\phi^2} \right\} \exp \left\{ -\frac{\phi^2}{2\sigma_\phi^2} \right\}$$

$$\propto \exp \left\{ -\frac{\phi^2}{2\sigma_\phi^2} \right\}.$$

Thus, the full conditional distribution of $\phi$ is $N(m_\phi, V_\phi^2)$, where

$$m_\phi = \frac{\sigma_\phi^2}{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1} \Delta_{i,j}} \left( \frac{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1} \Delta_{i,j}}{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1} \Delta_{i,j}} \right),$$

$$\sigma_\phi^2 = \frac{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1}^2 + \sigma_\phi^2}{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1}^2},$$

$$V_\phi^2 = \frac{1}{\sum_{i=1}^I \sum_{j=2}^{J_i} \Delta_{i,j-1}^2 / \sigma_\phi^2 + 1/\sigma_\phi^2}.$$
The full conditional distribution of $\sigma_{\varepsilon_1}^2$ satisfies
\[
p(\sigma_{\varepsilon_1}^2 | \cdot) \propto \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi \sigma_{\varepsilon_1}^2}} \exp \left\{ -\frac{(y_{i,1} - x_{i,1} \beta_i - \delta_i \varepsilon_{i,1})^2}{2\sigma_{\varepsilon_1}^2} \right\} \left( \sigma_{\varepsilon_1}^2 \right)^{(a_{\varepsilon_1} + 1)/2 + 1} \exp \left\{ -\frac{\sum_{i=1}^{I}(y_{i,1} - x_{i,1} \beta_i - \delta_i \varepsilon_{i,1})^2}{2 \sigma_{\varepsilon_1}^2} \right\}.
\]
Thus, the full conditional distribution of $\sigma_{\varepsilon_1}^2$ is Inverse-Gamma with shape $a_{\varepsilon_1} + I/2$ and scale $\sum_{i=1}^{I}(y_{i,1} - x_{i,1} \beta_i - \delta_i \varepsilon_{i,1})^2/2 + b_{\varepsilon_1}$.

For $\sigma_{\varepsilon}^2$ we have
\[
p(\sigma_{\varepsilon}^2 | \cdot) \propto \prod_{i=1}^{I} \frac{1}{\sigma_{\varepsilon_i}^{-1}} \exp \left\{ -\frac{\sum_{j=2}^{J_i}(\Delta_{i,j} - \phi \Delta_{i,j-1})^2}{2\sigma_{\varepsilon_i}^2} \right\} \left( \sigma_{\varepsilon_i}^2 \right)^{(a_{\varepsilon} + (\sum_{i=1}^{I} J_i - I)/2 + 1)} \exp \left\{ -\frac{\sum_{i=1}^{I} \sum_{j=2}^{J_i}(\Delta_{i,j} - \phi \Delta_{i,j-1})^2}{2 \sigma_{\varepsilon_i}^2} \right\}.
\]
Thus, the full conditional distribution of $\sigma_{\varepsilon}^2$ is Inverse-Gamma with shape $a_{\varepsilon} + (\sum_{i=1}^{I} J_i - I)/2$ and scale $\sum_{i=1}^{I} \sum_{j=2}^{J_i}(\Delta_{i,j} - \phi \Delta_{i,j-1})^2/2 + b_{\varepsilon}$. The full conditional distribution of $\gamma_l$ can be written as
\[
p(\gamma_l | \cdot) \propto \prod_{i=1}^{I} \exp \left\{ -\frac{(\beta_d - \gamma_l - \delta_i \varepsilon_{i,l})^2}{2\sigma_{\beta}^2[l]} \right\} \exp \left\{ -\frac{\gamma_l \mu_{\gamma_l[l]} - \gamma_l^2}{2\sigma_{\gamma_l}^2[l]} \right\}
\]
\[
\propto \frac{1}{v_{\gamma_l}} \exp \left\{ -\frac{(\gamma_l - m_{\gamma_l})^2}{2v_{\gamma_l}^2} \right\},
\]
where $m_{\gamma_l} = \frac{\sigma_{\beta}^2[l] \sum_{i=1}^{I}(\beta_d - \delta_i \varepsilon_{i,l}) + \sigma_{\gamma_l}^2[l] \mu_{\gamma_l[l]}}{\sigma_{\beta}^2[l] + \sigma_{\gamma_l}^2[l]}$ and $v_{\gamma_l}^{-2} = \frac{1}{\sigma_{\beta}^2[l]} + \frac{1}{\sigma_{\gamma_l}^2[l]}$. Thus, the full conditional distribution of $\gamma_l$ is $N(m_{\gamma_l}, v_{\gamma_l}^2)$.

For $\delta_{\beta}^{(i)}$ we have
\[
p(\delta_{\beta}^{(i)} | \cdot) \propto \prod_{l=1}^{L} \exp \left\{ -\frac{(\beta_d - \gamma_l - \delta_{\beta}^{(i)} d_{i,l})^2}{2\sigma_{\beta}^2[l]} \right\} p_{\beta}^{\{\delta_{\beta}^{(i)} = 1\}} (1 - p_{\beta})^{I\{\delta_{\beta}^{(i)} = 0\}},
\]
which gives us
\[
p(\delta_{\beta}^{(i)} = 1 | \cdot) \propto p_{\beta} \exp \left\{ -\sum_{l=1}^{L} \frac{(\beta_d - \gamma_l - d_{i,l})^2}{2\sigma_{\beta}^2[l]} \right\},
\]
\[
p(\delta_{\beta}^{(i)} = 0 | \cdot) \propto (1 - p_{\beta}) \exp \left\{ -\sum_{l=1}^{L} \frac{(\beta_d - \gamma_l)^2}{2\sigma_{\beta}^2[l]} \right\}.
\]

102
The full conditional distribution of $d_{il}$ satisfies

$$p(d_{il} | \cdot) \propto \exp \left[ - \frac{(\beta_{il} - \gamma_l - \delta_i^\beta d_{il})^2}{2\sigma_{\beta}^2[l]} \right] \exp \left[ - \frac{(d_{il} - \mu_{il}^l[l])^2}{2\sigma_{d}^2[l]} \right]$$

$$\propto \begin{cases} \frac{1}{\sqrt{2\pi\sigma_{\beta}^2[l]}} & \text{if } \delta_i^\beta = 0, \\ \frac{1}{\sqrt{2\pi\sigma_{d}^2[l]}} & \text{if } \delta_i^\beta = 1, \end{cases}$$

where $m_{il} = \frac{\sigma_{\beta}^2[l] \beta_{il} - \gamma_l + \sigma_{d}^2[l] \mu_{il}^l[l]}{\sigma_{\beta}^2[l] + \sigma_{d}^2[l]}$ and $v_{il}^2 = \frac{\sigma_{\beta}^2[l] \sigma_{d}^2[l]}{\sigma_{\beta}^2[l] + \sigma_{d}^2[l]}$.

The full conditional distribution of $\sigma_{\beta}^2[l]$ is given by

$$p(\sigma_{\beta}^2[l] | \cdot) \propto \prod_{i=1}^I \frac{1}{\sigma_{\beta}^2[l]} \exp \left[ - \frac{(\beta_{il} - \gamma_l - \delta_i^\beta d_{il})^2}{2\sigma_{\beta}^2[l]} \right] (\sigma_{\beta}^2[l])^{-(a_l+1)} e^{-b_l/\sigma_{\beta}^2[l]}$$

$$\propto (\sigma_{\beta}^2[l])^{-(a_l+1/2+1)} \exp \left\{ - \sum_{i=1}^I (\beta_{il} - \gamma_l - \delta_i^\beta d_{il})^2 \right\}.$$

Thus, the full conditional distribution of $\sigma_{\beta}^2[l]$ is Inverse-Gamma with shape $a_l + I/2$ and scale $\sum_{i=1}^I (\beta_{il} - \gamma_l - \delta_i^\beta d_{il})^2 / 2 + b_l$.

For $p_y$ we have

$$p(p_y | \cdot) \propto p_y^{\sum_{i=1}^I \delta_i^y - 1} (1 - p_y)^{\sum_{i=1}^I \delta_i^y = 0} p_y^{n_y-1} (1 - p_y)^{n_y-1}$$

$$\propto p_y^{n_y+u_y-1} (1 - p_y)^{I-n_y+v_y-1},$$

where $n_y = \#\{\delta_i^y = 1, i = 1, \ldots, I\}$. Thus, the full conditional distribution of $p_y$ is $\text{Beta}(n_y + u_y, I - n_y + v_y)$. Similarly, the full conditional distribution of $p_{\beta}$ is $\text{Beta}(n_{\beta} + u_{\beta}, I - n_{\beta} + v_{\beta})$, where $n_{\beta} = \#\{\delta_i^\beta = 1, i = 1, \ldots, I\}$.

### 7.3 Conditional Distributions of the Parameters without the Stage I and II Outlier Indicators $\delta_i^y$ and $\delta_i^\beta$

As seen in the previous chapters, for the tempering transition method the Stage I and II outlier indicators, $\delta_i^y$ and $\delta_i^\beta$, and the corresponding shifts, $c_i$ and $d_i$, are not
simulated together with the other parameters. Candidate values of the parameters are generated from their own full conditional distributions, which are derived as follows.

We have \( \eta_{i,j} = \phi_{i,j} - 1 + \varepsilon_{i,j} \), \( \varepsilon_{i,1} \sim N(0, \sigma_{\varepsilon_1}^2) \) and \( \varepsilon_{i,j} \sim N(0, \sigma_{\varepsilon_j}^2) \) for \( j = 2, \ldots, J_i \).

This leads to

\[
\begin{align*}
p(\eta_i) &= p(\eta_{i,1}) \prod_{j=2}^{J_i} p(\eta_{i,j}|\eta_{i,j-1}) \\
&\propto \frac{1}{\sigma_{\varepsilon_1}} \exp \left\{ -\frac{\eta_{i,1}^2}{\sigma_{\varepsilon_1}^2} \right\} \prod_{j=2}^{J_i} \frac{1}{\sigma_{\varepsilon_j}} \exp \left\{ -\frac{(\eta_{i,j} - \phi_{i,j-1})^2}{\sigma_{\varepsilon_j}^2} \right\} \\
&\propto \frac{1}{|\Sigma|} \exp \left\{ -\frac{1}{2} \eta_i^T \Sigma^{-1} \eta_i \right\},
\end{align*}
\]

where \( \Sigma_{11}^{-1} = \frac{1}{\sigma_{\varepsilon_1}^2} + \phi_{\varepsilon}^2 \), \( \Sigma_{jj}^{-1} = \frac{1 + \phi_{\varepsilon}^2}{\sigma_{\varepsilon_j}^2} \), for \( j = 2, \ldots, J_i - 1 \), \( \Sigma_{JJ_i}^{-1} = \frac{1}{\sigma_{\varepsilon_j}^2} \), and \( \Sigma_{jk}^{-1} = -\frac{\phi_{\varepsilon}}{\sigma_{\varepsilon_j}^2} \) for \( |j - k| = 1 \). Thus, \( \eta_i \sim N_{J_i}(0, \Sigma) \).

Also, we have

\[
y_i|\beta_i, \delta_i^y, c_i, \phi, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_j}^2 \sim N_{J_i}(x_i\beta_i + c_i, \Sigma). \tag{7.1}
\]

With \( P(\delta_i^y = 1|p_y) = p_y \), by integrating \( \delta_i^y \) out from Equation (7.1) it follows that

\[
y_i|\beta_i, p_y, c_i, \phi, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_j}^2 \sim p_y N_{J_i}(x_i\beta_i + c_i, \Sigma) + (1 - p_y) N_{J_i}(x_i\beta_i, \Sigma).
\]

Integrating \( c_i \) out yields

\[
y_i|\beta_i, p_y, \phi, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_j}^2 \sim p_y N_{J_i}(x_i\beta_i + \mu_c, \Sigma + \text{diag}(\sigma_c^2)) + (1 - p_y) N_{J_i}(x_i\beta_i, \Sigma).
\]

Similarly, we can obtain

\[
\beta_i|p_\beta, \gamma, \sigma_\beta^2 \sim p_\beta N_{L}(\gamma + \mu_d, \text{diag}(\sigma_{\beta}^2[1] + \sigma_d^2[1], \ldots, \sigma_{\beta}^2[L] + \sigma_d^2[L])) + (1 - p_\beta) N_{L}(\gamma, \text{diag}(\sigma_{\beta}^2[1], \ldots, \sigma_{\beta}^2[L])).
\]

Thus, the full conditional distribution of \( \beta_i \) satisfies

\[
p(\beta_i|\cdot) \propto p(y_i|\beta_i, p_y, \phi, \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_j}^2) p(\beta_i|p_\beta, \gamma, \sigma_\beta^2).
\]
For $\phi$ we have

$$p(\phi|\cdot) \propto \prod_{i=1}^{l} p(y_i|\beta_i, p_y, \phi, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_1}^2) \frac{1}{\sigma_{\phi}} \exp \left\{ -\frac{\phi^2}{2\sigma_{\phi}^2} \right\}.$$ 

The full conditional distribution of $\sigma_{\epsilon_1}^2$ satisfies

$$p(\sigma_{\epsilon_1}^2|\cdot) \propto \prod_{i=1}^{l} p(y_i|\beta_i, p_y, \phi, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_1}^2)(\sigma_{\epsilon_1}^2)^{-a_{\epsilon_1}+1}e^{-b_{\epsilon_1}/\sigma_{\epsilon_1}^2},$$

and the full conditional distribution of $\sigma_{\epsilon}^2$ is given by

$$p(\sigma_{\epsilon}^2|\cdot) \propto \prod_{i=1}^{l} p(y_i|\beta_i, p_y, \phi, \sigma_{\epsilon_1}^2, \sigma_{\epsilon_1}^2)(\sigma_{\epsilon}^2)^{-a_{\epsilon}+1}e^{-b_{\epsilon}/\sigma_{\epsilon}^2}.$$

For $\gamma$ we have

$$p(\gamma|\cdot) \propto \prod_{i=1}^{l} p(\beta_i|p_\beta, \gamma, \sigma_{\beta}^2)[\Sigma_\gamma]^{-1/2} \exp \left\{ -\frac{1}{2}(\gamma - \mu_\gamma)^\prime \Sigma_\gamma^{-1}(\gamma - \mu_\gamma) \right\},$$

where $\Sigma_\gamma = \text{diag}(\sigma_{\gamma_1}^2[1], \ldots, \sigma_{\gamma}^2[L])$.

The full conditional distribution of $\sigma_{\beta}^2$ is given by

$$p(\sigma_{\beta}^2|\cdot) \propto \prod_{i=1}^{l} p(\beta_i|p_\beta, \gamma, \sigma_{\beta}^2)\prod_{l=1}^{L}(\sigma_{\beta}^2[l])^{-a_l+1}e^{-b_l/\sigma_{\beta}^2[l]}.$$ 

Lastly, for $p_y$ and $p_\beta$ we have

$$p(p_y|\cdot) \propto \prod_{i=1}^{l} p(y_i|\beta_i, p_y, \phi, \sigma_{\epsilon_1}^2, \sigma_{\beta}^2)p_y^v(1-p_y)^{v_y-1},$$

and

$$p(p_\beta|\cdot) \propto \prod_{i=1}^{l} p(\beta_i|p_\beta, \gamma, \sigma_{\beta}^2)p_\beta^u(1-p_\beta)^{v_\beta-1}.$$ 

Note that, for the same reason mentioned in Section 4.3, we can use a log transformation for $\sigma_{\epsilon_1}^2$, $\sigma_{\epsilon}^2$ and $\sigma_{\beta}^2$ and a logit transformation for $p_y$ and $p_\beta$. 

105
7.4 Example

Here I consider a data set containing repeated measurements on nine handguns. The data were collected by Milek (1989) and analyzed in Cleveland (1994). The response is the average velocity of the cartridge over a distance of 12 feet and the predictor is the barrel length. The data are plotted in Figure 7.1. This figure shows that there exists some curvature in the responses for some of the handguns (for example, \(.30-30\) and \(.35REM\)). Following the suggestion by Cleveland, an inverse transformation is applied to the data set. Figure 7.2 plots the transformed data and the corresponding least-squares lines.

To fit the hierarchical Bayesian linear model of Section 7.1 to this data set, I assumed that there is no Stage I outlier, the Stage II outlier indicator is applied to the slope only and the measurement errors are independent. Thus, the Stage I model is given by:

\[
y_{i,j} = \beta_{i,1} + \beta_{i,2} x_{i,j} + \varepsilon_{i,j} \sim N(0, \sigma^2_e), \ i = 1, \ldots, I \text{ and } j = 1, \ldots, J_i.
\]

The Stage II model is given by:

\[
\beta_{i,1} \sim N(\gamma_1, \sigma^2_\beta[1]) \text{ and } \beta_{i,2} \sim N(\gamma_2 + \delta\beta d_i, \sigma^2_\beta[2]),
\]

\(i = 1, \ldots, I\). I specified \(a_e = 0.0001, b_e = 0.0001, \mu_\gamma = (0, 1.4)', \sigma^2_\gamma[1] = 10, \sigma^2_\gamma[2] = 1, \mu_d = 0, \sigma^2_d = 20, a_1 = a_2 = b_1 = b_2 = 0.0001, u_y = 1, v_y = 9, u_\beta = 2, \text{ and } v_\beta = 8.\)

7.4.1 Case 1

The plot of the least-squares line indicates that there are (at least) three reasonable explanations for the data, which may correspond to three posterior regions:

**Region 1.** There are no Stage II outliers, i.e., \(\delta_{i}^{\beta} = 0\) for all \(i\).

---

1 Reproduced from Cleveland (1994) with permission.
2 Reproduced from Cleveland (1994) with permission.
Figure 7.1: Scatter plot of the data.
Figure 7.2: (a): Scatter plot of the transformed data; (b): Least-squares lines.
**Region 2.** The handguns .22 LR, .30-30 and .35 REM are outliers and the other handguns are not outliers. Specifically, the least-squares line for .22 LR has a smaller slope and the lines for .30-30 and .35 REM appear to be much steeper than those of the other handguns.

**Region 3.** The handguns .30-30 and .35 REM are outliers and the others are not outliers.

Note that the least squares estimates of the slopes for the lines not identified as outliers in any of the three regions are all near 1.4. For this reason, 1.4 was chosen as a “null” value for the second component of the prior mean of $\mu_\gamma$. This prior specification is therefore, to some extent, data-driven.

I used the Gibbs sampler to generate samples of the parameters. Figure 7.3(a) plots the simulated samples of $\gamma_2$ and it appears that the Gibbs sampler can mix well with the exception of several iterations.

To implement the MCMC tempering transition method, the number $m$ of intermediate steps was set at 10 and the temperature parameter of each step was chosen to be $1/(2j)$ for $j = 1, \ldots, m$. Since we make the assumption of no Stage I outliers, only the parameters related to the Stage II outliers, $\gamma$, $\sigma_\beta^2$, and $p_\beta$, will change corresponding to different values of $\delta$. Thus, these parameters were updated via steps of the tempering transition method. The other parameters were updated via a random walk sampler. Figure 7.3(b) plots the generated MCMC samples of $\gamma_2$. The figure indicates that this method might not mix better than the Gibbs sampler in this case.

For the implementation of the algorithm of Section 3.1, due to the same reason as in the implementation of the tempering transition method, only the parameters related to the Stage II outliers, $\gamma$, $\sigma_\beta^2$, and $p_\beta$, were updated via steps involving a
mixture proposal. Gibbs samples were first generated for each of the preferred configurations corresponding to Region 1, Region 2 and Region 3. The simulated Gibbs samples for each preferred configuration were then used to build local Gaussian proposals for the parameter vector \((\gamma, -\log(\sigma_\beta^2), \log(p_\beta/(1-p_\beta)))\). These local proposals were assigned equal weight to create the mixture proposal. The candidate values of other parameters were generated using a random walk sampler. The generated samples of \(\gamma_2\) are plotted in Figure 7.3(c). It shows that this proposed algorithm can mix quite well.

For the implementation of the algorithm of Section 3.2, the candidate values of \(\beta_i, \delta_i^3, i = 1, \ldots, I, \sigma^2_\epsilon, \gamma\) and \(\sigma^2_\beta\) were generated using the procedure of Section 3.2.2. Candidate values of the shifts \(d_i, i = 1, \ldots, I\), were generated from the following proposal, which is motivated by the form of the full conditional distribution of \(d_i\),

\[
d_i = \begin{cases} N(0, \sigma^2_d) & \text{if } \delta_i^3 = 0, \\ N(\beta_{i,2} - \gamma_2, \sigma^2_\beta[2]) & \text{if } \delta_i^3 = 1. \end{cases}
\]

I used equal selection probability for each preferred configuration and set the probability of a Metropolis-Hasting step to \(p_{\text{MH}} = 0.5\), and the tuning parameters to \(r_0 = 0.99, r_1 = 0.99\). The generated samples of \(\gamma_2\) are plotted in Figure 7.3(d). This plot shows that this proposed algorithm can also mix well.

Overall, it seems that the proposed algorithms can mix better than the Gibbs sampler and the MCMC tempering transition method, but the evidence is not overwhelming.
Figure 7.3: Trace plots of the samples for $\gamma_2$ by each method based on the output for the last 10,000 iterations (every fifth). (a) the Gibb sampler; (b) the MCMC tempering transition method; (c) the proposed mixture-based algorithm; (d) the proposed hybrid algorithm.
7.4.2 Case 2

In this section I examine the performance of the algorithms after removing handgun .22 LR. This leaves the following two reasonable explanations for the reduced data set:

**Region 1.** There are no Stage II outliers, i.e., \( \delta_i^\beta = 0 \) for all \( i \).

**Region 2.** Handguns .30-30 and .35 REM are outliers and the others are not outliers.

The same procedure used in Case 1 was employed to implement each sampling method. For the implementation of the algorithm proposed in Section 3.2, two more chains were generated with \((r_0, r_1) = (0.99, 0.95)\) and \((r_0, r_1) = (0.95, 0.95)\). The samples of \( \gamma_2 \) generated by this algorithm with \((r_0, r_1) = (0.99, 0.99)\) are plotted in Figure 7.4, along with the samples generated by the other methods. This plot shows that it takes a while for the Gibbs sampler to move from one posterior region to another and that for the tempering transition method it is even more difficult to transition between the two posterior regions. The proposed algorithms perform quite well with regard to transitioning between the two isolated posterior regions, especially the algorithm in Section 3.2.

The estimates of \( p_i^\beta \) by each method are recorded in Table 7.1. The row labeled \( \hat{p}_i^{\beta,G} \) is for the Gibbs sampler, \( \hat{p}_i^{\beta,T} \) for the tempering transition method, \( \hat{p}_i^{\beta,M} \) for the proposed mixture-based algorithm, \( \hat{p}_i^{\beta,H_a} \), \( \hat{p}_i^{\beta,H_b} \) and \( \hat{p}_i^{\beta,H_c} \) for the proposed algorithm in Section 3.2 with \((r_0, r_1) = (0.99, 0.99)\), \((r_0, r_1) = (0.99, 0.95)\) and \((r_0, r_1) = (0.95, 0.95)\) respectively. This table clearly shows that the handguns .30-30 and .35 REM may have outlying slopes since all the estimates of \( \hat{p}_i^\beta \) for these two handguns are high with
a value around 0.42. It appears that the estimates \( \hat{p}_{i}^{\beta,T} \) for .30-30 and .35 REM are much higher than the estimates obtained by the other methods and their large standard errors reflect the difficulty of the tempering method to transition between the two posterior regions. The estimates by the Gibbs sampler are slightly higher than the estimates by the proposed algorithms and the corresponding standard errors are larger than those of the estimates obtained by the algorithm proposed in Section 3.2. Overall, the performance of the algorithm proposed in Section 3.2 appears to be superior both in terms of mixing speed and estimation accuracy.
Figure 7.4: Trace plots of the samples for $\gamma_2$ by each method based on the output for the last 10,000 iterations (every fifth). (a) the Gibb sampler; (b) the MCMC tempering transition method; (c) the proposed mixture-based algorithm; (d) the proposed hybrid algorithm.
<table>
<thead>
<tr>
<th>Type</th>
<th>.223 REM 2</th>
<th>.223 REM 1</th>
<th>7mm BR</th>
<th>.30-30</th>
<th>.357 MAG</th>
<th>.35 REM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{p}_i^{\beta,G}$</td>
<td>0.0168 (0.0008)</td>
<td>0.0196 (0.0011)</td>
<td>0.0182 (0.0006)</td>
<td>0.4454 (0.0215)</td>
<td>0.0698 (0.0067)</td>
<td>0.4426 (0.0228)</td>
</tr>
<tr>
<td>$\hat{p}_i^{\beta,T}$</td>
<td>0.0155 (0.0013)</td>
<td>0.0155 (0.0022)</td>
<td>0.0155 (0.0009)</td>
<td>0.5511 (0.0397)</td>
<td>0.1117 (0.0147)</td>
<td>0.5457 (0.0403)</td>
</tr>
<tr>
<td>$\hat{p}_i^{\beta,M}$</td>
<td>0.0212 (0.0007)</td>
<td>0.0191 (0.0013)</td>
<td>0.0191 (0.0034)</td>
<td>0.4324 (0.0248)</td>
<td>0.0814 (0.0154)</td>
<td>0.4254 (0.0253)</td>
</tr>
<tr>
<td>$\hat{p}_i^{\beta,H_a}$</td>
<td>0.0208 (0.0010)</td>
<td>0.0174 (0.0014)</td>
<td>0.0174 (0.0011)</td>
<td>0.4298 (0.0095)</td>
<td>0.0473 (0.0038)</td>
<td>0.4273 (0.0098)</td>
</tr>
<tr>
<td>$\hat{p}_i^{\beta,H_b}$</td>
<td>0.0195 (0.0011)</td>
<td>0.0188 (0.0013)</td>
<td>0.0188 (0.0017)</td>
<td>0.4199 (0.0088)</td>
<td>0.0420 (0.0029)</td>
<td>0.4188 (0.0090)</td>
</tr>
<tr>
<td>$\hat{p}_i^{\beta,H_c}$</td>
<td>0.0194 (0.0012)</td>
<td>0.0159 (0.0012)</td>
<td>0.0159 (0.0009)</td>
<td>0.4389 (0.0097)</td>
<td>0.0448 (0.0044)</td>
<td>0.4296 (0.0100)</td>
</tr>
</tbody>
</table>

Table 7.1: Estimates of $p_i^\beta$: $\hat{p}_i^{\beta,G}$ by the Gibbs sampler, $\hat{p}_i^{\beta,T}$ by the tempering transition method, and $\hat{p}_i^{\beta,M}$ by the proposed mixture-based algorithm, $\hat{p}_i^{\beta,H_a}$, $\hat{p}_i^{\beta,H_b}$ and $\hat{p}_i^{\beta,H_c}$ by the proposed hybrid algorithms. Standard errors are given in parentheses. The estimates are based on 45,000 iterations after a burn-in of size 5,000 iterations.
8.1 Conclusions

In this dissertation, two MCMC algorithms are developed based on an initial identification of possible isolated modes of a posterior distribution. These algorithms are motivated by the fact that most MCMC methods have difficulties in transitioning between isolated posterior regions when the posterior distribution has several isolated modes.

The proposed algorithms are described in Chapter 3. These algorithms are then implemented in the next four chapters to generate samples from the posterior distributions for four different models: a Bayesian univariate normal mixture model; a Bayesian univariate outlier accommodation model; a Bayesian linear regression model; and a hierarchical Bayesian regression model for repeated measures data. Direct comparisons are made against established algorithms including the Gibbs sampler and the MCMC tempering algorithm. In the examples in Chapters 4, 5 and 6, exact inferences can be obtained by direct numerical integration. In the example of Chapter 7, exact inferences cannot be obtained and use of the MCMC technology becomes really necessary. For all these cases the proposed algorithms outperform the standard competing algorithms in terms of mixing properties and inferential accuracy. MCMC
tempering can often explore the isolated modal regions. However, its mixing properties are inferior to those of the proposed algorithms. For the Gibbs sampler it is difficult to transition between the isolated modal regions. The examples in Chapters 5 and 6, in particular, indicate that this sampler can get trapped in local modal regions for long stretches, leading to inferential estimates that are highly unreliable.

8.1.1 Choice of the Tuning Parameters

To implement the algorithm proposed in Section 3.2, we need to specify the vector of selection probabilities for the preferred configurations $p_δ$, the probability of a Metropolis-Hasting step $p_{MH}$, and two kernel-specific parameter values $r_0$ and $r_1$. To obtain optimal values for these tuning parameters, we can first assign equal selection probabilities to the preferred configurations and assign reasonable values to $p_{MH}$ and to the $(r_0, r_1)$ pair. The parameters $r_0$ and $r_1$ should typically be assigned large values to facilitate transitioning between isolated modal regions. Then, preliminary MCMC chains are generated for each set of $p_{MH}$ and $(r_0, r_1)$ values. For each simulated chain, the acceptance rates corresponding to each preferred configuration are recorded. A larger acceptance rate implies that the corresponding posterior region may have larger posterior probability. Accordingly, more selection weight should be put on this configuration. Thus, we can adjust the selection probabilities to match the recorded acceptance rates and simulate a final chain for each set of $p_{MH}$ and $(r_0, r_1)$ values. Both the preliminary and final chains are then used to compare the algorithm’s performance based on different sets of tuning parameters.

To reduce chance variability in the comparisons, long chains are generated for each set of tuning parameters and subdivided into batches. Then, assuming that
the samples provide evidence of adequate mixing, functionals of model parameters
can be estimated from each batch for all sets of tuning parameters and the method
of Chen and Schmeiser (1993) can be applied to estimate the standard errors of all
batch estimates. These standard errors are then combined to produce a measure of
performance for each set of tuning parameters as follows.

Let \( I \) denote the number of quantities to be estimated. Let \( L \) denote the number
of sets of tuning parameters under consideration and let \( J \) denote the number of
batches. Let \( x_{ijl} \) denote the standard error of the point estimate of quantity \( i \) based
on batch \( j \) with the \( l \)-th set of tuning parameters. Then the geometric mean of the
standard errors for each set of tuning parameters, \( \rho_l, l = 1, \ldots, L, \) is given by

\[
\rho_l = \left( \prod_{i,j} x_{ijl} \right)^{1/(IJ)}.
\]

Taking the minimum value of \( \rho_l \) as a reference, the ratio of \( \rho_l \) to the reference is given
by

\[
\phi_l = \frac{\rho_l}{\min_l \{\rho_l\}}.
\]

Beside the set of tuning parameters corresponding to the minimum \( \rho_l \), for which the
ratio \( \phi_l \) equals one, other good sets of tuning parameters will be those yielding ratios
close to one.

For the example of Chapter 5, \( p_{\text{MH}} \) was set at 0.1 and 0.5. Three sets of \((r_0, r_1)\)
values were chosen: \((r_0 = 0.98, r_1 = 0.98)\), \((r_0 = 0.98, r_1 = 0.95)\) and \((r_0 = 0.95, r_1 =
0.95)\). Inferences on the model parameters, on the posterior probabilities of outlying-
ness for the first five observations, and on the posterior probabilities of outlyingness
for all the observations are of interest. Table 8.1 records the values of \( \phi_l \), where the
labels of “a” and “b” denote equal and adjusted selection probabilities respectively.
The column labeled “parameter” records the values of $\varphi_l$ concerning estimation of the model parameters, the column labeled “five $p_i$” records the values of $\varphi_l$ concerning estimation of the posterior outlier probabilities for the first five observations, and the column labeled “all $p_i$” records the values of $\varphi_l$ concerning estimation of the posterior outlier probabilities for all the observations. This table shows that using adjusted selection probabilities improves the stability of the estimates in the sets $(r_0, r_1) = (0.98, 0.98)$ and $(r_0, r_1) = (0.98, 0.95)$ with $p_{MH} = 0.5$, and that $p_{MH} = 0.5$ yields the smallest standard errors. Overall, it is good to choose $(r_0, r_1) = (0.98, 0.98)$ with $p_{MH} = 0.5$ and adjusted selection probabilities.

For the example in Chapter 6, $p_{MH}$ was set at 0.1 and 0.5. Three sets of $(r_0, r_1)$ values were chosen as follows: $(r_0 = 0.99, r_1 = 0.99)$, $(r_0 = 0.99, r_1 = 0.95)$ and $(r_0 = 0.95, r_1 = 0.95)$. Inferences on the model parameters, on the posterior probabilities of outlyingness for the last three observations, and on the posterior probabilities

<table>
<thead>
<tr>
<th>$p_{MH}$</th>
<th>$(r_0, r_1)$</th>
<th>$p_i$</th>
<th>$\varphi_l$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>(0.98, 0.98)</td>
<td>a</td>
<td>1.9666 2.0227 1.4032</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td>2.1199 2.1802 1.4552</td>
</tr>
<tr>
<td>(0.98, 0.95)</td>
<td>a</td>
<td>2.0873 2.1559 1.4782</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td>2.2168 2.2755 1.4830</td>
</tr>
<tr>
<td>(0.95, 0.95)</td>
<td>a</td>
<td>2.3333 2.3991 1.5824</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td>2.6036 2.6711 1.7086</td>
</tr>
<tr>
<td>0.5</td>
<td>(0.98, 0.98)</td>
<td>a</td>
<td>1.0457 1.0486 1.0611</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td><strong>1.0000</strong> 1.0000 1.0000</td>
</tr>
<tr>
<td>(0.98, 0.95)</td>
<td>a</td>
<td>1.1619 1.1445 1.1190</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td>1.1342 1.1286 1.0555</td>
</tr>
<tr>
<td>(0.95, 0.95)</td>
<td>a</td>
<td>1.0957 1.0920 1.0774</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>b</td>
<td>1.2713 1.3032 1.1506</td>
</tr>
</tbody>
</table>

Table 8.1: The values of $\varphi_l$ for the example in Chapter 5.
Table 8.2: The values of $\varphi_l$ for the example in Chapter 6.

<table>
<thead>
<tr>
<th>$p_{\text{MH}}$</th>
<th>$(r_0, r_1)$</th>
<th>$p_3$ parameter</th>
<th>all $p_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\varphi_l$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>three $p_i$</td>
<td>all $p_i$</td>
</tr>
<tr>
<td>0.1</td>
<td>(0.99, 0.99)</td>
<td>a 1.7835</td>
<td>1.1849</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.8741</td>
<td>1.1619</td>
</tr>
<tr>
<td></td>
<td>(0.99, 0.95)</td>
<td>a 1.8798</td>
<td>1.2122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.7313</td>
<td>1.1454</td>
</tr>
<tr>
<td></td>
<td>(0.95, 0.95)</td>
<td>a 1.9087</td>
<td>1.2055</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.8585</td>
<td>1.1842</td>
</tr>
<tr>
<td>0.5</td>
<td>(0.99, 0.99)</td>
<td>a 1.0672</td>
<td>1.1513</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.0135</td>
<td>1.0862</td>
</tr>
<tr>
<td></td>
<td>(0.99, 0.95)</td>
<td>a 1.0340</td>
<td>1.0387</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.0000</td>
<td>1.0593</td>
</tr>
<tr>
<td></td>
<td>(0.95, 0.95)</td>
<td>a 1.0235</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>b 1.0279</td>
<td>1.0602</td>
</tr>
</tbody>
</table>

of outlyingness for all the observations are of interest. Table 8.2 records the values of $\varphi_l$, where all the labels are the same as in the example of Chapter 5 except that the column labeled “three $p_i$” records the values of $\varphi_l$ concerning estimation of the posterior outlier probabilities for the last three observations, and the column labeled “all $p_i$” records the values of $\varphi_l$ concerning estimation the posterior outlier probabilities for all the observations. This table shows that, overall, using adjusted selection probabilities improves the stability of the estimates and $p_{\text{MH}} = 0.5$ achieves the smallest standard errors. Good choices are $(r_0, r_1) = (0.99, 0.99)$ or $(r_0, r_1) = (0.99, 0.95)$ with $p_{\text{MH}} = 0.5$ and adjusted selection probabilities.

For the example in Section 7.4.2, the same sets of $p_{\text{MH}}$ and $(r_0, r_1)$ values as in the example of Chapter 6 were chosen. Inferences on the model parameter $\gamma_2$, on the Stage II posterior outlier probabilities for the handguns .30-30 and .35 REM, and on the Stage II posterior outlier probabilities for all the handguns are of interest.
Table 8.3: The values of $\varphi_l$ for the example in Section 7.4.2.

<table>
<thead>
<tr>
<th>$p_{\text{MH}}$</th>
<th>$(r_0, r_1)$</th>
<th>$p_\delta$</th>
<th>$\gamma_2$</th>
<th>two $p_i$</th>
<th>all $p_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(0.99, 0.99)</td>
<td>a</td>
<td>1.9481</td>
<td>2.0305</td>
<td>1.0774</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>1.8521</td>
<td>2.0819</td>
<td>1.0540</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>(0.99, 0.95)</td>
<td>a</td>
<td>1.5481</td>
<td>1.6670</td>
<td>1.0009</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>2.1983</td>
<td>2.3607</td>
<td>1.1577</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.95, 0.95)</td>
<td>a</td>
<td>1.8295</td>
<td>1.9534</td>
<td>1.0540</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>2.3131</td>
<td>2.5113</td>
<td>1.1376</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>(0.99, 0.99)</td>
<td>a</td>
<td>1.2086</td>
<td>1.1619</td>
<td>1.0529</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>1.1828</td>
<td>1.1781</td>
<td>1.1407</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.99, 0.95)</td>
<td>a</td>
<td><strong>1.0000</strong></td>
<td>1.0430</td>
<td><strong>1.0000</strong></td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>1.0041</td>
<td><strong>1.0000</strong></td>
<td>1.0538</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.95, 0.95)</td>
<td>a</td>
<td>1.3922</td>
<td>1.3019</td>
<td>1.0495</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>1.0626</td>
<td>1.0790</td>
<td>1.0173</td>
<td></td>
</tr>
</tbody>
</table>

Table 8.3 records the values of $\varphi_l$, where the column labeled “$\gamma_2$” records the values of $\varphi_l$ concerning estimation of the model parameter $\gamma_2$, the column labeled “two $p_i$” records the values of $\varphi_l$ concerning estimation of the posterior outlier probabilities for the handguns .30-30 and .35 REM, and the column labeled “all $p_i$” records the values of $\varphi_l$ concerning estimation of the posterior outlier probabilities for all the handguns. This table shows that using adjusted selection probabilities improves the stability of the estimates in some sets of $p_{\text{MH}}$ and $(r_0, r_1)$ values and that $p_{\text{MH}} = 0.5$ yields the smallest standard errors. Overall, it is good to choose $(r_0, r_1) = (0.99, 0.95)$ with $p_{\text{MH}} = 0.5$.

A common feature of these three examples is that $p_{\text{MH}} = 0.5$ yields smaller standard errors of the estimates and using the adjusted selection probabilities makes the estimates more reliable when $p_{\text{MH}} = 0.5$. 

121
8.2 Future Work

My dissertation confirms that the performance of the proposed algorithms is highly
dependent on the correct identification of the isolated modes and the related clustering
of the observations. The model based clustering method (package MCLUST in R)
that is used in one of the proposed algorithms groups the observations based on a
normal mixture model via the EM algorithm. This method works effectively when the
normality assumption is satisfied but may not be otherwise appropriate. To surmount
this hurdle, it will help to develop distance measures based on relevant features of
the posterior distribution under a given model and use these distances to perform the
clustering.

I also intend to develop suitable diagnostics for assessing convergence of the pro-
posed algorithms. A promising approach that can be pursued is a generalization of the
Gelman and Rubin (1992) diagnostic. The basic premise of the original diagnostic is
that, in the presence of poor mixing, the simulated MCMC sequence depends highly
on the starting point. As a consequence, simulating multiple chains from overdis-
persed starting points and comparing within- and between-chain variability using
one-way ANOVA techniques provides a means of detecting lack of convergence. The
proposed generalization involves the simultaneous comparison of the performance of
several algorithms using two-way ANOVA techniques. The ANOVA models employed
to build the diagnostic should regard algorithm as the factor of principal interest and
block on the starting value.

It was noted that the proposed algorithms are highly dependent on the identifica-
tion of isolated modes. For some complex distributions, several possible modes may
not be discovered by the initial mode identification step of the algorithms. To remedy
this omission, some MCMC method capable of visiting all the posterior regions can be incorporated into the proposed algorithms. For example, the MCMC tempering transition method may be appropriate, since it is able to walk freely on the state space. If the added sampler is computationally expensive, fewer steps of this sampler should be entertained.

In addition, I intend to apply the hybrid algorithm with clustering initialization described in Section 3.2 for variable selection. High dimensional data sets arise in many fields such as genetics and image processing. Variable selection is very important to deal with these data sets. Many methods have been developed to perform variable selection, such as Gibbs variable selection suggested by Dellaportas et al. (1997) and stochastic search variable selection introduced by George and McCulloch (1993). The hybrid algorithm can also be used for variable selection. To do this an indicator variable is first assigned to each variable to denote whether it is or it is not included in the model, and possible sets of variables corresponding to promising models are identified through preliminary runs by some standard variable selection method. The preferred configurations corresponding to each identified set of promising variables can then be created and the procedure in Section 3.2.2 can be used to build the MH transition kernel.

A rough procedure was suggested in the previous section for selecting the tuning parameters. However, we need to refine this procedure for obtaining optimal values of the tuning parameters. Also, the examples shown in this dissertation do not have many isolated modes. In the future, I intend to apply the proposed algorithms to sample from mixture distributions with a large number of components. In addition, I will extend the algorithms to deal with the situations of a variable number of
components. A possible way to do this is to adopt a split-merge technique proposed by Jain and Neal (2004).
BIBLIOGRAPHY


