Hierarchical Hidden Markov Models with Applications to BiSulfite-Sequencing Data

Tusharkanti Ghosh

School of Mathematics and Statistics

The University of Glasgow

A thesis submitted for the degree of

Doctor of Philosophy

March 2018

Abstract

DNA methylation is an epigenetic modification with significant roles in various biological processes such as gene expression and cellular proliferation. Aberrant DNA methylation patterns compared to normal cells have been associated with a large number of human malignancies and potential cancer symptoms. In DNA methylation studies, an important objective is to detect differences between two groups under distinct biological conditions, for e.g., between cancer/ageing and normal cells. BiSulfite sequencing (BS-seq) is currently the gold standard for experimentally measuring genome-wide DNA methylation. Recent evolution in the BS-seq technologies enabled the DNA methylation profiles at single base pair resolution to be more accurate in terms of their genome coverages. The main objective of my thesis is to identify differential patterns of DNA methylation between proliferating and senescent cells. For efficient detection of differential methylation patterns, this thesis adopts the approach of Bayesian latent variable model. One such class of models is hidden Markov model (HMM) that can detect the underlying latent (hidden) structures of the model. In this thesis, I propose a family of Bayesian hierarchical HMMs for identifying differentially methylated cytosines (DMCs) and differentially methylated regions (DMRs) from BS-seq data which act as important indicators in better understanding of cancer and other related diseases. I introduce HMMmethState, a model-based hierarchical Bayesian technique for identifying DMCs from BS-seq data. My novel HMMmethState method implements hierarchical HMMs to

account for spatial dependence among the CpG sites over genomic positions of BS-seq methylation data.

In particular, this thesis is concerned with developing hierarchical HMMs for the differential methylation analysis of BS-seq data, within a Bayesian framework. In these models, aberrant DNA methylation is driven by two latent states: differentially methylated state and similarly methylated state, which can be interpreted as methylation status of CpG sites, that evolve over genomic positions as a first order Markov chain. I first design a (homogeneous) discrete-index hierarchical HMM in which methylated counts given the methylation status of CpG sites follow Beta-Binomial emission distribution specific to the methylation state. However, this model does not incorporate the genomic positional variations among the CpG sites, so I develop a (non-homogeneous) continuous-index hierarchical HMM, in which the transition probabilities between methylation status depend on the genomic positions of the CpG sites.

This Beta-Binomial emission model however does not take into account the correlation in the methylated counts of the proliferating and senescent cells, which has been observed in the BS-seq data analysis. So, I develop a hierarchical Normal-logit Binomial emission model that induces correlation between the methylated counts of the proliferating and senescent cells. Furthermore, to perform parameter estimation for my models, I implement efficient Markov Chain Monte Carlo (MCMC) based algorithms. In this thesis, I provide an extensive study on model comparisons and adequacy of all the models using Bayesian model checking. In addition, I also show the performances of all the models using Receiver Operating Characteristics

(ROC) curves. I illustrate the models by fitting them to a large BS-seq dataset and apply model selection criteria on the dataset in search of selecting the best model. In addition, I compare the performances of my methods with existing methods for detecting DMCs with competing methods. I demonstrate how the HMMmethState based algorithms outperform the existing methods in simulation studies in terms of ROC curves. I present the results of DMRs obtained using my method, i.e., the results of DMRs with the proposed HMMmethState that have been applied to the BS-seq datasets. The results of the hierarchical HMMs explain that I can certainly implement these methods under unconditioned settings to identify DMCs for high-throughput BS-seq data. The predicted DMCs can also help in understanding the phenotypic changes associated with human ageing.

Acknowledgements

I would sincerely like to thank my supervisors, Dr Mayetri Gupta and Dr Vincent Macaulay, for their guidance and patience over the course of my Ph.D. education. I would also like to extend my gratitude to Prof Peter Adams and the Adams' Lab for their help with the methylation data of this thesis. Further, I give my thanks to friends in the School of Mathematics and Statistics for their encouragement and support.

I am grateful to College of Science and Engineering for a tuition fee waiver.

Lastly, I would also like to thank my parents and my brother for their support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The recent arrival of ultra-high throughput, next generation sequencing (NGS) technologies has revolutionized the genetics and genomics fields by allowing rapid and inexpensive sequencing of the billions of bases in human and other genomes. The rapid deployment of NGS in a variety of sequencing-based experiments has resulted in fast accumulation of massive amounts of sequencing data. These technologies have enhanced the potential for understanding the workings of biological systems in depth and the development of personalized medicine and are having an impact on the types of questions that biologists can ask these days.

In the past few years, several pioneering studies have put the focus on epigenetics. Literally, the word *epigenetic* means *in addition to alterations in genetic sequence*. Epigenetics generally focuses on biological processes that regulate the activation of certain genes, i.e., how and when the genes are switched on or switched off, whereas epigenomics is involved in the analysis of epigenetic modifications across many genes in a cell or a multi-cellular organism. Epigenetic processes control normal organism functions. However, if they occur abnormally, there is the possibility of unfavourable health effects or diseases, such as cancer. The most significant epigenetic process, which has been studied extensively in the recent years due to the availability of high-throughput sequencing technology, is

DNA (deoxyribonucleic acid) methylation.

1.1 DNA methylation

Figure 1.1 shows DNA contains combinations of the four nucleotides which include cytosine(C) (pink), guanine(G) (green), thymine(T) (blue) and adenine(A) (orange). DNA methylation is a chemical modification of DNA resulting from the addition of a methyl (CH_3) group to a DNA nucleotide. DNA methylation is an epigenetic modification which regulates gene transcription and is recognized for their role in gene expression (Gopalakrishnan et al., 2008). The CpG sites are DNA dinucleotide positions of DNA where a C nucleotide is followed by a G nucleotide in the linear sequence of nucleotides along its $5' \rightarrow 3'$ direction. A CpG site (CpG dinucleotide) is defined to be methylated, if a CH_3 group is added to the C. In addition, DNA treatment with sodium bisulfite chemicals initiates conversion of unmethylated cytosine to Uracil (U) which is subsequently converted to T by DNA polymerase, whereas a methylated cytosine remains unaffected (Krueger et al., 2012).

1.1.1 Importance of DNA methylation

Cytosine methylation of DNA plays an active role in epigenetic mechanism to control gene expression, silencing or genomic imprinting (Li et al., 1993), both during the normal developmental stage and as well as in the adult (Law and Jacobsen, 2010). The occurrence of DNA methylation was first confirmed in human cancer in 1983. DNA methylation plays a key part in the development of cancer and is also an active regulator of gene transcription. It enables a single cell to develop into a complex multicellular organism (Smith and Meissner, 2013), in the formation of chromatin structure, which is another important epigenetic

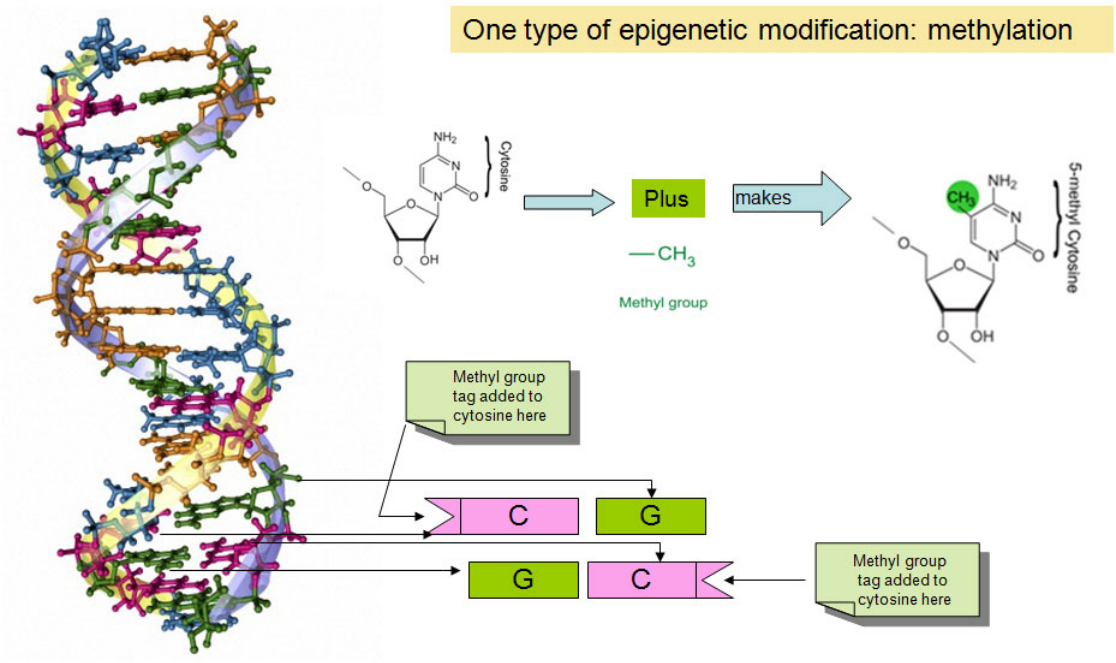


Figure 1.1: DNA methylation.

modification.

1.1.2 Differential methylation

Several studies have confirmed that genes with a promoter region that contains a high concentration of 5′-methylated Cs are transcriptionally silent. These studies mainly discuss the functional changes to the promoter regions that are differentially methylated between cancer/ageing and normal cells. Aberrant DNA methylation patterns are a hallmark feature of cancer (Das and Singal, 2004, Kulis and Esteller, 2010, Laird and Jaenisch, 1994) and have been widely associated with numerous diseases (Robertson, 2005). In this context, I shall use differentially methylated C (DMC) to denote a differentially methylated C and differentially methylated region (DMR) to denote a genomic region of adjacent DMCs. DNA hypermethylation is linked to the activation of genes and DNA hypomethylation (Esteller, 2002, Qu et al., 2014) has been associated with the development of cancer through various mechanisms.

1.2 Motivation

Many studies have confirmed correlation between promoter methylation and gene expression (Henrichsen et al., 2009, Moarii et al., 2015). In addition, the occurence of wide-ranging aberrantly methylated regions is a characteristic feature of various kinds of cancer (Ehrlich, 2002). Identifying DMRs in the genome is crucial for attaining deeper knowledge into the functioning of epigenetic processes, from the cellular level to multicellular organisms, i.e., eukaryotes. Understanding functional (regulatory) regions is one of the main challenges in epigenetics. One of the most essential steps during the epigenetic process is to determine how proteins interact with targeted DNA for the regulation of gene expression (Laurent et al., 2010).

To uncover the regulation mechanisms of the epigenetic process, one promising approach is to identify DMRs on the genome scale. The popular technology used to study the mechanism is BiSulfite sequencing (BS-seq). Epigenetic regulation is a routine mechanism in cancer for silencing the expression of tumor suppressor genes (Blair and Yan, 2012) and actively participates in various cellular processes, such as, gene expression and regulation (Newell-Price et al., 2000).

One of the most important applications in the field of epigenetics is the epigenetic modifications to the genome of cancer cells that do require an alteration in the nucleotide sequence. Understanding the pattern of epigenetic mechanisms seems likely to be effective in the future for cancer detection, therapy and prevention. BS-seq procedure is one of the most reliable technologies to profile genome-wide DNA methylation reads in eukaryotes. In contrast to the rapid development of numerous pre-processing, alignment and mapping softwares, tools for analysing the generated methylation reads and implementing a flexible pipeline to identify differential DNA methylation patterns in two groups, e.g., cancer and control samples, are comparatively limited.

1.3 Bayesian latent variable framework for the analysis of differential methylation

In this thesis, I have focussed on the application of latent variable Bayesian techniques to epigenomics, in particular to detect DMRs. These DMRs are usually studied by performing BS-sequencing, an experimental prcedure that applies high-throughput methods on bisulfite-induced DNA to ascertain the methylation status at each CpG site. The epigenomic profiles of differential methylation of DNA are examined to assess the regulatory roles of differentially expressed genes which are generally identified with predominant hypomethylation.

Due to the rapid development of BS-seq technology, several algorithms have been designed to analyse the data and identify the DMRs of interest, but the algorithms are mostly restricted to Fisher's exact test or Wald's test. On the other hand, I focussed on developing a method that is suitable for modelling the data observations with respect to the data-generating process and the underlying genomic structure.

I incorporate the genomic location in the model to help identify the DMRs of interest. I designed a family of hierarchical hidden Markov models (HMMs), HMMmethState, that treat the genome as a sequence of latent states, classified as DMCs or similarly methylated Cs (SMCs). I implemented Markov Chain Monte Carlo (MCMC) based algorithms using Forward-sum recursions, Gibbs samplers and the Metropolis-Hastings (M-H) algorithms to estimate the latent states and the model parameters.

In addition, I have explored several characteristics of my proposed method to study its performance. The main advantage of HMMmethState is the inclusion of the Bayesian approach to parameter and state estimation. Furthermore, my proposed choice of the Binomial distribution to model the distributions at the first stage of the hierarchical model of methylated counts reports the approximately random process during the sampling of the two types of reads, i.e., methylated or unmethylated.

1.4 Thesis outline

This thesis is motivated by questions in epigenetics and aims to analyse BS-seq data applied to the study of differential methylation patterns. In Chapter 2, I introduce basic statistical concepts that form the basis of my analysis. I provide a brief description of the Bayesian framework and MCMC based algorithms

that form the basis of my research. In addition, I also describe a family of models within a hierarchical HMM framework, which aim to characterize systems that are dependent on an underlying structure. The generic Bayesian estimation techniques involving MCMC based algorithms and also the Forward-sum recursion are also described in Chapter 2. Implementation and further developments of the MCMC based algorithms and Forward-sum recursions are described in details in Chapters 4 and 5, tailored to the nature of the problem in question. Furthermore, I give a brief description of the implementation of MCMC based algorithms and convergence tools. Finally, I describe the model selection criteria and also posterior predictive analysis within the Bayesian framework, which form a substantial part of my analysis. However, not all model selection criteria are directly applicable to my problems and I provide an explanation of those difficulties in Chapter 6.

Chapter 3 gives an introduction to the high-throughput sequencing technology that generates the data analysed in the remaining part of this thesis and existing differential methylation caller approaches. It describes the sequencing procedure and the subsequent steps involved in the processing of BS-seq methylation data. Furthermore, it also provides an overview of the BS-seq tool, Bismark, that performs the alignments of bisulfite-treated reads to a reference genome for further analysis. In addition, I also give brief descriptions of existing differential methylation caller approaches that aim to detect DMCs in the genome. Finally, I also introduce the structure of the BS-seq datasets that are used for the analysis in this thesis.

In Chapter 4, I propose two HMMmethState models (*BBDM* and *BBCM*) that I developed using a hierarchical Beta-Binomial emission distribution and explain its association with the data-generating process. I also provide a detailed description of the Bayesian estimation procedure for estimating the model parameters and hidden states, which subsequently enables the identification of DMCs.

In Chapter 5, I develop an extension to the HMMmethState models. I propose an improved emission distribution as the Beta-Binomial emission distribution fails to capture the observed correlation between the methylated counts of the two cell types (senescent and proliferating cells). I implement the extended HMMmethState models (*NLBD* and *NLBC*) using a hierarchical bivariate Normal-Binomial distribution and explain their associations with the data-generating process. Here also, I provide a detailed description of Bayesian estimation procedure for estimating the model parameters and hidden states, which subsequently enables the identification of the DMCs.

Chapter 6 provides a description of a simulation study of all the HMMmethState models and compare the model performances based on the selection criteria and ROC curves. In addition, I include the visual exploration and assessment of some features of the model using posterior predictive checks. I also discuss the comparison among the model selection criteria and posterior predictive p values of the competing HMMmethState methods. Furthermore, I also discuss the the performance of my HMMmethState model and assess my proposed algorithms by comparing with two existing differential methylation caller approaches described in Chapter 3. Finally, I present the results of HMMmethState on chromosomal datasets and compare them with the two competing methods.

Chapter 2

Statistical Concepts and Methods

In this chapter, I introduce a number of statistical concepts and methods used and implemented throughout this thesis. In Section 2.1, I introduce the Bayesian techniques used in Chapters 4 and 5, in particular Markov Chain Monte Carlo (MCMC) based inference. These MCMC methods are combined with the hidden Markov models (HMMs) from Section 2.2 to create Bayesian HMMs introduced in Chapters 4 and 5. The concepts of identifiability and label switching are outlined in Sections 2.2.5 and 2.2.6, respectively. In addition, I also discuss Bayesian model checking and model selection (Section 2.3) implemented in Chapter 6.

2.1 Bayesian framework

In a Bayesian framework, the parameters are considered as random variables whereas classical framework treats parameters to be unspecified but fixed. For a given model specification, the data D can be modelled based on the parameters $\boldsymbol{\theta}$, a vector of random quantities with prior distribution $p(\boldsymbol{\theta})$. According to Bayes' theorem (Bayes, 1763), the (posterior) distribution of the parameters $\boldsymbol{\theta}$ given the data D is proportional to the product of likelihood $\mathbf{L}(\boldsymbol{\theta}|D) = p(D|\boldsymbol{\theta})$, i.e., the probability of the data D given the parameters $p(\boldsymbol{\theta})$ and the prior distribution

$p(\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.1}$$

assuming $\boldsymbol{\theta}$ is continuous.

Generally, calculating the posterior distribution and its moments is not possible in complex or high-dimensional problems as integrating over $\boldsymbol{\theta}$ requires computing high-dimensional integrals with no closed form solution. However, to tackle these complications, numerical integration is required, which can be done by Monte Carlo methods. The posterior distribution $p(\boldsymbol{\theta}|D)$ can be sampled from $p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ using the MCMC methodology.

2.1.1 Markov chain Monte Carlo

The MCMC method simulates samples from the posterior distribution of the model parameters when the likelihood is tractable. The objective is to construct a Markov chain whose stationary distribution is the posterior distribution of interest (target distribution). The new sample is simulated based on the current sampled value only and hence the samples form a Markov chain. The chain is run for sufficiently long, as convergence to the stationary distribution is not attained immediately. A burn-in period is thus required, where the initial simulated values are discarded (Brooks and Gelman, 1998). Furthermore, to save storage memory and reduce the autocorrelation between samples, sometimes, only every i^{th} $(i > 1)$ updated sample of the parameter is stored. This process is termed thinning of the chain.

In the next two subsections, I discuss the two most popular MCMC implementations used to simulate values of the parameters from their posterior distribution.

2.1.2 Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm is an MCMC technique which was first introduced by Metropolis et al. (1953) and later developed by Hastings (1970). The M-H algorithm is often used where the posterior distribution cannot easily be directly sampled from. It implements a rejection sampling method based on the target distribution to sample the parameter from the posterior distribution through an acceptance and rejection step. A candidate sample is simulated from a proposal distribution conditional on the updated draw of the previous state, and subsequently it is accepted or rejected based on an acceptance probability that depends on the posterior density and the proposal density.

To illustrate the algorithm, let the proposal density for a candidate draw θ' given the current update θ in the sequence of the samples be denoted by $q(\theta, \theta')$. The M-H algorithm simulates new candidate values of the parameter from the proposal distribution (candidate distribution) and accepts them as the next update in the Markov chain according to the acceptance probability,

$$\alpha(\theta, \theta') = min\left(1, \frac{p(\theta'|D)q(\theta', \theta)}{p(\theta|D)q(\theta, \theta')}\right). \tag{2.2}$$

The proposed value θ' is only accepted if the acceptance probability $\alpha(\theta, \theta')$ is greater than a realized value u of the uniform random variable U on the interval $[0, 1]$, such that $U \sim \text{Uniform}(0,\ 1)$. However, if the proposed value is rejected, then the next sampled value is set to be the current one.

2.1.3 Gibbs sampler

Gibbs sampler (Geman and Geman, 1984) is a special case of the M-H algorithm. The parameter vector of the Markov chain is split into components and then each component of the parameter vector is updated sequentially. The Gibbs sampler samples each component from its distribution conditional on the remaining com-

ponents and the data, (the full conditional) one at a time.

I describe the sampling steps of the algorithm for a parameter vector $\boldsymbol{\theta}$ with S components $(\theta_1, \ldots, \theta_S)$ and full conditionals $p(\theta_s|D, \boldsymbol{\theta}_{-s})$, where $\boldsymbol{\theta}_{-s} = \{\theta_1, \ldots, \theta_{s-1}, \theta_{s+1}, \ldots, \theta_S\}$ as below:

1. Set $i = 0$ and initialize starting values: $\boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_S^0)$.

2. Simulate,

$$\begin{aligned}
&\theta_1^{i+1} \text{ from } p(\theta_1^{i+1}|D, \theta_2^i, \ldots, \theta_S^i) \\
&\quad \vdots \\
&\theta_s^{i+1} \text{ from } p(\theta_s^{i+1}|D, \theta_1^{i+1}, \ldots, \theta_{s-1}^{i+1}, \theta_{s+1}^i, \ldots, \theta_S^i) \\
&\quad \vdots \\
&\theta_S^{i+1} \text{ from } p(\theta_S^{i+1}|D, \theta_1^{i+1}, \ldots, \theta_{S-1}^{i+1}).
\end{aligned} \tag{2.3}$$

3. Set $i = i + 1$.

4. Repeat 2 and 3 beyond convergence and discard burn-in.

After a number of iterations, the Markov chain that converges to the target distribution and then the updated values are sampled from the desired posterior distribution.

The M-H algorithm does not require the information related to the full conditional distributions, in contrast to the Gibbs sampler. However, the M-H algorithm often requires the tuning of parameters in the proposal distribution in order to speed up the convergence to stationarity. On the other hand, the Gibbs sampler automatically determines the proposal distributions from which the updated samples are always accepted.

2.1.4 Assessing MCMC convergence

Convergence is often examined by running parallel chains with different initial values to assess whether all the chains converge to the same target distribution by using Potential Scale Reduction Factors (PSRFs). A PSRF is a measure which evaluates the convergence of multiple parallel MCMC chains as proposed by Gelman and Rubin (1992). The calculation of the PSRF for each parameter, θ, requires m parallel sequences, each of length n. Let θ_{ij} be the i^{th} sample of the j^{th} chain, $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Let $\bar{\theta}_j$ and s_j^2 be the sample posterior mean and variance of the j^{th} parallel chain. Let $\bar{\theta}$ be the overall sample posterior mean. To calculate the PSRF of each of the model parameters, one computes the between-sequence, B, and within-sequence, W, variances:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\theta}_j - \bar{\theta}\right)^2, \quad \text{where} \quad \bar{\theta}_j = \frac{1}{n} \sum_{i=1}^{n} \theta_{ij}, \quad \bar{\theta} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_j \tag{2.4}$$

$$W = \frac{1}{m} \sum_{j=1}^{m} s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_{ij} - \bar{\theta}_j\right)^2. \tag{2.5}$$

One can then estimate $\text{Var}(\theta|D)$, the marginal posterior variance of the parameter, by a weighted average of W and B:

$$\widehat{\text{Var}}(\theta|D) = \frac{n-1}{n} W + \frac{1}{n} B. \tag{2.6}$$

This quantity overestimates the marginal posterior variance of the parameter, $\text{Var}(\theta|\mathbf{y})$, while W underestimates it for finite n. From this the PSRF can be calculated as follows:

$$\hat{R} = \sqrt{\frac{\widehat{\text{Var}}(\theta|D)}{W}}, \tag{2.7}$$

where the value decreases to 1 as $n \to \infty$. Large values of $\hat{R}$ indicate a lack of convergence and values of less than 1.05 or 1.1 generally indicate convergence.

2.2 Hidden Markov models

In this section, I introduce hidden Markov models (HMMs) (Rabiner, 1989) that describe observations emerging conditionally from an underlying discrete and unobserved process. Each observation is associated with a latent or hidden state which yields classification of the data into distinct clusters/groups. In the context of finite mixture models, the hidden or latent states are assumed to be independent and identically distributed (i.i.d.) random variables. However, in a HMM, they are represented by an unobservable Markov chain. HMMs induce long-range conditional dependencies in the observed data by imposing Markovian conditioning on the latent states and have many applications in pattern recognition, high-throughput sequencing data and bioinformatics (Durbin et al., 1998, Koski, 2001).

A HMM is a bivariate stochastic process comprising an observed process and a hidden (unobserved) process. The unobserved process is assumed to be a first-order Markov chain with a finite number of hidden states, whereas, the observable random variables conditional on the hidden states generate a conditionally independent sequence, which is termed as the emission sequence, where the conditional (emission) distribution of the observable random variable depends only on the corresponding hidden state. In most standard cases, the HMM is generally assumed to be homogeneous if the Markov chain in the hidden process is homogeneous, i.e., in the underlying Markov chain, the transition probabilities are constant over time. However, non-homogeneous transition probabilities can also be incorporated in the hidden process. The concept of a non-homogeneous hidden Markov process will be formally introduced in Section 4.1.4 through a continuous-index hidden Markov process.

Let us define, $\mathbf{X} = (X_1, \cdots, X_T)$ be the sequence of observable random variables, such that $\mathbf{x} = (x_1, \cdots, x_T)$ are the realizations of $\mathbf{X}$ and $\mathbf{Z} = (Z_1, \cdots, Z_T) \in$

$\mathbb{Z}_K^T$ be the sequence of hidden states, where $\mathbb{Z}_K^T = \overbrace{\mathbb{Z}_K \otimes \cdots \otimes \mathbb{Z}_K}^{\text{T terms}}$ and $\mathbb{Z}_K = \{1, \ldots, K\}$, such that $\mathbb{Z}_K^T$ is the set of all possible hidden states. Now, the hidden process can be derived from the first-order Markovian property as,

$$P(Z_t = j|\mathbf{Z}_{1:t-1}) = P(Z_t = k|Z_{t-1} = j),\ t = 2, \ldots, T \text{ and } j, k = 1, \ldots, K, \quad (2.8)$$

where $\mathbf{Z}_{1:t-1} = (Z_1, \ldots, Z_{t-1})$.
The three main sets of parameters of an HMM correspond to the initial state distribution, the transition probability matrix and the emission distribution (Rabiner, 1989). The initial state parameters, transition parameters and emission parameters are denoted by $\boldsymbol{\pi}$, $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$, respectively. They are as follows.

- Let us consider the initial state distribution $P(Z_1 = k) = \pi_k$ for $k = 1, \ldots, K$, with initial state probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \cdots, \pi_K)$, such that $\sum_{k=1}^K \pi_k = 1$. π_k is the prior probability of state k at the first step in the chain.

- The transition probabilities between the states $\tau_{jk}(t) = P(Z_t = k|Z_{t-1} = j)$ for $j, k = 1, 2, \ldots, K$ and $t = 2, \ldots, T$ are given by the matrix $\boldsymbol{\tau}$.

- The emission probability of the (discrete) observation x_t conditional on the hidden state Z_t and the emission parameter can be written as:

$$\begin{aligned} b_k(t) &= P(X_t = x_t|\boldsymbol{\theta}, Z_t = k) \\ &= P(x_t|\boldsymbol{\theta}, Z_t = k),\ k = 1, \ldots, K. \end{aligned} \quad (2.9)$$

  For notational simplicity, I re-write $P(x_t|\boldsymbol{\theta}, Z_t) = P_{Z_t}(x_t|\boldsymbol{\theta}) = b_{Z_t}(t)$, which is termed as the emission distribution at index t for $t = 1, \ldots, T$ conditional on the state Z_t.

I simplify my notation, by defining, $\mathbf{e}_{s:T} = (e_s,\ e_{s+1}, \ldots, e_T)$, where $\mathbf{e}_{s:T}$ is a vector with $(T-s)+1$ elements and s is any positive integer, such that $s \leq T$.

More generally, the hidden state sequence $\mathbf{Z} = (Z_1, \ldots, Z_T)$ can also be assumed to be a Markov process of m^{th} order, such that the conditional distribution of Z_t given all the past values $\mathbf{Z}_{1:t-1}$ depends only on the preceding m values, i.e., $\mathbf{Z}_{t-1:t-m} = (Z_{t-m}, Z_{t-m+1}, \ldots, Z_{t-1})$. When $m = 0$, the HMM boils down to a finite mixture model.

2.2.1 Computing the likelihood

Let the set of all parameters be generically denoted by $\boldsymbol{\zeta} = (\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\tau})$ where $\boldsymbol{\theta}$ denotes the emission parameters and $\boldsymbol{\pi}$, $\boldsymbol{\tau}$ denote the initial state and transition parameters, respectively, such that $\boldsymbol{\pi} = \{\pi_k :\ k = 1, 2, \ldots, K\}$ and $\boldsymbol{\tau} = \{\tau_{jk}(t) :\ j, k = 1, 2, \ldots, K\}$. The joint probability of the sequence of observable random variables $\mathbf{X}$ and the sequence of the hidden states $\mathbf{Z}$ conditional on the model parameters $\boldsymbol{\zeta}$ is

$$P(\mathbf{X}, \mathbf{Z}\,|\boldsymbol{\zeta}) = \pi_{Z_1} P_{Z_1}(\mathbf{X}_1|\boldsymbol{\theta}) \prod_{t=2}^{T} \tau_{Z_{(t-1)}, Z_t}(t) P_{Z_t}(\mathbf{X}_t|\boldsymbol{\theta}). \tag{2.10}$$

If $\mathbf{X}$ and $\mathbf{Z}$ were observed, (2.10) would give the complete data likelihood. To emphazise that only $\mathbf{X}$ is directly observed (to be $\mathbf{x}$), I shall hereafter write $P(\mathbf{X} = \mathbf{x}, \mathbf{Z}\,|.)$ as $P(\mathbf{x}, \mathbf{Z}\,|.)$, in a slight abuse of notation and similarly in posterior distributions and likelihoods, throughout my thesis.

Then, the likelihood of the observed data values $\mathbf{x}$ of $\mathbf{X}$ given the HMM mo-

del parameter $\boldsymbol{\zeta}$ can be expressed as,

$$\begin{aligned} L_{\mathbf{x}}(\boldsymbol{\zeta}) &= P(\mathbf{X} = \mathbf{x}|\boldsymbol{\zeta}) \\ &= \sum_{Z_1,\ldots,Z_T} \pi_{Z_1} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}) \prod_{t=2}^{T} \tau_{Z_{(t-1)},Z_t}(t) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}), \end{aligned} \tag{2.11}$$

the probability of the observed data values $\mathbf{x}$ of $\mathbf{X}$ for the HMM model parameters $\boldsymbol{\theta}$ which is the sum over all the K^T possible state sequences of the complete data likelihood.

The direct computation of the likelihood expression in (2.11), being the sum over all K^T possible realisations of $\mathbf{Z}$, is infeasible and must be avoided. With 1^{st} order Markovian dependencies of the hidden states in (2.11), it is straightforward to compute the likelihood using a recursive forward summation (Rabiner, 1989, Scott, 2002) procedure described below.

2.2.2 Forward-sum recursion

In this section, I introduce a forward variable at each index of the hidden state sequence and these forward variables are processed to compute the terms of the likelihood using a recursive method. The forward probability can be expressed as,

$$\alpha_k(t) = P\left(\mathbf{X}_{1:t} = \mathbf{x}_{1:t}; Z_t = k|\boldsymbol{\zeta}\right), \text{ for } k = 1, 2, \ldots, K. \tag{2.12}$$

Interestingly, the forward probability $\alpha_k(t)$ can also be viewed as the partial likelihood of the first t observed values of $\mathbf{x}_{1:t}$ of $\mathbf{X}_{1:t}$ with hidden state $Z_t = k$, i.e., $\alpha_k(t)$ is the joint probability of observing the data at the first t indices and being in state k at the t^{th} index.

For $t = 1$, I can write,

$$\begin{aligned}\alpha_k(1) &= P\left(X_1 = x_1, Z_1 = k|\boldsymbol{\zeta}\right)\\ &= \pi_k P\left(x_1|Z_1 = k, \boldsymbol{\theta}\right)\\ &= \pi_k P_k\left(x_1|\boldsymbol{\theta}\right)\\ &= \pi_k b_k(1).\end{aligned} \tag{2.13}$$

I can derive a recursive procedure to calculate $\alpha_k(t)$ for $t = 2, \ldots, T$ and $k = 1, 2, \ldots, K$ as below.

$$\begin{aligned}\alpha_k(t) &= P\left(\mathbf{X}_{1:t} = \mathbf{x}_{1:t}, Z_t = k|\boldsymbol{\zeta}\right)\\ &= \sum_{l \in \mathbb{Z}_K} P\left(\mathbf{x}_{1:t}, Z_{t-1} = l, Z_t = k|\boldsymbol{\zeta}\right)\\ &= \sum_{l \in \mathbb{Z}_K} P\left(\mathbf{x}_{1:t-1}, Z_{t-1} = l|\boldsymbol{\zeta}\right) P\left(x_t, Z_t = k|\mathbf{X}_{1:t-1}, Z_{t-1} = l; \boldsymbol{\zeta}\right)\\ &= \sum_{l \in \mathbb{Z}_K} P\left(\mathbf{x}_{1:t-1}, Z_{t-1} = l|\boldsymbol{\zeta}\right) P\left(x_t|\mathbf{x}_{1:t-1}, Z_{t-1} = l, Z_t = k; \boldsymbol{\zeta}\right)\\ &\times P\left(Z_t = k|Z_{t-1} = l\right)\\ &= \sum_{l \in \mathbb{Z}_K} \alpha_l(t-1) P(x_t|Z_t, \boldsymbol{\theta}) P\left(Z_t = k|Z_{t-1} = l\right)\\ &= b_k(t) \sum_{l \in \mathbb{Z}_K} \alpha_l(t-1) P\left(Z_t = k|Z_{t-1} = l\right).\end{aligned} \tag{2.14}$$

Now, the likelihood can be derived from (2.14),

$$\begin{aligned}L_{\mathbf{x}}(\boldsymbol{\zeta}) &= P(\mathbf{X}_{1:T} = \mathbf{x}_{1:T}|\boldsymbol{\zeta})\\ &= \sum_{k \in \mathbb{Z}_K} \pi_{Z_1} P_{Z_1}(x_1|\boldsymbol{\theta}) \prod_{t=2}^{T} \tau_{Z_{(t-1)}, Z_t}(t) P_{Z_t}(x_t|\boldsymbol{\theta})\\ &= \sum_{k \in \mathbb{Z}_K} \alpha_k(T).\end{aligned} \tag{2.15}$$

For this reason, the forward sum recursion is often referred as the likelihood recursion.

2.2.3 Bayesian parameter and state estimation

The Bayesian parameter and state estimation of HMMs uses the strategy of a two-stage Gibbs sampler which simulates from the joint posterior distribution of the HMM parameters and hidden states by alternating between sampling the hidden states $\mathbf{Z}$ given $\boldsymbol{\zeta}$ and $\mathbf{x}$ from the conditional posterior distribution $p(\mathbf{Z}|\boldsymbol{\zeta}, \mathbf{x})$ and sampling the HMM parameters $\boldsymbol{\zeta}$ given the complete data $(\mathbf{x}, \mathbf{Z})$ from the conditional posterior distribution $p(\boldsymbol{\zeta}|\mathbf{Z}, \mathbf{x})$.

The HMM model parameters and hidden states are thus sampled from their corresponding full conditional distributions. I can sample the hidden states from the conditional posterior distribution in two ways. The first method is called the *Direct Gibbs* sampler (Scott, 2002). It is just a general version of the Gibbs sampler. Now, to sample from the conditional posterior distribution of the hidden states, the *Direct Gibbs* sampler treats every state as an individual parameter then simulates each state Z_t from its full conditional distribution for all $t = 2, \ldots, T$,

$$P(Z_t = k|\mathbf{Z}_{-t}, \mathbf{x}, \boldsymbol{\zeta}) \propto \tau_{Z_{t-1},k}\tau_{k,Z_{t+1}}P(x_t|Z_t, \boldsymbol{\theta}). \tag{2.16}$$

The *Direct Gibbs* sampler step can be coupled with Gibbs sampler to generate samples of the HMM parameters $\boldsymbol{\zeta}$ from the conditional posterior distribution $p(\boldsymbol{\zeta}|\mathbf{Z}, \mathbf{x})$, which returns values of parameter updates and hidden states at every iteration. Consider the i^{th} iteration, for $i = 1, \ldots, I$ to sample an update of the parameter $\boldsymbol{\zeta}^{(i)}$ and $\mathbf{Z}^{(i)}$ from an MCMC process whose limiting distribution is $p(\boldsymbol{\zeta}, \mathbf{Z}|\mathbf{x})$. Since the Gibbs sampler directly depends on the parameters and their dimensionality, considering every single Z_t as a separate parameter increases the dimension of parameters and can cause algorithmic inefficiency.

The idea of the *Forward Gibbs* sampler was first conceived by Chib (1996) and later it was developed by Scott (2002). Here all the states were treated as one block for updating and then implementing the forward sum recursions as described in Section 2.2.2. The important characteristic of the *Forward Gibbs* sampler is that it can directly sample the hidden states $\mathbf{Z}$ from the conditional posterior distribution $p(\mathbf{Z}|\boldsymbol{\zeta}, \mathbf{x})$, whereas the *Direct Gibbs* sampler can only sample from the full conditionals of each Z_t (2.16). One needs to update one block of hidden states parameter $\mathbf{Z}$ rather than updating all the T hidden states separately.

2.2.4 Backward sampling

The goal of this step is to update the states from the posterior distribution of all the states conditional on data and parameters. This Backward sampling procedure requires computation of the Forward-sum probabilities described in Section 2.2.2.

The conditional posterior distribution of the hidden states $\mathbf{Z}$ given $\boldsymbol{\zeta}$ and $\mathbf{x}$ can be written as,

$$\begin{aligned} P(\mathbf{Z}\,|\boldsymbol{\zeta}, \mathbf{x}) &= P(\mathbf{Z}_{1:T}\,|\mathbf{x}_{1:T}, \boldsymbol{\zeta}) \\ &= P(Z_T|\mathbf{x}_{1:T}, \boldsymbol{\zeta}) \cdots P(Z_t|\mathbf{x}_{1:T}; Z_{t+1:T}; \boldsymbol{\zeta}) \cdots P(Z_1|\mathbf{x}_{1:T}; Z_{2:T}; \boldsymbol{\zeta}). \end{aligned} \tag{2.17}$$

The t^{th} term in (2.17) can be written as,

$$P(Z_t|\mathbf{x}_{1:T}; Z_{t+1:T}; \boldsymbol{\zeta}) \propto P(Z_t|\mathbf{x}_{1:t}; \boldsymbol{\zeta}) P(\mathbf{x}_{t+1:T}; \mathbf{Z}_{t+1:T}|\mathbf{x}_{1:t}, Z_t; \boldsymbol{\zeta}). \tag{2.18}$$

The states $Z_t,\ t = (1, \ldots, T)$ can now be updated using a *backward sampling* imputation step:

$$\begin{aligned}
&\text{Sample } Z_T \text{ from } P\left(Z_T = k|\mathbf{x}_{1:T}; \boldsymbol{\zeta}\right) = \frac{\alpha_k(T)}{\sum_k \alpha_k(T)}. \\
&\qquad\qquad \vdots \\
&\text{Sample } Z_t \text{ from } P\left(Z_t = k|\mathbf{x}_{1:T}; \mathbf{Z}_{t+1:T}; \boldsymbol{\zeta}\right) \propto \alpha_k(t) P\left(Z_{t+1}|Z_t = k\right). \qquad (2.19) \\
&\qquad\qquad \vdots \\
&\text{Sample } Z_1 \text{ from } P\left(Z_1 = k|\mathbf{x}_{1:T}; \mathbf{Z}_{2:T}; \boldsymbol{\zeta}\right) \propto \alpha_k(1) P\left(Z_2|Z_1 = k\right). \qquad (2.20)
\end{aligned}$$

I first sample Z_T and use this updated information recursively in order to sample the remaining states. Thus, after sampling Z_T, the remaining hidden states $\mathbf{Z}_{-T} = (Z_{T-1}, \ldots, Z_1)$ can be sampled by going backwards and updating the general t^{th} term Z_t from $P(Z_t|\mathbf{x}_{1:T}; Z_{t+1:T}; \boldsymbol{\zeta})$ for $t = T-1, \ldots, 1$.

An advantage of using Backward sampling compared to *Direct Gibbs* is that it allows more rapid mixing as the Markov chain has fewer components. Subsequently, the dependence of every hidden state on its previous updated value can be significantly minimized by directly sampling from $P(\mathbf{Z}|\mathbf{x}, \boldsymbol{\zeta})$. The emission and transition parameters $\boldsymbol{\zeta}$ are also sampled using either a M-H or Gibbs sampler conditional on the updates states.

2.2.5 Identifiability and label switching

A mixture model is a special case of a HMM, where the hidden states are assumed to be independent. In general, HMMs including mixture models often suffer from the label switching problem when the parameters are estimated using MCMC techniques. There have been many approaches developed in the recent past in order to tackle the label switching problem. The values of the parameters adjust themselves to suitable modal values and cause label switching. Label switching

emerges mostly when one has exchangeable priors for all the parameters. The symmetric nature of the priors can cause non-identifiability. Non-identifiability means that more than one set of parameter values can lead to the same likelihood. It can be proved that the parameters of a HMM/mixture model are identifiable (Leroux, 1992).

In Bayesian mixture models/HMMs literature, popular approaches for dealing with the label switching problem include constraints on the prior distributions of the parameters which cause rejection of the proposed values of the parameters that do not comply with the prior constraint assumptions (Richardson and Green, 1997). The label switching problem in HMMs can also be tackled by choosing the initial values of the MCMC updates empirically using empirical measures from the data or by using method of moments even if I have uninformative or weakly informative priors. In addition, a decision theoretic approach was proposed in Stephens (2000) using the Kullback-Leibler divergence method that minimises the expected posterior loss under a class of loss functions and calculate the marginal distributions of the parameters. I can also fix the label switching by ordering the means in my prior specification.

In this thesis, I implement an efficient classification based relabelling algorithm in HMMs proposed by Cron and West (2011). The idea of the algorithm can be explained as below:

- Given the current parameter draw $\boldsymbol{\zeta}$, define the corresponding hidden states $\hat{\mathbf{Z}}$ with T elements, such that $\hat{\mathbf{Z}} = \left(\hat{Z}_1, \cdots, \hat{Z}_T\right)$, where $\hat{\mathbf{Z}}$ assigns each observation to its modal component under the current set of classification probabilities.
- $\hat{\mathbf{Z}}^R = \left(\hat{Z}_1^R, \cdots, \hat{Z}_T^R\right)$ as the corresponding hidden states (classification vector) with elements $\hat{Z}_t^R$.

- Loss function: The misclassifications that $\hat{\mathbf{Z}}$ implies relative to $\hat{\mathbf{Z}}^R$ leads to a natural, intuitive loss function, such that, permuting the component labels in $\mathbf{Z}$ to maximize the match with $\hat{\mathbf{Z}}^R$ minimizes the misclassification.

- Define a $K \times K$ misclassification matrix $\mathbf{C}$,

$$C_{hj} = |\{(\hat{Z}_t^R = h \wedge \hat{Z}_t = j)\}|, \ (j, \ h = 1, \ldots, K) \text{ and } t = 1, \ldots, T. \quad (2.21)$$

  This matrix contains full information on sample and component classifications to compare the current MCMC state $\hat{\mathbf{Z}}$ with a reference $\hat{\mathbf{Z}}^R$, and can be computed even with very large sample sizes.

- C_{hj} counts misclassified observations MCMC component j is matched with reference component h, thus a column permutation is required to minimize $tr(\mathbf{C})$.

- This technique applied in Cron and West (2011) can be implemented efficiently using the Hungarian algorithm (Munkres, 1957).

2.2.6 Relabelling algorithm

The online relabelling algorithm proposed by Cron and West (2011) can be implemented completely on-line. It computes the optimal component permutations to minimize referenced misclassification costs at each MCMC iterate. The summary of the algorithm is provided as below:

- Calculate $\hat{\mathbf{Z}}$ given the current MCMC iterate $\boldsymbol{\zeta}$.

- Calculate the misclassifcation cost matrix $\mathbf{C}$.

- Apply the Hungarian algorithm to match the optimal permutation of component indices denoted by $\sigma(1:K)$, in the current MCMC draw.

- Permute $\boldsymbol{\zeta}_{1:K} \rightarrow \boldsymbol{\zeta}_{\sigma(1:K)}$ accordingly.

- Move to the next MCMC iterate.

2.3 Bayesian model checking and selection

Model checking is critical to any statistical analysis in that it tries to verify that the model assumptions are reasonable and sufficient. A well-performed Bayesian analysis must therefore adhere to some competent model assessment techniques, so that the model provides plausible descriptions of the data. If there is more than one relevant model, model selection then becomes of interest to statisticians. These concepts and definitions are described in the following sections.

2.3.1 Posterior predictive model checking

To understand whether the model captures the data in a Bayesian context, I will perform some model adequacy assessment, called posterior predictive checking. A Bayesian model fit can be examined using the posterior predictive distribution (Gelman and Meng, 1998) and test statistics that can be a function of both the data and parameters. These test statistics are termed as discrepancy variables (Gelman and Stern, 2000) to highlight the purpose of assessing the discrepancy between the model and data, in contrast to checking the accuracy of the model.

The basic technique implemented by Gelman and Stern (2000) for checking the model fit is to simulate replicated data from the posterior predictive distribution.

The posterior predictive distribution is described as below.

I have earlier defined the likelihood $L_{\mathbf{x}}(\boldsymbol{\zeta})$, and the posterior distribution be $p(\boldsymbol{\zeta}|\mathbf{x})$. I also define $\boldsymbol{\zeta}^{(1)}, \cdots, \boldsymbol{\zeta}^{(I)}$ to be I simulated draws from the posterior distribution. Next, I generate data $\mathbf{x}^{(i)}$ according to the model assumptions based on the parameter updates $\boldsymbol{\zeta}^{(i)}$ at every MCMC iteration i, for $i = 1, \ldots, I$.

Now, I obtain I sets of replicated data $\left\{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(I)}\right\}$, and compare those with the observed data by using a discrepancy test-statistic. The posterior predictive distribution for the replicated data $\mathbf{x}^{rep}$ is

$$p(\mathbf{x}^{rep}|\mathbf{x}) = \int p(\mathbf{x}^{rep}|\boldsymbol{\zeta}, \mathbf{x}) p(\boldsymbol{\zeta}|\mathbf{x}) d\boldsymbol{\zeta}. \tag{2.22}$$

The computational steps of posterior predictive checking can be described as follows:

- Simulate $\boldsymbol{\zeta}$ from the posterior distribution $p(\boldsymbol{\zeta}|\mathbf{x})$.
- Simulate $\mathbf{x}^{rep}$ from the predictive distribution $p(\mathbf{x}^{rep}|\boldsymbol{\zeta}, \mathbf{x})$.
- Next, compare the data $\mathbf{x}$ to the replicated datasets $\mathbf{x}^{rep}$ using a discrepancy test-statistic.

If the model fit is reasonable, then the replicated data $\mathbf{x}^{rep}$ simulated under the specific model assumptions should be similar to the observed data, i.e., the observed data should look plausible under the posterior predictive distribution.

The discrepancy between model and data can be measured by quantifying a discrepancy variable $\mathbf{T}(\mathbf{x}, \boldsymbol{\zeta})$, which is a scalar summary of the model parameters and data.

2.3.2 Posterior predictive p-values

As already explained at the beginning, the discrepancy test statistics can be functions of the parameters since they are calculated from the simulated posterior draws of the parameters at every iteration. The posterior predictive p-value can be defined as the probability that the discrepancy test statistic based on the replicated data and posterior draws of the parameters exceeds the discrepancy test statistic based on the observed data and posterior draws of the parameters,

as denoted by,

$$p_d = P(\mathbf{T}(\mathbf{x}^{rep}, \boldsymbol{\zeta}) \geq \mathbf{T}(\mathbf{x}, \boldsymbol{\zeta})|\mathbf{x}), \tag{2.23}$$

where the probability in Equation (2.23) is calculated over the posterior draws of $\boldsymbol{\zeta}$ and the posterior predictive distribution of $\mathbf{x}^{rep}$, which is basically the joint distribution, $p(\boldsymbol{\zeta}, \mathbf{x}^{rep}|\mathbf{x})$. For practical purposes, I calculate the posterior predictive distribution using posterior draws of $\boldsymbol{\zeta}$ for I simulations and then generate $\mathbf{x}^{rep(i)}$ from the predictive distribution for each posterior draw of $\boldsymbol{\zeta}^{(i)}$ for $i = 1, \ldots, I$, i.e., I generate I replicated draws of $\mathbf{x}^{rep}$ from the joint posterior distribution $p(\boldsymbol{\zeta}, \mathbf{x}^{rep}|\mathbf{x})$. Thus, to estimate the posterior predictive p-value, I must compute the proportion of times in which the discrepancy test statistic based on the replicated data and posterior draws of the parameters exceeds the discrepancy test statistic based on the observed data and posterior draws of the parameters,

$$p_d = \frac{1}{I} \sum_{i=1}^{I} \mathbf{I}\Big(\mathbf{T}(\mathbf{x}^{rep}, \boldsymbol{\zeta}) \geq \mathbf{T}(\mathbf{x}, \boldsymbol{\zeta})|\mathbf{x}\Big), \tag{2.24}$$

where $\mathbf{I}(.)$ is the indicator function.

A visual check can be performed by a scatter plot of the realised values $\mathbf{T}(\mathbf{x}, \boldsymbol{\zeta})$ against the replicated values $\mathbf{T}(\mathbf{x}^{rep}, \boldsymbol{\zeta})$. For a good fit, about half the points would be expected to fall above the line of equality and half to fall below it.

2.3.3 Deviance Information Criterion

The Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002) is a popular tool for Bayesian model comparison and is defined in terms of deviance,

$$DIC_1 = -2 \log P(\mathbf{x} \,\big|\, \tilde{\boldsymbol{\zeta}}) + 2p_{DIC_1}, \tag{2.25}$$

where $\tilde{\boldsymbol{\zeta}}$ are the posterior means, i.e., $\tilde{\boldsymbol{\zeta}} = E_{post}(\boldsymbol{\zeta}|\mathbf{x})$ where $E_{post}(\boldsymbol{\zeta}|\mathbf{x})$ can be estimated from $\widehat{E}_{post}(\boldsymbol{\zeta}|\mathbf{x}) = \frac{1}{I}\sum_{i=1}^{I}\boldsymbol{\zeta}^{(i)}$ and p_{DIC_1} is the effective number of parameters which can be found from,

$$p_{DIC_1} = 2\left(\log P\left(\mathbf{x}|\tilde{\boldsymbol{\zeta}}\right) - E_{post}\left(\log P(\mathbf{x}|\boldsymbol{\zeta})\right)\right). \tag{2.26}$$

For model comparison, DIC with the lowest numerical value indicates the best performing model. The expectation term in Equation (2.26) is the posterior expectation of $\log P(\mathbf{x}|\boldsymbol{\zeta})$ which can be estimated from the average of $\log P(\mathbf{x}|\boldsymbol{\zeta})$ over the posterior draws of $\boldsymbol{\zeta}$, so that, the computed version of p_{DIC_1} using the simulated draws of $\boldsymbol{\zeta}$ can be expressed as,

$$p_{DIC_1} = 2\left(\log P(\mathbf{x}|\tilde{\boldsymbol{\zeta}}) - \frac{1}{I}\sum_{i=1}^{I}\log P(\mathbf{x}|\boldsymbol{\zeta}^{(i)})\right). \tag{2.27}$$

An alternative version of p_{DIC_1} (Gelman et al., 2014) is,

$$p_{DIC_{1\,alt}} = 2\,\mathrm{Var}_{post}\left(\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}\right)\right)\right). \tag{2.28}$$

Even though p_{DIC_1} is numerically stable, the alternative version of $p_{DIC_{1\,alt}}$ has the benefit of always being positive. I have only used the alternative definition of effective number of free parameters, i.e., $p_{DIC_{1\,alt}}$. Thus, DIC can be re-evaluated as,

$$DIC_1 = -2\log P(\mathbf{x}|\tilde{\boldsymbol{\zeta}}) + 2p_{DIC_{1\,alt}}, \tag{2.29}$$

where $P(\mathbf{x}\Big|\tilde{\boldsymbol{\zeta}}) = L_{\mathbf{x}}(\tilde{\boldsymbol{\zeta}})$.
Now, $\text{Var}_{post}\left(\log(P(\mathbf{x}|\boldsymbol{\zeta}))\right)$ can be estimated from $\widehat{\text{Var}}_{post}\left(\log(P(\mathbf{x}|\boldsymbol{\zeta}))\right)$, such that

$$\widehat{\text{Var}}_{post}\left(\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}\right)\right)\right) = \frac{1}{I-1}\sum_{i=1}^{I}\left[\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}^{(i)}\right)\right) - \frac{1}{I}\sum_{i=1}^{I}\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}^{(i)}\right)\right)\right]^2, \tag{2.30}$$

where $\boldsymbol{\zeta}^{(i)}$ is the i^{th} posterior draw of the parameter $\boldsymbol{\zeta}$ and $\tilde{\boldsymbol{\zeta}}$ is the posterior estimate (mean) of $\boldsymbol{\zeta}$, averaged over the total number of iterations after burn-in.

Furthermore, the computations and definitions have also been explored for the family of latent variable models (Celeux et al., 2006) which also includes mixture models and HMMs. However, various studies have advised against the use of the DIC with data augmentation for comparing latent variable models. Li et al. (2012) claim that DIC must not be used with data-augmentation as the augmented data is non-regular and does not validate the asymptotic properties that are required for the DIC.
The DIC expression for complete-data $(\mathbf{x}, \mathbf{Z})$ can be defined as,

$$DIC_2 = -2\log P(\mathbf{x}, \mathbf{Z}|\tilde{\boldsymbol{\zeta}}) + 2p_{DIC_{2alt}}, \tag{2.31}$$

where $P(\mathbf{x}, \mathbf{Z}|\tilde{\boldsymbol{\zeta}}) = L_{\mathbf{x},\mathbf{Z}}(\tilde{\boldsymbol{\zeta}})$.
Now, $\text{Var}_{post}\left(\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}\right)\right)\right)$ can be estimated from $\widehat{\text{Var}}_{post}\left(\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}\right)\right)\right)$,

$$\widehat{\text{Var}}_{post}\left(\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}\right)\right)\right) = \frac{1}{I-1}\sum_{i=1}^{I}\left[\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}^{(i)}\right)\right) - \frac{1}{I}\sum_{i=1}^{I}\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}^{(i)}\right)\right)\right]^2. \tag{2.32}$$

One of the most used version of computing the DIC is based on the conditional likelihood $L_{\mathbf{x}}(\tilde{\boldsymbol{\zeta}}, \mathbf{Z})$ in the context of HMM in the recent literature. Here, the latent variables $\mathbf{Z}$ are considered as an additional parameter in the construction

of DIC (Celeux et al., 2006). Now, the definition of DIC based on conditional likelihood can be computed as

$$DIC_3 = -2\log P(\mathbf{x}|\tilde{\boldsymbol{\zeta}}, \mathbf{Z}) + 2p_{DIC_{3alt}}, \tag{2.33}$$

where $P(\mathbf{x}|\tilde{\boldsymbol{\zeta}}, \mathbf{Z}) = L_{\mathbf{x}}(\tilde{\boldsymbol{\zeta}}, \mathbf{Z})$.
Now, $\text{Var}_{post}\left(\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}\right)\right)\right)$ can be estimated from $\widehat{\text{Var}}_{post}\left(\log\left(P\left(\mathbf{x}, \mathbf{Z}|\boldsymbol{\zeta}\right)\right)\right)$,

$$\begin{aligned}\widehat{\text{Var}}_{post}\left(\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}, \mathbf{Z}\right)\right)\right) = \frac{1}{I-1}\sum_{i=1}^{I}\Bigg[&\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}^{(i)}, \mathbf{Z}^{(i)}\right)\right) \\ &- \frac{1}{I}\sum_{i=1}^{I}\log\left(P\left(\mathbf{x}|\boldsymbol{\zeta}^{(i)}, \mathbf{Z}^{(i)}\right)\right)\Bigg]^2.\end{aligned} \tag{2.34}$$

For Poisson model comparisons, Millar (2009) concludes that the DIC computed based on the conditional likelihood, obtained by conditioning on the latent variables and parameters $(\mathbf{x}|\mathbf{Z}, \boldsymbol{\zeta})$, usually prefers the Poisson-Gamma model instead of the Poisson-log-Normal model, even though the latter is the base model from which the data are generated from. Contrary to this, Millar (2009) also established the fact that DIC calculated using the integrated likelihood, i.e., $L(\tilde{\boldsymbol{\zeta}})$, obtained by integrating out the latent variables appears to perform well in comparison to the conditional likelihood. The DIC performance using the integrated likelihood is not unexpected as the standard asymptotic properties for validating the DIC are based on the integrated likelihood based DIC. However, in my applications in Chapter 6, I have computed the last versions of DIC, i.e., DIC_2 and DIC_3. Chan and Grant (2016) set $\tilde{\boldsymbol{\zeta}}$ to be the posterior mode of $\boldsymbol{\zeta}$. However, in my applications, I have used the posterior mean of $\boldsymbol{\zeta}$.

2.3.4 Widely Applicable Information Criterion

The Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) is another measure for selecting the most appropriate model and it has an edge over

DIC as it does not depend on the posterior point estimates of the parameters but rather averages over the posterior distribution. WAIC is given by

$$WAIC = -2\text{lppd} + 2p_{WAIC}, \tag{2.35}$$

where lppd (log-pointwise predictive density) can be estimated from

$$\text{computed lppd} = \sum_{t=1}^{T} \log\left(\frac{1}{I}\sum_{i=1}^{I} P(x_t|Z_t^{(i)}, \boldsymbol{\zeta}^{(i)})\right). \tag{2.36}$$

p_{WAIC} is the effective number of free parameters and can be computed as

$$p_{WAIC} = \sum_{t=1}^{T} \text{Var}_{post}\left(\log\left(P\left(x_t|Z_t, \boldsymbol{\zeta}\right)\right)\right), \tag{2.37}$$

where $\text{Var}_{post}\left(\log\left(P\left(x_t|Z_t, \boldsymbol{\zeta}\right)\right)\right)$ can be estimated from $\widehat{\text{Var}}_{post}\left(\log\left(P\left(x_t|Z_t, \boldsymbol{\zeta}\right)\right)\right)$.

$$\begin{aligned}\widehat{\text{Var}}_{post}\left(\log\left(P\left(x_t|Z_t, \boldsymbol{\zeta}\right)\right)\right) = \frac{1}{I-1}\sum_{i=1}^{I}\Bigg[&\log\left(P\left(x_t|Z_t^{(i)}, \boldsymbol{\zeta}^{(i)}\right)\right) \\ &-\frac{1}{I}\sum_{i=1}^{I}\log\left(P\left(x_t|Z_t^{(i)}, \boldsymbol{\zeta}^{(i)}\right)\right)\Bigg]^2.\end{aligned} \tag{2.38}$$

Likewise DIC, the model with the smallest WAIC is to be preferred.

Chapter 3

BiSulfite-Sequencing Data and Differential Methylation Callers

In this chapter, I describe a new sequencing technology, BiSulfite-sequencing (BS-seq), that can determine DNA methylation profiles with higher resolution and greater sensitivity. We also give an overview of some available algorithms that can detect differential methylation patterns from BS-seq data.

3.1 BS-sequencing procedure

BS-seq is a high-throughput sequencing procedure that can ascertain DNA methylation patterns. It employs standard sequencing methods on bisulfite-treated genomic DNA to ascertain the methylation status at each CpG site. In the BS-seq technique, DNA is treated with bisulfite chemicals which convert the non-methylated Cs to Us and subsequently to Ts, but the methylated cytosines remain unaffected (Figure 3.1). Now, the converted DNA fragments are aligned using an appropriate alignment tool to read the methylation status of a nucleotide base. The total number of aligned reads determines the accuracy of the estimated methylation levels at each CpG site. Typically, the reads of the methylation status calculated by the BS-seq data are in percentages. Here, the percentage

measure computes the proportion of actual C bases in the reads that are aligned with respect to a given C base in the reference genome (e.g., hg19 assembly) multiplied by 100.

The explanations that a percentage measure is provided as a methylation score are as follows:

1. the likely sequencing errors in the high-throughput bisulfite sequencing experiments;
2. due to incomplete bisulfite conversions of the cytosines;
3. the most probable case, the heterogeneity of samples and the fact that most of the genome is diploid.

In general, BS-seq experiments have both test and control samples. The test samples are mostly obtained from the disease tissue (e.g., cancer tissue) whereas the control samples can be obtained from a healthy tissue (e.g., proliferating tissue).

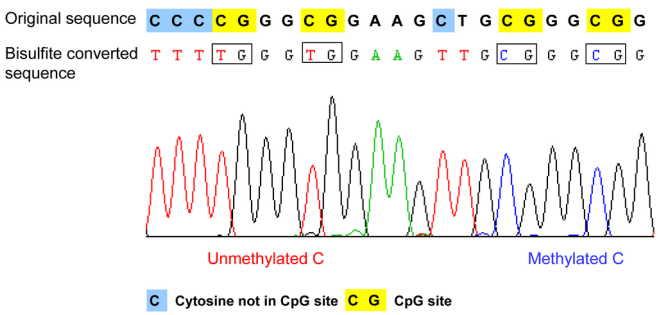


Figure 3.1: Bisulfite sequencing result of a single read. Figure taken from Yingying and Jeltsch (2010). After bisulfite conversion, the unmethylated cytosines are converted to thymines, and the methylated cytosines remain as cytosines. The methylated cytosine and unmethylated cytosine can be distinguished according to the sequencing result. Original sequence: DNA sequence before bisulfite treatment.

3.2 BS-sequencing tools

Several tools are available for the analysis of epigenomic datasets, especially BS-seq datasets. Tools to accurately analyze bisulfite-induced DNA are periodically being improved. These tools not only differ to a notably large extent in terms of their alignment technique, robustness and computational cost but also in the amount of information they generate. The latest tools also produce exhaustive methylation output, which in turn allow the end user to investigate the epigenomic effects of methylation more swiftly due to their ease of use. Two considerations are pivotal when ascertaining the methylation state of a read from a BS-seq experiment.

- The sequence of the read must be correctly derived entirely from a bisulfite-converted sequence in the original genome.
- The read must be mapped correctly to its corresponding position of the reference genome.

The methylation state of genomic positions involving Cs in the reference genome sequence can be inferred once a dataset of best alignments has been assigned. In the next subsection, we provide a brief description of one such popular tool, Bismark, which has been extensively used in recent BS-seq data analysis for its robust alignment procedure and methylation calling performance.

3.2.1 Bismark

Bismark (Krueger and Andrews, 2011) is a versatile tool for the analysis of BS-seq data. It carries out both read mapping and methylation calling in a single step. Furthermore, the methylation state of each C position in the read can be determined by this software. The main objective of Bismark bisulfite mapping is to find a unique alignment by simultaneously running four alignment processes

as the strand identity of a bisulfite read may be unknown, in advance.
The alignment techniques of Bismark can be explained as below.

1. Bisulfite reads are converted into a C-to-T and a G-to-A version (which is an equivalent version of C-to-T on the reverse strand).

2. Employing four parallel instances of the short read aligner Bowtie, each read is aligned to equivalently pre-converted forms of the human reference genome, Figure 3.2 (A).

3. The strand origin of a bisulfite read can be uniquely determined using these read mapping.

Before the alignment process begins, residual Cs are converted in *silico* into a fully bisulfite-converted form. Mapping conducted using this technique accounts for partial methylation precisely. Furthermore, Figure 3.2 (B) shows that the methylation state of each C position in the read is determined using Bismark. Most previous BS-seq tools were mainly mapping applications. Thus, a huge amount of post-processing were required to extract the methylation information. However, Bismark generates an output of bisulfite mapping which can be explored further by researchers.

3.3 Differential methylation calling

In the past few years, several statistical tools have been developed for the analysis of BS-seq data. MethVisual (Zackay and Steinhoff, 2010) is an R/Bioconductor package, which has been developed for visualization and exploratory statistical analysis of BS-seq data. BiQ Analyzer HT (Lutsik et al., 2011) implements a locus-specific analysis and visualization of BS-seq data. Streamlined Analysis and Annotation Pipeline for Reduced Representation Bisulfite Sequencing (SAAP-RRBS) (Sun et al., 2012) is mainly designed for implementation of methylation summary statistics and annotation of CpG sites. However, few tools have been

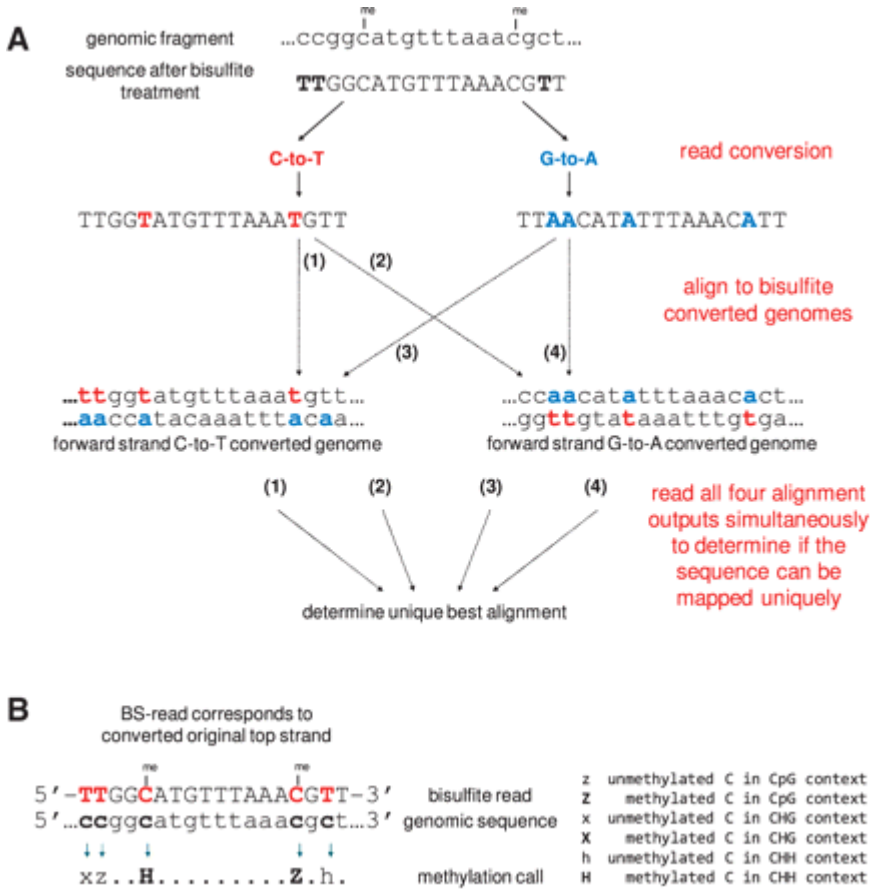


Figure 3.2: Bismark's approach (Krueger and Andrews, 2011) to mapping of bisulfite reads and calling methylation. (A) Reads from a BS-Seq experiment are transformed into a C-to-T and a G-to-A version and are then aligned to equivalently converted versions of the reference genome. The best alignment is then assessed from the four parallel alignment processes [in this example, the best alignment has no mismatches and comes from thread (1)]. (B) The methylation state of each C position in the read is determined by comparing the read sequence with the corresponding genomic sequence.

developed for the analysis of differential methylation. The most popular approach is to perform Fisher's exact test in a specific CpG window (Challen et al., 2012). BSmooth has been developed as a pipeline to detect DMRs in whole-genome BS data (Hansen et al., 2012). BSmooth essentially leans on smoothing the methylation values sample-wise and then testing for group differences via CpG-wise t-tests. DMRs are explained as adjacent CpG sites with absolute t-statistics above a defined cut-off value. The BiSeq package (Hebestreit et al., 2013) in Bioconductor is based on an algorithm which can detect DMRs. The package takes already aligned BS-seq data from one or multiple samples. The BiSeq package provides useful classes and functions to handle and analyze targeted BS-seq data such as reduced-representation BS-seq (RRBS) data. In particular, it implements an algorithm to detect DMRs. The package takes already aligned BS-seq data from one or multiple samples. The MethylSeekR package (Bioconductor/R) (Burger et al., 2013) is a computational tool that can identify the active regulatory regions based on the transcription binding which leads to defined reduction in DNA methylation. This package can accurately identify such functional regions from BS-seq data.

3.3.1 MethylKit

MethylKit (Akalin et al., 2012) is an all-inclusive R/Bioconductor package primarily designed to deal with sequencing data from RRBS data. In addition to that, it can also manage whole-genome bisulfite sequencing (WGBS) data and other variations of RRBS provided the proper data input format is created for the analysis. This R/Bioconductor package can efficiently tackle the high-throughput BS-seq data structure for the annotation and subsequent analysis of DNA methylation. The advantage of using methylKit is that it only requires a methylation score per base for any analysis. The main features of *methylKit* and the sequential relationship between them are as follows:

1. Reading the methylation calls from sorted Bismark alignments: Methyla-

tion percentage calls can be determined from sorted Sequence Alignment Map (SAM) format (Li et al., 2009) or Binary Alignment Map (BAM) alignment files from Bismark aligner and can be read into memory.

(a) Reading methylation call files. The data can be read into methylKit in two possible ways:

i. The methylKit can read the methylation scores from a typical methylation call text file as shown in Table 3.1.

ii. It can also read SAM format or BAM alignment files that are generated from Bismark.

(b) When a SAM file is provided, it processes the alignment file to obtain percent methylation scores and then methylKit can read that information into a flat file database, Table 3.2.

2. Merging samples from both groups: Most of the BS-seq data have test (e.g., cancer tissue) and control (e.g., normal tissue) samples and biological replicates. Merging samples from both the control and treatment group is an important database manipulation. Since I am interested in CpG sites, it is essential to merge reads on both strands of a CpG dinucleotide as it gives better coverage. Table 3.2 shows a methylBase object (flat file database) for differential methylation analysis using methylKit.

3. Differential Methylation calculation: In methylKit, two main methods have been implemented to identify differential methylation patterns across all regions.

(a) Logistic regression: In logistic regression, the number of methylated Cs and unmethylated Cs at a given region are specified for each sample. The logistic regression model is fitted in such a way that it can compare the fraction of methylated Cs for the treatment and control groups. The null hypothesis is that the methylation levels are the same in both

groups. Rejection of the null hypothesis is the same as declaring that CpG site or CpG region a DMC or DMR. On the other hand, if the null hypothesis is not rejected, it means that there is no statistically significant difference in methylation level between the two groups.

(b) Fisher's Exact test: Fisher's exact test is used to compare the fraction of methylated Cs in treatment and control samples in the absence of replicates.

The R/Bioconductor package implementation of methylKit decides between the choice of tests (Fisher's exact or logistic regression based test) based on the biological replicates per group. If there is only one sample at each CpG dinucleotide for both the groups, i.e., no biological replicate, then Fisher's exact test can be used. However, if there are multiple samples at each CpG dinucleotide for both the groups, i.e., there exists biological replicates, then the logistic regression based test is employed. Furthermore, multiple samples from the biological replicates can be pooled together to create one merged sample at each CpG dinucleotide for both the groups by summing the number of Ts and number of Cs across replicates with respect to their CpG sites in each group. Subsequently, Fisher's exact test can then be applied. In addition, methylKit also implements the sliding linear model (SLIM) method to adjust p-values to q-values (Wang et al., 2011) and eventually corrects for the problem of multiple hypothesis testing.

3.3.2 DSS

1. DSS-single (Wu et al., 2015) is mainly designed for the detection of DMRs from WGBS data for two groups without replicates. The BS-seq methylation data as explained by the authors is described below.

   Let X_{tj} be the methylated count and N_{tj} be the total count at the t^{th} CpG site and j^{th} treatment group for $t = 1, \ldots, T$ and $j = 1, 2$. The true

	chrBase	chr	base	strand	coverage	freqC	freqT
1	chr21.9826907	chr21	9826907	F	96	18.75	81.25
2	chr21.9853326	chr21	9853326	F	16	87.50	12.50
3	chr21.9853296	chr21	9853296	F	18	88.89	11.11
4	chr21.9860126	chr21	9860126	F	83	100.00	0.00
5	chr21.9906663	chr21	9906663	R	14	92.86	7.14

Table 3.1: A typical methylation call text file includes a unique identifier (chrBase), chromosome name (chr), strand information (F denotes forward and R denotes reverse strand), read coverage (coverage), percent of C (methylated cytosines) bases (freqC) and percent of T (unmethylated cytosines which eventually transformed into thymines (Ts) after bisulfite treatment) bases (freqT) at that particular genomic base.

underlying methylation proportion is denoted by p_{tj}. Feng et al. (2014) showed that it is reasonable to assume that X_{tj} follows a Beta-Binomial distribution, which encapsulates both the biological and technical variations in the counts. The Beta distribution is parametrized by its mean (μ_{tj}) and dispersion (φ_{tj}), where φ_{tj} denotes the biological variation among replicates in the same treatment group.

A log-normal prior is imposed on φ_{tj} in order to gather information from all CpG sites in estimating the site-specific dispersions. The mean of the beta distribution is assumed to vary across the genome. To incorporate the spatial correlation in the methylation levels, it has been assumed $\mu_{tj} = f_j(l_t)$, where l_t denotes the genomic co-ordinate of the t^{th} CpG site and f_j is a smoothing function. A simple moving average procedure is applied on the collapsed counts to estimate f_j. The final hierarchical structure for modelling the BS-seq data under this set up is given below.

$$\begin{aligned} X_{tj}|N_{tj}, p_{tj} &\sim Bin(N_{tj}, p_{tj}) \\ p_{tj}|\mu_{tj}, \varphi_{tj} &\sim Beta(\mu_{tj}, \varphi_{tj}) \\ \varphi_{tj} &\sim \log N(m_{j0}, r_{j0}^2), \end{aligned} \tag{3.1}$$

	chr	start	end	strand	coverage1	numCs1	numTs1	coverage2	numCs2	numTs2	coverage3	numCs3	numTs3	coverage4	numCs4	numTs4
1	chr21	9853296	9853296	+	17	10	7	333	268	65	18	16	2	395	341	54
2	chr21	9853326	9853326	+	17	12	5	329	249	79	16	14	2	379	284	95
3	chr21	9860126	9860126	+	39	38	1	83	78	5	83	83	0	41	40	1
4	chr21	9906604	9906604	+	68	42	26	111	97	14	23	18	5	37	33	4
5	chr21	9906616	9906616	+	68	52	16	111	104	7	23	14	9	37	27	10
6	chr21	9906619	9906619	+	68	59	9	111	109	2	22	18	4	37	29	8

Table 3.2: The flat database text file includes chromosome name (chr), start position (start), end position (end), count of read coverage of treatment group 1 (coverage1), count of methylated Cs of treatment group 1(numCs1), count of unmethylated Cs of treatment group 1(numTs1), count of read coverage of treatment group 2 (coverage2), count of methylated Cs of treatment group 2(numCs2), count of unmethylated Cs of treatment group 2(numTs2), count of read coverage of control group 1 (coverage3), count of methylated Cs of control group 1(numCs3), count of unmethylated Cs of control group 1(numTs3), count of read coverage of control group 2 (coverage4), count of methylated Cs of control group 2 (numCs4), count of unmethylated Cs of control group 2 (numTs4) at that particular genomic base. The positive strand of DNA contains the information for creating the sequence of a protein whereas the negative strand contains the complementary sequence according to base-pair rules (as A binds with T and C with G in a dinucleotide base). The negative strand is usually not transcribed into RNA and subsequently translated into protein.

where m_{j0} and r_{j0}^2 are hyperparameters. In (3.1), the parameters (μ_{tj}, φ_{tj}) have the following relationship compared to the conventional Beta(α, β) parametrization:

$$\mu = \frac{\alpha}{\alpha + \beta},$$
$$\varphi = \frac{1}{\alpha + \beta + 1}.$$

Table 3.3 shows the data inputs for the DMR detection algorithm of DSS.

2. Statistical test procedure: After estimating the hyperparameters through an empirical Bayes (EB) procedure (Feng et al., 2014), DML (differentially methylated loci) or DMC can be identified by a hypothesis test: $H_0 : \mu_{t1} = \mu_{t2}$ for the equality of the mean methylation levels at each CpG site. Wald's test is employed to compare the mean methylation levels at each CpG site, and p-values are evaluated from the test statistics.

3. In an extension to DSS-single, Park and Wu (2016) developed DSS-general to model BS-seq data under a more general multifactor experimental design. The data input more or less remains the same as described by Wu et al. (2015) except that the idea of the treatment group is extended to a generalized multifactor dataset.

3.4 Data

In this thesis, I have analysed a dataset from a study of methylation changes in human ageing provided by the Adams' lab, Beatson Cancer Research Institute, Glasgow. It contains the pooled dataset of three biological replicates for proliferating and senescent IMR90 cells. The BS-seq data has information about

	chr	pos	N	X
1	chr18	3014904	26	2
2	chr18	3031032	33	12
3	chr18	3031044	33	13
4	chr18	3031065	48	24
5	chr18	3031069	17	4
6	chr18	3031082	93	37

Table 3.3: The text file includes chromosome name (chr), genomic position (CpG site), count of read coverage of one group (N), count of methylated Cs of treatment one group(X).

methylation for all the chromosomes. The longest chromosome, i.e, Chromosome-1 contains $4,590,977$ CpG sites while the shortest Chromsome-Y contains $27,562$ CpG sites. On average, this dataset covers 1.8 million CpG sites per chromosome with an average sequencing depth of 10. Table 3.4 displays the data format of Cruickshanks et al. (2013). In the following chapters, I present my own met-

	chr	pos	x_p	y_p	x_s	y_s
1	chr21	9411551	16	35	6	53
2	chr21	9411552	22	51	9	74
3	chr21	9411783	6	21	1	23
4	chr21	9411784	11	29	6	39
5	chr21	9412098	8	11	8	10
6	chr21	9412099	18	13	11	13

Table 3.4: The text file includes chromosome name (chr), genomic position (pos: CpG site), count of methylated Cs of proliferating cells (x_p), count of unmethylated Cs of proliferating cells (y_p), count of methylated Cs of senescent cells (x_s), count of unmethylated Cs of senescent cells (y_s).

hod for detecting DMCs based on this dataset, which takes a different approach compared to the existing methods discussed in this chapter.

Chapter 4

Hierarchical Hidden Markov Models with Applications to BS-Seq Data

In this chapter, I propose Bayesian latent variable models for predicting DMCs on the basis of BS-seq data using a hierarchical hidden Markov model (HMM) framework. I developed Bayesian latent variable models that aim to incorporate many features of the data under a hierarchical framework. HMMs are quite popular in the analysis of biological datasets. A suitable model for this type of analysis is the HMM, where the evolution of a latent characteristic of interest is represented by an unobserved Markov chain. By imposing Markovian conditioning on the latent states, the model class becomes richer than mixture models where the states are considered to be independent, since the Markovian property of HMM can induce long-range conditional dependencies in the observed data.

I employ Bayesian techniques to estimate the HMM parameters and infer the hidden states. In the following sections, I describe the model assumptions and structure of my HMMs and demonstrate an efficient method for applying MCMC

techniques to both simulated as well as real data.

4.1 Model assumptions

The main objective is to infer genomic Cs with different levels of methylation between distinct cell types. I can denote whether a CpG site is differentially methylated or not using a latent variable.

BS-sequencing of methylated samples generates counts of methylated and unmethylated Cs. At present, I ignore the genomic position of each CpG site, and the fact that adjacent CpG sites are not equally spaced. To study and analyse methylation patterns, I assume two methylation states, i.e., a similarly methylated state and a differentially methylated state, corresponding to similar and differential methylation of CpG sites, respectively.

Let the BS-seq data (observed) be denoted by $\mathbf{x} = (\mathbf{x}^p,\ \mathbf{x}^s)$, such that $\mathbf{x}^p = (x_1^p, \cdots, x_T^p)$ and $\mathbf{x}^s = (x_1^s, \cdots, x_T^s)$, where x_t^p and x_t^s $(t = 1, \ldots, T)$ are the methylated counts of proliferating and senescent cells, respectively, for the t^{th} CpG site as described in Section 3.4. Furthermore, let $\mathbf{n} = (\mathbf{n}^p,\ \mathbf{n}^s)$, such that $\mathbf{n}^p = (n_1^p, \cdots, n_T^p)$ and $\mathbf{n}^s = (n_1^s, \cdots, n_T^s)$, where n_t^p and n_t^s $(t = 1, \ldots, T)$ are the total number of counts (methylated Cs and unmethylated Cs) of proliferating and senescent cells respectively, at the t^{th} CpG site. For each observation, I assume an unobserved state Z_t, $(t = 1, \ldots, T)$, where Z_t represents the t^{th} hidden state such that $Z_t = 1$, if the methylation levels in proliferating and senescent cells are the same at the t^{th} CpG site and $Z_t = 2$, if there is differential methylation between the two cell types at the t^{th} CpG site, where $\mathbf{Z} = (Z_1, \cdots, Z_T)$. Since the process of BS-seq involves the random sampling of two types of reads- methylated and unmethylated counts, the data will follow an independent bivariate Binomial distribution.

I assume x_t^p and x_t^s are the realizations of the pair of random variables X_t^p and X_t^s such that X_t^p and X_t^s independently follow Binomial distributions with parameters (n_t^p, p_t^p) and (n_t^s, p_t^s) respectively, such that,

$$X_t^p | p_t^p \sim Bin(n_t^p, p_t^p), \qquad t = 1, \ldots, T, \tag{4.1}$$

and

$$X_t^s | p_t^s \sim Bin(n_t^s, p_t^s), \qquad t = 1, \ldots, T, \tag{4.2}$$

where p_t^p and p_t^s are the probability parameters of methylation for proliferating and senescent cells, respectively, at the t^{th} CpG site. For notational simplicity, let the pair of random variables X_t^p and X_t^s be denoted by $\mathbf{X}_t = (X_t^p,\ X_t^s)$ and the pair of proportion parameters p_t^p and p_t^s be denoted by $\mathbf{p}_t = (p_t^p,\ p_t^s)$ for $t = 1, \ldots, T$. I also assume $\mathbf{X} = (\mathbf{X}^p,\ \mathbf{X}^s)$, where $\mathbf{X}^p = (X_1^p, \cdots, X_T^p)$ and $\mathbf{X}^s = (X_1^s, \cdots, X_T^s)$.

Now, I describe the emission densities. I implement a Beta-Binomial hierarchical model conditional on the true underlying methylation proportions and hidden states.

The underlying methylation proportions at CpG site t in state k for proliferating and senescent cells can be defined as p_t^{pk} and p_t^{sk}, respectively. Let $\mathbf{p}_t^k = (p_t^{pk}, p_t^{sk}),\ t = 1, \ldots, T$. Then,

$$X_t^p | p_t^p, Z_t = k \sim Bin(n_t^p, p_t^{pk}) \text{ and } X_t^s | p_t^s, Z_t = k \sim Bin(n_t^p, p_t^{sk}),\ t = 1, \ldots, T, \tag{4.3}$$

are independently distributed, where k takes a value of 1 or 2.

Equation (4.3) can be written more compactly as

$$X_t^p|p_t^{pk} \sim Bin(n_t^p, p_t^{pk}),\ \text{and}\ X_t^s|p_t^{sk} \sim Bin(n_t^p, p_t^{sk}),\ t = 1, \ldots, T. \tag{4.4}$$

The true underlying methylation proportions at the CpG site indexed t in state k for proliferating and senescent cells are assumed to follow Beta distributions at the second stage of the hierarchical model:

$$p_t^p|Z_t = 1 \sim Beta(\alpha, \beta), \qquad p_t^s|Z_t = 2 \sim Beta(\alpha, \beta) \tag{4.5}$$

$$p_t^s|Z_t = 1 \sim Beta(\gamma_1, \delta_1), \qquad p_t^s|Z_t = 2 \sim Beta(\gamma_2, \delta_2). \tag{4.6}$$

Define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ \boldsymbol{\theta}_2)$, where $\boldsymbol{\theta}_1 = (\alpha,\ \beta)$ and $\boldsymbol{\theta}_2 = (\gamma_1,\ \delta_1,\ \gamma_2,\ \delta_2)$.

Now, if the methylation pattern in proliferating and senescent cells is the same for state 1, i.e., $k = 1$, I assume $p_t^{p1} = p_t^{s1} = p_t^*$, say, for the unobserved state $Z_t = 1$ in the t^{th} CpG site, such that, $\mathbf{p}_t^1 = (p_t^*,\ p_t^*)$.

Similarly, if the methylation in proliferating and senescent cells is different for state 2, I assume, $\mathbf{p}_t^2 = (p_t^{p2},\ p_t^{s2})$ for the unobserved state $Z_t = 2$ in the t^{th} CpG site.

4.1.1 Binomial emission distributions of the model

The emission probability of the observation $\mathbf{x}_t = (x_t^p,\ x_t^s)$ conditional on the hidden state Z_t can be written as:

$$b_k(t) = P(\mathbf{x}_t|\mathbf{p}_t^k, Z_t = k),\ k = 1,\ 2 \text{ and } t = 1, \ldots, T. \tag{4.7}$$

Let X be a discrete random variable and follows Binomial distribution with parameters n and p. Then the probability mass function (p.m.f.) of the realized

value x of X is defined below.

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, \ldots, n$$
$$\equiv Bin(x; n, p). \tag{4.8}$$

The emission distributions of the model are as follows:

- The emission probability of the observation $\mathbf{x}_t = (x_t^p,\ x_t^s)$ conditional on the hidden state $Z_t = 1$ is given by,

$$b_1(t) = Bin(x_t^p; n_t^p, p_t^*) \times Bin(x_t^s; n_t^s, p_t^*). \tag{4.9}$$

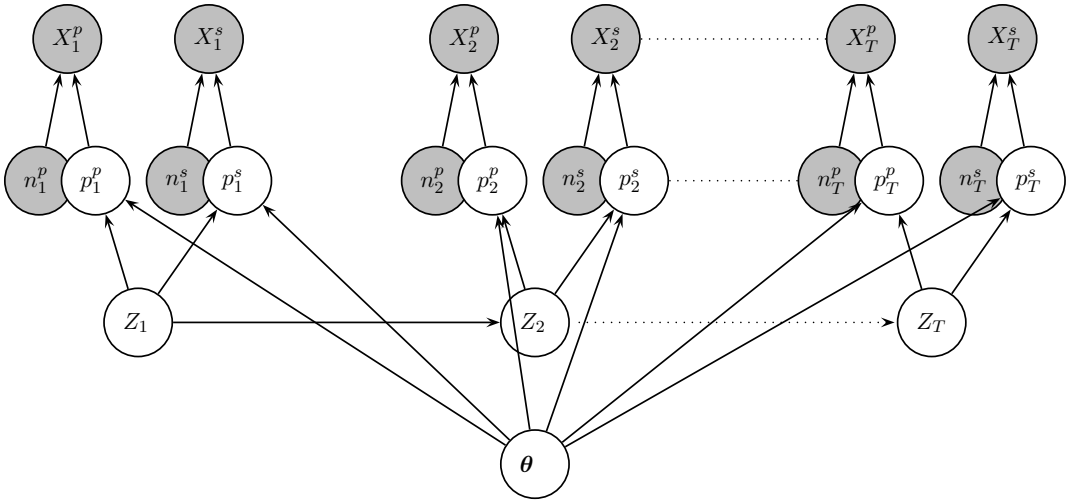
- Similarly, the emission probability of $\mathbf{x}_t = (x_t^p,\ x_t^s)$ conditional on the hidden state $Z_t = 2$ is given by,

$$b_2(t) = Bin(x_t^p; n_t^p, p_t^p) \times Bin(x_t^s; n_t^s, p_t^s). \tag{4.10}$$

4.1.2 Beta-Binomial emission distributions of the model

I use a two-level hierarchical model and assume Beta priors on p_t^p, p_t^s and p_t^*. Now, in the bivariate Beta-Binomial density, the effect of the nuisance parameters $p_t^p,\ p_t^s,\ p_t^*$ can be integrated out with respect to the relevant states due to conjugacy, leaving the hyperparameters of the conjugate prior distributions as the only parameters. This leads to computational efficiency in the proposed model (Figure 4.1).

By marginalizing the second level hierarchical model parameters, the emission



For: $t = 1, \ldots, T$

$$X_t^p \sim \begin{cases} \text{Bin}(n_t^p, p_t^p), & \text{if } Z_t = 1 \\ \text{Bin}(n_t^p, p_t^p), & \text{if } Z_t = 2 \end{cases} \qquad X_t^s \sim \begin{cases} \text{Bin}(n_t^s, p_t^s), & \text{if } Z_t = 1 \\ \text{Bin}(n_t^s, p_t^s), & \text{if } Z_t = 2 \end{cases}$$

$$p_t^p \sim \begin{cases} \text{Beta}(\alpha, \beta), & \text{if } Z_t = 1 \\ \text{Beta}(\gamma_1, \delta_1), & \text{if } Z_t = 2 \end{cases} \qquad p_t^s \sim \begin{cases} \text{Beta}(\alpha, \beta), & \text{if } Z_t = 1 \\ \text{Beta}(\gamma_2, \delta_2), & \text{if } Z_t = 2 \end{cases}$$

Figure 4.1: Graphical representation of the Beta-Binomial emission model. The grey circles refer to the fixed values of the total counts and data respectively, while the white circles refer to emission hyperparameters and hidden states that are inferred.

distributions can be written as:

$$\begin{aligned} P(x_t^p, x_t^s|\alpha,\beta; Z_t=1) &= \int_0^1 Bin(x_t^p; n_t^p, p_t^*) Bin(x_t^s; n_t^s, p_t^*) Beta(p_t^*; \alpha, \beta) dp_t^* \\ &= \binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{1}{\mathbf{B}(\alpha,\beta)}\int_0^1 \Big(p_t^{*(x_t^p+x_t^s+\alpha-1)} \\ &\qquad \times (1-p_t^*)^{(n_t^p+n_t^s-x_t^p-x_t^s+\beta-1)}\Big) dp_t^* \\ &= \binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{\mathbf{B}(x_t^p+x_t^s+\alpha, n_t^p+n_t^s-x_t^p-x_t^s+\beta)}{\mathbf{B}(\alpha,\beta)}, \end{aligned} \tag{4.11}$$

and,

$$\begin{aligned} P(x_t^p, x_t^s|\gamma_1,\delta_1,\gamma_2,\delta_2; Z_t=2) &= \int_0^1 Bin(x_t^p; n_t^p, p_t^p) Beta(p_t^p; \gamma_1, \delta_1) dp_t^p \\ &\quad \times \int_0^1 Bin(x_t^s; n_t^s, p_t^s) Beta(p_t^s; \gamma_2, \delta_2) dp_t^s \\ &= \binom{n_t^p}{x_t^p}\frac{1}{\mathbf{B}(\gamma_1,\delta_1)}\int_0^1 p_t^{s(x_t^p+\gamma_1-1)}(1-p_t^p)^{(n_t^p-x_t^p+\delta_1-1)} dp_t^p \\ &\quad \times \binom{n_t^s}{x_t^s}\frac{1}{\mathbf{B}(\gamma_2,\delta_2)}\int_0^1 p_t^{s(x_t^s+\gamma_2-1)}(1-p_t^s)^{(n_t^s-x_t^s+\delta_2-1)} dp_t^s \\ &= \binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{\mathbf{B}(x_t^p+\gamma_1, n_t^p-x_t^p+\delta_1)}{\mathbf{B}(\gamma_1,\delta_1)} \\ &\quad \times \frac{\mathbf{B}(x_t^s+\gamma_2, n_t^s-x_t^s+\delta_2)}{\mathbf{B}(\gamma_2,\delta_2)}, \end{aligned} \tag{4.12}$$

where $\mathbf{B}(a,b)$ is the Beta function,i.e., $\int_0^1 u^{a-1}(1-u)^{b-1}du$, where $a,\ b>0$.

4.1.3 Homogeneous transition model

The initial state distribution at the first CpG site is denoted as $P(Z_1=k)=\pi_k$ for $k=1,2$, with initial probabilities $\boldsymbol{\pi}=(\pi_1,\pi_2)$.

The transition probabilities between the states $\tau_{jk} = P(Z_t = k|Z_{t-1} = j)$ are given by the matrix $\boldsymbol{\tau}$. So, $\tau_{12} = 1 - \tau_{11}$ and $\tau_{22} = 1 - \tau_{21}$.

I denote the transition counts from state 1 to state 1, state 1 to state 2, state 2 to state 1, state 2 to state 2 as $t_{11}, t_{12}, t_{21}, t_{22}$ respectively. t_1 and t_2 are the total counts of state 1s and state 2s respectively. That is,
$t_{kl} = \sum_{t=2}^{T} I(Z_{t-1} = k, Z_t = l)$ and $t_k = \sum_{t=1}^{T} I(Z_t = k)$

The probability of the initial state Z_1 given π_1 is

$$P(Z_1|\pi_1) \propto {\pi_1}^{\mathbf{I}(Z_1=1)}(1-\pi_1)^{\mathbf{I}(Z_1=2)}. \tag{4.13}$$

The probability for the sequence of the hidden states $\mathbf{Z}_{2:T}$ conditional on the initial state Z_1 and the transition parameters is

$$\begin{aligned} P(\mathbf{Z}_{2:T}|Z_1, \boldsymbol{\tau}) &\propto P(Z_2|Z_1, \boldsymbol{\tau})P(Z_3|Z_2, \boldsymbol{\tau}) \dots P(Z_T|Z_{T-1}, \boldsymbol{\tau}) \\ &\propto \tau_{11}^{t_{11}}(1-\tau_{11})^{t_{12}}\tau_{21}^{t_{21}}(1-\tau_{21})^{t_{22}} \\ &\propto Bin(t_{11}; t_{11}+t_{12}, \tau_{11})Bin(t_{21}; t_{21}+t_{22}, \tau_{21}). \end{aligned} \tag{4.14}$$

4.1.4 Non-homogeneous transition model

In reality, there are unequal gaps between CpG sites in BS-seq data, which motivates me to introduce a non-homogeneous transition model leading to a continuous-index HMM. The only modification required for this model is to assume that the underlying methylation status to be a latent stochastic process emitting over a continuous genomic index, represented by $Z(c)$ for $c > 0$. $[Z(c),\ c > 0]$ is a continuous-index Markov process, assuming values in a finite state space 1, 2, i.e., a two state Markov process, such that, if, $Z(c) = 1(2)$, a similarly methylated state (differentially methylated state) is signalled for the CpG site. I define Ψ_t as the genomic distance (in base pairs) between two adjacent CpG sites at

genomic positions ψ_t and ψ_{t-1}, i.e., $\Psi_t = \psi_t - \psi_{t-1}$, such that, $\boldsymbol{\Psi} = (\Psi_1, \cdots, \Psi_T)$, where Ψ_1 is initialized to be 0.

In Section 4.1.3, the underlying structure of the methylation status was assumed to be a latent stochastic process emitting over a discrete genomic index, represented by Z_t for $t = 1, \ldots, T$.

I define a non-homogeneous transition probability $\tau_{jk}(\Psi_t)$ for $t = 2, \ldots, T$, as

$$P\Big(Z(\psi_t)|Z(\psi_{t-1}), \ldots, Z(\psi_1), X_{1:t-1}\Big) = P\Big(Z(\psi_t)|Z(\psi_{t-1})\Big) = \tau_{jk}(\Psi_t), \quad j, k = 1, 2, \tag{4.15}$$

the process was in state k at genomic position ψ_t conditional on the process being in state j at genomic position ψ_{t-1}. (4.15) clearly indicates that the transition probability depends on the gapped distance of the genomic positions between two adjacent CpG sites indexed by t and $(t-1)$.

For notational simplicity, I shall this time onwards refer to $Z(\psi_t)$ as Z_t. The probability of staying in a state is subsequently assumed to be linear with respect to the genomic index for an infinitesimal interval. The two-state hidden Markov process at genomic index t can be parameterized with transition rate parameters λ_1 and λ_2, where λ_1 and λ_2 are the transition rate parameters from a similarly methylated state to a differentially methylated state and from a differentially methylated state to a similarly methylated state, respectively.

The intensity matrix $\boldsymbol{\nu}$ of the transition rate parameters λ_1 and λ_2 is then given

by

$$\boldsymbol{\nu} = \begin{pmatrix} \lambda_1 & -\lambda_1 \\ -\lambda_2 & \lambda_2 \end{pmatrix}. \tag{4.16}$$

The transition probability matrix $\boldsymbol{\tau}(t)$ over genomic interval Ψ_t is calculated by the matrix exponential of $\boldsymbol{\nu}$ multiplied by Ψ_t, i.e., $\boldsymbol{\tau}(t) = \exp(\boldsymbol{\nu}\Psi_t)$. Hence, $\boldsymbol{\tau}(t)$ is represented by,

$$\boldsymbol{\tau}(t) = \begin{pmatrix} \tau_{11}(t) & \tau_{12}(t) \\ \tau_{21}(t) & \tau_{22}(t) \end{pmatrix}, \tag{4.17}$$

where the non-homogeneous transition probabilities at CpG site t over genomic interval Ψ_t are given by,

$$\begin{aligned} \tau_{11}(t) &= \tfrac{\lambda_2}{\lambda_1+\lambda_2} + \tfrac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t}, \\ \tau_{12}(t) &= \tfrac{\lambda_1}{\lambda_1+\lambda_2} - \tfrac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t}, \\ \tau_{21}(t) &= \tfrac{\lambda_2}{\lambda_1+\lambda_2} - \tfrac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t}, \\ \tau_{22}(t) &= \tfrac{\lambda_1}{\lambda_1+\lambda_2} + \tfrac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t}. \end{aligned} \tag{4.18}$$

Here also, the initial state distribution at the first CpG site is denoted as $P(Z_1 = k) = \pi_k$ for $k = 1, 2$, with initial probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2)$.
The initial state distribution of Z_1 is taken to be uniform.

$$P(Z_1 = k) = \pi_k = 0.5, \text{ for } k = 1,\ 2. \tag{4.19}$$

Now, the probability of the sequence of the hidden states $\mathbf{Z}_{2:T}$ given the initial state Z_1 and transition rate parameters can be factorized as,

$$\begin{aligned} P(\mathbf{Z}_{2:T}\,|Z_1, \boldsymbol{\tau}) &\propto P\left(Z_2|Z_1, \boldsymbol{\tau}(2)\right) P\left(Z_3|Z_2, \boldsymbol{\tau}(3)\right) \ldots P\left(Z_T|Z_{T-1}, \boldsymbol{\tau}(T)\right) \\ &\propto \prod_{t=2}^{T} \Big(\tau_{11}(t)^{\mathbf{I}(Z_{t-1}=1, Z_t=1)} \tau_{12}(t)^{\mathbf{I}(Z_{t-1}=1, Z_t=2)} \\ &\quad \times \tau_{21}(t)^{\mathbf{I}(Z_{t-1}=2, Z_t=1)} \tau_{22}(t)^{\mathbf{I}(Z_{t-1}=2, Z_t=2)} \Big). \end{aligned} \tag{4.20}$$

4.1.5 Beta-Binomial hierarchical HMMs

In this section, I describe the two hierarchical Beta-Binomial HMMs by combining the Beta-Binomial emission probability distributions and transition probability distributions.

- Model BBDM: this model combines the Beta-Binomial emission probability model in (4.11) and (4.12) with the homogeneous discrete-index transition probability model in (4.13) and (4.14).
- Model BBCM: this model combines the same Beta-Binomial emission probability model with the non-homogeneous continuous-index transition probability model in (4.19) and (4.20).

To simplify the notation in the following sections, I represent the hidden states as $\mathbf{Z}$, π_1 as the initial state transition parameter and $(\alpha, \beta, \gamma_1, \delta_1, \gamma_2, \delta_2)$ as emission parameters for both the models $BBDM$ and $BBCM$ even though the behaviour of the hidden states and the parameters are different in the two models. In addition, I describe the general version of the likelihood for model M where M represents the true model, i.e., $M = BBDM,\ BBCM$. However, to simplify the notational subscripts of the parameters, I use $M = D,\ C$, where D denotes $BBDM$ and C denotes $BBCM$. Thus, the transition parameters for model M are

assumed to be $\boldsymbol{\tau}^{(M)}$, such that $\boldsymbol{\tau}^{(D)} = (\tau_{11}, \tau_{21})$ for $BBDM$ and $\boldsymbol{\tau}^{(C)} = (\lambda_1, \lambda_2)$ for $BBCM$. Similarly, initial state parameters for model M are assumed to be $\boldsymbol{\pi}^{(M)} = (\pi_1^{(M)}, \pi_2^{(M)})$. The transition probability matrix is denoted by $\boldsymbol{\tau}^{(M)}(t)$ for model M, where $\tau_{kl}^{(M)}(t)$ is the $(k,l)^{th}$ element of $\boldsymbol{\tau}^{(M)}(t)$, such that the process was in state l at genomic index t conditional on the process being in state k at genomic index $t-1$.

4.1.6 Computing the likelihoods

In this section, let the set of all parameters and hyperparameters be generically denoted by $\boldsymbol{\zeta}^{(M)} = (\boldsymbol{\theta}^{(M)}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\pi}^{(M)})$ for both the models as described in Section 4.1.5 where $\boldsymbol{\theta}^{(M)} = (\boldsymbol{\theta}_1^{(M)}, \ \boldsymbol{\theta}_2^{(M)})$, such that $\boldsymbol{\theta}_1^{(M)} = (\alpha, \beta)$ and $\boldsymbol{\theta}_2^{(M)} = (\gamma_1, \delta_1, \gamma_2, \delta_2)$ and $\boldsymbol{\tau}^{(M)}$ for model M. The joint probability distribution of the observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and the sequence of the hidden states $\mathbf{Z} = (Z_1, \ldots, Z_T)$ for model M conditional on the model parameters $\boldsymbol{\zeta}^{(M)}$ is the complete data likelihood of the observations and the states:

$$P(\mathbf{x}, \mathbf{Z} \,|\boldsymbol{\zeta}^{(M)}) = \pi_{Z_1}^{(M)} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(M)}) \prod_{t=2}^{T} \tau_{Z_{(t-1)},Z_t}^{(M)}(t) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(M)}) \tag{4.21}$$

$$= \pi_{Z_1}^{(M)} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(M)}) \tau_{Z_1,Z_2}^{(M)}(2) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(M)}) \cdots$$
$$\times \ \tau_{Z_{(T-1)},Z_T}^{(M)}(T) P_{Z_T}(\mathbf{x}_T|\boldsymbol{\theta}^{(M)}), \tag{4.22}$$

where $P_k(\mathbf{x}_t|\boldsymbol{\theta}^{(M)}) = P(\mathbf{x}_t|Z_t = k; \boldsymbol{\theta}^{(M)})$, $\pi_k^{(M)} = P(Z_1 = k)$ and $\tau_{kl}^{(M)}(t) = P(Z_t = l|Z_{t-1} = k; \boldsymbol{\tau}^{(M)})$ for $k, l = 1, \ 2$.
Basically, (4.11) and (4.12) provide $P_k(\mathbf{x}_t|\boldsymbol{\theta}^{(M)})$, such that,

$$\begin{aligned} P_1(\mathbf{x}_t|\boldsymbol{\theta}_1^{(M)}) &= P(\mathbf{x}_t|Z_t = 1; \boldsymbol{\theta}_1^{(M)}) \\ &= P(x_t^p, x_t^s|\alpha, \beta; Z_t = 1) \\ &= \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \frac{\mathbf{B}(x_t^p + x_t^s + \alpha, n_t^p + n_t^s - x_t^p - x_t^s + \beta)}{\mathbf{B}(\alpha, \beta)} \end{aligned} \tag{4.23}$$

and

$$\begin{aligned} P_2(\mathbf{x}_t|\boldsymbol{\theta}_2^{(M)}) &= P(\mathbf{x}_t|Z_t=2;\boldsymbol{\theta}_2^{(M)}) \\ &= P(x_t^p, x_t^s|\gamma_1,\delta_1,\gamma_2,\delta_2;Z_t=2) \\ &= \binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{\mathbf{B}(x_t^p+\gamma_1, n_t^p-x_t^p+\delta_1)}{\mathbf{B}(\gamma_1,\delta_1)}\frac{\mathbf{B}(x_t^s+\gamma_2, n_t^s-x_t^s+\delta_2)}{\mathbf{B}(\gamma_2,\delta_2)}. \end{aligned} \tag{4.24}$$

Now, the joint probability for the observed methylation data $\mathbf{x}$ and the sequence of the hidden states (methylation status) $\mathbf{Z}$ can be obtained from the emission quantities (4.23) and (4.24) and the hidden states probability expressions from (4.13) and (4.14) for model $BBDM$ and (4.19) and (4.20) for model $BBCM$. So, (4.22) can be rewritten specific to model $BBDM$ as

$$\begin{aligned} P(\mathbf{x},\mathbf{Z}\,|\boldsymbol{\zeta}^{(D)}) &= \pi_{Z_1}^{(D)} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(D)})\prod_{t=2}^{T}\tau_{Z_{(t-1)},Z_t}^{(D)}(t)P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(D)}) \\ &= \pi_{Z_1}^{(D)} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(D)})\tau_{Z_1,Z_2}^{(D)}(2)P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(D)})\ldots\tau_{Z_{(T-1)},Z_T}^{(D)}(T)P_{Z_T}(\mathbf{x}_T|\boldsymbol{\theta}^{(D)}) \\ &= \prod_{t=1}^{T}\Bigg(\left[\binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{\mathbf{B}(x_t^p+x_t^s+\alpha, n_t^p+n_t^s-x_t^p-x_t^s+\beta)}{\mathbf{B}(\alpha,\beta)}\right]^{\mathbf{I}[Z_t=1]} \\ &\quad\times\left[\binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\binom{n_t^p}{x_t^p}\binom{n_t^s}{x_t^s}\frac{\mathbf{B}(x_t^p+\gamma_1, n_t^p-x_t^p+\delta_1)}{\mathbf{B}(\gamma_1,\delta_1)}\right. \\ &\quad\left.\times\frac{\mathbf{B}(x_t^s+\gamma_2, n_t^s-x_t^s+\delta_2)}{\mathbf{B}(\gamma_2,\delta_2)}\right]^{\mathbf{I}[Z_t=2]}\Bigg) \\ &\quad\times \pi_1^{\mathbf{I}[Z_1=1]}(1-\pi_1)^{\mathbf{I}[Z_1=2]}Bin(t_{11};t_{11}+t_{12},\tau_{11})Bin(t_{21};t_{21}+t_{22},\tau_{21}). \end{aligned} \tag{4.25}$$

Equation (4.22) can be rewritten specific to model $BBCM$,

$$
\begin{aligned}
P(\mathbf{x}, \mathbf{Z} \,|\boldsymbol{\zeta}^{(C)}) &= \pi^{(C)}_{Z_1} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(C)}) \prod_{t=2}^{T} \tau^{(C)}_{Z_{(t-1)},Z_t}(t) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(C)}) \\
&= \pi^{(C)}_{Z_1} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(D)}) \tau^{(C)}_{Z_1,Z_2}(2) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(C)}) \ldots \tau^{(C)}_{Z_{(T-1)},Z_T}(T) P_{Z_T}(\mathbf{x}_T|\boldsymbol{\theta}^{(C)}) \\
&= \prod_{t=1}^{T} \Bigg(\left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \frac{\mathbf{B}(x_t^p + x_t^s + \alpha, n_t^p + n_t^s - x_t^p - x_t^s + \beta)}{\mathbf{B}(\alpha, \beta)} \right]^{\mathbf{I}[Z_t=1]} \\
&\quad \times \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \frac{\mathbf{B}(x_t^p + \gamma_1, n_t^p - x_t^p + \delta_1)}{\mathbf{B}(\gamma_1, \delta_1)} \right. \\
&\quad \left. \times \frac{\mathbf{B}(x_t^s + \gamma_2, n_t^s - x_t^s + \delta_2)}{\mathbf{B}(\gamma_2, \delta_2)} \right]^{\mathbf{I}[Z_t=2]} \Bigg) \\
&\quad \times [0.5]^{\mathbf{I}(Z_1=1)} [0.5]^{\mathbf{I}(Z_1=2)} \\
&\quad \times \prod_{t=2}^{T} \Bigg[\left(\frac{\lambda_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=1)} \\
&\quad \times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} - \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=2)} \\
&\quad \times \left(\frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=1)} \\
&\quad \times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=2)} \Bigg],
\end{aligned}
\tag{4.26}
$$

where (4.25) and (4.26) are the complete data likelihoods for models $BBDM$ and $BBCM$, respectively.

Then, the likelihood of the observed methylation data $\mathbf{x}$ given the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ for model M can be expressed as,

$$
\begin{aligned}
L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}) &= P(\mathbf{x}|\boldsymbol{\zeta}^{(M)}) \\
&= \sum_{Z_1,\ldots,Z_T} \pi^{(M)}_{Z_1} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(M)}) \prod_{t=2}^{T} \tau^{(M)}_{Z_{(t-1)},Z_t}(t) P_{Z_t}(\mathbf{x}_t|\boldsymbol{\theta}^{(M)}),
\end{aligned}
\tag{4.27}
$$

where Equation (5.18) is the probability of the observed methylation data $\mathbf{x}$ conditional on the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ and thus can be written as the sum over all the 2^T possible state sequences of the complete data likelihood.

4.1.7 Choice of Priors

I use Uniform priors for the emission hyperparameters and non-informative Beta conjugate prior densities for the transition parameters.

The prior for the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ can be decomposed into three parts: i) priors of the emission hyperparameters $\boldsymbol{\theta}^{(M)}$; ii) priors of the initial state parameters $\boldsymbol{\pi}^{(M)}$; iii) priors of the transition parameters $\boldsymbol{\tau}^{(M)}$. I assume

$$p(\boldsymbol{\zeta}^{(M)}) = p(\boldsymbol{\theta}^{(M)})\ p(\boldsymbol{\pi}^{(M)})\ p(\boldsymbol{\tau}^{(M)}), \tag{4.28}$$

where $p(\chi)$ denotes the prior for χ.
For model M, the priors for the emission hyperparameters $\boldsymbol{\theta}^{(M)}$ are assumed to be independent:

$$p(\boldsymbol{\theta}^{(M)}) = p(\alpha)\ p(\beta)\ p(\gamma_1)\ p(\delta_1)\ p(\gamma_2)\ p(\delta_2). \tag{4.29}$$

The priors of the Beta (emission) hyperparameters for Model M are taken to be uniform and they are expressed as,

$$\begin{aligned}
\alpha &\sim U(a_\alpha, b_\alpha) \\
\beta &\sim U(a_\beta, b_\beta) \\
\gamma_1 &\sim U(a_{\gamma_1}, b_{\gamma_1}) \\
\delta_1 &\sim U(a_{\delta_1}, b_{\delta_1}) \\
\gamma_2 &\sim U(a_{\gamma_2}, b_{\gamma_2}) \\
\delta_2 &\sim U(a_{\delta_2}, b_{\delta_2}),
\end{aligned} \tag{4.30}$$

where $U(a, b)$ is the *Uniform* distribution with density $f(y|a, b) \propto \frac{1}{b-a}$, for $a \leq y \leq b$.

For model $BBDM$, the priors for the initial state parameters $\boldsymbol{\pi}^{(D)}$ are likewise assumed to be independent:

$$p(\boldsymbol{\pi}^{(D)}) = p(\pi_1). \tag{4.31}$$

Similarly, the priors for the transition parameters $\boldsymbol{\tau}^{(D)}$ are assumed to be independent:

$$p(\boldsymbol{\tau}^{(D)}) = p(\tau_{11})\ p(\tau_{21}). \tag{4.32}$$

The priors for initial state and transition probabilities $(\pi_1,\ \tau_{11},\ \tau_{21})$ are each assumed to be $Beta(\eta_1, \eta_2)$. For model $BBCM$, the priors for the transition parameters $\boldsymbol{\tau}^{(C)}$ are also independent:

$$p(\boldsymbol{\tau}^{(C)}) = p(\lambda_1)\ p(\lambda_2). \tag{4.33}$$

The priors for $\boldsymbol{\tau}^{(C)} = (\lambda_1,\ \lambda_2)$ are assumed to be uniform and they can be expressed as,

$$\begin{aligned} \lambda_1 &\sim U(a_{\lambda_1}, b_{\lambda_1}) \\ \lambda_2 &\sim U(a_{\lambda_2}, b_{\lambda_2}). \end{aligned} \tag{4.34}$$

4.1.8 Joint posterior distribution

The joint unnormalized posterior distribution for model M is given by,

$$p(\boldsymbol{\zeta}^{(M)}|\mathbf{x}) \propto \mathbf{L}(\boldsymbol{\zeta}^{(M)}) p(\boldsymbol{\zeta}^{(M)}). \tag{4.35}$$

4.2 Parameter and state estimation

I explore a fully Bayesian approach for estimating the parameters and the hidden states in my model. I construct an MCMC-based algorithm to examine the joint posterior distribution of Beta-Binomial HHMM. The Bayesian approach to estimate the model parameters and hidden states provides us with the capability of drawing inference directly from the posterior distributions. It also takes into account any prior information, including constraints on the parameters, to be incorporated in the data analysis. I have chosen conjugate priors for the Binomial proportions which permits only estimating the hyperparameters and thus reducing the dimension of the parameter space and increasing the computational efficiency by integrating out the parameters in the middle of the hierarchy.

In this MCMC-based algorithm, I have developed an augmented Gibbs sampler to obtain the posterior samples. The augmented Gibbs sampler cycles among updating the values of the emission hyperparameters, initial state and transition parameters and the hidden states. The samples of the hyperparameters are simulated from the posterior distributions conditional on the states using a M-H within Gibbs sampler as no closed form can be obtained from the posterior distribution of the hyperparameters.

The hidden states are sampled from their posterior distributions conditional on the hyperparameters. However, the direct computation of the likelihood $L(\boldsymbol{\zeta}^{(M)})$ must be avoided due to high computational cost. I introduce a recursive method that considers all the hidden states as one block and then updates their posterior distribution which in turn allows us to sample every state directly from the joint density. This technique enables more rapid mixing as the Markov chain contains a smaller number of parameters and also the dependency of every hidden state on its preceding sampled value can be significantly diminished (Liu et al., 1994).

4.2.1 Outline of the augmented Gibbs algorithm

In this section, I describe the details of the augmented Gibbs sampling scheme for one iteration implemented to sample from the posterior distributions of the HMM parameters $\boldsymbol{\zeta}^{(M)}$ for model M.

1. I calculate the full likelihood of model M conditional on the current values of the HMM parameters $\boldsymbol{\zeta}^{(M)}$ using the forward sum recursion described in Section 2.2.2. In my model M, I can re-construct the forward probability as,

$$\alpha_k^{(M)}(t) = P\left(\mathbf{x}_{1:t}; Z_t = k|\boldsymbol{\zeta}^{(M)}\right), \tag{4.36}$$

where $k = 1,\ 2$ denotes the similarly methylated state and differentially methylated state respectively. The quantity $\alpha_k^{(M)}(t)$ can also be viewed as the partial likelihood up to genomic position t, such that genomic position t is in state k for $t = 1, \ldots, T$ and $k = 1,\ 2$ which can be written as

$$\alpha_k^{(M)}(t) = \sum_{Z_1,\ldots,Z_t} \pi_{Z_1}^{(M)} P_{Z_1}(\mathbf{x}_1|\boldsymbol{\theta}^{(M)}) \prod_{s=2}^{t} \tau_{Z_{(s-1)},Z_s}^{(M)}(s) P_{Z_s}(\mathbf{x}_s|\boldsymbol{\theta}^{(M)}). \tag{4.37}$$

Using the forward sum recursion, the partial likelihood is given by,

$$\alpha_k^{(M)}(t) = b_k^{(M)}(t) \sum_{l=1}^{2} \alpha_l(t-1)\tau_{kl}^{(M)}(t), \quad t = 2, \ldots, T. \tag{4.38}$$

Here, $b_k^{(M)}(t) = P_k(\mathbf{x}_t|\boldsymbol{\theta}^{(M)})$. I have already derived expressions of $P_k(\mathbf{x}_t|\boldsymbol{\theta}^{(M)})$ in (4.23) and (4.24), respectively. For $t = 1$, I can write,

$$\alpha_{M_k}(1) = \pi_k b_k(1). \tag{4.39}$$

The full likelihood of the entire sequence can be expressed as,

$$L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}) = \sum_{k=1}^{2} \alpha_k^{(M)}(T), \tag{4.40}$$

where $L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)})$ is the full likelihood for model M.

2. After computing the partial likelihoods and the full likelihood using the forward sum recursion, I employ a backward sampling procedure to sample the hidden states $\mathbf{Z}$. The probability that the genomic position t is in state k given the sampled states at genomic positions $t+1, \ldots, T$, observed methylation data $\mathbf{x}$ and the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ is given by $P\left(Z_t = k | \mathbf{x}_{1:T}; \mathbf{Z}_{t+1:T}; \boldsymbol{\zeta}^{(M)}\right)$.

   The hidden states $Z_t,\ t = 1, \ldots, T$ can now be updated using a *backward* sampling imputation step:

$$\begin{aligned}
&\text{Sample } Z_T \text{ from } P\left(Z_T = k | \mathbf{x}_{1:T}; \boldsymbol{\zeta}^{(M)}\right) = \frac{\alpha_k^{(M)}(T)}{\sum_k \alpha_k^{(M)}(T)} \\
&\qquad\qquad \vdots \\
&\text{Sample } Z_t \text{ from } P\left(Z_t = k | \mathbf{x}_{1:T}; \mathbf{Z}_{t+1:T}; \boldsymbol{\zeta}^{(M)}\right) \propto \alpha_k^{(M)}(t) P\left(Z_{t+1} | Z_t = k\right) \\
&\qquad\qquad \vdots \\
&\text{Sample } Z_1 \text{ from } P\left(Z_1 = k | \mathbf{x}_{1:T}; \mathbf{Z}_{2:T}; \boldsymbol{\zeta}^{(M)}\right) \propto \alpha_k^{(M)}(1) P\left(Z_2 | Z_1 = k\right)
\end{aligned} \tag{4.41}$$

   In practical computations, the expressions for $\alpha_k^{(M)}(t)$ require reformulation using logarithms in order to avoid computational underflow.

3. Next, I update the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameters $\boldsymbol{\tau}^{(M)}$ conditional on the current values of the emission hyperparameters $\boldsymbol{\theta}^{(M)}$ and the hidden states $\mathbf{Z}$ and the observed methylation data

$\mathbf{x}$.

(a) For model $BBDM$, the initial state parameters $\boldsymbol{\pi}^{(D)} = (\pi_1,\ 1-\pi_1)$ and transition parameters $\boldsymbol{\tau}^{(D)} = (\tau_{11},\ \tau_{21})$ can be updated using a Gibbs sampler due to conjugacy in the full conditional posterior distributions.

(b) For model $BBCM$, the transition rate parameters $\boldsymbol{\tau}^{(C)} = (\lambda_1,\ \lambda_2)$ can be updated using a M-H algorithm.

4. For Model M, the emission hyperparameters $\boldsymbol{\theta}^{(M)}$ conditional on the current values of the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameters $\boldsymbol{\tau}^{(M)}$, the hidden states $\mathbf{Z}$ and the observed methylation data $\mathbf{x}$ can be updated using a M-H procedure.

4.2.2 Further details of the augmented Gibbs sampler

I use a Gibbs sampler to update all the parameters of interest, i.e., HMM parameters $\boldsymbol{\zeta}^{(M)}$ and the hidden states $\mathbf{Z}$. The essential steps of the augmented Gibbs sampler are as follows:

1. I sample the hidden state path $\mathbf{Z}$ from the full conditional posterior distribution $p(\mathbf{Z}|\mathbf{x}, \boldsymbol{\zeta}^{(M)})$ given $\boldsymbol{\zeta}^{(M)} = (\boldsymbol{\theta}^{(M)},\ \boldsymbol{\pi}^{(M)},\ \boldsymbol{\tau}^{(M)})$ and the observed methylation data $\mathbf{x}$. For this step, I employ the data-augmentation based Forward-Sum Backward Sampling (FSBS) procedure (Scott, 2002) instead of evaluating the likelihood expression, as described in Section 4.2.1.

2. I sample the emission hyperparameters $\boldsymbol{\theta}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)})$ given the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition parameters $\boldsymbol{\tau}^{(M)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. However, in this step, it is enough to sample $\boldsymbol{\theta}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z})$ using a M-H algorithm given the updated hidden states $\mathbf{Z}$ and observed methylation data

$\mathbf{x}$, since,

$$p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}) = p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}). \tag{4.42}$$

3. In this step, I sample the HMM initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition parameters $\boldsymbol{\tau}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(M)})$ given the emission model parameters $\boldsymbol{\theta}^{(M)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$.

   (a) For model $BBDM$, I sample the HMM initial state parameters $\boldsymbol{\pi}^{(D)}$ and transition parameters $\boldsymbol{\tau}^{(D)}$ from the full conditional posterior distribution $p(\boldsymbol{\pi}^{(D)}, \boldsymbol{\tau}^{(D)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(D)})$ given the emission model parameters $\boldsymbol{\theta}^{(D)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. Due to the Beta-Binomial conjugacy of the full conditional posterior distribution, it is enough to sample from $p(\boldsymbol{\pi}^{(D)}, \boldsymbol{\tau}^{(D)}|\mathbf{Z})$, since,

$$p(\boldsymbol{\pi}^{(D)}, \boldsymbol{\tau}^{(D)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(D)}) = p(\boldsymbol{\tau}^{(D)}|\mathbf{Z}). \tag{4.43}$$

   (b) For model $BBCM$, I sample the HMM transition parameters $\boldsymbol{\tau}^{(C)}$ from the full conditional posterior distribution $p(\boldsymbol{\tau}^{(C)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(C)}, \boldsymbol{\Psi})$ given the emission model parameters $\boldsymbol{\theta}^{(C)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. It is enough to sample from $p(\boldsymbol{\tau}^{(C)}|\mathbf{Z}, \boldsymbol{\Psi})$, since

$$p(\boldsymbol{\tau}^{(C)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(C)}, \boldsymbol{\Psi}) = p(\boldsymbol{\tau}^{(C)}|\mathbf{Z}, \boldsymbol{\Psi}), \tag{4.44}$$

I now describe the sampling steps of the emission hyperparameters (2), the initial state and transition parameters $(3.(a),\ (b))$ for both the models.

4.2.3 Sampling from conditional posterior distributions

4.2.3.1 Emission hyperparameters

In this section, I elaborate in details the sampling steps of the emission hyperparameters from their full conditional posterior distributions. I first write the full conditional posterior density of the HMM model emission hyperparameters $\boldsymbol{\theta}^{(M)}$:

$$\begin{aligned} p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}) &= p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}) \\ &\propto L_{\mathbf{x},\mathbf{Z}}(\boldsymbol{\theta}^{(M)}) p(\boldsymbol{\theta}^{(M)}), \end{aligned} \tag{4.45}$$

$L_{\mathbf{x},\mathbf{Z}}(\boldsymbol{\theta}^{(M)})$ denotes the complete data likelihood.

I sample the emission hyperparameters $(\alpha,\ \beta,\ \gamma_1,\ \delta_1, \gamma_2,\ \delta_2)$ from their full conditional posterior distributions as follows:

- Sample $\alpha|\beta, \mathbf{x}, \mathbf{Z}$ from

$$p(\alpha|\beta, \mathbf{x}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \frac{\mathbf{B}(x_t^p + x_t^s + \alpha, n_t^p + n_t^s - x_t^p - x_t^s + \beta)}{\mathbf{B}(\alpha, \beta)} \right]^{\mathbf{I}[Z_t=1]} \tag{4.46}$$

$$\times \frac{1}{b_\alpha - a_\alpha}.$$

- Sample $\beta|\alpha, \mathbf{x}, \mathbf{Z}$ from

$$p(\beta|\alpha, \mathbf{x}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \frac{\mathbf{B}(x_t^p + x_t^s + \alpha, n_t^p + n_t^s - x_t^p - x_t^s + \beta)}{\mathbf{B}(\alpha, \beta)} \right]^{\mathbf{I}[Z_t=1]} \tag{4.47}$$

$$\times \frac{1}{b_\beta - a_\beta}.$$

- Sample $\gamma_1|\delta_1, \gamma_2, \delta_2, \mathbf{x}, \mathbf{Z}$ from

$$p(\gamma_1|\delta_1, \gamma_2, \delta_2, \mathbf{x}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \times \frac{\mathbf{B}(x_t^p + \gamma_1, n_t^p - x_t^p + \delta_1)}{\mathbf{B}(\gamma_1, \delta_1)} \frac{\mathbf{B}(x_t^s + \gamma_2, n_t^s - x_t^s + \delta_2)}{\mathbf{B}(\gamma_2, \delta_2)} \right]^{\mathbf{I}[Z_t=2]} \times \frac{1}{b_{\gamma_1} - a_{\gamma_1}}. \tag{4.48}$$

- Sample $\delta_1|\gamma_1, \gamma_2, \delta_2, \mathbf{x}, \mathbf{Z}$ from

$$p(\delta_1|\gamma_1, \gamma_2, \delta_2, \mathbf{x}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \times \frac{\mathbf{B}(x_t^p + \gamma_1, n_t^p - x_t^p + \delta_1)}{\mathbf{B}(\gamma_1, \delta_1)} \frac{\mathbf{B}(x_t^s + \gamma_2, n_t^s - x_t^s + \delta_2)}{\mathbf{B}(\gamma_2, \delta_2)} \right]^{\mathbf{I}[Z_t=2]} \times \frac{1}{b_{\delta_1} - a_{\delta_1}}. \tag{4.49}$$

- Sample $\gamma_2|\delta_1, \gamma_1, \delta_2, \mathbf{x}, \mathbf{Z}$ from

$$p(\gamma_2|\delta_1, \gamma_1, \delta_2, \mathbf{x}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \times \frac{\mathbf{B}(x_t^p + \gamma_1, n_t^p - x_t^p + \delta_1)}{\mathbf{B}(\gamma_1, \delta_1)} \frac{\mathbf{B}(x_t^s + \gamma_2, n_t^s - x_t^s + \delta_2)}{\mathbf{B}(\gamma_2, \delta_2)} \right]^{\mathbf{I}[Z_t=2]} \times \frac{1}{b_{\gamma_2} - a_{\gamma_2}}. \tag{4.50}$$

- Sample $\delta_2|\gamma_1\delta_1, \gamma_2, \mathbf{x}, \mathbf{Z}$ from

$$p(\delta_2|\gamma_1\delta_1, \gamma_2, \mathbf{X}, \mathbf{Z}) = \prod_{t=1}^{T} \left[\binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \binom{n_t^p}{x_t^p} \binom{n_t^s}{x_t^s} \times \frac{\mathbf{B}(x_t^p + \gamma_1, n_t^p - x_t^p + \delta_1)}{\mathbf{B}(\gamma_1, \delta_1)} \frac{\mathbf{B}(x_t^s + \gamma_2, n_t^s - x_t^s + \delta_2)}{\mathbf{B}(\gamma_2, \delta_2)} \right]^{\mathbf{I}[Z_t=2]} \times \frac{1}{b_{\delta_2} - a_{\delta_2}}, \tag{4.51}$$

where $a_{\theta^{(M)}}$, $b_{\theta^{(M)}}$ for $\theta^{(M)} \in \{\alpha,\ \beta,\ \gamma_1,\ \delta_1, \gamma_2,\ \delta_2\}$ are fixed values of the Uniform emission hyperpriors.

The emission hyperparameters are sampled from their full conditionals using the M-H algorithm as the conditional posterior densities of $\boldsymbol{\theta}^{(M)}$ do not have any closed form. I propose new emission hyperparameter values of $\boldsymbol{\theta}'^{(M)} = (\alpha', \beta', \gamma_1', \delta_1', \gamma_2', \delta_2')$ given the current emission hyperparameter values $\boldsymbol{\theta}^{t(M)} = (\alpha^t, \beta^t, \gamma_1^t, \delta_1^t, \gamma_2^t, \delta_2^t)$ using symmetric random walk updates. To guarantee that the proposed values (indicated by primes) of the hyperparameters $\boldsymbol{\theta}^{(M)}$ are non-negative, I choose the following truncated Normal proposal densities left-truncated at zero:

$$\begin{aligned} \alpha' &\sim Trunc.N(\alpha^t, \sigma_\alpha^2) \\ \beta' &\sim Trunc.N(\beta^t, \sigma_\beta^2) \\ \gamma_1' &\sim Trunc.N(\gamma_1^t, {\sigma_{\gamma 1}}^2) \\ \delta_1' &\sim Trunc.N(\delta_1^t, {\sigma_{\delta 1}}^2) \\ \gamma_2' &\sim Trunc.N(\gamma_2^t, {\sigma_{\gamma 2}}^2) \\ \delta_2' &\sim Trunc.N(\delta_2^t, {\sigma_{\delta 2}}^2), \end{aligned} \tag{4.52}$$

where $\sigma_\alpha,\ \sigma_\beta,\ \sigma_{\gamma_1},\ \sigma_{\delta_1},\ \sigma_{\gamma_2},\ \sigma_{\delta_2}$ are the tuning proposal parameters which can be adjusted in order to improve the convergence properties of the MCMC-based

augmented Gibbs sampler. The new value for α is accepted with acceptance probability $\min(1,\ r_\alpha)$, where the M-H ratio, r_α, is:

$$\begin{aligned} r_\alpha &= \frac{p\left(\alpha'|\beta,\mathbf{x},\mathbf{Z}\right) q\left(\alpha^t|\alpha'\right)}{p\left(\alpha^t|\beta,\mathbf{x},\mathbf{Z}\right) q\left(\alpha'|\alpha^t\right)} \\ &= \frac{p\left(\alpha'|\beta,\mathbf{x},\mathbf{Z}\right)\left(1-\Phi\left(\alpha^t\right)\right)}{p\left(\alpha^t|\beta,\mathbf{x},\mathbf{Z}\right)\left(1-\Phi\left(\alpha'\right)\right)}, \end{aligned} \tag{4.53}$$

where $q(a'|a^t)$ is the truncated Normal proposal density with proposed value a' given the current value a^t. Now (4.46) can be substituted in (4.53) to get the full expression for r_α. Similarly, I can update the value of β with acceptance probability $\min(1,\ r_\beta)$ just by replacing α with β in (4.53).

Again, the new value for γ_1 is accepted with acceptance probability $\min(1,\ r_{\gamma_1})$. Now, the M-H ratio r_{γ_1} can be written as follows:

$$\begin{aligned} r_{\gamma_1} &= \frac{p\left(\gamma_1'|\delta_1,\gamma_2,\delta_2,\mathbf{x},\mathbf{Z}\right) q\left(\gamma_1^t|\gamma_1'\right)}{p\left(\gamma_1^t|\delta_1,\gamma_2,\delta_2,\mathbf{x},\mathbf{Z}\right) q\left(\gamma_1'|\gamma_1^t\right)} \\ &= \frac{p\left(\gamma_1|\delta_1',\gamma_2,\delta_2,\mathbf{x},\mathbf{Z}\right)\left(1-\Phi\left(\gamma_1^t\right)\right)}{p\left(\gamma_1|\delta_1^t,\gamma_2,\delta_2,\mathbf{x},\mathbf{Z}\right)\left(1-\Phi\left(\gamma_1'\right)\right)}. \end{aligned} \tag{4.54}$$

(4.48) can be substituted in (4.54) for the detailed expression of r_{γ_1}. Similarly, I can update the values of $\delta_1,\ \gamma_2,\ \delta_2$ with acceptance probabilities $\min(1,\ r_{\delta_1})$, $\min(1,\ r_{\gamma_2})$, $\min(1,\ r_{\delta_2})$ just by replacing γ_1 with $\delta_1,\ \gamma_2,\ \delta_2$ respectively in (4.54).

4.2.3.2 Initial state and transition probabilities

I describe in detail the sampling steps of the initial state and transition probabilities for model *BBDM* conditional on $\mathbf{Z}$. π_1, τ_{11} and τ_{21} given $\mathbf{Z}$ are independent. So, I can write the full conditional posterior distribution of the HMM model initial state and transition parameters $(\boldsymbol{\pi}^{(D)}, \boldsymbol{\tau}^{(D)})$ as

$$p(\boldsymbol{\pi}^{(D)}, \boldsymbol{\tau}^{(D)}|\mathbf{Z}) = p(\pi_1|Z_1)p(\tau_{11}|\mathbf{Z}_{2:T})p(\tau_{21}|\mathbf{Z}_{2:T}). \tag{4.55}$$

I assign a Beta prior for the initial state and transition probabilities, i.e., $\pi_1,\ \tau_{11},\ \tau_{21} \sim Beta(\eta_1, \eta_2)$ independently, but η_1 and η_2 are both set to 1 to give a noninformative prior to τ_{11} and τ_{21}, respectively, namely the Uniform distribution $U(0, 1)$.

Now, I sample the initial state and transition parameters $(\pi,\ \tau_{11},\ \tau_{21})$ as follows:

- Sample $\pi_1|Z_1$, i.e.,

$$\pi_1 \sim Beta(2, 1),\ \text{if } Z_1 = 1 \tag{4.56}$$

  and

$$\pi_1 \sim Beta(1, 2),\ \text{if } Z_1 = 2. \tag{4.57}$$

- Sample $\tau_{11}|\mathbf{Z}_{2:T}$, i.e.,

$$\tau_{11} \sim Beta(t_{11} + 1, t_{12} + 1). \tag{4.58}$$

- Sample $\tau_{21}|\mathbf{Z}_{2:T}$, i.e.,

$$\tau_{21} \sim Beta(t_{21} + 1, t_{22} + 1). \tag{4.59}$$

Thus, the initial state and the transition parameters $(\pi,\ \tau_{11},\ \tau_{21})$ are sampled directly from their full conditional posterior distributions simply by using Gibbs sampler as the full conditionals have closed form due to Beta-Binomial conjugacy.

4.2.3.3 Transition rate parameters

In this section, I describe the sampling steps of the transition rate parameters for model *BBCM* from their full conditional posterior distributions. I first write the

full conditional posterior density of the HMM model transition rate parameters $\boldsymbol{\tau}^{(C)}$,

$$
\begin{aligned}
p(\boldsymbol{\tau}^{(C)}|\mathbf{x},\mathbf{Z},\boldsymbol{\Psi}) = p(\boldsymbol{\tau}^{(C)}|\mathbf{Z},\boldsymbol{\Psi}) \\
\prod_{t=2}^{T}\Bigg[\left(\frac{\lambda_2}{\lambda_1+\lambda_2}+\frac{\lambda_1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)\Psi_t}\right)^{\mathbf{I}(Z_{t-1}=1,Z_t=1)} \\
\times\left(\frac{\lambda_1}{\lambda_1+\lambda_2}-\frac{\lambda_1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)\Psi_t}\right)^{\mathbf{I}(Z_{t-1}=1,Z_t=2)} \\
\times\left(\frac{\lambda_2}{\lambda_1+\lambda_2}-\frac{\lambda_2}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)\Psi_t}\right)^{\mathbf{I}(Z_{t-1}=2,Z_t=1)} \\
\left(\frac{\lambda_1}{\lambda_1+\lambda_2}+\frac{\lambda_2}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)\Psi_t}\right)^{\mathbf{I}(Z_{t-1}=2,Z_t=2)}\Bigg],
\end{aligned}
\tag{4.60}
$$

where $\boldsymbol{\Psi}=(\Psi_1,\cdots,\Psi_t)$, such that, Ψ_t is the genomic distance between two adjacent CpG sites indexed at $t-1$ and t.

I sample the transition rate parameters $(\lambda_1,\ \lambda_2)$ from their full conditional posterior distributions as follows.

- Sample $\lambda_1|\lambda_2, \mathbf{Z}_{2:T}, \boldsymbol{\Psi}$ from

$$
\begin{aligned}
p(\lambda_1|\lambda_2, \mathbf{Z}_{2:T}, \boldsymbol{\Psi}) \\
&\propto \prod_{t=2}^{T} \Bigg[\left(\frac{\lambda_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=1)} \\
&\times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} - \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=2)} \\
&\times \left(\frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=1)} \\
&\left(\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=2)} \Bigg] \\
&\times \frac{1}{b_{\lambda_1} - a_{\lambda_1}}.
\end{aligned}
\tag{4.61}
$$

- Sample $\lambda_2|\lambda_1, \mathbf{Z}_{2:T}, \boldsymbol{\Psi}$ from

$$
\begin{aligned}
p(\lambda_2|\lambda_1, \mathbf{Z}_{2:T}, \boldsymbol{\Psi}) \\
&\propto \prod_{t=2}^{T} \Bigg[\left(\frac{\lambda_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=1)} \\
&\times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} - \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=2)} \\
&\times \left(\frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=1)} \\
&\left(\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=2)} \Bigg] \\
&\times \frac{1}{b_{\lambda_2} - a_{\lambda_2}},
\end{aligned}
\tag{4.62}
$$

where $a_{\lambda_1},\ b_{\lambda_1},\ a_{\lambda_2},\ b_{\lambda_2}$ are fixed values of the Uniform transition rate priors.

The transition rate parameters $\boldsymbol{\tau}^{(C)} = (\lambda_1,\ \lambda_2)$ are sampled from their full conditionals using M-H algorithm as the conditional posterior densities of $\boldsymbol{\tau}^{(C)}$ do not have any closed form. I propose new transition rate parameter values of $\boldsymbol{\tau}'^{(C)} = (\lambda_1', \lambda_2')$ given the current transition rate parameter values $\boldsymbol{\tau}^{t(C)} = (\lambda_1^t, \lambda_2^t)$ using symmetric random walk updates. To guarantee that the proposed values of the transition rate parameters $\boldsymbol{\tau}^{(C)}$ are non-negative, I again choose the truncated Normal proposal densities left-truncated at zero:

$$\begin{aligned} \lambda_1' &\sim Trunc.N(\lambda_1^t, \sigma_{\lambda_1}^2) \\ \lambda_2' &\sim Trunc.N(\lambda_2^t, \sigma_{\lambda_2}^2), \end{aligned} \tag{4.63}$$

where $\sigma_{\lambda_1},\ \sigma_{\lambda_2}$ are the tuning proposal parameters.

4.2.4 Summary of the augmented Gibbs sampler algorithm steps

1. Initialize all the emission hyperparameters $\boldsymbol{\theta}^{(M)}$ for model M.
   (a) For model *BBDM*,
   initialize initial state and transition parameters $(\pi_1, \tau_{11}, \tau_{21})$ and,
   (b) For model *BBCM*,
   initialize transition rate parameters $\boldsymbol{\tau}^{(C)} = (\lambda_1, \lambda_2)$.
2. Compute the state-specific emission distributions,
   $P(\mathbf{x}_t | Z_t = k, \boldsymbol{\theta}_k^{(M)})$ for $k = 1, 2$ and $t = 1, \ldots, T$.
3. Compute $\alpha_k^{(M)}(t)$ for $k = 1, 2$ and $t = 1, \ldots, T$.
4. Sample backwards $Z_T, \ldots, Z_1$ using backward sampling (Scott, 2002).
5. Sample $\boldsymbol{\theta}^{(M)}$ using Component-wise M-H algorithm (Metropolis et al., 1953) as described in 4.2.3.1.

6. Sample the transition parameters:

   (a) For model *BBDM*,
   sample $\tau_{ij} \sim \text{Beta}\left(\text{t}_{ij}+1, \sum_{k\neq j}^{2} t_{ik}+1\right)$ for $i,\ j=1,\ 2$, such that, $k \neq j$,
   and $\pi_k|Z_1=k \sim \text{Beta}\left(1+\mathbf{I}(\text{Z}_1=k), 1+\mathbf{I}(Z_1=k')\right)$ for $k,\ k'=1,\ 2$.

   (b) For model *BBCM*,
   sample λ_1 and λ_2 using M-H algorithm.

7. Implement the relabelling algorithm as described in Section 2.2.6.

8. Repeat steps (2)-(7) until convergence.

4.2.5 Updating the predicted states

Finally, I note the method used to identify the SMCs (state1s) and DMCs (state2s) in the chromosome. Define, $\mathbf{Z}^{(1)}, \ldots, \mathbf{Z}^{(I)}$ to be I Gibbs draws (after burn-in) of the joint hidden states, where $\mathbf{Z}^{(i)} = (Z_1^{(i)}, \ldots, Z_T^{(i)})$. The estimate of the posterior probability that t^{th} genomic position is similarly methylated is given by $\widehat{P}(Z_t=1|\mathbf{x}) = \frac{1}{I}\sum_{i=1}^{I} I(Z_t^{(i)}=1)$. To decide whether a CpG site is differentially methylated or not, I specify a threshold value on these posterior probabilities. If $\widehat{P}(Z_t=1|\mathbf{x}) > 0.5$, I predict the t^{th} CpG site to be similarly methylated or if $\widehat{P}(Z_t=1|\mathbf{x}) < 0.5$, I call that t^{th} CpG site to be differentially methylated.

4.3 Simulation studies

In this section, I perform simulation studies to compare the performance of my proposed models (*BBDM*, *BBCM*) in identifying the DMCs in artificial datasets. The simulation studies were designed to examine the performance and robustness of both models under different situations, such as model misspecification and varying levels of noise in the data.

4.3.1 Data generation

100 datasets were generated with $T = 10000$ observations each under different situations to check the robustness of my models. The data generation was done in 3 steps for both the models:

1. (a) For Model *BBDM*, the sequence of the hidden states $\mathbf{Z} = (Z_1, \ldots, Z_T)$ was simulated using a Markov Chain with true fixed transition probabilities τ_{11}, τ_{21} and initial state probability π_1; (b) For model *BBDM*, the sequence of the hidden states $\mathbf{Z} = (Z_1, \ldots, Z_T)$ was simulated using a continuous-index Markov chain with true fixed transition rate parameters λ_1 and λ_2.

2. The nuisance parameters p_t^*, p_t^p and p_t^s for each $t = 1, \ldots, 10000$ were sampled from Beta distributions with true fixed emission hyperparameters conditional on the state labels $Z_t = k$, $k = 1, 2$. p_t^* was sampled from a Beta distribution with fixed state 1 hyperparameters $(\alpha,\ \beta)$, whereas p_t^p and p_t^s are sampled from Beta distributions with fixed state 2 hyperparameters $(\gamma_1,\ \delta_1,\ \gamma_2,\ \delta_2)$.

3. The methylated counts of proliferating and senescent cells of each CpG site x_t^p and x_t^s for $t = 1, \ldots, 10000$ were sampled from Binomial distributions with parameters n_t^p and n_t^s taken from the real data and probability of methylations from the corresponding sampled values of p_t^*, p_t^p and p_t^s conditional on Z_t. Since the total counts (methylated counts + unmethylated counts) at each CpG site were taken from the real data, it made my simulation study design biologically realistic in this regard. I have studied 3 potential cases in the following simulations.

1. **Moderately overlapped**

   (a) For model *BBDM*, the data are generated in such a way that the data classified by the simulated states overlap with each other, thus making it difficult to correctly predict the states. The data are generated

from the Beta-Binomial HMM with similar modes for data of both the states, i.e., state 1 hyperparameters ($\alpha = 3,\ \beta = 4$) and state 2 hyperparameters ($\gamma_1 = 3.2,\ \delta_1 = 3.9,\ \gamma_2 = 4,\ \delta_2 = 5$). The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order Markov Chain with an initial state probability for state 1, $\pi_1 = 0.34$, and transition probabilities $\tau_{11} = 0.87$, $\tau_{21} = 0.068$.

(b) For model *BBCM*, the data are generated as for *BBDM* except that the hidden states $\mathbf{Z}$ are simulated from a 1^{st} order continuous-index Markov chain with transition rate parameters $\lambda_1 = 0.22$ and $\lambda_2 = 0.22$.

2. **Well separated**

(a) For model *BBDM*, the data are generated in such a way that the data classified by the simulated states are well separated from each other, making it easier to correctly predict the states. The data ($\mathbf{x}^p,\ \mathbf{x}^s$) are generated from the Beta-Binomial HMM with well-separated modes for data of both the states, i.e., state 1 hyperparameters ($\alpha = 1.2,\ \beta = 8.8$) and state 2 hyperparameters ($\gamma_1 = 5.5,\ \delta_1 = 4.5,\ \gamma_2 = 8.5,\ \delta_2 = 1.5$). The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order Markov Chain with an initial state probability for State-1 $\pi_1 = 0.34$ and transition probabilities $\tau_{11} = 0.87$, $\tau_{21} = 0.068$.

(b) For model *BBCM*, the data are generated as for *BBDM* except that the hidden states $\mathbf{Z}$ are simulated from a 1^{st} order continuous-index Markov chain with transition rate parameters $\lambda_1 = 0.278$ and $\lambda_2 = 0.28$.

3. **Realistic**

(a) For model *BBDM*, the data are generated using the real data study estimates on this dataset. In this case, the data classified by the simulated states slightly overlap with each other. Thus, it would be interesting to test the performance of my model in the case of a more realistic situation. The data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated from the Beta-Binomial HMM with less well-separated modes for data of both the states, i.e., state 1 hyperparameters $(\alpha = 5.2,\ \beta = 2.65)$ and state 2 hyperparameters $(\gamma_1 = 1.36,\ \delta_1 = 3.25,\ \gamma_2 = 1.07,\ \delta_2 = 5.3)$, thus causing some amount of overlapping. The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order Markov Chain with an intial state probability for State-1 $\pi_1 = 0.34$ and transition probabilities $\tau_{11} = 0.87$, $\tau_{21} = 0.068$.

(b) For model *BBCM*, the data are generated using the real data study estimates on this dataset. The data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated from the Beta-Binomial HMM with less well-separated modes for data of both the states, i.e., state 1 hyperparameters $(\alpha = 11.62,\ \beta = 5.10)$ and state 2 hyperparameters $(\gamma_1 = 1.19,\ \delta_1 = 1.90,\ \gamma_2 = 0.78,\ \delta_2 = 1.82)$. The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order continuous-index Markov chain with transition rate parameters $\lambda_1 = 0.534$ and $\lambda_2 = 0.11$.

Other than the *realistic* case, there was no strong additional reason in the choice of true values for the transition parameters. The true values of the transition parameters were more or less similar for all the three cases. The main objective was to generate the data in such a way that the data classified by the simulated states must satisfy the data generation conditions.

4.3.2 Priors

- I use weakly informative and independent Uniform priors for the emission hyperparameters. The prior distributions of the emission hyperparameters $\boldsymbol{\theta}^{(M)} = (\alpha, \beta, \gamma_1, \delta_1, \gamma_2, \delta_2)$ for model M are as follows:

$$\begin{aligned} \alpha &\sim U(0, 2000) \\ \beta &\sim U(0, 2000) \\ \gamma_1 &\sim U(0, 2000) \\ \delta_1 &\sim U(0, 2000) \\ \gamma_2 &\sim U(0, 2000) \\ \delta_2 &\sim U(0, 2000). \end{aligned} \tag{4.64}$$

- I use weakly informative and independent Uniform priors for the transition rate parameters $\boldsymbol{\tau}_C = (\lambda_1,\ \lambda_2)$ for model *BBCM* and they are both $U(0, 2000)$.

4.3.3 Consistency of model parameters estimation

I generated 100 datasets under both models (*BBDM* and *BBCM*). These datasets of size $10,000$ CpG sites were generated for each parameter setting described in Section 4.3.1, which subsequently were estimated using the augmented Gibbs sampler described in Section 4.2. Each simulated dataset was then fitted to the models *BBDM* and *BBCM* for each case with $60,000$ MCMC iterations (with $20,000$ as burn-in) after which the posterior samples for each model parameter were assessed for convergence.

After attaining convergence, to estimate the quality of the estimation of the model parameters, I estimated the Root Mean Square Error (RMSE) of each of the model parameters. The RMSE in my simulation studies for any parameter ϵ

was determined from

$$RMSE(\hat{\epsilon}) = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\epsilon}_j - \epsilon_{true})^2}, \text{ for } J = 100, \tag{4.65}$$

where ϵ_{true} is the true value of the parameter and in the j^{th} simulated dataset $\hat{\epsilon}_j$ is its posterior mean estimate.

In Table 4.1, I presented the range of estimated RMSE of the model parameters for each case. In each case, the estimated RMSE was small, demonstrating good estimation of the model parameters except for the *moderately overlapped* case in both the models. Here, the transition parameters for the *moderately overlapped* case in both the models showed inconsistent estimation. Since the data for both the states are generated using a similar set of true values of the hyperparameters, it fails to distinguish between the 2 states. As a result, the RMSE for the model parameters for the *moderately overlapped* case in both the models are large. However, for the *well separated* case in both the models, the RMSE for the model parameters are much smaller ranging between (0.0002, 0.009) and (0.0006, 0.01), respectively. Similarly, for the *realistic* case in both the models the RMSE for model parameters ranged between (0.0006, 0.0093) and (0.0008, 0.009), respectively. Clearly, the values of RMSE for the model parameters for the *well separated* case in both the models are generally the lowest.

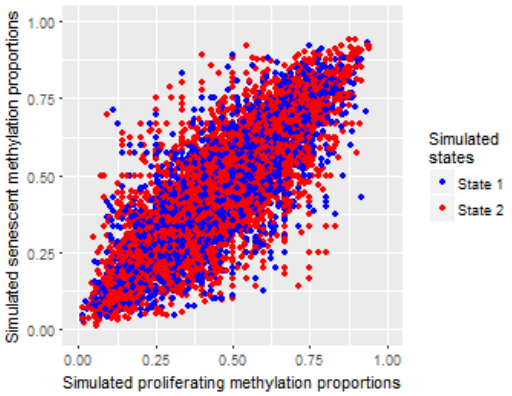
The posterior state-membership for all the CpG sites are assigned using a cut-off value of 0.5 as discussed in Section 4.2.5. The misclassification rate for all the 3 cases was calculated by comparing the simulated and predicted states at each genomic position. The misclassification rate is the proportion of mismatches between the simulated and the predicted states. The average misclassification rate is then the average of the misclassification rates based on 100 simulated datasets. In the *moderately overlapped* case, the average misclassification rates for *BBDM*

Model	Case	Average Misclass. rate	Range of RMSE
BBDM	*Moderately overlapped*	0.6673	(0.05, 1.091)
	Well separated	0.0042	(0.0002, 0.009)
	Realistic	0.0242	(0.0006, 0.0093)
BBCM	*Moderately overlapped*	0.2782	(0.08, 1.72)
	Well separated	0.0196	(0.0006, 0.01)
	Realistic	0.0664	(0.0008, 0.009)

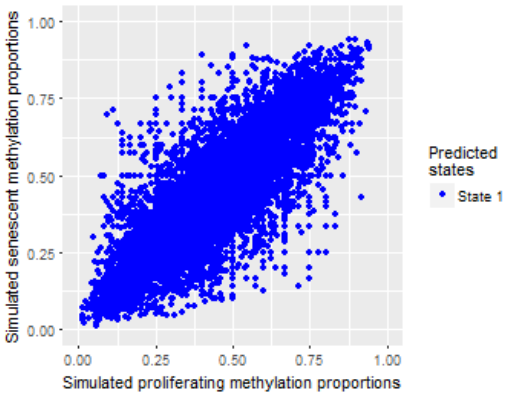
Table 4.1: Simulation study: Average misclassification rate and range of RMSE for models: BBDM and BBCM based on 100 simulated datasets.

and *BBCM* are 0.6673 and 0.2782, respectively (Table 4.1). The corresponding misclassification rates for the *well separated* case are 0.0042 and 0.0196 and for the *realistic* case 0.0242 and 0.0664, respectively. The misclassification rates for the *well separated* case in both models are much lower than the *realistic* case.

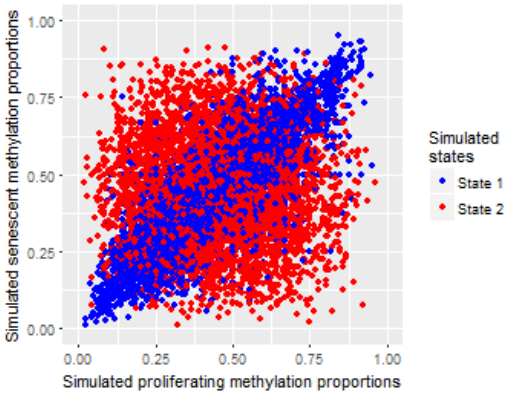
For one of the randomly selected simulation studies out of 100, I also present the scatter plots of the data generation for all the cases. In Figures 4.2, 4.3, 4.4, I showed visually how well I have selected the true values of the parameters for each of the cases, explained in Section 4.3.1. The scatter plots of simulated methylation proportions between proliferating and senescent cells classified by the true states, for both the models, validate the choice of true values of the parameters and the data generation procedure. From Figures 4.2a, 4.2c for the *moderately overlapped* case in both models, it can be seen from the scatter plots that the simulated methylation proportions between proliferating and senescent cells classified by the true states overlap with each other. It can also be observed that the simulated proportions between two cell types are much more scattered for *BBCM* compared to *BBDM* in the moderately overlapped case. Hence, model *BBCM* is able to classify the hidden states better than *BBDM*. Furthermore, from Figures 4.3a, 4.3c for the *well separated* case in both models, the scatter plots display



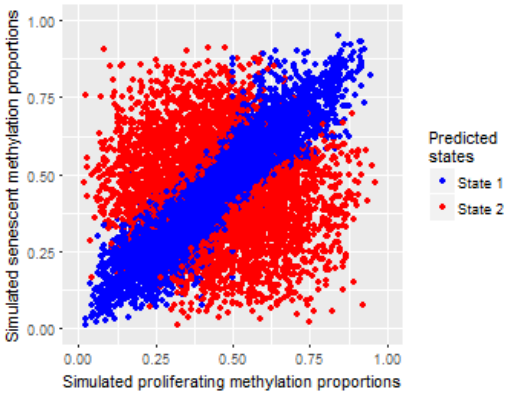
(a) Scatter plot for *BBDM* classified by simulated states.



(b) Scatter plot for *BBDM* classified by predicted states.

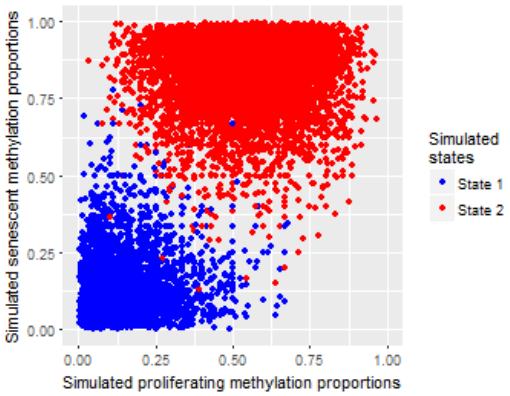


(c) Scatter plot for *BBCM* classified by simulated states.

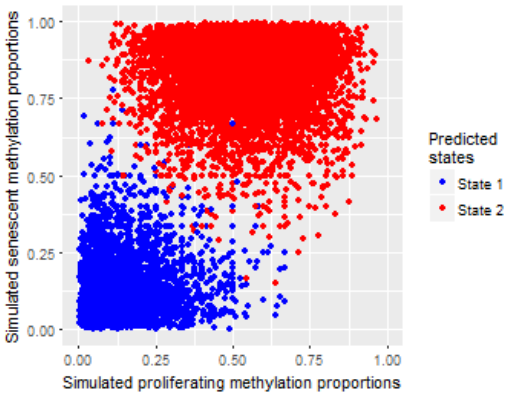


(d) Scatter plot for *BBCM* classified by predicted states.

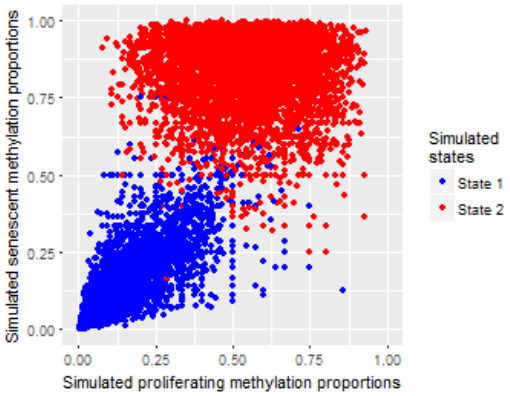
Figure 4.2: For the *moderately overlapped* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBCM*.



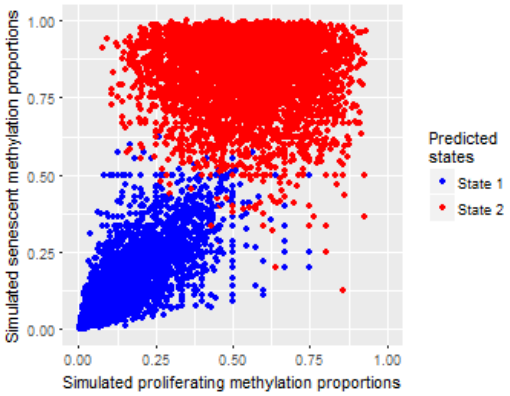
(a) Scatter plot for *BBDM* classified by simulated states.



(b) Scatter plot for *BBDM* classified by predicted states.

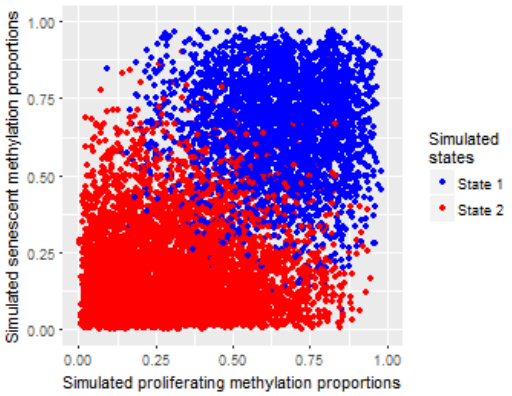


(c) Scatter plot for *BBCM* classified by simulated states.

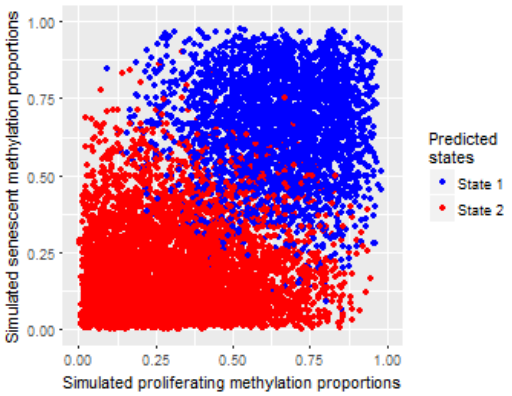


(d) Scatter plot for *BBCM* classified by predicted states.

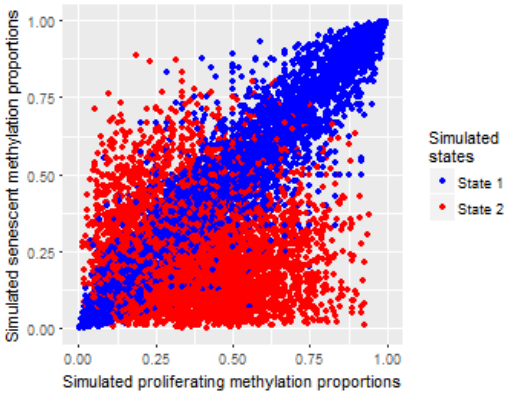
Figure 4.3: For the *well separated* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBCM*.



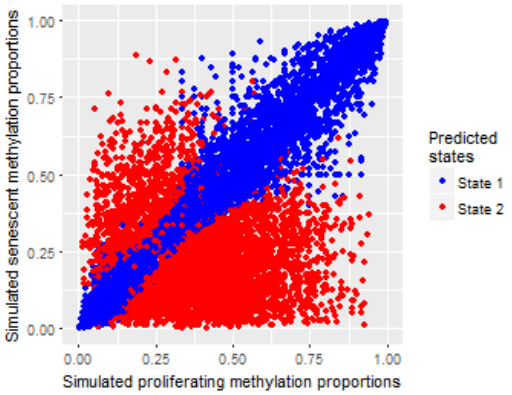
(a) Scatter plot for *BBDM* classified by simulated states.



(b) Scatter plot for *BBDM* classified by predicted states.



(c) Scatter plot for *BBCM* classified by simulated states.



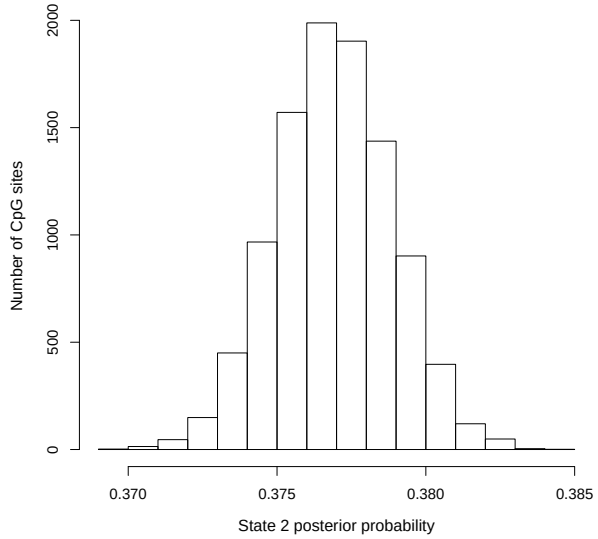
(d) Scatter plot for *BBCM* classified by predicted states.

Figure 4.4: For the *realistic* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *BBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *BBCM*.

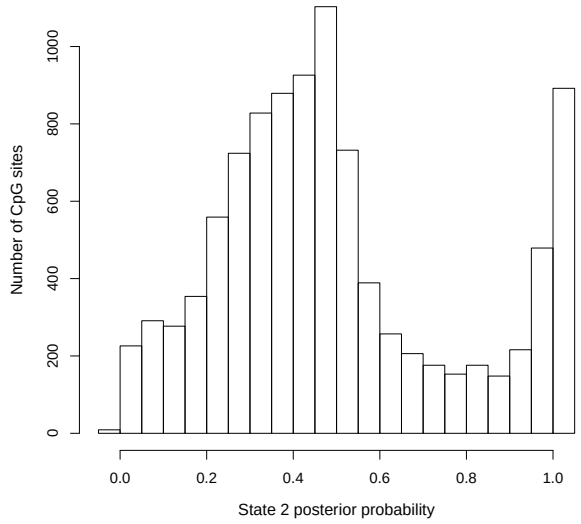
that the simulated methylation proportions between proliferating and senescent cells classified by the true states are well separated. The *realistic* case in both the models also exhibit similar kind of pattern displayed by the *moderately overlapped* case and they can be verified from from Figures 4.4a, 4.4c. In addition, I have also displayed the prediction power of my algorithm in both the models by comparing the scatter plots of simulated methylation proportions between proliferating and senescent cells classfied by predicted states in Figures 4.2b and 4.2d for the *moderately overlapped* case, Figures 4.3b and 4.3d for the *well separated* case and Figures 4.4b and 4.4d for the *realistic* case, respectively.

For this randomly selected simulation study, the histograms of state 2 posterior probabilities for all the cases are also plotted in Figures (the histogram of state 1 posterior probabilities is just an inverse image of state 2). For *moderately overlapped* case in both the models as can be seen from the histograms (Figures 4.5a, 4.5b), the medians of the posterior probabilities for state 2 are close to 0.5. Thus, it gets extremely difficult to classify the correct states. For *well separated* cases and *realistic* cases, the states are strongly classified as extreme posterior state-membership probabilities close to 0 or 1 can be obtained. In Figures 4.6a, 4.6b and Figure 4.7a, 4.7b, the posterior probabilities (mostly very close to 0 or 1) admit little uncertainty in the state reconstruction. The histograms for the *moderately overlapped* case in both the models are symmetric whereas the histograms of the *well separated* and *realistic* cases are either U-shaped, J-shaped or reflected J-shaped depending on the distributions of the state labels (state 1 and state 2).

I examine the performance of all the three cases for both the models using receiver operating characteristic (ROC) curves based on the simulation study design. The ROC curve explains the relationship between the false positive rate (FPR) and true positive rate (TPR) of inferred methylation status at each CpG site. The TPR, also termed as sensitivity, is the proportion of correctly identi-

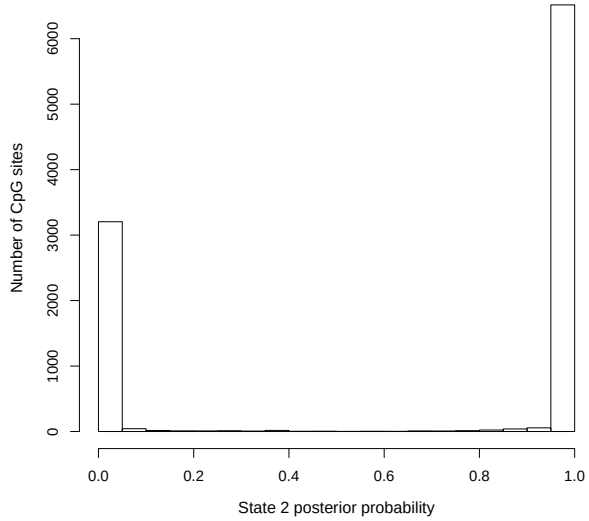


(a) Histogram for *BBDM*

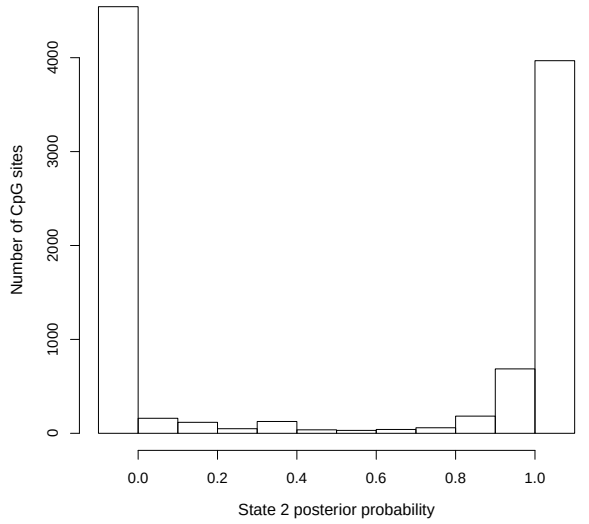


(b) Histogram for *BBCM*

Figure 4.5: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *moderately overlapped* case: (a) *BBDM* and (b) *BBCM* based on one randomly selected simulation.

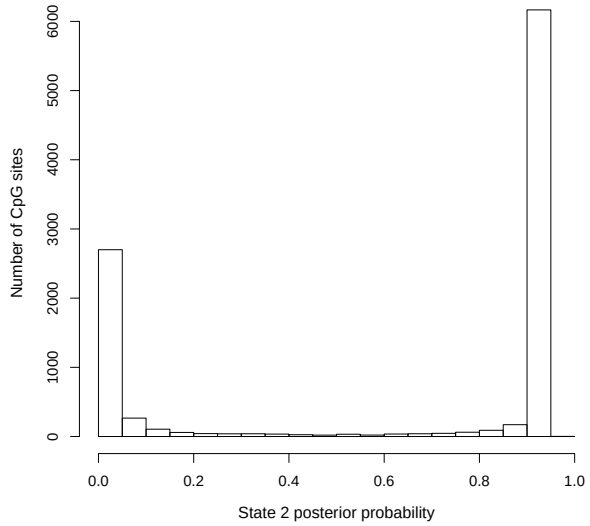


(a) Histogram for *BBDM*

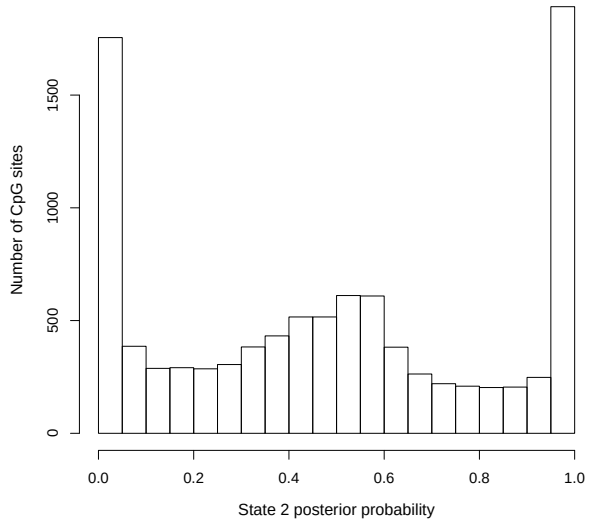


(b) Histogram for *BBCM*

Figure 4.6: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *well separated* case: (a) *BBDM* and (b) *BBCM* based on one randomly selected simulation.

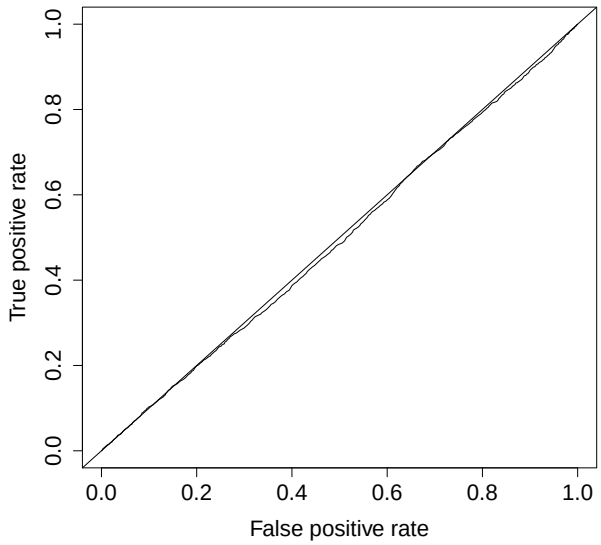


(a) Histogram for *BBDM*

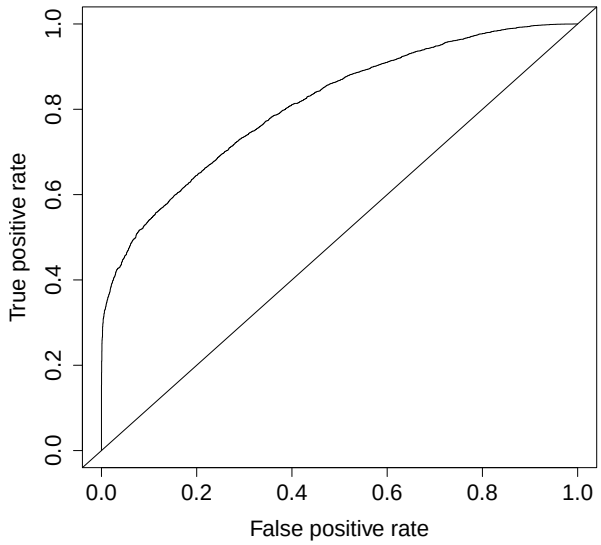


(b) Histogram for *BBCM*

Figure 4.7: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *realistic* case: (a) *BBDM* and (b) *BBCM* based on one randomly selected simulation.

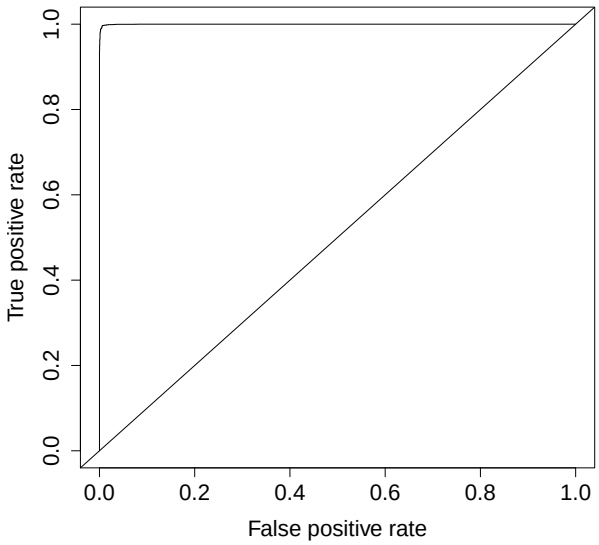


(a) ROC curve for *BBDM*

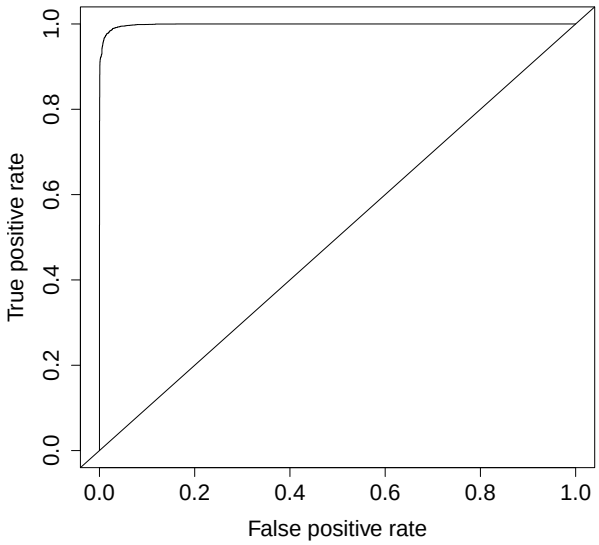


(b) ROC curve for *BBCM*

Figure 4.8: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the ROC curves for the *moderately overlapped* case: (a) *BBDM* and (b) *BBCM*.

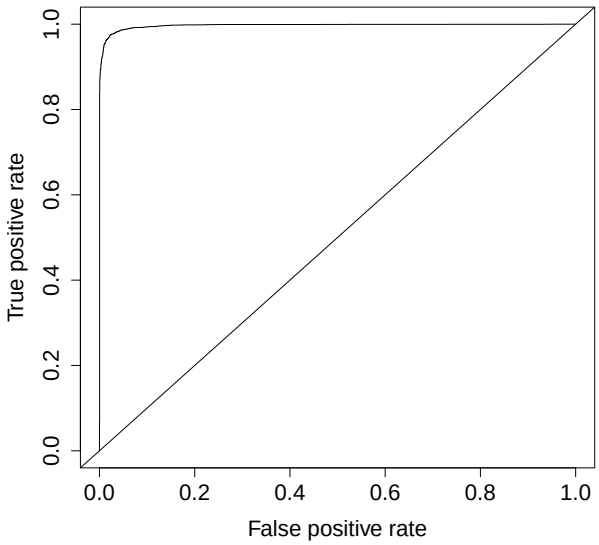


(a) ROC curve for *BBDM*

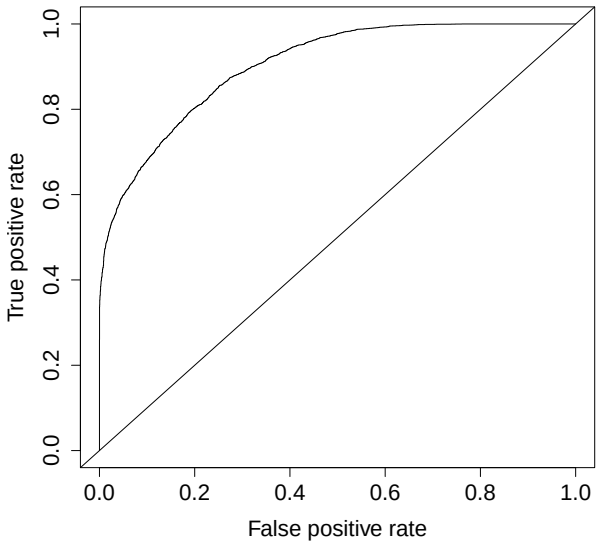


(b) ROC curve for *BBCM*

Figure 4.9: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the ROC curves for the *well separated* case: (a) *BBDM* and (b) *BBCM*.



(a) ROC curve for *BBDM*



(b) ROC curve for *BBCM*

Figure 4.10: For the simulation study of *BBDM* and *BBCM*, the 2 panels depict the ROC curves for the *realistic* case: (a) *BBDM* and (b) *BBCM*.

fied differentially methylated CpG sites. The FPR is the proportion of similarly methylated CpG sites which are incorrectly classified. The obtained ROC curves are also plotted in Figures 4.8, 4.9, 4.10, respectively. In a ROC curve the TPR (sensitivity) is plotted in function of the FPR (1-specificity). A test with perfect discrimination (no overlap in the two distributions) has a ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test. Figures 4.8a, 4.8b display the poor performance in the *moderately overlapped* case for both the models. Clearly, from Figures 4.9a, 4.9b, 4.10a, 4.10b, the *well separated* case for both the models outperforms the *realistic* case by a small margin.

4.4 Real data study

In this section, I illustrate the potential of my proposed BS-seq methylation data analysis method. I have applied the extended HMMmethState method to analyze BS-seq data from Cruickshanks et al. (2013). I analysed a 90.23458 Mb region ($0.060034 - 90.294609$ Mb on chromosome 16) of $2,165,796$ CpG sites.

4.4.1 Inference via MCMC

In this section, I assess the results obtained using MCMC techniques and the convergence properties of MCMC chains using various diagnostics for the real data Chromosome 16. I tried to make sure that the MCMC chains run long enough such that the samples of the parameters could be regarded as a good representation of their respective posterior distributions. I ran the augmented Gibbs sampler for $10,000$ iterations for 3 parallel chains, thinning the chains and saving every 10^{th} value of the updates, to reduce autocorrelation between consecutive updates and save storage space. The parameters of the proposal distributions were also tuned to get coherent M-H updates. I checked the posterior samples using

various convergence diagnostics to establish the fact that the posterior samples of the parameters represented their corresponding posterior distributions.To assess the convergence and mixing properties of the MCMC updates, I ran 3 MCMC chains intializing with different starting points and also did a burn-in the first 300 MCMC updates. Optimizing the tuning parameter of the proposal distribution played a significant role in obtaining less correlated consecutive draws, thus enhancing the efficiency of the M-H algorithm. The proposal distributions of the hyperparameters were tuned appropriately in order to obtain optimal acceptance rate which resulted in acceptance rates in the range of $(0.25,\ 0.43)$. In addition, I checked a few other convergence diagnostics to reaffirm my claim in the convergence of the parameter estimates. I used PSRF (Gelman and Rubin, 1992) on the 3 MCMC chains with dispersed starting values to examine the convergence for each of the parameter to the same target distribution. The PSRF values are only slightly above 1 (Tables 2 and 3), which is consistent with convergence of the chains.

I carried out all the convergence diagnostics discussed and no evidence of non-convergence was obtained from any of the diagnostics. From the noisy traceplots, it can be easily inferred that the 3 chains mixed properly. Traceplots and PSRF plots are presented in Appendix 7.2.3. The estimates of the posterior mean, posterior standard deviation (S.D.), 95% credible intervals are presented in Table 2 for emission hyperparameters and Table 3 for transition parameters of both the models for all the 3 chains.

The number of iterations for the simulation and real studies are not the same because in some cases (for example, *moderately overlapped*), the transition parameters were taking longer to converge to their appropriate stationary distributions compared to the *realistic* and *well separated* cases. Thus, to make the number of iterations conformable with the other two cases, I have used $60,000$ iterations

and 20,000 burn-in. Thinning was not implemented in the simulation studies.

Furthermore, for both the model boxplots (Figures 4.11, 4.13), y-axis denotes the difference between methylation proportions of senescent and proliferating in various categories. Hypo refers to the category when the proportions of senescent cells is greater than proliferating cells and vice-versa for Hyper. Note how the DMCs are hypomethylated on average (since the median difference is negative), consistent with the biological presumption (Cruickshanks et al., 2013). Also the Non-DMCs are indeed similarly methylated since the boxplot is centred on zero. Additionally, the histograms of state 2 posterior probabilities for model *BBDM* show strong classification of states in Figure 4.12 whereas the classification of the states is moderately weak in the case of model *BBCM* as shown in Figure 4.14, as many of the state 2 posterior probabilities vary between 0.2 and 0.8.

4.5 Discussion

In this chapter, I have described my proposed HMMmethState method for identifying DMCs from BS-seq methylation data. I described the structure of my models and their association with the data-generating process. In addition, I developed an efficient augmented Gibbs sampling method for applying MCMC based techniques to datasets with hidden states, with the help of Forward-sum recursion. I designed my proposed method HMMmethState: a HMM, where I assumed that the data followed an independent bivariate Binomial distribution conditional on the true underlying methylation proportions and hidden states, i.e., the methylation status of the CpG sites at the first stage of the hierarchical model. The underlying methylation proportions for each CpG site were assumed to be centrally clustered around a state-specific mean with a state-specific variance at the second stage of the hierarchical model. The advantage of using the Binomial distribution at the first stage was that it involved CpG site-specific va-

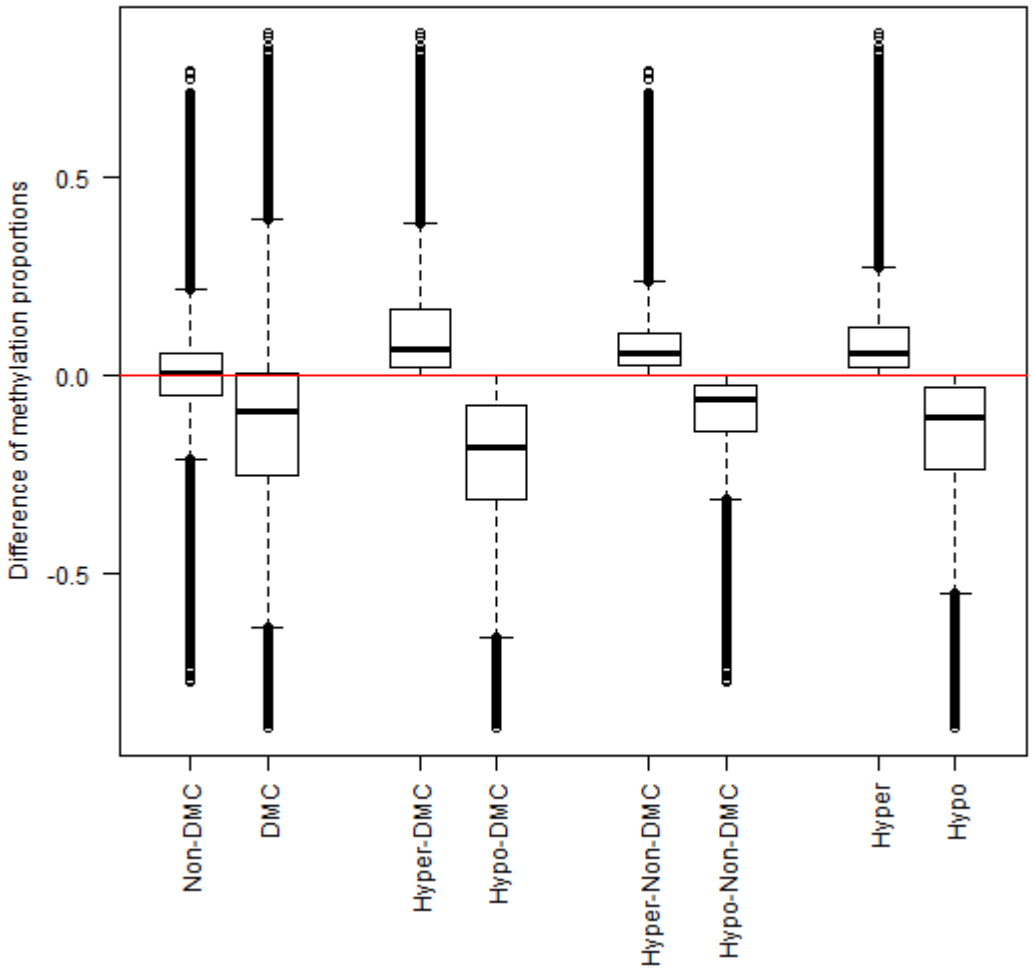


Figure 4.11: For the real study of *BBDM*, boxplots of the difference of methylation proportions between proliferating and senescent cells classified by various categories defined in the text.

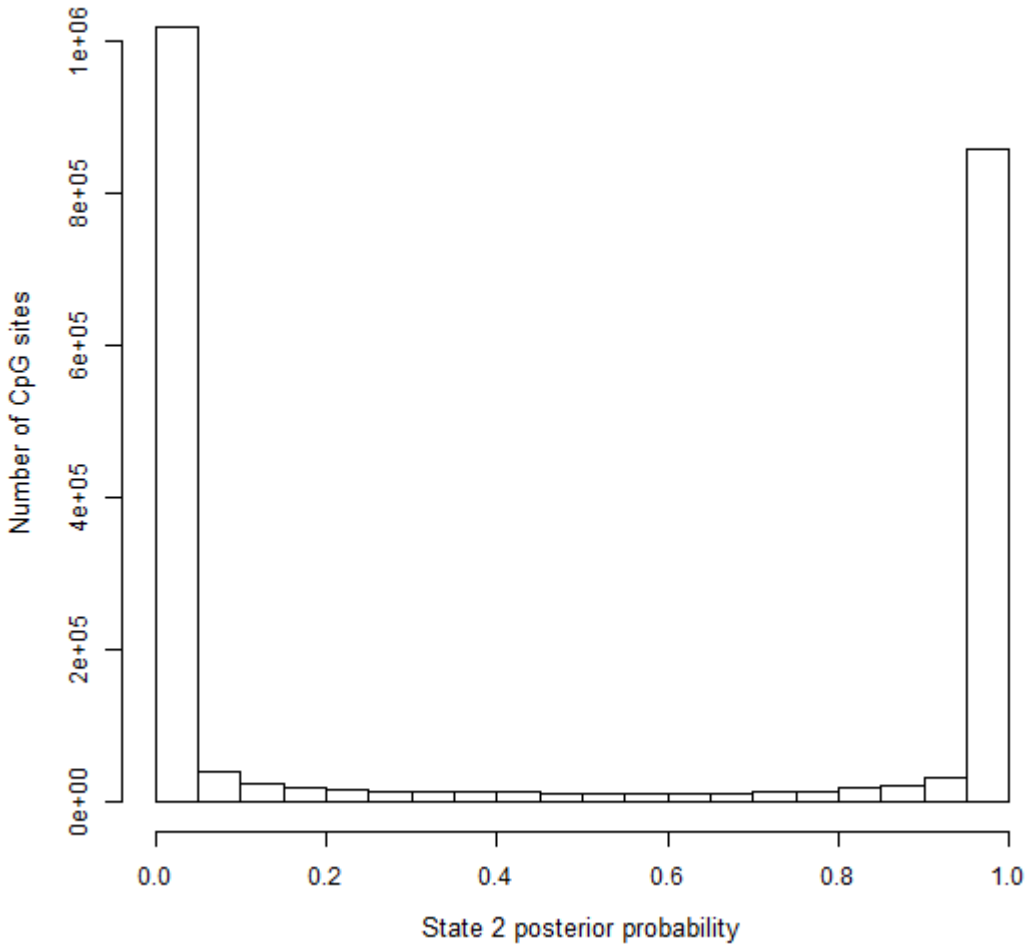


Figure 4.12: For the real study of *BBDM*, histogram of posterior state 2 probabilities.

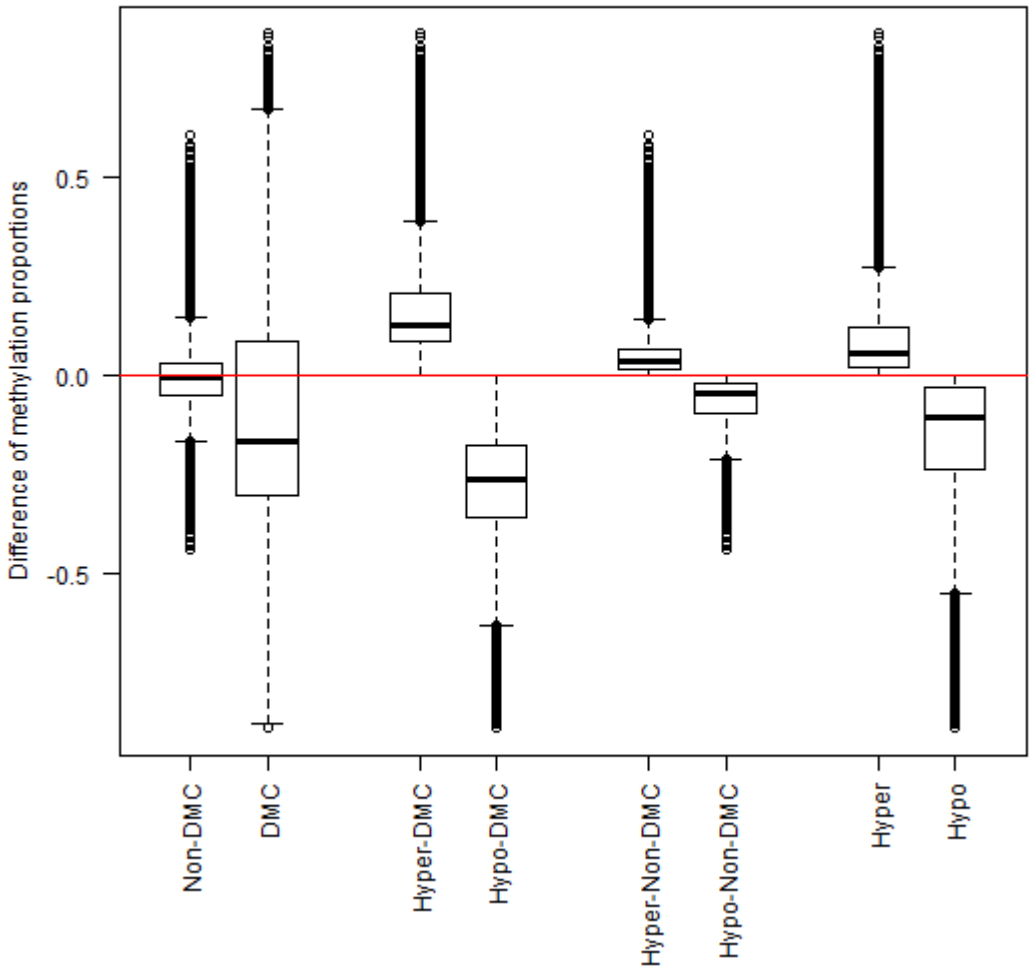


Figure 4.13: For the real study of *BBCM*, boxplots of the difference of methylation proportions between proliferating and senescent cells classified by various categories defined in the text.

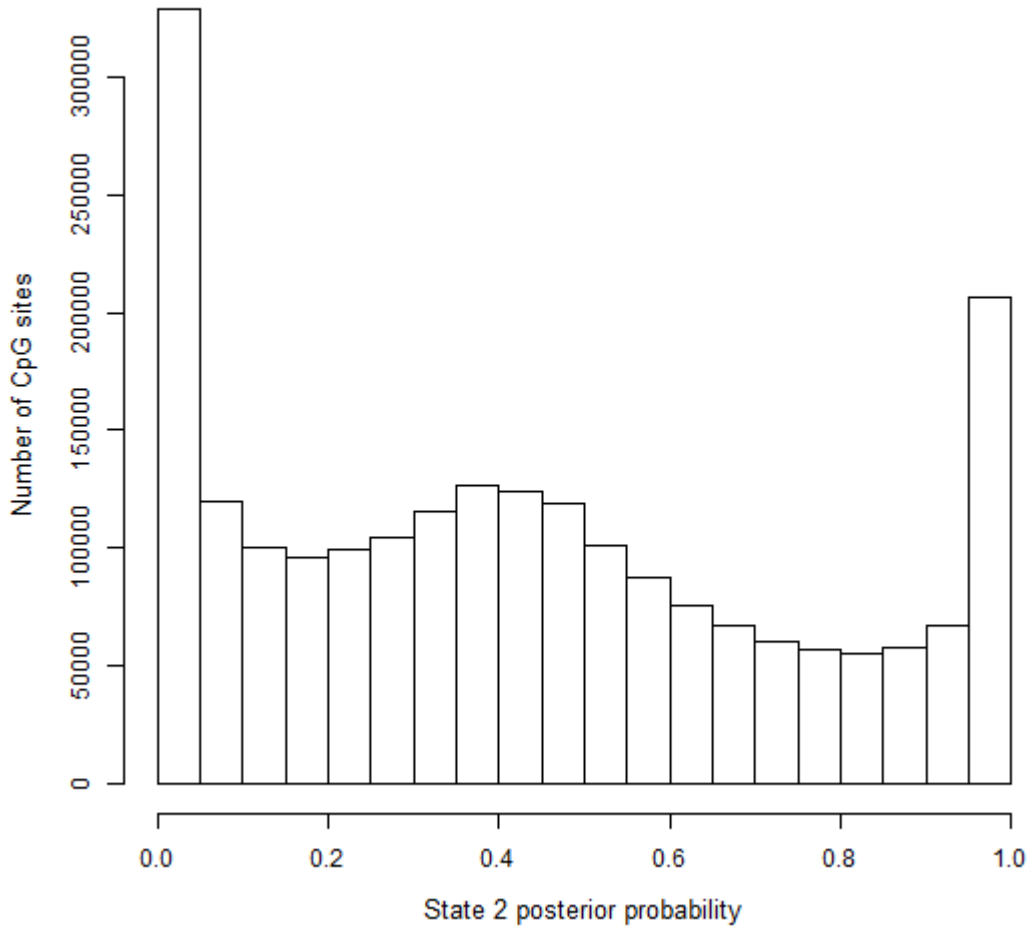
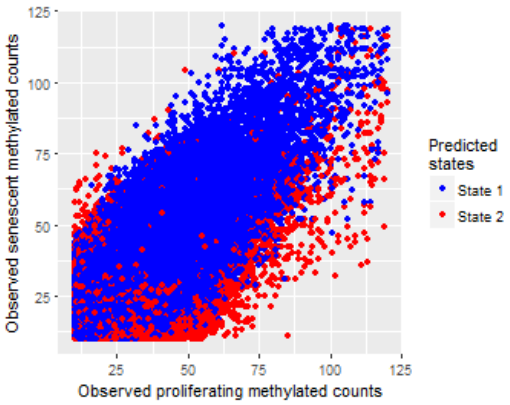
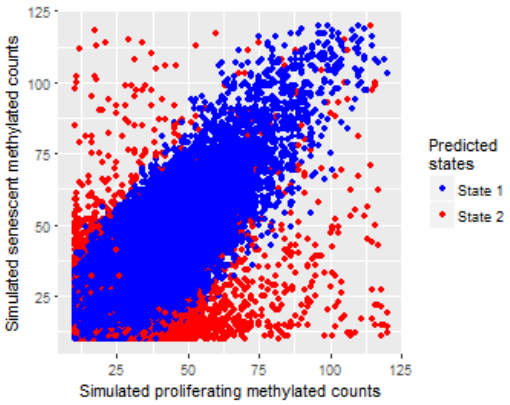


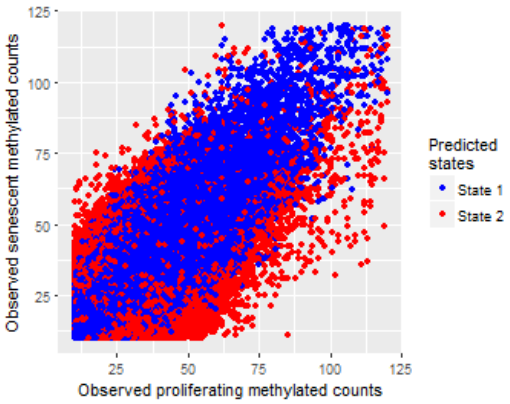
Figure 4.14: For the real study of *BBCM*, histogram of posterior state 2 probabilities.



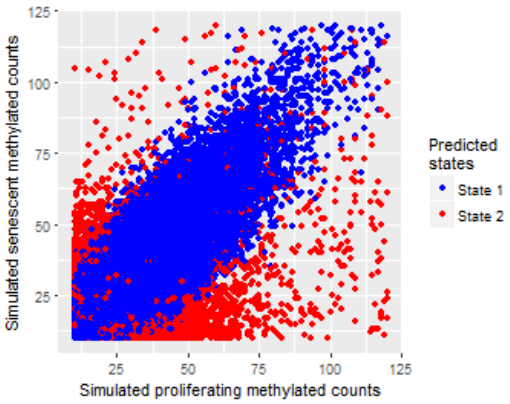
(a) *BBDM*: Observed



(b) *BBDM*: Simulated



(c) *BBCM*: Observed



(d) *BBCM*: Simulated

Figure 4.15: For the the real study, the 4 panels depict the scatter plots: (a) A scatter plot of observed methylated counts for the proliferating and senescent cells classified by the predicted states for *BBDM*. (b) A scatter plot of simulated methylated counts (generated using the posterior mean estimates) for the proliferating and senescent cells classified by the predicted states for *BBDM*. (c) A scatter plot of observed methylated counts for the proliferating and senescent cells classified by the predicted states for *BBCM*. (d) A scatter plot of simulated methylated counts (generated using the posterior mean estimates) for the proliferating and senescent cells classified by the predicted states for *BBCM*.

riances, i.e., the data point (the methylated counts of proliferating and senescent cells) for each CpG site was generated using the information obtained from the total count and the true underlying methylation proportion parameter for each CpG site. Thus, the hierarchical model became more capable of describing the variability among CpG sites within each state. In order to obtain computational simplicity, I implemented a simple version of the bivariate Binomial hierarchical HMM with a collapsed distributional structure due to Beta-Binomial conjugacy. I have presented the scatter plots of the methylation counts of proliferating cells against senescent cells for the observed data and for the fitted models in Figure 4.15. The visual posterior predictive checking, using the posterior mean estimates of the model parameters in Figure 4.15 indicates that the fitted models fail to capture the correlation between the methylated counts of proliferating and senescent cells. Figures 4.15a and 4.15c show the scatterplots of the observed data classified by the predicted states. On the other hand, Figures 4.15b and 4.15d show scatterplots of the fitted data classified by the predicted states. From Figures 4.15b and 4.15d, it can be interpreted that the correlation between the methylated counts of proliferating and senescent cells cannot be captured by the HMMmethState models. The main drawback of the bivariate Beta-Binomial emission model is that it cannot induce correlation between the methylated counts of these two cell types. Although it offers a natural interpretation to the distribution of the data due to its collapsed hierarchical structure (Beta-Binomial conjugacy), it fails to accomodate some features of the observations. In the next chapter, I present an extended version of HMMmethState models which can incorporate a correlation parameter into the model for more robust inference.

Chapter 5

Model Extensions

In this chapter, I propose extensions of the HMM-based HMMmethState models proposed in Chapter 4 for predicting DMCs in BS-seq data. The bivariate Beta-Binomial emission model seems reasonably adequate to model the BS-seq data but fails to capture certain features, primarily the correlation between the methylated counts of the two cell types. This is visible in the visual posterior predictive checking analysis in Chapter 4 (Figure 4.15). The scatterplots in Figure 4.15 show that the data exhibit strong correlation between the methylated counts of the proliferating and senescent cells which was not allowed by the fitted models. Failure to properly address the correlation between the methylated counts of the two cell types may result in misleading inference. Thus, to incorporate the correlation feature in the data, I propose a bivariate Normal distribution at the 2^{nd} stage of the HMMmethState model which introduces a correlation parameter.

5.1 Model assumptions

Let, as previously denoted in Chapter 4, x_t^p and x_t^s denote the methylated counts in proliferating and senescent cells of the pair of random variables X_t^p and X_t^s at the t^{th} CpG site, such that X_t^p and X_t^s independently follow Binomial distributions with parameters $(n_t^p, logit^{-1}(q_t^p)$ and $(n_t^s, logit^{-1}(q_t^s)$ respectively, where $\mathbf{X}_t =$

(X_t^p, X_t^s), $x_t = (x_t^p, x_t^s)$ and $\mathbf{X} = (\mathbf{X}_1, \cdots, \mathbf{X}_T)$, $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_T)$, such that

$$X_t^p | q_t^p \sim Bin\Big(n_t^p, logit^{-1}(q_t^p)\Big), \qquad t = 1, \ldots, T \tag{5.1}$$

and

$$X_t^s | q_t^s \sim Bin\Big(n_t^s, logit^{-1}(q_t^s)\Big), \qquad t = 1, \ldots, T, \tag{5.2}$$

where q_t^p and q_t^s are the logit transforms of probability parameters p_t^p and p_t^s as explained in Section 4.1 of methylation of proliferating and senescent cells, respectively, at the t^{th} CpG site, such that,

$$q_t^c = \log\left(\frac{p_t^c}{1 - p_t^c}\right), \quad c = p,\ s.$$

n_t^p and n_t^s are the total number of reads from the two cell types. For notational simplicity, let the pair of logit parameters q_t^p and q_t^s be denoted by $\mathbf{Q}_t = (q_t^p,\ q_t^s)$ for $t = 1, \ldots, T$.

For state $k = 1, 2$, the underlying logits will be written as

$$\mathbf{Q}_t^k = (q_t^{pk},\ q_t^{sk}),\ k = 1, 2, \tag{5.3}$$

i.e., $\mathbf{Q}_t^k$ is the pair of auxiliary parameters for state k at each CpG site, where q_t^{pk} and q_t^{sk} denote the underlying auxiliary logit parameters for state k at t^{th} CpG site of proliferating and senescent cells, respectively.

In order to account for the variability of the mean among CpG sites in the same state, I constructed a hierarchical model where I have state-specific auxiliary logit parameters for each CpG site and these auxiliary logit parameters are eventually clustered around a state-specific mean with a state-specific variance and correlation.

Thus, the structure of the hierarchical bivariate Normal-Logit-Binomial emission

model is:

$$X_t^p|Z_t = k \sim Bin\Big(n_t^p, logit^{-1}(q_t^{pk})\Big) \text{ and } X_t^s|Z_t = k \sim Bin\Big(n_t^s, logit^{-1}(q_t^{sk})\Big),$$
$$\mathbf{Q}_t^k|Z_t = k \sim BVN(\boldsymbol{\theta}_k),\ k = 1,\ 2 \text{ and } t = 1, \ldots, T, \tag{5.4}$$

where $\boldsymbol{\theta}_k = (\mathbf{M}_k, \boldsymbol{\Sigma}_k)$ and $BVN(.)$ is the bivariate Normal distribution, such that,

$$\mathbf{M}_1 = \begin{bmatrix} \mu_* \\ \mu_* \end{bmatrix};\ \mathbf{M}_1 = \begin{bmatrix} \mu_p \\ \mu_s \end{bmatrix};\ \boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_*^2 & \sigma_*^2\rho_* \\ \rho_*\sigma_*^2 & \sigma_*^2 \end{bmatrix};\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_p^2\rho_2 & \sigma_p\sigma_s\rho_2 \\ \sigma_s\sigma_p\rho_2 & \sigma_s^2\rho_2 \end{bmatrix}. \tag{5.5}$$

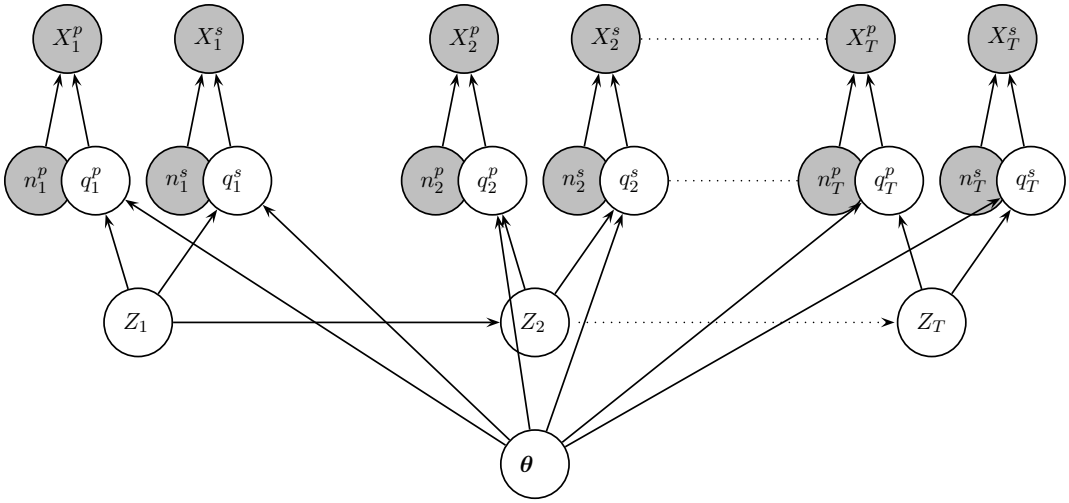
For notational simplicity, I denote the bivariate Normal state-specific parameters as follows: $\boldsymbol{\theta}_1 = (\mu_*, \sigma_*^2, \rho_*)$; $\boldsymbol{\theta}_2 = (\mu_p, \mu_s, \sigma_p^2, \sigma_s^2, \rho_2)$.
Now, if the methylation levels in proliferating and senescent cell are the same at the t^{th} CpG site, i.e., $Z_t = 1$, then

$$\mathbf{Q}_t^1 \sim BVN(\boldsymbol{\theta}_1). \tag{5.6}$$

Similarly, if the methylation in proliferating and senescent cells are different at the t^{th} CpG site, i.e., $Z_t = 2$, then

$$\mathbf{Q}_t^2 \sim BVN(\boldsymbol{\theta}_2). \tag{5.7}$$

I will further explain the bivariate-Binomial emission model in Section 5.1.1 and hierarchical structure of the bivariate-Binomial-Normal-Logit emission model (Figure 5.1) along with the auxiliary emission parameters in Section 5.1.2.



For: $t = 1, \ldots, T$

$$X_t^p \sim \begin{cases} \text{Bin}(n_t^p, logit^{-1}(q_t^p)), & \text{if } Z_t = 1 \\ \text{Bin}(n_t^p, logit^{-1}(q_t^p)), & \text{if } Z_t = 2 \end{cases} \qquad X_t^s \sim \begin{cases} \text{Bin}(n_t^s, logit^{-1}(q_t^s)), & \text{if } Z_t = 1 \\ \text{Bin}(n_t^s, logit^{-1}(q_t^s)), & \text{if } Z_t = 2 \end{cases}$$

$$q_t^p, q_t^s \sim \begin{cases} \text{BVN}(\boldsymbol{\theta}_1), & \text{if } Z_t = 1 \\ \text{BVN}(\boldsymbol{\theta}_2), & \text{if } Z_t = 2 \end{cases}$$

Figure 5.1: Graphical representation of the bivariate Normal-logit emission model. The grey circles refer to the fixed values of the total counts and data respectively, while the white circles refer to auxiliary emission parameters, hyperparameters and hidden states that are inferred.

5.1.1 Binomial emission distributions of the model

Define the emission probability $P(\mathbf{x}_t \,|\mathbf{Q}_t^k, Z_t = k) = b_k(t)\,,$ where $k = 1, 2$. The emission probability of the pair of observation $\mathbf{x}_t = (x_t^p,\ x_t^s)$ conditional on the hidden state $Z_t = k,\ (k = 1,\ 2)$ is given by

$$\begin{aligned} b_k(t) &= P\left(\mathbf{x}_t \,\middle|\mathbf{Q}_t^k, Z_t = k\right) \\ &= Bin\left(x_t^p; n_t^p, \frac{e^{q_t^{pk}}}{1+e^{q_t^{pk}}}\right) Bin\left(x_t^s; n_t^s, \frac{e^{q_t^{sk}}}{1+e^{q_t^{sk}}}\right),\ k = 1,\ 2. \end{aligned} \tag{5.8}$$

5.1.2 Auxiliary emission parameters

To classify the states, I need to have different properties of (q_t^{p1}, q_t^{s1}) and (q_t^{p2}, q_t^{s2}) $(t = 1, \ldots, T)$, the auxiliary parameters (inverse-logit parameters) of proliferating and senescent CpG site for both the states. I have introduced auxiliary emission parameters and used a 3-stage hierarchical Bayesian model assuming bivariate Normal state-dependent conditional priors on these parameters. I have defined $\mathbf{Q}_t^k = (q_t^{pk},\ q_t^{sk}) \sim BVN(\boldsymbol{\theta}_k)$.

Stage I: Methylation counts of proliferating and senescent cells sampled from Bivariate Binomial Emission distributions with state-dependent auxiliary parameters:

$$\mathbf{x}_t \,|\mathbf{Q}_t^k \propto P\left(\mathbf{X}_t|\mathbf{Q}_t^k\right), \tag{5.9}$$

where $P\left(\mathbf{x}_t \,|\mathbf{Q}_t^k\right) = P(\mathbf{x}_t \,|\mathbf{Q}_t^k, Z_t = k),\ \ k = 1, 2$ is the bivariate Binomial emission distribution for state k.

Stage II: Auxiliary emission parameters are generated from bivariate Normal prior distributions conditional on the differentially methylated (similarly methylated) states:

$$\mathbf{Q}_t^k \,|Z = k \sim BVN(\boldsymbol{\theta}_k), \tag{5.10}$$

such that,

$$\phi(\mathbf{Q}_t^k, \mathbf{M}_k, \boldsymbol{\Sigma}_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{Q}_t^k - \mathbf{M}_k)^T {\boldsymbol{\Sigma}_k}^{-1} (\mathbf{Q}_t^k - \mathbf{M}_k) \right)}{2\pi \left|\boldsymbol{\Sigma}_k\right|}. \tag{5.11}$$

where $\phi(.)$ denotes the bivariate Normal density. I can further simplify the equations specific to each state, $k = 1, 2$,

$$\begin{aligned}\phi(\mathbf{Q}_t^1, \mathbf{M}_1, \boldsymbol{\Sigma}_1) = \frac{1}{2\pi\sigma_*^2\sqrt{1-\rho_*^2}} \times \\ \exp\Bigg[-\frac{1}{2\sigma_*^2(1-\rho_*^2)} \Big\{ \left(q_t^{p1} - \mu_*\right)^2 - 2\rho_* \left(q_t^{p1} - \mu_*\right)\left(q_t^{s1} - \mu_*\right) \\ + \left(q_t^{s1} - \mu_*\right)^2 \Big\} \Bigg]\end{aligned}$$

and

$$\begin{aligned}\phi(\mathbf{Q}_t^2, \mathbf{M}_2, \boldsymbol{\Sigma}_2) = \frac{1}{2\pi\sigma_p\sigma_s\sqrt{1-\rho^2}} \times \\ \exp\Bigg[-\frac{1}{2(1-\rho_2^2)} \Bigg\{ \frac{\left(q_t^{p2} - \mu_p\right)^2}{\sigma_p^2} - \frac{2\rho_2 \left(q_t^{p2} - \mu_p\right)\left(q_t^{s2} - \mu_s\right)}{\sigma_p\sigma_s} \\ + \frac{\left(q_t^{s2} - \mu_s\right)^2}{\sigma_s^2} \Bigg\} \Bigg].\end{aligned}$$

Stage III: Global Hyperparameters $\boldsymbol{\theta}_k,\ k = 1, 2$ follow hyper-prior distributions $p(\boldsymbol{\theta}_k)$:

$$\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k). \tag{5.12}$$

5.1.3 Normal-Logit-Binomial hierarchical HMM models

In this section, I define the two hierarchical Normal-Logit-Binomial HMM models by combining the Normal-Logit-Binomial emission probability distributions and transition probability distributions. The transition probability models were explained in details in Sections 4.1.3 and 4.1.4, respectively.

- Model NLBDM: this model combines the Normal-Logit-Binomial emission probability model in Section 5.1.1 and homogeneous transition probability model in Section 4.1.3 through (4.13) and (4.14).
- Model NLBCM: this model combines the Normal-Logit-Binomial emission probability model in Section 5.1.1 and non-homogeneous continuous-index transition probability model in Section 4.1.4 through (4.19) and (4.20).

For notational consistency in this chapter, I have also assumed the same set of HMM transition model parameters.

5.1.4 Computing the likelihood

In this section, I describe the general version of the likelihood for model M where M represents the true model, i.e., $M = NLBDM,\ NLBCM$.
Let the set of all parameters and hyperparameters be generically denoted by $\boldsymbol{\zeta}^{(M)} = (\boldsymbol{\eta}^{(M)}, \boldsymbol{\theta}^{(M)}, \boldsymbol{\tau}^{(M)})$ for both the models as described in Section 5.1.3 where $\boldsymbol{\eta}^{(M)} = (\boldsymbol{\eta}_1^{(M)},\ \boldsymbol{\eta}_2^{(M)})$ such that, $\boldsymbol{\eta}_k^{(M)} = \left(\boldsymbol{\eta}_k^{(M)}(1), \cdots, \boldsymbol{\eta}_k^{(M)}(T)\right)$. Then, I denote $\boldsymbol{\eta}_k^{(M)}(t) = \mathbf{Q}_t^k = (q_t^{pk}, q_t^{sk})$ for $k = 1,\ 2$ and $t = 1, \ldots, T$; where $\boldsymbol{\theta}^{(M)} = (\boldsymbol{\theta}_1^{(M)},\ \boldsymbol{\theta}_2^{(M)})$, such that $\boldsymbol{\theta}_1^{(M)} = (\mu_*, \sigma_*^2, \rho_*)$ and $\boldsymbol{\theta}_2^{(M)} = (\mu_p, \mu_s, \sigma_p^2, \sigma_s^2, \rho_2)$ and $\boldsymbol{\tau}^{(M)}$ for model M. The joint probability distribution of the observations $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and the sequence of the hidden states $\mathbf{Z} = (Z_1, \ldots, Z_T)$ for model M conditional on the model parameters $\boldsymbol{\zeta}^{(M)}$ can be interpreted as the

complete data likelihood of the observations and the states:

$$\begin{aligned} P(\mathbf{x}, \mathbf{Z} \,|\boldsymbol{\zeta}^{(M)}) &= \pi_{Z_1}^{(M)} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}_{Z_1}^{(M)}(1)\right) \prod_{t=2}^{T} \tau_{Z_{(t-1)},Z_t}^{(M)}(t) P_{Z_t}\left(\mathbf{x}_t|\boldsymbol{\eta}_{Z_t}^{(M)}(t)\right) \\ &= \pi_{Z_1}^{(M)} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}_{Z_1}^{(M)}(1)\right) \tau_{Z_1,Z_2}^{(M)}(2) P_{Z_2}\left(\mathbf{x}_2|\boldsymbol{\eta}_{Z_2}^{(M)}(2)\right) \cdots \\ &\quad \cdots \tau_{Z_{(T-1)},Z_T}^{(M)}(T) P_{Z_T}\left(\mathbf{x}_T|\boldsymbol{\theta}_{Z_T}^{(M)}(T)\right), \end{aligned} \tag{5.13}$$

where $P_k\left(\mathbf{x}_t|\boldsymbol{\eta}_k^{(M)}(t)\right) = P\left(\mathbf{x}_t|Z_t = k; \boldsymbol{\eta}_k^{(M)}(t)\right)$, $\pi_k^{(M)} = P(Z_1 = k)$ and $\tau_{kl}^{(M)}(t) = P(Z_t = l|Z_{t-1} = k; \boldsymbol{\tau}^{(M)})$ for $k, l = 1,\ 2$.
Basically, (5.8) provides $P_k\left(\mathbf{x}_t|\boldsymbol{\eta}_k^{(M)}(t)\right)$, such that,

$$\begin{aligned} P_1\left(\mathbf{x}_t|\boldsymbol{\eta}_1^{(M)}(t)\right) &= P\left(\mathbf{x}_t|Z_t = 1; \boldsymbol{\eta}_1^{(M)}(t)\right) \\ &= P(x_t^p, x_t^s|q_t^{p1}, q_t^{s1}; Z_t = 1) \\ &= Bin\left(x_t^p; n_t^p, \frac{e^{q_t^{p1}}}{1+e^{q_t^{p1}}}\right) Bin\left(x_t^s; n_t^s, \frac{e^{q_t^{s1}}}{1+e^{q_t^{s1}}}\right) \end{aligned} \tag{5.14}$$

and

$$\begin{aligned} P_2\left(\mathbf{x}_t|\boldsymbol{\eta}_2^{(M)}(t)\right) &= P\left(\mathbf{x}_t|Z_t = 2; \boldsymbol{\eta}_2^{(M)}(t)\right) \\ &= P(x_t^p, x_t^s|q_t^{p2}, q_t^{s2}; Z_t = 2) \\ &= Bin\left(x_t^p; n_t^p, \frac{e^{q_t^{p2}}}{1+e^{q_t^{p2}}}\right) Bin\left(x_t^s; n_t^s, \frac{e^{q_t^{s2}}}{1+e^{q_t^{s2}}}\right). \end{aligned} \tag{5.15}$$

Now, the detailed joint probability distribution expression for the observed methylation data $\mathbf{x}$ and the sequence of the hidden states (methylation status) $\mathbf{Z}$ can be obtained from the emission quantities (5.14) and (5.15) and the hidden states probability expressions from (4.13), (4.14) for model $NLBDM$ and (4.19), (4.20)

for model $NLBCM$. So, (5.13) can be re-written specific to model $NLBDM$ as

$$\begin{aligned}
P(\mathbf{x}, \mathbf{Z} \,|\boldsymbol{\zeta}^{(D)}) &= \pi_{Z_1}^{(D)} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}_{Z_1}^{(D)}(1)\right) \prod_{t=2}^{T} \tau_{Z_{(t-1)},Z_t}^{(D)}(t) P_{Z_t}\left(\mathbf{x}_t|\boldsymbol{\eta}_{Z_t}^{(D)}(t)\right) \\
&= \pi_{Z_1}^{(D)} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}_{Z_1}^{(D)}(1)\right) \tau_{Z_1,Z_2}^{(D)}(2) P_{Z_2}\left(\mathbf{x}_t|\boldsymbol{\eta}_{Z_2}^{(D)}(2)\right) \cdots \\
&\quad \cdots \tau_{Z_{(T-1)},Z_T}^{(D)}(T) P_{Z_T}\left(\mathbf{x}_T|\boldsymbol{\eta}_{Z_T}^{(D)}(T)\right) \\
&= \prod_{t=1}^{T} \Bigg(\left[Bin\left(x_t^p; n_t^p, \frac{e^{q_t^{p1}}}{1+e^{q_t^{p1}}}\right) Bin\left(x_t^s; n_t^s, \frac{e^{q_t^{s1}}}{1+e^{q_t^{s1}}}\right) \right]^{\mathbf{I}[Z_t=1]} \\
&\quad \times \left[Bin\left(x_t^p; n_t^p, \frac{e^{q_t^{p2}}}{1+e^{q_t^{p2}}}\right) Bin\left(x_t^s; n_t^s, \frac{e^{q_t^{s2}}}{1+e^{q_t^{s2}}}\right) \right]^{\mathbf{I}[Z_t=2]} \Bigg) \\
&\quad \times {\pi_1}^{\mathbf{I}[Z_1=1]} (1-\pi_1)^{\mathbf{I}[Z_1=2]} Bin(t_{11}; t_{11}+t_{12}, \tau_{11}) \\
&\quad (t_{21}; t_{21}+t_{22}, \tau_{21}). \qquad (5.16)
\end{aligned}$$

Similarly, equation (5.13) can be re-written specific to model $NLBCM$ as

$$
\begin{aligned}
P(\mathbf{x}, \mathbf{Z} \,|\boldsymbol{\zeta}^{(C)}) &= \pi^{(C)}_{Z_1} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}^{(C)}_{Z_1}(1)\right) \prod_{t=2}^{T} \tau^{(C)}_{Z_{(t-1)},Z_t}(t) P_{Z_t}\left(\mathbf{x}_t|\boldsymbol{\eta}^{(C)}_{Z_t}(t)\right) \\
&= \pi^{(C)}_{Z_1} P_{Z_1}\left(x_1|\boldsymbol{\eta}^{(C)}_{Z_1}(1)\right) \tau^{(C)}_{Z_1,Z_2}(2) P_{Z_2}\left(\mathbf{x}_t|\boldsymbol{\eta}^{(C)}_{Z_2}(2)\right) \cdots \\
&\quad \cdots \tau^{(C)}_{Z_{(T-1)},Z_T}(T) P_{Z_T}\left(\mathbf{x}_T|\boldsymbol{\eta}^{(C)}_{Z_T}(T)\right) \\
&= \prod_{t=1}^{T} \Bigg(\left[Bin\left(x^p_t; n^p_t, \frac{e^{q^{p1}_t}}{1+e^{q^{p1}_t}}\right) Bin\left(x^s_t; n^s_t, \frac{e^{q^{s1}_t}}{1+e^{q^{s1}_t}}\right) \right]^{\mathbf{I}[Z_t=1]} \\
&\quad \times \left[Bin\left(x^p_t; n^p_t, \frac{e^{q^{p2}_t}}{1+e^{q^{p2}_t}}\right) Bin\left(x^s_t; n^s_t, \frac{e^{q^{s2}_t}}{1+e^{q^{s2}_t}}\right) \right]^{\mathbf{I}[Z_t=2]} \Bigg) \\
&\quad \times [0.5]^{\mathbf{I}(Z_1=1)} [0.5]^{\mathbf{I}(Z_1=2)} \\
&\quad \times \prod_{t=2}^{T} \Bigg[\left(\frac{\lambda_2}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=1)} \\
&\quad \times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} - \frac{\lambda_1}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=1, Z_t=2)} \\
&\quad \times \left(\frac{\lambda_2}{\lambda_1+\lambda_2} - \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=1)} \\
&\quad \times \left(\frac{\lambda_1}{\lambda_1+\lambda_2} + \frac{\lambda_2}{\lambda_1+\lambda_2} e^{-(\lambda_1+\lambda_2)\Psi_t} \right)^{\mathbf{I}(Z_{t-1}=2, Z_t=2)} \Bigg].
\end{aligned}
\tag{5.17}
$$

Then, the likelihood of the observed methylation data $\mathbf{x}$ given the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ for model M can be expressed as

$$
\begin{aligned}
L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}) &= P(\mathbf{x}|\boldsymbol{\zeta}^{(M)}) \\
&= \sum_{Z_1,\ldots,Z_T} \pi^{(M)}_{Z_1} P_{Z_1}\left(\mathbf{x}_1|\boldsymbol{\eta}^{(M)}_{Z_1}(1)\right) \prod_{t=2}^{T} \left[\tau^{(M)}_{Z_{(t-1)},Z_t}(t) P_{Z_t}\left(\mathbf{x}_t|\boldsymbol{\eta}^{(M)}_{Z_t}(t)\right) \right].
\end{aligned}
\tag{5.18}
$$

5.1.5 Conditional Bivariate Normal Priors of the auxiliary emission parameters

I consider bivariate Normal priors for auxiliary emission parameters conditional on the global emission hyperparameters

$$\boldsymbol{\eta}_k^{(M)}(t)|\boldsymbol{\theta}_k^{(M)} \sim BVN(\boldsymbol{\theta}_k^{(M)}),\ k = 1, 2 \text{ and } t = 1, \ldots, T. \tag{5.19}$$

5.1.6 Choice of priors

The priors for the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition parameters $\boldsymbol{\tau}^{(M)}$ for both models remain the same as described in Section 4.1.7.

The prior for the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ can be decomposed into four parts: i) bivariate Normal priors for auxiliary parameters conditional on the global emission hyperparameters; ii) priors for the emission hyperparameters $\boldsymbol{\theta}^{(M)}$; iii) priors of the initial state parameters $\boldsymbol{\pi}^{(M)}$; iv) priors of the transition parameters $\boldsymbol{\tau}^{(M)}$:

$$p(\boldsymbol{\zeta}^{(M)}) = p(\boldsymbol{\eta}^{(M)}|\boldsymbol{\theta}^{(M)})\ p(\boldsymbol{\theta}^{(M)})\ p(\boldsymbol{\pi}^{(M)})\ p(\boldsymbol{\tau}^{(M)}), \tag{5.20}$$

For model M, the priors for the global emission hyperparameters $\boldsymbol{\theta}^{(M)}$ can be written as,

$$p(\boldsymbol{\theta}^{(M)}) = p(\mu_*)\ p(\sigma_*^2)\ p(\rho_*)\ p(\mu_p)\ p(\mu_s)\ p(\Sigma_2). \tag{5.21}$$

The priors of the bivariate Normal hyperparameters (emission) for model M are assumed to be uniform and they are as expressed as,

$$\begin{aligned}
\mu_* &\sim U(a_{\mu_*}, b_{\mu_*}) \\
\sigma_*^2 &\sim U(a_{\sigma_*^2}, b_{\sigma_*^2}) \\
\rho_* &\sim U(a_{\rho_*}, b_{\rho_*}) \\
\mu_p &\sim U(a_{\mu_p}, b_{\mu_p}) \\
\mu_s &\sim U(a_{\mu_s}, b_{\mu_s}) \\
\Sigma_2 &\sim IW(\nu_0, \Omega_0^{-1})
\end{aligned} \tag{5.22}$$

where $U(a, b)$ is the *Uniform* distribution with density on $(a,\ b)$ $f(y|a, b) \propto \frac{1}{(b-a)}$, for $a \leq y \leq b$ and $IW(\nu_0, \Omega_0^{-1})$ is the bivariate Inverse-Wishart distribution with density as

$$f(\mathbf{Y}|\nu_0, \Omega_0^{-1}) \propto |\mathbf{Y}|^{-\frac{\nu_0+2+1}{2}} \exp\left\{-\frac{1}{2} tr\left(\Omega_0 \mathbf{Y}^{-1}\right)\right\},$$

such that $\mathbf{Y}$ is a 2×2 matrix and the elements of Ω_0 (which is also a 2×2 matrix) and ν_0 are fixed constants.

5.1.7 Joint posterior distribution

The joint posterior distribution for model M is given by,

$$p(\boldsymbol{\zeta}^{(M)}|\mathbf{x}) \propto L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}) p(\boldsymbol{\zeta}^{(M)}), \tag{5.23}$$

up to a normalization constant.

5.2 Parameter and state estimation

In this section, I implement an MCMC-based algorithm to estimate the hidden states and parameters analogous to that in Section 4.2. The priors for the logit auxiliary parameters described in Section 5.1.5 are not conjugate which makes the MCMC algorithm more computationally intensive.

The augmented Gibbs sampler that I develop in this chapter sequentially updates the values of auxiliary parameters, then the global emission hyperparameters conditional on the data (auxiliary parameters), transition parameters and the hidden states. The samples of the auxiliary emission parameters are simulated from their conditional distributions using M-H (within Gibbs) samplers as no closed form can be obtained from the conditional posterior distributions of the auxiliary emission parameters. The samples of the global emission hyperparameters are then sampled using a mix of direct samplers and one M-H (within Gibbs) step. The updating scheme of the remaining parameters, i.e., the states $\mathbf{Z}$ and the transition parameters $\boldsymbol{\tau}^{(M)}$ for model M remain the same as in Chapter 4.

5.2.1 Outline of the augmented Gibbs algorithm

In this section, I outline the steps of the augmented Gibbs-M-H sampling scheme for one iteration implemented to sample from the posterior distributions of the HMM model parameters $\boldsymbol{\zeta}^{(M)}$ for model M.

1. I calculate the full likelihood of model M conditional on the HMM model parameter $\boldsymbol{\zeta}^{(M)}$ using the forward sum recursion. The details of the forward sum recursion procedure have been described in Section 2.2.2. In my model M, I can re-construct the forward probability as

$$\alpha_k^{(M)}(t) = P\left(\mathbf{x}_{1:t}; Z_t = k|\boldsymbol{\zeta}^{(M)}\right), \tag{5.24}$$

where $k = 1,\ 2$ denotes the similarly methylated state and differentially methylated state, respectively. The quantity $\alpha_k^{(M)}(t)$ can also be viewed as the partial likelihood up to genomic position t, such that genomic position t is in state k for $t = 1, \ldots, T$ and $k = 1,\ 2$ which can be written as

$$\alpha_k^{(M)}(t) = \sum_{Z_1,\ldots,Z_t} \pi_{Z_1}^{(M)} P_{Z_1}\left(x_1|\boldsymbol{\eta}_{Z_1}^{(M)}(1)\right) \prod_{s=2}^{t} \tau_{Z_{(s-1)},Z_s}^{(M)}(s) P_{Z_s}\left(x_s|\boldsymbol{\eta}_{Z_s}^{(M)}(s)\right). \tag{5.25}$$

Using the forward sum recursion, the partial state based likelihood is given by

$$\alpha_k^{(M)}(t) = b_k^{(M)}(t) \sum_{l=1}^{2} \alpha_l(t-1)\tau_{kl}^{(M)}(t), \quad t = 2, \ldots, T. \tag{5.26}$$

Here, $b_k^{(M)}(t) = P_k\left(\mathbf{x}_t|\boldsymbol{\eta}_k^{(M)}(t)\right)$. I have already derived expressions for $P_k(\mathbf{x}_t|\boldsymbol{\eta}_t^{(M)})$ in (5.14) and (5.15). For $t = 1$, I can write

$$\alpha_{M_k}(1) = \pi_k b_k(1). \tag{5.27}$$

The full likelihood of the entire sequence can be expressed as,

$$L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}) = \sum_{k=1}^{2} \alpha_k^{(M)}(T), \tag{5.28}$$

where $L(\boldsymbol{\zeta}^{(M)})$ is the full likelihood for model M.

2. After computing the state-based partial likelihoods and the full likelihood using forward sum recursion, I employ a backward sampling procedure to sample the hidden states $\mathbf{Z}$. The steps of the backward sampling have been described in detail before (step (2) of Section 4.2.1).

3. Next, I update the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model para-

meters $\boldsymbol{\tau}^{(M)}$, conditional on the current values of the emission hyperparameters $\boldsymbol{\eta}^{(M)}$, the sequence of the hidden states $\mathbf{Z}$ and the observed methylation data $\mathbf{x}$. Again, the steps have been described in detail before (step (3) of Section 4.2.1).

4. For model M, the auxiliary emission parameters $\boldsymbol{\eta}^{(M)}$, conditional on the current values of the global emission model parameters $\boldsymbol{\theta}^{(M)}$, the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameters $\boldsymbol{\tau}^{(M)}$, the sequence of the hidden states $\mathbf{Z}$ and the observed methylation data $\mathbf{x}$, can be updated using a M-H procedure.

5. For model M, the global emission model parameters $\boldsymbol{\theta}^{(M)}$, conditional on the current values of the auxiliary emission parameters $\boldsymbol{\eta}^{(M)}$, the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameters $\boldsymbol{\tau}^{(M)}$, and the hidden states $\mathbf{Z}$ and the observed methylation data $\mathbf{x}$ can be updated using a mix of Gibbs sampler and M-H sampling.

5.2.2 Further details of the augmented Gibbs sampler

The key steps of the augmented Gibbs sampler are as follows:

1. I sample the hidden state path $\mathbf{Z}$ from the full conditional posterior distribution $p(\mathbf{Z}|\mathbf{x}, \boldsymbol{\zeta}^{(M)})$ given $\boldsymbol{\zeta}^{(M)} = (\boldsymbol{\theta}^{(M)},\ \boldsymbol{\pi}^{(M)},\ \boldsymbol{\tau}^{(M)})$ and observed methylation data $\mathbf{x}$. For this step, I employ the data-augmentation based FSBS procedure as described in Section 5.2.1.

2. I sample the auxiliary emission parameters $\boldsymbol{\eta}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\theta}^{(M)})$ given the global emission model parameter $\boldsymbol{\theta}^{(M)}$, the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameter $\boldsymbol{\tau}^{(M)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. However, in this step, I sample $\boldsymbol{\eta}^{(M)}$ from the full conditional distribution $p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(M)})$ using a M-H algorithm given the updated hidden

states $\mathbf{Z}$, observed methylation data $\mathbf{x}$ and the global emission model parameter $\boldsymbol{\theta}^{(M)}$, since,

$$p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\theta}^{(M)}) = p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(M)}). \tag{5.29}$$

3. I sample the global emission hyperparameters $\boldsymbol{\theta}^{(M)}$ from the conditional posterior distribution $p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\eta}^{(M)})$ given the auxiliary emission model parameter $\boldsymbol{\eta}^{(M)}$, the initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition model parameter $\boldsymbol{\tau}^{(M)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. In this step, it is enough to sample $\boldsymbol{\theta}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\theta}^{(M)}|\mathbf{Z}, \boldsymbol{\eta}^{(M)})$ using a M-H algorithm given the updated hidden states $\mathbf{Z}$ and the auxiliary emission model parameter $\boldsymbol{\eta}^{(M)}$, since,

$$p(\boldsymbol{\theta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\eta}^{(M)}) = p(\boldsymbol{\theta}^{(M)}|\mathbf{Z}, \boldsymbol{\eta}^{(M)}). \tag{5.30}$$

4. In this step, I sample the HMM model initial state parameters $\boldsymbol{\pi}^{(M)}$ and transition parameters $\boldsymbol{\tau}^{(M)}$ from the full conditional posterior distribution $p(\boldsymbol{\pi}^{(M)}, \boldsymbol{\tau}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\eta}^{(M)}, \boldsymbol{\theta}^{(M)})$ given the auxiliary emission parameter $\boldsymbol{\eta}^{(M)}$, global emission model hyperparameter $\boldsymbol{\theta}^{(M)}$, updated hidden states $\mathbf{Z}$ and observed methylation data $\mathbf{x}$. Again, the steps have been described in detail before (step (3) of Section 4.2.2).

I now describe the sampling steps of the auxiliary emission parameters (2), global emission hyperparameters (3) for both the models $NLBDM$ and $NLBCM$. I have already explained the sampling steps of the initial state and transition parameters (4.(a), (b)) for both the models analogous to Section 4.2.2 (3.(a), (b)) of Chapter 4 in Sections 4.2.3.2 and 4.2.3.3, respectively.

5.2.3 Sampling steps from conditional posterior distributions

5.2.3.1 Auxiliary emission parameters

In this section, I elaborate on the sampling steps of the auxiliary emission parameters from their full conditional distributions. I first write the full conditional distribution of the HMM model auxiliary emission parameters $\boldsymbol{\eta}^{(M)}$,

$$\begin{aligned} p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\tau}^{(M)}, \boldsymbol{\theta}^{(M)}) &= p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(M)}) \\ &\propto L_{\mathbf{x},\mathbf{Z}}(\boldsymbol{\eta}^{(M)}) p(\boldsymbol{\eta}^{(M)}|\boldsymbol{\theta}^{(M)}). \end{aligned} \tag{5.31}$$

(5.31) is proportional to the complete data likelihood $L_{\mathbf{x},\mathbf{Z}}(\boldsymbol{\eta}^{(M)})$ times the second-stage conditional prior $p(\boldsymbol{\eta}^{(M)}|\boldsymbol{\theta}^{(M)})$ for the auxiliary emission parameter.

Again, (5.31) can be further simplified as the product of full conditional distributions of bivariate auxiliary emission parameters for model M,

$$p(\boldsymbol{\eta}^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}^{(M)}) = p(\boldsymbol{\eta}_1^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}_1^{(M)}) p(\boldsymbol{\eta}_2^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}_2^{(M)}). \tag{5.32}$$

Since, $\boldsymbol{\eta}_1^{(M)}$ are state 1 auxiliary emission parameters and $\boldsymbol{\eta}_2^{(M)}$ are state 2 auxiliary emission parameters, I can further re-write (5.32) as,

$$p(\boldsymbol{\eta}_k^{(M)}|\mathbf{x}, \mathbf{Z}, \boldsymbol{\theta}_k^{(M)}) = \prod_{t=1}^{T} p(q_t^{pk}, q_t^{sk}|\mathbf{x}_t, \mathbf{Z}_t, \boldsymbol{\theta}_k^{(M)}). \tag{5.33}$$

Now, the conditional posterior distribution of the state-specific $\boldsymbol{\eta}_k^{(M)}(t)$ given $Z_t = k$ can be written as the product of single-point data likelihood and the conditional prior for the auxiliary emission parameter at the t^{th} CpG site,

$$p(\boldsymbol{\eta}_k^{(M)}(t)|\mathbf{x}_t, \mathbf{Z}_t = k, \boldsymbol{\theta}_k^{(M)}) \propto L(\boldsymbol{\eta}_k^{(M)}(t)|\mathbf{x}_t, \mathbf{Z}_t) p(\boldsymbol{\eta}_k^{(M)}(t)|\boldsymbol{\theta}_k^{(M)}).$$

If $Z_t = k'$, i.e., $\{t : Z_t = k\}$ is an empty set as no observation is associated with the hidden state k, then the conditional posterior distribution of state-specific $\boldsymbol{\eta}_k^{(M)}(t)$ is just proportional to its conditional prior:

$$p(\boldsymbol{\eta}_k^{(M)}(t)|\mathbf{x}_t, \mathbf{Z}_t = k', \boldsymbol{\theta}_k^{(M)}) \propto p(\boldsymbol{\eta}_k^{(M)}(t)|\boldsymbol{\theta}_k^{(M)}). \tag{5.34}$$

5.2.3.2 Global emission hyperparameters

The full conditionals of the global emission hyperparameters can be developed from the Bivariate conditional priors of the auxiliary emission parameters which in this case act as the likelihoods as mentioned in (5.19) due to non-informative prior distributions for model M:

- Sample $\mu_*|\sigma_*^2, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z}$ from $N\left(\frac{\sum_{t=1}^{T}\left(q_t^{p1}+q_t^{s1}\right)\mathbf{I}[Z_t=1]}{2t_1}, \frac{(1+\rho_*)\sigma_*^2}{2t_1}\right)$.

$$\begin{aligned}
p(\mu_*|\sigma_*^2, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z}) &= \prod_{t=1}^{T}\left[\phi(\mathbf{Q}_t^1, \mathbf{M}_1, \boldsymbol{\Sigma}_1)\right]^{\mathbf{I}[Z_t=1]} \times \frac{1}{b_{\mu_*} - a_{\mu_*}} \\
&\propto \prod_{t=1}^{T}\exp\left[-\frac{1}{2}\left((\mathbf{Q}_t^1 - \mathbf{M}_1)^T {\boldsymbol{\Sigma}_1}^{-1}(\mathbf{Q}_t^1 - \mathbf{M}_1)\right)\right]^{\mathbf{I}[Z_t=1]} \\
&= \exp\left[-\frac{1}{2\sigma_*^2(1-\rho_*^2)}\sum_{t=1}^{T}\left\{\left(q_t^{p1} - \mu_*\right)^2\right.\right. \\
&\quad \left.\left. -2\rho_*\left(q_t^{p1} - \mu_*\right)\left(q_t^{s1} - \mu_*\right) + \left(q_t^{s1} - \mu_*\right)^2\right\}\right]\mathbf{I}[Z_t = 1] \\
&\propto \exp\left[-\frac{1}{2\sigma_*^2(1-\rho_*^2)}\sum_{t=1}^{T}\left(\mu_*^2(2-2\rho_*)\right.\right. \\
&\quad \left.\left. -2\mu_*\left[\left(q_t^{p1} + q_t^{s1}\right)(1-\rho_*)\right]\right)\right]\mathbf{I}[Z_t = 1] \\
&\propto \exp\left[-\frac{2t_1}{2\sigma_*^2(1+\rho_*)}\left(\mu_* - \frac{\sum_{t=1}^{T}\left(q_t^{p1} + q_t^{s1}\right)\mathbf{I}[Z_t = 1]}{2t_1}\right)\right].
\end{aligned}$$

i.e.,

$$\mu_*|\sigma_*^2, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \sim N\left(\frac{\sum_{t=1}^{T}\left(q_t^{p1}+q_t^{s1}\right)\mathbf{I}[Z_t=1]}{2t_1}, \frac{(1+\rho_*)\,\sigma_*^2}{2t_1}\right). \quad (5.35)$$

- Sample $\sigma_*^2|\mu_*, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z}$ from $IG\left[t_1-1, \frac{\sum_{t=1}^{T}\left\{\left(q_t^{p1}-\mu_*\right)^2-2\rho_*\left(q_t^{p1}-\mu_*\right)\left(q_t^{s1}-\mu_*\right)+\left(q_t^{s1}-\mu_*\right)^2\right\}\mathbf{I}[Z_t=1]}{2(1-\rho_*^2)}\right]$.

$$\begin{aligned} p(\sigma_*^2|\mu_*, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z}) &= \prod_{t=1}^{T}\left[\phi(\mathbf{Q}_t^1, \mathbf{M}_1, \boldsymbol{\Sigma}_1)\right]^{\mathbf{I}[Z_t=1]} \times \frac{1}{b_{\sigma_*^2}-a_{\sigma_*^2}} \\ &\propto \sigma_*^{2-(t_1-1)-1}\exp\Bigg[-\frac{1}{2\sigma_*^2(1-\rho_*^2)}\sum_{t=1}^{T}\Big\{\left(q_t^{p1}-\mu_*\right)^2 \\ &\quad -2\rho_*\left(q_t^{p1}-\mu_*\right)\left(q_t^{s1}-\mu_*\right)+\left(q_t^{s1}-\mu_*\right)^2\Big\}\mathbf{I}[Z_t=1]\Bigg]. \end{aligned}$$

So,

$$\sigma_*^2|\mu_*, \rho_*, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \sim IG\left[t_1-1, \frac{\sum_{t=1}^{T}\left\{\left(q_t^{p1}-\mu_*\right)^2-2\rho_*\left(q_t^{p1}-\mu_*\right)\left(q_t^{s1}-\mu_*\right)+\left(q_t^{s1}-\mu_*\right)^2\right\}\mathbf{I}[Z_t=1]}{2(1-\rho_*^2)}\right].$$

- Sample $\rho_*|\mu_*, \sigma_*^2, \boldsymbol{\eta}^{(M)}, \mathbf{Z}$ from

$$p(\rho_*|\mu_*, \sigma_*^2, \boldsymbol{\eta}^{(M)}, \mathbf{Z}) = \prod_{t=1}^{T}\left[\phi(\mathbf{Q}_t^1, \mathbf{M}_1, \boldsymbol{\Sigma}_1)\right]^{\mathbf{I}[Z_t=1]} \times \frac{1}{b_{\rho_*}-a_{\rho_*}}.$$

- Sample $\mathbf{M}_2 \left| \boldsymbol{\Sigma}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \right.$ from $BVN \left[\begin{pmatrix} \frac{\sum_{t=1}^{T} q_t^{p2} \mathbf{I}[Z_t=2]}{t_2} \\ \\ \frac{\sum_{t=1}^{T} q_t^{s2} \mathbf{I}[Z_t=2]}{t_2} \end{pmatrix}, \begin{pmatrix} \frac{\sigma_p^2}{t_2} & \frac{\sigma_p \sigma_s \rho_2}{t_2} \\ \\ \frac{\sigma_p \sigma_s \rho_2}{t_2} & \frac{\sigma_s^2}{t_2} \end{pmatrix} \right]$.

$$
\begin{aligned}
p(\mathbf{M}_2 \left| \sigma_p^2, \boldsymbol{\Sigma}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \right.) &= \prod_{t=1}^{T} \left[\phi(\mathbf{Q}_t^2, \mathbf{M}_2, \boldsymbol{\Sigma}_2) \right]^{\mathbf{I}[Z_t=2]} \times \frac{1}{b_{\mu_p} - a_{\mu_p}} \times \frac{1}{b_{\mu_s} - a_{\mu_s}} \\
&\propto \prod_{t=1}^{T} \exp \left[-\frac{1}{2} \left(\left(\mathbf{Q}_t^2 - \mathbf{M}_2 \right)^T {\boldsymbol{\Sigma}_2}^{-1} \left(\mathbf{Q}_t^2 - \mathbf{M}_2 \right) \right) \right]^{\mathbf{I}[Z_t=2]} \\
&= \exp \left[-\frac{1}{2} \sum_{t=1}^{T} \left(\left(\mathbf{Q}_t^2 - \mathbf{M}_2 \right)^T {\boldsymbol{\Sigma}_2}^{-1} \left(\mathbf{Q}_t^2 - \mathbf{M}_2 \right) \right) \right] \mathbf{I}[Z_t = 2] \\
&\propto \exp \left[-\frac{1}{2} \left(-2 \mathbf{M}_2 {\boldsymbol{\Sigma}_2}^{-1} t_2 \overline{\mathbf{Q}^2} + t_2 {\mathbf{M}_2}^T {\boldsymbol{\Sigma}_2}^{-1} \mathbf{M}_2 \right) \right] \\
&\propto \exp \left[-\frac{t_2}{2} \left({\mathbf{M}_2}^T {\boldsymbol{\Sigma}_2}^{-1} \mathbf{M}_2 - 2 \mathbf{M}_2 {\boldsymbol{\Sigma}_2}^{-1} \overline{\mathbf{Q}^2} + \overline{\mathbf{Q}^2}^T {\boldsymbol{\Sigma}_2}^{-1} \overline{\mathbf{Q}^2} \right) \right] \\
&\propto \exp \left[-\frac{t_2}{2} \left(\left(\mathbf{M}_2 - \overline{\mathbf{Q}^2} \right)^T {\boldsymbol{\Sigma}_2}^{-1} \left(\mathbf{M}_2 - \overline{\mathbf{Q}^2} \right) \right) \right]
\end{aligned}
$$

i.e.,

$$
\mathbf{M}_2 \left| \boldsymbol{\Sigma}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \right. \sim BVN \left[\begin{pmatrix} \frac{\sum_{t=1}^{T} q_t^{p2} \mathbf{I}[Z_t=2]}{t_2} \\ \\ \frac{\sum_{t=1}^{T} q_t^{s2} \mathbf{I}[Z_t=2]}{t_2} \end{pmatrix}, \begin{pmatrix} \frac{\sigma_p^2}{t_2} & \frac{\sigma_p \sigma_s \rho_2}{t_2} \\ \\ \frac{\sigma_p \sigma_s \rho_2}{t_2} & \frac{\sigma_s^2}{t_2} \end{pmatrix} \right], \tag{5.36}
$$

where $\overline{\mathbf{Q}^2} = \left[\frac{1}{t_2} \left(\sum_{t=1}^{T} q_t^{p2} \mathbf{I}[Z_t = 2],\ \sum_{t=1}^{T} q_t^{s2} \mathbf{I}[Z_t = 2] \right) \right]^T$.

- Sample $\Sigma_2|\mathbf{M}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z}$ from $IW\left(\nu_0 + t_2, \left[\Omega_0 + \left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)\left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)^T\right]^{-1}\right)$.

$$
\begin{aligned}
p(\boldsymbol{\Sigma}_2|\mathbf{M}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z}) &= \prod_{t=1}^{T}\left[\phi(\mathbf{Q}_t^2, \mathbf{M}_2, \boldsymbol{\Sigma}_2)\right]^{\mathbf{I}[Z_t=2]} \times p(\boldsymbol{\Sigma}_2) \\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{t_2}{2}} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\left(\left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)^T {\boldsymbol{\Sigma}_2}^{-1}\left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)\right)\right]\mathbf{I}[Z_t = 2] \\
&\quad \times |\boldsymbol{\Sigma}_2|^{-\frac{(\nu_0+2+1)}{2}} \exp\left\{-\frac{1}{2} tr\left(\Omega_0 {\Sigma_2}^{-1}\right)\right\} \\
&\propto |\boldsymbol{\Sigma}_2|^{-\frac{(\nu_0+t_2+2+1)}{2}} \\
&\quad \times \exp\left\{-\frac{1}{2} tr\left(\left[\Omega_0 + \left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)\left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)^T\right]{\boldsymbol{\Sigma}_2}^{-1}\right)\right\}\mathbf{I}[Z_t = 2]
\end{aligned}
$$

i.e.,

$$
\Sigma_2|\mathbf{M}_2, \boldsymbol{\eta}^{(M)}, \mathbf{Z} \sim IW\left(\nu_0 + t_2, \left[\Omega_0 + \left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)\left(\mathbf{Q}_t^2 - \mathbf{M}_2\right)^T\right]^{-1}\right). \tag{5.37}
$$

5.2.4 Summary of the Augmented Gibbs sampler algorithm steps

1. Initialize all auxiliary emission parameters ($\boldsymbol{\eta}^{(M)}$), hyperparameters ($\boldsymbol{\theta}^{(M)}$) for model M.
   (a) For model *NLBDM*,
   initialize all transition parameters $\boldsymbol{\tau}^{(D)} = (\pi_1, \tau_{11}, \tau_{21})$ and,
   (b) For model *NLBCM*,
   initialize all transition rate parameters $\boldsymbol{\tau}^{(C)} = (\lambda_1, \lambda_2)$.
2. Compute the state-specific emission distributions,
   $P(\mathbf{x}_t|Z_t = k, \boldsymbol{\eta}_k^{(M)})$ for $k = 1,\ 2$ and $t = 1, \ldots, T$.
3. Compute $\alpha_k^{(M)}(t)$ for $k = 1,\ 2$ and $t = 1, \ldots, T$.

4. Sample backwards $Z_T, \ldots, Z_1$ using backward sampling (Scott, 2002).

5. Sample $(q_t^{p1}, q_t^{s1}, q_t^{p2}, q_t^{s2})$ using M-H algorithm as explained in Section 5.2.3.1.

6. Sample (μ_*, σ_*^2) and $(\mu_p, \mu_s, \sigma_p^2, \sigma_s^2, \rho_2)$ using direct sampler and ρ_* using M-H sampler.

7. For model M, sample the transition parameters as described in step (7) of Section 4.2.4.

8. Implement the relabelling algorithm as described in Section 2.2.6.

9. Repeat steps (2)-(8) until convergence.

5.3 Simulation study

In this section, I describe the simulation study design, which plays the same role for the models of this chapter as that in Section 4.3 did for the models in Chapter 4.

5.3.1 Data generation

100 datasets were generated with $T = 10000$ observations each under different situations to check the robustness of my models. The data generations were done for 3 cases in each of the model as described in detail before (Section 4.3.1).

1. **Moderately overlapped**

   (a) For model *NLBDM*, the data $(\mathbf{x}^p, \mathbf{x}^s)$ are generated from the Normal-logit-Binomial HMM with exactly same modes for data of both the states, i.e., state 1 hyperparameters $(\mu_* = 0.2, \sigma_*^2 = 1.2, \rho_* = 0.7)$ and state 2 hyperparameters $(\mu_p = 0.2, \mu_s = 0.2, \sigma_p^2 = 1.2, \sigma_s^2 = 1.2, \rho_2 = 0.7)$. The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order Markov Chain

with an initial state probability for state 1, $\pi_1 = 0.34$, and transition probabilities $\tau_{11} = 0.87$, $\tau_{21} = 0.068$, as before.

(b) For model *NLBCM*, the data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated as for *NLBDM* except that the hidden states $\mathbf{Z}$ are simulated from a 1^{st} order continuous-index Markov Chain with transition rate parameters $\lambda_1 = 0.27$ and $\lambda_2 = 0.27$.

2. **Well separated**

(a) For model *NLBDM*, the data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated from the Normal-logit-Binomial HMM with well-separated modes for data of both the states, i.e., state 1 hyperparameters $(\mu_* = -2.95, \sigma_*^2 = 0.7, \rho_* = 0.65)$ and state 2 hyperparameters $(\mu_p = 2.3, \mu_s = 3.2, \sigma_p^2 = 0.85, \sigma_s^2 = 1.2, \rho_2 = 0.75)$. The hidden states $\mathbf{Z}$ are simulated as for the *moderately overlapped NLBDM* case.

(b) For model *NLBCM*, the data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated as for *NLBDM* with the hidden states $\mathbf{Z}$ simulated as for the *moderately overlapped NLBCM* case.

3. **Realistic**

(a) For model *NLBDM*, the data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated from the Normal-logit-Binomial HMM with less well-separated modes for data of both the states comparable to the real data, i.e., state 1 hyperparameters $(\mu_* = 0.326, \sigma_*^2 = 1.806, \rho_* = 0.964)$ and state 2 hyperparameters $(\mu_p = -0.676, \mu_s = -1.65, \sigma_p^2 = 1.77, \sigma_s^2 = 2.364, \rho_2 = 0.97)$, thus causing some amount of overlapping. The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order Markov Chain with an initial state probability for state-1 $\pi_1 = 0.39$ and transition probabilities $\tau_{11} = 0.97$, $\tau_{21} = 0.02$.

(b) For model *NLBCM*, the data $(\mathbf{x}^p,\ \mathbf{x}^s)$ are generated from the Normal-logit-Binomial HMM with less well-separated modes for data of both

the states comparable to the real data, i.e., state 1 hyperparameters ($\mu_* = 2.77, \sigma_*^2 = 1.62, \rho_* = 0.73$) and state 2 hyperparameters ($\mu_p = -0.66, \mu_s = -1.58, \sigma_p^2 = 3.64, \sigma_s^2 = 4.71, \rho_2 = 0.86$). The hidden states $\mathbf{Z}$ are simulated from a 1^{st} order continuous-index Markov Chain with transition rate parameters $\lambda_1 = 0.20$ and $\lambda_2 = 0.128$.

5.3.2 Priors for the global emission hyperparameters

I have already observed that the full conditionals of global emission hyperparameters become independent of the prior choices for these hyperparameters as they were chosen to uninformative except for one global emission hyperparameter for state 1, i.e., ρ_*.

$$\rho_* \sim U(-1, 1) \tag{5.38}$$

5.3.3 Consistency of model parameters estimation

I generated 100 datasets under models *NLBDM* and *NLBCM*. These datasets of size $10,000$ CpG sites were generated for each parameter setting described in Section 5.3.1 and fitted using the augmented Gibbs sampler described in Section 5.2. Each simulated dataset was fitted to the models *NLBDM* and *NLBCM* for each case with $60,000$ MCMC iterations (with $20,000$ as burn-in) after which the posterior samples for each model parameter were assessed for convergence. In Table 5.1, I present the results of the range of estimated RMSE of the model parameters for each case. In each case, the estimated RMSE was small for the *well separated* and *realistic* cases and much larger for the *moderately overlapped* case, comparable to the corresponding results for *BBDM* and *BBCM* in Section 4.3.3.

Model	Case	Average Misclass. rate	Range of RMSE
NLBDM	*Moderately overlapped*	0.6617	(0.03, 1.026)
	Well separated	0.0014	(0.0001, 0.005)
	Realistic	0.0255	(0.0005, 0.0098)
NLBCM	*Moderately overlapped*	0.5091	(0.08, 1.66)
	Well separated	0.0043	(0.0004, 0.009)
	Realistic	0.0963	(0.0006, 0.009)

Table 5.1: Simulation study: Average misclassification rate and range of RMSE for models: NLBDM and NLBCM based on 100 simulated datasets.

5.4 Real data study

In this section, I fit the two models *NLBDM* and *NLBCM* discussed in this chapter to the same real data from chromosome 16 as in Chapter 4.

5.4.1 Inference via MCMC

I examined the results obtained using MCMC techniques and the convergence properties of the estimates using various diagnostics for the real data Chromosome 16 described in Section 4.4.1. The proposal distribution of the ρ_* for both the models *NLBDM* and *NLBCM* was tuned appropriately in order to obtain optimal acceptance rates. The acceptance rates of ρ_* for models *NLBDM* and *NLBCM* were 0.22 and 0.37, respectively. All necessary convergence diagnostics were carried out and no evidence of non-convergence was obtained from any of the diagnostics. I present the traceplots and PSRF plots in Appendix 7.2.3. In addition, I also present the estimates of the posterior mean, S.D., 95% credible intervals in Tables 4 and 5 for emission hyperparameters and Table 6 for transition

parameters of both the models for all the 3 chains.

5.5 Comparison with Chapter 4

In this section, I have compared the simulation and real data results of this chapter with Chapter 4.

In the *moderately overlapped* case, the average misclassification rates for *NLBDM* and *NLBCM* are 0.6617 and 0.5091, respectively (Table 5.1). The corresponding misclassification rates for the *realistic* case are 0.0255 and 0.0969, respectively, whereas the misclassfication rates for the *well separated* case in both models are much smaller than the *realistic* case (Table 5.1). For one of the randomly selected simulation studies out of 100, I also present the scatter plots (Figures 5.2 5.3 5.4) of the methylation proportions between the two cell types for all the cases classified by the true states and predicted states analogous to the corresponding Figures 4.2, 4.3, 4.4 for models *BBDM* and *BBCM* described in Section 4.3.3. These scatter plots (Figures 5.2 5.3 5.4) provide some improvement in the correlation between the simulated methylation proportions between proliferating and senescent cells. In addition, I also plot the histograms (Figures 5.5a, 5.5b, 5.6a, 5.6b, 5.7a, 5.7b) for all the cases. The histograms (Figures 5.5a, 5.5b) for the *moderately overlapped* case in both the models are symmetric whereas the histograms of the *well separated* (Figures 5.6a, 5.6b) and *realistic* (Figures 5.7a, 5.7b cases are either U-shaped, J-shaped or reflected J-shaped. Furthermore, the ROC curves are plotted for all the cases (Figures 5.8a, 5.8b 5.9a, 5.9b, 5.10a, 5.10b). Figures 5.8a, 5.8b display the weak performance in the *moderately overlapped* case for both the models. The *well separated* (Figures 5.9a, 5.9b) case for both the models beats the *realistic* (Figures 5.10a, 5.10b) case by a narrow margin.

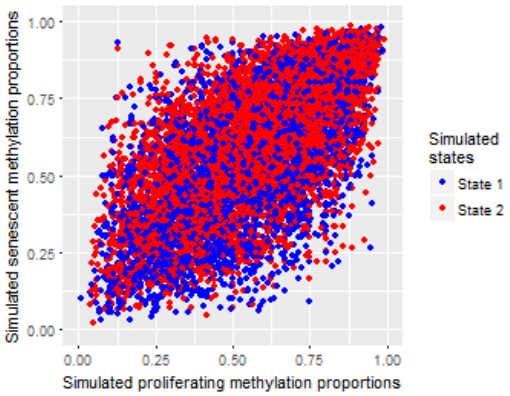
In the real data study, the boxplots (Figures 5.11, 5.13) validating the biolo-

gical presumption, for both models are plotted. These boxplots are analogous to the boxplots in Chapter 4 (Figures 4.11, 4.13) . Additionally, the histograms of state 2 posterior probabilities for model *BBDM* show strong classification of states in Figure 5.12 whereas the classification of the states is moderately poor in the case of model *BBCM* as shown in Figure 5.14, comparable to the corresponding histograms (Figures 5.12, 5.14) for *BBDM* and *BBCM* in Section 4.5.

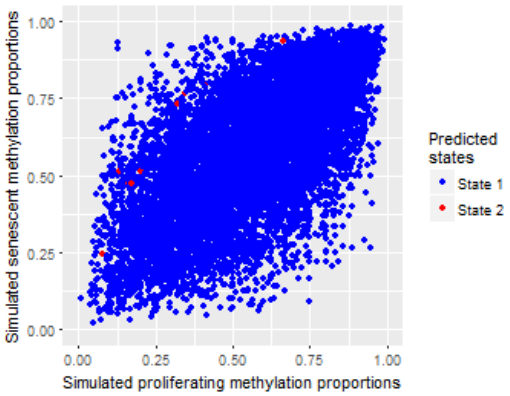
5.6 Summary

In this Chapter, I described my extended HMMmethState method for identifying DMCs from BS-seq methylation data. I implemented both the models *NLBDM* and *NLBCM*, to capture the correlation in the bivariate data between the methylated counts of senescent and proliferating cells. I have extended the original Beta-Binomial emission model described in Chapter 4 to a bivariate Normal-logit emission model, where the underlying bivariate logit parameters at the 2^{nd} stage of the hierarchical model are assumed to be normally distributed at each CpG site and they are clustered around a state-specific mean with a state-specific variance and state-specific correlation between the two cell types.

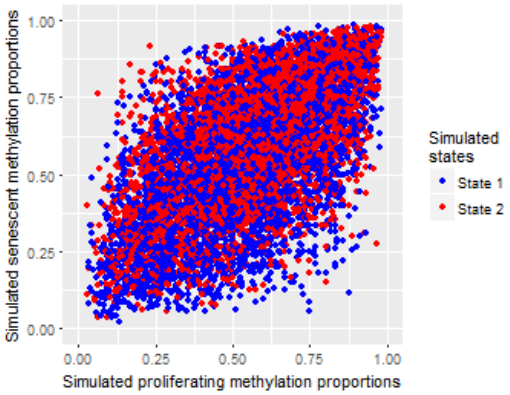
I have also visually illustrated my claim that both the models *NLBDM* and *NLBCM* can capture the correlation between the methylated counts of both the cells. I have presented the scatter plots of the methylation counts of proliferating cells against senescent cells for the observed data and for the fitted models in Figure 5.15. The visual posterior predictive checking using the posterior mean estimates of the parameters of the fitted models in Figure 5.15 clearly indicates that there is a correlation between the methylated counts of proliferating and senescent cells. Figures 5.15a, 5.15c show the scatterplots of the observed data classified by the predicted states. On the other hand, Figures 5.15b, 5.15d show



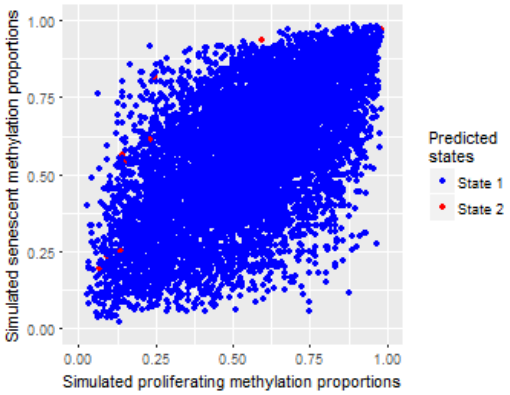
(a) Scatter plot for *NLBDM* classified by simulated states.



(b) Scatter plot for *NLBDM* classified by predicted states.

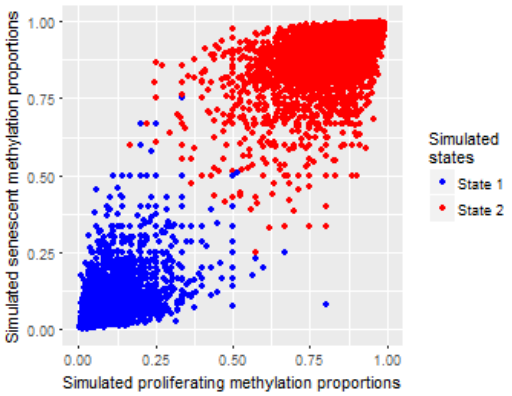


(c) Scatter plot for *NLBCM* classified by simulated states.

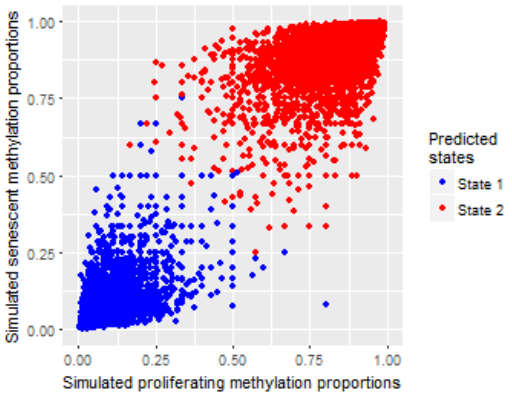


(d) Scatter plot for *NLBCM* classified by predicted states.

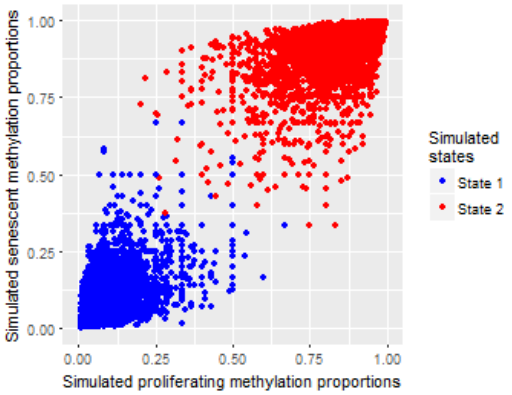
Figure 5.2: For the *moderately overlapped* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBCM*.



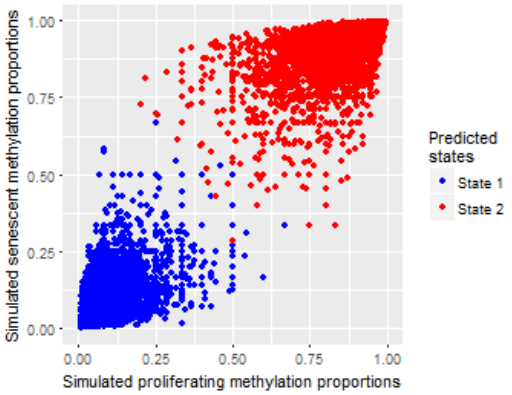
(a) Scatter plot for *NLBDM* classified by simulated states.



(b) Scatter plot for *NLBDM* classified by predicted states.

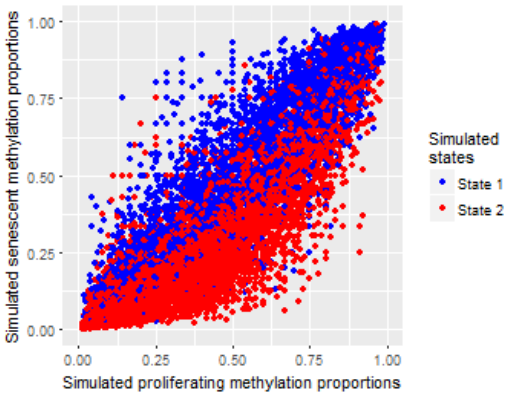


(c) Scatter plot for *NLBCM* classified by simulated states.

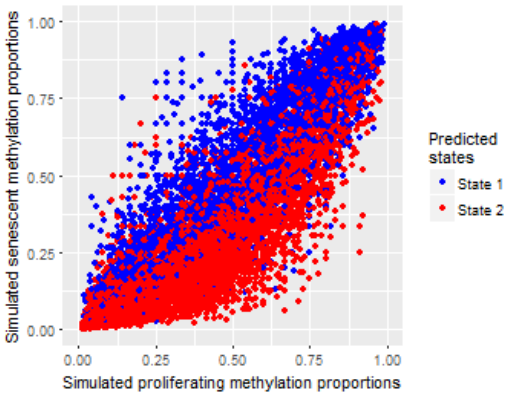


(d) Scatter plot for *NLBCM* classified by predicted states.

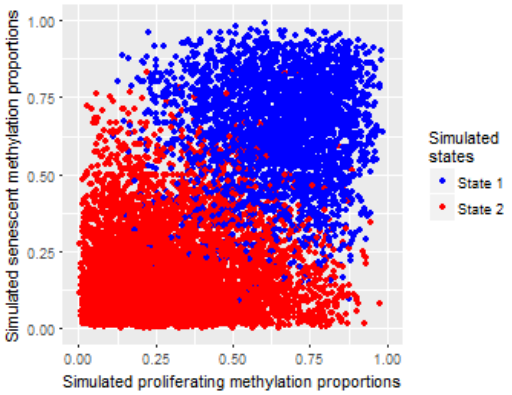
Figure 5.3: For the *well separated* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBCM*.



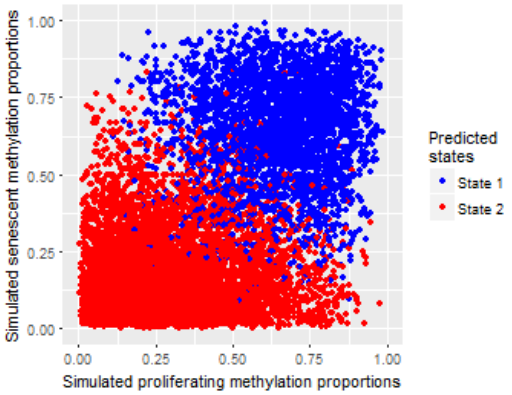
(a) Scatter plot for *NLBDM* classified by simulated states.



(b) Scatter plot for *NLBDM* classified by predicted states.

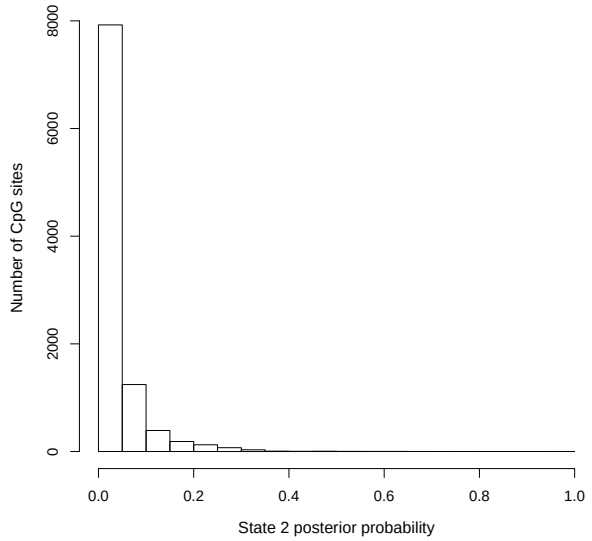


(c) Scatter plot for *NLBCM* classified by simulated states.

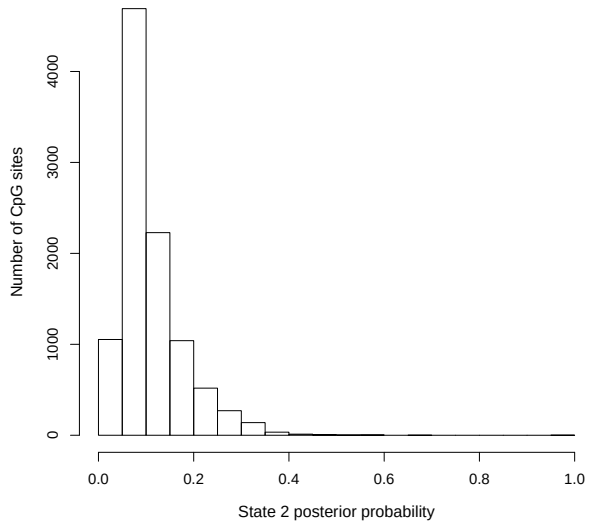


(d) Scatter plot for *NLBCM* classified by predicted states.

Figure 5.4: For the *realistic* case. (a) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBDM*. (b) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBDM*. (c) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the simulated states for *NLBCM*. (d) A scatter plot of simulated methylation proportions between proliferating and senescent cells classified by the predicted states for *NLBCM*.

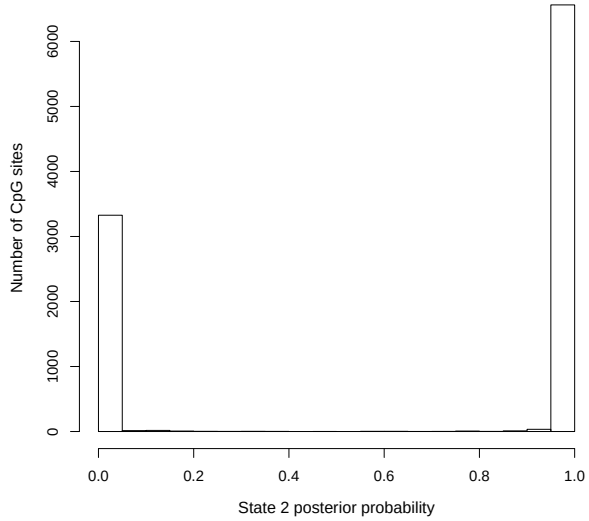


(a) Histogram for *BBDM*

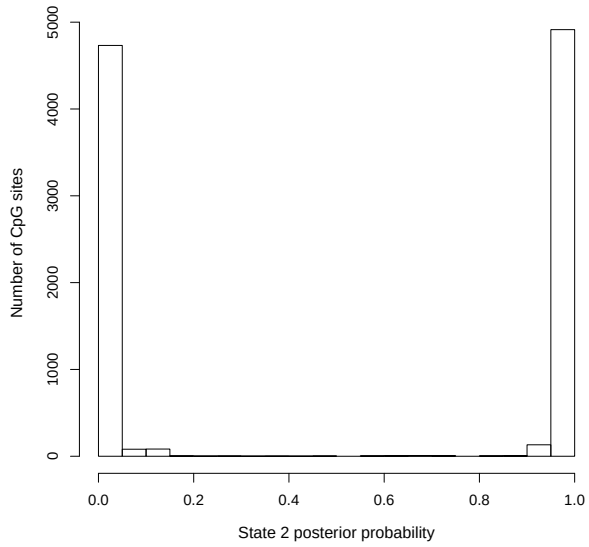


(b) Histogram for *BBCM*

Figure 5.5: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *moderately overlapped* case: (a) *NLBDM* and (b) *NLBCM* based on one randomly selected simulation.

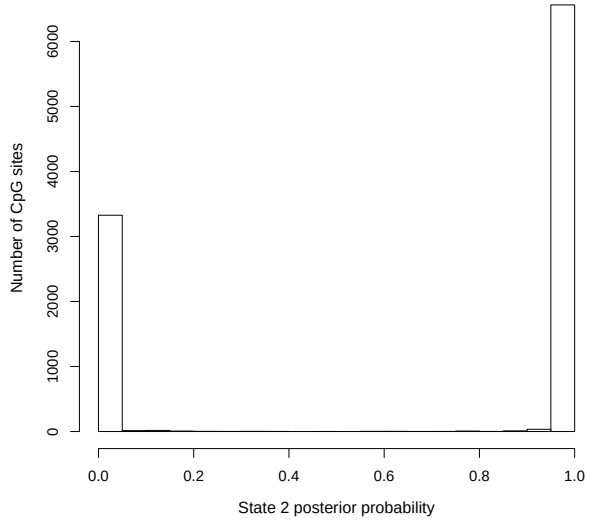


(a) Histogram for *BBDM*

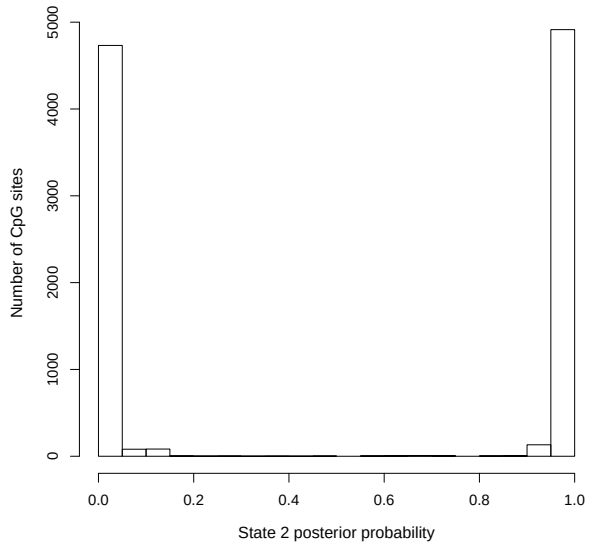


(b) Histogram for *BBCM*

Figure 5.6: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *well separated* case: (a) *NLBDM* and (b) *NLBCM* based on one randomly selected simulation.

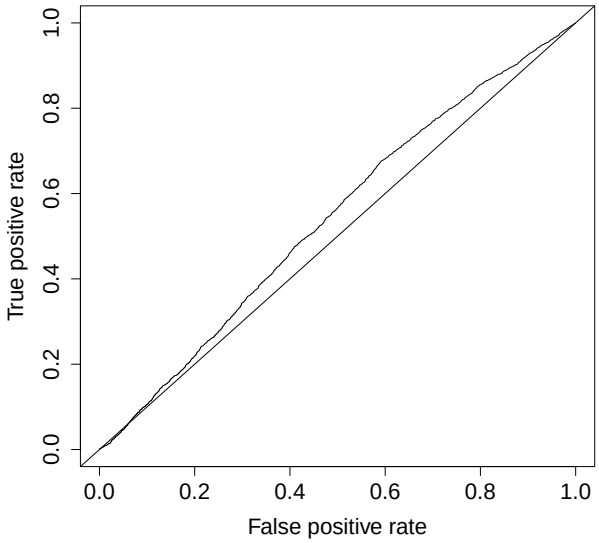


(a) Histogram for *BBDM*

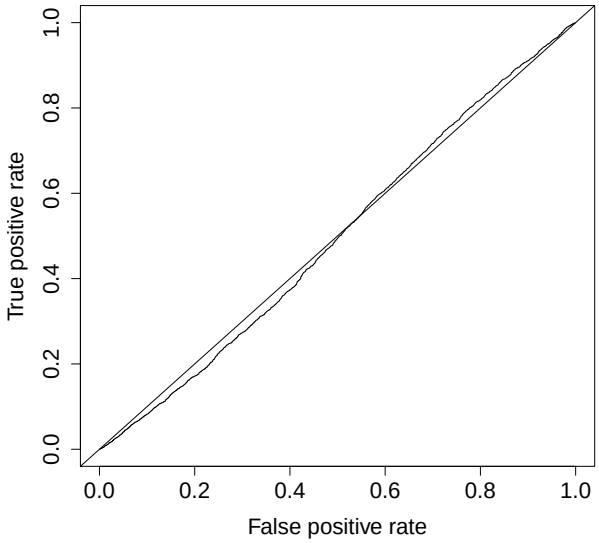


(b) Histogram for *BBCM*

Figure 5.7: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the histogram of posterior state 2 probabilities for the *realistic* case: (a) *NLBDM* and (b) *NLBCM* based on one randomly selected simulation.

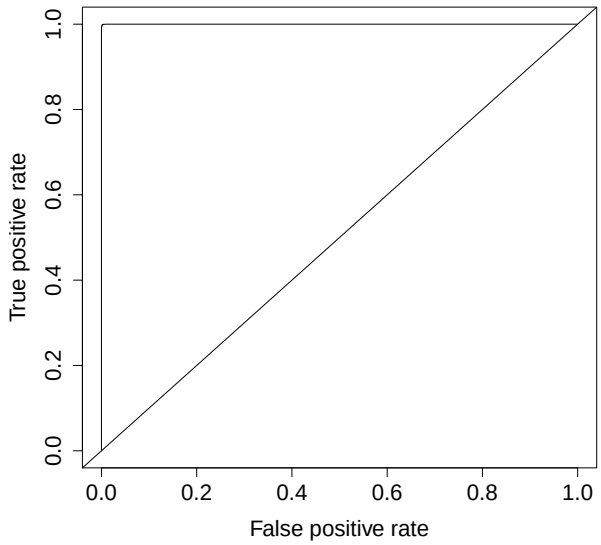


(a) ROC curve for *NLBDM*

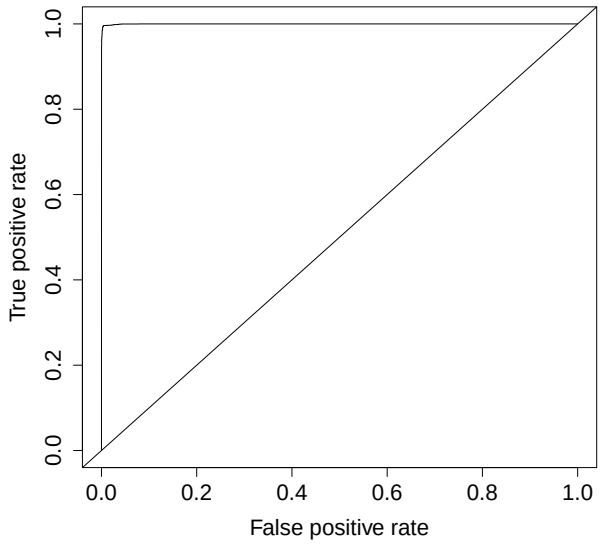


(b) ROC curve for *NLBCM*

Figure 5.8: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the ROC curves for the *moderately overlapped* case: (a) *NLBDM* and (b) *NLBCM*.

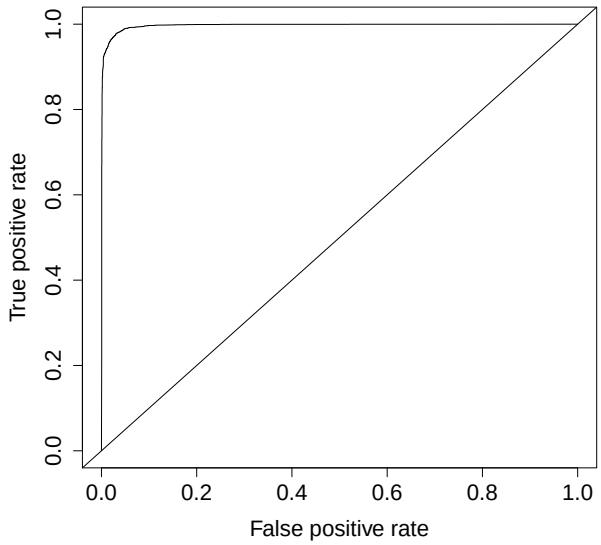


(a) ROC curve for *NLBDM*

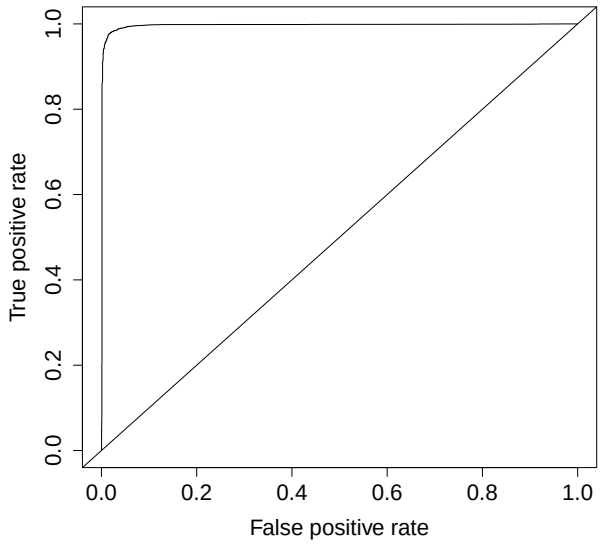


(b) ROC curve for *NLBCM*

Figure 5.9: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the ROC curves for the *well separated* case: (a) *NLBDM* and (b) *NLBCM*.



(a) ROC curve for *NLBDM*



(b) ROC curve for *NLBCM*

Figure 5.10: For the simulation study of *NLBDM* and *NLBCM*, the 2 panels depict the ROC curves for the *realistic* case: (a) *NLBDM* and (b) *NLBCM*.

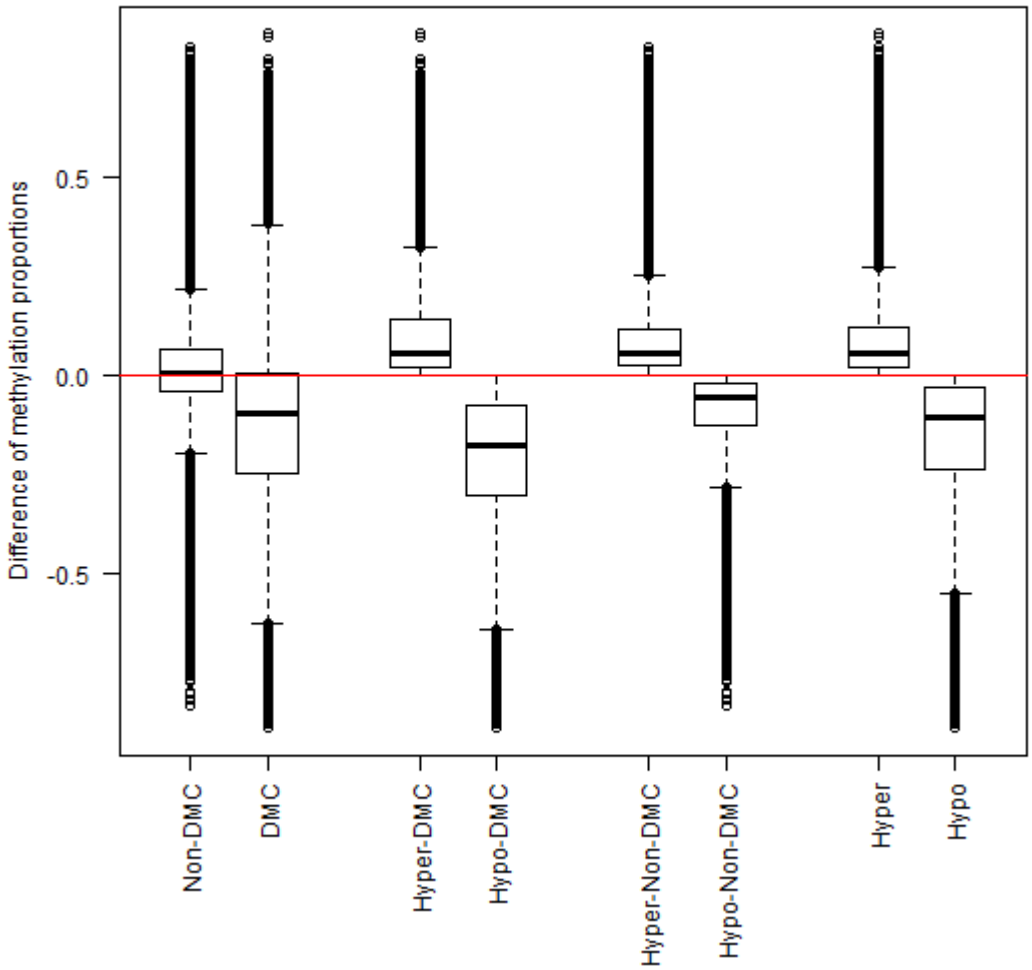


Figure 5.11: For the real study of *NLBDM*, boxplots of the difference of methylation proportions between proliferating and senescent cells classified by various categories defined in the text.

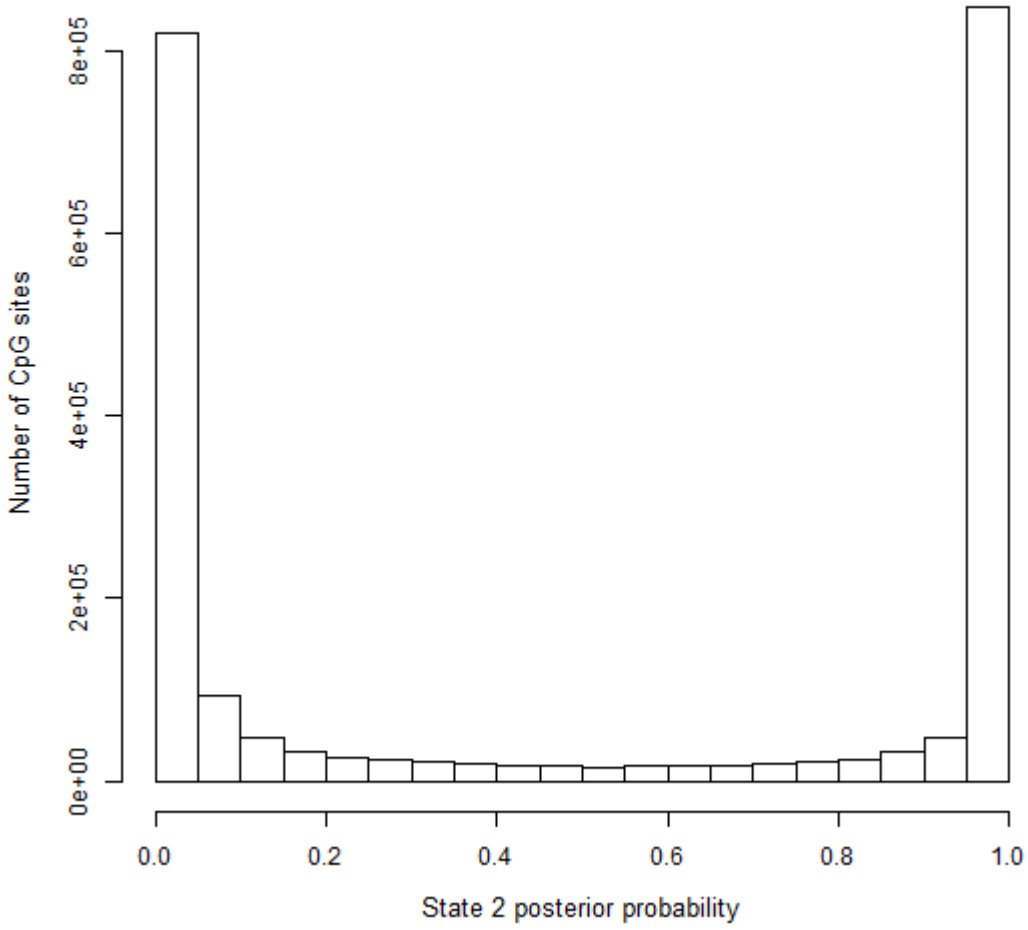


Figure 5.12: For the real study of *NLBDM*, histogram of posterior state 2 probabilities.

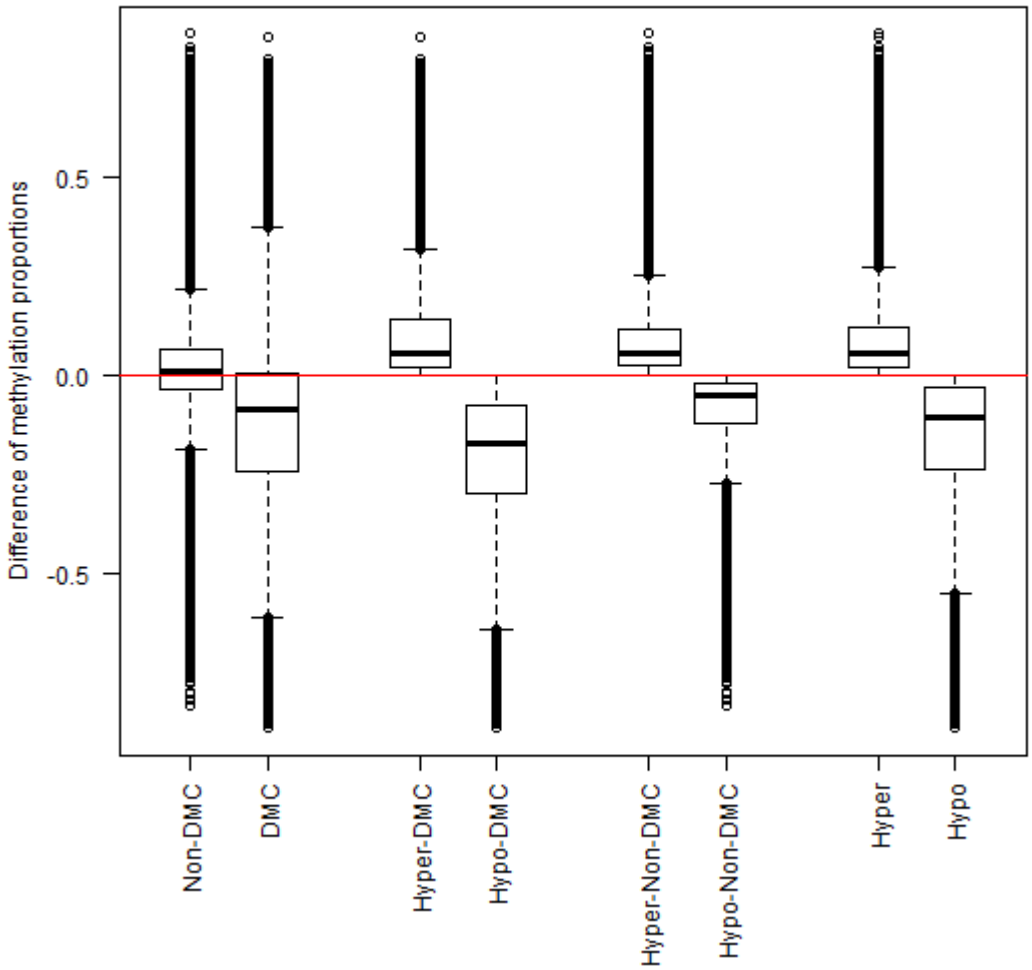


Figure 5.13: For the real study of *NLBCM*, boxplots of the difference of methylation proportions between proliferating and senescent cells classified by various categories defined in the text.

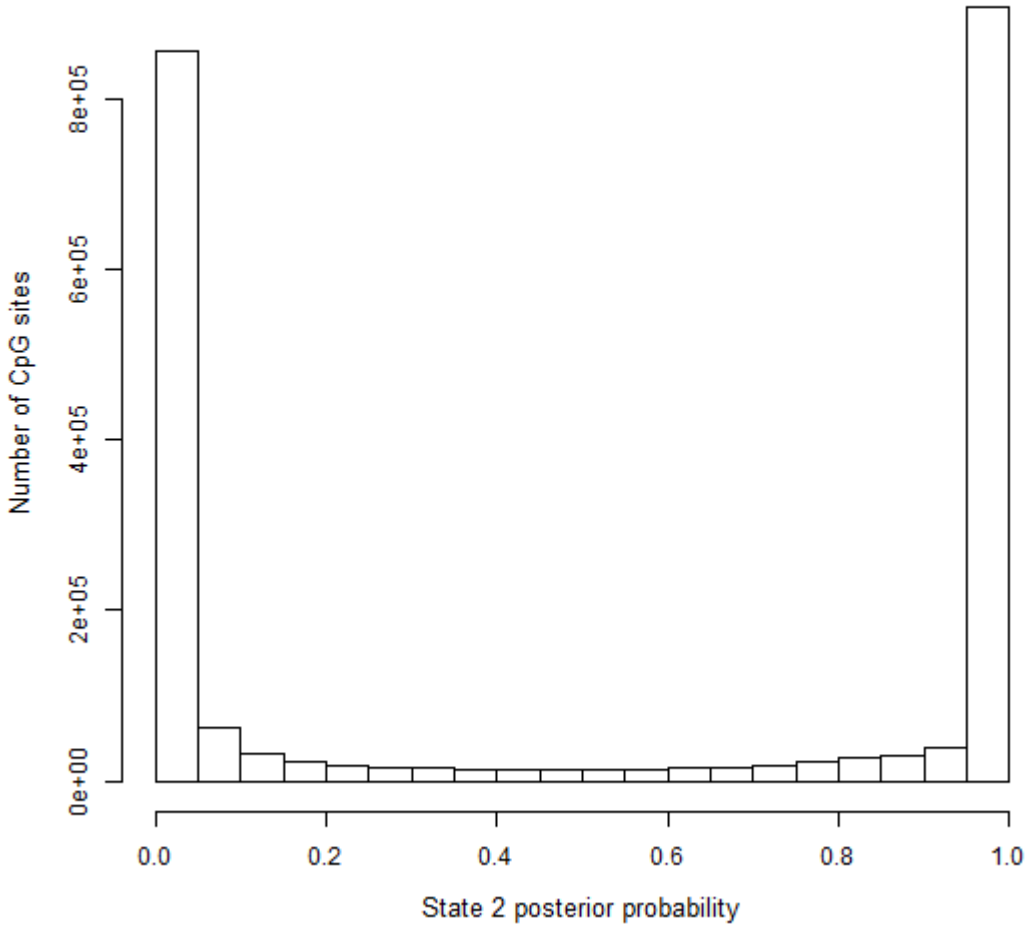
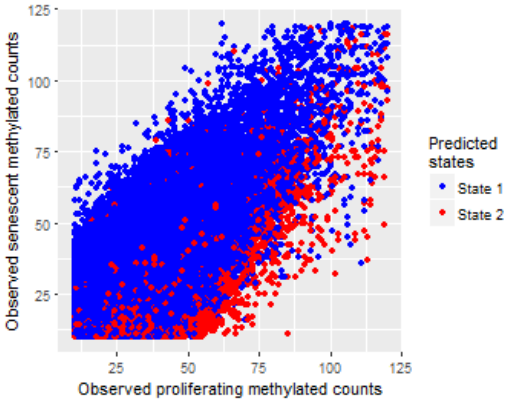
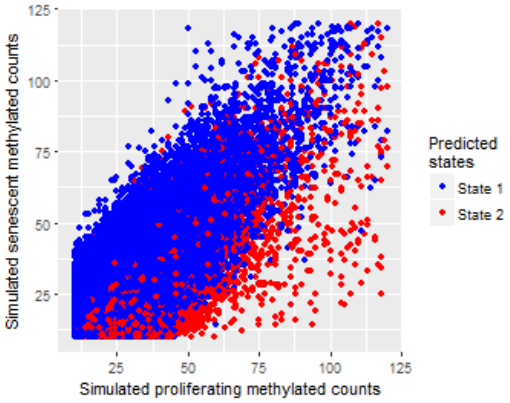


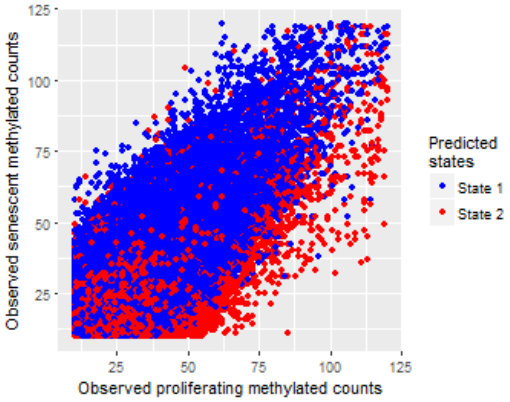
Figure 5.14: For the real study of *NLBCM*, histogram of posterior state 2 probabilities.



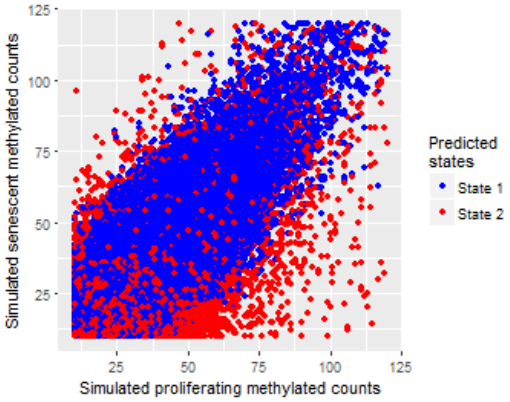
(a) *NLBDM*: Observed



(b) *NLBDM*: Simulated



(c) *NLBCM*: Observed



(d) *NLBCM*: Simulated

Figure 5.15: For the the real study, the 4 panels depict the scatter plots: (a) A scatter plot of observed methylated counts for the proliferating and senescent cells classified by the predicted states for *NLBDM*. (b) A scatter plot of simulated methylated counts (generated using the posterior mean estimates) for the proliferating and senescent cells classified by the predicted states for *NLBDM*. (c) A scatter plot of observed methylated counts for the proliferating and senescent cells classified by the predicted states for *NLBCM*. (d) A scatter plot of simulated methylated counts (generated using the posterior mean estimates) for the proliferating and senescent cells classified by the predicted states for *NLBCM*.

scatterplots of the fitted data classified by the predicted states, displaying that the correlation between the methylated counts of proliferating and senescent cell can be better captured by the extended HMMmethState models: *NLBDM* and *NLBCM*. Furthermore, from these initial visual posterior predictive scatterplots, I concluded that, the Normal-logit-Binomial emission model is an improvement over Beta-Binomial emission model in capturing the correlation between methylated counts of proliferating and senescent cells.

Although the extended HMMmethState models *NLBDM* and *NLBCM* give a reasonable description of the data-generating process, it is still essential to determine the appropriateness of the models to data, more widely. In the following chapter, I will check the practical fit to whole-genome data and whether my selection of it is important to check the practical fit and whether my selection of the Normal-logit-Binomial emission model over the Beta-Binomial emssion model, as described in Chapter 4 is justified or not.

Chapter 6

Assessment of HMMmethState and Biological Results

In this chapter, I perform model adequacy checks and model comparisons to assess the suitability of my proposed HMMmethState models: BBDM, BBCM, NLBDM and NLBCM. I have described the concepts of model assessment within a Bayesian framework in Section 2.3. I use the posterior predictive model checking techniques and model selection criteria to examine the adequacy and fit of the models.

I also assess the performance of my models and compare their efficacies in identifying DMRs/DMCs with other existing methods. While it would be ideal to know the true state of methylation in order to compare the performance of the newly proposed models, unfortunately, an ideal BS-seq test data set with known methylation status at each CpG site does not exist. Even though several studies have been put into developing gold-standard datasets which can be used for comparison purposes, I need a well-founded dataset, such that the data (methylated and unmethylated counts at each CpG site) as well as the *missing data*, i.e., the methylation status at each CpG site, and both the data and the *missing data*

are derived from a realistic approach. Thus, these comparisons are carried out by means of a simulation study where the true methylation status is known, as well as application to real data where this information is missing. I also use an alternative surrogate data to assess the results.

In the following sections, I perform cross comparison under different models. I simulate data under each model and subsequently estimate the model parameters and hidden states to decide whether in each case the correct true model was the most accurate or not. I compare the model assessment and model selection results for all the models when applied to a real dataset. Furthermore, I investigate and explore various ways of assessing the efficacies of DMC calling methods using simulated datasets. In addition, I also present and compare the DMCs and DMRs obtained using my proposed method: HMMmethState and existing methods.

6.1 Simulation study

In this section, the data generation assumptions considered for comparing my proposed HMMmethState models are described. Four simulation studies were performed to compare the predictive accuracies of the hidden states and performances among the four HMMmethState models namely *BBDM*, *BBCM*, *NLBDM*, *NLBCM*. For each simulation study, the data was generated from model M (M:*BBDM*, *BBCM*, *NLBDM*, *NLBCM*) using the posterior estimates (means) of the transition parameters and emission hyperparameters based on a subset of a real dataset (Chromosome 21) of $10,000$ CpG sites. To ensure the data generated in the simulation studies exhibit prominent features of the real data, methylated counts $(\mathbf{x}^p, \mathbf{x}^s)$ were generated for each CpG site using a Binomial distribution, where the total counts $(\mathbf{n}^p, \mathbf{n}^s)$ were taken from the real dataset of Chromosome 21. For each simulation study, the methylated counts $(\mathbf{x} = \mathbf{x}^p, \mathbf{x}^s)$

True base model	Emission hyperparameters								Transition parameters		
BBDM	α	β	γ_1	δ_1	γ_2	δ_2			π_1	τ_{11}	τ_{21}
	5.19	2.678	1.356	3.228	1.107	5.48			0.334	0.867	0.067
BBCM	α	β	γ_1	δ_1	γ_2	δ_2				λ_1	λ_2
	11.62	5.10	1.19	1.90	0.78	1.82				0.534	0.11
NLBDM	μ_*	$\sigma^2{}_*$	ρ_*	μ_p	μ_s	$\sigma^2{}_p$	$\sigma^2{}_s$	ρ_2	π_1	τ_{11}	τ_{21}
	0.33	1.892	0.964	-0.65	-1.61	1.73	2.30	0.968	0.38	0.97	0.0175
NLBCM	μ_*	$\sigma^2{}_*$	ρ_*	μ_p	μ_s	$\sigma^2{}_p$	$\sigma^2{}_s$	ρ_2		λ_1	λ_2
	2.77	1.62	0.73	-0.66	-1.58	3.64	4.71	0.86		0.20	0.128

Table 6.1: Simulation study: parameters for generation of data using HMMmethState models.

and the true hidden states $\mathbf{Z}$ were generated with the parameter values given in Table 6.1. The transition parameters and emission hyperparameters as provided in Table 6.1 were chosen to match the posterior estimates for the realistic cases as described in Sections 4.3.1 and 5.3.1, respectively.

For each of the four competing models, 100 datasets were simulated. Then, each simulated dataset was fitted with the four competing HMMmethState models: *BBDM*, *BBCM*, *NLBDM*, *NLBCM* and compared using different model selection criteria and performance.

6.1.1 Model selection criteria

In this section, the ability of model selection criteria to distinguish among the four HMMmethState models is discussed. The model selection criteria implemented are DIC_1 (Spiegelhalter et al., 2002), DIC_3 (Celeux et al., 2006) as described in Section 2.3.3 and *WAIC* (Gelman et al., 2014, Watanabe, 2010) as described in 2.3.4. Table 6.2 presents the proportion of times that DIC_1, DIC_3 and *WAIC* selected each of the four compteing HMMmethState models for each true model (base model). It can be observed from Table 6.2 that DIC_1, DIC_3 and *WAIC* mostly select the correct model, except in the case of DIC_3 for model *NLBCM*. It can be further noted that the DIC_3 values for models *NLBCM* and *NLBDM* with respect to the true base model *NLBCM* were similar and the differences

between these values were small. Hence, this kind of model selection criterion rejects the information about relative model selection accuracy contained in the differences between the DIC_3 values of models *NLBCM* and *NLBDM*.

Model	Chosen model	BBDM	BBCM	NLBDM	NLBCM
Base model					
BBDM	DIC_1	**0.71**	0	0.29	0
	DIC_3	**0.8**	0.0	0.2	0.0
	WAIC	**0.63**	0.05	0.26	0.06
	misclass. prob.	**0.0195**	0.1941	0.0201	0.0415
BBCM	DIC_1	0	**1**	0	0
	DIC_3	0.0	**1.0**	0.0	0.0
	WAIC	0.03	**0.67**	0.09	0.21
	misclass. prob.	0.0816	**0.0660**	0.0743	0.0679
NLBDM	DIC_1	0	0.27	**0.73**	0
	DIC_3	0.0	0.4	**0.6**	0.0
	WAIC	0.02	0.03	**0.88**	0.7
	misclass. prob.	0.3503	0.3075	**0.0263**	0.1659
NLBCM	DIC_1	0.2	0.19	0.08	**0.53**
	DIC_3	0.4	0.3	0.3	**0.0**
	WAIC	0.01	0.06	0.11	**0.82**
	misclass. prob.	0.1186	0.4215	0.0341	**0.0956**

Table 6.2: Performance of model selection criteria and sensitivity based on the simulation study.

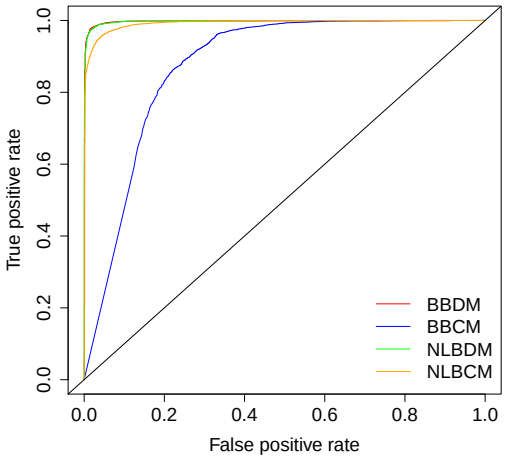
6.1.2 ROC curves

In the previous section, I examined the ability of HMMmethState models in selecting the true base model. Now, in this section, I review the performance of the HMMmethState models using receiver operating characteristic (ROC) curves based on the simulation study design described in Section 6.1. The ROC curve of the model-based method explains the relationship between the false positive rate (FPR) against true positive rate (TPR) of methylation status at each CpG site. The TPR can be described as the proportion of correctly identified differentially methylated CpG sites. The FPR can then be described as the proportion of

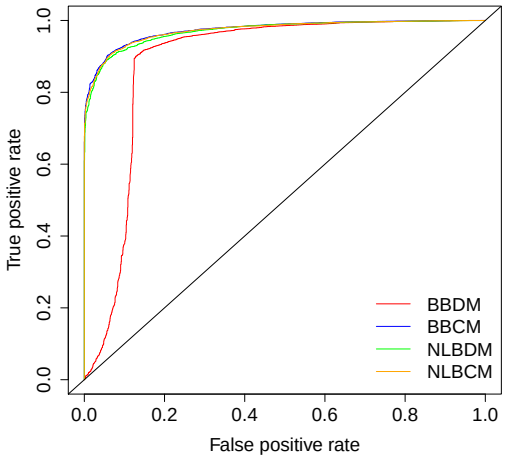
similarly methylated CpG sites which are incorrectly selected by the method due to classification error. I present the results of the misclassification rates of the model with repect to the true base models in Table 6.2. It can be clearly observed from Table 6.2 that the misclassification rates of HMMmethState models are the lowest when the data are generated from the true (base) models except in the case for base model *NLBCM*, where the misclassification rate of model *NLBDM* is the lowest. Overall, the performance of model *NLBDM* is the best in terms of misclassification rates as it ranged between (0.0201, 0.0743) irrespective of the true base models. These performances of HMMmethState models can also be validated visually using the ROC curves. Figures 6.1 shows the ROC curves for the models *BBDM* (red line), *BBCM* (blue line), *NLBDM* (green line), *NLBCM* (yellow line) with area under the ROC curves suggestive of the relative accuracies of the models in identifying the status of methylated CpG sites averaged over 100 repetitions. While the performance of model *NLBDM* efficiently overtakes the performance of the other models, *NLBCM* also attains a higher area under the curves than the other two competing models. However, the pertinent question arises, which among these models is the best one and on what basis? I next study the choice of the best model in modelling the real data.

6.2 Real data analysis ($0.060034 - 90.294609$ Mb on chromosome 16)

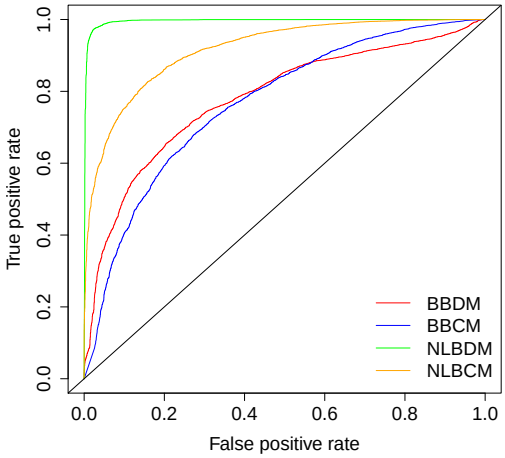
The performance of the 4 HMMmethState models namely *BBDM*, *BBCM*, *NLBDM*, *NLBCM* are assessed with respect to their corresponding true base models in the previous simulation study section 6.1. In this section, I assess the adequacy and appropriateness of the competing HMMmethState models and compare the models using different model selection criteria.



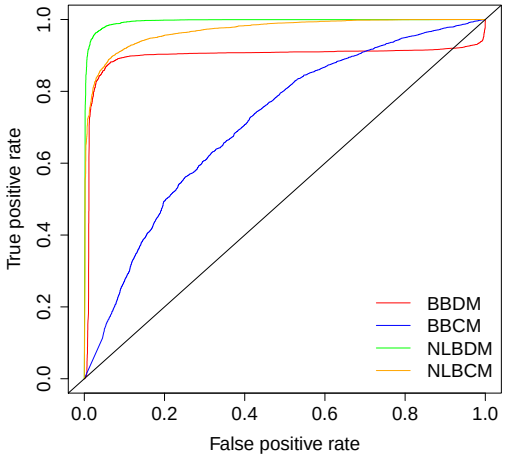
(a) Base model: *BBDM*



(b) Base model: *BBCM*



(c) Base model: *NLBDM*



(d) Base model: *NLBCM*

Figure 6.1: For the simulation study of four true base HMMmethState model setups: (a) *BBDM*, (b) *BBCM*, (c) *NLBDM* and (d) *NLBCM*, four panel depict the ROC curves for each HMMmethState models in comparison to the true base HMMmethState models based on 100 simulated datasets.

6.2.1 Posterior predictive model checking

The fit of the model can be studied using the log-posterior predictive distributions of the data (Gelman and Meng, 1998), which are commonly used as the

discrepancy statistics for finite mixture models. Thus, the use of such kind of discrepancy statistics can also be extended for HMMs (Scott, 2002) As already described in details in Section 2.3.2, the discrepancy statistics can be functions of both the parameters and the data, that assess the discrepancy between the model and the data rather than correctness of the model. In order to explore the relevant characteristics of the BS-seq methylation data, I use the most pragmatic version of the discrepancy test statistic, log-posterior predictive distributions as used in Gelman and Stern (2000) and Jonghyun et al. (2014), which include all the features of my model parameters.

Let us denote $\mathbf{x}^{rep}$ as the replicated data simulated from the posterior predictive distribution $p(\mathbf{x}^{rep}|\mathbf{x})$ and then the discrepancy test-statistics can be described for model M as,

$$\mathbf{T}(\mathbf{x}, \boldsymbol{\zeta}^{(M)}) = \log L_{\mathbf{x}}(\boldsymbol{\zeta}^{(M)}), \tag{6.1}$$

where $\boldsymbol{\zeta}^{(M)}$ denotes the HMM parameters for model M.
The posterior predictive p-value for model M can also be explained as the probability that the replicated data is more extreme than the observed data, which is defined as below,

$$p^{(M)} = P(\mathbf{T}(\mathbf{x}^{rep}, \boldsymbol{\zeta}^{(M)}) \geq \mathbf{T}(\mathbf{x}, \boldsymbol{\zeta}^{(M)})|\mathbf{x}). \tag{6.2}$$

To study the plausability of HMMmethState models: *BBDM*, *BBCM*, *NLBDM*, *NLBCM*, I compute the posterior predictive p-values of all the models. The posterior predictive p-value can be interpreted as the measure to evaluate the discrepancies between the model and the data. I calculate p-values using the MCMC samples for $i = 1, \ldots, I$, $[(\mathbf{x}^{rep})^{(i)}, \boldsymbol{\zeta}^{(M)(i)}]$ and $(\mathbf{x}^{rep})^{(i)}$ are generated from $p(\mathbf{x}|\boldsymbol{\zeta}^{(M)(i)})$ for model M. I performed the posterior predictive checks for all the competing models. I present the scatterplots in Figure 6.2 of the MCMC

Model		BBDM	BBCM	NLBDM	NLBCM
	DIC_1	20162274	23754504	6400362623	9994570530
	p_{DIC_1}	5.897	254064	3194513526	4991464492
	DIC_3	20098146	23229972	6400362180	9994568928
	p_{DIC_3}	5.816	246928	3194512813	4991460281
	$WAIC$	19258033	22787917	14960221	15137114
	p_{WAIC}	8.64	19.65	1614579	1688645
	$p^{(M)}$	0.0000000	0.0000000	0.6119829	0.7517832

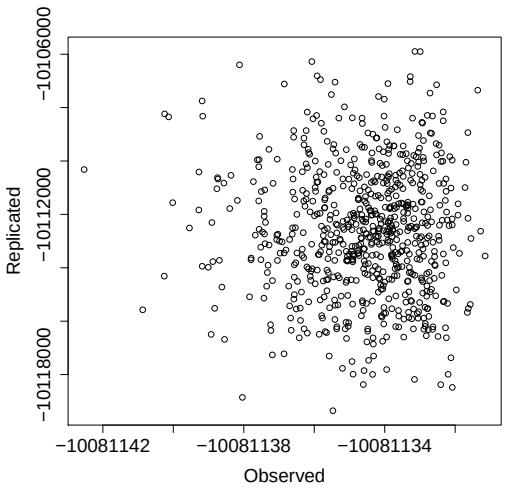
Table 6.3: Model comparisons.

simulated paired values of $\mathbf{T}(\mathbf{x}, \boldsymbol{\zeta}^{(M)})$ and $\mathbf{T}(\mathbf{x}^{rep}, \boldsymbol{\zeta}^{(M)})$ for the four models after burn-in.

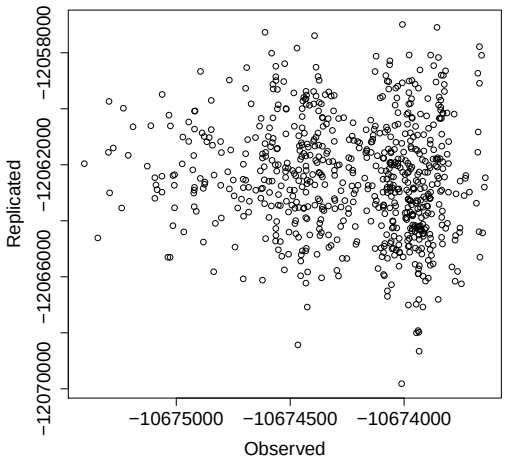
I performed the posterior predictive checks for all the competing models and the scatter plots based on the test statistics for model M (6.1) are displayed in Figure 6.2. The last row in Table 6.3 show the posterior predictive p-values, which demonstrate the discrepancies between the models and the data. The p-values of *BBDM* and *BBCM* being 0 clearly indicate a lack of model fit. The posterior predictive p-values for *NLBDM* and *NLBCM* do not show any evidence of discrepancies between the model and the data. It can also be observed from Figure 6.2 that the replicated log-posterior densities are higher than the observed log-posterior densities over the MCMC draws. Thus, the posterior checking I conducted in this section, indicates that models *BBDM*, *BBCM* are not adequate for the data whereas *NLBDM*, *NLBCM* would be better suited for BS-seq methylation data.

6.2.2 Model selection

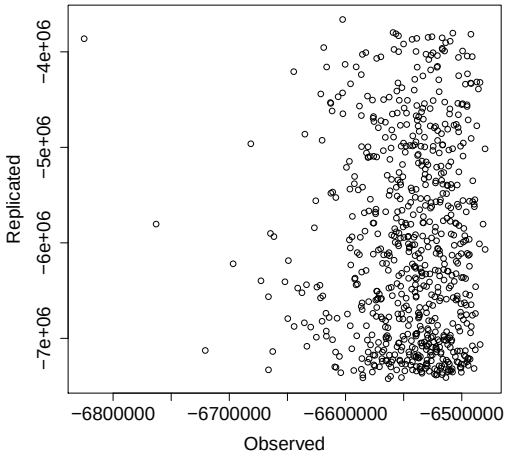
Table 6.3 comprises the model selection criteria estimates from the real data analysis. I have presented the values of the effective number of parameters for both versions of DICs, i.e., DIC_1 and DIC_3 for HMMmethState model in Table 6.3 to show the variations of these values especially for models *NLBDM* and *NLBCM*.



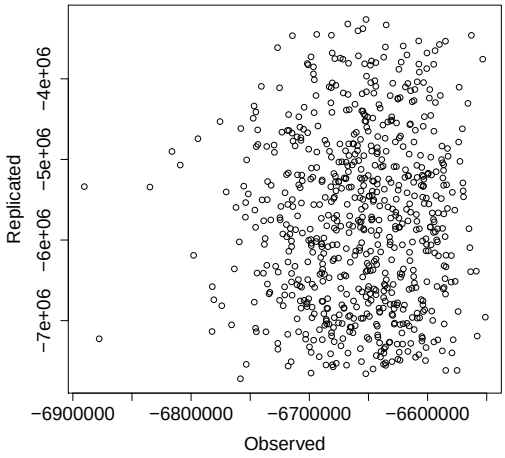
(a) *BBDM*



(b) *BBCM*



(c) *NLBDM*



(d) *NLBCM*

Figure 6.2: For the real study of four HMMmethState models: (a) *BBDM*, (b) *BBCM*, (c) *NLBDM* and (d) *NLBCM*, four panel depict the scatter plots of log-posterior densities for the observed and replicated data based on thinned MCMC draws.

Although the values of two versions of DIC are different, both these DICs, i.e., DIC_1 and DIC_3 select the same model. I observed this characteristic even in the simulation study of model selection criteria as described in Section 6.1.1. DIC

values for models *NLBDM* and *NLBCM* are largely affected due to the variations in the values of effective dimension (effective number of free parameters) compared to models *BBDM* and *BBCM*, where the effective dimension values are quite stable. As one can see from Figure 6.2, the y-axis (observed log-posterior densities based on $10,000$ MCMC iterations) of the scatter plots for models *NLBDM* and *NLBCM* show huge variations compared to models *BBDM* and *BBCM*. Table 6.3 displays the estimated effective dimension of values of *WAIC* which are quite stable for all the HMMmethState models. I have also observed that the estimated effective dimension values for models *NLBDM* and *NLBCM* are approximately close to the total number of data points (around 2.16 million data points for Chromosome 16).

WAIC based model selection has an edge over DIC based model selection specifically for models with mixture and hierarchical structures as explained in Gelman et al. (2013). The point estimates of the model sometimes do not make sense as the number of parameters increases with the sample size for hierarchical HMMs. I also found the efficacy of *WAIC* in this section. As it can be observed from Table 6.3 that with the increase in the effective number of parameters, the DIC values were inconclusive especially for models *NLBDM* and *NLBCM*. Due to the non-conjugate structure of models *NLBDM* and *NLBCM*, the number of parameters (auxiliary parameters) increases with the sample size for HHMM, which is evidently quite high for real data. DIC values can be quite distinct to each other for the pair of models *BBDM*, *BBCM* and models *NLBDM*, *NLBCM* as they depend on the effective dimension, i.e., effective dimension of auxiliary parameters for models *NLBDM* and *NLBCM*, which is essential to the idea of DIC (Celeux et al., 2006). Several authors including Celeux et al. (2006) and Plummer (2008) have suggested that DIC might not be appropriate in the context of hierarchical missing data models. Although Celeux et al. (2006) discussed in details different variations of DICs for missing data models. I have only described three versions

of DICs in Section 2.3.3. DIC expressions for latent variable models were previously explored by Richardson (2002) which was again discussed by Celeux et al. (2006), Hooten and Hobbs (2015) and later which had been theoretically justified by Watanabe (2010). For the purpose of model selection, I applied *WAIC* expressions for selecting the best models based on real data.

6.3 Comparison with other methods

Two existing methods methylKit and DSS, that analyse BS-seq data in order to detect DMCs/DMRs were discussed in Chapter 3. To compare the performance of my models to other competing methods, I implemented an extensive simulation study based on the HMMmethState models to compare the performance of each of my true HMMmethState models with methylKit and DSS.

6.3.1 Simulation study

To examine the robustness of my proposed method, i.e., HMMmethState, with other existing methods, I performed a simulation study. In this section, I investigate the performance of each of the models of HMMmethState with DSS and methylKit.

The simulation procedure remains the same as described in Section 6.1. For each simulation study, the methylated counts $\mathbf{x} = (\mathbf{x}_p, \mathbf{x}_s)$ and true underlying methylation status for each CpG site, i.e., $\mathbf{Z}$, hidden states, were generated with the parameter values provided in Table 6.1. The simulation study involved 100 replications under each simulation setting.

I fit four HMMmethState models to the data generated using the true base models. Furthermore, I simultaneously fit DSS and methylKit models to the data and compare them with each of the HMMmethState models (true base models)

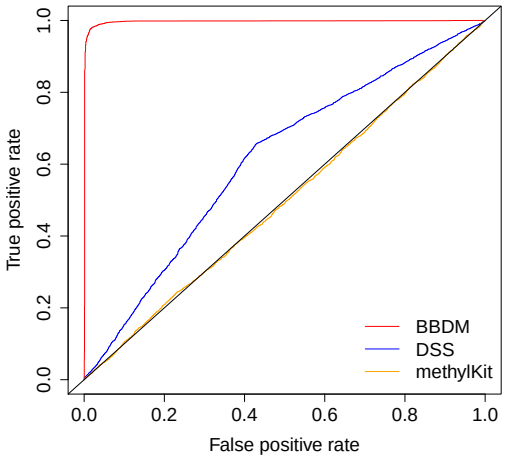
that generated the data.

6.3.2 ROC curves

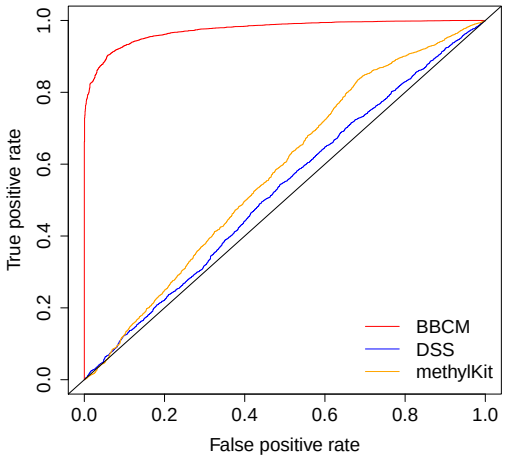
To compare the performance of each of the HMMmethState models with the other differential methylation caller methods, I inspect their ROC curves. For each simulation setup of each of the HMMmethState models, I plot the ROC curves of each of the HMMmethState (true base) models and then compare with DSS and methylKit averaged over 100 repetitions. Figure 6.3 shows the ROC curves for the HMMmethState methods (red line), DSS (blue line) and methylKit (orange line) with areas under the ROC curves indicating the accuracies of the competing methods in identifying the DMCs. For the ROC curves in Figure 6.3, HMMmethState clearly achieves the highest area under the ROC curves than the competing methods: DSS and methylKit irrespective of the HMMmethState models, thus inferring their high reliability in identifying DMCs.

6.4 Simulating data from a mixture model

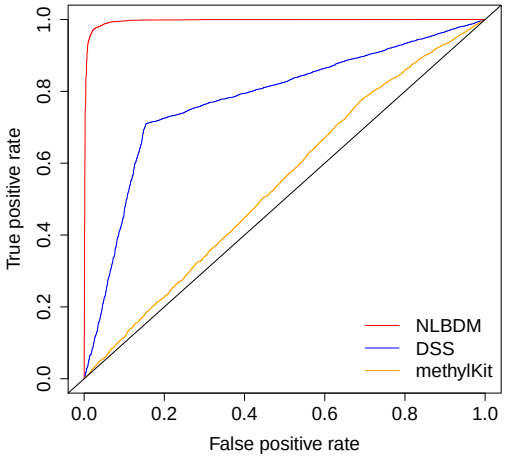
To assess the reliability of HMMmethState, computationally-derived data that mimic the experimental observations with known underlying structure of the hidden states was simulated in Section 6.3.1 using HMMmethState model as the true base model. Although my HMMmethState method models the properties of the data reasonably well as described in Section 6.3.2, the performance should not be examined on data simulated using the same models, since it tends to provide an undue advantage to the true base model and the comparison would eventually become biased. To implement a more objective analysis, I searched for a conceptually distinct simulator that has similarities to the data assumption criteria used in the competing methods: methylKit and DSS. Clearly, both these models do not consider Markovian dependence between two adjacent hidden states in their model specifications, since the spatial dependence among the CpG sites is



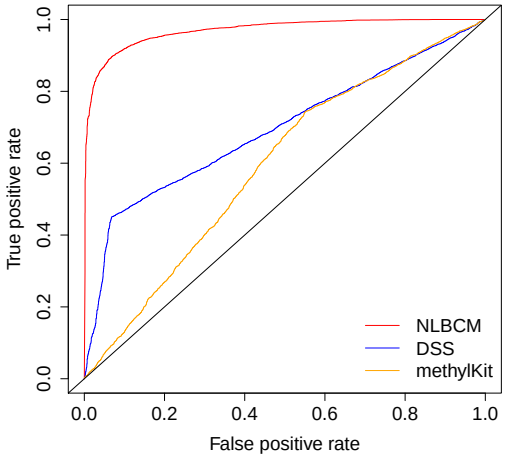
(a) Base model: *BBDM*



(b) Base model: *BBCM*



(c) Base model: *NLBDM*



(d) Base model: *NLBCM*

Figure 6.3: For the simulation study of four true base HMMmethState model setups: (a) *BBDM*, (b) *BBCM*, (c) *NLBDM* and (d) *NLBCM*, four panel depict the ROC curves for methylKit and DSS in comparison to the true base HMMmethState models.

not taken into account.

One possible model that could emulate BS-seq methylation data and yet retain

non-Markovian and independence assumptions of the hidden states is the mixture model. To test the robustness of HMMmethState to the data generated using the mixture model, I simulated methylated counts from a model similar to model *NLBDM* as described in Chapter 5 with the only difference being that the hidden states $\mathbf{Z}$ are assumed to be independent. The joint mixture model (Model *MM*) for $\mathbf{X}$ can be defined by its probability distribution:

$$P(\mathbf{x}|\boldsymbol{\eta}_1^{(MM)}, \boldsymbol{\eta}_2^{(MM)}) = \pi^{(MM)} P\left(\mathbf{x}|\boldsymbol{\eta}_1^{(MM)}, \mathbf{Z}\right) + \left(1 - \pi^{(MM)}\right) P\left(\mathbf{x}|\boldsymbol{\eta}_2^{(MM)}, \mathbf{Z}\right), \tag{6.3}$$

where $\pi^{(MM)}$ is the mixture proportion of state 1 component and $P\left(\mathbf{x}|\boldsymbol{\eta}_1^{(MM)}, \mathbf{Z}\right)$ and $P\left(\mathbf{x}|\boldsymbol{\eta}_2^{(MM)}, \mathbf{Z}\right)$ are state-dependent densities of state 1 and state 2 respectively. I have already defined the generic notations of the emission parameters in Section 5.1.4.

Furthermore, $P\left(\mathbf{x}|\boldsymbol{\eta}_k^{(MM)}, \mathbf{Z}\right)$ for state k can be written as,

$$P\left(\mathbf{x}|\boldsymbol{\eta}_k^{(MM)}, \mathbf{Z}\right) = \prod_{t=1}^{T} P\left(\mathbf{x}_t|\boldsymbol{\eta}_k^{(MM)}(t)\right)^{\mathbf{I}[Z_t=k]}, \ k = 1, \ 2. \tag{6.4}$$

It is obvious that in a mixture model, the hidden states $\mathbf{Z}$, which influence the mixture component to be picked for each observation, are independent of each other rather than related through a Markov process (as in the case of HMM).

The structure of the hierarchical mixture model can be written as follows:

$$\begin{aligned} &X_t^p|Z_t = k \sim Bin\left(n_t^p, logit^{-1}(q_t^{pk})\right) \text{ and } X_t^s|Z_t = k \sim Bin\left(n_t^s, logit^{-1}(q_t^{sk})\right), \\ &\mathbf{Q}_t^k|Z_t = k \sim BVN(\boldsymbol{\theta}_k^{(MM)}), \ k = 1, \ 2 \text{ and } t = 1, \ldots, T, \end{aligned} \tag{6.5}$$

where $\boldsymbol{\theta}_k^{(MM)} = (\mathbf{M}_k, \boldsymbol{\Sigma}_k)$ and $BVN(.)$ is the bivariate Normal distribution as described in Section 5.1. The structure of the hierarchical mixture model remains the same as of *NLBDM* and *NLBCM* except that the hidden states are assumed to be non-Markovian and independent.

The likelihood function given the data $\mathbf{x}$ and the hidden states $\mathbf{Z}$ is given by,

$$\begin{aligned} L_{\mathbf{x},\mathbf{Z}}\left(\boldsymbol{\eta}_1^{(MM)}, \boldsymbol{\eta}_2, \pi^{(MM)}\right) &= \prod_{t=1}^{T}\Bigg[\left\{\pi^{(MM)} P\left(\mathbf{x}_t|\mathbf{Q}_t^1\right)^{\mathbf{I}[Z_t=1]}\right\} \\ &\quad \times \left\{\left(1-\pi^{(MM)}\right) P\left(\mathbf{x}_t|\mathbf{Q}_t^2\right)^{\mathbf{I}[Z_t=2]}\right\}\Bigg] \\ &= \left\{\pi^{(MM)}\right\}^{t_1}\left\{1-\pi^{(MM)}\right\}^{t_2} \\ &\quad \times \prod_{t=1}^{T}\left[P\left(\mathbf{x}_t|\mathbf{Q}_t^1\right)\right]^{\mathbf{I}[Z_t=1]} \prod_{t=1}^{T}\left[P\left(\mathbf{x}_t|\mathbf{Q}_t^2\right)\right]^{\mathbf{I}[Z_t=2]}. \end{aligned} \tag{6.6}$$

The probability for the sequence of the hidden states $\mathbf{Z}$ conditional on the mixture proportion $\pi^{(MM)}$ is:

$$P(\mathbf{Z}|\pi^{(MM)}) = \left\{\pi^{(MM)}\right\}^{t_1}\left\{1-\pi^{(MM)}\right\}^{t_2}. \tag{6.7}$$

Similar to my previous approaches as described in Equation 4.55 for the state transition probabilities, I choose a Uniform prior for $\pi^{(MM)}$ such that $\pi^{(MM)} \sim Beta(1,1)$ independently and sample the mixture proportion $\pi^{(MM)}$ conditional on the hidden states $\mathbf{Z}$ as below.

$$\pi^{(MM)}|\mathbf{Z} \sim Beta(t_1+1, t_2+1). \tag{6.8}$$

The conditional bivariate Normal priors of the auxiliary emission parameters $\boldsymbol{\eta}$ and the priors for the global hyperparameters $\boldsymbol{\theta}$ are explained in Sections 5.19 and 5.19, respectively.

The full conditional posterior distributions of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ are described in detail in Sections 5.2.3.1 and 5.2.3.2, respectively.

The full conditional posterior probability of Z_t at state k, $k = 1, 2$ are:

$$P(Z_t = 1|\mathbf{x}, \boldsymbol{\eta}^{(MM)}, \boldsymbol{\theta}^{(MM)}, \pi^{(MM)}) \propto \pi^{(MM)} P\left(\mathbf{x}_t|\mathbf{Q}_t^1\right) \tag{6.9}$$

and

$$P(Z_t = 2|\mathbf{x}, \boldsymbol{\eta}^{(MM)}, \boldsymbol{\theta}^{(MM)}, \pi^{(MM)}) \propto \left(1 - \pi^{(MM)}\right) P\left(\mathbf{x}_t|\mathbf{Q}_t^1\right). \tag{6.10}$$

6.4.1 Summary of the Gibbs sampler algorithm steps

1. Initialize all auxiliary parameters $\boldsymbol{\eta}^{(MM)}$, hyperparameters $(\boldsymbol{\theta}^{(MM)})$, $\mathbf{Z}$.
2. Initialize mixture proportion $\pi^{(MM)}$.
3. Update $\boldsymbol{\eta}^{(MM)}$ from the full conditional posterior distributions in Section 5.2.3.1.
4. Update $(\boldsymbol{\theta}^{(MM)})$ from the full conditional posterior distributions in Section 5.2.3.2.
5. Update $\pi^{(MM)}$ from the full conditional posterior distribution in Section 6.8.
6. Implement the relabelling algorithm as described in Section 2.2.6.
7. Repeat steps (3)-(6) until convergence.

To resemble the real data, for model *MM*, the data are generated using the real data study posterior estimates. The posterior estimates of the parameters obtained using MCMC techniques and convergence properties of the estimates as

explained in Section are given below:

$$\begin{aligned}
\mu_* &= 0.26 \\
\sigma_*^2 &= 1.54 \\
\rho_* &= 0.91 \\
\mu_p &= -0.82 \\
\mu_s &= -1.95 \\
\sigma_p &= 1.68 \\
\sigma_s &= 2.27 \\
\rho_2 &= 0.96 \\
\pi^{(MM)} &= 0.42
\end{aligned} \tag{6.11}$$

In this case, the simulation procedure was replicated 100 times, such that 100 datasets were generated with $T = 10000$ observations. I fit four HMMmethState models to the mixture model *MM* data. Furthermore, I fit *methylKit* and *DSS* models to the data. In Figure 6.4, I plotted the ROC curves. For the ROC plot, all the HMMmethState methods attain the higher area under the ROC curve compared to the competing methods. In addition, Model *NLBDM* again outperforms all the competing methods and validates its high reliability in identifying DMCs.

6.5 Real data analysis across all chromosomes (Cruickshanks et al., 2013)

In this section, I demonstrate the results of applying the proposed HMMmethState models on BS-seq methylation data from Cruickshanks et al. (2013). I analyse the whole data for detecting DMRs across all chromosomes described in Sections 6.5.2 and 6.5.3, respectively.

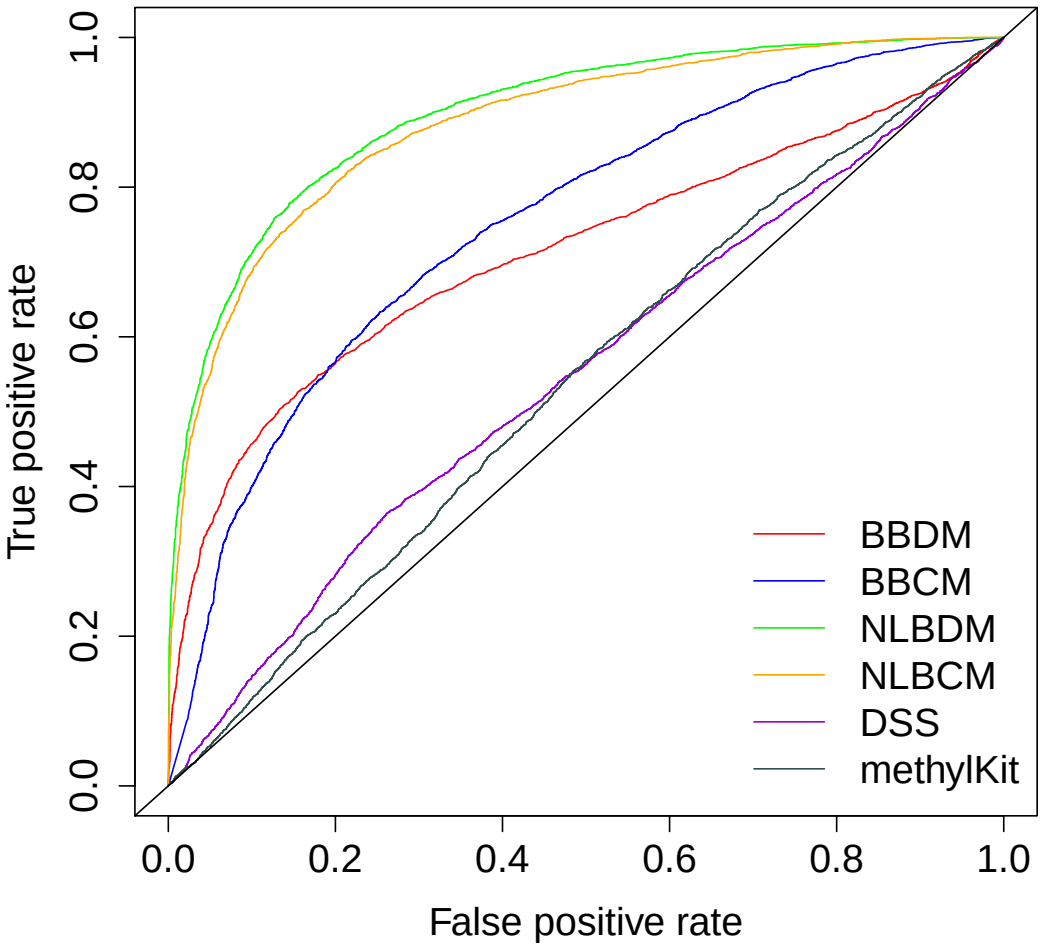


Figure 6.4: For the mixture model *MM* simulation study, ROC curves for models *BBDM*, *BBCM*, *NLBDM*, *NLBCM*, *methylKit* and *DSS*.

6.5.1 Implementations of HMMmethState models with methylKit, DSS

In this section, I compare the results of HMMmethState analysis with the results obtained from DSS and methylKit based on their publicly available R/Bioconductor package implementations. I have applied HMMmethState models to analyze four chromosomal datasets: Chromosomes-3, 9, 14, 22 and model selection was done using WAIC. I only considered $20,000$ contiguous CpG sites (randomly selected) of each chromosome for comparisons.

In this subsection, I have fitted four HMMmethState models: *BBDM*, *BBCM*, *NLBDM* and *NLBCM* to the data (Chromosome 16). The WAIC favors the *NLBDM* model as the best among the four HMMmethState models, which implies positional variations of the CpG sites does not affect the methylation status prediction of the neighboring CpG sites. In the rest of this subsection, I illustrate the HMMmethState results based on all the four HMMmethState models.

I first applied the algorithms based on the four HMMmethState models and then compared the predicted states with the competing methods *methylKit* (Akalin et al., 2012) and *DSS* (Wu et al., 2015) as described in Sections 3.3.1 and 3.3.2 respectively. I presented the results of the DMCs identified by each of the four HMMmethState models, *methylKit* and *DSS* in the following subsections. Figure 6.5 presents a Venn diagram that summarizes the results for DSS, methylKit and HMMmethState models.

- Chromosome 3: I have applied the HMMmethState, methylKit and DSS methods to $20,000$ CpG sites of Chromosome 3. The WAIC favors the *NLBCM* model as the best among the four HMMmethState models for this dataset, which implies positional variations of the CpG sites affects the methylation status prediction of the neighboring CpG sites. The sets

of DMCs identified by the methods HMMmethState, methylKit and DSS are summarized in Figure 6.5a. The method HMMmethState discovered 19,999 CpG sites. In contrast, the methods DSS and methylKit detected only 13,796 DMCs and 14,790 DMCs respectively. A closer examination sheds light on the differing sets of DMCs identified by HMMmethState, methylKit and DSS. Of the DMCs detected by HMMmethState, as many as 13,795 DMCs were also identified by DSS. And, all the DMCs detected by methylKit were also identified by HMMmethState.

- Chromosome 9: All methods were applied to 20,000 CpG sites of Chromosome 9. The WAIC favors the *NLBDM* model as the best among the four HMMmethState models for this dataset, which implies positional variations of the CpG sites does not affect the methylation status prediction of the neighboring CpG sites. The HMMmethState technique identified 13,565 DMCs. The overlapping set of DMCs are summarized in Figure 6.5b and reveal a greater lack of concordance among the methods than the Chromosome 3 dataset. Only 6,117 CpG sites are identified as DMCs by all three methods. This low level of agreement is a result of the low overlap that methylKit has with the other methods.

- Chromosome 14: I have applied the HMMmethState, methylKit and DSS methods to 20,000 CpG sites of Chromosome 14. I have found that the WAIC favours the *NLBDM* model, which is a hierarchical correlated HMM without CpG sites dependence. The sets of DMCs identified by the methods HMMmethState, methylKit and DSS are summarized in Figure 6.5c. Here, HMMmethState method identifies 17,190 CpG sites as DMCs. methylKit and DSS detect 12,633 and 11,447 DMCs respectively. 6,843 DMCs are identified by all the three methods.

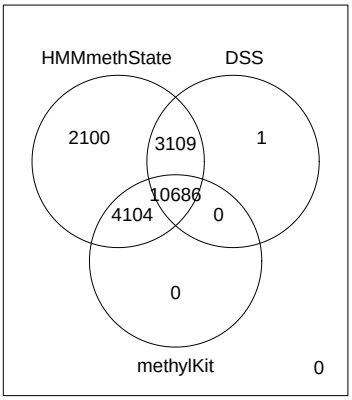
- Chromosome 22: Here, WAIC favours the *NLBDM* model as well. 8795 DMCs are identified by all the three methods. Of the 13,785 DMCs iden-

tified by DSS, HMMmethState detected 12,717 DMCs and of the 17,884 identified by HMMmethState, methylKit successfully detected 11,873 DMCs. Figure 6.5d presents a Venn diagram that summarizes the results for DSS, methylKit and HMMmethState.

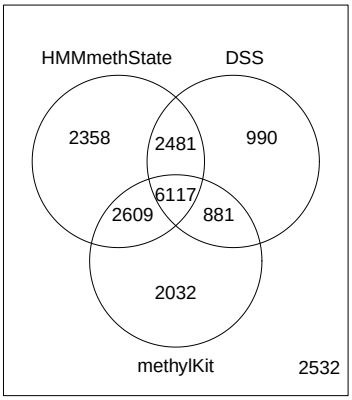
6.5.2 Spatial dependence comparison among chromosomes

I analyze the BS-seq data of each chromosome separately and subsequently investigate the posterior distribution of the chromosome-specific parameters using the MCMC based algorithm explained in Section 5.2.2, since all the models selected were either *NLBDM* or *NLBCM*.

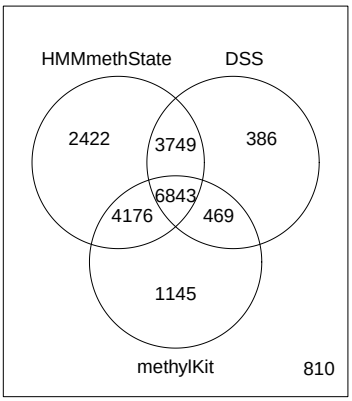
To analyze the difference between models NLBDM and NLBCM, I investigate patterns among the chromosomal datasets in which one model fits better than the other. One significant pattern is related to the way the two models tackle the spatial dependence in the data. Their difference is quite evident in the posterior distributions of the estimated probabilities or the deviations of certains data points. Model NLBDM fits better than model NLBCM in most datasets. Selecting model NLBDM implies that the positional variations among the CpG sites do not affect the spatial dependence among the CpG sites. The chromosomal datasets that select model NLBCM are: Chromosome 1, 2, 3, 5, 6, 8 and Y. The remaining chromosomal datasets select NLBDM for model fit based on *WAIC*. Furthermore, I estimate the credible intervals of HMMmethState model parameters selected using *WAIC* for each chromosomal datasets. Tables 6.4 and 6.5 present the WAIC picked model for each chromosome. Figures 6.6, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12 and 6.13 display the credible intervals of the WAIC picked model emission parameters for each chromosome. Since NLBDM and NLBCM have the same set of emission parameters, I plotted the credible intervals of the parameter in the same graph. Clearly, there is a pattern of consistency in the credible intervals of the parame-



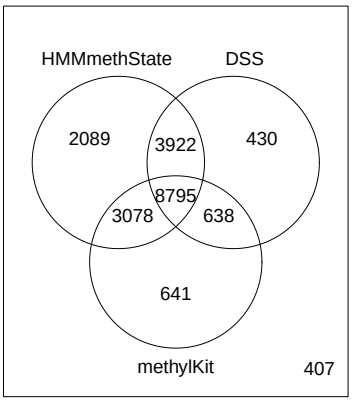
(a) *NLBCM*- Chr3



(b) *NLBDM*- Chr9



(c) *NLBDM*- Chr14



(d) *NLBDM*- Chr22

Figure 6.5: Venn diagrams for the DMCs identified by the methods HMMmethState, DSS and methylKit in the real data analysis of 20,000 CpG sites of 4 chromosomes. HMMmethState model setups: (a) *NLBCM*- Chr3, (b) *NLBDM*- Chr9, (c) *NLBDM*- Chr14 and (d) *NLBDM*- Chr22.

ters. There is a significant variation in the credible intervals of *NLBDM* model parameters to the *NLBCM* model parameters due to the change of assumptions in the respective transition models. The credible interval regions of the WAIC

picked model M: ($NLBDM$, $NLBCM$) parameters are quite consistent in nature and the two different model parameters can even be visually segregated from Figures 6.6 to 6.13.

6.5.3 Defining DMR windows

In many instances, it might be sensible to summarize the differential methylation status of each CpG site over tiling windows instead of the single base pair resolution. As an example, D Smith et al. (2012) studied methylation profiles with RRBS experiments on gametes and zygote and subsequently summarized methylation data information over 100 bp resolution windows across the genome. These results uncovered a unique segment of DMRs maintained in early embryo. Employing tiling window techniques could be useful when methylation pattern of a region determines its whole funtional dynamics and also help in understanding the role of gene-expression in differential methylation.

I implemented a simple technique in HMMmethState for defining DMRs based on the predicted methylation status at each CpG site. The method I implemented to calculate the start and end region of these 500 bp windows is slightly different from the conventional 500 bp equispaced tiling windows. The start and end of the region are the start and end position of the CpG site of each window. In 500 bp equispaced window, the chromosome is divided into 500 bp regions where the difference between the start and end region is exactly 500. But in my case, the first start of the region is the first position of the CpG site and first end of the region is the highest nucleotide position of the CpG site within the 500 sliding bp window. The next start of the region is the position of the CpG site which is just after the CpG nucleotide position of the preceding end region. I had to deal with the genomic positions of the CpG in a different manner than a conventional 500 bp window because the positions of the CpG sites are not equispaced. That is why I have used the sliding 500 bp technique to account for the contiguous CpG

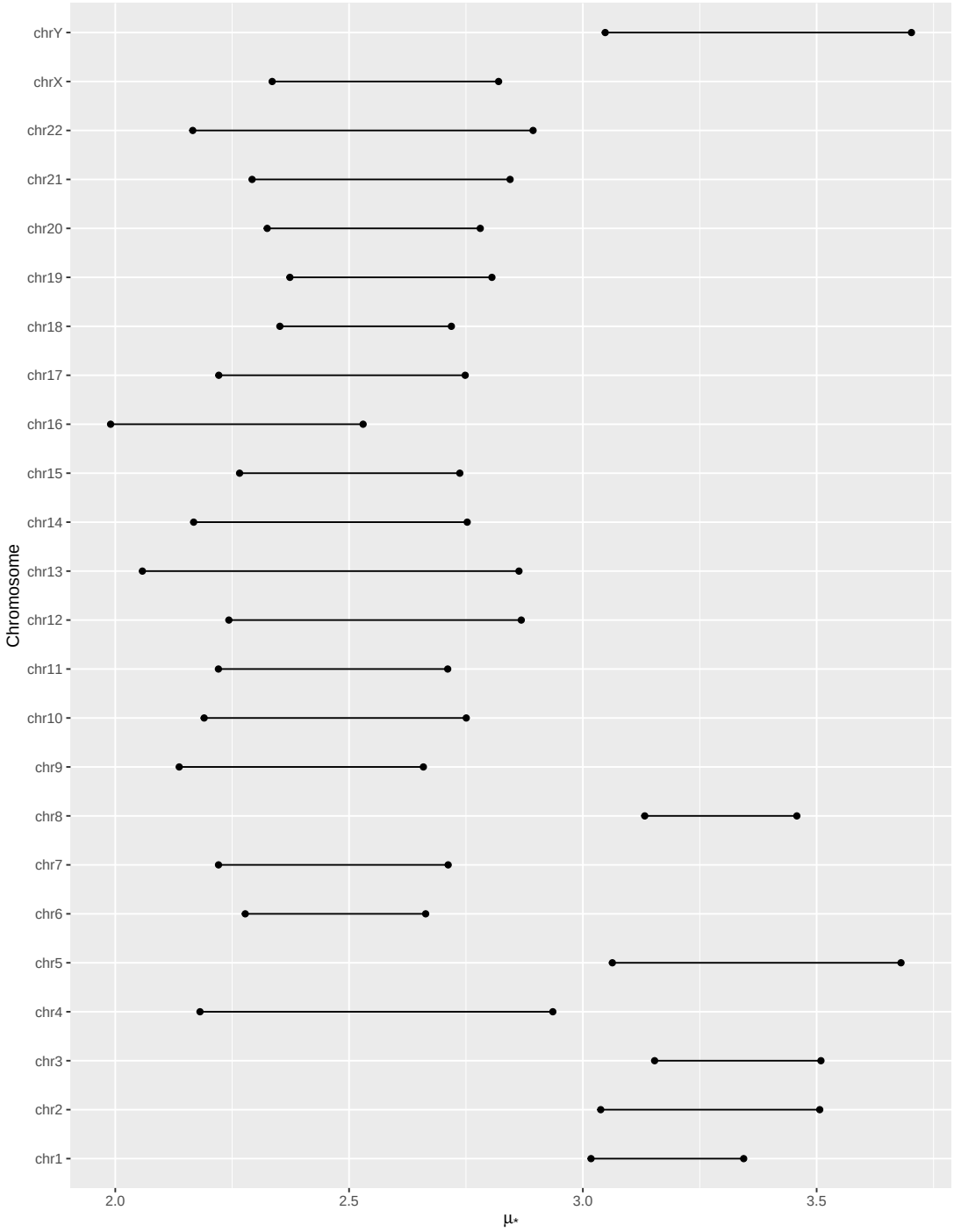


Figure 6.6: 95% horizontal posterior credible interval plots for WAIC picked model parameter μ_*. x-axis: range of the credible intervals; y-axis: Chromosome.

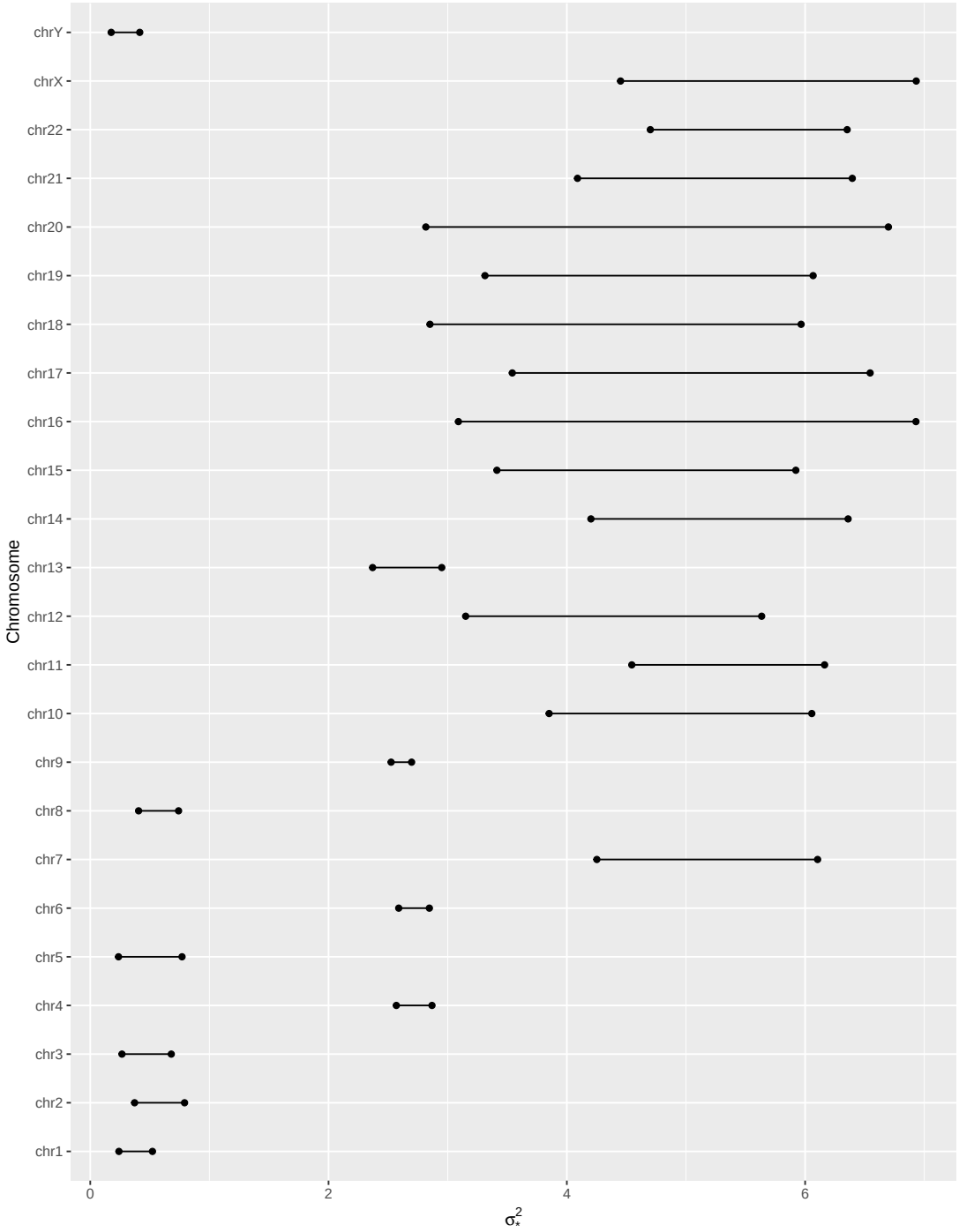


Figure 6.7: 95% horizontal posterior credible interval plots for WAIC picked model parameter σ_*^2. x-axis: range of the credible intervals; y-axis: Chromosome.

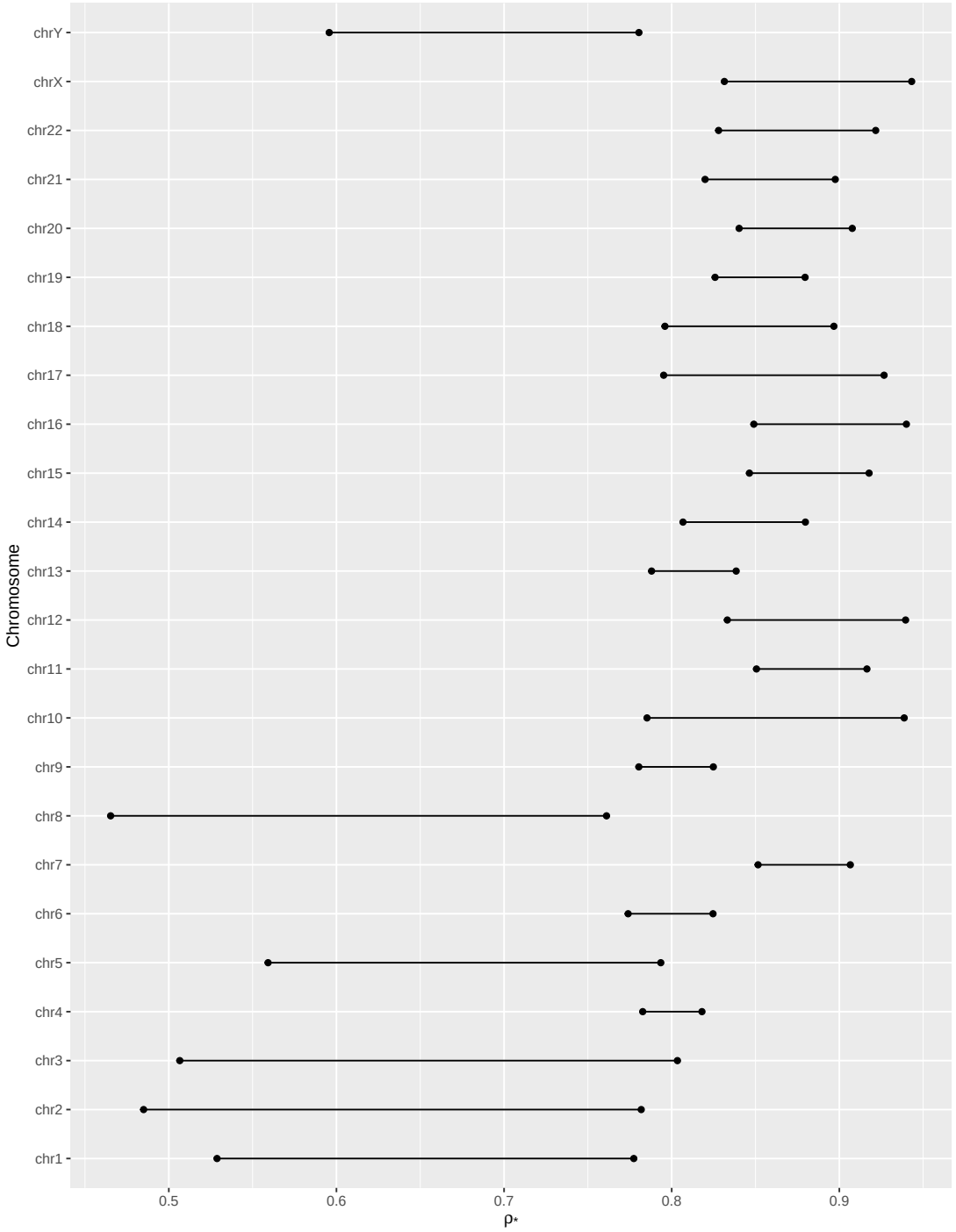


Figure 6.8: 95% horizontal posterior credible interval plots for WAIC picked model parameter ρ_*. x-axis: range of the credible intervals; y-axis: Chromosome.

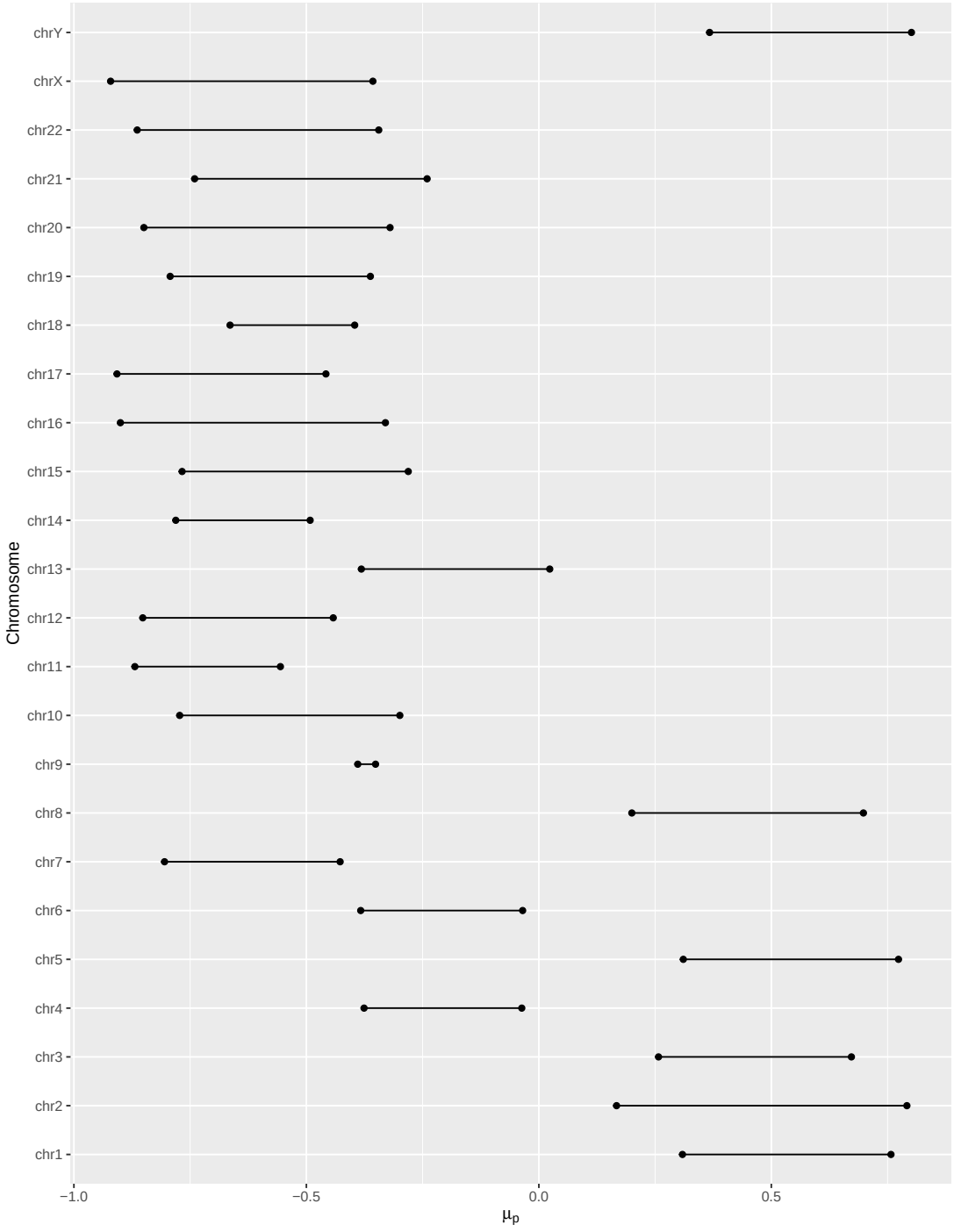


Figure 6.9: 95% horizontal posterior credible interval plots for WAIC picked model parameter μ_p. x-axis: range of the credible intervals; y-axis: Chromosome.

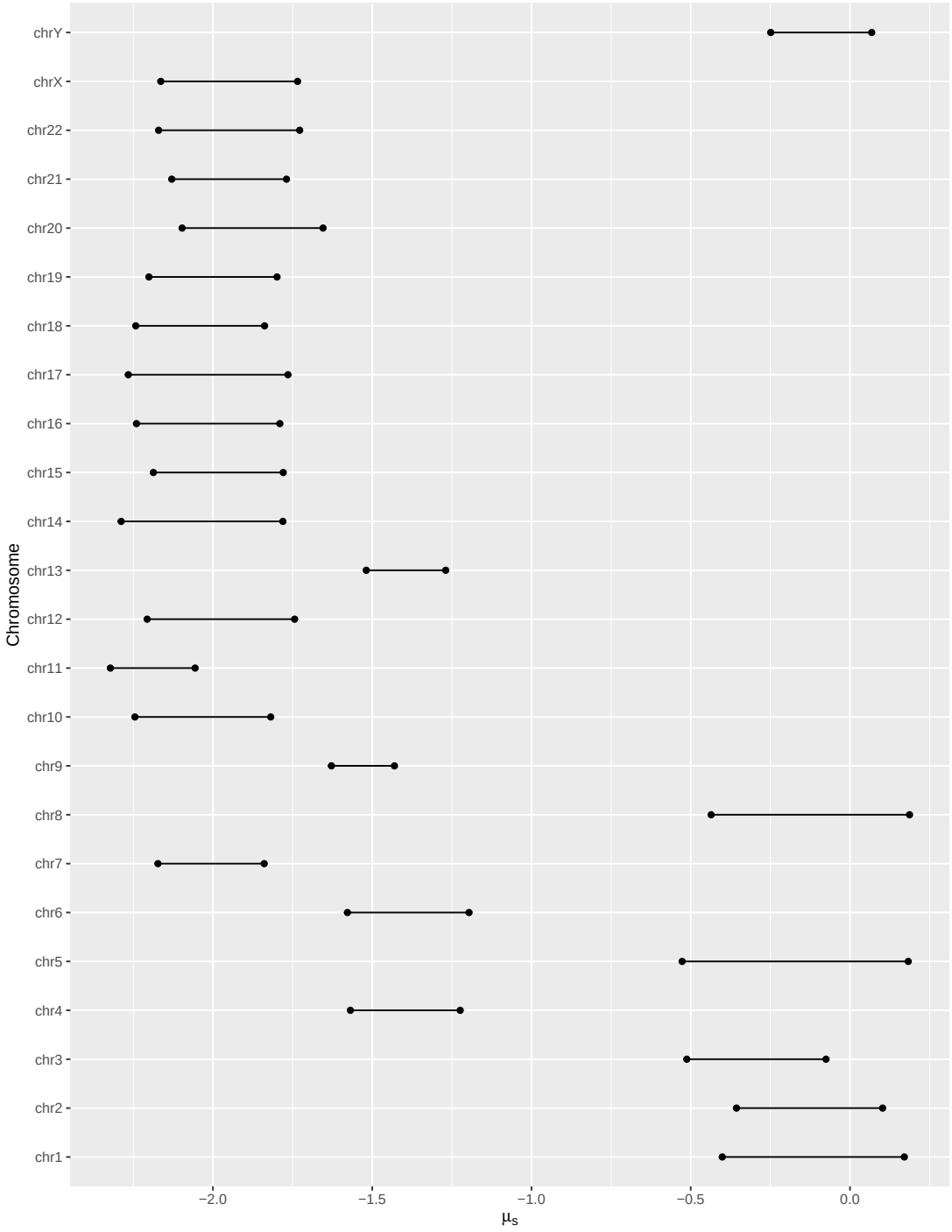


Figure 6.10: 95% horizontal posterior credible interval plots for WAIC picked model parameter μ_s. x-axis: range of the credible intervals; y-axis: Chromosome.

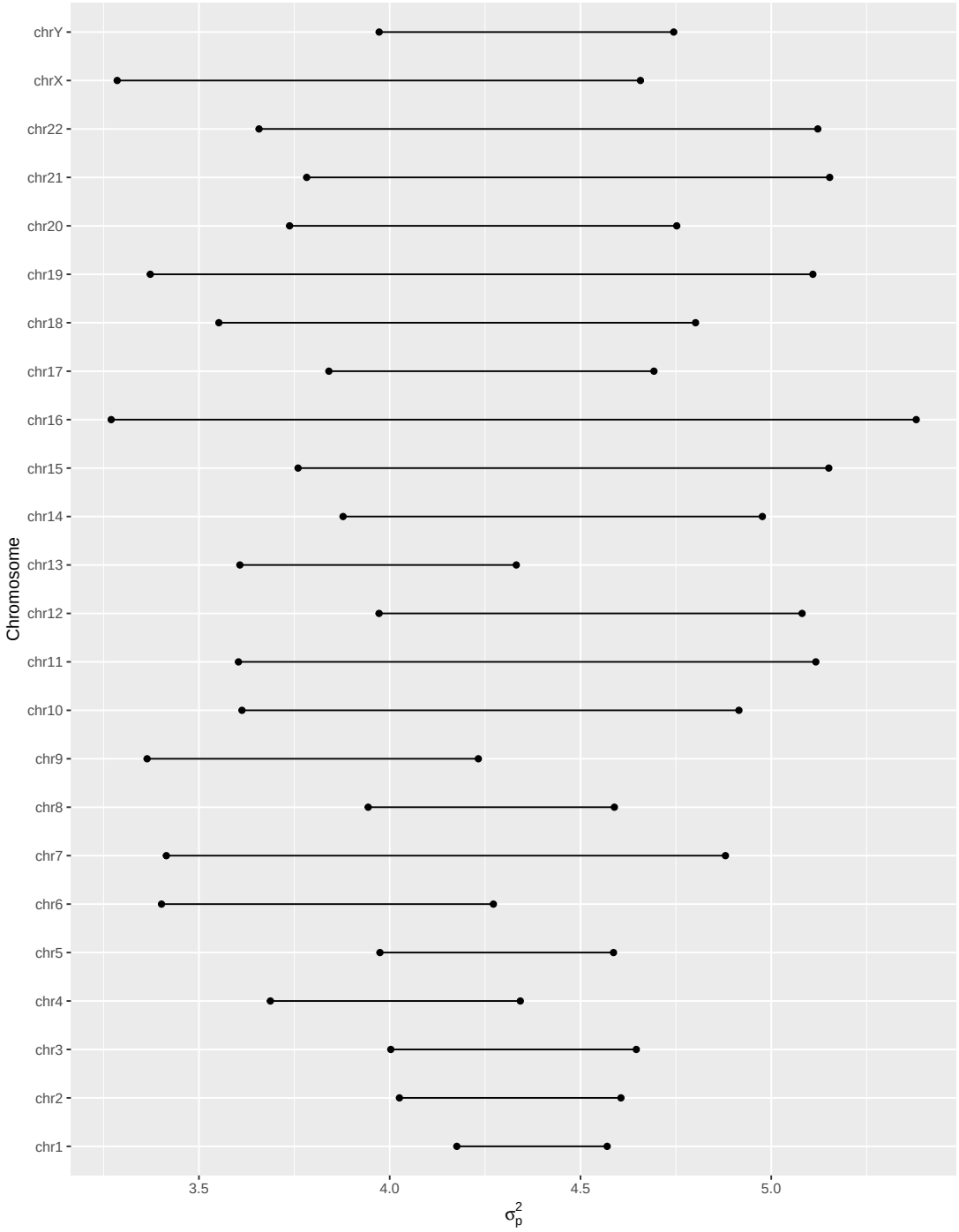


Figure 6.11: 95% horizontal posterior credible interval plots for WAIC picked model parameter σ_p^2. x-axis: range of the credible intervals; y-axis: Chromosome.

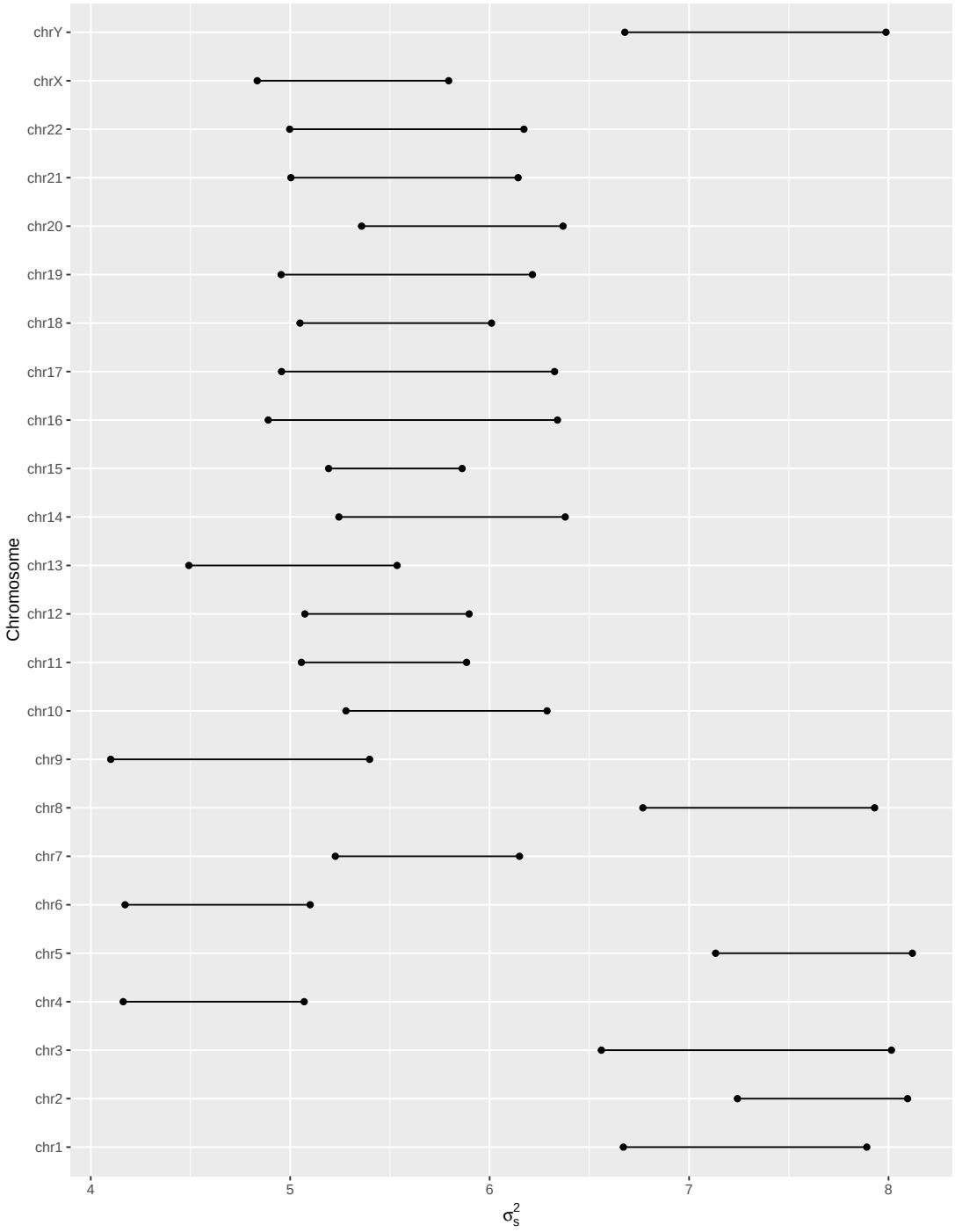


Figure 6.12: 95% horizontal posterior credible interval plots for WAIC picked model parameter σ_s^2. x-axis: range of the credible intervals; y-axis: Chromosome.

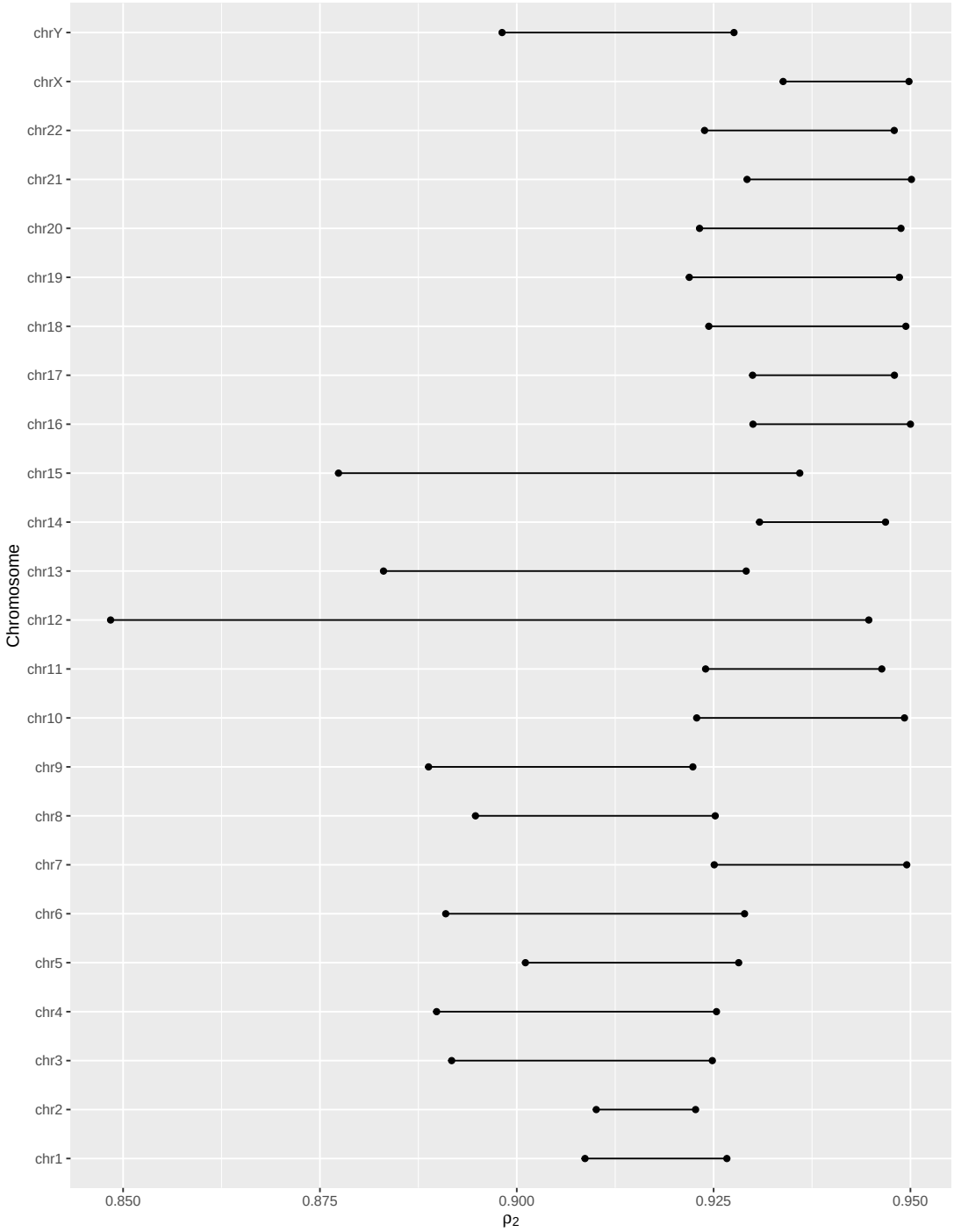


Figure 6.13: 95% horizontal posterior credible interval plots for WAIC picked model parameter ρ_2. x-axis: range of the credible intervals; y-axis: Chromosome.

sites.

Furthermore, to identify a DMR, the proportion of CpG sites identified as DMCs in a 500 bp region must exceed the threshold value of 0.5. Similarly, I also classified SMRs, such that, if the proportion of CpG sites identified as DMCs in a 500 bp region is less than the threshold value of 0.5, then I term the region to be SMR.

In addition, I further classified DMRs into partial DMRs (pDMRs) and strong DMRs (sDMRs) and SMRs into partial SMRs (pSMRs) and strong SMRs (sSMRs). They are described as follows:

- If the proportion of DMCs in a 500 bp region is greater than or equal to 0.8, then I call the region to be sDMR.
- If the proportion of DMCs in a 500 bp region lies between 0.5 and 0.8, I term the region to be pDMR.
- If the proportion of DMCs in a 500 bp region is less than or equal to 0.2, then I call the region to be sSMR.
- If the proportion of DMCs in a 500 bp region lies between 0.2 and 0.5, I term the region to be pSMR.

Tables 6.4 and 6.5 show the different classes of DMRs and SMRs identified WAIC picked model for each chromosome.

6.6 Computational time

The HMMmethState method was coded in R and C++ and run on a Linux machine with a 2.50 GHz processor. The MCMC simulations of my proposed method reach convergence within $10,000$ iterations, so I burn-in the first $3,000$

WAIC picked(Chromosome)	Methylation level	DMR	SMR
NLBCM(chr1)	p	18414	24257
	s	177067	124898
NLBCM(chr2)	p	20541	33724
	s	101018	63719
NLBCM(chr3)	p	7694	9367
	s	245441	30859
NLBDM(chr4)	p	9359	11474
	s	151089	104712
NLBCM(chr5)	p	30745	173739
	s	12713	48127
NLBCM(chr6)	p	14076	42259
	s	108328	88176
NLBDM(chr7)	p	7481	8103
	s	97295	124600
NLBCM(chr8)	p	1805	23140
	s	8619	183256
NLBDM(chr9)	p	7083	8867
	s	93010	69436
NLBDM(chr10)	p	8352	9957
	s	89284	95384
NLBDM(chr11)	p	5857	6583
	s	82848	106059
NLBDM(chr12)	p	8114	10219
	s	94488	88007

Table 6.4: DMR identified by WAIC picked HMMmethState model for each chromosome (Chromosome-1-12).

samples and thin at every 10^{th} iteration. Although it takes a longer time compared to other methods like DSS and methylKit, HMMmethState mostly achieves higher accuracy of DMC identification than other methods due to its robustness that allows for spatial genomic dependence over the genomic positions of the CpG sites. Different HMMmethState methods take different computation times.

WAIC picked(Chromosome)	Methylation level	DMR	SMR
NLBDM(chr13)	p	5192	6727
	s	77864	53668
NLBDM(chr14)	p	4648	5788
	s	63939	61002
NLBDM(chr15)	p	5188	6864
	s	55909	58681
NLBDM(chr16)	p	4804	5068
	s	57117	60326
NLBDM(chr17)	p	6634	9037
	s	6634	9037
NLBD(chr18)	p	3949	5068
	s	60795	44492
NLBDM(chr19)	p	5062	6675
	s	19723	63415
NLBDM(chr20)	p	3329	4351
	s	55359	32707
NLBDM(chr21)	p	2387	2845
	s	27558	22041
NLBDM(chr22)	p	4051	4568
	s	22224	27621
NLBDM(chrX)	p	9021	59678
	s	88552	61281
NLBCM(chrY)	p	543	1233
	s	1212	1579

Table 6.5: DMR identified by WAIC picked HMMmethState model for each chromosome (Chromosome-13-22,X,Y).

For a chromosome with approximately 1.8 million CpG sites, NLBDM takes to run approximately 76 hours, whereas, NLBCM takes around 109 hours. BBDM and BBCM take approximately 67 hours and 92 hours respectively with proper MCMC convergence. Besides, the computational time of all the HMMmethState models are insignificant compared to the time and resource required to perform

experiments to obtain BS-seq data by biologists. In addition, for a large memory, say 32 GB, and 40 cores, the HMMmethState analysis can be run in parallel computing for individual chromosomes to save the computational cost.

6.7 Summary

In this Chapter, I conducted a thorough investigation of the features of my proposed HMMmethState models and justified their strength and limitations in identifying DMCs in BS-seq data. I assessed my models and showed that the Normal-logit-Binomial emission model adequately fits the data and that the correlation between the methylated counts of proliferating and senescent cells cannot be explained by the Beta-Binomial emission model irrespective of the transition models. This claim can further be corroborated with the results of the estimates of *WAIC*. I have also examined the reliability of my results where HMMmethState models are applied to both simulated and real data and simultaneously compare their performances with existing differential methylation caller methods. The differential methylation identification methods are based on certain model assumptions and they have their own advantages and disadvantages. The performance study of these methods even using an independent simulator could be a matter of dispute as the performance can be tilted towards the true base model or a model similar to the true base model. Thus, it is hard to conclude that *HMMmethState* models perform better than *DSS* and *methylKit* as there is no available gold standard BS-seq methylation dataset (training, test and validation) where the methylation status of each CpG site is known.

In an attempt to assess the performance of my proposed method- HMMmethState, I implemented a simulation study design based on reasonable model of the underlying process. Through simulated datasets, I compare the performance of each of the HMMmethState models with two popular methods, illustrating the

reliability of my method in the identification of DMCs. I have also conducted detailed investigations of the features of the models and justified their ability to identify DMCs in BS-seq data, specific to the chromosomal datasets. As a first step, I checked the MCMC simulations of the hidden states and the model parameters converged to the stationary posterior distribution to ensure the reliability of my estimates. I then chose the chromosomal data specific models based on WAIC model selection criterion, which are then used for further analyses, i.e., DMC prediction.In addition, I applied two DMC callers to the same datasets for comparing their results to my results. Since the true DMCs are not known for these datasets, I concluded that the DMCs identified by my method were reliable.

Chapter 7

Conclusions and Further Work

The key contribution of this thesis has been to develop models that can identify DMCs in the BS-seq data. I propose HMMmethState, a method based on Bayesian hierarchical HMMs for identifying DMCs between proliferating and senescent cells for BS-seq methylation data. My proposed approach also employs hierarchical HMMs to account for the spatial dependence among the CpG sites based on their genomic positions. The HMMmethState models can also be applied to any other sequencing experiment of two treatment groups. In this chapter, I highlight my thesis contributions and then provide a brief outline of some possible directions that can be implemented as a basis for further research.

7.1 Contributions of this thesis

The thesis contributions can be catergorized into two parts and they are as follows:

1. Methodological advances: In Section 7.1.1, I discuss the main goals I have achieved in my methodological work and the importance of the implementation of HMMmethState models in detecting DMCs in BS-seq data.

2. Biological Advances: In Section 7.1.2, I also discuss biological advantages in my methodology that offers improved performance over other existing

methods.

7.1.1 Methodological advances

The primary objective of this thesis was to create HHMMs for BS-seq methylation data within a Bayesian framework. To this end, I have examined four HHMMs with state-dependent emission distributions for methylated counts, given the methylation status of the CpG sites. I have further shown how the positional variations of the CpG sites can be incorporated in the methylation state of the CpG sites to account for the spatial dependence between the genomic positions of the CpG sites.

7.1.1.1 The HMMmethState method

Taking a broad view, the main contributions of Chapters 4 and 5 has been towards an improved understanding of the potential HHMMs to assess the methylation data from BS-seq experiments and also critically examining the strengths and limitations of the HMMmethState models.

In particular, Chapters 4, 5 and 6 explain the significance of developing and comparing the four versions of HMMmethState- *BBDM*, *BBCM*, *NLBDM*, *NLBCM* for analysing BS-seq methylation data. The four versions of the HMMmethState models were implemented by combining the emission and transition models as shown in Table 7.1. The models ED and EC, [where E: BB, NLB] can be distinguished by their transition models as discussed in Sections 4.1.3 and 4.1.4 respectively. The only alteration required for EC from ED is to assume that the status of DMC is represented by an unobservable Markov chain instead of an unobservable discrete Markov chain. The sole idea behind implementing EC is to capture the positional variations of CpG dinucleotide bases and whether it has any significant effect over ED in detecting the DMCs. The proposed approach of

HMMmethState models are described in details in Chapters 4 and 5.

In Chapter 4, I proposed the first two models of the HMMmethState framework-

Emission Model	BB	NLB
Transition model		
D	BBDM	NLBDM
C	BBCM	NLBCM

Table 7.1: Description of HMMmethState models.

BBDM and BBCM, as shown in Figure 4.1. In Section 4.1.2, I proposed Beta-Binomial emission models and subsequently combined with transition model T: C, D for the implementations of BBDM and BBCM. The reason I model the methylated counts using a Binomial emission distribution at the first stage of the hierarchical model is due to the process of BS-seq which subsequently involves the random sampling of methylated and unmethylated reads.The underlying true methylation proportions (2^{nd} stage of the hierarchical model) are assumed to follow a Beta distribution. In order to accelerate computational simplicity, Beta-Binomial emission distribution becomes a natural choice with collapsed distributional structure due to Beta-Binomial conjugacy. However, there is substantive potential for improvement in the structure of BBDM and BBCM, especially on emission probability functions that eventually play a key role in computing the likelihood functions.

To improve upon my emission model, I develop a hierarchical emission model that considers correlation between proliferating and senescent methylated proportions. From visual posterior predictive checks, it has been observed that there is strong evidence of correlation between proliferating and senescent methylated proportions. I develop a hierarchical bivariate Normal-Binomial emission model to account for the correlation in the bivariate underlying true methylation proportions in Chapter 5 and subsequently combined with transition model T: C, D for

the implementations of *NLBDM* and *NLBCM*. Again, the primary structure (1^{st} stage of the hierarchical model) of the model, i.e., the methylated counts follow a Binomial distribution remains the same as in Chapter 4. I modify the underlying true methylation proportions as functions of logit variables for each CpG site, which ultimately act as auxiliary parameters. Unlike Beta-Binomial conjugacy, the Bayesian Bivariate Normal-Binomial emission model does not have a collapsed structure, thus it involves computational complexity in estimating the emission hyperparameters and auxiliary parameters. Furthermore, to perform parameter estimation for my models, I implement efficient MCMC based algorithms. In Chapters 4 and 5, I examine the convergence properties of the posterior distributional quantities for simulation and real studies.

7.1.1.2 Significance of transition model for model comparison

In Chapter 6, I have done an extensive analysis on model comparison. The Normal-logit-Binomial emission model outperformed the Beta-Binomial emission model in most real datasets. Furthermore, in Chapter 6, I have also observed that a particular chromosomal dataset is modelled by either *NLBDM* or *NLBCM* depending on the effect of positional variations among CpG sites. Even though the spatial dependence assumption is taken into account by considering Markovian dependence over the latent states, the effect of positional variations in identifying DMCs can only be observed in a particular dataset. I have only selected the models based on WAIC computations as it has been specifically formulated for hierarchical or mixture models. WAIC appeared to perform consistently well compared to two different DIC versions in simulation studies as described in Section 6.1.1.

7.1.2 Biological Advances

In Chapter 6, I compare the performances of my methods with existing methods for detecting DMCs/DMRs. I subsequently illustrate the advantages of HMM-

methState by appying to simulated data and comparing it with two of the most popular packages (R/Bioconductor packages) *DSS* and *methylKit*. I demonstrate how the HMMmethState based algorithms outperform the existing methods in simulation studies in terms of sensitivity and specificity. In addition, I have also applied HMMmethState to a published dataset (Cruickshanks et al., 2013) and presented my findings. I presented the results of DMCs and DMRs obtained using my methods, i.e., the results of DMCs/DMRs with the proposed HMMmethState that have been applied to the BS-seq datasets.

The main biological contributions of HMMmethState can be explained as follows:

1. It can robustly identify DMCs from BS-seq data.

2. It can automatically update DMRs based on the results on DMCs and can futher classify into pDMRs (partial DMRs) and sDMRs (strong DMRs) which can help biologists in better understanding of the functional genomic regions of interests.

3. It can also be applied to both whole-genome and targeted BS-seq methylation data.

The results of the HMMmethState models explain that I can certainly implement these methods under unconditioned settings to identify DMCs/DMRs for high-throughput BS-seq data. The predicted DMCs/DMRs can also help in understanding the phenotypic changes associated with human ageing.

7.2 Further Work

The HMMmethState models I developed and assessed in this thesis provide an efficient way of identifying DMCs in BS-seq data. However, there still remains

scope for improving the models which can work better in understanding the specific biological questions. In the following sections, I briefly outline the scope for further work in this area.

7.2.1 Bivariate Beta-Binomial correlated emission distribution

The idea behind choosing and constructing a Bivariate Beta-Binomial distribution is to induce correlation between proliferating and senescent methylation proportion parameters as it is theoretically very complicated to construct a Bivariate Binomial distribution with a correlation parameter. I construct a Bivariate Beta distribution as a prior for two correlated proportions from the Bivariate Binomial distribution. The approach used in Chapter 4 can be modified to account for the correlation within paired samples. The bivariate Beta distribution can be assumed as a prior distribution on proliferating methylation proportions and senescent methylation proportions. Here, I use Variable-in-common and transformation-based constructions as explained by (Olkin and Trikalinos, 2015) and (Oleson, 2010).

The joint full conditional distribution of $(p_t^p,\ p_t^s)$ for $t = 1, \ldots, T$ is given by,

$$p(p_t^p,\ p_t^s|.) \propto \frac{(p_t^p)^{x_t^p+a-1}(1-p_t^p)^{n_t^p-x_t^p+b+c-1}(p_t^s)^{x_t^s+b-1}(1-p_t^s)^{n_t^s-x_t^s+a+c-1}}{(1-p_t^p p_t^s)^{(a+b+c)}}. \quad (7.1)$$

This is the form of a generalized Beta distribution where $a,\ b,\ c\ >\ 0$.

It will be interesting to examine the results based on correlated Beta-Binomial emission model and whether it has the ability to outperform the HMMmethState models.

7.2.2 Ad hoc label-switching technique

In Chapters 4 and 5, I have efficiently implemented a relabelling algorithm which perform quite switfly in my augmented Gibbs sampler. However, I can also fix label switching by ordering the means in my prior specification. Another way of tackling this problem is by using informative priors, However in this case there is a limitation. If the priors I use are informative as well as exchangeable then they still might cause label switching. Thus in order to counter label switching, I impose informative prior constraints based on the nature of the hidden state labels on the parameters.

In a nutshell, informative priors can still cause label switching either due to the exchangeable properties or when the modes are not clearly separated in the model. Thus label switching problem can be tackled by using a constraint on the prior of the parameter.

In my approach, I use uninformative and exchangeable priors in my model parameters. I impose a constraint on the state 2 hyperparameters such that, if,

- For Chapter 4,

$$\left| \frac{\gamma_1}{(\gamma_1 + \delta_1)} - \frac{\gamma_2}{(\gamma_2 + \delta_2)} \right| < 0.008. \tag{7.2}$$

- For Chapter 5,

$$|\mu_p - \mu_s| < 0.35. \tag{7.3}$$

I swap the state labels in order to avoid label switching, i.e., I perform an online relabelling at every MCMC run. The reason I swap the labels is because from (7.2) and (7.3), it is evident that the Beta and Normal prior means of the methylation levels of the proliferating and senescent tend to be similar as the absolute

difference between them is getting closer to 0. In my assumption, I had already stated that state 2 indicates DMC, i.e, it is fair to assume that the proportions of methylation levels of proliferating and senescent must be significantly different and the proportions of these cells can only be reflected through their means or modes. However, the choice of this cut-off value varies from distribution to distribution and a lot of simulation experiments are required in order to choose the best cut-off value. The cut-off values are extremely sensitive even by a small margin. This kind of assumption can only be valid for 2 state labels in a HMM.

7.2.3 Merging contiguous DMCs

For practical situations, it might be desirable to summarize DMRs over tiling windows. For this reason, I have defined DMR windows by tiling vast genomic regions in Section 6.5.3. However, I can also create a new form of DMR. If the contiguous CpG sites are all DMCs, I can call it a DMR window. This kind of DMR windows might also be useful for biologists who wish to correlate information about gene-expression and differential methylation. In Figure 7.1, I have also presented an Integrative Genomics Viewer (IGV) snapshot of my DMRs. In this Figure 7.1, I merge the contiguous DMCs to form a new block of DMR. I also plan to extend HMMmethState by including other biological sources of dependence among CpG sites. For HMMmethState models, I have already considered the spatial dependence assumption among CpG sites based on genomic position. However, including other sources of biological variations like gene-expression information, promoter region, promoter-enhancer-promoter interactions might improve my current method in understanding the differential methylation pattern. I will also focus on extending HMMmethState to BS-seq data under general multiple experimental design. In addition, I am developing an R package which implements my proposed HMMmethState method.

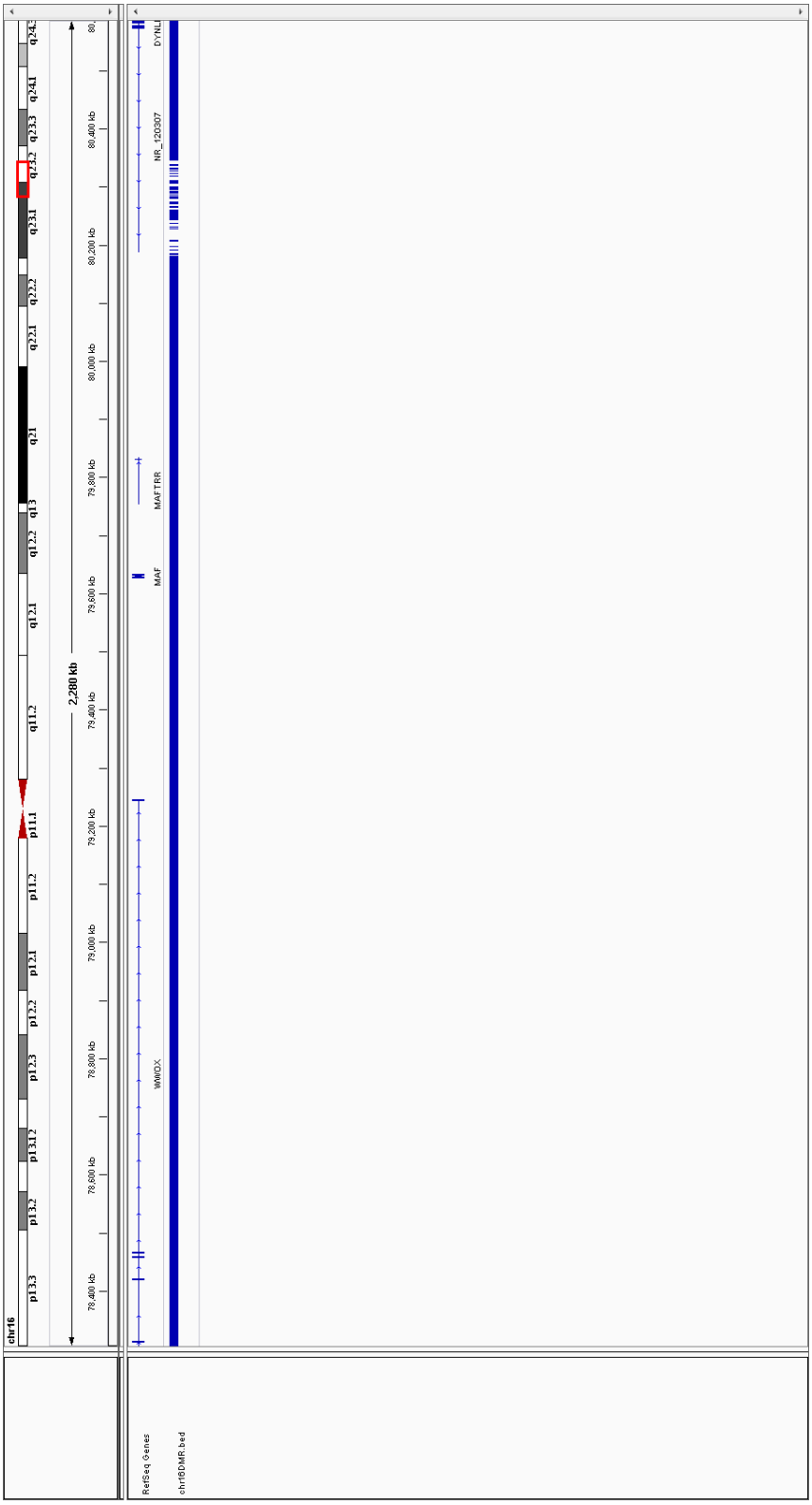


Figure 7.1: IGV snapshot of a segment of Chromosome 16, where the blue blocks corresponding to $chr16DMR.bed$ are DMRs detected using HMMmethState method.

Appendix A1

BBDM	Parameter	α	β	γ_1	δ_1	γ_2	δ_2
MCMC Chain 1	Posterior Mean	6.790	0.770	0.672	1.327	0.550	1.990
	Posterior S.D.	0.015	0.001	0.001	0.002	0.001	0.004
	95% Credible Interval	(6.760, 6.819)	(0.768, 0.772)	(0.670, 0.674)	(1.323, 1.332)	(0.548, 0.552)	(1.982, 1.998)
MCMC Chain 2	Posterior Mean	6.791	0.770	0.672	1.327	0.550	1.990
	Posterior S.D.	0.015	0.001	0.001	0.002	0.001	0.004
	95% Credible Interval	(6.761, 6.823)	(0.768, 0.772)	(0.670, 0.674)	(1.322, 1.332)	(0.548, 0.552)	(1.981, 1.998)
MCMC Chain 3	Posterior Mean	6.790	0.770	0.672	1.327	0.550	1.990
	Posterior S.D.	0.014	0.001	0.001	0.002	0.001	0.004
	95% Credible Interval	(6.762, 6.819)	(0.768, 0.772)	(0.670, 0.674)	(1.322, 1.332)	(0.548, 0.552)	(1.982, 1.998)
Convergence Diagnostics	PSRF factor	1.004	1.002	1.001	1.000	1.000	1.001
BBCM	Parameter	α	β	γ_1	δ_1	γ_2	δ_2
MCMC Chain 1	Posterior Mean	0.443	0.301	3.110	1.645	0.760	0.790
	Posterior S.D.	0.001	0.000	0.009	0.006	0.002	0.003
	95% Credible Interval	(0.441, 0.444)	(0.300, 0.301)	(3.092, 3.127)	(1.634, 1.656)	(0.757, 0.763)	(0.784, 0.796)
MCMC Chain 2	Posterior Mean	0.443	0.301	3.110	1.645	0.760	0.790
	Posterior S.D.	0.001	0.000	0.009	0.006	0.002	0.003
	95% Credible Interval	(0.442, 0.444)	(0.300, 0.302)	(3.093, 3.126)	(1.633, 1.657)	(0.756, 0.764)	(0.784, 0.796)
MCMC Chain 3	Posterior Mean	0.443	0.301	3.110	1.645	0.760	0.790
	Posterior S.D.	0.001	0.000	0.009	0.006	0.002	0.003
	95% Credible Interval	(0.441, 0.444)	(0.300, 0.301)	(3.091, 3.127)	(1.634, 1.656)	(0.757, 0.763)	(0.784, 0.795)
Convergence Diagnostics	PSRF factor	1.000	1.003	1.000	1.001	1.007	1.003

Table 2: Posterior summaries of emission hyperparameters for real data study.

	Model	BBDM			BBCM	
	Parameter	π_1	τ_{11}	τ_{21}	λ_1	λ_2
MCMC Chain 1	Posterior Mean	0.550	0.940	0.071	0.250	0.557
	Posterior S.D.	0.0005	0.0006	0.0003	0.001	0.002
	95% Credible Interval	(0.549, 0.551)	(0.940, 0.940)	(0.070, 0.072)	(0.248, 0.252)	(0.553, 0.561)
MCMC Chain 2	Posterior Mean	0.550	0.940	0.071	0.250	0.557
	Posterior S.D.	0.0005	0.0006	0.0003	0.001	0.002
	95% Credible Interval	(0.549, 0.551)	(0.940, 0.940)	(0.070, 0.072)	(0.248, 0.252)	(0.553, 0.561)
MCMC Chain 3	Posterior Mean	0.550	0.940	0.071	0.250	0.557
	Posterior S.D.	0.0005	0.0006	0.0003	0.001	0.002
	95% Credible Interval	(0.549, 0.551)	(0.940, 0.940)	(0.071, 0.071)	(0.248, 0.252)	(0.553, 0.561)
Convergence Diagnostics	PSRF factor	1	1	1	1	1

Table 3: Posterior summaries of transition parameters for real data study.

NLBDM	Parameter	μ^*	σ_*^2	ρ^*
MCMC Chain 1	Posterior Mean	2.371	4.214	0.877
	Posterior S.D.	0.094	0.684	0.017
	95% Credible Interval	(2.167 2.518)	(3.150 5.742)	(0.851 0.915)
MCMC Chain 2	Posterior Mean	2.362	4.153	0.879
	Posterior S.D.	0.100	0.678	0.019
	95% Credible Interval	(2.144 2.521)	(3.142 5.611)	(0.852 0.919)
MCMC Chain 3	Posterior Mean	2.363	4.191	0.879
	Posterior S.D.	0.093	0.684	0.018
	95% Credible Interval	(2.177 2.513)	(3.147 5.552)	(0.851 0.917)
Convergence Diagnostics	PSRF factor	1.006	1.000	1.001
NLBCM	Parameter	μ^*	σ_*^2	ρ^*
MCMC Chain 1	Posterior Mean	2.742	2.061	0.759
	Posterior S.D.	0.052	0.049	0.014
	95% Credible Interval	(2.638 2.835)	(1.973 2.160)	(0.734 0.785)
MCMC Chain 2	Posterior Mean	2.743	2.066	0.760
	Posterior S.D.	0.056	0.050	0.014
	95% Credible Interval	(2.633 2.841)	(1.980 2.168)	(0.734 0.786)
MCMC Chain 3	Posterior Mean	2.743	2.064	0.760
	Posterior S.D.	0.055	0.053	0.013
	95% Credible Interval	(2.630 2.838)	(1.973 2.168)	(0.734 0.786)
Convergence Diagnostics	PSRF factor	1.000	1.003	1.000

Table 4: Posterior summaries of state 1 emission hyperparameters for real data study.

NLBDM	Parameter	μ_p	μ_s	σ_p^2	σ_s^2	ρ_2
MCMC Chain 1	Posterior Mean	-0.703	-2.040	4.556	5.553	0.940
	Posterior S.D.	0.103	0.095	0.411	0.291	0.005
	95% Credible Interval	(-0.878 -0.491)	(-2.212 -1.853)	(3.696 5.263)	(5.008 6.129)	(0.931 0.949)
MCMC Chain 2	Posterior Mean	-0.700	-2.034	4.562	5.549	0.940
	Posterior S.D.	0.111	0.098	0.407	0.316	0.005
	95% Credible Interval	(-0.875 -0.467)	(-2.213 -1.852)	(3.763 5.288)	(5.013 6.152)	(0.931 0.949)
MCMC Chain 3	Posterior Mean	-0.706	-2.025	4.554	5.526	0.940
	Posterior S.D.	0.112	0.100	0.425	0.318	0.005
	95% Credible Interval	(-0.887 -0.453)	(-2.215 -1.825)	(3.668 5.259)	(4.958 6.171)	(0.931 0.949)
Convergence Diagnostics	PSRF factor	1.003	1.003	1.002	1.001	1.001
NLBCM	Parameter	μ_p	μ_s	σ_p^2	σ_s^2	ρ_2
MCMC Chain 1	Posterior Mean	-0.543	-1.563	4.767	5.670	0.909
	Posterior S.D.	0.125	0.146	0.270	0.568	0.007
	95% Credible Interval	(-0.800 -0.323)	(-1.855 -1.292)	(4.310 5.362)	(4.615 6.757)	(0.895 0.919)
MCMC Chain 2	Posterior Mean	-0.552	-1.562	4.740	5.657	0.908
	Posterior S.D.	0.131	0.166	0.260	0.552	0.006
	95% Credible Interval	(-0.806 -0.311)	(-1.888 -1.264)	(4.306 5.277)	(4.684 6.774)	(0.895 0.919)
MCMC Chain 3	Posterior Mean	-0.558	-1.571	4.746	5.646	0.908
	Posterior S.D.	0.133	0.160	0.272	0.544	0.007
	95% Credible Interval	(-0.817 -0.314)	(-1.887 -1.277)	(4.301 5.302)	(4.668 6.721)	(0.895 0.919)
Convergence Diagnostics	PSRF factor	1.001	1.003	1.003	1	1.001

Table 5: Posterior summaries of state 2 emission hyperparameters for real data study.

	Model	NLBDM			NLBCM	
	Parameter	π_1	τ_{11}	τ_{21}	λ_1	λ_2
MCMC Chain 1	Posterior Mean	0.594	0.985	0.018	0.250	0.557
	Posterior S.D.	0.018	0.000	0.002	0.005	0.012
	95% Credible Interval	(0.563 0.631)	(0.985 0.986)	(0.014 0.022)	(0.241 0.259)	(0.534 0.578)
MCMC Chain 2	Posterior Mean	0.594	0.985	0.018	0.250	0.558
	Posterior S.D.	0.018	0.000	0.002	0.005	0.012
	95% Credible Interval	(0.563 0.631)	(0.985 0.986)	(0.014 0.022)	(0.241 0.259)	(0.535 0.578)
MCMC Chain 3	Posterior Mean	0.592	0.985	0.018	0.250	0.558
	Posterior S.D.	0.019	0.000	0.002	0.005	0.012
	95% Credible Interval	(0.563 0.634)	(0.985 0.986)	(0.014 0.022)	(0.241 0.259)	(0.534 0.578)
Convergence Diagnostics	PSRF factor	1	1.007	1.001	1	1.000

Table 6: Posterior summaries of transition parameters for real data study.

Model	Emission hyperparameters								Transition parameters		
BBDM	α	β	γ_1	δ_1	γ_2	δ_2			π_1	τ_{11}	τ_{21}
	0.088	0.071	0.050	0.093	0.066	1.091			0.086	0.062	0.067
BBCM	α	β	γ_1	δ_1	γ_2	δ_2				λ_1	λ_2
	0.132	0.121	0.080	1.720	0.873	0.174				0.092	0.106
NLBDM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2	π_1	τ_{11}	τ_{21}
	0.081	0.056	0.030	0.076	1.026	0.042	0.071	0.053	0.075	0.063	0.044
NLBCM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2		λ_1	λ_2
	0.080	0.132	0.091	1.660	0.142	0.176	0.552	0.198		0.124	0.121

Table 7: Simulation study: RMSE values of the parameters for *moderately overlapped* case.

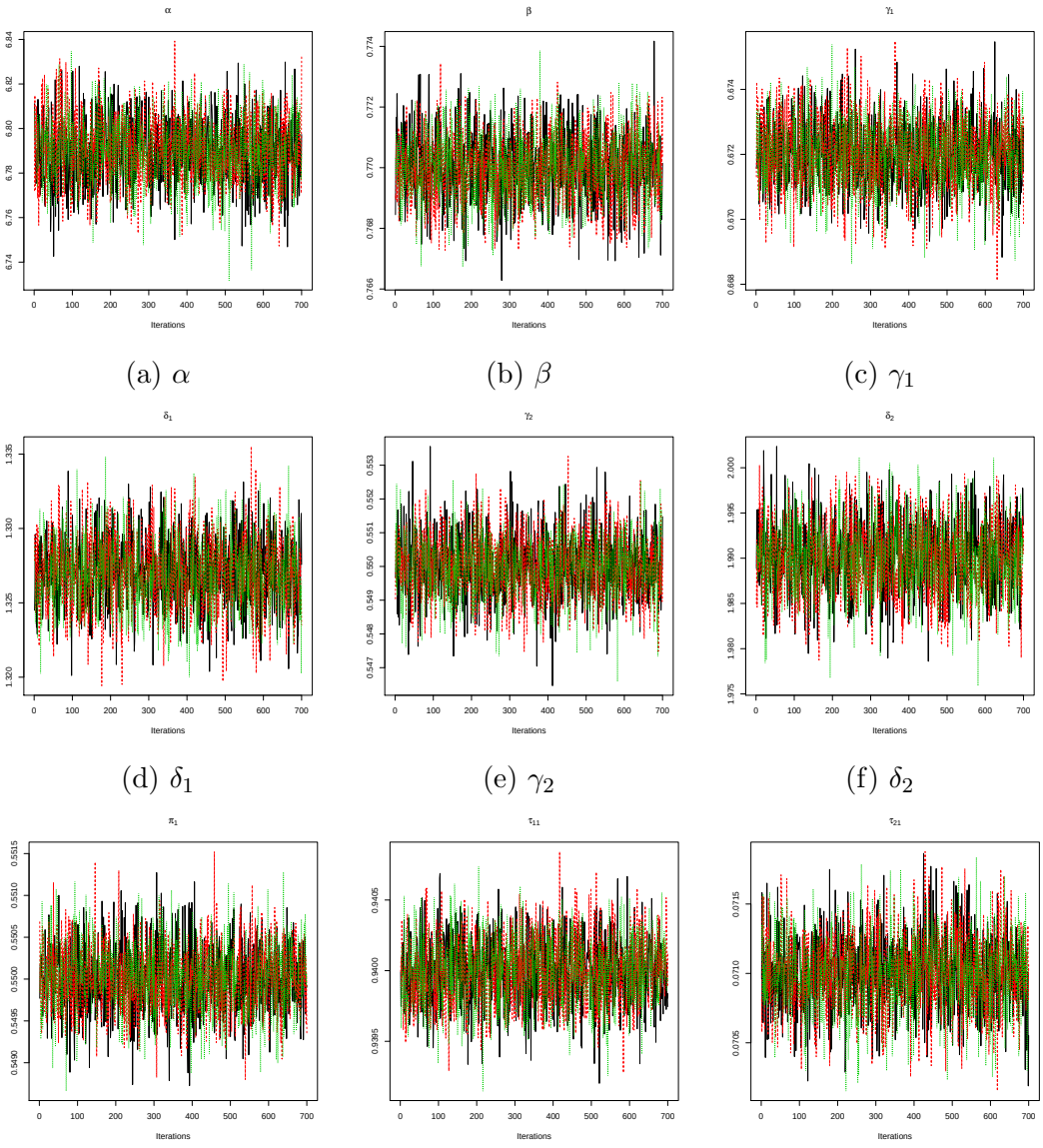
Model	Emission hyperparameters								Transition parameters		
BBDM	α	β	γ_1	δ_1	γ_2	δ_2			π_1	τ_{11}	τ_{21}
	0.0004	0.0003	0.0002	0.0006	0.0002	0.0090			0.0004	0.0003	0.0002
BBCM	α	β	γ_1	δ_1	γ_2	δ_2				λ_1	λ_2
	0.0007	0.0006	0.0100	0.0008	0.0006	0.0082				0.0009	0.00087
NLBDM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2	π_1	τ_{11}	τ_{21}
	0.0002	0.0006	0.0001	0.0003	0.0009	0.0048	0.0032	0.0050	0.0002	0.0018	0.0009
NLBCM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2		λ_1	λ_2
	0.0005	0.0006	0.0004	0.0081	0.0047	0.0019	0.00067	0.0009		0.0090	0.0083

Table 8: Simulation study: RMSE values of the parameters for *well separated* case.

Model	Emission hyperparameters								Transition parameters		
BBDM	α	β	γ_1	δ_1	γ_2	δ_2			π_1	τ_{11}	τ_{21}
	0.0087	0.0091	0.0006	0.0072	0.0039	0.0093			0.0008	0.0006	0.0015
BBCM	α	β	γ_1	δ_1	γ_2	δ_2				λ_1	λ_2
	0.0042	0.0017	0.0008	0.0090	0.0067	0.0009				0.0075	0.0042
NLBDM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2	π_1	τ_{11}	τ_{21}
	0.0073	0.0052	0.0005	0.0098	0.0034	0.0061	0.0008	0.0042	0.0009	0.0028	0.0054
NLBCM	μ_*	${\sigma^2}_*$	ρ_*	μ_p	μ_s	${\sigma^2}_p$	${\sigma^2}_s$	ρ_2		λ_1	λ_2
	0.0052	0.0074	0.0006	0.0018	0.0008	0.0006	0.0079	0.0086		0.0022	0.0046

Table 9: Simulation study: RMSE values of the parameters for *realistic* case.

Appendix A2

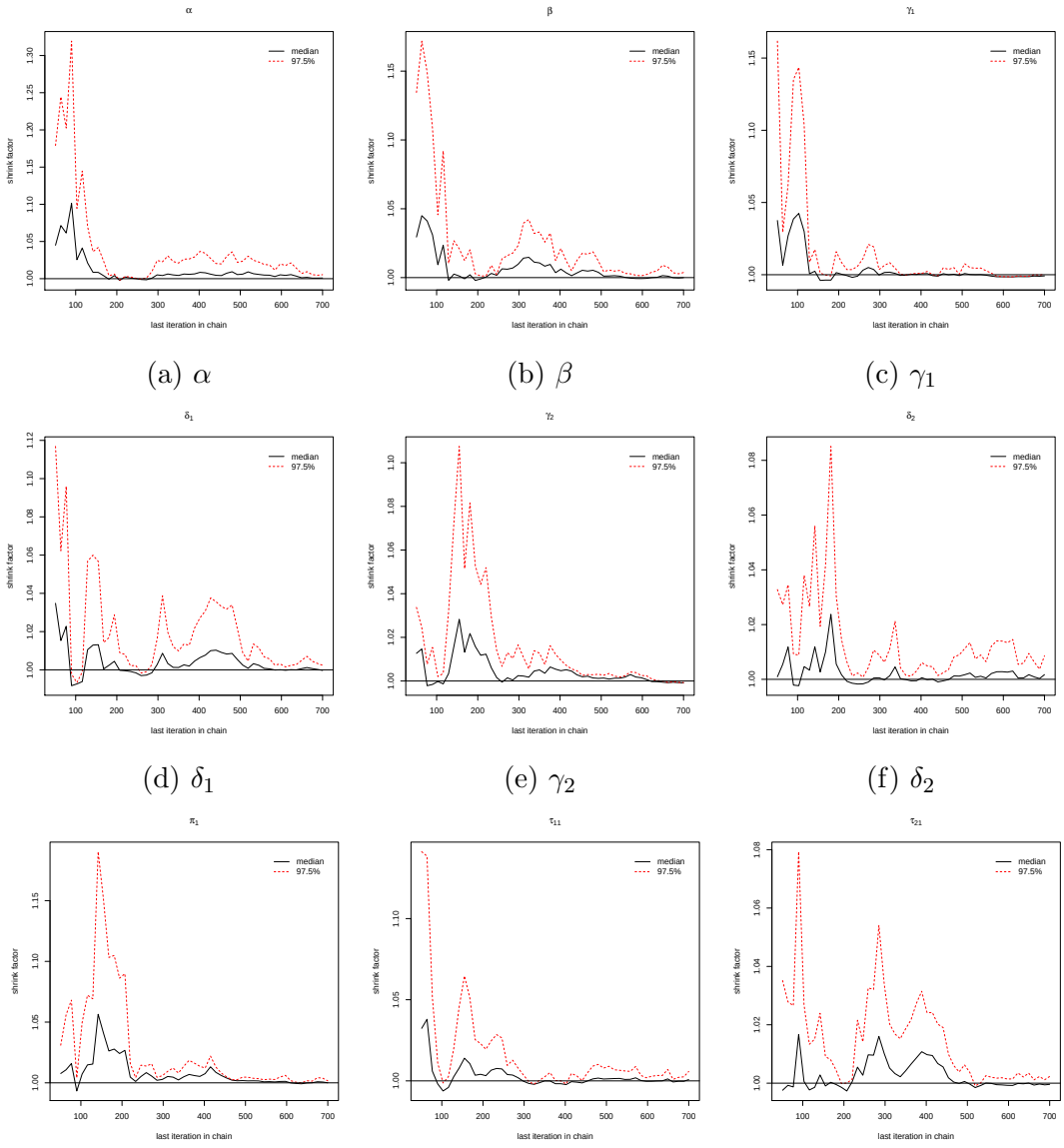


(a) α (b) β (c) γ_1

(d) δ_1 (e) γ_2 (f) δ_2

(g) π_1 (h) τ_{11} (i) τ_{21}

Figure 2: Trace plots of BBDM model parameters applied to the Chromosome 16 data.

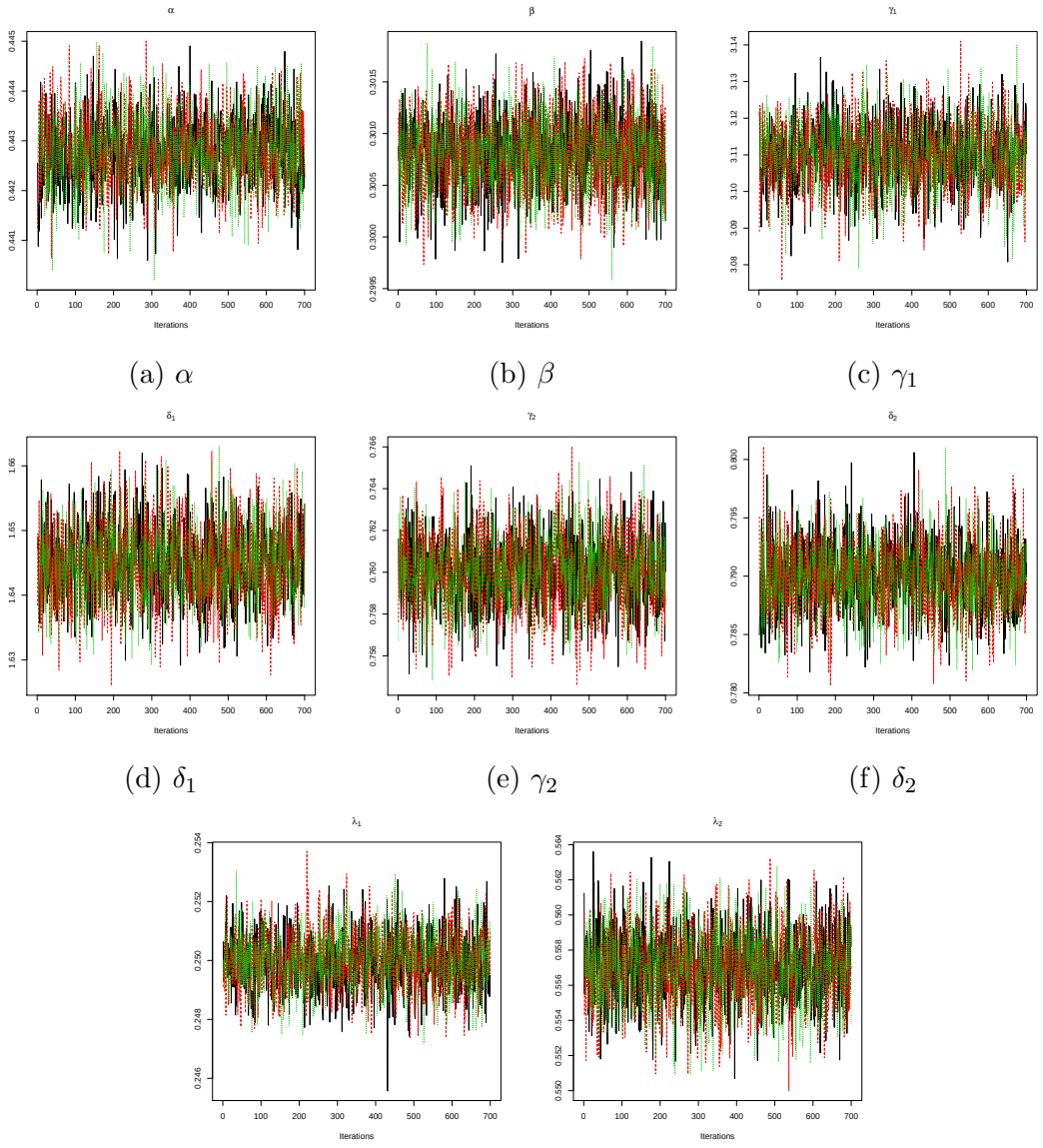


(a) α (b) β (c) γ_1

(d) δ_1 (e) γ_2 (f) δ_2

(g) π_1 (h) τ_{11} (i) τ_{21}

Figure 3: Gelman and Rubin's shrink factor plot of BBDM model parameters applied to the Chromosome 16 data.

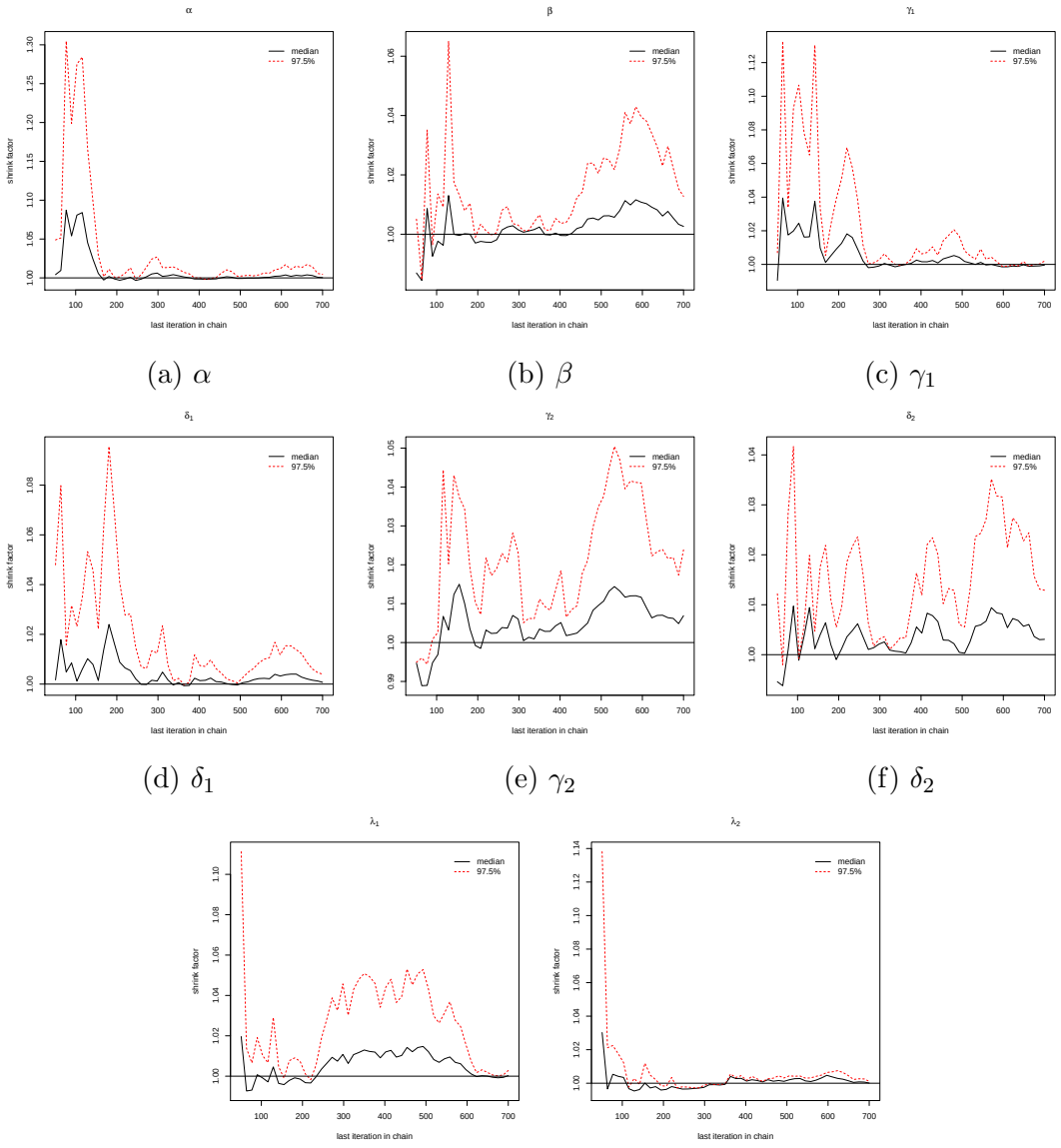


(a) α (b) β (c) γ_1

(d) δ_1 (e) γ_2 (f) δ_2

(g) λ_1 (h) λ_2

Figure 4: Trace plots of BBCM model parameters applied to the Chromosome 16 data.



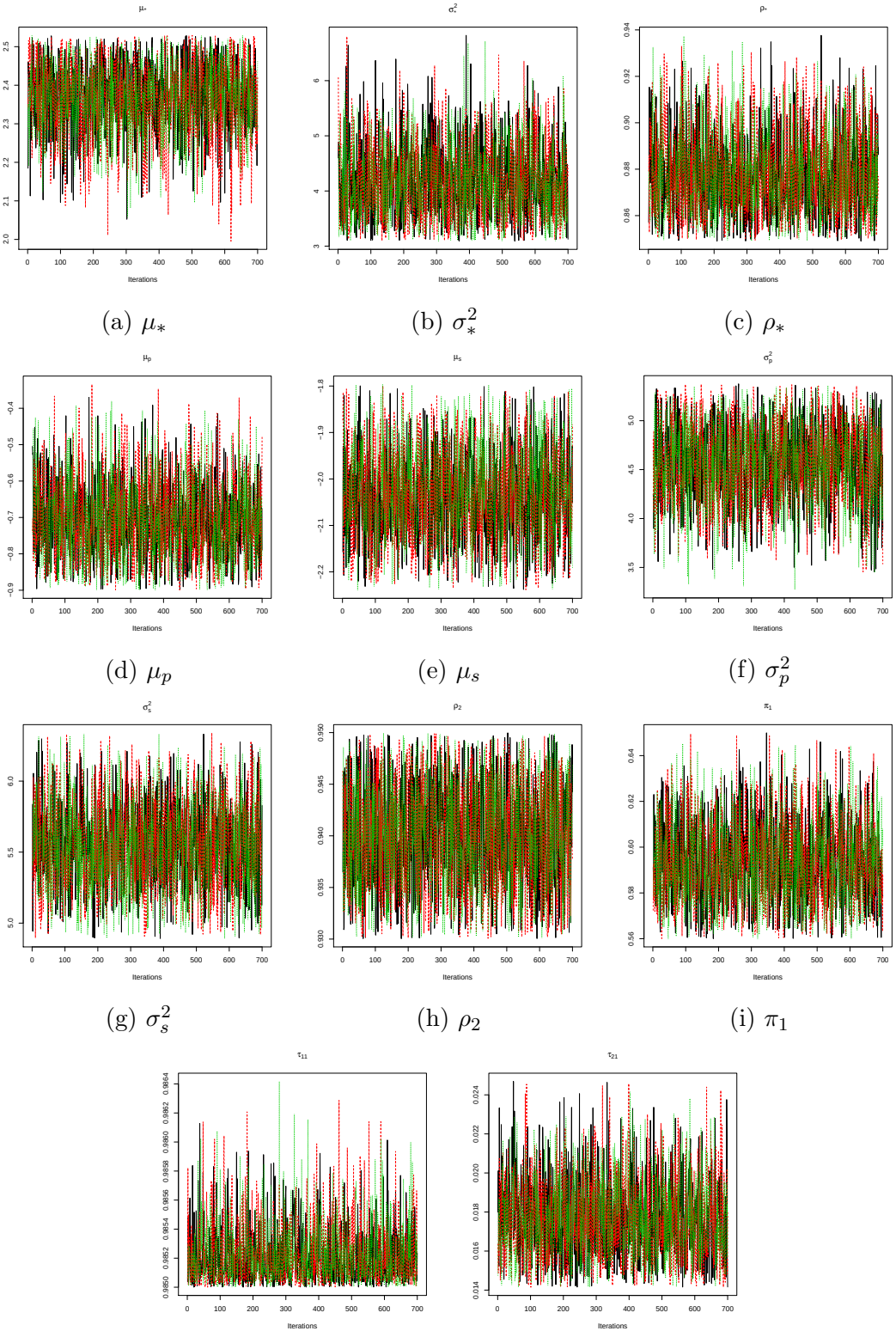
(a) α (b) β (c) γ_1

(d) δ_1 (e) γ_2 (f) δ_2

(g) λ_1 (h) λ_2

Figure 5: Gelman and Rubin's shrink factor plot of BBCM model parameters applied to the Chromosome 16 data.

Appendix A3



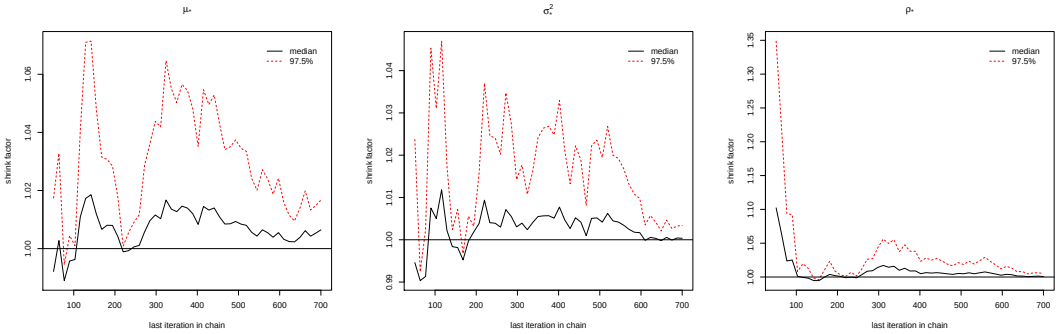
(a) μ_* (b) σ_*^2 (c) ρ_*

(d) μ_p (e) μ_s (f) σ_p^2

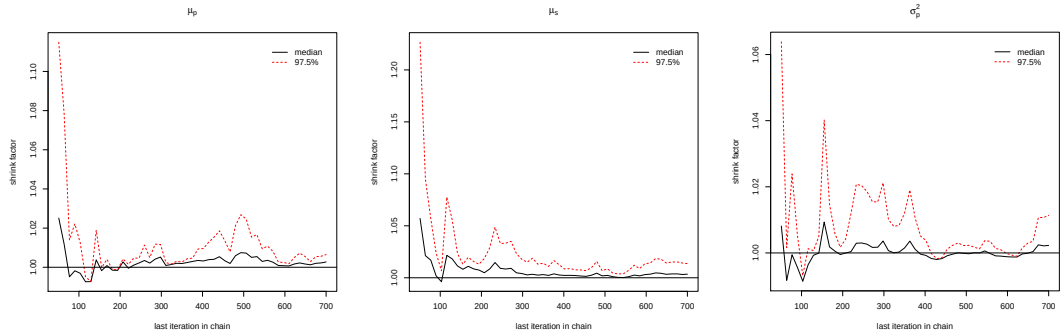
(g) σ_s^2 (h) ρ_2 (i) π_1

(j) τ_{11} (k) τ_{21}

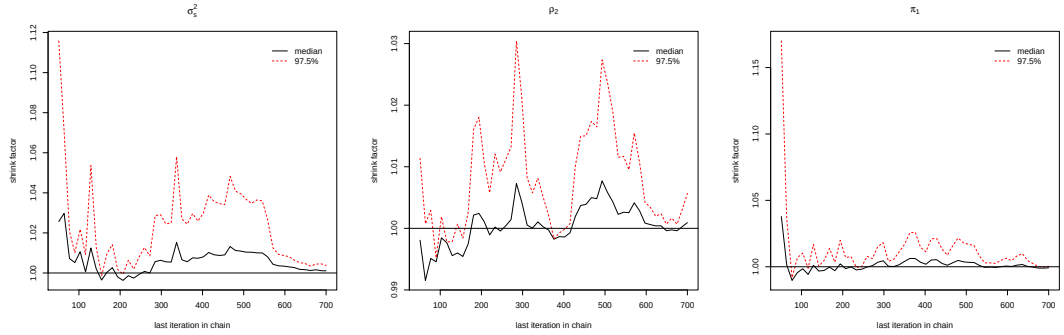
Figure 6: Trace plots of NLBDM model parameters applied to the Chromosome 16 data.



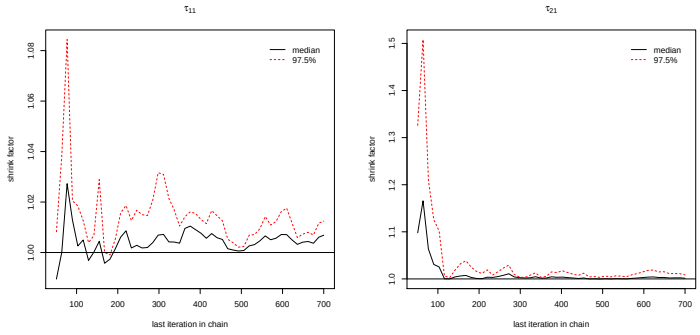
(a) μ_* (b) σ_*^2 (c) ρ_*



(d) μ_p (e) μ_s (f) σ_p^2

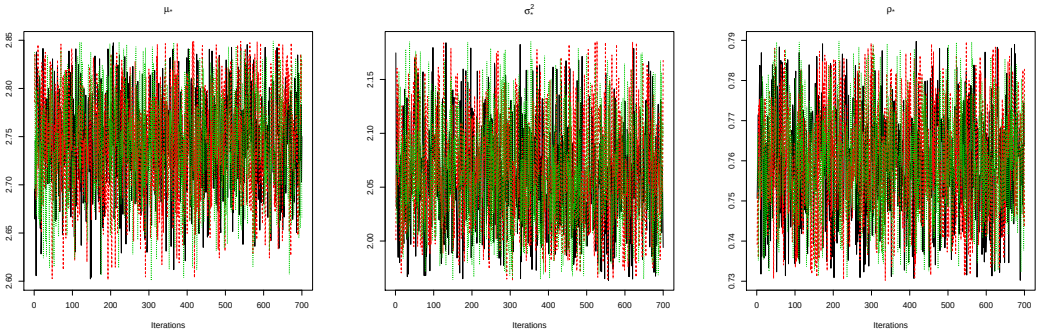


(g) σ_s^2 (h) ρ_2 (i) π_1

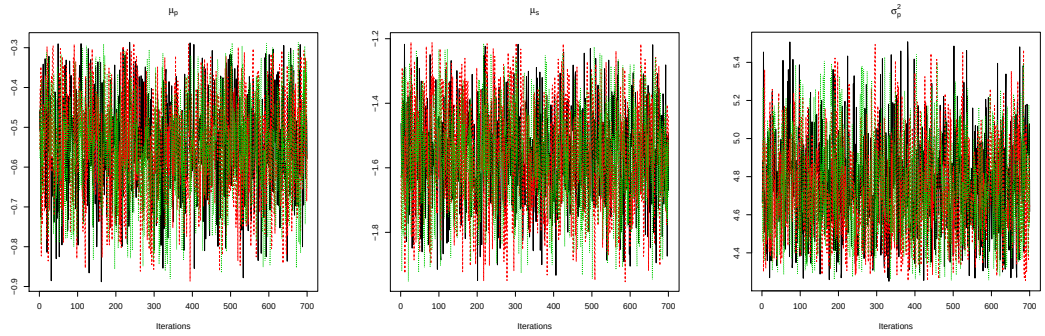


(j) τ_{11} (k) τ_{21}

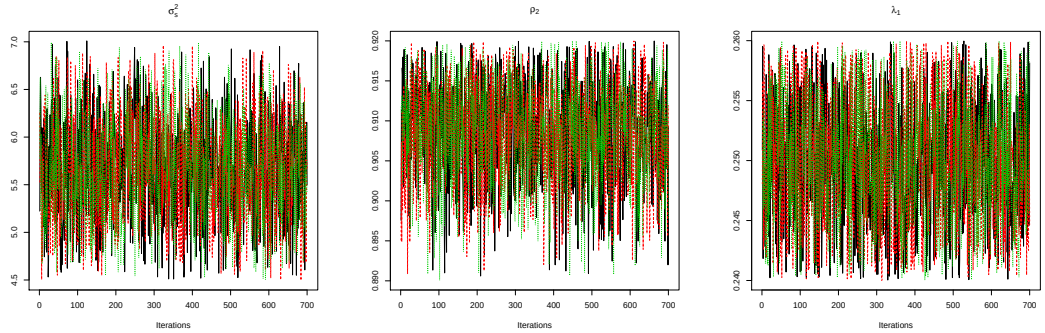
Figure 7: Gelman and Rubin's shrink factor plot of NLBDM model parameters applied to the Chromosome 16 data.



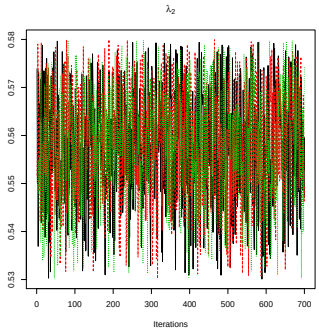
(a) μ_* (b) σ_*^2 (c) ρ_*



(d) μ_p (e) μ_s (f) σ_p^2

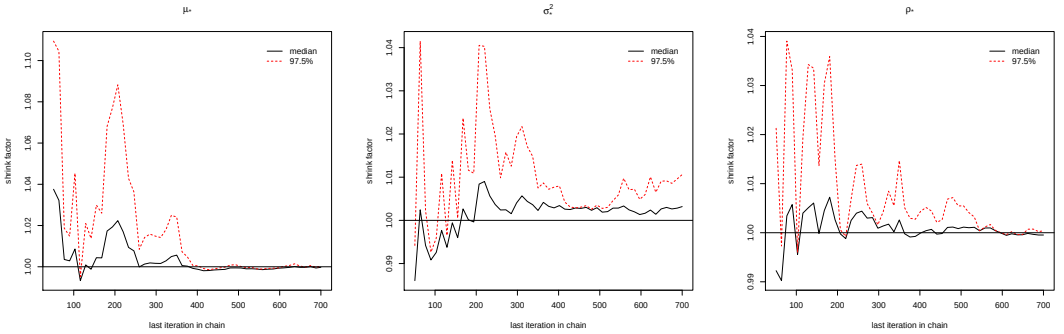


(g) σ_s^2 (h) ρ_2 (i) λ_1

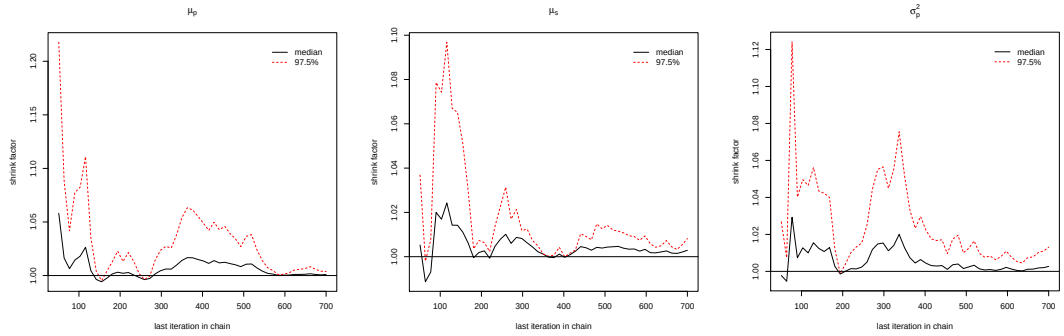


(j) λ_2

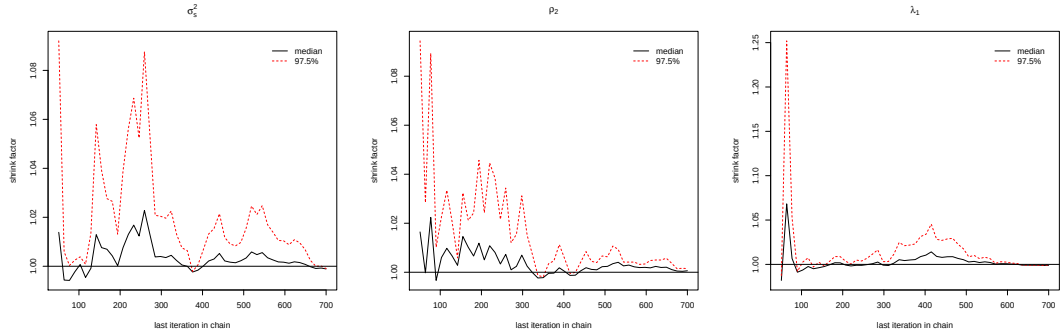
Figure 8: Trace plots of NLBCM model parameters applied to the Chromosome 16 data.



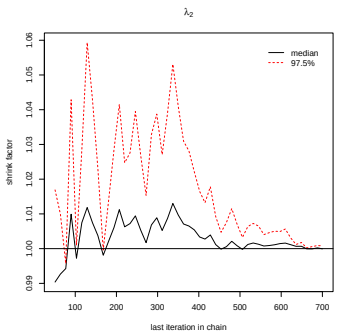
(a) μ_* (b) σ_*^2 (c) ρ_*



(d) μ_p (e) μ_s (f) σ_p^2



(g) σ_s^2 (h) ρ_2 (i) λ_1



(j) λ_2

Figure 9: Gelman and Rubin's shrink factor plot of NLBCM model parameters applied to the Chromosome 16 data.

References

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., and Mason, C. E. (2012). methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome Biology*, 13(10):R87. 36, 158

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418. 9

Blair, L. P. and Yan, Q. (2012). Epigenetic mechanisms in commonly occurring cancers. *DNA AND CELL BIOLOGY*, 31(Supplement 1):. S–49S–61. 5

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455. 10

Burger, L., Gaidatzis, D., Schbeler, D., and Stadler, M. (2013). Identification of active regulatory regions from dna methylation data. *Nucleic Acids Res.*, 41:e155. 36

Celeux, G., Forbes, F., Robert, C., and Titterington, D. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1(4):651–674. 28, 29, 142, 149, 150

Challen, G., Sun, D., Jeong, M., Luo, M., Jelinek, J., Berg, J., Bock, C., Vasanthakumar, A., Gu, H., Xi, Y., Liang, S., Lu, Y., Darlington, G., Meissner, A.,

Issa, J., Godley, L., Li, W., and Goodell, M. (2012). Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.*, 44(2):23–31. 36

Chan, J. C. and Grant, A. L. (2016). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics Data Analysis*, 100:847 – 859. 29

Chib, S. (1996). Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75:79–97. 20

Cron, A. J. and West, M. (2011). Efficient classification-based relabeling in mixture models. *Am Stat.*, 65(1):16–20. 22, 23

Cruickshanks, H., McBryan, A., Nelson, D., VanderKraats, N., Shah, P., van Tuyn, J., Rai, T., Brock, C., Donahue, G., Dunican, D., Drotar, M., Meehan, R., Edwards, J., Berger, S., and Adams, P. (2013). Senescent cells harbour features of the cancer epigenome. *Nature Cell Biology*, 15(12):1495–1506. ix, xiv, 42, 89, 91, 156, 180

D Smith, Z., M Chan, M., S Mikkelsen, T., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of dna methylation in the early mammalian embryo. 484:339–44. 162

Das, P. M. and Singal, R. (2004). Establishing, maintaining and modifying dna methylation patterns in plants and animals. *Nat. Rev. Genet.*, 22(22):4632–4642. 4

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids. 14

Ehrlich, M. (2002). Dna methylation in cancer: too much, but also too little. *Oncogene*, 21(35):5400. 4

Esteller, M. (2002). Cpg island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene*, 21. 4

Feng, H., Conneely, K. N., and Wu, H. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*, 42(8):e69. 39, 41

Gelman, A., Hwang, J., and Vehtari, A. (2013). Understanding predictive information criteria for bayesian models. 24. 149

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016. 27, 142

Gelman, A. and Meng, X.-L. (1998). chapter for gilks, richardson, and spiegelhalter book: Markov chain monte carlo in practice, chapman hall/crc. 24, 145

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511. 13, 90

Gelman, A., M. X. and Stern, H. (2000). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, 6:733–807. 24, 146

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741. 11

Gopalakrishnan, S., Van Emburgh, B., and Robertson, K. (2008). Dna methylation in development and human disease. *Mutat. Res.*, 647((1-2)):30–38. 2

Hansen, K., Langmead, B., and Irizarry, R. (2012). Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.*, 13(R83). 36

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109. 11

Hebestreit, K., Dugas, M., and Klein, H. (2013). Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13):1647–1653. 36

Henrichsen, C. N., Chaignat, E., and Reymond, A. (2009). Copy number variants, diseases and gene expression. *Hum Mol Genet*, 18. 4

Hooten, M. B. and Hobbs, N. T. (2015). A guide to bayesian model selection for ecologists. *Ecological Monographs, by the Ecological Society of America*, 85(1):3–28. 150

Jonghyun, Y., Tao, W., and Guanghua, X. (2014). Bayesian hidden markov models to identify rnaprotein interaction sites in par-clip. *Biometrics*, 70(2):430–440. 146

Koski, T. (2001). Hidden markov models for bioinformatics. 14

Krueger, F. and Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. *Bioinformatics*, 27(11):1571–1572. xi, 33, 35

Krueger, F., Kreck, B., Franke, A., and Andrews, S. (2012). Dna methylome analysis using short bisulfite sequencing data. *Nat. Meth.*, 9(2):145–151. 2

Kulis, M. and Esteller, M. (2010). Dna methylation and cancer. *Adv Genet*, 70. 4

Laird, P. W. and Jaenisch, R. (1994). Dna methylation and cancer. *Hum Mol Genet*, 3. 4

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., and Ong, C. T. (2010). Dynamic changes in the human methylome during differentiation. *Genome Res*, 20. 4

Law, J. and Jacobsen, S. (2010). Dna methylation and cancer. *Dna methylation and cancer. J*, 11:204–220. 2

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models. *Stochastic processes and their applications*, 40:127–143. 22

Li, E., Beard, C., and Jaenisch, R. (1993). Role for dna methylation in genomic imprinting. *Nature*, 366. 2

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079. 37

Li, Y., Zeng, T., and Yu, J. (2012). Robust deviance information criterion for latent variable models. *SMU Economics and Statistics Working Paper Series*. 28

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40. 59

Lutsik, P., Feuerbach, L., Arand, J., Lengauer, T., Walter, J., and Bock, C. (2011). Biq analyzer ht: locus-specific analysis of dna methylation by high-throughput bisulfite sequencing. *Nucleic Acids Res.*, 39:551–556. 34

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092. 11, 71

Millar, R. B. (2009). Comparison of hierarchical bayesian models for overdispersed count data using dic and bayes' factors. *Biometrics*, 65(3):962–969. 29

Moarii, M., Boeva, V., Vert, J.-P., and Reyal, F. (2015). Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics*, 16(1):873. 4

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the Society of Industrial and Applied Mathematics*, (5):32–38. 23

Newell-Price, J., Clark, A. J. L., and King, P. (2000). Dna methylation and silencing of gene expression. *Trends Endocrinol Metab*, 11. 5

Oleson, J. J. (2010). Bayesian credible intervals for binomial proportions in a single patient trial. *Stat Methods Med Res*, 19(6):559–74. 181

Olkin, I. and Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics and Probability Letters*, 96:54–60. 181

Park, Y. and Wu, H. (2016). Differential methylation analysis for bs-seq data under general experimental design. *Bioinformatics*, 32(10):1446–1453. 41

Plummer, M. (2008). Penalized loss functions for bayesian model comparison. *Biostatistics*, 9(3):523–539. 149

Qu, Y., Lennartsson, A., Gaidzik, V. I., Deneberg, S., and Karimi, M. (2014). Differential methylation in cn-aml preferentially targets non-cgi regions and is dictated by dnmt3a mutational status and associated with predominant hypomethylation of hox genes. *Epigenetics*, 9. 4

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. 14, 15, 17

Richardson, S. (2002). Discussion of the paper by spiegelhalter et al. journal of the royal statistical society b 64:626227. *Biostatistics*, 9(3):523–539. 150

Richardson, S. and Green, P. (1997). Journal of the royal statistical society: Series b (statistical methodology). *Nat. Rev. Genet.*, 59(4):731–792. 22

Robertson, K. (2005). Dna methylation and human disease. *Nat. Rev. Genet.*, 6:597–610. 4

Scott, S. L. (2002). Bayesian methods for hidden markov models. *J. Am. Stat. Assoc.*, 97:337–351. 17, 19, 20, 62, 71, 119, 146

Smith, Z. D. and Meissner, A. (2013). Dna methylation: roles in mammalian development. *Nat Rev Genet*, 14. 2

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. 26, 142

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809. 22

Sun, Z., Baheti, S., Middha, S., Kanwar, R., Zhang, Y., Li, X., Beutler, A., Klee, E., Asmann, Y., Thompson, E., and Kocher, J. (2012). Saap-rrbs: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing. *Bioinformatics*, 28:2180–2181. 34

Wang, H., Tuominen, L., and Tsai, C. (2011). Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, 27(2):225–231. 38

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594. 29, 142, 150

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*, 43(21):e141. 38, 41, 158

Yingying, Z. and Jeltsch, A. (2010). The application of next generation sequencing in dna methylation analysis. 1. 32

Zackay, A. and Steinhoff, C. (2010). Methvisual - visualization and exploratory statistical analysis of dna methylation profiles from bisulfite sequencing. *BMC Res. Notes*, 3:337. 34