RICE UNIVERSITY

Application of Bayesian Modeling to
High-throughput Genomic Data and Clinical Trial
Design

by

Yanxun Xu

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

Approved, Thesis Committee:

Dennis D. Cox, Chair
Professor of Statistics

Yuan Ji, Co-Chair
Director, Cancer Research Informatics,
NorthShore University HealthSystem

Peng Qiu
Assistant Professor, Department of
Bioinformatics and Computational
Biology, The University of Texas MD
Anderson Cancer Center

David W. Scott
Noah Harding Professor of Statistics

Luay K. Nakhleh
Associate Professor of Computer Science
and Biochemistry and Cell Biology

Houston, Texas
December, 2013
ABSTRACT

Application of Bayesian Modeling to High-throughput Genomic Data and Clinical Trial Design

by

Yanxun Xu

My dissertation is concerned about the development of Bayesian models for big data and clinical trial designs. The methodologies are illustrated in two different contexts: high-throughput next generation sequencing (NGS) data and Bayesian adaptive designs.

NGS experiments generate millions of short reads, which provide valuable information for various aspects of cellular activities and biological functions. So far, NGS techniques have been applied in quantitatively measurement of diverse platforms, such as RNA expression, DNA copy number variation (CNV) and DNA methylation. Although NGS is powerful and largely expedites biomedical research in various fields, challenge remains due to the high modality of disparate high-throughput data, high variabilities in the data acquisition, and high dimensionality of the resulting data. Bayesian paradigm has the potential to fill in these gaps, because of its coherent and principled way of combining prior information with data and its versatility with
multiple models. In the big data part, I will propose Bayesian approaches to address
above challenges so that we can take full advantage of the NGS technology. This
part includes three topics: (1) BM-Map: a Bayesian mapping of multireads for NGS
data, (2) a Bayesian graphical model for integrative analysis of TCGA data, and (3)
a nonparametric Bayesian Bi-clustering for NGS count data.

In the trial design part, I will propose a latent Gaussian process (LGP) to model
discrete longitudinal data with recurrent or cyclic patterns. The model is designed
to monitor clinical trials for chronic and repetitive diseases such as mental disorders.
In those trials, patients are followed for a long period of time (e.g., 35 weeks) with
frequent disease relapses. LGP describes the underlying latent process that gives rise
to the observed longitudinal binary outcomes defined as recurring efficacy responses
to treatment. The posterior consistency property of the proposed model is studied.
We propose a monitoring rule based on posterior stopping probabilities, with which
trials may be terminated early.
Acknowledgments

My utmost gratitude goes to my advisor, Dr. Yuan Ji, whose encouragement, guidance and support from the initial to the final level enabled me to develop an understanding of Bayesian statistics and inspired me step by step to bring forth these projects and beyond in the very Bayesian view. Without his help and guidance, this work could never be well accomplished. He is not only an outstanding professor with wide knowledge and extraordinary academic achievements, but also a great friend to me. His way of thinking and teaching, his perspectives of life and research have left me happiest circumstances during my Ph.D study and will also light up my life in future. I could not have imagined having a better advisor and mentor for my Ph.D study.

My great appreciation should be then expressed to Dr. Peng Qiu, who has generously supported me in the last year of my Ph.D study, both financially and academically. His help and instruction always appear timely, especially on my several presentations in multiple conferences and my job interview talk.

I am also grateful to Dr. Dennis D. Cox and Dr. David W. Scott, for their encouragements, their precious time to be my committee members and their impressive lectures and helps during my three years at Rice. They are the mentors that every student would like to turn to for help along the way. And thank Dr. Luay K. Nakhleh for his valuable time to be in my committee and insightful suggestions and comments during my proposal and defense.
In addition, I am quite lucky to be in Department of Statistics at Rice University. The whole department is like a big family and I am very glad to meet many smart, kindhearted and accommodating faculties, staffs, friends and classmates who make my life more colorful and meaningful at Rice. Further thanks shall then go to all my collaborators on these projects in this dissertation including Dr. Peter Müller, Dr. Han Liang, Dr. Riten Mitra, Dr. Juhee Lee, etc.

I offer my regards and blessings to all of those who supported me in any respect during the completion of my dissertation. To my wonderful Zuzu, I shall express my special thanks for her tireless help, her faith in me and our great friendship.

Lastly, and most importantly, I wish to thank my parents. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this dissertation.
Contents

Abstract ii
Acknowledgments iv
List of Illustrations x
List of Tables xvi

1 Introduction 1

1.1 Background .................................................. 1
  1.1.1 Bayes’ Theorem ........................................ 1
  1.1.2 Next-Generation Sequencing ............................. 2
  1.1.3 Gaussian Process ....................................... 5
1.2 Outline and Contributions of the Dissertation ............... 6

2 BM-Map: Bayesian Mapping of Multireads for Next-
Generation Sequencing Data 9

  2.1 Literature Review and Motivation .......................... 10
  2.2 Yeast RNA-Seq data ....................................... 13
  2.3 Methodology ............................................... 14
    2.3.1 Probability model .................................... 14
    2.3.2 Markov chain Monte Carlo simulations ................ 18
  2.4 Simulation studies ........................................ 19
3 A Bayesian Graphical Model for Integrative Analysis of TCGA Data

3.1 Literature Review and Motivation .................................................. 32
3.2 TCGA Ovarian Cancer Data ............................................................. 33
3.3 Probability Model ................................................................. 34
  3.3.1 Sampling Model ............................................................... 34
  3.3.2 Markov Chain Monte Carlo (MCMC) Simulations ...................... 36
3.4 Simulation Study ................................................................. 37
3.5 Ovarian Cancer Data Analysis .................................................... 38

4 Nonparametric Bayesian Bi-Clustering for Next Generation Sequencing Count Data

4.1 Literature Review and Motivation .................................................. 42
4.2 ChIP-Seq Data ................................................................. 45
4.3 Methodology ................................................................. 45
  4.3.1 Probability Model ............................................................... 45
  4.3.2 Markov Chain Monte Carlo Simulations .................................. 49
  4.3.3 Posterior Inference ............................................................ 51
LGP Model 105

Bibliography 118
Illustrations

2.1 (Colored) A flow chart of the main steps in the RNA-Seq approach. Our proposed method, BM-MAP, considers mapping the multireads after the general read alignment is finished. The available tools for each step are listed on the right side of the chart. Currently, there is only one method, the proportional method, that deals with the mapping of the multireads. The drawbacks of the proportional method are discussed in chapter 2.1. ................................. 11

2.2 (Colored) Results summarizing the false discovery rates (FDRs) of the three methods in the simulation studies. Left panel: stacked FDRs over the 21 simulated cases in chapter 2.4. A method name followed by “-1” and “-2” respectively represents the stacked 21 FDRs for mapping multireads to locations 1 and 2. Right panel: FDRs for selected simulation cases in which there are large differences in the FDRs among the three methods. ................................. 22

2.3 Histogram of the differences in mapping the multireads of the yeast genome data between the BM-Map method and the Prop method. ................................. 25
2.4 (Colored) Three examples of multireads mapped to two hits based on
the probabilities of the BM-Map and Prop methods. In each panel
(row), the two plots correspond to two competing genomic locations
for a multiread. Plotted are the posterior means of $\beta_{kt}$, the
probability of hidden nucleotide variation. These are estimated based
on the error mismatch profiles between the unique reads and the
genomic locations. The red dots/lines indicate where the multireads
have mismatches. Top panel: the multiread has mismatches at bases
(23, 24) to location 1 and (12, 13) to location 2. Middle panel: the
multiread has mismatches at bases (7, 22) to location 1 and (34, 35)
to location 2. Bottom panel: the multiread has mismatches at bases
(1, 12) to both locations. The probabilities of mapping based on the
BM-Map and proportional methods are presented in each plot.

2.5 Results comparing the RPKMs from three methods. Shown are the
$log_2$ absolute differences in the RPKMs between the BM-Map and
the Naive method (left panel) and between the BM-Map and the
proportional method (right panel).
3.1 The simulation truth versus the estimated graph for three simulated data set. Edge colors black and red represent positive and negative relationships. The solid line represents that the edge exists. The red dotted lines indicate that the corresponding edges do not exist. The number next to each edge represents either the true value or the posterior mean of the autologistic coefficients $\beta$’s, 0 for the edges do not exist. The estimated graph based on posterior inference is identical to the simulation truth.

3.2 Smooth scatter plots of pairwise relationship among platforms $C$, $M$ and $E$. The upper panel is for gene “ERLIN2”, the low panel is for gene “PIR”. The red line in each smooth scatter plot is the lowess smoother. Dots correspond to the raw expression measurements from the level three TCGA data.

3.3 Posterior estimated graphs for genes “ERLIN2” and “PIR”. Black edges represent positive relationships and red edges represent negative relationships. The number next to each edge is the posterior mean of $\beta_{ij}$.

4.1 An illustration of the proposed NoB-LCP model with 9 HMs and 10 genomic locations. There are two active HM sets and an idle HM set, including HMs 5, 8, 9. In the two active HM sets, cells in off-white are idle genomic locations. The rest of the cells marked with the same color in the same column form local clusters of genomic locations (rows). Different colors indicate different values of parameters $\tilde{\theta}_{dg}$. 
4.2 Heatmaps of the HM sets in the simulation truth versus the identified HM sets under the NoB-LCP model. The first row block of each active HM set is the idle genomic location cluster for that HM set. The remaining blocks are active genomic location clusters. The division of genomic location clusters is indicated by white horizontal lines.

4.3 Heatmaps of HMs in two biclusters of the simulated data identified by the plaid model. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.

4.4 Heatmaps of HMs in six biclusters for the simulated data identified by QUBIC. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.

4.5 Heatmaps of HMs in two active HM sets for the simulated data identified by the model with regular Pólya urn priors without zero-enrichment. The division of genomic locations is indicated by white horizontal lines.

4.6 Heatmaps of three active HM sets for ChIP-Seq data. White horizontal lines indicate division of location clusters.

5.1 Three random samples drawn from the Gaussian process $\tau_{ij}$ with four different covariance matrices. The corresponding covariance is given below each plot. The decrease in length scale from (a) to (b) produces more rapidly fluctuating functions. The periodic properties of the covariance function in (c) and (d) can clearly be seen.
5.2 Sample curves of latent variable values for 20 randomly selected patients in scenarios 1-4. The red curve in each plot represents the mean function \( \mu(t) \) of the GP. The horizontal blue line represents the threshold \( a_h \). 

5.3 The posterior estimates of \( m \) in all cases for scenarios 1-3 in the simulation.

5.4 A set of 20 realizations of latent variables for two cases in scenario 2. The left panel presents the latent variable values for one patient with the first 32 time points observed. The right panel presents one patient with the first 29 time points observed. The red solid line represents the truth. Black solid lines represent 20 realizations at observed time points, and blue dashed lines represent predicted values at future time points.

5.5 A set of 20 realizations of latent variables for two cases in scenario 4. The left panel presents the latent variable values for one patient with the first 32 time points observed. The right panel presents one patient with the first 29 time points observed. The red solid line represents the truth. Black solid lines represent 20 realizations at observed time points, and blue dashed lines represent predicted values at future time points.
5.6 The true DDR $T_i$ and the true mean functions $\mu_i(t)$ in the control and experimental arms in all six scenarios for the lupus trial simulation. The red line represents the true mean in the control arm, while the blue line represents the true mean in the experimental arm. The black horizontal line represents the threshold $a_h = 0$. . . . . . . 90
2.1 A list of yeast genes with largest changes in the normalized RPKMs between the BM-Map and the proportional methods. The normalized RPKM diff is defined as $|\text{RPKM}_{\text{prop}} - \text{RPKM}_{\text{BM-Map}}| / \text{RPKM}_{\text{BM-Map}}$. The ORF and Gene names are from the saccharomyces genome database (http://www.yeastgenome.org). All the five genes listed below are duplicate genes, hence possessing sequences similar to other genes. They all play important biological functions in yeast.

4.1 The true mean counts for active genomic location clusters, $\tilde{\theta}_{dy}$, in the simulated data.

4.2 The number of genomic locations in each genomic location cluster for active HM sets.

4.3 Comparisons of the location cluster membership estimated by the NoB-LCP model with the true location cluster membership.

5.1 True probability, empirical probability and average posterior predictive probability of response at future time point $t_{33}$, $t_{34}$ and $t_{35}$. The posterior mean DDR, $\hat{T}$, is also presented.
5.2 True Values of $m_i$, $\beta_i$ and $T_i$, $i = 1, 2$ for all six scenarios in the simulation.

5.3 Results for the simulations based on the lupus trial. AD: average trial duration (weeks); MD: maximum duration (weeks); AP: average number of patients in each group.
Chapter 1

Introduction

1.1 Background

1.1.1 Bayes’ Theorem

Bayes (1763) [1] studied the parameters of a Binomial distribution based on observed data and derived the famous Bayes’ theorem for the specific case in his paper. Later, Laplace (1812) [2] extended the theorem to any probability distribution. Statistically, Bayes’ rule gives the relationship between the probabilities of events $A$ and $B$, $P(A)$ and $P(B)$, and the conditional probabilities $P(A \mid B)$ and $P(B \mid A)$:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Assuming that we have a model $H$ with a set of unknown parameters $\theta$ to model data $D_n = (y_1, \ldots, y_n)$. The prior belief of unknown parameters $\theta$ given $H$ is expressed as $P(\theta \mid H)$. According to Bayes’ rule, we have

$$P(\theta \mid D_n, H) = \frac{P(D_n \mid \theta, H)P(\theta \mid H)}{P(D_n \mid H)}.$$

The predictive posterior distribution for a future data point $y_{n+1}$ is

$$P(y_{n+1} \mid D_n, H) = \int P(y_{n+1} \mid \theta, D_n, H)P(\theta \mid D_n, H) \, d\theta.$$
1.1.2 Next-Generation Sequencing

Next-generation sequencing (NGS) technology generates millions of short reads at low cost, which provide valuable information for various aspects of cellular activities and biological functions.

RNA-seq reads mapping

RNA-Seq is an NGS application that generates millions of short RNA reads. By mapping and counting individual short sequence reads to specific genomic locations on the source genome, gene expression can be quantified as the number of mapped reads, which is considered the “digital” expression as opposed to the “analog” expression of relative transcript abundance in microarrays. Due to its low cost and high precision, RNA-Seq has become the primary tool for sequencing all the RNAs in species ranging from yeast to human [3, 4, 5, 6, 7, 8]. Many statistical challenges lie ahead in various steps of the RNA-Seq, preventing the technology from reaching its full potential. A key step in the RNA-Seq is read mapping, which infers the origin of short reads on a reference genome. Two sources of variations in read alignment complicate the accuracy of this step.

(i) The first source is sequencing errors [9, 10, 11] from upstream analysis, which occasionally occur during the process of generating short reads. In particular, bases within a read are often sequenced with mistakes due to machine and systemic errors. The error rates can be represented by the quality score underlying each base. These sequencing errors will cause mismatches between the short reads and the genome, which should not be counted.

(ii) The second source is called hidden nucleotide variations, such as a mutation
or SNP [12, 13, 14]. The main cause for this type of variations is because the short reads are typically mapped to a public reference genome rather than the sample genome from which the reads are generated. Hence, variations between the two genome versions (i.e., SNPs) cause mismatches between the reads and the reference genome.

In the reads mapping, a significant number of reads are mapped to more than one genomic location with similar fidelities, and these reads are called multireads. Importantly, multireads disproportionally come from the genes with similar sequences (e.g., duplicated genes) and essentially determine their expression levels. The alignment of multireads is highly susceptible to the two sources of variations above, making it difficult to map them to appropriate locations.

**Integrating RNA expression, CNV and DNA methylation**

Gene expression is a critical genetic process in which DNA is transcribed to RNA. Perturbation of transcription directly affects mRNA expression and hence the subsequent protein production, leading to pathological states. Genetic variations such as copy-number variations (CNVs) and DNA methylations frequently contribute to disrupted gene expression. CNVs result in an abnormal number of copies of DNA and thus change the gene expression level and associated phenotypes. For example, a higher copy number of CCL3L1 has been associated with lower susceptibility to HIV infection [15], and a low copy number of FCGR3B can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders [16]. DNA methylation is a biochemical modification that adds a methyl group to the 5 position of the cytosine pyrimidine ring or the number 6 nitrogen of the adenine purine ring. There is strong evidence that abnormal hypermethylation at the gene
promoter region results in transcriptional silencing of tumor suppressor genes. Also, aberrant DNA methylation patterns have been associated with a large number of human malignancies such as cancer, lupus, and a range of birth defects [17]. Therefore, elucidating tumor-specific methylation changes will shed light on potential clinical applications in cancer diagnosis, prognosis and therapeutics [18].

Thanks to the next generation sequencing technologies, the measurement of diverse platforms, such as RNA expression, DNA copy number variation (CNV) and DNA methylation can be quantitatively measured. Based on their measurements, current literature mainly focuses on the pair-wise integration, between CNVs and mRNA or between methylation and mRNA. Since both CNVs and DNA methylation play important roles in mRNA expression, an integrated analysis that models all three platforms together is most appropriate. We will consider a Bayesian network model to integrate RNA expression, CNV and DNA methylation.

**Revealing histone modification patterns by ChIP-Seq**

ChIP-Seq integrates chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (Seq) to identify genome-wide expression patterns of DNA-binding proteins. ChIP-seq is a technique to study the interactions of DNA and protein of interests. ChIP-Seq data record the counts of sequence tags mapped onto non-overlapping positions that cover the genome. Histones are proteins that package DNA into structural units called nucleosomes. Through post-translational modifications, histones play key roles in transcription [19, 20], chromosomal segregation [21], and DNA repair. Combinations of such histone modifications (HMs) are known as the “histone code”, which modulates chromatin structure to regulate gene expression [22, 23]. For example, combinations of HMs have been linked to cancer prognosis [24]
and clinical decisions [25]. By applying HM-specific antibodies, ChIP-Seq experiments can record the counts of DNA fragments that include a certain HM. And the fragments are mapped to specific locations across the whole genome. A large count of DNA fragments indicates high occurrence of the targeted HM.

Recently, several HM patterns have been shown to be associated with various classes of regulatory elements. Such HM patterns are known as chromatin signatures [26]. For example, distinct and predictive chromatin signatures are used to characterize active promoters and enhancers [27, 28]. These results lead us to look for more such patterns. We expect that regulatory elements with similar functionality are likely to share similar patterns of some subset of HMs. We conjecture that annotating genomic location on the basis of such patterns could be a promising step towards deciphering the histone code.

1.1.3 Gaussian Process

The Gaussian process (GP) has been frequently applied to research areas in finance, engineering, cognitive research, and others. Examples of such applications include work in machine learning [29], neural networks [30], and batched data [31]. Initially introduced in O’Hagan and Kingman (1978) [32], GP priors have also been used in Bayesian inference for regression and classification problems, as reviewed by Williams (1998) [33]. GP models are considered nonparametric in curve fitting as they are not dependent on any functions to describe the shapes of the curves. At the same time, GP models are relatively easy to compute since they are based on multivariate Gaussian distributions.

A Gaussian process is a stochastic process $a(t)$, for which any $n$-finite variates
\[ a_n = \{a(t_1), \ldots, a(t_n)\} \] has a multivariate Gaussian distribution given by

\[
P(a_n \mid \mu, C) \propto |C|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (a_n - \mu)^T C^{-1} (a_n - \mu) \right\}
\]

for any \( n \) and collection of input \( \{t_1, \ldots, t_n\} \). Vector \( \mu \) is the mean with dimension \( n \) and \( C \) represents the \( n \times n \) covariance matrix, and is often parameterized as a covariance function \( C^{uv}(t_u, t_v; \Theta) \), where \( \Theta \) is a set of hyperparameters. In other words, a Gaussian process is one which every finite-dimensional joint distribution is multivariate Gaussian. Since we are expressing the correlations between different points in the input space through the covariance function, it is crucial to study the characteristics of the GP, such as its smooth properties and differentiability [34].

The most significant computational cost of implementing a GP model when dealing with a vast number of time points is the inversion of the covariance matrix \( C \). Such an inversion is required to make any predictions and, most significantly, in Bayesian inference to evaluate the gradient of the log posterior distribution over \( \Theta \) in every Markov chain Monte Carlo (MCMC) iteration.

### 1.2 Outline and Contributions of the Dissertation

In chapter 2, a Bayesian algorithm to map the multireads of NGS data is proposed. A key step in NGS applications (e.g., RNA-Seq) is to map short reads to correct genomic locations of the source genome. While most reads are mapped to a unique location, a significant proportion of reads align to multiple genomic locations with equal or similar numbers of mismatches; these are called multireads. The ambiguity in mapping the multireads may lead to bias in downstream analyses. Currently, the industry standard is to discard the multireads, resulting in loss of valuable information, especially for the genes with similar sequences. To refine the read mapping,
we develop a Bayesian model that computes the posterior probability of mapping a multiread to each competing location. The probabilities are used for downstream analyses, such as the quantification of gene expression. We show through simulation studies and a yeast RNA-Seq analysis that the Bayesian method yields better mapping than the current leading methods. We provide a C++ program for downloading that is being packaged into a user-friendly software.

In chapter 3, we integrate three TCGA data sets including measurements on matched DNA copy numbers (C), DNA methylation (M), and mRNA expression (E) over 500+ ovarian cancer samples. The integrative analysis is based on a Bayesian graphical model treating the three types of measurements as three vertices in a network. The graph is used as a convenient way to parameterize and display the dependence structure. Edges connecting vertices infer specific types of regulatory relationships. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation. We apply the graphical model to each of the genes in the TCGA data independently and provide a comprehensive list of inferred profiles. Examples are provided based on simulated data as well.

Chapter 4 presents a Bayesian nonparametric biclustering approach for ChIP-Seq data. Histone modifications (HMs) play important roles in transcription through post-translational modifications. Combinations of HMs, known as chromatin signatures, encode specific messages for gene regulation. We therefore expect that inference on possible clustering of HMs and an annotation of genomic locations on the basis of such clustering can provide new insights on the functions of regulatory elements and their relationships to combinations of HMs. We propose a nonparametric Bayesian
local clustering Poisson model (NoB-LCP) to facilitate posterior inference on two-dimensional clustering of HMs and genomic locations. The NoB-LCP clusters HMs into HM sets and lets each HM set define its own clustering of genomic locations. Furthermore, it probabilistically excludes HMs and genomic locations that are irrelevant to clustering. By doing so, the proposed model effectively identifies important sets of HMs and groups regulatory elements with similar functionality based on HM patterns.

In chapter 5, we present a latent Gaussian process (LGP) to model discrete longitudinal data with recurrent or cyclic patterns. The model is applied to monitor clinical trials for chronic diseases such as mental disorders. In those trials, patients are followed for a long period of time (e.g., 35 weeks) with frequent relapses. LGP describes the underlying latent process that gives rise to the observed longitudinal binary outcomes defined as recurring efficacy responses to treatment. The posterior consistency property of the proposed model is studied. Posterior inference is conducted with a hybrid Markov chain Monte Carlo algorithm. Based on the observed outcomes, LGP can forecast patient-specific future responses. We propose a monitoring rule based on posterior stopping probabilities, with which trials may be terminated early. Simulation studies are conducted under various clinical scenarios, and a case study is reported based on a real-life trial.
Chapter 2

BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data

In this chapter, we develop a Bayesian model that computes the posterior probability of mapping a multiread to each competing location on the reference genome. A crucial step in next-generation sequencing (NGS) (e.g., RNA-Seq) is to map short reads to correct genomic locations of the source genome. While most reads are mapped to a unique location, a significant proportion of reads align to multiple genomic locations with equal or similar numbers of mismatches; these are called multireads. The ambiguity in mapping the multireads may lead to bias in downstream analyses. Currently, the industry standard is to discard the multireads, resulting in loss of valuable information, especially for the genes with similar sequences. To refine the read mapping, we develop a Bayesian model that computes the posterior probability of mapping a multiread to each competing location. The probabilities are used for downstream analyses, such as the quantification of gene expression.

This chapter proceeds as follows. Chapter 2.1 reviews the literature on aligning the short reads and introduces the motivation of this work. Chapter 2.2 introduces the details of yeast RNA-Seq data we use to demonstrate our methodology. The probability model and Markov chain Monte Carlo (MCMC) sampling algorithms are presented in chapter 2.3. We conduct simulation studies to evaluate the performance of the proposed BM-Map method in chapter 2.4. Chapter 2.5 presents the result of yeast RNA-Seq data analysis.
2.1 Literature Review and Motivation

Next-generation sequencing (NGS) technology produces vast amount of sequence data at low cost and provides enormous opportunities for the life sciences. As introduced in chapter 1, an key step in RNA-Seq is to map millions of short reads to a reference genome.

Several published works have been developed for aligning the short reads to the reference genome. For example, Li, Ruan and Durbin (2008) [35] considered mapping short DNA sequences based on the quality scores and developed a software package called MAQ. Another popular program is Bowtie [36], an ultrafast, memory-efficient alignment program that aligns short reads to large genomes. Other representative works include the SOAP by Li et al. (2009) [37], the RMAP by Smith, Xuan, and Zhang (2008) [38], and the SHRiMP by Rumble et al. (2009) [39]. Despite the afore-mentioned sources of variations in read mapping, most mappable short reads (>75%) based on available methods and software (e.g., Bowtie) align to a single genomic location with relatively high precision. These reads are called unique reads. However, a significant number of reads are mapped to more than one genomic location with similar fidelities, and these reads are called multireads. Importantly, multireads disproportionately come from the genes with similar sequences (e.g., duplicated genes) and essentially determine their expression levels. The alignment of multireads is highly susceptible to the two sources of variations above, making it difficult to map them to appropriate locations.

We aim to improve the mapping of the multireads, as a refinement step after the general reads alignment is completed. Figure 2.1 demonstrates where our proposed method stands in the entire process of the NGS data analyses.

Currently, the industry standard is to discard the multireads in subsequent anal-
Figure 2.1: (Colored) A flow chart of the main steps in the RNA-Seq approach. Our proposed method, BM-MAP, considers mapping the multireads after the general read alignment is finished. The available tools for each step are listed on the right side of the chart. Currently, there is only one method, the proportional method, that deals with the mapping of the multireads. The drawbacks of the proportional method are discussed in chapter 2.1.
yses such as gene expression quantification. This practice generates a large bias in estimating the expression levels of duplicated genes. As an initial attempt, Mortazavi et al. (2008)[6] proposed a proportional alignment method in which unique reads are first mapped, and then multireads are aligned to equally similar loci in proportion to the number of corresponding mapped unique reads. The key idea of the proportional method is that the individual numbers of unique reads are used to infer the probabilities of mapping the multireads. While the proportional method provides a simple and valuable solution to the mapping of the multireads, it fails to account for the mismatch profiles between the unique reads and the genomic locations.

For example, using the proportional method a multiread will be mapped to a genomic location with a high probability as long as that location possesses a large number of unique reads, even when the unique reads are relatively poorly matched to the location. In other words, the proportional method ignores the matching quality between the unique reads and the reference genome, hence it is unable to account for the source of variation (ii) listed in chapter 1.

We propose a Bayesian mapping of multireads (BM-Map) approach that computes a posterior probability of mapping each multiread to a genomic location. Unlike the proportional method which only considers the equally best aligned genomic locations, the BM-Map evaluates genomic locations with unequal numbers of mismatches to a multiread. More importantly, the BM-Map utilizes three sources of information when mapping the multireads: the sequencing error profiles, the likelihood of hidden nucleotide variations, and the expression levels of competing genomic locations. In contrast, the proportional method only uses the last source of information. The key idea of the BM-Map is to use the base-level error rates and the observed mismatch profiles from unique reads to estimate the error rate due to the hidden nucleotide
variations in a hierarchical model. In the end, the BM-Map assigns multireads to competing genomic locations based on posterior probabilities.

2.2 Yeast RNA-Seq data

We demonstrate our methodology with a yeast RNA-Seq data set from Nagalaskshmi et al. (2008)[3], in which a total of 22.4 million reads, each with $K = 35$ bases, are generated using the Solexa Genome Analyzer. We apply Bowtie (version 10.0.1; [36]) to process the initial read alignment, allowing a read to be mapped to multiple locations. To include as much information as possible, we consider the multireads that are mapped to a genomic location with up to three mismatches (industry standard is up to two mismatches). Consequently, we refer to a genomic location as a *hit* if a read is mapped to that location with no more than three mismatches. A top hit is a genomic location to which a read is mapped with the least number of mismatches. The following is assumed in preprocessing the data: (1) a read is considered *mappable* if it has a top hit with no more than two mismatches; (2) given the top hit of a mappable read, other hits with no more than two extra mismatches and no more than three total mismatches are defined as additional competing genomic locations (we did not include any hits with more than three total mismatches because Bowtie only outputs hits with up to three mismatches); (3) reads with more than five competing locations are excluded because these reads likely originate from repetitive elements in the genome. With these criteria, we obtain 6,912,733 mappable reads, and among them, 5,256,339 are unique reads (those with only one hit) and 1,656,394 are multireads (with more than one hit).
2.3 Methodology

2.3.1 Probability model

We propose statistical models for a single multiread and apply the same models independently to each of the remaining multireads. This strategy allows us to analyze the ~1.6 million multireads in parallel, achieving reasonable computational speed. For a given multiread, suppose that it is mappable to \( T \) genomic locations, indexed by \( t = 1, \ldots, T \). Each genomic location corresponds to a segment of the genome with the same length as the multiread. For the yeast data, the length \( K = 35 \). We denote \( U_{kt} \equiv \{U_{l_k,t}, l_k = 1, \ldots, n_{kt}\} \) as the set of \( n_{kt} \) unique reads that overlap with location \( t \) at the \( k \)-th base, for \( t = 1, \ldots, T \) and \( k = 1, \ldots, K \). Therefore, the set \( U_t = \{U_{kt}\}_{k=1}^K \) contains all the unique reads that overlap with the genomic location \( t \). Note that here “overlap” means partial match. For example, if the location \( t \) spans from genome position 101 to 135, any overlapping unique read will have a starting base or ending base in \([101, 135]\).

[Mismatch profiles] Given the multiread, the observed data consist of the base-level mismatch profiles between the genomic location \( t \) and the multiread, and between the genomic location \( t \) and the unique reads \( U_t \), for \( t = 1, \ldots, T \). Below we introduce the labels for these mismatch profiles.

- For the multiread and location \( t \), a mismatch could occur at each of the \( K \) bases. Let \( \{e_{kt}^M = 1\} \) or \( \{e_{kt}^M = 0\} \) respectively denote that there is a mismatch or perfect match between the multiread and location \( t \) at the \( k \)-th base. Here the superscript \( ^M \) indicates that the mismatch is between a multiread and a genomic location. Then the observed data at location \( t \) for the multiread are given by a \( K \)-dimensional mismatch vector \( e_t^M = \{e_{kt}^M\}_{k=1}^K \).
• For a unique read $U_{ik,t} \in U_t$, let $\{e_{ik,t}^U = 1\}$ or $\{e_{ik,t}^U = 0\}$ respectively denote that there is a mismatch or perfect match between the unique read $U_{ik,t}$ and location $t$ at the $k$-th base. Similarly, we use the superscript $U$ to indicate that the mismatch is between a unique read and a genomic location. Define the vector of mismatch indicators at the $k$-th base for all the overlapping unique reads in $U_{kt}$ as $e_{kt}^U = \{e_{ik,t}^U\}_{i \in U_{kt}}$. Lastly, define $e_t^U = \{e_{kt}^U\}_{k=1}^K$ as the vector of mismatch indicators for all the overlapping unique reads in $U_t$ with location $t$, which is the observed data at location $t$ for its overlapping unique reads.

In summary, the full data are $\{e_t^M, e_t^U\}_{t=1}^T$.

[Parameters] We want to estimate the probability of mismatch between the $k$-th base of the genomic location $t$ and the corresponding base of a read, unique or multiple. Let that corresponding base be $r_k$. Then we denote this probability as $q_{r_k,t}$. For a multiread, $r_k = k$ since the multiread and the location fully match. For a unique read, $r_k$ usually differs from $k$ since the unique read only partially matches with the location. We will write $q_{r_k,t}$ as a function of $\alpha_{r_k}$, the sequencing error rate at base $r_k$ of the read, and $\beta_{kt}$, the probability of hidden nucleotide variations at base $k$ of the location $t$. Note that $\alpha_{r_k}$ is a read-specific parameter and $\beta_{kt}$ is a location-specific parameter. Specifically, let

$$q_{r_k,t} = \alpha_{r_k} + \beta_{kt}(1 - \alpha_{r_k}), \quad (2.3.1)$$

Model (2.2.1) essentially follows the probability law of two independent joint events. To see this, use $A$ to label the event that $\{\text{there is a sequencing error}\}$ and $B$ to label the event that $\{\text{there is a hidden nucleotide variation}\}$. Then using our notation, we have that the probability of mismatch $q = Pr(A \cup B)$, the probability of sequencing error $\alpha = Pr(A)$, and the probability of hidden nucleotide variation
\[ \beta = Pr(B). \] Model (2.2.1) says that \( q = \alpha + \beta - \alpha \beta, \) which is the probability law of two independent events \( Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A)Pr(B). \)

In our subsequent analysis of the yeast data, we fix the values of \( \alpha_{rk} \) and estimate \( \beta_{kt}. \) Recall that \( \alpha_{rk} \) is the probability of a sequencing error at the \( r_k \)-th base of a short read. We can reliably estimate it based on the sample mean of the observed sequencing error rates using the millions of unique reads in the RNA-Seq data. Due to the large sample size, rather than imposing a prior distribution on the \( \alpha_{rk}, \) we decide to fix them at their observed sample means.

**[Likelihood, prior, and posterior]** We write the likelihood contribution from the unique reads and the multireads separately. First, the contribution to the likelihood from the unique reads at location \( t \) is given by

\[
L(e^U_t) = \prod_{k=1}^K \prod_{l_h=1}^{n_{kt}} q_{rk,t}^{e^U_{lk,t}} \{1 - q_{rk,t}\}^{1-e^U_{lk,t}},
\]

where \( n_{kt} \) is the number of unique reads that overlap with the genomic location \( t \) at its \( k \)-th base. Second, define \( \{Z^M = t\} \) as the event that the multiread is mapped to genome location \( t. \) Then the contribution to the likelihood from the multiread at location \( t \) is (recall that for the multiread, \( r_k = k \))

\[
L(e^M_t) \equiv Pr\{e^M_t | Z^M = t\} = \prod_{k=1}^K q_{k,t}^{e^M_{k,t}} \{1 - q_{k,t}\}^{1-e^M_{k,t}}.
\]

Then the full likelihood is

\[
\prod_{t=1}^T L(e^U_t)L(e^M_t),
\]

in which the unknown parameters are \( \beta_{kt} \) and \( Z^M \) (the values of \( \alpha_{rk} \) are fixed).

We assume that the prior for \( \beta_{kt} \) is given by

\[
\beta_{kt} \sim B(a, b),
\]
where $B(a, b)$ represents a beta distribution with the density function proportional to $x^{a-1}(1-x)^{b-1}, \ a > 0, \ b > 0$. In our analysis for the yeast RNA-Seq data, $a = b = 1$. In addition, we assume that the prior $Pr(Z^M = t)$ is proportional to the number of unique reads mapped to location $t$. Note that this construction of $Pr(Z^M = t)$ follows the main idea in Mortazavi et al. (2008).

We want to estimate the posterior probability of mapping the multiread to location $t$, given by
\[
p^M(t) = Pr[Z^M = t \mid e^M_t, e^U_t].
\]
We can easily express $p^M(t)$ in terms of the likelihood (2.3.3) and the posterior distribution of $\beta_{kt}$ given $e^U_t$. Denoting $\beta(t) = \{\beta_{1t}, \ldots, \beta_{Kt}\}$, the vector of probabilities of mismatches at all $K$ bases of genome location $t$, we have
\[
p^M(t) = \int \left[ Pr\{Z^M = t \mid e^M_t, \beta(t)\} f[\beta(t) \mid e^U_t] \right] d\beta(t),
\]
The above equation says that the posterior probability $p^M(t)$ equals the integral of part 1 with respect to the posterior of $\beta(t)$, where part 2 is the posterior distribution of $\beta(t)$ given the observed mismatch profiles for all the unique reads. We will numerically evaluate this integral by drawing random samples from the posterior of $\beta(t)$ via Markov chain Monte Carlo (MCMC) simulations, described in Section 2.2. Suppose an MCMC sample is denoted as $\{\beta^{(s)}(t), s = 1, \ldots, S\}$ for $t = 1, \ldots, T$. We apply Bayes’ theorem to part 1 and obtain
\[
[part 1]_{\beta(t)=\beta^{(s)}(t)} = \frac{P\{e^M_t \mid Z^M = t, \beta^{(s)}(t)\}}{\sum_{t' = 1}^T P\{e^M_{t'} \mid Z^M = t', \beta^{(s)}(t)\}},
\]
which can be easily evaluated based on (2.3.3). Finally we obtain an MCMC estimate.
\( \hat{p}^M(t) \) given by
\[
\hat{p}^M(t) = \frac{1}{S} \sum_{s=1}^{S} \left[ \text{part 1} \right] | \beta(t) = \beta^{(s)}(t) \] = \frac{1}{S} \sum_{s=1}^{S} \frac{P\{e_i^M \mid Z^M = t, \beta^{(s)}(t)\}}{\sum_{t'=1}^{T} P\{e_i^M \mid Z^M = t', \beta^{(s)}(t)\}}. \tag{2.3.5}
\]

### 2.3.2 Markov chain Monte Carlo simulations

We augment the parameter space \([40]\) and employ a simple Gibbs sampler to simulate random numbers from the marginal posterior distributions of the unknown parameters \( \{\beta_{kt}\} \). The basic idea is to introduce a latent binomial variable with a conditional distribution defined by
\[
u_{kt} \mid \beta_{kt} \sim Bin\left\{ n_{kt}, \frac{(1 - \alpha_r) \beta_{kt}}{\alpha_r + (1 - \alpha_r) \beta_{kt}} \right\}.
\]

With the augmented binomial distribution, we can easily show that \( [\beta_{kt} \mid u_{kt}] \) follows a beta distribution
\[
B\{u_{kt} + a, n_{kt} - \sum_{t_k=1}^{n_{kt}} e_{l_k,t}^{(1)} + b\}.
\]

Alternating the random sampling of \([u_{kt} \mid \beta_{kt}]\) and \([\beta_{kt} \mid u_{kt}]\) in a Gibbs sampler, we obtain imputed values of \((u^{(s)}_{kt}, \beta^{(s)}_{kt})\) in the \(s\)-th iteration of the Gibbs sampler. We use the MCMC samples \((\beta^{(1)}_{kt}, \ldots, \beta^{(S)}_{kt})\) to calculate (2.3.5) and evaluate the posterior probability \( \hat{p}^M(t) \).
The proposed Gibbs sampler is as follows.

- **Step 1**: Let $\beta_{kt}^{(1)} = \epsilon$, where $\epsilon$ is an arbitrary small probability close to zero.

- **Step 2**: In the $s$-th iteration, sample $u_{kt}^{(s)}$ from

  $$Bin\left\{ n_{kt}, \frac{(1 - \alpha_r)\beta_{kt}^{(s-1)}}{\alpha_r + (1 - \alpha_r)\beta_{kt}^{(s-1)}} \right\}.$$

- **Step 3**: Sample $\beta_{kt}^{(s)}$ from $B(u_{kt}^{(s)} + a, n_{kt} - \sum_{l_k=1}^{n_{kt}} e_{l_k,t} + b)}$.

- **Step 4**: Iterate steps 1 to 3 $S$ number of times, for a large integer $S$.

For the special case in which $n_{kt} = 0$, set $u_{kt}^{(s)} = 0$ and $\beta_{kt}^{(s)} = \beta_{kt}^{(s-1)}$.

In our analysis, the number of iterations $S$ was set to 1,000 with the first 200 iterations as burn-in. The Markov chain converged fast and mixed well.

### 2.4 Simulation studies

We conducted simulation studies to evaluate the performance of the proposed BM-Map method in comparison with two other approaches.

#### 2.4.1 Simulation setup

Without loss of generality, we considered multireads potentially mapped to $T = 2$ genomic locations. We designed seven scenarios with different combinations of three factors that would affect read mapping: 1) $Diff = (Yes; No)$, defined as whether there was a true sequence difference between the two genomic locations in the reference genome; 2) $Mut = (Yes, No)$, defined as whether there was a hidden nucleotide variation (e.g., mutation) at a base belonging to one of the two genomic locations; 3) $Exp = (Yes, No)$, defined as whether the expression levels, measured as the numbers of
mapped short reads, between the two genomic locations were the same. Enumeration of the three factors would give us eight scenarios. However, the scenario in which all three factors were false was of no interest and did not provide any information about to where the multireads should be mapped. Hence, that scenario was not considered in this simulation. When \( \text{Diff} \) was Yes, we assumed that there was a sequence difference at base 18 between the two genomic locations. When \( \text{Mut} \) was Yes, we assumed that the mutation rate at base 16 of genomic location 1 was \( \beta_{16,1} = 0.9 \); when \( \text{Mut} \) was No, the rate was 0. When \( \text{Exp} \) was Yes, i.e., the expression levels were the same for the two locations, we assumed that the same number of unique reads originated from both genomic locations, and the number could have been 4, 10, or 100. When \( \text{Exp} \) was No, we assumed different numbers of unique reads from locations 1 and 2: (4,3), (10, 5), or (100,10). Therefore, combining the three sets of sample sizes for all seven scenarios, we obtained a total of 21 possible simulation cases. For each of the cases, we generated 200 multireads with a mismatch probability at the \( k \)-th base equal to \( q_{k,t}, m = 1, \ldots, 200, \text{ and } t = 1,2 \), where \( q_{k,t} = \alpha_k + (1 - \alpha_k)\beta_{kt} \). Finally, the values of \( \alpha_k \)'s were fixed at the sample means of the mismatch rates of base \( k \) using all the unique reads in the yeast data. The probability \( \beta_{kt} \) was 0 unless when \( \text{Mut} \) was true, in which case \( \beta_{16,1} = 0.9 \). The first \( 200 \times N_1/(N_1 + N_2) \) (round to an integer) multireads were assumed to originate from genomic location 1, where \( N_1 \) and \( N_2 \) were the numbers of the unique reads originated from locations 1 and 2, respectively.

### 2.4.2 Simulation results

We compared three methods for each of the 21 cases in the simulation studies.

- **BM-Map** – the proposed Bayesian method.
• Prop – the proportional method (Mortazavi et al., 2008) in which reads are mapped to the location with the fewest number of mismatches. When there are ties, the multireads are mapped to each tied location with a probability proportional to the number of unique reads mapped to that location.

• Rand – a reference method that assigns the multireads to each of the genome locations with a uniform probability.

Figure 2.2 summarizes the false discovery rates (FDRs) for the three methods after they were applied to the 200 simulated multireads originated from two genomic locations. Here the FDR is defined as the percentage of multireads being falsely mapped to a genomic location among all the multireads mapped to that location. The left panel presents the stacked 21 FDRs for mapping the multireads to genomic locations 1 and 2 under each of the three methods. The lower the bar, the better the overall performance. The BM-Map method has much lower bars than both the Prop and Rand. The right panel shows the FDRs of a list of representative cases (4,5,6,7,8,9,16,17,21). In other cases, the BM-Map and Prop methods performed equally well. For each presented case in the right panel, the six vertical bars represent the FDRs of the three methods, with the first three bars for location 1 and the next three for location 2. Almost in all the cases, the BM-Map has much smaller FDRs than the other two methods. Highlighted is case 21, in which all three factors are true with the numbers of unique reads being 100 and 10 for the two genomic locations. While the BM-Map has an FDR (0.01) at genomic location 1 comparable to that of the Prop (0.00), it has a much smaller FDR at genomic location 2 (0.06 vs. 0.32). Further examination shows that the Prop method mapped many multireads to location 2 that belonged to location 1. This is due to the fact that when there is a mismatch between
a read and the location 1 at the 16th base, the Prop method could not tell whether the mismatch was due to mutation or sequence difference. Because there was a high mutation rate 0.9 \( (\beta_{16,1} = 0.9) \), most mismatches between the multireads and the location 1 at this base were due to the mutation. The BM-Map was able to learn and borrow the information based on the mismatch profiles of the unique reads, and thus correctly map most of the multireads.

Figure 2.2: (Colored) Results summarizing the false discovery rates (FDRs) of the three methods in the simulation studies. Left panel: stacked FDRs over the 21 simulated cases in chapter 2.4. A method name followed by “-1” and “-2” respectively represents the stacked 21 FDRs for mapping multireads to locations 1 and 2. Right panel: FDRs for selected simulation cases in which there are large differences in the FDRs among the three methods.

### 2.5 Yeast RNA-Seq analysis

We present results for the analysis of the yeast RNA-Seq data described in Section 2.2.
2.5.1 Read mapping

In mapping the multireads to the yeast genome, we first identified the genomic locations with the fewest number of mismatches as the top hits. We required that the number of mismatches between the best hits and the short read be no larger than two. Additional hits are included according to the three criteria listed in Section 2.2. There are a total of 5,256,339 unique reads. The numbers of multireads with 2 to 5 hits were respectively (1,494,678,147,142,12,079,2,495).

We applied the BM-Map method to map the multireads. For each multiread, we first identified the set of unique reads for each of its candidate genomic locations. Matching the sequences of the multiread to its candidate genomic locations, we obtained the mismatch profiles \( \{ e_t^M \}_{t=1}^T \). Matching the sequences of the unique reads and their corresponding genomic locations, we obtained the mismatch profiles \( \{ e_t^U \}_{t=1}^T \). With these profiles, for each multiread indexed by \( m \), we applied the Gibbs sampler outlined in Section 2 and computed the posterior probabilities that the multiread \( m \) is mapped to location \( t \), \( \hat{p}_m(t) \) (this was \( \hat{p}_M(t) \) before without the index \( m \)). These probabilities are used to compute gene expression in downstream analyses.

For comparison, we applied the proportional method (Mortazavi et al., 2008) to map the multireads as well. Let \( n.loci \in \{2,3,4,5\} \) be the number of hits for a multiread. Let \( \hat{r}_m(t) \) denote the probability of mapping multiread \( m \) to location \( t \) using the proportional method. For each multiread \( m \) with \( n.loci \) candidates, we compute the difference in the probabilities of read mapping between the BM-Map method and the proportional method as

\[
D_m(p, r) = \sum_{t=1}^{n.loci-1} |\hat{p}_m(t) - \hat{r}_m(t)|.
\]

For example, for a multiread with \( n.loci = 2 \) hits, the proportional method might yield
\[ \hat{r}_m(1) = 0.6 \text{ and } \hat{r}_m(2) = 0.4 \text{ while the BM-Map method might yield } \hat{p}^M(1) = 0.3 \text{ and } \hat{p}^M(2) = 0.7 \; \text{we would have } D_m(p, r) = |0.6 - 0.3| = 0.3. \] Figure 2.3 presents the histogram of the log \( D_m(p, r) \) for those multireads with at least one unique read mapped to each hit, when \( n.loci = 2 \). Figure 2.4 demonstrates three representative examples of mapping multireads. The top two panels demonstrate cases in which the BM-Map and proportional methods gave contrasting results, mapping the multireads to opposing genomic locations. The bottom panel is an example in which the two methods agree. In each plot, we present the posterior mean of \( \beta_{kt} \), the probability of a hidden nucleotide variation for the \( k \)-th base at genomic location \( t \), as a function of base \( k \). A large posterior mean of \( \beta_{kt} \) implies that a large number of unique reads have mismatches at the \( k \)-th base of location \( t \), which are due to hidden nucleotide variations. We present these important phenomena in Figure 2.4.

- [Top panel] If a hit has a high probability of a hidden nucleotide variation at a base where the multiread has a mismatch, the mismatch will be down-weighted because it could be caused by a mutation. Consequently, the probability of mapping the multiread to that hit will increase due to improved matching. This is the case for the left plot in the top panel (high \( \beta_{24,1} \)).

- [Middle panel] In contrast, if a hit has a high probability of a hidden nucleotide variation at a base where the multiread has a perfect match, the perfect match will be down-weighted and the probability of mapping the multiread to that hit will decrease. This is the case for the right plot in the top panel (high \( \beta_{15,1} \) and \( \beta_{16,1} \)) and the left plot in the middle panel (high \( \beta_{2,1} \) and \( \beta_{3,1} \)).

- [Bottom panel] The bottom panel in Figure 2.4 presents a “null” case in which the BM-Map method and the proportional method give the same mapping
probabilities. In both plots, the probabilities of hidden nucleotide variations are negligible at all bases of both genomic locations. Therefore the probability of mapping the multiread is based on the numbers of mismatches between the multiread and the locations, and the number of unique reads on each location.

In summary, the first two examples above highlight the importance of borrowing the matching information between the unique reads and the genomic locations in refining the mapping of the multireads. This is a key advantage of the BM-Map method comparing over the proportional method.
Figure 2.4: (Colored) Three examples of multireads mapped to two hits based on the probabilities of the BM-Map and Prop methods. In each panel (row), the two plots correspond to two competing genomic locations for a multiread. Plotted are the posterior means of $\beta_{kt}$, the probability of hidden nucleotide variation. These are estimated based on the error mismatch profiles between the unique reads and the genomic locations. The red dots/lines indicate where the multireads have mismatches. Top panel: the multiread has mismatches at bases (23, 24) to location 1 and (12, 13) to location 2. Middle panel: the multiread has mismatches at bases (7, 22) to location 1 and (34, 35) to location 2. Bottom panel: the multiread has mismatches at bases (1, 12) to both locations. The probabilities of mapping based on the BM-Map and proportional methods are presented in each plot.
2.5.2 Gene expression quantification

The final goal in processing the yeast RNA-Seq data is to quantify gene expression. A popular approach is to compute the number of reads that map per kilobase of exon model per million mapped reads (RPKM). A larger RPKM indicates larger gene expression. A standard algorithm for computing the RPKM is given below. We compare the RPKMs based on three read mapping approaches: the BM-Map method, the proportional method, and the industry standard. The industry standard simply discards all the multireads in computing the RPKM. We call this the naive method. In the BM-Map and proportional methods, we use the probabilities of mapping a multiread to each genomic location as the count number in the first step of computing the RPKM.
Algorithm for computing the RPKM:

1. Count the number of reads (multireads and unique reads) mapped to the genes in millions. Call that number $m$.

2. Count the number of bases in all the exons of the gene in kilo-bases. Call that number $l$.

3. Count the total number of reads that have been mapped to the entire genome in millions. Call that number $t$.

4. $\text{RPKM} = \frac{m}{l/t}$.

We get one RPKM value per gene. In the BM-Map and proportional methods, we use the probabilities of mapping a multiread to each genomic location as the count number in the first step.

There are 5,862 known yeast genes, which result in 5,862 RPKMs for each of the three methods. Figure 2.5 compares the pairwise RPKMs among the three methods. The RPKMs of over 40% of the genes from the BM-Map method are different from those from the naive method (left panel), because the naive method ignores the multireads when computing the RPKM. The proportional method and the BM-Map method yield identical RPKMs for 5,049 genes out of the 5,813 genes with non-zero RPKMs from both methods. This is because for most of the 5,049 yeast genes, there are no multireads. However, when there are multireads mapped to the genes, the two methods can be very different (right panel) and have a difference up to $2^{10}$ RPKMs. We note that the lack of multireads in many yeast genes is mainly due to the relatively high sequence uniqueness in the yeast genome. This would not be the case, for example, in the human genome where gene duplication is much more
Table 2.1: A list of yeast genes with largest changes in the normalized RPKMs between the BM-Map and the proportional methods. The normalized RPKM diff is defined as $|\text{RPKM}_{\text{prop}} - \text{RPKM}_{\text{BM-Map}}| / \text{RPKM}_{\text{BM-Map}}$. The ORF and Gene names are from the saccharomyces genome database (http://www.yeastgenome.org). All the five genes listed below are duplicate genes, hence possessing sequences similar to other genes. They all play important biological functions in yeast.

<table>
<thead>
<tr>
<th>ORF name</th>
<th>Gene name</th>
<th>Protein product description</th>
<th>RPKM diff* %</th>
</tr>
</thead>
<tbody>
<tr>
<td>YLR134W</td>
<td>PDC5</td>
<td>Pyruvate deCarboxylase</td>
<td>37.5%</td>
</tr>
<tr>
<td>YPL036W</td>
<td>PMA2</td>
<td>Plasma membrane ATPase</td>
<td>24.6%</td>
</tr>
<tr>
<td>YMR121C</td>
<td>RPL15B</td>
<td>Ribosomal protein of the large subunit</td>
<td>21.2%</td>
</tr>
<tr>
<td>YJL052W</td>
<td>TDH1</td>
<td>Triose-phosphate deHydrogenase</td>
<td>20.9%</td>
</tr>
<tr>
<td>YPL081W</td>
<td>RPS9A</td>
<td>Ribosomal protein of the small subunit</td>
<td>11.1%</td>
</tr>
</tbody>
</table>

As a further evaluation of the impact of the BM-Map method on gene expression quantification, we present in Table 2.1 the normalized differences between the RPKMs from the BM-Map method and the proportional method, defined as $|\text{RPKM}_{\text{prop}} - \text{RPKM}_{\text{BM-Map}}| / \text{RPKM}_{\text{BM-Map}}$. Table 2.1 shows that the RPKMs of some genes based on the BM-Map method vary with those from the proportional method by more than 20%.
Figure 2.5: Results comparing the RPKMs from three methods. Shown are the $\log_2$ absolute differences in the RPKMs between the BM-Map and the Naive method (left panel) and between the BM-Map and the proportional method (right panel).
Chapter 3

A Bayesian Graphical Model for Integrative Analysis of TCGA Data

In this chapter, we propose a Bayesian graphical model to integrate gene-specific TCGA data including measurements on matched DNA copy numbers (C), DNA methylation (M), and mRNA expression (E) over 500 ovarian cancer samples. This Bayesian graphical model treats the three types of measurements as three nodes in a graph. Edges between the nodes infer the type of regulatory relationships in genomics. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the gene expression is sensitive to methylational variation but not copy number variation.

This chapter proceeds as follows. In chapter 3.1, we review the literature and introduce the motivation. In chapter 3.2, we give a brief overview of the ovarian cancer data to which we apply our integration analysis. In chapter 3.3, we propose our Bayesian graphical models to integrate all three platforms with MCMC simulation details. Chapter 3.4 presents several simulation studies to evaluate the performance of the proposed graphical model. In chapter 3.5, we report results based on the analysis of the TCGA ovarian cancer data introduced in chapter 3.2.
3.1 Literature Review and Motivation

Gene expression is a critical genetic process in which DNA is transcribed to RNA. Perturbation of transcription directly affects mRNA expression and hence the subsequent protein production, leading to pathological states. Genetic variations such as copy-number variations (CNVs) and DNA methylations frequently contribute to disrupted gene expression.

Current literature mainly focuses on the pair-wise integration, between CNVs and mRNA or between methylation and mRNA. Bussey et al. [41] computed Pearson’s correlation coefficients and tested the significance of correlations using false discovery rate (FDR) control. Waaijenborg et al. [42] proposed a penalized canonical correlation analysis to study genome-wide association between DNA copy number and mRNA expression. Menezes et al. [43] modeled the relationship of DNA copy number and mRNA expression by a linear model based on a modified correlation coefficient and an explorative Wilcoxon test. Choi et al. [44] described a Bayesian double-layered mixture model which directly modeled the stochastic nature of CNVs and identified abnormally expressed genes due to aberrant copy number. Etcheverry et al. [45] investigated the effect of methylation on mRNA expression in glioblastoma, and identified 13 genes that display an inverse correlation between methylation and mRNA expression using Pearson’s correlation coefficient.

Since both CNVs and DNA methylation play important roles in mRNA expression, an integrated analysis that models all three platforms together is most appropriate. Denoting with $C$, $M$, and $E$ the three platforms used to measure CNVs, methylation, and mRNA expression, we integrate data from all three platforms and present inference results as graphs that include $C$, $M$, and $E$ as three vertices. In particular, we propose a Bayesian graphical model which imposes a probability distribution on the
unknown networks and apply an autologistic prior to learn the dependence structure of three platforms through a graph. The vertices of the graph represent the platforms, and the presence or absence of edges indicates the presence or absence of conditional dependence between the platforms. For example, an edge between $M$ and $E$ and a lack of edge between $C$ and $E$ implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation. In this application, the use of a 3-node graphical model to represent the dependence structure of C, M and E is mainly chosen for convenience and for ease of display.

3.2 TCGA Ovarian Cancer Data

Ovarian cancer is ranked as the fifth leading cause of death related to reproductive cancer in women. The Cancer Genome Atlas (TCGA) Research Network has examined more than 500 tumor samples and thousands of genes. The data is publically available online [46]. Special effort has been directed to produce matched measurements on DNA copy number ($C$), DNA methylation ($M$), and mRNA expression ($E$) for all the genes across the tumor samples. Taking advantage of this effort, we use the level 3 data of measurements on ($C, M, E$) for each gene with matched tumor samples. Specifically, let $y_{itg}$ denote the measurement for gene $g$, on sample $t$, with platform $i$. Here $i = 1, 2, 3$ represents $C$, $M$, and $E$ respectively, $t$ indexes the $T = 534$ tumor samples, and $g$ indexes the $N = 9283$ genes.
3.3 Probability Model

3.3.1 Sampling Model

We apply the proposed model for individual genes separately and thus drop the index $g$ in subsequent discussion. For a single gene, the data is arranged in a $3 \times T$ matrix $Y = [y_{it}], i = 1, 2, 3$ and $t = 1, 2, \ldots, T$. We assume independence of measurements $y_{it}$ across samples. The proposed model introduces latent trinary indicators $e_{it} \in \{-1, 0, 1\}$. The indicators have an interpretation as under-, regular and over-expression of the corresponding measurement. Using $e_{it}$ we apply the mixture model proposed by Parmigiani et al. (2002) [47] for $y_{it}$. In words, we assume a mixture model with uniform, normal and uniform components corresponding to under-, regular and over-expression. The model is

\[
(y_{it} - \alpha_t - \mu_i) \mid e_{it}, \theta_{it} \sim
\]

\[
I[e_{it} = -1]U(y_{it} \mid -k_{i-}, 0) + I[e_{it} = 0]N(y_{it} \mid 0, \sigma_i^2) +
\]

\[
I[e_{it} = 1]U(y_{it} \mid 0, k_{i+}),
\]

(3.3.1)

where $I[\cdot]$ is the indicator function, $U(A)$ denotes a uniform distribution over the set $A$, and $N(\cdot, \cdot)$ denotes the normal distribution. The vector $\theta_{it} = (\alpha_t, \mu_i, \sigma_i^2, k_{i-}, k_{i+})$ collects all the other parameters. For example, $\alpha_t$ and $\mu_i$ are the random effects of sample $t$ and platform $i$. We subsequently convert the trinary variable $e_{it}$ to a binary variable $z_{it}$ with $p(e_{it} \mid z_{it} = 0) = \delta_{-1}(e_{it})$, and

\[
p(e_{it} = 0 \mid \pi_i, z_{it} = 1) = \pi_i, \quad p(e_{it} = 1 \mid \pi_i, z_{it} = 1) = 1 - \pi_i.
\]

This conversion is devised to set up the following graphical model.

Denote $V = \{1, 2, 3\}$ the set of three vertices representing $C$, $M$, and $E$. We use a graph on these three nodes to characterize the dependence structure across the three
platforms. A graph is a pair $G = \{V, S\}$ where $S$ is a set of undirected edges $\{i, j\}$, $i, j \in V$. A graph $G$ can be used to describe the conditional independence structure of a set of variables indexed by $V$, for example the binary indicators $\{z_{it}, i \in V\}$ in the case of our application. The absence of an edge $\{i, j\}$ indicates conditional independence of $z_{it}, z_{jt}$ given the remaining variables $z_{kt}, k \neq i, k \neq j$. In the case of the three platforms the set of remaining variables reduces to just the third platform.

Any joint probability model $p(z_{1t}, z_{2t}, z_{3t})$ that respects the dependence structure $G$ can be written as (Besag, 1974 [48]):

$$p(z_t \mid \beta, G) = p(0 \mid \beta, G) \times \exp \left\{ \sum_{i=1}^{3} \beta_i z_{it} + \sum_{\{i, j\} \in V, i < j} \beta_{ij} z_{it} z_{jt} \right\}$$

where $z_t = (z_{1t}, z_{2t}, z_{3t})$ and $\beta = (\beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{23}, \beta_{13})$. Coefficients $\beta_{ij}$ are non-zero only when the corresponding edge is included in the graph. Model (3.3.2) is known as the autologistic model.

Caragea and Kaiser [49] and Hughes et al. [50] proposed a centered parametrization of the autologistic model and argued that the centered version improves mixing of the Markov chain Monte Carlo (MCMC) posterior simulation and simplifies prior specification. The centered version is used in the form of

$$p(z_t \mid \beta, G) = p(0 \mid \beta, G) \times \exp \left\{ \sum_{i=1}^{3} \beta_i z_{it} + \sum_{\{i, j\} \in V, i < j} \beta_{ij} (z_{it} - \nu_i) (z_{jt} - \nu_j) \right\}$$

(3.3.3)

where $\nu_i = \exp(\beta_i)/\{1 + \exp(\beta_i)\}$.

The joint model factors as

$$p(Y, e, z, \pi, \theta, G) = p(Y \mid e, \theta)p(e \mid z, \pi)p(z \mid \beta, G)p(\theta)p(\beta \mid G)p(G)$$

(3.3.4)
We introduce the priors $p(\theta)p(\beta | G)p(G)$ next. Let $Ga(a, b)$ denote a gamma distribution with mean $a/b$. We assume conditionally conjugate priors

\begin{align*}
\mu_i &\sim N(0, \tau_\mu), \quad \frac{1}{\sigma_i^2} \sim Ga(\gamma_\sigma, \lambda_\sigma), \\
\frac{1}{k_{i-}} &\sim Ga(\gamma_{k_{i-}}, \lambda_{k_{i-}}), \quad \frac{1}{k_{i+}} \sim Ga(\gamma_{k_{i+}}, \lambda_{k_{i+}}), \\
\beta_s &\sim N(0, \sigma_\beta^2), \quad \pi_i \sim U(0, 1),
\end{align*}

where $\beta_s$ stands for the coefficients $\beta_i, \beta_{ij}$ in (3.3.3). For the sample random effects $\alpha_t$'s, we assume $\alpha_t \sim N(0, \tau_\alpha)$ subject to identifiability constraint $\sum_t \alpha_t = 0$. Lastly, we define a prior $p(G)$ as a uniform distribution over all possible graphs. With 3 vertices, we only need to consider up to 8 graphs. Each of the subgraphs is given a prior probability of 1/8.

### 3.3.2 Markov Chain Monte Carlo (MCMC) Simulations

We carry out posterior inference for model (3.3.4) using MCMC simulations. Each iteration of the MCMC scheme includes the following transition probabilities. We start by generating $z_{it}$ from its complete conditional posterior. Following the update of $z$, we generate values for $e$ from complete conditional posterior $p(e | Y, \alpha, z)$. If $z_{it} = 0$, the update is deterministic, $e_{it} = -1$. If $z_{it} = 1$, the update requires a Bernoulli draw for $e_{it} = 0$ versus $e_{it} = 1$. The update of parameters $\theta$ is straightforward. Resampling $G$ and the regression coefficients $\beta$ could be challenging in larger graphs, essentially because of the difficult evaluation of the normalization constant $p(0 | \beta, G)$ in (3.3.3) (see, e.g. [51]). However, here $p(G)$ is only supported over 8 possible graphs, making the evaluation of the normalization constant straightforward. Thus, resampling $G$ and $\beta$ reduces to straightforward trans-dimensional MCMC as in [52].
3.4 Simulation Study

To evaluate the proposed model, we examine the performance of our model with 3 simulated data sets, each with $T = 300$ samples, one true graph and a single gene. For each simulation, a true graph $G$ is first generated as follows. For a pair of vertices $\{i, j\}$, we include the edge with probability 0.5. For each imputed edge $\{i, j\}$, we generate values of $\beta_{ij} \sim N(\mu_1, 0.5^2)$ with $\mu_1 \sim U(-3, 3)$. We generate $\beta_i \sim N(\mu_2, 0.5^2)$ with $\mu_2 \sim U(-0.5, 0.5)$. Then, we generate $z$ for $T = 300$ samples. Since $p(e_{it} | z_{it} = 0) = \delta_{-1}(e_{it})$, and $p(e_{it} = 0 | \pi_i, z_{it} = 1) = \pi_i$, $p(e_{it} = 1 | \pi_i, z_{it} = 1) = 1 - \pi_i$, we first generate $\pi_i \sim U(0.25, 0.75)$ and then generate $e$. Furthermore, we let $\mu_i = 0, \sigma_i = 0.316, k_{i-} = 5.556, k_{i+} = 5.556$ for each node, and generate $\alpha_i \sim N(0, 0.1^2)$ subject to the identifiability criterion $\sum_{t} \alpha_i = 0$. Lastly, the hyper-parameters are $\tau_\alpha = 1, \tau_\mu = 1, \gamma_\sigma = 2, \lambda_\sigma = 0.1, \gamma_{k_+} = 10, \lambda_{k_+} = 50, \gamma_{k_-} = 10, \lambda_{k_-} = 50, \sigma_\beta^2 = 10$.

We implement our model to compute the posterior summaries for each simulated data set. The posterior estimates are obtained by MCMC posterior simulation with 5,000 iterations, of which 2,000 are burn-in. Since graph $G$ is modeled as a random variable, we report the inference $\xi = P(G = G_0 \mid data)$, where $G_0$ is the true graph in the simulation. For the three data sets $\xi = 0.82, 0.86, 1$, respectively. We also report parameter estimates $\bar{\beta} = E(\beta \mid Y)$ denoting the posterior mean for the autologistic coefficients.

From Figure 3.1, we can see that the estimated graph match the simulation truth for all three data sets. Here the estimated graph is the graph with highest posterior probability. We denote the positive and negative edges by black lines and red lines, respectively. The sign of $\beta_{ij}$ has an intuitively appealing interpretation related to the effect of the $j$-th platform on the probability of presence of $i$-th platform, keeping the other platform fixed. Let $z_{-ij} = z \backslash \{z_i, z_j\}$. We can show that $\beta_{ij}$ is the log odds
Figure 3.1: The simulation truth versus the estimated graph for three simulated data set. Edge colors black and red represent positive and negative relationships. The solid line represents that the edge exists. The red dotted lines indicate that the corresponding edges do not exist. The number next to each edge represents either the true value or the posterior mean of the autologistic coefficients $\beta$’s, 0 for the edges do not exist. The estimated graph based on posterior inference is identical to the simulation truth.

ratio of $z_i$ and $z_j$ through simple algebra, where $\beta_{ij} > 0$ implies that $p(z_i = 1 \mid z_j = 1, z_{-ij}) > p(z_i = 1 \mid z_j = 0, z_{-ij})$. See Figure 3.1 for the values of $\beta$’s.

3.5 Ovarian Cancer Data Analysis

We apply our model and inference method to one gene at a time using the ovarian cancer data described in chapter 3.2, aiming to recover the unknown dependence structure among the three platforms for each gene, and display it as a three-vertices graph. We carry out inference using the described MCMC posterior simulation and
ran 5,000 iterations with 2,000 burn-in. We obtain a posterior estimate $\hat{G}$ of the unknown graph with the largest posterior probability.

We can get the posterior probability of each subgraph for each gene. There are 142 genes whose $Pr(G = \hat{G} \mid data) > 0.4$. When the cutoff is set to 0.6, there are 61 genes. For cutoff = 0.8, there are only 13 genes. From these 13 gene, we select two genes “ERLIN2” and “PIR” randomly to demonstrate the results.

Figure 3.2 shows smooth scatter plots of the data for the two selected genes. Figure 3.3 displays the estimated graph for them. From these two figures, we can see that the actual trend exhibited in the scatter plot is consistent with our model estimation. For example, there is an obvious positive correlation between mRNA expression and CNVs for ERLIN2 in Figure 3.2 and the posterior mean given by our model for the mRNA expression-CNVs edge in Figure 3.3 is 7.30, indicating a strong positive correlation between the two platforms, which corresponds well with what we observed in Figure 3.2. This matching pattern is also observed for other cases. Overall, our model estimation corresponds well with the association observed among the platforms.
Figure 3.2: Smooth scatter plots of pairwise relationship among platforms $C$, $M$ and $E$. The upper panel is for gene “ERLIN2”, the low panel is for gene “PIR”. The red line in each smooth scatter plot is the lowess smoother. Dots correspond to the raw expression measurements from the level three TCGA data.
Figure 3.3: Posterior estimated graphs for genes “ERLIN2” and “PIR”. Black edges represent positive relationships and red edges represent negative relationships. The number next to each edge is the posterior mean of $\beta_{ij}$. 
Chapter 4

Nonparametric Bayesian Bi-Clustering for Next Generation Sequencing Count Data

In this chapter, we propose a nonparametric Bayesian local clustering Poisson model (NoB-LCP) to facilitate posterior inference on two-dimensional clustering for discrete count data. We consider data from ChIP-Seq experiments, which are applications of next generation sequencing (NGS) technology. The data is a matrix of histone modification (HM) counts, with rows representing genomic locations and columns representing HMs. Traditional one dimension clustering methods focus on clustering either HMs or genomic locations, but not both. Since learning the regulatory mechanisms of co-localized HMs is very crucial for understanding the biological processes of HM regulation, it is necessary for us to consider a two-dimensional clustering, here we call it biclustering.

This chapter proceeds as follows. We review the literature and introduce the motivating application in chapter 4.1. We introduce the data set in the motivating application in chapter 4.2. In chapter 4.3, we present probability models and computational methods for posterior inference. We present a simulation study in chapter 4.4, and in Section chapter 4.5, we report inference results on the ChIP-Seq data.

4.1 Literature Review and Motivation

We consider data from ChIP-Seq experiments, which are applications of next generation sequencing (NGS) technology and will be introduced later. The sequencing data
is a matrix of HM counts, with rows representing genomic locations and columns representing HMs. Traditional one-dimensional clustering techniques aim to partition either the HMs or genomic locations. While useful, such clustering methods are often inadequate to identify co-localized HMs that are important factors in deciding functions of genomic regions. In addition, how genomic regions cluster should depend on which subset of HMs we focus on. Different HM sets might partition genomic locations in different ways, which might indicate different cellular or chromatin states.

These considerations lead us to consider two-dimensional clustering. Getz (2000) [53] presented a coupled two-way clustering approach that employs hierarchical clustering to each separate dimension, combining the clustering results along each dimension in a problem-specific manner. Later, Cheng and Church (2000) [54] introduced the concept of biclustering to find biclusters within a data matrix. They proposed a quantitative measure as a guide to search for biclusters in gene expression data. Lazzeroni and Owen (2002) [55] developed the plaid model that describes gene expression data as a sum of biclusters. In their model, each bicluster contains a group of genes expressed similarly within a given set of samples, indicating the presence of a particular biological process. Turner et al. (2005) [56] proposed an improved algorithm for fitting the plaid model. Li et al. (2009) [57] reported an effective and computationally efficient biclustering algorithm, QUBIC, to identify overlapping biclusters by employing a combination of qualitative measures of gene expression data and a combinatorial optimization technique.

We extend these approaches to incorporate two important new features: first, we develop models for discrete count data as opposed to continuous measurements. Second, we introduce full model-based inference that defines a posterior probabil-
ity model for the random partitions, including a full probabilistic description of the associated uncertainties. Specifically, we propose a nonparametric Bayesian local clustering Poisson model (NoB-LCP) to close this gap in the existing literature. The proposed method builds on Lee et al. (2012) [58] who developed bi-directional clustering for continuous protein activation data. The proposed NoB-LCP model clusters any two HMs (columns) together if they give rise to the same partition of genomic locations. That is, the partitions of genomic locations (rows) are nested within clusters of HMs, with a separate partition of locations for each HM cluster. This definition of HM clusters based on inducing the same (nested) clustering of genomic locations distinguishes the proposed model from most currently used models, including Bayesian nonparametric approaches, that define clusters based on common parameters in the sampling model. We will refer to the column clusters as “HM clusters” and to the row clusters as “location clusters”. Location clusters can be used to define different functional signatures that are characterized by subsets of HMs, while HM clusters suggest unique combinatory patterns that annotate chromatin states. One advantage of nonparametric Bayesian clustering is that it provides model-based posterior probability models for the random partitions. It entirely avoids the problem of specifying the number of clusters in advance. Another key difference between NoB-LCP and other biclustering methods is that we allow that some HMs and some genomic locations might not meaningfully cluster with the other HMs or locations. In practice, experimental data usually include noisy rows and/or columns that are irrelevant to the scientific problem being addressed. Excluding them significantly increases the power of detecting meaningful signals in the remaining rows and columns.
4.2 ChIP-Seq Data

We consider a ChIP-Seq experiment for CD4+ T lymphocytes [59, 60], in which 39 types of HMs, including 18 acetylations, 20 methylations, and a special histone modification H2A.Z, are reported. We focus on genomic locations with at least one enriched HM for meaningful inference and use the peak-calling program SICER [61] to decide enrichment. SICER parameters were set to W\_SIZE=200, GAP\_SIZE=600, EVALUE=1000, FRAG\_SIZE=150. Also any adjacent windows with unchanged SICER calls for the 39 HM counts are merged to create larger regions.

4.3 Methodology

4.3.1 Probability Model

The ChIP-Seq data is arranged in an $N \times G$ matrix $Y = [y_{ig}]$ with each element $y_{ig}$ representing the read count for HM $g$ in genomic location $i$, $i = 1, 2, \ldots, N$ and $g = 1, 2, \ldots, G$. Here, genomic locations are defined as windows of 200 base pairs.

We start the model construction with a random partition of HMs $\{1, \ldots, G\}$ into non-overlapping subsets $C_q$ as $\{1, \ldots, G\} = \bigcup_{q=0}^{Q} C_q$. The unusual indexing starting with $q = 0$ is in anticipation of the upcoming discussion. The number $Q + 1$ of subsets is random itself. It is part of the random partition $\{C_0, \ldots, C_Q\}$. In the following discussion we find it convenient to index the partition equivalently by cluster membership indicators $c_g$, $g = 1, \ldots, G$ with $c_g = q$ if $g \in C_q$. Under the proposed model some HMs are singled out as not giving rise to a nested partition of genomic locations. We refer to these HMs as the “idle HMs”, and to the remaining ones as “active HMs”. We use the special cluster $C_0$ to combine the idle HMs, i.e., $c_g = 0$ for all idle HMs. Assume that there are $G' < G$ active HMs and $(G - G')$ idle HMs. We
propose a zero-enriched Pólya urn ([62]) prior for \( c = (c_1, c_2, \ldots, c_G)^T \):

\[
P(c) = \pi_0^G (1 - \pi_0)^{G-G'} \frac{\alpha^Q \prod_{q=1}^Q \Gamma(p_q)}{\prod_{g=1}^G \Gamma(\alpha + g - 1)},
\]

where \( p_q \) is the number of HMs in HM set \( q \) and \( \alpha \) is the total mass parameter of the Pólya urn scheme. Under this model, \( c_g = 0 \) with probability \( (1 - \pi_0) \), i.e., HM \( g \) falls into the idle HM set with probability \( (1 - \pi_0) \). When \( c_g \) is non-zero, HM \( g \) is either assigned to an existing active HM set \( q \) with probability proportional to \( p_q \), or assigned to a new singleton active HM set with probability proportional to \( \alpha \). We refer to (4.3.1) as a nonparametric Bayesian prior model. The Pólya urn is traditionally considered a nonparametric Bayesian model since it can be constructed as the partition that is implied by the ties under i.i.d. sampling from a probability measure with a Dirichlet process prior. See, for example, a recent review by Lee et al. (2013) [63].

Next, we consider clustering of genomic locations for each of the \( Q \) active HM sets. Recall that the partition of locations is nested within HM sets, i.e., we want to allow for a different set of location clusters with respect to each HM cluster. We define \( r_q = (r_{q1}, r_{q2}, \ldots, r_{qN})^T \) to be the \( N \) cluster labels \( r_{qi} \in \{0, \ldots, D_q\} \) that describe the partition of genomic locations corresponding to the \( q \)-th HM set. Again we allow for a special cluster \( r_{qi} = 0 \) of inactive genomic locations that do not meaningfully co-cluster with other loci with respect to the \( q \)-th HM set. We assume that \( r_q \) includes \( D_q \) active location clusters with \( r_{qi} = d \) indicating that locus \( i \) is assigned to active location cluster \( d \), and \( r_{qi} = 0 \) indicating that genomic location \( i \) is assigned to the idle location cluster. Let \( r = (r_1^T, \ldots, r_Q^T)^T \). We assume independent zero-enriched Pólya urn priors for each \( r_q \) given by

\[
P(r \mid c) = \prod_{q=1}^Q P(r_q) \quad \text{and} \quad P(r_q) = \pi_1^{m_q} (1 - \pi_1)^{N-m_q} \frac{\beta^{D_q} \prod_{d=1}^{D_q} \Gamma(n_{qd})}{\prod_{i=1}^{m_q} \Gamma(\beta + i - 1)}. \]

\[(4.3.2)\]
Note that $Q$ is random and depends on $c$. In (4.3.2), for a given active HM set $q$, $n_{qd}$ is the number of genomic locations in the active location cluster $d(> 0)$ and 
\[ m_q = \sum_{d=1}^{D} n_{qd}. \]
In addition, $\beta$ is the total mass parameter of the Pólya urn. The cluster label $r_{qi}$ is allowed to be 0 with probability $(1 - \pi_1)$, characterizing the idle location cluster.

The described prior probability model can be characterized as a partition of HMs and a nested partition of locations, nested within each (active) cluster of HMs. In words, we identify subsets of HMs that are characterized by the fact that genomic locations cluster into the same subsets with respect to all HMs in a HM cluster. These subsets will provide important information on the co-location patterns of HMs and actionable target HMs for diagnosis and prognosis. In addition, the resulting clusterings of genomic regions can be examined and integrated with other information (e.g., transcription binding sites) to potentially achieve better understanding of gene regulation.

Given $c$ and $r$, we now define a sampling model for the observed counts $y_{ig}$. Let $\text{Poi}(\theta)$ denote a Poisson distribution with mean $\theta$. We start with a Poisson sampling model for the count data, i.e.,
\[ y_{ig} \sim \text{Poi}(\theta_{ig}). \]

The prior probability model for $\theta_{ig}$ makes use of the clustering. Let $\text{Gamma}(a, b)$ denote a gamma distribution with mean $a/b$. We define $P(\theta_{ig} \mid c, r)$ as follows. Assume $c_g = q$ and $r_{qi} = d$. The model gives meaning to the partition of locations by assuming a shared rate $\tilde{\theta}_{dg}$ for all locations in the same location cluster, i.e., $\theta_{ig} = \tilde{\theta}_{dg}$ for all $i$ with $r_{qi} = d$. But HMs in the same HM cluster share the same partition of locations only, not the same rate, i.e., $\theta_{jh} = \tilde{\theta}_{dh} \neq \tilde{\theta}_{dg}$ for all $(h, j)$ with $c_h = q$ and
$$r_{qj} = d \text{ and } h \neq g.$$ We assume

$$\tilde{\theta}_{dg} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k_{0g}, \lambda_{0g}),$$

For the idle genomic locations in the active HM sets, i.e., $$r_{qi} = 0$$ with $$q > 0$$, we assume a priori $$\theta_{ig} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k_{1g}, \lambda_{1g})$$. For idle HMs, i.e., $$c_g = 0$$, we assume $$\theta_{ig} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k_{2g}, \lambda_{2g})$$ for all locations $$i$$. Note that taking a Poisson sampling model with parameter $$\theta_{ig}$$ and a gamma prior for $$\theta_{ig}$$, we equivalently constructed a negative binomial sampling model for the count data, which provides additional variabilities to account for potential over dispersion.

Finally, denoting with Beta($$a, b$$) a beta distribution with parameters ($$a, b$$), we assume conditionally conjugate priors

$$\pi_0 \sim \text{Beta}(a_0, b_0), \quad \pi_1 \sim \text{Beta}(a_1, b_1).$$

The beta hyperprior on $$\pi_0$$ and $$\pi_1$$ is important to allow for inference about the number of active HMs and locations, as it allows adjustment of the priors $$p(c \mid \pi_0)$$ and $$p(r_q \mid c, \pi_1)$$ to adapt to the level of noise in the data. See, for example [64] for a discussion of this multiplicity correction feature.

Figure 4.1 is a graphical illustration of the proposed NoB-LCP model. It demonstrates the core idea of how we define local clusters. In Figure 4.1, we assume that 9 HMs belong to two active HM sets and and an idle HM set, including HMs 5, 8 and 9. In the two active HM sets, cells in off-white are idle genomic locations. The rest of cells marked with the same color in the same column form local clusters of genomic locations (rows). Different colors indicate different values of parameters $$\tilde{\theta}_{dg}$$. Within each local cluster, the colors are the same across the genomic locations but different across different HMs. We define an active HM set as the set of HMs that partition the genomic locations in the same way, regardless of the actual values of $$\tilde{\theta}_{dg}$$. This
highlights the important difference between NoB-LCP and other clustering methods that often assume common values of $\theta_d$ for items in the same cluster. In other words, in Figure 4.1, the cells in each local cluster would be marked in the same color across both genomic locations and HMs.

In summary, the joint model is:

$$P(Y, c, r, \theta, k, \lambda, \pi_0, \pi_1) = P(Y | \theta)P(\theta | c, r, k, \lambda)P(r | c)P(c)P(\pi_0)P(\pi_1).$$

(4.3.3)

### 4.3.2 Markov Chain Monte Carlo Simulations

We carry out posterior inference using MCMC simulation. Letting $[x | y, z]$ generically denote a transition density that updates an unknown parameter $x$ conditional on currently imputed values for $y$ and $z$, we propose a Gibbs sampler that iterates over the following sampling steps that draw random values from the transition densities:

$$[r | Y, c, \pi_1], [c | Y, r, \pi_0], [\theta | Y, c, r], [\pi_0 | c], [\pi_1 | c, r]$$

We start by generating $r_{qi}$, $q = 1, \ldots, G'$, $i = 1, \ldots, N$, from its full conditional posterior distribution. When resampling $r_{qi}$ and $c_g$, we marginalize over $\theta$.

Let $Q$ denote the currently imputed number of active HM clusters. A challenge in constructing a valid transition probability arises when $c_g = Q + 1$ is considered, i.e., when we consider placing $g$ into a new, $(Q + 1)$-th, singleton HM cluster. The problem is that a proposal $c_g = Q + 1$ gives rise to a new partition $r_{Q+1}$ of locations. We use the pseudo prior mechanism of Carlin and Chib (1995) [65] to construct an MCMC scheme. We introduce a set of auxiliary variables $\tilde{r}_g = (\tilde{r}_{ig}, i = 1, 2, \ldots, N)$, $g = 1, 2, \ldots, G$, and augment the probability model with a pseudo prior $P(\tilde{r}_g | \pi_1)$. Let $p_{1g}(r | \pi_1)$ denote the conditional posterior of the location partition with respect
Figure 4.1: An illustration of the proposed NoB-LCP model with 9 HMs and 10 genomic locations. There are two active HM sets and an idle HM set, including HMs 5, 8, 9. In the two active HM sets, cells in off-white are idle genomic locations. The rest of the cells marked with the same color in the same column form local clusters of genomic locations (rows). Different colors indicate different values of parameters $\tilde{\theta}_{dg}$. 
to a singleton HM cluster \( \{g\} \). We define \( P(\tilde{r}_g \mid \pi_1) = p_{1g}(\tilde{r}_g \mid \pi_1) \). Think of \( \tilde{r}_g \) as a potential genomic location partition with respect to a singleton HM set \( \{g\} \). In other words, when a new singleton HM set is proposed for \( c_g \), the proposal distribution for the genomic location clusters under this new HM set is determined by imputed value \( \tilde{r}_g \). Lastly we draw \( \theta, \pi_0 \) and \( \pi_1 \) whose full conditional posterior distributions are in closed forms. More MCMC technical details are included in the Appendix.

### 4.3.3 Posterior Inference

A practical challenge related to posterior inference is the need to summarize a distribution over random partitions. Medvedovic et al. (2004) [66] initially addressed this problem by estimating posterior probabilities that any two HMs are clustered together. They evaluated probabilities \( H_{gh} = P(c_g = c_h \mid \text{data}) \) of pair-wise co-clustering, and used \( H \) as a distance matrix for a (deterministic) hierarchical clustering algorithm. Alternatively, Dahl (2006) [67] proposed a point estimate of a random partition under a Dirichlet process mixture model by reporting a least-squares partition. Specifically, the least-squares clustering \( c^{LS} \) is the observed clustering \( c \) which minimizes the Frobenius distance (\( L_2 \) norm for matrices) between \( S^c \) and \( H \), where \( S^c \) is an association \( G \times G \) matrix whose \( (g, g') \) element is an indicator that HM \( g \) is clustered with HM \( g' \). We include HMs in the idle HM set by letting \( s^c_{g,g'} = 0 \) for all \( g' \) if \( c_g = 0 \). Following Dahl (2006) [67], we propose a least-square summary

\[
\text{c}^{LS} = \arg \min_c || S^c - H ||^2,
\]

as a point estimate of the clustering of HMs, which minimizes the sum of the squared deviation of association matrix \( S \) from the matrix \( H \) of the posterior pairwise co-clustering probabilities. Given \( c^{LS} \), we compute \( r^{LS}_q \), the least square estimate of the clustering for genomic locations, based on the same formulation.
4.4 Simulation Studies

4.4.1 Simulation setup

We conducted simulation studies to evaluate the performance of the proposed NoB-LCP model. We compared posterior inference with the simulation truth and with inference under two alternative clustering methods, the plaid model and the QUBIC. Furthermore, to show the importance of zero-enriched Pólya urn priors which allow some HMs or genomic locations to be idle, we performed a sensitivity analysis by using regular Pólya urn priors without zero-enrichment as the prior for the random partitions of HMs and genomic locations. It means that we let \( \pi_0 = 1 \) and \( \pi_1 = 1 \) in (4.3.1) and (4.3.2) respectively.

We simulated a data matrix \( \mathbf{Y} \) with \( N = 300 \) genomic locations and \( G = 18 \) HMs. We let 13 out of 18 HMs belong to two active HM sets, in which HMs 1-7 belonged to set 1 and HMs 8-13 to set 2. The remaining 5 HMs, HMs 14–18, belonged to the idle HM set. We assumed that the active HM set 1 partitioned the genomic locations into four location clusters including one idle location cluster, i.e., \( D_1 = 3 \), and that the active HM set 2 partitioned the genomic locations into three location clusters including one idle location cluster, i.e., \( D_2 = 2 \). We generated location cluster labels, \( r_{qi} \), for each active HM set assuming that a genomic location belonged to one of the location clusters with equal probability. In keeping with the definition of the idle HM set \( (q = 0) \), we did not generate location clusters with respect to the idle HMs with \( c_g = 0 \). We fixed \( \tilde{\theta}_{dg} \) for all the active location clusters for each of the 13 HMs residing in the active HM set as listed in Table 4.1. Finally, denoting with \( \text{NB}(\text{mean}=a, \text{size}=b) \) a negative binomial distribution with \( \text{mean}=a \), \( \text{variance}=a + a^2/b \) and with \( \text{Unif}(0,1) \) a Uniform distribution on \((0,1)\),
Table 4.1: The true mean counts for active genomic location clusters, $\hat{\theta}_{dg}$, in the simulated data.

<table>
<thead>
<tr>
<th></th>
<th>HM 1</th>
<th>HM 2</th>
<th>HM 3</th>
<th>HM 4</th>
<th>HM 5</th>
<th>HM 6</th>
<th>HM 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster 1</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>cluster 2</td>
<td>15</td>
<td>7</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>cluster 3</td>
<td>13</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>HM 8</th>
<th>HM 9</th>
<th>HM 10</th>
<th>HM 11</th>
<th>HM 12</th>
<th>HM 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster 1</td>
<td>9</td>
<td>15</td>
<td>9</td>
<td>11</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>cluster 2</td>
<td>11</td>
<td>11</td>
<td>15</td>
<td>9</td>
<td>13</td>
<td>7</td>
</tr>
</tbody>
</table>

the remaining $\theta_{ig}$ were independently generated from NB(mean = $\mu$, size = 1), where $\mu \sim$ Unif(0, 10), including the idle genomic locations in the active HM sets and all the genomic locations in the idle HM set. The NB distribution was chosen to examine the sensitivity of posterior inference with respect to deviations from the assumed Poisson sampling model.

4.4.2 Simulation Results

The left panel of Figure 4.2 shows the heatmaps of $y_{ig}$ under the simulation truth. After rearranging the HMs and the genomic locations within each active HM set according to the simulation truth, we can clearly observe the local clustering patterns in the data. In the active HM sets, the idle genomic locations, which are located in the first row block, do not show a noticeable pattern: the colors are more or less randomly
scattered. In contrast, the active genomic locations in the columns corresponding to active HM sets show clear patterns and the colors are more homogeneous within each location cluster. In the idle HM set, since the genomic locations do not cluster, the corresponding color mapping exhibits large variability.

We applied the proposed NoB-LCP model to the simulated data. In the MCMC posterior simulation, we initialized the HMs allocation variable \( c \) using the clustering result from hierarchical clustering by cutting the dendrogram to achieve two active HM sets and one idle HM set. HMs 2, 4, 5 and 6 belonged to active HM set 1, HMs 8, 11 and 13 belonged to active HM set 2 and the remaining belonged to the idle HM set. The initial values and priors of \( \pi_0 \) and \( \pi_1 \) were set to 0.5 and Beta(1, 1), respectively. We fixed parameters \( k_{0g} \) and \( \lambda_{0g} \) by setting the mean of \( \tilde{\theta}_{dg} \) equal to \( g \)-th column mean of \( Y \) and setting the variance of \( \tilde{\theta}_{dg} \) equal to 10. Finally, \( k_{1g}, \lambda_{1g}, k_{2g} \) and \( \lambda_{2g} \) were computed by setting the mean of \( \theta_{ig} \) equal to \( g \)-th column mean of \( Y \) and variance equal to 50. After 10,000 MCMC iterations with 5,000 burn-in, the Markov chains converged and mixed well. We conducted convergence diagnostics using the R package \textit{coda} and found no evidence for convergence problems. Traceplots and empirical autocorrelation plots (not shown) for the imputed parameters indicate a well mixing Markov chain. For example, the empirical autocorrelation of \( \pi_0 \) and \( \pi_1 \) is practically zero beyond lag 2. The simulation was carried out on a MacBook Pro laptop with 2.53 GHz Intel Core and 8GB memory. Computation was completed in 2.5 hours.

The least-squares summary of the posterior on \( c \) was \( c^{LS} = (1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0) \). Conditional on \( c^{LS} \), we further calculated the least-squares estimates of genomic location clusters for active HM sets, \( r_q^{LS}, q = 1, 2 \). Figure 4.2 right panel shows that the NoB-LCP model correctly detected the two active HM sets in the
Simulation Truth

Posterior Estimate

Figure 4.2: Heatmaps of the HM sets in the simulation truth versus the identified HM sets under the NoB-LCP model. The first row block of each active HM set is the idle genomic location cluster for that HM set. The remaining blocks are active genomic location clusters. The division of genomic location clusters is indicated by white horizontal lines.

Simulation data: HMs 1-7 belonged to the active HM set 1 and HMs 8-13 belonged to the active HM set 2, the remaining HMs belonged to the idle HM set, consistent with the simulation truth. Tables 4.2 and 4.3 show that there are five estimated active genomic location clusters and one idle genomic location cluster for HM set 1, where clusters \{0, 1, 2, 3\} dominate and largely overlap with the four true genomic location clusters. And the model identified four active genomic location clusters and one idle genomic location cluster with respect to HM cluster 2, where clusters \{0, 1, 2\} dominated and largely overlapped with the three true genomic location clusters of true HM set 2.

For comparison, the two alternative methods, the plaid model and QUBIC, were applied to the same simulated data. Figure 4.3 shows the heatmaps of HMs in two
<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>active HM set 1</strong></td>
<td>89</td>
<td>75</td>
<td>63</td>
<td>64</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td><strong>active HM set 2</strong></td>
<td>91</td>
<td>101</td>
<td>18</td>
<td>89</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4.2: The number of genomic locations in each genomic location cluster for active HM sets.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>c_{1}^{LS}</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>c^{TRUE}</strong></td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>d=0</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d=1</td>
<td>3</td>
<td>68</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d=2</td>
<td>5</td>
<td>0</td>
<td>61</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d=3</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>60</td>
<td>0</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Comparisons of the location cluster membership estimated by the NoB-LCP model with the true location cluster membership.
biclusters identified by the plaid model. The first bicluster included 18 genomic locations of HMs 8, 9, 10, 11 and 12, all of which belonged to true genomic location cluster \( d = 0 \) of true HM set 2. The second bicluster included 37 genomic locations of HMs 14 and 15, which belonged to the idle HM set under the simulation truth. The QUBIC method detected 23 biclusters, 17 of which only included one single HM and the other six included two HMs. Figure 4.4 shows the heatmaps of HMs in the six biclusters with two HMs. Some of those six biclusters included idle HMs such as HMs 14, 16 and 17, and others included either idle genomic locations, active genomic locations, or multiple active location sets. For example, bicluster 1 included 15 genomic locations of HMs 4 and 7, among which 7 belonged to the true genomic location cluster \( d = 1 \) of true HM set 1, and 5 belonged to the true genomic location cluster \( d = 2 \) of true HM set 1; bicluster 2 included 21 genomic locations of HMs 8 and 10, among which 17 belonged to the true genomic location cluster \( d = 2 \) of true HM set 2.

Next we replaced the zero-enriched Pólya urn priors in (4.3.1) and (4.3.2) with regular Pólya urn priors. And we used the same hyperparameters and initialized the parameters as before, except for \( c \). We initialized \( c \) by letting HMs 1-13 belong to active HM set 1 and HMs 14-18 belong to active set 2. After 10,000 iterations of MCMC simulation with 5,000 burn-in, the Markov chains converged and mixed well.

The least-squares summary of the posterior on \( c \) was \( c^{LS} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2) \). Conditional on \( c^{LS} \), we further calculated the least-squares estimates of genomic location clusters for active HM sets, \( r_q^{LS}, q = 1, 2 \). Figure 4.5 shows the heatmaps of two detected active HM sets. Compared to the simulation truth, the model with regular Pólya urn priors failed to differentiate the two active HM sets. In addition, many small and meaningless genomic location clusters nested within two
Figure 4.3: Heatmaps of HMs in two biclusters of the simulated data identified by the plaid model. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.
Figure 4.4: Heatmaps of HMs in six biclusters for the simulated data identified by QUBIC. The division of genomic locations is indicated by white horizontal lines. Below the white line is the detected bi-cluster.
4.5 ChIP-Seq Data Analysis

We present local clustering results for the ChIP-Seq data described in chapter 4.2. For demonstration purpose, we apply our NoB-LCP model to clustering of promoters and insulators, both of which are important regulatory elements. Information on active HM sets can be observed.
the genomic location for promoters was obtained from the UCSC Genome Browser [68]. Read counts were recorded for all genomic locations and all HMs. The insulator information was obtained from the CTCFBSDB [69], a CTCF binding site database to identify insulators.

We consider a small subset of the ChIP-Seq data covering randomly selected 50 genomic locations in promoter regions and 50 genomic locations in insulator regions. The data is a $100 \times 39$ matrix with genomic locations as rows and HMs as columns. To fit the NoB-LCP model, $c$ is initialized by the clustering determined by a (deterministic) hierarchical clustering algorithm. We chose parameters $k_{0g}$ and $\lambda_{0g}$ by fixing the prior variance of $\tilde{\theta}_{dg}$ at $\text{Var}(\tilde{\theta}_{dg}) = 10$, and matching the mean of $\tilde{\theta}_{dg}$ with the column means of the data matrix. Similarly, $k_{1g}$, $\lambda_{1g}$, $k_{2g}$ and $\lambda_{2g}$ are chosen by fixing the prior variance at 50, and matching the prior mean of $\theta_{ig}$ with the column means. Finally, $\pi_0$ and $\pi_1$ are initially set to 0.5 and we used $a_0 = b_0 = a_1 = b_1 = 1$, i.e., uniform hyperpriors. After 10,000 iterations with a 5,000 burn-in for MCMC posterior simulation, we evaluated convergence diagnostics (R package *coda*) and found no evidence for practical convergence problems. The chain mixed well.

We compute the least-squares estimates $c^{LS}$ and $r^{LS}$ to summarize posterior inference. The NoB-LCP model identifies 3 active HM sets, each of which partitions genomic locations differently. Figure 4.6 shows the heatmaps of all active HM sets. These three sets are candidates of co-localized HMs that relate to gene transcription. In addition, the heatmap shows genomic location clusters nested in each active HM set.

Posterior inference distinguishes different types of regulatory elements and clusters similar types together reasonably well. For example, active HM set 1 includes the following HMs: H4K12ac, H3K79me2, H3K79me3. Genomic location clusters 1 and
Figure 4.6: Heatmaps of three active HM sets for ChIP-Seq data. White horizontal lines indicate division of location clusters.
5 in active HM set 1 include only promoter regions, in which H4K12ac, H3K79me2 and H3K79me3 clearly show relatively high expression in Figure 4.6. Our results are consistent with previous findings that H4K12ac counts are elevated in the promoter and transcribed regions of active genes [60], H3K79me2 and H3K79me3 are important histone markers for the prediction of promoter regions [70, 71]. Out of the 12 HMs in active HM set 2, all of them are acetylations; out of the 21 HMs in active HM set 3, 15 of them are methylations. From this fact, we can conjecture that the same types of histone modifications (methylations, acetylations, etc.) are more likely to be clustered together.

In addition, highly correlated HM patterns can be identified by our model. For example, active HM set 2 includes the following HMs: H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K18ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, and H4K91ac, which were reported to have relatively high correlation according to Wang et al. (2008) [60].

For comparison, we applied the plaid model and QUBIC to the same ChIP-Seq data. The plaid model did not report any biclusters. QUBIC found 57 biclusters, but none of them provide us clear divisions of regulatory elements. In addition, it is not easy to extract useful information from so many biclusters.

Finally, we used a qq-plot to validate the assumed sampling model. Assuming a Poisson/gamma hierarchical sampling model, we have implicitly defined a negative binomial marginal sampling model. The negative binomial model allows larger variabilities in modeling the counts. We made a qq-plot of the empirical c.d.f. of the observed ChIP-Seq data versus simulated data sampled from the imputed negative binomial distribution. We can see a linear relationship between two quantiles, suggesting that the hierarchical sampling model is well calibrated (Figure not shown).
Chapter 5

A Latent Gaussian Process Model with Application to Monitoring Clinical Trials

In this chapter, we propose a latent Gaussian process (LGP) to model discrete longitudinal data with recurrent or cyclic patterns. We consider modeling longitudinal discrete data that show recurrent or cyclic patterns. For example, a data vector of interest could be binary values observed on consecutive time points, and the vector values alternate between 1’s and 0’s. This type of data arises from clinical trials in many therapeutic areas, such as auto-immune diseases like multiple sclerosis or Schizophrenia. Typically, treatments need to be applied multiple times in hopes to slow down disease progression. When multiple drugs are to be evaluated, it is challenging to quantify and compare the efficacy of the drugs due to the temporal and periodic nature of disease progression and treatment response. For example, when two drugs are compared in a clinical trial involving one new treatment and a standard treatment, therapeutic effects are usually assessed by comparing the relapse rates of treatments within a time framework. However, that strategy might not be ideal since a treatment with a slightly inflated relapse rate might still be preferred if it allows patients to stay in disease remission for a longer period of time. In this chapter, we consider an important design issue for trials of the same nature.

This chapter proceeds as follows. Chapter 5.1 reviews the literature and introduces our motivating examples. In chapter 5.2, we present probability models and computational methods based on the posterior distributions. In chapter 5.3, we pro-
pose inference for predicting future responses and introduce stopping rules as part of
the trial design. We examine the performance of LGP through extensive simulation
studies in chapter 5.4. In chapter 5.5, we report numerical results based on the lupus
trial. In chapter 5.6, we derive posterior consistency results.

5.1 Literature Review and Motivation

In this chapter, we consider an important design issue for trials in which the data
show recurrent patterns. For each patient, multiple responses to treatments and mul-
tiple disease relapses are expected. One goal is to demonstrate that one treatment
allows patients to stay in disease remission longer than the other, and a second goal is
to predict future response status for individual patients based on observed outcomes.
A typical patient response vector may look like \((1, 1, 0, 0, 1, 1, 1, 0, 0, \ldots)\) with binary
outcomes at multiple time points, where “1” represents a response and “0” a nonre-
response. In addition, each binary outcome is associated with a time point at which
the outcome is recorded. Therefore, the data is summarized as \(\{e(t_k), k = 1, \ldots, K\}\)
where \(e(t_k) \in \{0, 1\}\) is binary. In Comi et al. (2012) [72], a response \(\{e(t_k) = 1\}\) is
defined as a combination of cognitive improvements assessed by psychiatric tests. An
important observation of this type of trial is that the underlying disease progression
is a continuous process over time, although measurements of outcome can only be
taken at discrete time points, e.g., once a month.

To model a latent process, Zeger et al. (1985) [73] considered an extension of lo-
gistic regression to binary longitudinal observations. However, such method was only
applied to stationary binary series. Czado and Song (2008) [74] developed a state
space mixed model for binary longitudinal observations using the standard linear
state-space formulation [75]. While useful, the linear state-space models are lim-
ited and may not work well on the aforementioned trials with nonlinear and cyclic responses. More importantly, standard models for longitudinal data cannot accommodate all the requirements in our case, including 1) to model cyclic binary responses, 2) to forecast for each patient based on the cyclic pattern, 3) to compare population patterns between different conditions (treatment v.s. control), 4) to reflect the underlying continuous disease progression mechanism. To this end, we consider a latent stochastic process with cyclic features to describe the relapsing nature of the disease. Specifically, let \( a_j(t) = \mu(t) + \tau_j(t) \) be a sum of the mean process \( \mu(t) \) describing the treatment effect over time, and a cyclic process \( \tau_j(t) \) describing the subject-\( j \)-specific recurrent disease progression. For example, the treatment effect \( \mu(t) \) could be increasing over time due to the continuous usage of the drug but \( \tau_j(t) \), to be modeled as the latent Gaussian process (LGP), could be cyclic mimicking the recurrent disease progression as a result of the battle between the disease-causing antigens and the disease-fighting immune cells. We model the binary outcome at a time point \( t_k \) as an indicator \( e_j(t_k) = I\{a_j(t_k) > a_h\} \) with a fixed and arbitrary threshold value \( a_h \). This construction can be considered a stochastic-process version of the latent probit model in Albert and Chib (1993) [76]. We defer the construction of \( a_j(t) \) to Section 5.2.

The key idea of our proposed LGP is to model observed longitudinal discrete outcomes by thresholding the latent variates that follow a GP prior. We consider two motivating examples. The first example is a randomized, double-blind, placebo-controlled clinical trial aiming to evaluate the efficacy and safety of a 200-mcg dose of a new drug (drug name masked for copyright protection) in patients with systemic lupus erythematosus (SLE). We will refer to this trial the “lupus trial” hereinafter. SLE is a long-term auto-immune disease in which the body mistakenly attacks healthy
organs, such as the skin and brain. There is no cure for the disease, and immuno-
therapies only reduce symptoms and must be applied frequently. Patients respond to
treatments quickly, typically within weeks, although the disease also relapses quickly.
A clinical response is based on the SLE responder index, which is defined by a combi-
nation of four different cognitive test scores. In the presence of control, each patient
is randomized and followed for 35 weeks, during which time multiple relapses and
responses might be observed. The trial objective is to compare the response rates of
the new treatment and the control. It is desirable to terminate the trial early since
the 35-week follow-up time period is long; for example, whenever there is substantial
evidence that the new treatment is no better than the control, the trial should be
stopped for both ethical and financial considerations. The total sample size is 200
patients, who are randomized 1:1 between the two treatments.

A second similar example is a placebo-controlled trial of oral laquinimod for mul-
tiple sclerosis [72]. The efficacy and safety of laquinimod in patients with relapsing-
remitting multiple sclerosis were evaluated in a randomized, double-blind, phase III
trial with 1,106 patients randomly assigned in a 1:1 ratio to receive placebo or laquin-
imod. The primary end point was whether or not the disease had relapsed during
the study period. An event was counted as a relapse if the patient’s symptoms were
accompanied by objective neurologic changes according to predefined criteria.

A common feature in both trials is that multiple responses and relapses are ex-
pected for each patient during the follow-up period, and therefore simple monotonic
parametric dose-response models such as logistic regression are no longer suitable.
The observed data for each patient will be a binary vector $\mathbf{e} = (e_1, e_2, \ldots, e_K)$ for $K$
time points. The indicators are temporally correlated.
5.2 Probability Model

5.2.1 Latent Gaussian Process

In the aforementioned lupus trial, patients are randomized between two arms, with arm 1 being the standard control and arm 2 the new treatment. Multiple interim analyses are proposed to monitor the progress of the trial and to compare the two treatments. At the time of interim analysis, assume that \( N_1 \) and \( N_2 \) patients have been assigned to arms 1 and 2, respectively (\( N_1 = N_2 \) if equal randomization). For each patient \( j \) in group \( i \), we assume that disease outcomes have been measured at \( K_{ij} \) time points, denoted as \( t_{ij} = (t_1, t_2, \ldots, t_{K_{ij}})' \). Here a time point refers to the duration of follow-up. Let \( e_{ij}(t_k) \) be the binary response at time point \( t_k \), simplified as \( e_{ijk} \). If the \( j \)th patient in group \( i \) responded at time \( t_k \), then \( e_{ijk} = 1 \); otherwise, \( e_{ijk} = 0 \). Therefore, the observed data are tuple \( e^K = \{e_{ijk}\}, i = 1, 2; j = 1, 2, \ldots, N_i; k = 1, 2, \ldots, K_{ij} \).

Define a stochastic process

\[
a_{ij}(t) = \mu_i(t) + \tau_{ij}(t) \tag{5.2.1}
\]

where \( \mu_i(t) \) is the mean process that describes the effects of drug \( i \) and \( \tau_{ij}(t) \) is a zero-mean GP that induces temporal correlation within the same patient.

Denoting the latent variable \( a_{ijk} \equiv a_{ij}(t_k) \), we assume \( e_{ijk} = I(a_{ijk} > a_h) \), i.e.,

\[
e_{ijk} = \begin{cases} 
1 & \text{if } a_{ijk} > a_h \\
0 & \text{otherwise}
\end{cases},
\]

where \( a_h \) is an arbitrary threshold and \( I(\cdot) \) is the indicator function.

In (5.2.1), the choice of \( \mu_i(t) \) depends on the underlying disease and drug mechanism. Based on prior knowledge, for example, if we believe that the efficacy of a
drug is increasing over time, we could simply use a linear function $\mu_i(t) = \beta_{i0} + \beta_{i1}t$
with a positive slope; if we believe the efficacy increases first then decreases due to
drug resistance, we could use a quadratic function. To accommodate different shapes
of response curves, we choose the polynomial regression model to describe the drug
mean effects over time, i.e.,

$$\mu_i(t) = \beta_{i0} + \beta_{i1}t + \cdots + \beta_{i,m_i}t^{m_i},$$

(5.2.2)

where $\mathbf{\beta}_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{i,m_i})'$ are regression coefficients of time for each of the two
treatment arms ($i = 1, 2$). In practice, the degree $m_i$ of the polynomial is unknown,
and we let $m_i = \{0, 1, \ldots, M\}$ be a random variable taking integer values between 0
and a large number $M$. It is important to allow $(m_i, \mathbf{\beta}_i)$ to vary for different arms
$i$ so that the drug effects can be easily compared by posterior inference using these
parameters.

Note that $\mathbf{\beta}_i$ describes the global response pattern of each arm $i$. For the depen-
dence across time within each patient $j$, we assume that $\tau_{ij}(t)$ follows a zero-mean
GP that induces the cyclic correlation of $a_{ij}(t)$. Let

$$\tau_{ij}(t) \mid \Theta \sim GP(\mathbf{0}_{K_{ij}}, \mathbf{C}(t_{ij}; \Theta)),$$

(5.2.3)

where $\mathbf{0}_{K_{ij}}$ is a $K_{ij}$-dimension 0-vector, and $\mathbf{C}(t_{ij}; \Theta)$ is a $K_{ij} \times K_{ij}$ covariance matrix
indexed by $\Theta$.

Based on (5.2.1), (5.2.2) and (5.2.3), the proposed latent Gaussian process func-
tional regression model is given by

$$P(\mathbf{a}^K \mid \mathbf{\beta}, \mathbf{m}, \Theta) \propto \prod_{i=1}^{2} \prod_{j=1}^{N_i} |\mathbf{C}(t_{ij})|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{a}_{ij}^K - \mathbf{X}_{ij} \mathbf{\beta}_i)' \mathbf{C}(t_{ij})^{-1} (\mathbf{a}_{ij}^K - \mathbf{X}_{ij} \mathbf{\beta}_i) \right\},$$

(5.2.4)

where $\mathbf{a}_{ij}^K = (a_{ij1}, \ldots, a_{ijK_{ij}})'$, $\mathbf{\beta} = \{\beta_i\}$, $\mathbf{m} = \{m_i\}$, $i = 1, 2$ and the design matrix
$\mathbf{X}_{ij}$ of the polynomial (5.2.2) is
The covariance matrix $C(t_{ij})$ is usually parameterized to induce within-patient dependence over time. The most popular parametrization [33] of the stationary covariance function $C(t_{ij}; \Theta)$ is a $K_{ij} \times K_{ij}$ matrix with the $uv$-th element $C_{uv}$ defined by

$$C_{uv}(\Theta) = \theta_1^2 \exp \left\{ -r^2(t_u - t_v)^2 \right\} + \delta_{uv}J^2, \quad u, v = 1, \ldots, K_{ij}. \quad (5.2.5)$$

In (5.2.5) $\Theta = (\theta_1, r, \delta_{uv})$, $C_{uv}(\Theta)$ is the covariance of the GP for time points $t_u$ and $t_v$, $\delta_{uv} = I(u = v)$, and $J$ is the variance on the diagonal reflecting the amount of jitter [77], which usually takes on a small value (e.g., $J = 0.1$). This construction of the covariance matrix yields a strong correlation for observations at time points close to each other and has been widely adopted [33].

However, if the system is recurrent, which is the case here for the lupus trial, formulation (5.2.5) is not suitable since it is aperiodic. Instead, we should consider other forms of covariance functions with periodicity. For example, a periodic model with known wavelength $\theta_2$ is given by

$$C_{uv}(\Theta) = \theta_1^2 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u - t_v)}{\theta_2} \right) \right\} + \delta_{uv}J^2, \quad (5.2.6)$$

and $\Theta = (\theta_1, \theta_2, r, \delta_{uv})$. This covariance function provides strong correlations not only for adjacent time points but also for those separated from each other by a distance $\theta_2$ or its multiplications. We will use (5.2.6) as the model for our trial applications. As an illustration, Figure 5.1 shows some realizations of $\tau_{ij}(t)$ with different covariance
functions $C$. The realizations display some periodic patterns although their shapes can be quite different. This implies that the class of GP models with covariance (5.2.6) is general and can describe curves with a variety of different shapes.

Define the column vector $a^K = \{a_{ijk}; i = 1, 2; j = 1, 2, \ldots, N_i; k = 1, 2, \ldots, K_{ij}\}'$. The likelihood function is given by

$$P(e^K | a^K) = \prod_{i=1}^{2} \prod_{j=1}^{N_i} \prod_{k=1}^{K_{ij}} P(e_{ijk} | a_{ijk})$$

$$= \prod_{i=1}^{2} \prod_{j=1}^{N_i} \prod_{k=1}^{K_{ij}} \{I(a_{ijk} > a_h)I(e_{ijk} = 1) + I(a_{ijk} \leq a_h)I(e_{ijk} = 0)\}.$$  

(5.2.7)

As a summary, the hierarchical model factors as

$$P(e^K, a^K, m, \beta, \Theta) \propto P(e^K | a^K) P(a^K | \beta, m, \Theta) P(\beta | m) P(m) P(\Theta).$$  

(5.2.8)

The first factor is the sampling model (5.2.7). The second factor is the latent Gaussian process functional regression model (5.2.4). The remaining factors are priors of $m, \beta$ and $(\theta_1, \theta_2, r)$. We assume a uniform prior on $P(m_i)$. For example, given a large integer $M$, $P(m_i) = 1/(M + 1)$ for $m_i = 0, 1, \ldots, M$. We assume

$$\beta_i \sim \text{Gaussian}(\mu_{m_i+1}, \sigma_0^2 I_{m_i+1})$$

$$\theta_1 \sim \text{Gaussian}(\mu_1, \sigma_1^2), \ r \sim \text{Gaussian}(\mu_2, \sigma_2^2), \ \theta_2 \sim \text{Gaussian}(\mu_3, \sigma_3^2),$$

where $\mu_{m_i+1}$ is $(m_i + 1)$-dimension mean vector, and $I_{m_i+1}$ is an $(m_i + 1) \times (m_i + 1)$ dimension identity matrix. We assume vague priors for $\beta_i, \theta_1, r, \theta_2$ by imposing large values of variances ($\sigma_0^2, \sigma_1^2, \sigma_2^2, \sigma_3^2$).

We will show that the LGP model in (5.2.8) possesses desirable theoretical and numerical properties, such as consistency, the ability to forecast for individual patients and to stop trials early.
Figure 5.1: Three random samples drawn from the Gaussian process $\tau_{ij}$ with four different covariance matrices. The corresponding covariance is given below each plot. The decrease in length scale from (a) to (b) produces more rapidly fluctuating functions. The periodic properties of the covariance function in (c) and (d) can clearly be seen.
5.2.2 MCMC Simulations

We carry out posterior inference based on MCMC simulation. The proposed Gibbs sampler proceeds by iterating over the following transition probabilities

\[
P(a^K | e^K, \beta, m, \Theta), P(m | a^K, \Theta), P(\beta | a^K, m, \Theta), P(\Theta | a^K, \beta, m).
\]

We start by generating \(a^K\) from their full conditional posterior distributions according to Geweke (1991) \[78\]. Through a cycle of Gibbs steps we also generate random draws of \(m, \beta\) and \(\Theta\). When updating \(m\), we sample from the full conditional posterior distributions, marginalized with respect to \(\beta\) (See Appendix B.1). Given \(m\), sampling of \(\beta\) is straightforward because of the fixed dimensionality and Gaussianity. This avoids the need for trans-dimensional MCMC. The challenging step is to sample \(\Theta\), the full conditional of which is analytically intractable. However, we can obtain the approximations using the most probable values such as Evidence maximization \[79\] or other non-Gibbs MCMC samplers. The most commonly used and straightforward method is the random-walk Metropolis sampling, but Bernardo et al. (1999) \[77\] argued that simple Metropolis algorithms were not efficient and therefore advocated for the use of hybrid Monte Carlo \[80\]. This idea was implemented by Barber and Williams (1997) \[81\] who analyzed classification problems and achieved better results using hybrid Monte Carlo compared to Metropolis algorithms.

With the same motivation, we propose a hybrid MCMC algorithm to sample the hyperparameters \(\Theta\) in the covariance function \(C(\Theta)\). The idea behind hybrid MCMC is to augment the parameter space and draw Metropolis proposals with improved samplers. Suppose we want to draw Monte Carlo samples from a proposed density \(P(\Theta) \propto \exp\{-E(\Theta)\}\), where \(\Theta = (\theta_1, \ldots, \theta_n)\). In physics terminology, \(\Theta\) can be regarded as a position vector and \(E(\Theta)\) the potential energy function. As a data
augmentation step, we then introduce another set of variables called “momentum variables”, \( \mathbf{w} = \{ w_1, w_2, \ldots, w_n \} \), one \( w_i \) for each \( \theta_i \), with a kinetic energy function, 
\[
P(\mathbf{w}) = \frac{1}{Z_K} \exp(-K(\mathbf{w})) = (2\pi)^{-n/2} \exp(-\frac{1}{2} \sum_i w_i^2).
\]
The momentum variables are introduced to make random walks continue in a consistent direction until a region of low probability is encountered. The total energy function, known as Hamiltonian, is 
\[
H(\Theta, \mathbf{w}) = E(\Theta) + K(\mathbf{w}) = E(\Theta) + \frac{1}{2} \sum_i w_i^2.
\]
The canonical distribution defined by this energy function is 
\[
P(\Theta, \mathbf{w}) = \frac{1}{Z_H} \exp\{ -H(\Theta, \mathbf{w}) \} = P(\Theta)P(\mathbf{w}).
\]
Random samples of \((\Theta, \mathbf{w})\) are jointly proposed by appropriate probabilities (see Appendix B.1 for details). Samples from the marginal distribution for \( \Theta \) can be obtained by ignoring the values \( \mathbf{w} \). Specifically, we use a discretized approximation called the 
leapfrog algorithm [82], details of which are described in Appendix B.1.

To summarize, the proposed MCMC algorithm includes a Gibbs sampler for \( \mathbf{a^K} \), \( \mathbf{m} \), \( \mathbf{b} \) and a hybrid Monte Carlo algorithm for sampling \( \Theta \). The algorithm converges well with an initial burn-in of 2,000 iterations and a total of 10,000 iterations with a lag of 10 in our simulation studies.

### 5.3 Posterior Inference

#### 5.3.1 Forecast

An important and useful feature of the LGP model is the ability to forecast, i.e., making predictions on \( e_{ij}(t_{K_{ij}+s}) \) for future time points \( t_{K_{ij}+s} \), \( s = 1, 2, \ldots, S_{ij} \). The forecast uses the posterior predictive distribution \( p(e_{ij}(t_{K_{ij}+s}) \mid e^K) \), where \( e^K \) are observed data at \( K_{ij} \) past time points. In particular, define
\[ q_{ij} \equiv Pr \{ e_{ij}(t_{K_{ij}+s}) = 1 \mid e^K \} = Pr \{ a_{ij}(t_{K_{ij}+s}) > a_h \mid e^K \} = \int I \{ a_{ij}(t_{K_{ij}+s}) > a_h \} \ p \{ a_{ij}(t_{K_{ij}+s}) \mid a^K_{ij}, m_i, \beta_i, \Theta \} \times p \{ a^K_{ij}, m_i, \beta_i, \Theta \mid e^K \} \ da_{ij}(t_{K_{ij}+s}) \ da^K_{ij} \ dm_i \ d\beta_i \ d\Theta. \quad (5.3.1) \]

The key component in \( q_{ij} \) is the conditional distribution

\[ p \{ a_{ij}(t_{K_{ij}+s}) \mid a^K_{ij}, m_i, \beta_i, \Theta \}, \] for all future points \( t_{K_{ij}+s}, s = 1, 2, \ldots, S_{ij} \). Therefore, we consider the vector

\[ a^{(S-K)}_{ij} = \left( a_{ij}(t_{K_{ij}+1}), a_{ij}(t_{K_{ij}+2}), \ldots, a_{ij}(t_{K_{ij}+S_{ij}}) \right)', \] the latent Gaussian variables for future time points. By definition of the GP model, the joint distribution of

\[ p \left\{ a^K_{ij}, a^{(S-K)}_{ij} \mid m_i, \beta_i, \Theta \right\} \] is a multivariate Gaussian, and so is the conditional

\[ p \{ a_{ij}(t_{K_{ij}+s}) \mid a^K_{ij}, m_i, \beta_i, \Theta \}. \] Denote its conditional CDF by \( \Phi_{S-K|K} (\cdot \mid a^K_{ij}, m_i, \beta_i, \Theta) \).

Then the Monte Carlo estimate of \( q_{ij} \) is given by

\[ \hat{q}_{ij} = \frac{1}{B} \sum_{b=1}^{B} \left( 1 - \Phi_{S-K|K} \left( a_h \mid a^{(b)}_{ij}, m_i^{(b)}, \beta_i^{(b)}, \Theta^{(b)} \right) \right), \quad (5.3.2) \]

where \( (b) \) denotes the MCMC sample of the \( b \)-th iteration and \( B \) is the total number of MCMC samples kept for posterior inference.

Note that the forecast is subject-\( j \)-specific. That is, based on the observed outcomes for patient \( j \), the forecast could be different due to the subject-specific \( \tau_{ij} \) in our original LGP model (5.2.1). This is useful in practice allowing individual patients and their physicians to prepare for future disease relapses.

### 5.3.2 Trial Monitoring Rules

In Section 5.2.1, we assume that the treatment mean effect for arm \( i \) is \( \mu_i(t) = \beta_{i0} + \beta_{i1} t + \cdots + \beta_{i,m_i} t^{m_i} \) with a random degree \( m_i \). For example, when \( m_i = 1 \) and
slope $\beta_{11}$ is positive, the mean effect $\mu_i(t)$ is a simple linear function increasing with time; if $m_i = 2$ with $\beta_{i2} < 0$, the mean effect $\mu_i(t)$ first increases and then decreases, potentially due to drug resistance. Estimating $\beta_i$ for two arms $i = 1, 2$ allows us to compare the overall drug effects and monitor the trial accordingly. For example, if the experimental arm is considered to perform no better than the control arm in terms of increasing disease remission time, the trial should be stopped early due to ethical and logistic reasons.

Denote the duration of the disease remission (DDR) as

$$T_i(\beta_i, m_i) = \{ t : \text{E} \{ a_{ij}(t) \} = \mu_i(t) > a_h \},$$

where $\{ \}$ is the Lebesgue measure on the real line. That is, $T_i$ is the length of time intervals in which $\mu_i(t)$ is above $a_h$ for arm $i$, i.e., patients stay in remission. The arm with longer DDR is more desirable.

The proposed monitoring criterion uses the posterior probability

$$\eta = \Pr(T_2 > T_1 + \delta \mid e^K)$$

$$= \int I \{ T_2(\beta_2, m_2) > T_1(\beta_1, m_1) + \delta \} \ p(\beta, m \mid e^K) \ d\beta \ dm, \quad (5.3.4)$$

where $\delta \geq 0$ is a threshold that defines the desirable increment of the efficacy outcome for the experimental arm over the control. The value of $\delta$ is fixed by investigators, and should reflect the minimum clinically meaningful improvement.

The Monte Carlo estimate of $\eta$ is given by

$$\hat{\eta} = \frac{1}{B} \sum_{b=1}^{B} I \left\{ T_2(\beta_{2b}, m_{2b}) > T_1(\beta_{1b}, m_{1b}) + \delta \right\}, \quad (5.3.5)$$

Let $\xi_U$ be an upper probability boundary above which the trial will be terminated early and the experimental treatment declared superior if $\Pr(T_2 > T_1 + \delta \mid e^K) \geq \xi_U$. 

Similarly, let $\xi_L$ be a lower boundary below which the trial will be terminated early due to futility if $\Pr(T_2 > T_1 + \delta \mid e^K) \leq \xi_L$.

Our proposed trial-monitoring rules are given as

\[
\text{Decision} = \begin{cases} 
\text{stop and declare arm 2 superior} & \text{if } \hat{\eta} \geq \xi_U, \\
\text{continue enrolling patients} & \text{if } \xi_L < \hat{\eta} < \xi_U, \\
\text{stop and declare arm 2 not superior} & \text{if } \hat{\eta} \leq \xi_L.
\end{cases}
\]

Cutoffs $\xi_U$ and $\xi_L$ can be calibrated based on simulations. The values $\xi_U = 0.95$ and $\xi_L = 0.05$ provided desirable performance in the simulation for the lupus trial.

### 5.4 Model Assessment With Sensitivity Analysis

When making forecasts on $e_{ij}(t_{K_{ij}+s})$ for future time points, three factors can affect the forecasting accuracy: 1) the choice of the mean functions $\mu_i(t)$ and covariance functions $\tau_{ij}(t)$, 2) the number of observed time points $K_{ij}$ at the time of forecast, and 3) the choice of the threshold $a_h$. To evaluate the impact of these model components, we perform a sensitivity analysis based on simulation. To simplify the simulation setup, we only consider data from one arm and focus on the accuracy of posterior estimation and forecast instead of trial characteristics. Therefore the index $i$ is dropped in the discussion within this section.

#### 5.4.1 Simulation Setup

We assumed that the sample size was 100 patients for a single arm. We first generated values $\mathbf{a}^K$ given $\mu(t)$ and $\tau_j(t)$ at a total of 35 time points $t_1, t_2, \ldots, t_{35}$, out of which the first $K_j \leq 32$ time points were used to predict the responses of patient $j$ at the last three time points. After generating $\mathbf{a}^K$, we generated $\mathbf{e}^K$ given the threshold $a_h$ of our choice. Without generality, we assumed that all the patients had the same
K_j = K for j = 1, \ldots, 100. We specified various choices of K = 20, 23, 26, 29, 32 to study their effects on forecasting in each scenario. To show the robustness to the threshold \( a_h \), we assumed \( a_h = 0 \) when we generated the simulation data and set \( a_h = 0 \) or \( a_h = 0.5 \) when we fitted models to the simulated data.

We considered five scenarios. In scenarios 1-3, we generated data by using different polynomial mean functions \( \mu(t) \) and the covariance function (5.2.6) with \( \theta_1 = 1, \theta_2 = 3.5 \) and \( r = 2 \). For ease of exposition, we display (5.2.6) again:

\[
C_{uv}(\Theta) = \theta_1^2 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u-t_v)}{\theta_2} \right) \right\} + \delta_{uv} J^2.
\]

In scenario 1, the true mean \( \mu(t) = \beta_0 \), where \( \beta_0 = -0.8 \). In scenario 2, the true \( \mu(t) = \beta_0 + \beta_1 t \), where \( \beta_0 = -0.8, \beta_1 = 0.4 \). In scenario 3, the true \( \mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2 \), where \( \beta_0 = -1, \beta_1 = 3.5, \beta_2 = -1 \).

In words, scenarios 1-3 represent constant mean, linear mean, and quadratic mean respectively. We fitted the LGP model (5.2.8) to the simulated data and assumed vague priors \( \beta \sim \text{Normal}(0, 10^2) \), \( \theta_1, \theta_2, r \sim \text{Normal}(0, 10^2) \).

In scenario 4, we generated data from the trigonometric mean function \( \mu(t) = \alpha + \sin(\beta_0 \pi t) \) with \( \alpha = -0.8, \beta_0 = 1.5 \), and the covariance function (5.2.5) with \( \theta_1 = 1 \) and \( r = 3 \). We then fitted the model using the same trigonometric mean function but with unknown parameters, and the covariance function (5.2.5) to the simulated data and estimated unknown parameters \( \alpha, \beta_0, \theta_1 \) and \( r \), for which we also assumed vague priors \( \alpha, \beta_0, \theta_1, r \sim \text{Normal}(0, 10^2) \).

In scenarios 1-4, we took periodicity into consideration by placing the trigonometric function in the fitted model. Scenarios 1-3 had a trigonometric function in the covariance function, while scenario 4 assumed a trigonometric mean function. To show the differences between these two configurations, in Figure 5.2 we plotted simulated \( \alpha \) for 20 randomly selected replications under scenarios 1-4. We can see that the overall mean in the top right panel increases with time and each patient curve
has a different periodic behavior in scenario 2. In contrast, in the bottom right panel for scenario 4, the overall mean has an obvious periodic pattern and each individual curve has its own variability around the mean.

As discussed above, putting a trigonometric function in the mean function $\mu(t)$ or covariance function $\tau_j(t)$ leads to different group and individual behaviors. Since both models are periodic, we designed scenario 5 to investigate whether one model can predict for the other. For this purpose, we generated data from the true model with mean $\mu(t) = \sin(\beta_0 \pi t)$, where $\beta_0 = 1$ and $\tau_j(t)$ from a GP with zero mean and covariance function $C$ in (5.2.5) with $\theta_1 = 1$ and $r = 3$, which does not have a trigonometric component. Then we fitted model (5.2.8) with a polynomial mean function and the covariance function having trigonometric components.

5.4.2 Simulation Results

For each of the five scenarios described above, we specified five different values for $K = 20, 23, 26, 29, 32$, and two threshold values $a_h = 0$ or 0.5. Therefore, we obtained a total of $5 \times 2 \times 5 = 50$ cases. For each case, we implemented the proposed LGP model with 10,000 MCMC iterations with a burn-in of 2,000 iterations. The convergence of the MCMC algorithm was diagnosed by standard methods in R package coda. All the chains converged quickly and mixed well.

Figure 5.3 shows the posterior estimates of the degree of polynomial $m$ in scenarios 1-3. Recall that $m$ follows a discrete uniform prior taking values in the set \{0, 1, 2, \ldots, M\}; here we set $M = 5$. We can see that the degree of the polynomial from the true model dominates the posterior estimates. In scenarios 4-5 there are no true $m$ values since the true $\mu(t)$ is not a polynomial function.

For each case, we computed the average posterior predictive probability of response
at a future time point $t_s$ over 100 patients, defined as

$$q_s = \frac{1}{100} \sum_{j=1}^{100} Pr\{e_j(t_s) = 1 \mid e^K\},$$

where $Pr\{e_j(t_s) = 1 \mid e^K\}$ was calculated according to (5.3.1) and (5.3.2). We used $t_s = 33, 34, \text{ and } 35$, since the lupus trial protocol plans to make forecasts at these three future time points. Also, we computed the posterior mean DDR

$$\hat{T} = \frac{1}{B} \sum_{b=1}^{B} l\{t : \mu(t; \beta^{(b)}, m^{(b)}) > a_h\}.$$ 

Table 5.1 summarizes the results.
Table 5.1: True probability, empirical probability and average posterior predictive probability of response at future time point $t_3$, $t_4$ and $t_5$. The posterior mean DDR $\hat{T}$ is also presented.

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Truth $q_{33}$</th>
<th>Truth $q_{34}$</th>
<th>Truth $q_{35}$</th>
<th>$\hat{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 32, a_h = 0$</td>
<td>Simulation 0.2212</td>
<td>0.2398</td>
<td>0.2491</td>
<td>0</td>
</tr>
<tr>
<td>$K = 32, a_h = 0.5$</td>
<td>Simulation 0.2205</td>
<td>0.2389</td>
<td>0.2488</td>
<td>0</td>
</tr>
<tr>
<td>$K = 29, a_h = 0$</td>
<td>Simulation 0.2683</td>
<td>0.2637</td>
<td>0.2613</td>
<td>0</td>
</tr>
<tr>
<td>$K = 29, a_h = 0.5$</td>
<td>Simulation 0.2712</td>
<td>0.2667</td>
<td>0.2640</td>
<td>0</td>
</tr>
<tr>
<td>$K = 26, a_h = 0$</td>
<td>Simulation 0.2734</td>
<td>0.2691</td>
<td>0.2663</td>
<td>0</td>
</tr>
<tr>
<td>$K = 26, a_h = 0.5$</td>
<td>Simulation 0.2726</td>
<td>0.2687</td>
<td>0.2665</td>
<td>0</td>
</tr>
<tr>
<td>$K = 23, a_h = 0$</td>
<td>Simulation 0.2445</td>
<td>0.2443</td>
<td>0.2442</td>
<td>0</td>
</tr>
<tr>
<td>$K = 23, a_h = 0.5$</td>
<td>Simulation 0.2450</td>
<td>0.2447</td>
<td>0.2446</td>
<td>0</td>
</tr>
<tr>
<td>$K = 20, a_h = 0$</td>
<td>Simulation 0.2377</td>
<td>0.2376</td>
<td>0.2376</td>
<td>0</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>Simulation 0.2399</td>
<td>0.2399</td>
<td>0.2399</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>Truth $q_{33}$</th>
<th>Truth $q_{34}$</th>
<th>Truth $q_{35}$</th>
<th>$\hat{T}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 32, a_h = 0$</td>
<td>Simulation 0.7459</td>
<td>0.7568</td>
<td>0.7717</td>
<td>18.07</td>
</tr>
<tr>
<td>$K = 32, a_h = 0.5$</td>
<td>Simulation 0.7482</td>
<td>0.7604</td>
<td>0.7755</td>
<td>18.70</td>
</tr>
<tr>
<td>$K = 29, a_h = 0$</td>
<td>Simulation 0.7778</td>
<td>0.7920</td>
<td>0.8053</td>
<td>19.00</td>
</tr>
<tr>
<td>$K = 29, a_h = 0.5$</td>
<td>Simulation 0.7800</td>
<td>0.7944</td>
<td>0.8075</td>
<td>19.14</td>
</tr>
<tr>
<td>$K = 26, a_h = 0$</td>
<td>Simulation 0.7684</td>
<td>0.7826</td>
<td>0.7960</td>
<td>18.76</td>
</tr>
<tr>
<td>$K = 26, a_h = 0.5$</td>
<td>Simulation 0.7747</td>
<td>0.7878</td>
<td>0.8010</td>
<td>18.89</td>
</tr>
<tr>
<td>$K = 23, a_h = 0$</td>
<td>Simulation 0.7906</td>
<td>0.8034</td>
<td>0.8156</td>
<td>18.92</td>
</tr>
<tr>
<td>$K = 23, a_h = 0.5$</td>
<td>Simulation 0.7979</td>
<td>0.8106</td>
<td>0.8227</td>
<td>19.35</td>
</tr>
<tr>
<td>$K = 20, a_h = 0$</td>
<td>Simulation 0.7732</td>
<td>0.7852</td>
<td>0.7967</td>
<td>17.98</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>Simulation 0.7718</td>
<td>0.7840</td>
<td>0.7957</td>
<td>18.27</td>
</tr>
<tr>
<td>Scenario</td>
<td>Truth Q_{33}</td>
<td>Truth Q_{34}</td>
<td>Truth Q_{35}</td>
<td>Truth $\hat{T}$</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>0.3676</td>
<td>0.2557</td>
<td>0.1599</td>
<td>28.723</td>
</tr>
<tr>
<td>$K = 32, a_h = 0$</td>
<td>0.42</td>
<td>0.30</td>
<td>0.19</td>
<td>27.55</td>
</tr>
<tr>
<td>$K = 32, a_h = 0.5$</td>
<td>0.4212</td>
<td>0.3201</td>
<td>0.2317</td>
<td>29.506</td>
</tr>
<tr>
<td>$K = 29, a_h = 0$</td>
<td>0.4763</td>
<td>0.3723</td>
<td>0.2731</td>
<td>29.893</td>
</tr>
<tr>
<td>$K = 29, a_h = 0.5$</td>
<td>0.4594</td>
<td>0.3538</td>
<td>0.2546</td>
<td>29.728</td>
</tr>
<tr>
<td>$K = 26, a_h = 0$</td>
<td>0.6648</td>
<td>0.5795</td>
<td>0.4873</td>
<td>31.261</td>
</tr>
<tr>
<td>$K = 26, a_h = 0.5$</td>
<td>0.6465</td>
<td>0.5625</td>
<td>0.4742</td>
<td>30.962</td>
</tr>
<tr>
<td>$K = 23, a_h = 0$</td>
<td>0.4690</td>
<td>0.3814</td>
<td>0.2999</td>
<td>29.547</td>
</tr>
<tr>
<td>$K = 23, a_h = 0.5$</td>
<td>0.3744</td>
<td>0.2866</td>
<td>0.2105</td>
<td>28.762</td>
</tr>
<tr>
<td>$K = 20, a_h = 0$</td>
<td>0.7575</td>
<td>0.7160</td>
<td>0.6731</td>
<td>30.931</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>0.7483</td>
<td>0.7068</td>
<td>0.6646</td>
<td>30.869</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>Truth Q_{33}</td>
<td>Truth Q_{34}</td>
<td>Truth Q_{35}</td>
<td>Truth $\hat{T}$</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>--------------</td>
<td>--------------</td>
<td>---------------</td>
</tr>
<tr>
<td>$K = 32, a_h = 0$</td>
<td>0.3118</td>
<td>0.1660</td>
<td>0.0823</td>
<td>-</td>
</tr>
<tr>
<td>$K = 32, a_h = 0.5$</td>
<td>0.3123</td>
<td>0.1664</td>
<td>0.0825</td>
<td>-</td>
</tr>
<tr>
<td>$K = 29, a_h = 0$</td>
<td>0.2458</td>
<td>0.1228</td>
<td>0.0614</td>
<td>-</td>
</tr>
<tr>
<td>$K = 29, a_h = 0.5$</td>
<td>0.2458</td>
<td>0.1228</td>
<td>0.0614</td>
<td>-</td>
</tr>
<tr>
<td>$K = 26, a_h = 0$</td>
<td>0.2640</td>
<td>0.1366</td>
<td>0.0683</td>
<td>-</td>
</tr>
<tr>
<td>$K = 26, a_h = 0.5$</td>
<td>0.2687</td>
<td>0.1400</td>
<td>0.0699</td>
<td>-</td>
</tr>
<tr>
<td>$K = 23, a_h = 0$</td>
<td>0.2528</td>
<td>0.1314</td>
<td>0.0677</td>
<td>-</td>
</tr>
<tr>
<td>$K = 23, a_h = 0.5$</td>
<td>0.2522</td>
<td>0.1310</td>
<td>0.0675</td>
<td>-</td>
</tr>
<tr>
<td>$K = 20, a_h = 0$</td>
<td>0.2641</td>
<td>0.1387</td>
<td>0.0710</td>
<td>-</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>0.2632</td>
<td>0.1378</td>
<td>0.0705</td>
<td>-</td>
</tr>
</tbody>
</table>
After obtaining the posterior imputation of the latent variables $\mathbf{a}^K$ and posterior predictive values of $\mathbf{a}^{(35-K)}$ for the new time points, we plotted 20 realizations of the latent variables for one randomly selected patient in case ($K = 32, a_h = 0$) for scenario 2 and case ($K = 29, a_h = 0$) for scenario 4 as shown in Figures 5.4 and 5.5, respectively. The left panel presents the latent variable values given the first 32 observed time points and the right panel presents those with the first 29 time points observed. It can be seen that our model fits the simulated data well in both panels.

In scenario 5, we generated simulated data using the model with a trigonometric mean function and covariance function (5.2.5) without trigonometricity, but fitted the model with the polynomial mean function and the trigonometric covariance function (5.2.6) to the simulated data. As seen in Table 5.1, the predictions are poor. This implies that the model choice is important in terms of the placement of the

<table>
<thead>
<tr>
<th>Scenario 5</th>
<th>Truth</th>
<th>Empirical</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 32, a_h = 0$</td>
<td>0.2104</td>
<td>0.23</td>
<td>0.1544</td>
</tr>
<tr>
<td>$K = 32, a_h = 0.5$</td>
<td>0.1720</td>
<td>0.22</td>
<td>0.0679</td>
</tr>
<tr>
<td>$K = 29, a_h = 0$</td>
<td>0.1599</td>
<td>0.19</td>
<td>0.0181</td>
</tr>
<tr>
<td>$K = 29, a_h = 0.5$</td>
<td>0.1447</td>
<td>0.0287</td>
<td>0.0711</td>
</tr>
<tr>
<td>$K = 26, a_h = 0$</td>
<td>0.0068</td>
<td>0.0000</td>
<td>0.0002</td>
</tr>
<tr>
<td>$K = 26, a_h = 0.5$</td>
<td>0.0103</td>
<td>0.0000</td>
<td>0.0003</td>
</tr>
<tr>
<td>$K = 23, a_h = 0$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$K = 23, a_h = 0.5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$K = 20, a_h = 0$</td>
<td>0.0017</td>
<td>0.0012</td>
<td>0.0028</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>0.0011</td>
<td>0.0024</td>
<td>0.0035</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$K = 20, a_h = 0.5$</td>
<td>0.9960</td>
<td>0.9950</td>
<td>0.9950</td>
</tr>
</tbody>
</table>
trigonometric functions, in the mean or covariance of the GP.

Next we checked the robustness of the choice of $a_h$. As shown in Table 5.1, the estimations remained consistently well when we varied the values of $a_h$ ($a_h = 0$ and $a_h = 0.5$). Thus, the model is robust to choice of $a_h$.

Finally we examined the impact of different $K$'s on the forecast. From Table 5.1 and Figure 5.3, we can see that the fewer the observed time points, i.e. the smaller the value of $K$, the worse the prediction. For example, in scenario 3, when $K = 32$ and $a_h = 0$, the degree of polynomial of the true model is successfully recovered 99% of times; but when $K = 20$ and $a_h = 0$, this percentage drops to 85%. Also, the true $q_{33} = 0.3676$, while the posterior estimate of $q_{33}$ is 0.4179 when ($K = 32, a_h = 0$) and 0.7575 when ($K = 20, a_h = 0$). This also echoes the consistency results in Section 5.6 to be shown next; that is, the larger the number of time points observed, the better the estimation.

Next in the lupus trial example, one concern is how to choose the proper time point to start the forecast, which also depends on the drug effects. Based on previous clinical experiments and prior knowledge, we can make assumptions about the mean of the the drug effects and the start time point can be determined by simulation and calibration. Examining the simulation results observed so far, we decided to use $K = 23$ as the starting time for forecast.

5.5 Trial Example

Using the lupus trial as an example, we simulated a large number of clinical trials on computer and applied the proposed monitoring rules to examine the operating characteristics of our method. According to the original trial setup, we set the duration of the simulated trials at 35 weeks and the maximum number of patients to be 200,
Figure 5.2: Sample curves of latent variable values for 20 randomly selected patients in scenarios 1-4. The red curve in each plot represents the mean function $\mu(t)$ of the GP. The horizontal blue line represents the threshold $a_h$. 
Figure 5.3: The posterior estimates of $m$ in all cases for scenarios 1-3 in the simulation.
Figure 5.4: A set of 20 realizations of latent variables for two cases in scenario 2. The left panel presents the latent variable values for one patient with the first 32 time points observed. The right panel presents one patient with the first 29 time points observed. The red solid line represents the truth. Black solid lines represent 20 realizations at observed time points, and blue dashed lines represent predicted values at future time points.

Figure 5.5: A set of 20 realizations of latent variables for two cases in scenario 4. The left panel presents the latent variable values for one patient with the first 32 time points observed. The right panel presents one patient with the first 29 time points observed. The red solid line represents the truth. Black solid lines represent 20 realizations at observed time points, and blue dashed lines represent predicted values at future time points.
equally randomized between the standard arm and experimental arm. We assumed that patients were recruited over time and the number of patients enrolled weekly for each arm was a random number following a discrete uniform in \( \{2, 3, 4\} \). Patients were observed and followed each week, when their responses \( e \)'s were recorded. The experimental treatment (arm 2) is considered more effective than the control (arm 1) if the DDR in (5.3.3) for the treatment arm \( T_2 \) is at least 2 weeks longer than that of the control arm \( T_1 \), i.e., the minimum difference between the two arms must be \( \delta = 2 \).

We considered six scenarios and simulated 100 trials for each scenario. The binary response outcomes were generated over time using the following scheme. We first specified true mean functions (see Table 5.2 for different \( \mu_i(t) \)) and the true covariance function (5.2.6) with \( \theta_1 = 1, \theta_2 = 3.5 \) and \( r = 2 \). We then generated the LGP \( a(t_k) \) and \( e(t_k) \) at 35 weeks of follow up, \( k = 1, 2, \ldots, 35 \).

For all six scenarios, the true DDR \( T_i \) according to the true \( \mu_i(t) \) were calculated, and the configurations including \( m_i, \beta_i \) and \( T_i, i = 1, 2 \) are displayed in Table 5.2. Figure 5.6 shows the true mean functions of the standard and experimental arms in all six scenarios, in which we also mark the true \( T_i \) values. The arm with a larger \( T_i \) value is more effective since it leads to a longer DDR. For example, in scenario 1, \( T_1 = 20.616 \) and \( T_2 = 27.616 \), so arm 2, the experimental treatment, is better.

We fitted the proposed model (5.2.8) to the simulated data and assumed vague priors \( \beta \sim \text{Normal}(\mathbf{0}, 10^2 \mathbf{I}), \theta_1, \theta_2, r \sim \text{Normal}(0, 10^2) \). The trial monitoring started at week \( K = 23 \) and was based on the proposed monitoring rules with \( \delta = 2 \). Table 5.3 summarizes the operating characteristics of the LGP method for all six scenarios, including the average duration of trials, the maximum duration of trials, the average number of patients studied and stopping probabilities.
Table 5.2: True Values of $m_i$, $\beta_i$ and $T_i$, $i = 1, 2$ for all six scenarios in the simulation.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Group</th>
<th>$m_i$</th>
<th>$\beta_i$</th>
<th>$T_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>Standard</td>
<td>2</td>
<td>(-2, 3.5, -1)</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>3</td>
<td>(-1.4, 7.5, -5.3, 1)</td>
<td>27.6</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>Standard</td>
<td>3</td>
<td>(-1.5, 7.5, -5.3, 1)</td>
<td>25.9</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>2</td>
<td>(-1, 3.5, -1)</td>
<td>28.7</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>Standard</td>
<td>3</td>
<td>(-2.4, 7.5, -5.3, 1)</td>
<td>15.4</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>2</td>
<td>(-2.4, 3.5, -1)</td>
<td>16.3</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>Standard</td>
<td>3</td>
<td>(-2, 7.5, -5.3, 1)</td>
<td>19.7</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>2</td>
<td>(-1, 3.5, -1)</td>
<td>28.7</td>
</tr>
<tr>
<td>Scenario 5</td>
<td>Standard</td>
<td>3</td>
<td>(-1.28, 3.5, -1)</td>
<td>26.7</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>3</td>
<td>(-1.2, 3.6, -1)</td>
<td>28.6</td>
</tr>
<tr>
<td>Scenario 6</td>
<td>Standard</td>
<td>2</td>
<td>(-0.39, 0.3)</td>
<td>22.0</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>2</td>
<td>(-1.1, 1)</td>
<td>24.0</td>
</tr>
</tbody>
</table>
Figure 5.6: The true DDR $T_i$ and the true mean functions $\mu_i(t)$ in the control and experimental arms in all six scenarios for the lupus trial simulation. The red line represents the true mean in the control arm, while the blue line represents the true mean in the experimental arm. The black horizontal line represents the threshold $a_h = 0$. 
Table 5.3: Results for the simulations based on the lupus trial. AD: average trial duration (weeks); MD: maximum duration (weeks); AP: average number of patients in each group.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Group</th>
<th>$T_i$</th>
<th>AD</th>
<th>MD</th>
<th>AP</th>
<th>Stopping probabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Standard</td>
<td>20.6</td>
<td>24.9</td>
<td>29</td>
<td>76.1</td>
<td>97% (superiority)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>27.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Standard</td>
<td>25.9</td>
<td>29.72</td>
<td>35</td>
<td>87.43</td>
<td>52% (superiority)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>28.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Standard</td>
<td>15.4</td>
<td>29.06</td>
<td>35</td>
<td>87.68</td>
<td>60% (futility)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>16.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Standard</td>
<td>19.7</td>
<td>23.31</td>
<td>26</td>
<td>70.65</td>
<td>100% (superiority)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>28.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Standard</td>
<td>26.7</td>
<td>29.11</td>
<td>35</td>
<td>86.91</td>
<td>25% (superiority) and 30% (futility)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>28.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Standard</td>
<td>22.0</td>
<td>30.90</td>
<td>35</td>
<td>91.59</td>
<td>16% (superiority) and 19% (futility)</td>
</tr>
<tr>
<td></td>
<td>Experimental</td>
<td>24.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In scenario 1, $m_1 = 2$ and $m_2 = 3$, and from Figure 5.6, patients in the experimental arm relapse more often than patients in the control arm. However, the experimental arm is still preferred since it has longer DDR as $T_2 = 27.616 > (20.616 + 2) = (T_1 + \delta)$. Thus, scenario 1 explored the ability of the proposed LGP method to stop if the experimental arm has a longer DDR but more frequent relapses. The average trial duration was 24.9 weeks and the maximum trial duration was 29 weeks, which indicated that the trial stopped quickly after it was monitored at week 23. The average number of patients was 76.1 per group and the maximum was 100 patients.
Among 100 simulated trials, 97 were stopped early due to superiority.

In scenario 2, $m_1 = 3, m_2 = 2$ and $T_1 = 25.939, T_2 = 28.723$. Here $T_2 - T_1 \approx 2.8$ which is close to $\delta = 2$. Among the 100 trials simulated, 52 stopped early due to superiority. In this scenario, the average trial duration was 29.72 weeks and the maximum trial duration was 35 weeks. The average number of patients was 87.43 per group.

In scenario 3, $m_1 = 3, m_2 = 2$ and $T_1 = 15.414, T_2 = 16.279$. Among the 100 simulated trials, 60 stopped early due to futility as expected. In this scenario, the average trial duration was 29.06 weeks. The average number of patients was 87.68 per group. In scenario 4, $m_1 = 3, m_2 = 2$ and $T_1 = 19.736, T_2 = 28.723$, implying that arm 2 was much better than arm 1. All 100 trials stopped early due to superiority of arm 2. The average trial duration was 23.31 weeks and the maximum trial duration was 26 weeks. The average number of patients was 70.65 per group.

In the last two scenarios, different arms had the same $m_i$ value. In scenario 5, $m_1 = m_2 = 2$. The average trial duration was 29.11 weeks, and the average number of patients was 86.91 per group. Among the 100 trials simulated, 25 stopped due to superiority, 30 stopped due to futility, and 45 did not stop early. In scenario 6, $m_1 = m_2 = 1$ and $T_1 = 22, T_2 = 24$. The DDR in arm 2 was exactly 2 weeks longer than in arm 1, making it difficult to stop early. Among 100 simulated trials, 65 didn’t stop early, 16 stopped early due to superiority, and 19 stopped early due to futility. The average trial duration was 30.9 weeks and the maximum trial duration was 35 weeks. The average number of patients was 91.59 per group.

In summary, the proposed LGP model and trial monitoring rules exhibited desirable operating characteristics in all six scenarios.
5.6 Posterior Consistency

Finally, we discuss theoretical properties of the LGP model and show that it is consistent as the number of time points in each observation goes to infinity. For simplicity in the following discussion, we use $n$ to denote the number of time points in an observation. The proposed LGP model can be summarized as follows:

$$
e_{t_k} \mid p(t_k) \sim \text{Bernoulli}(p(t_k)), \quad k = 1, \ldots, n,$$

$$p(t) = H(a(t)) \equiv Pr(a(t) > a_h), t \in [0, T_E]$$

$$a(\cdot) \sim GP(\mu(\cdot), C(\cdot, \cdot)), \quad (5.6.1)$$

where $t_k$'s are fixed in advance on the compact set $T = [0, T_E] \subset \mathcal{R}$ in which $T_E$ is the length of the trial, and $H$ is a known, strictly increasing, Lipschitz continuous cumulative function. Note that $a(\cdot)$ is a Gaussian process parameterized by its mean function $\mu: T \to \mathcal{R}$ and its covariance function $C: T^2 \to \mathcal{R}$, denoted by $GP(\mu, C)$. Additionally, we assume that the underlying true probability function is $p_0(t) = H(a_0(t))$, and $a_0(t)$, the true GP, has a continuously differentiable sample path on $T$. Without loss of generality, we assume a linear mean function $\mu(t) = \beta_0 + \beta_1 t$. The proof for the case of $m$-polynomial is easily extended. In addition, we assume that the true covariance function has the form $C_0(t_u, t_v; \Theta) = C_0(|t_u - t_v|)$, where $C_0$ is a positive multiple of a nowhere-zero density function of one real variable. In our model, we consider a simple form of covariance function (5.2.6) without the jitter part. Then $C_0(t) = \theta_1^2 \exp \left\{-r^2 \sin^2 \left(\frac{\pi}{\theta_2} t\right)\right\}$. We rewrite $C_0(t) \equiv C_0(t; \theta_1, r, \theta_2)$ to prepare for the coming discussion.

Posterior consistency using GP priors has been extensively studied, see [83, 84, 85, 86, 87], among others. However, little is known about the type of LGP models proposed in this paper. We first present posterior consistency results based on a
general modeling framework under a wide class of GP priors $a(\cdot)$ and link functions $H$, which can be any strictly increasing, Lipschitz continuous function. Our proposed LGP model in (5.6.1) is simply a special case of the general results. The proofs of all lemmas and theorems stated in this section are given in Appendix B.2.

To start, let $\mathcal{F}$ be the set of Borel measurable functions defined on $\mathbb{R}$, the real line. For now, assume that we have a topology on $\mathcal{F}$ that measures the distance between any two functions in $\mathcal{F}$. We also follow Ghosal et al. (1999) [88], to define a probability measure on the functional space. That is, we operate under the probability space $(\mathcal{F}, \sigma_\mathcal{F}, P_\mathcal{F})$, where $\sigma_\mathcal{F}$ is the $\sigma$-field of $\mathcal{F}$ and $P_\mathcal{F}$ is the probability measure for functions. Suppose $p(t) \in \mathcal{F}$, and let $P_p$ stand for the probability measure corresponding to $p(t)$ on the probability space for the real line, $(\mathbb{R}, \mathcal{B}, P)$, where $\mathcal{B}$ and $P$ are the Borel set and the probability measure on the real line, respectively. To clarify, we use $(\mathcal{F}, \sigma_\mathcal{F}, P_\mathcal{F})$ and $(\mathbb{R}, \mathcal{B}, P)$ to denote the probability spaces for random functions and random variables, respectively. Let $(e_{t_1}, \ldots, e_{t_n})$ be $n$ random discrete variables, each of which follows $e_{t_k} \mid p(t_k) \sim \text{Bern}(p(t_k))$. We assign prior distribution $\Pi$ for $p(t)$ and the posterior distribution given $e_{t_1}, \ldots, e_{t_n}$, denoted by $\Pi(\cdot \mid e_{t_1}, \ldots, e_{t_n})$. We will show that if $e_{t_k}$’s are generated conditional on the true function $p_0(t)$, i.e., $e_{t_k} \mid p_0(t_k) \sim \text{Bern}(p_0(t_k))$, then under the hierarchical model $e(\cdot) \mid p(\cdot) \sim \text{Bern}(p(\cdot))$; $p(\cdot) \sim \Pi$ and regulatory conditions, the posterior of $p(\cdot)$ (on the functional probability space) converges to $p_0$.

Below we introduce the metric on $\mathcal{F}$, three lemmas, and the main theorem.

**Definition 1** Let $\mathcal{F}$ be the set of Borel measurable functions defined on $\mathbb{R}$. For any $f_0 \in \mathcal{F}$, denote the Kullback-Leibler neighborhood $K_\epsilon(f_0) = \{f : \int f_0 \log(f_0/f) < \epsilon\}$. Let $\Pi$ be a prior on $\mathcal{F}$, we say $f_0$ is in the K-L support of $\Pi$ if $\Pi(K_\epsilon(f_0)) > 0$ for all...
Lemma 1  Let \(([0, T_E], \mathcal{B}, P)\) be a probability space on the real line, and let \(\mathcal{F}\) be the set of all real-valued Borel measurable functions \(f: [0, T_E] \to [0, 1]\). Let \(P(t) = \frac{l(t)}{T_E}\) be a probability measure, where \(l(t)\) is the Lebesgue measure on \([0, T_E]\). Define
\[
d(f, g) = \inf\{\varepsilon : P(\{t : |f(t) - g(t)| > \varepsilon\}) < \varepsilon\}.
\]
Then \(d(\cdot, \cdot)\) induces a metric, and \(f_n\) converges to \(f\) in probability if and only if \(\lim_{n \to \infty} d(f_n, f) = 0\).

Lemma 2  Let \(0 < a, b < 1\). Then
\[
 a \log \frac{a}{b} + (1 - a) \log \frac{1 - a}{1 - b} \leq (a - b)^2.
\]

Lemma 3  Assume the compact supports of \(\theta_1, r, \theta_2\) are \(B_1, B_2\) and \(B_3\) respectively. Let \(B = B_1 \times B_2 \times B_3\), a compact support of \((\theta_1, r, \theta_2)\). Define two functions \(\rho_1(\theta_1, r, \theta_2) = C_0(0; \theta_1, r, \theta_2)\) and \(\rho_2(\theta_1, r, \theta_2) = -C_0''(0; \theta_1, r, \theta_2)\), where \(C_0''\) is the second derivative of \(C_0\). Let \(a(\cdot)\) be the GP on \(\mathcal{T}\) in (5.6.1), where \(\mathcal{T}\) is a bounded subset of \(\mathcal{R}\). Then \(a(\cdot)\) has differentiable sample paths and the derivative process \(a'(\cdot)\) is also a GP. Further, there exist constants \(A\) and \(d_1, d_2\) such that
\[
\sup_{(\theta_1, r, \theta_2) \in B} \Pr\left\{\sup_{t \in [0, T_E]} |a(t)| > M_n \mid \theta_1, r, \theta_2\right\} \leq Ae^{-d_1 n},
\]
\[
\sup_{(\theta_1, r, \theta_2) \in B} \Pr\left\{\sup_{t \in [0, T_E]} |a'(t)| > M_n \mid \theta_1, r, \theta_2\right\} \leq Ae^{-d_2 n},
\]
where \(M_n = O(n^{1/2})\).
Next we present the main results in Theorem 1. Choi and Schervish (2007) [85] proposed a Consistency Theorem (see Appendix B.3) as an extension of Schwartz’s theorem [89] to independent but non-identically distributed cases. We make use of such an extension to achieve the posterior consistency for our proposed LGP by verifying the two conditions in Choi and Schervish (2007) [85].

Theorem 1 Suppose that the values of the covariate $t_k$’s are in $[0, T_E]$. Assume that the mean function $\mu(\cdot)$ and $C(\cdot, \cdot)$ in the Gaussian process prior (5.6.1) satisfy the following conditions:

1. $C(\cdot, \cdot)$ is nonsingular,

2. $C(t_u, t_v)$ has the form $C_0(|t_u - t_v|)$, where $C_0(t)$ is a positive multiple of a nowhere zero density function on $\mathcal{R}$ and four times continuously differentiable on $\mathcal{R},$

3. the mean function $\mu(t)$ is continuously differentiable in $[0, T_E]$, and

4. there exist $0 < \delta < \frac{1}{2}$ and $b_1, b_2 > 0$ such that

$$Pr(\theta_1^2 > n^\delta) < b_1 \exp(-b_2 n), Pr(r^2 > n^\delta) < b_1 \exp(-b_2 n),$$

$$Pr(\theta_2^2 < n^{-\delta}) < b_1 \exp(-b_2 n).$$

Let $P_0^n$ denote the joint distribution of $\{e_{t_k}\}_{k=1}^n$ assuming that $p_0(t) = H(a_0(t))$ is the true probability function, where $a_0(t)$ is the true GP that has a continuously differentiable sample path on $\mathcal{T}$. Then for every $\epsilon > 0$, letting $S_{\epsilon}^1(p_0) = \{p \in \mathcal{F} : d(p, p_0) > \epsilon\}$, as $n \to \infty$

$$\Pi\{S_{\epsilon}^1(p_0) \mid e_{t_1}, \ldots, e_{t_n}\} \to 0 \quad [P_0^\infty]$$

(5.6.2)
In addition, for every $\epsilon > 0$, letting $S_\epsilon^2(p_0) = \{ p \in \mathcal{F} : \int |p(t) - p_0(t)| dt > \epsilon \}$, as $n \to \infty$

\[ \Pi \{ S_\epsilon^2 \mid e_{t_1}, \ldots, e_{t_n} \} \to 0 \quad [P_0^\infty] \quad (5.6.3) \]

In general, if more hyperparameters are introduced in the covariance function, we can verify posterior consistency by assuming compact supports of hyperparameters and proper continuity conditions as in Theorem 1.
Chapter 6

Conclusions and Future Work

This dissertation has developed several Bayesian models for a variety of application problems in high-throughput genomic data and clinical trial designs. These application problems include mapping multireads to reference genome, integrating multiple platforms for TCGA data, clustering count data for ChIP-Seq data and proposing a monitoring rule for recurrent longitudinal clinical data.

Chapter 2 proposes a read mapping method that utilizes the full information contained in NGS data. Specifically, the proposed BM-Map method maps the multireads by taking into account the sequencing error profiles and the information related to the mapping of unique reads. The Bayesian paradigm works very well and yields desirable results in our simulation studies and the analysis of yeast whole genome. Computation is a challenging for analyzing NGS data with millions of short reads. The C++ source code we developed is available from


We have released a software that implements the BM-Map method and we will continue to develop it. Although not many genes in yeast were found to have different RPKMs between the BM-Map and proportion methods through our RNA-Seq analysis, we show in Table 2.1 that the ones that did exhibit differences were important to the viability of the yeast. In addition, through careful examination of our analysis results we found that our method shows significant improvements when hidden nucleotide variations are present in the competing mapping loci. Therefore,
the proposed methodology is expected to have a larger effect in the species with high polymorphism frequencies or in cross-reference situations where RNA-Seq reads from one species without available genome sequence are mapped to the genome of a closely related species as a surrogate reference. As another possible future direction, although our study is primarily based on RNA-Seq data, the proposed Bayesian framework can be easily extend to other NGS applications such as DNA-resequencing and Chip-Seq.

In Chapter 3, we propose a Bayesian graphical model to describe the dependence structure of three genetic phenomena, CNVs, DNA methylation, and mRNA expression [90]. The inferred graph gives a clear representation of the regulatory relationships involving the three genetic features. For example, the mRNA expression of gene ERLIN2 is sensitive to copy number changes but robust to DNA methylation, while the mRNA expression of gene PIR is sensitive to both copy number changes and DNA methylation. We are in the process of making a comprehensive list of these relationships using the entire TCGA data, expanding the effort to include more cancer types and more features such as microRNA and protein expression. And we are in the process of developing a software that implements our network model.

Chapter 4 proposes a nonparametric Bayesian local clustering Poisson model for a count data matrix. The NoB-LCP model detects local clustering patterns by performing simultaneous clustering on columns and rows of a data matrix. Idle local clusters are introduced to better separate noisy HMs and location from the actual signals in the genomics data. Through simulation studies and the analysis of ChIP-Seq data we demonstrate the effectiveness of our model in grouping regulatory elements with similar functionality based on HMs patterns. We used zero-enriched Pólya urn priors to model random partitions of HMs and genomic locations. Although partitions do not allow overlap between the partitioning subsets in one imputation of the parame-
ters, posterior inference could still report positive (marginal) posterior probability for membership in multiple clusters for the same HM (reporting such probabilities also requires a resolution of the label switching problem).

In Chapter 4, we assume that all bi-clusters identified do not overlap, that means one HM can only function in one regulatory process [91]. As a potential future work direction, it is possible that one HM could participate in multiple regulation processes. Alternatively, we could use feature allocation models, such as the Indian buffet processes [92] as priors for a random allocation of HMs to subsets, including membership in multiple subsets.

Chapter 5 proposes a Bayesian LGP model for modeling time-dependent discrete data from clinical trials. Through the posterior estimates, we can predict the probability of efficacy response for each patient at future time points. The proposed trial monitoring rules allow for early termination of a trial if one arm is considered more effective than the other. The Bayesian paradigm works very well and yields desirable results in our simulation studies.

Although our current models are set up for binary outcomes, they can be easily extended to other applications with ordinal or categorical outcomes. Furthermore, the LGP provides a framework for the inclusion of patient- or group-specific covariates such as patients’ weights, ages, etc., which can be easily implemented by expanding the columns of the design matrix $X_{ij}$ in (5.2.4). This will be a future direction of this research.
Appendix A: MCMC Details of NoB-LCP Model

Joint pdf

\[
p(Y, c, r, \theta, \pi_0, \pi_1) = p(\pi_0)p(\pi_1)p(c)p(r \mid c)p(\theta \mid c, r, k, \lambda)p(Y \mid \theta)
\]

\[
= p(\pi_0)p(\pi_1)p(c) \prod_{q=1}^{Q} \left[ p(r_q) \prod_{d=1}^{D_q} \left\{ \prod_{g \in V_q} (p(\hat{\theta}_{dq} \mid k_{0g}, \lambda_{0g}) \prod_{i \in R_{qd}} p(y_{ig} \mid \hat{\theta}_{dq})) \right\} \right] \\
\times \prod_{g \in V_q} \prod_{i \in R_{1g}} (p(\theta_{1g} \mid k_{1g}, \lambda_{1g})p(y_{ig} \mid \theta_{1g})) \\
\times \prod_{g \in V_0} \prod_{i=1}^{N} (p(\theta_{ig} \mid k_{2g}, \lambda_{2g})p(y_{ig} \mid \theta_{ig}))
\]

where \( V_q = \{ g \mid c_g = q, q = 1, \ldots, G \} \) is the set of HMs in a HM set \( q, q = 0, \ldots, Q \), and \( R_{qd} = \{ i \mid r_{qi} = d, i = 1, \ldots, N \} \) is the set of genomic locations in genomic location cluster \( d \) corresponding to HM set \( q \) for \( q = 1, \ldots, Q \) and \( d = 1, \ldots, D_q \). We include \( \lambda \) and \( k \) in the conditioning sets to indicate the relevant (fixed) hyperparameters.

The prior probability distribution of \( c \) and \( r_q \) are a zero-enriched Pólya urnscheme given in Equations (4.3.1) and (4.3.2) of the main paper.

Full conditional

1. Update \( \theta \)

(a) For active HMs (\( c_g > 0 \)) and active genomic locations (\( r_{qi} > 0 \), \( q =\)
1, \ldots, Q \text{ and } d = 1, \ldots, D_q.

\[ \tilde{\theta}_{dg} \mid k_{0g}, \lambda_{0g}, c, r_q, y_g \sim \text{Gamma}(k_{0g} + \sum_{i \in R_{qd}} y_{ig}, \lambda_{0g} + n_{qd}) \]

(b) For active HMs ($c_g > 0$) and idle genomic locations ($r_{qi} = 0$), $i = 1, \ldots, N$,

\[ \theta_{ig} \mid k_{1g}, \lambda_{1g}, c, r_q, y_g \sim \text{Gamma}(k_{1g} + y_{ig}, \lambda_{1g} + 1) \]

(c) For idle HM set ($c_g = 0$), $i = 1, \ldots, N$,

\[ \theta_{ig} \mid k_{2g}, \lambda_{2g}, c, y_g \sim \text{Gamma}(k_{2g} + y_{ig}, \lambda_{2g} + 1) \]

2. Update $\pi_0$

\[ \pi_0 \mid c \sim \text{Beta}(a_0 + G', b_0 + G - G') \]

3. Update $\pi_1$

\[ \pi_1 \mid r \sim \text{Beta}(a_1 + \sum_q \sum_i I(r_{qi} > 0), b_1 + NQ - \sum_q \sum_i I(r_{qi} > 0)) \]

where $I$ is an indicator function: $I(r_{qi} > 0) = 1$ if $r_{qi} > 0$; $I(r_{qi} > 0) = 0$ if $r_{qi} = 0$.

4. Update $r_q$

Update $r_q$ for active HM sets, $q = 1, \ldots, Q$ and $i = 1, \ldots, N$.

Remove $\theta_i^q$, define $m_q^-, \theta_i^{q-}, D_q^-, r_q^-, n_q^-$ and $R_{qd}^-$, and integrate with respect to
\[ p(r_{qi} = d \mid c, r_q^-, k_0, \lambda_0, y) \propto \]
\[
\begin{cases}
\pi_1 \left[ \frac{\beta + m_q}{\beta + n_g} \right] \prod_{g \in V_q} \frac{1}{y_g} \prod_{g \in V_q} \frac{\Gamma(k_{0y} + \sum_{l \in R_{gd}} y_{lg})}{\Gamma(k_{0y} + \sum_{l \in R_{yd}} y_{lg})} 	imes \\
\times \frac{\lambda_{0y} + n_g}{\lambda_{0y} + n_{gd}} \frac{\Gamma(k_{0y} + \sum_{l \in R_{yd}} y_{lg})}{\Gamma(k_{0y} + \sum_{l \in R_{yd}} y_{lg})} & d = 1, \ldots, D_q^-,
\end{cases}
\]
\[
\begin{cases}
\pi_1 \left[ \frac{\beta + m_q}{\beta + n_g} \right] \prod_{g \in V_q} \frac{\Gamma(k_{1y} + \sum_{l \in R_{yd}} y_{lg})}{\Gamma(k_{1y})} \frac{\lambda_{1y}}{\lambda_{1y} + 1} \frac{\Gamma(k_{1y} + \sum_{l \in R_{yd}} y_{lg})}{\Gamma(k_{1y})} & d = D_q + 1
\end{cases}
\]
\[
(1 - \pi_1) \prod_{g \in V_q} \frac{\Gamma(k_{1y} + \sum_{l \in R_{yd}} y_{lg})}{\Gamma(k_{1y})} \frac{\lambda_{1y}}{\lambda_{1y} + 1} \frac{\Gamma(k_{1y} + \sum_{l \in R_{yd}} y_{lg})}{\Gamma(k_{1y})} & d = 0
\]

5. Update \( c \)

Remove \( c_g \), and define \( G^-, Q^-, p^-, c^- \) and \( n^- \). Sample \( c_g \) as follows: \( c_g \in \{0, 1, \ldots, Q^-, (Q^- + 1)\} \). Note that \( c_g = 0 \) implies becoming idle, \( 1 \leq c_g \leq Q^- \) joining one of the existing HM sets, and \( c_g = Q^- + 1 \) starting a new singleton HM set.

\[
p(c_g = q \mid y_g) \propto \begin{cases}
p(c_g = 0)p(y_g \mid c_g = 0) & \text{for } q = 0 \\
p(c_g = q)p(y_g \mid c_g) & q = 1, \ldots, Q^- \\
p(c_g = Q^- + 1) \sum_r p(r \mid c_g = Q^- + 1)p(y_g \mid r) & q = Q^- + 1
\end{cases}
\]

The marginalization of \( r \) is difficult and computationally intensive. To avoid this problem, we consider a pseudo prior \( p(\tilde{r}_g \mid y_g) \) and let \( \tilde{r}_{Q^-+1} = \tilde{r}_g \). Finally, after canceling \( \prod_{g' = 1}^G p(\tilde{r}_g \mid y_g) \), we have the following:

\[
p(c_g = q \mid y_g) \propto \begin{cases}
p(c_g = 0)p(y_g \mid c_g = 0) & \text{for } q = 0 \\
p(c_g = q)p(y_g \mid c_g) & q = 1, \ldots, Q^- \\
p(c_g = Q^- + 1)p(r_{Q^-+1})p(y_g \mid r_{Q^-+1}) & q = Q^- + 1
\end{cases}
\]
For joining an existing cluster, \( q = 1, \ldots, Q \),

\[
p(c_g = q \mid c^-, \text{rest}) \propto \pi_0 \left( \frac{p_q}{\pi_0 + G'} \right) \prod_{d=1}^{D_q} \left[ \frac{\lambda_{0g}^{k_0g} \Gamma(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}{\Gamma(k_{0g}) (n_{qd} + \lambda_{0g})^{(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}} \prod_{i \in R_{qd}} \frac{1}{y_{ig}!} \right] \times \\
\times \prod_{i \in R_{q0}} \frac{\Gamma(k_{1g} + y_{ig})}{\Gamma(k_{1g})} \frac{(\lambda_{1g})^{k_{1g}}}{(\lambda_{1g} + 1)^{k_{1g} + y_{ig} y_{ig}}}.
\]

for starting a new (singleton) cluster

\[
p(c_g = Q^- + 1 \mid c^-, \text{rest}) \propto \pi_0 \left( \frac{\alpha}{\alpha + G'} - \pi_1 \right)^{m_g} (1 - \pi_1)^{(N - m_g)} \beta^{D_q} \prod_{d=1}^{D_q} \Gamma(n_{qd}) \prod_{i=1}^{m_g} \frac{1}{\beta + i - 1} \times \\
\times \prod_{d=1}^{D_q} \left[ \frac{\lambda_{0g}^{k_0g} \Gamma(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}{\Gamma(k_{0g}) (n_{qd} + \lambda_{0g})^{(k_{0g} + \sum_{i \in R_{qd}} y_{ig})}} \prod_{i \in R_{qd}} \frac{1}{y_{ig}!} \right] \times \\
\times \prod_{i \in R_{q0}} \frac{\Gamma(k_{1g} + y_{ig})}{\Gamma(k_{1g})} \frac{(\lambda_{1g})^{k_{1g}}}{(\lambda_{1g} + 1)^{k_{1g} + y_{ig} y_{ig}}}.
\]

and for joining the inactive cluster

\[
p(c_g = 0 \mid c^-, \text{rest}) \propto (1 - \pi_0) \prod_{i=1}^{N} \frac{\Gamma(k_{2g} + y_{ig})}{\Gamma(k_{2g})} \frac{(\lambda_{2g})^{k_{2g}}}{(\lambda_{2g} + 1)^{k_{2g} + y_{ig} y_{ig}}}.
\]
Appendix B: Details of MCMC Sampling and Proofs in LGP Model

Appendix B.1: MCMC Details

Joint Distributions

\[
P(e^K, a^K, m, \beta, \Theta) = \prod_{i=1}^{N_i} \prod_{j=1}^{K_{ij}} \left\{ P(e_{ijk} \mid a_{ijk}) = I(a_{ijk} > a_h)I(e_{ijk} = 1) + I(a_{ijk} \leq a_h)I(e_{ijk} = 0) \right\} \}
\times (2\pi)^{-K_{ij}/2} |C(t_{ij})|^{-1/2}
\times \exp \left\{ -\frac{1}{2} (a^K_{ij} - X_{ij}\beta_i)C(t_{ij})^{-1}(a^K_{ij} - X_{ij}\beta_i) \right\}
\times (2\pi)^{-m_{i+1}/2} |\sigma_0^2 I_{m_i+1}|^{-1/2}
\times \exp \left\{ -\frac{1}{2} (\beta_i - \mu_{0i}^{m+1})'(\sigma_0^2 I_{m_i+1})^{-1}(\beta_i - \mu_{0i}^{m+1}) \right\}
\times P(m) P(\Theta)
\]

where \( a^K_{ij} = (a_{ij1}, \ldots, a_{ijK_{ij}})' \), \( \beta_i = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{i,m_i})' \), \( I_{m_i+1} \) is an \((m_i + 1) \times (m_i + 1)\) dimension identity matrix, and
\[ \mathbf{X}_{ij} = \begin{pmatrix} 1 & t_1 & \cdots & t_{m_1}^i \\ 1 & t_2 & \cdots & t_{m_2}^i \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{K_{ij}} & \cdots & t_{K_{ij}}^i \end{pmatrix}. \]

Full Conditional

Algorithm for simulating \( a_{ij}^K \)

\[
P(a_{ij}^K \mid \mathbf{e}^K, \beta, m, \Theta) \propto \left\{ \prod_{k=1}^{K_{ij}} P(e_{ijk} \mid a_{ijk}) = I(a_{ijk} > a_h)I(e_{ijk} = 1) \\
+ I(a_{ijk} \leq a_h)I(e_{ijk} = 0) \right\} \\
\times (2\pi)^{-\frac{K_{ij}}{2}}|C(t_{ij})|^{-\frac{1}{2}} \\
\times \exp \left\{ -\frac{1}{2}(a_{ij}^K - \mathbf{X}_{ij}\beta_i)'C(t_{ij})^{-1}(a_{ij}^K - \mathbf{X}_{ij}\beta_i) \right\}
\]

To sample this truncated multivariate normal distribution, we use the method of [78] to compose a cycle of Gibbs steps through the components of \( a_{ij} \).

To draw the samples from multivariate normal distribution subject to linear inequality restrictions:

\[ x \sim N(\mu, \Sigma), \quad a < Dx < b. \]

This is equivalent to:

\[ z \sim N(0, T), \quad \alpha < z < \beta. \]

where

\[ T = DD', \quad \alpha = a - D\mu, \quad \beta = b - D\mu, \]
and we then take \( x = \mu + D^{-1} z \).

Suppose that in the non-truncated distribution \( N(0, T) \),

\[
E[z_i \mid z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n] = \sum_{j \neq i} c_{ij} z_j.
\]

Then in the truncated normal distribution of (6), the distribution of \( z_i \) conditional on \( \{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n\} \) has the construction

\[
z_i = \sum_{j \neq i} c_{ij} z_j + h_i \epsilon_i, \quad \epsilon_i \sim T N[\alpha_i - \sum_{j \neq i} c_{ij} z_j]/h_i, (\beta_i - \sum_{j \neq i} c_{ij} z_j)/h_i].
\]

Denote the vectors of coefficients \( c^i = (c_{i1}, \ldots, c_{i,i-1}, c_{i,i+1}, \ldots, c_{in})', i = 1, \ldots, n \).

From the conventional theory of the conditional multivariate normal distribution, \( c^i = -(T^{ii})^{-1} T^{i,<i} \) and \( h_i^2 = (T^{ii})^{-1} \), where \( T^{ii} \) is the element in row \( i \) and column \( i \) of \( T^{-1} \), and \( T^{i,<i} \) is row \( i \) of \( T^{-1} \) with \( T^{ii} \) deleted.

Therefore, to sample \( a_{ij} \) from the posterior conditional probabilities, we compose a cycle of \( K_{ij} \) Gibbs steps through the components of \( a_{ij} \). In the \( k \)th step of this cycle, \( a_{ijk} \) is simulated from \( a_{ijk} \mid e^K, a_{ijq}(q \neq k), \beta_i, m_i, \Theta \), which is a univariate normal distribution truncated to one region.

**Sample the degree of polynomial \( m_i \)**

\[
P(m_i \mid a^K, \Theta) \propto P(a^K \mid m, \Theta)P(m_i) = P(m_i) \int P(a^K \mid \beta, m, \Theta)P(\beta \mid m) \, d\beta
\]

\[
\propto P(m_i) \frac{|A_{m_i}|^{\frac{1}{2}}}{|\sigma_0^2 I_{m_i+1}|^{\frac{1}{2}}}
\times \exp \left\{ \frac{1}{2} b_{m_i}' A_{m_i} b_{m_i} - \frac{1}{2} \mu_{m_i+1}' (\sigma_0^2 I_{m_i+1})^{-1} \mu_{m_i+1} \right\},
\]
where
\[
A_{m_i} = \left\{ \sum_{j=1}^{N_j} X'_{ij} C(t_{ij})^{-1} X_{ij} + (\sigma_i^2 I_{m_i+1})^{-1} \right\}^{-1},
\]
\[
b_{m_i} = \left\{ \sum_{j=1}^{N_j} X'_{ij} C(t_{ij})^{-1} a^K_{ij} + (\sigma_0^2 I_{m_i+1})^{-1} \mu_{m_i+1} \right\}.
\]

Since we only consider \((M + 1)\) possible models, we can compute \(r_{m_i} = P(m_i | a^K, \Theta)\), \(m_i = 0, 1, \ldots, M\). Let \(w_{m_i} = r_{m_i} / \sum_{m=0}^{M} r_{m_i}\), then we draw a random number \(u\) from \(\text{Uniform}(0, 1)\). If \(u < w_0, m_i = 0\); if \(u < w_0 + w_1, m_i = 1\); if \(u < w_0 + w_1 + w_2, m_i = 2\); if \(u < w_0 + w_1 + w_2 + w_3, m_i = 3\); etc.

**Posterior conditional distribution of \(\beta_i\)**

The posterior conditional distribution of \(\beta_i\) is a multivariate normal distribution.

\[
P(\beta_i | a^K, \Theta) \sim \text{MVN}(\mu^\beta_i, \left\{ \sum_{j=1}^{N_j} X'_{ij} C(t_{ij})^{-1} X_{ij} + (\sigma_0^2 I_{m_i+1})^{-1} \right\}^{-1}),
\]

where
\[
\mu^\beta_i = \left\{ \sum_{j=1}^{N_j} X'_{ij} C(t_{ij})^{-1} X_{ij} + (\sigma_i^2 I_{m_i+1})^{-1} \right\}^{-1}
\]
\[
\times \left\{ \sum_{j=1}^{N_j} X'_{ij} C(t_{ij})^{-1} a^K_{ij} + (\sigma_0^2 I_{m_i+1})^{-1} \mu_{m_i+1} \right\}.
\]

**Hybrid Monte Carlo algorithm to sample \(\Theta\)**

We apply Hybrid Monte Carlo method to sample \(\Theta\). Suppose we want to sample from the canonical distribution for a set of variables, \(\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}\), with respect to
the potential energy function $E(\Theta)$. Assuming that $E(\Theta)$ is differentiable with respect to $\theta_i$, this canonical distribution is

$$P(\Theta) = \frac{1}{Z_E} \exp\{-E(\Theta)\}.$$

We then introduce another set of variables, $w = \{w_1, w_2, \ldots, w_n\}$, one $w_i$ for each $\theta_i$, with a kinetic energy function, $P(w) = \frac{1}{Z_K} \exp\{-K(w)\} = (2\pi)^{-n/2} \exp(-\frac{1}{2} \sum_i w_i^2)$. The total energy function, known as Hamiltonian, is $H(\Theta, w) = E(\Theta) + K(w) = E(\Theta) + \frac{1}{2} \sum_i w_i^2$. The canonical distribution defined by this energy function is $P(\Theta, w) = \frac{1}{Z_H} \exp\{-H(\Theta, w)\} = P(\Theta)P(w)$.

The unknown parameters ($\theta_1, \theta_2, \ldots, \theta_n$) can be sampled from the marginal distribution for $\Theta$ by ignoring the values we obtained for $w$. In practice, the differentiation equations that describe the Hamiltonian equilibrium through time need to be discretized. In our case, to obtain the posterior samples of our unknown parameters, we use leapfrog discretization. In order to perform a leapfrog discretization, the derivative of the log of the posterior probability with respect to hyper-parameters is needed.

A single leapfrog iteration calculates approximations $\hat{\theta}$ and $\hat{w}$ at time $\tau + \epsilon$ from $\hat{\theta}$ and $\hat{w}$ at time $\tau$ as follows:

$$\hat{w}_i(\tau + \frac{\epsilon}{2}) = \hat{w}_i(\tau) - \frac{\epsilon}{2} \frac{\partial E}{\partial \hat{\theta}_i} \{\hat{\theta}_i(\tau)\}$$

$$\hat{\theta}_i(\tau + \epsilon) = \hat{\theta}_i(\tau) + \epsilon \hat{w}_i(\tau + \frac{\epsilon}{2})$$

$$\hat{w}_i(\tau + \epsilon) = \hat{w}_i(\tau + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \frac{\partial E}{\partial \hat{\theta}_i} \{\hat{\theta}_i(\tau + \epsilon)\}.$$
The hybrid Monte Carlo algorithm

Given values for the magnitude of the leapfrog stepsize, \( \epsilon_0 \), and the number of leapfrog steps, \( L \), the dynamical transitions of the hybrid Monte Carlo algorithm operate as follows:

- Randomly choose a direction, \( \lambda \), for the trajectory with the two values \( \lambda = +1 \) representing a forward trajectory, and \( \lambda = -1 \) representing a backward trajectory, both being equally likely.

- Starting from the current state, \( (\theta, w) = (\hat{\theta}(0), \hat{w}(0)) \), perform \( L \) leapfrog steps with a stepsize of \( \epsilon = \lambda \epsilon_0 \), resulting in the state \( (\hat{\theta}(\epsilon L), \hat{w}(\epsilon L)) = (\theta^*, w^*) \).

- Regard \( (\theta^*, w^*) \) as a candidate for the next state, as in the Metropolis algorithm: accepting it with probability

\[
A\{((\theta, w), (\theta^*, w^*))\} = \min\{1, \exp\{-(H(\theta^*, w^*) - H(\theta, w))\}\},
\]

otherwise letting the new state be the same as the old one.

For our problem,

\[
E(\Theta) = \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{N_i} \left\{ (a_{ij}^K - X_{ij}\beta_i)'C(t_{ij})^{-1}(a_{ij}^K - X_{ij}\beta_i) + \log(|C(t_{ij})|) \right\} - \log P(\Theta),
\]

\[
\frac{\partial E(\Theta)}{\partial \Theta} = -\frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{N_i} (a_{ij}^K - X_{ij}\beta_i)'C(t_{ij})^{-1} \frac{\partial C(t_{ij})}{\partial \Theta} C(t_{ij})^{-1}(a_{ij}^K - X_{ij}\beta_i)
\]

\[
+ \frac{1}{2} \sum_{i=1}^{2} \sum_{j=1}^{N_i} tr \left\{ C(t_{ij})^{-1} \frac{\partial C(t_{ij})}{\partial \Theta} \right\} - \frac{\partial \log P(\Theta)}{\partial \Theta}.
\]
Since $C_{uv} = C(t_u, t_v) = \theta_1^2 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u - t_v)}{\theta_2} \right) \right\} + \delta_{uv} J^2$, and we have the formula $\frac{\partial C^{-1}}{\partial u} = -C^{-1} \frac{\partial C}{\partial u} C^{-1}$, therefore,

$$C'_{\theta_1}(t_i, t_j) = \frac{\partial C}{\partial \theta_1} = 2\theta_1 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u - t_v)}{\theta_2} \right) \right\},$$

$$C'_{\theta_2}(t_i, t_j) = \frac{\partial C}{\partial \theta_2} = \theta_2^2 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u - t_v)}{\theta_2} \right) \right\} \left( -2r \sin \left\{ \frac{\pi(t_u - t_v)}{\theta_2} \right\} \right),$$

$$C'_{r}(t_i, t_j) = \frac{\partial C}{\partial r} = 2r \theta_1 \exp \left\{ -r^2 \sin^2 \left( \frac{\pi(t_u - t_v)}{\theta_2} \right) \right\} \sin \left\{ \frac{\pi(t_u - t_v)}{\theta_2} \right\} \cos \left\{ \frac{\pi(t_u - t_v)}{\theta_2} \right\} \left\{ \frac{\pi(t_u - t_v)}{\theta_2^2} \right\}.$$

**Appendix B.2: Proofs of Lemmas and Main Theorem**

**Proof of Lemma 1**

Clearly $d(f, g) \geq 0$ and $d(f, g) = d(g, f)$. Also, we can easily show $d(f, g) = 0$ if and only if $f = g$ a.s.. To prove $d(\cdot, \cdot)$ induces a metric, the last condition we need to verify is triangle inequality $d(f, h) \leq d(f, g) + d(g, h)$ for each $f, g, h \in \mathcal{F}$.

Define $Q_{fg} = \{ \epsilon : P(\{ t : |f(t) - g(t)| > \epsilon \}) < \epsilon \}$, then $d(f, g) = \inf Q_{fg}$. So we need to show

$$\inf Q_{fh} \leq \inf Q_{fg} + \inf Q_{gh}.$$

Assuming $\epsilon_1 \in Q_{fg}$ and $\epsilon_2 \in Q_{gh}$, then

$$P(\{ t : |f(t) - h(t)| > \epsilon_1 + \epsilon_2 \}) \leq P(\{ t : |f(t) - g(t)| > \epsilon_1 \}) + P(\{ t : |g(t) - h(t)| > \epsilon_2 \}) \leq \epsilon_1 + \epsilon_2.$$
So if $\varepsilon_1 \in Q_{fg}$ and $\varepsilon_2 \in Q_{gh}$, then $\varepsilon_1 + \varepsilon_2 \in Q_{fh}$, which implies $\inf Q_{fh} \leq \inf Q_{fg} + \inf Q_{gh}$.

Lastly we show $f_n$ converges to $f$ in probability if and only if $\lim_{n \to \infty} d(f_n, f) = 0$. First, assume that $\lim_{n \to \infty} d(f_n, f) = 0$. Then for every $\varepsilon > 0$ there exists $N$ such that for all $n \geq N$, $d(f_n, f) \leq \varepsilon$, which is equivalent to $P(\{t : |f_n - f| > \varepsilon\}) < \varepsilon$. Hence, $f_n$ converges to $f$ in probability. Finally, assume that $f_n$ converges to $f$ in probability. Then for every $\varepsilon > 0$, $\lim_{n \to \infty} P(\{t : |f_n - f| > \varepsilon\}) = 0$. So, for every $\varepsilon > 0$, there exists $N$ such that for all $n \geq N$, $d(f_n, f) \leq \varepsilon$, which completes the proof.

**Proof of Lemma 2**

It follows easily from applying Taylor’s expansion to $\log(a), \log(b), \log(1 - a)$ and $\log(1 - b)$.

**Proof of Lemma 3**

Our mean function $\mu(t) = \beta_0 + \beta_1 t$ is continuously differentiable in $[0, T_E]$. For sufficiently large $M$,

\[
Pr(\sup_t |a(t)| > M) \leq Pr(\sup_t |a(t) - \mu(t)| > M - \sup_t |\mu(t)|) \\
\leq Pr(\sup_t |a(t) - \mu(t)| > M/2)
\]

Thus, without generality, we assume that the mean function is identically zero.

From the result of Theorem 5 in Ghosal and Roy (2006) [86], there exist constants $A_w, c_w, d^w(\theta_1, r, \theta_2)$ such that

\[
Pr \left\{ \sup_{t \in [0, T_E]} |D^w a(t)| > M_n \mid \theta_1, r, \theta_2 \right\} \leq A_w e^{-c_w d^w n},
\]

where $w = 0, 1$ and $c_w > 0$. Through calculations (details not shown), $d^0 = 1/C_0(0; \theta_1, \theta_2) = \frac{1}{\theta_1^2}$, $d^1 = -1/C_0'(0; \theta_1, \theta_2) = -\frac{1}{2\pi^2 \theta_1^3 r^2 / \theta_2^2}$. Furthermore, since
we assume \( \rho_1(\theta_1, r, \theta_2) \) and \( \rho_2(\theta_1, r, \theta_2) \) are continuous on the compact set \( B \), they are uniformly bounded. Thus, there exist universal constant \( S_1, S_2, S_3 \) and \( S_4 \) such that

\[
0 < S_1 \leq \sup_{(\theta_1, r, \theta_2) \in B} |\rho_1| \leq S_2 \\
0 < S_3 \leq \sup_{(\theta_1, r, \theta_2) \in B} |\rho_2| \leq S_4
\]

Consequently,

\[
\begin{align*}
\sup_{(\theta_1, r, \theta_2) \in B} &Pr\left\{ \sup_{t \in [0, T_E]} |a(t)| > M_n | \theta_1, r, \theta_2 \right\} \leq Ae^{-d_1n} \\
\sup_{(\theta_1, r, \theta_2) \in B} &Pr\left\{ \sup_{t \in [0, T_E]} |a'(t)| > M_n | \theta_1, r, \theta_2 \right\} \leq Ae^{-d_2n},
\end{align*}
\]

where \( d_1 = c_0/S_2 \), \( d_1 = c_1/S_4 \) and \( A = \max(A_0, A_1) \).

**Proof of Theorem 1**

The first condition of the Consistency Theorem in Choi and Schervish (2007) [85] we need to verify is prior positivity of neighborhoods. If the prior satisfies this condition, the probability of every Kullback-Leibler (KL) neighborhood of true density function is positive. The density of \( e_{tk} \) with respect to the counting measure on \( \{0, 1\} \), say \( Q \), is given by \( f(e_{tk}) = p(t_k)^{e_{tk}} (1 - p(t_k))^{1-e_{tk}} \). For simplicity, the index \( t_k \) is dropped here. So we have \( f(e) = p^e (1-p)^{1-e} \). The corresponding true density function is \( f_0(e) = p_0^e (1-p_0)^{1-e} \) and \( \int f_0 \log(f_0/f) \, dQ = p_0 \log(p_0/p) + (1-p_0) \log((1-p_0)/(1-p)) \).

To show that the probability of every KL neighborhood of true density function is positive, we first show that \( \Pi\{ f : \int f_0 \log(f_0/f) \, dQ < \epsilon \} > 0 \) for all \( \epsilon > 0 \), where \( \Pi \) is the prior for \( f \), or equivalently, \( \Pi\{ p : p_0 \log(p_0/p) + (1-p_0) \log((1-p_0)/(1-p)) < \epsilon \} > 0 \) for all \( \epsilon > 0 \). Following Lemma 2, \( p_0 \log(p_0/p) + (1-p_0) \log((1-p_0)/(1-p)) \leq ||p-p_0||_\infty^2 \), where \( ||p-p_0||_\infty^2 = \sup_{t \in T_E} |p(t) - p_0(t)|^2 \). Therefore, it suffices to show that \( \Pi(p : ||p-p_0||_\infty < \epsilon) > 0 \) for every \( \epsilon > 0 \). Since our link function \( H \) is assumed to
be bounded and Lipschitz continuous, it suffices to show that $\Pi(a : ||a-a_0||_\infty < \epsilon) > 0$ for every $\epsilon > 0$.

The prior distribution of $a$ is $a \sim GP(\mu, C)$, where $||\mu||_\infty < A_2$ and $||\mu'||_\infty < A_3$ for some constants $A_2$ and $A_3$ under our assumptions. Without loss of generality we assume $\mu \equiv 0$. Choi and Schervish (2004) [84] examined the general result on the uniform support for a Gaussian process prior with zero mean. It is easily extended to support $[0, T_E]$ and our mean assumption. Therefore the positivity of neighborhoods holds.

The second condition is the existence of tests. We construct a similar sieve as in Choi and Schervish (2004) [84] and then construct a test for each element of the sieve.

$$SI_n = \{ p(\cdot) : p(t) = H(a(t)), ||D^w a||_\infty < M_n \}.$$ 

Let $p_1$ be a continuous function on $[0, T_E]$. Let $h_k = 1$ if $p_1(t_k) \geq p_0(t_k)$ and $-1$ otherwise. Let $r > 0$, $m_n = n^{1/2}$ and $I_n$ be the indicators of set $U_1 = \{ \sum_{k=1}^n h_k(e_0(t_k) - p_0(t_k)) > 2m_n \sqrt{n} \}$, where $e_0(t_k) \sim \text{Bernoulli}(p_0(t_k))$. For all $p_1$ that satisfy

$$\sum_{k=1}^n |p_1(t_k) - p_0(t_k)| > rn, \quad (B.2.1)$$

by Bernstein’s inequality, we have

$$E_{P_0}(I_n) = P_0 \left\{ \sum_{k=1}^n h_k(e_0(t_k) - p_0(t_k)) > 2m_n \sqrt{n} \right\}$$

$$= 2 \exp\left(-\frac{1}{2 n + 2m_n \sqrt{n}}\right) \leq 2 \exp(-2m_n^2). \quad (B.2.2)$$

Also, let us assume $e(t_k) \sim \text{Bernoulli}(p(t_k))$. For all sufficiently large $n$ such that
\[ m_n / \sqrt{n} < 4/r \] and all \( p \) satisfying \( \|p - p_1\|_\infty < r/4 \), we have

\[ E_P (1 - I_n) = P \left\{ \sum_{k=1}^{n} h_k(e(t_k) - p_0(t_k) \leq 2m_n \sqrt{n} \right\} 
\]

\[ = P \left\{ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h_k(e(t_k) - p(t_k)) + \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h_k(p(t_k) - p_1(t_k)) 
+ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h_k(p_1(t_k) - p_0(t_k)) \leq 2m_n \right\} 
\]

\[ \leq P \left\{ \frac{1}{\sqrt{n}} \sum_{k=1}^{n} h_k(e(t_k) - p(t_k)) \leq r \sqrt{n}^2 - r \sqrt{n} + 2m_n \right\} 
\]

\[ \leq 2 \exp(-\frac{1}{2} \frac{r^2}{1 + r} n), \quad \text{(B.2.3)} \]

where the second to last inequality is by Bernstein’s inequality.

We showed how to construct consistent test functions when the inequality (B.2.1) holds. Choi (2007) [93] examined inequality (6.0.1) held with metric \( d \) introduced in Lemma 1. By (B.2.2), (B.2.3) and Lemma 3, the second condition of Consistency Theorem in Choi (2007) [93] is verified.

So far the two conditions of Consistency Theorem in Choi and Schervish (2007) [85] have been verified. We have showed that the posterior probability

\[ \sup_{(\theta_1, r, \theta_2) \in B} \Pi \{ S_{e\epsilon}^2 \mid e, \theta_1, r, \theta_2 \} \to 0 \quad [P^n_{0}], \quad \text{(B.2.4)} \]

where \( S_{e\epsilon}^2 = \{ p : \int |p(t) - p_0(t)| dt > \epsilon \} \) and \( e = (e_{t_1}, \ldots, e_{t_n})' \).

Finally, let us consider the goal of our theorem. By Fubini’s Theorem,

\[ \Pi \{ S_{e\epsilon}^2 \mid e \} = \int_{B} \Pi \{ S_{e\epsilon}^2 \mid e, \theta_1, r, \theta_2 \} \ d\Pi (\{(\theta_1, r, \theta_2) \mid e\}) 
= \int_{B_3} \int_{B_2} \int_{B_1} \Pi \{ S_{e\epsilon}^2 \mid e, \theta_1, r, \theta_2 \} \ d\Pi(\theta_1 \mid e) \ d\Pi(r \mid e) \ d\Pi(\theta_2 \mid e). \]

Since the supermum of the conditional probability in (B.2.4) converges to 0 in \( P^n_{0} \) probability, the marginal posterior probability converges to 0 in \( P^n_{0} \) probability regardless of the asymptotic distribution of \( \Pi(\theta_1 \mid e), \Pi(r \mid e) \) and \( \Pi(\theta_2 \mid e) \). This is
formalized as follows:

$$
\Pi\{S^2_e \mid e\} = \int_{B_3} \int_{B_2} \int_{B_1} \Pi\{S^2_e \mid e, \theta_1, r, \theta_2\} \, d\Pi(\theta_1 \mid e) \, d\Pi(r \mid e) \, d\Pi(\theta_2 \mid e)
$$

$$
\leq \sup_{(\theta_1, r, \theta_2) \in B} \Pi\{S^2_e \mid e, \theta_1, r, \theta_2\} \int_{\theta_1 \in B_1} d\Pi(\theta_1 \mid e) 
\times \int_{r \in B_2} d\Pi(r \mid e) \int_{\theta_2 \in B_3} d\Pi(\theta_2 \mid e) \to 0 \quad [P_0^\infty]
$$

So, (5.6.3) is proved. From Lemma 1, (5.6.2) is also proved.

**Appendix B.3**

**Theorem 1 (Choi and Schervish, 2007)** Let \( \{Z_i\}_{i=1}^\infty \) be independently distributed with densities \( \{f_i(\cdot; \theta)\}_{i=1}^\infty \), with respect to a common \( \sigma \)-finite measure, where the parameter \( \theta \) belongs to an abstract measurable space \( \Theta \). The densities \( f_i(\cdot; \theta) \) are assumed to be jointly measurable. Let \( \theta_0 \in \Theta \) and let \( P_{\theta_0} \) stand for the joint distribution of \( \{Z_i\}_{i=1}^\infty \) when \( \theta_0 \) is the true value of \( \theta \). Let \( \{U_n\}_{i=1}^\infty \) be a sequence of subsets of \( \Theta \). Let \( \theta \) have prior \( \Pi \) on \( \Theta \). Define

$$
\Lambda_i(\theta_0, \theta) = \log \frac{f_i(Z_i; \theta_0)}{f_i(Z_i; \theta)},
$$

$$
K_i(\theta_0, \theta) = \mathbb{E}_{\theta_0}(\Lambda_i(\theta_0, \theta)),
$$

$$
V_i(\theta_0, \theta) = \text{Var}_{\theta_0}(\Lambda_i(\theta_0, \theta)).
$$

**(A1) Prior positivity of neighborhoods.**

Suppose that there exists a set \( B \) with \( \Pi(B) > 0 \) such that

1. \( \sum_{i=1}^\infty \frac{V_i(\theta_0, \theta)}{\Delta^2} < \infty, \forall \theta \in B, \)

2. For all \( \epsilon > 0 \), \( \Pi(B \cap \{\theta : K_i(\theta_0, \theta) < \epsilon \text{ for all } i\}) > 0. \)
(A2) Existence of tests

Suppose that there exist test functions \( \{\Phi_n\}_{i=1}^{\infty} \), sets \( \{\Theta_n\}_{i=1}^{\infty} \) and constants \( C_1, C_2, c_1, c_2 > 0 \) such that

1. \( \sum_{i=1}^{\infty} \mathbb{E}_{\theta_0} \Phi_n < \infty, \)

2. \( \sup_{\theta \in U \cap \Theta_n} \mathbb{E}_{\theta} (1 - \Phi_n) \leq C_1 e^{-c_1 n}, \)

3. \( \Pi(\Theta_n^C) \leq C_2 e^{-c_2 n}. \)

Then

\[ \Pi(\theta \in U_n^C | Z_1, \ldots, Z_n) \rightarrow 0 \quad a.s. \quad [P_{\theta_0}]. \]
Bibliography


